ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ (Δ.Π.Μ.Σ.)
ΣΤΗ ΔΙΟΙΚΗΣΗ ΕΠΙΧΕΙΡΗΣΕΩΝ ΓΙΑ ΣΤΕΛΕΧΗ

MSc Διπλωματική Εργασία

ΑΝΑΛΥΤΙΚΗ ΔΕΔΟΜΕΝΩΝ ΣΤΙΣ ΕΠΙΧΕΙΡΗΣΕΙΣ: ΜΕΛΕΤΗ ΠΕΡΙΠΤΩΣΗΣ ΜΕ ΔΕΔΟΜΕΝΑ ΠΩΛΗΣΕΩΝ

υπό

ΙΩΑΝΝΗ Π. ΒΑΣΙΛΕΙΑΔΗ

Υποβλήθηκε ως προϋπόθεση για την εκπλήρωση των απαιτήσεων για την απόκτηση του μεταπτυχιακού διπλώματος στη Διοίκηση Επιχειρήσεων

Ιανουάριος 2019

Ευχαριστίες

Καταρχάς θα ήθελα να ευχαριστήσω τον επιβλέποντα της διπλωματικής μου εργασίας και καθηγητή μου, κ. Κωνσταντίνο Ταραμπάνη, που με εμπιστεύτηκε και μου ανέθεσε την εκπόνηση της παρούσας διπλωματικής εργασίας.

Επιπρόσθετα, θα ήθελα να ευχαριστήσω τον επίκουρο λέκτορα κ. Ευάγγελο Καλαμπόκη για τον χρόνο που μου αφιέρωσε προσφέροντάς μου, πολύτιμη βοήθεια και καθοριστικές συμβουλές. Επίσης είμαι ευγνώμων στα υπόλοιπα μέλη της εξεταστικής επιτροπής της διπλωματικής εργασίας μου, τον καθηγητή κ. Ανδρέα Γεωργίου και τον αναπληρωτή καθηγητή κ. Ιάσονα Παπαθανασίου για την προσεκτική ανάγνωση της εργασίας μου.

Ακόμα θα ήθελα να ευχαριστήσω τον βοηθό του εργαστηρίου, Συμεών Κοκοβίδη για την βοήθεια που μου προσέφερε στην εκτέλεση και επεξήγηση των αλγορίθμων στην πλατφόρμα του Kaggle.

Ένα πολύ μεγάλο ευχαριστώ σε όλους τους φίλους και τις φίλες μου, τους συμφοιτητές και τις συμφοιτήτριές μου για τις στιγμές χαλάρωσης και διασκέδασης που μου πρόσφεραν τα δύο αυτά χρόνια, αλλά και για την στήριξή τους κατά την εκπόνηση της παρούσας εργασίας.

Τέλος θα ήθελα να ευχαριστήσω τους γονείς μου και τον αδερφό μου, στους οποίους και αφιερώνω την παρούσα διπλωματική εργασία, για την ολόψυχη υποστήριξη τους όλα αυτά τα χρόνια.

Γιάννης Βασιλειάδης

Περίληψη

Το επιχειρηματικό περιβάλλον έχει αλλάξει δραματικά τις τελευταίες δεκαετίες με την ευρεία υιοθέτηση της τεχνολογίας και την εμφάνιση του ηλεκτρονικού επιχειρείν. Η αξία και η σημασία, της εκ των προτέρων πληροφόρησης αποτελεί τον ακρογωνιαίο λίθο του επιχειρηματικού σχεδιασμού και ένας σωστά καθορισμένος στόχος πρόβλεψης μπορεί να καθοδηγήσει αποτελεσματικότερα τις πωλήσεις.

Ως πρόβλεψη πωλήσεων ορίζεται, η μελλοντική προβολή της αναμενόμενης ζήτησης, λαμβάνοντας υπ'όψιν ένα δεδομένο σύνολο περιβαλλοντικών συνθηκών. Ο βασικός στόχος των επιχειρήσεων δεν είναι μόνο η ακρίβεια, αλλά η αποτελεσματική και αποδοτική κάλυψη της προβλεπόμενης ζήτησης. Η πρόβλεψη πωλήσεων είναι ιδιαίτερα χρήσιμη για τα περισσότερα τμήματα μιας επιχείρησης, όπως τα τμήματα μάρκετινγκ, πωλήσεων, χρηματοοικονομικών, παραγωγής και εφοδιασμού (logistic).

Η επιστήμη που ασχολείται με την πρόβλεψη πωλήσεων είναι η επιστήμη των δεδομένων (Data Science), η οποία έχει ως αντικείμενο την εξαγωγή σημαντικών πληροφοριών από αδόμητα ή δομημένα δεδομένα. Υπάρχει πληθώρα τεχνικών πρόβλεψης πωλήσεων, οι τρεις ευρείες κατηγορίες είναι οι χρονοσειρές (time series), η παλινδρόμηση (regression) και η κρίσιμη (judgemental).

Σε αυτή την εργασία θα ασχοληθούμε με τον διαγωνισμό στην πλατφόρμα Kaggle για την πρόβλεψη πωλήσεων για λογαριασμό της ROSSMANN, της δεύτερης μεγαλύτερης αλυσίδας φαρμακείων στη Γερμανία. Το Kaggle είναι μία διαδικτυακή πλατφόρμα, όπου επιχειρήσεις κοινοποιούν σύνολα δεδομένων και καλούν τους αναλυτές μέσω διαγωνισμών να τα διερευνήσουν ή να πραγματοποιήσουν κάποια πρόβλεψη.

Τα δοθέντα δεδομένα της ROSSMANN αφορούσαν 1115 καταστήματά της για δυόμιση περίπου έτη. Ερευνήθηκαν οι έτοιμες λύσεις που ήταν δημοσιοποιημένες στην πλατφόρμα και επιλέχθηκαν δύο, οι οποίες ήταν κατανοητές και αξιόλογες. Η μία λύση αφορά την διερεύνηση των δεδομένων, ενώ η άλλη πραγματοποιεί την πρόβλεψη πωλήσεων χρησιμοποιώντας τη μέθοδο χρονοσειρών Prophet και τη μέθοδο παλινδρόμησης XGBoost.

Σκοπός της παρούσας διπλωματικής είναι η κατανόηση της χρησιμότητας της πρόβλεψης πωλήσεων και των μεθόδων που χρησιμοποιούνται σε αυτές.

Πίνακας Περιεχομένων

Κατάλογος πινάκων

Κατάλογος Σχημάτων

1. Εισαγωγή

1.1. Δήλωση προβλήματος

Στη σημερινή ανταγωνιστική παγκόσμια οικονομία, οι επιχειρήσεις πρέπει να προσαρμόζονται διαρκώς στις συνεχώς μεταβαλλόμενες αγορές. Επομένως, η πρόβλεψη των μελλοντικών γεγονότων στην αγορά, είναι ζωτικής σημασίας για την επιτυχία των επιχειρηματικών δραστηριοτήτων. Το επιστημονικό πεδίο που πραγματοποιεί προβλέψεις χρησιμοποιώντας σύνολα δεδομένων είναι το Data Science, το οποίο έχει ως αντικείμενο την εξαγωγή σημαντικών πληροφοριών από αδόμητα ή δομημένα δεδομένα.

Οι επιχειρήσεις λιανικού εμπορίου είναι αναγκασμένες να χρησιμοποιούν αποτελεσματικά τους πόρους τους και να λαμβάνουν σωστές στρατηγικές αποφάσεις για το μέλλον, προκειμένου να επιβιώσουν και να αυξήσουν τα έσοδά τους, σε ένα τόσο ανταγωνιστικό περιβάλλον. Δεδομένου ότι όλες οι προβλέψεις εμπεριέχουν κάποιο ποσοστό αβεβαιότητας, οι επιχειρήσεις πρέπει να προβούν σε εκτιμήσεις με στόχο την αύξηση της ακρίβειας. Οι επιχειρήσεις πρέπει να πραγματοποιούν προβλέψεις που αφορούν πολλές μεταβλητές, όπως απαιτήσεις πρώτων υλών, βέλτιστα επίπεδα αποθεμάτων, απαιτήσεις δανεισμού και απαιτήσεις προσωπικού. Ωστόσο, για να εκτιμηθεί οποιαδήποτε από αυτές, πρέπει πρώτα να προβλεφθεί το επίπεδο ζήτησης της αγοράς και κατά συνέπεια, οι προσδοκώμενες πωλήσεις της εταιρείας. Έτσι, οι προβλέψεις της ζήτησης της αγοράς αποτελούν απαραίτητο στοιχείο για όλες τις άλλες εκτιμήσεις που απαιτούνται από κάποιον οργανισμό. Οι ακριβείς προβλέψεις επιτρέπουν την κατάλληλη διαχείριση των δραστηριοτήτων της εταιρείας (όπως το επίπεδο παραγωγής, τη χρηματοδότηση, την έρευνα και την ανάπτυξη, τις προμήθειες και το μάρκετινγκ) και διευκολύνουν την επίτευξη των στόχων τους (Mentzer and Bienstock (1998)). Επίσης, η πρόβλεψη των πωλήσεων έχει ιδιαίτερη σημασία για τις εταιρείες που λαμβάνουν στρατηγικές αποφάσεις σχετικά με τις μελλοντικές επενδύσεις τους. Για παράδειγμα, η πρόβλεψη πωλήσεων σε συνδυασμό με το περιθώριο κέρδους χρησιμοποιούνται για την εκτίμηση του μελλοντικού εισοδήματος της εταιρείας και μαζί με τις προβλέψεις του κύκλου εργασιών της, γίνεται αξιολόγηση των μελλοντικών περιουσιακών στοιχείων της (Curtis et al. (2014)).

Η λιανική πώληση μπορεί να ορίζεται από το σύνολο της εμπορίας αγαθών και υπηρεσιών απευθείας στους τελικούς χρήστες, ως εκ τούτου όμως μπορεί να θεωρηθεί και ως μια γέφυρα μεταξύ του παραγωγού και του τελικού καταναλωτή. Η ικανότητα

των διευθυντικών στελεχών να προβλέπουν τις μελλοντικές πωλήσεις συσχετίζεται με την αυξημένη ικανοποίηση του πελάτη, τη μείωση των αποθεμάτων, τα αυξημένα έσοδα από τις πωλήσεις και τα αποδοτικότερα και αποτελεσματικότερα σχέδια παραγωγής (Chen and Ou (2011)). Οι Barksdale και Hilliard (1975) μελέτησαν σε καταστήματα λιανικής πώλησης, τη σχέση μεταξύ των αποθεμάτων και των πωλήσεων και κατέληξαν στο συμπέρασμα, ότι η επιτυχής διαχείριση των αποθεμάτων εξαρτάται κυρίως από την ακριβή πρόβλεψη των λιανικών πωλήσεων. Η ακριβής πρόβλεψη των μελλοντικών λιανικών πωλήσεων, μπορεί να συμβάλει στην αύξηση της αποδοτικότητας και της αποτελεσματικότητας των εργασιών που πραγματοποιούνται στις λιανικές επιχειρήσεις και στις αλυσίδες εφοδιασμού. Επομένως, οι προβλέψεις διαδραματίζουν καθοριστικό ρόλο στη διαχείριση των επιχειρήσεων και στον στρατηγικό ους σχεδιασμό. Οι αποφάσεις διαχείρισης που λαμβάνονται σε όλα τα επίπεδα μιας επιχείρησης, συνδέονται άμεσα ή έμμεσα με τις προβλέψεις που έχει πραγματοποιήσει. Χωρίς χρήσιμες προβλέψεις, οι δραστηριότητες προγραμματισμού και ελέγχου δεν μπορούν να υλοποιηθούν αποτελεσματικά. Επιπρόσθετα, μία λανθασμένη πρόβλεψη μπορεί να επηρεάσει αρνητικά την δυναμικότητα της επιχείρησης για την επίτευξη των στόχων της, καθώς μπορεί να την οδηγήσει σε προβλήματα, όπως με μία λανθασμένη πρόβλεψη μεγάλης ζήτησης, θα προκαλέσει διατήρηση μεγάλων αποθεμάτων και κατά συνέπεια αύξηση του κόστους, ή αντιθέτως με μία λανθασμένη πρόβλεψη μικρής ζήτησης, θα επιφέρει μικρά αποθέματα, άρα αδυναμία ανταπόκρισης στη πραγματική ζήτηση, η οποία μπορεί με τη σειρά της να οδηγήσει σε απώλεια του μεριδίου αγοράς (Agrawal and Schorling (1997)).

Σε πλήρη συμφωνία με τα προαναφερθέντα, η μελέτη περίπτωσης που θα εξεταστεί σε αυτή την εργασία, αφορά την πρόβλεψη πωλήσεων στην αλυσίδα καταστημάτων της ROSSMANN. Η εταιρεία, ως μεγάλη αλυσίδα καταστημάτων, με 2110 σημεία πώλησης στη Γερμανία και άλλα 1680 σε πέντε χώρες της Ευρώπης, αντιμετωπίζει σημαντικό πρόβλημα στη διαχείριση της εφοδιαστικής αλυσίδας και του προγραμματισμού της απασχόλησης των εργαζομένων.

Αυτό είχε ως αποτέλεσμα οι διευθυντές των καταστημάτων να αφιερώνουν τον περισσότερο χρόνο τους σε αυτά τα προβλήματα και να μην αφοσιώνονται στο σημαντικότερο κομμάτι της επιχείρησης, τους πελάτες. Αυτό είχε ως επίπτωση, η ROSSMANN να διοργανώσει έναν διαγωνισμό, με χρηματικό έπαθλο, για τους τρεις πρώτους νικητές, ώστε να βρει την καλύτερη δυνατή πρόβλεψη, η οποία θα της δώσει τη δυνατότητα να γνωρίζει εκ των προτέρων τις ημερήσιες πωλήσεις των καταστημάτων της και κατά συνέπεια θα μπορεί να ελέγξει καλύτερα την εφοδιαστική

αλυσίδα και να προγραμματίσει τα ωράρια των εργαζομένων. Ο διαγωνισμός αυτός διοργανώθηκε στο Kaggle, το οποίο είναι μία διαδικτυακή πλατφόρμα όπου επιχειρήσεις κοινοποιούν σύνολα δεδομένων και καλούν τους αναλυτές μέσω διαγωνισμών να τα διερευνήσουν ή να πραγματοποιήσουν κάποια πρόβλεψη.

1.2. Ο σκοπός και ο στόχος της έρευνας

Στην παρούσα εργασία αρχικά θα αναπτύξουμε την σημασία, της εκ των προτέρων πληροφόρησης στον επιχειρηματικό σχεδιασμό, καθώς και τα πεδία εφαρμογής της πρόβλεψης των πωλήσεων σε όλα τα τμήματα των επιχειρήσεων. Επίσης θα παραθέσουμε τις κύριες κατηγορίες των τεχνικών πρόβλεψης και θα γίνει μία σύντομη περιγραφή τους. Παρουσιάζοντας στο κείμενο διάφορες μελέτες περίπτωσης, θα μας δοθεί η ευκαιρία να κατανοήσουμε καλύτερα τα προαναφερθέντα μέσα από πραγματικά παραδείγματα.

Στη συνέχεια θα μελετήσουμε την περίπτωση της ROSSMANN, της δεύτερης μεγαλύτερης αλυσίδας φαρμακείων στη Γερμανία. Η οποία μέσω της πλατφόρμας Kaggle, διοργάνωσε έναν διαγωνισμό, στον οποίο ζητούσε την πρόβλεψη ημερήσιων πωλήσεων για έξι συνεχείς εβδομάδες. Η ROSSMANN παρείχε τα δεδομένα δυόμιση περίπου ετών από 1115 καταστήματά της στη Γερμανία. Στα αρχεία δεδομένων που δόθηκαν στον διαγωνισμό, συμπεριλαμβανόταν τόσο ενδογενή δεδομένα, όσο και εξωγενή.

Στον διαγωνισμό έγιναν 3003 υποβολές prediction για το έπαθλο, των οποίων οι λύσεις δεν δημοσιεύτηκαν. Ενώ προς κοινή χρήση δόθηκαν πολλά kernels με λύσεις, εκ των οποίων 75 ήταν σε γλώσσα προγραμματισμού R και 68 σε γλώσσα Python και είχαν τουλάχιστον μία ψήφο κοινού. Υπήρχαν και εκατοντάδες άλλα kernels, τα οποία δεν είχαν καμία ψήφο κοινού, οπότε δεν δόθηκε ιδιαίτερη σημασία. Από όλες αυτές τις λύσεις, μετά από έρευνα διάλεξα τις δύο με τις περισσότερες ψήφους κοινού, θεωρώντας αυτές ως τις πιο ενδιαφέρουσες και κατανοητές. Στη συνέχεια τις παρουσίασα αναλυτικά. Η μία λύση ήταν του Christian Thiele, σε γλώσσα προγραμματισμού R και ως αντικείμενο είχε τη διερεύνηση των δεδομένων χρησιμοποιώντας πίνακες και διάφορες οπτικοποιήσεις. Η δεύτερη ήταν της Elena Petrova, η οποία χρησιμοποίησε την γλώσσα Python και αρχικά παρουσίασε μία μικρή διερεύνηση και επεξεργασία των δεδομένων και έπειτα πραγματοποίησε την πρόβλεψη πωλήσεων με δύο μεθόδους. Η πρώτη μέθοδος ήταν η Prophet, η οποία είναι μία

μέθοδος ανάλυσης χρονοσειρών, ενώ η δεύτερη ήταν η Ακραία Βαθμιδωτή Ενίσχυση (Extreme Gradient Boosting – XGBoost), η οποία είναι μέθοδος παλινδρόμησης.

Ο βασικός στόχος της εργασίας είναι να κατανοήσουμε την όλη λειτουργία του Data Science και ποια είναι η χρήση του σε επιχειρηματικά προβλήματα. Επίσης θα προσπαθήσουμε να κατανοήσουμε, ποιος είναι ο ρόλος της διερεύνησης των δεδομένων, της μηχανικής μάθησης και των μεθόδων πρόβλεψης, χρησιμοποιώντας μεγάλες βάσεις δεδομένων που προέρχονται από επιχειρήσεις, ώστε να βελτιώσουμε τη διαδικασία λήψης αποφάσεων αυτών των επιχειρήσεων. Για το λόγο αυτό, επικεντρωθήκαμε σε μία μόνο μελέτη περίπτωσης, αυτή της ROSSMANN, και την αναλύσαμε διεξοδικά. Θέλαμε να εστιάσουμε πρακτικά στη χρήση των τεχνολογιών του Data Science και όχι θεωρητικά, ώστε να έχουμε μία εμπεριστατωμένη άποψη.

1.3. Δομή της εργασίας

Στο δεύτερο κεφάλαιο, παρουσιάζεται βιβλιογραφική αναφορά για τη πρόβλεψη πωλήσεων, στην οποία αναφέρεται ο τρόπος με τον οποίο γίνεται η διαχείριση των προβλέψεων και ο ρόλος τους στον σχεδιασμό των πωλήσεων και λειτουργίας. Στη συνέχεια επισημαίνουμε την ιδιαίτερη χρησιμότητα της πρόβλεψης πωλήσεων για τα περισσότερα τμήματα μιας επιχείρησης, όπως τα τμήματα μάρκετινγκ, πωλήσεων, χρηματοοικονομικών, παραγωγής και εφοδιασμού (logistic) και με συγκεκριμένες μελέτες περίπτωσης παραθέτουμε πραγματικές εφαρμογές. Κατόπιν αναφέρουμε τα εργαλεία διαχείρισης των προβλέψεων πωλήσεων, στα οποία ανήκουν οι διάφορες τεχνικές που χρησιμοποιούνται, η προσέγγιση των προβλέψεων που γίνεται από τις επιχειρήσεις και οι μετρήσεις απόδοσης και ακρίβειας των προβλέψεων. Στο τελευταίο τμήμα του κεφαλαίου, παραθέτονται κάποια καίρια ερωτήματα που γεννώνται από τις προβλέψεις πωλήσεων και αναφέρεται ο λόγος κατά τον οποίο ο προγραμματισμός των προβλέψεων είναι μία επαναληπτική διαδικασία.

Στο τρίτο κεφάλαιο περιγράφουμε τη μεθοδολογία που θα ακολουθήσουμε κατά την αντιμετώπιση της μελέτης περίπτωσης της ROSSMANN, ενώ στο τέταρτο κεφάλαιο ερευνάμε την εταιρεία της ROSSMANN, παρουσιάζοντας την ιστορία της, τα προϊόντα της, το ηλεκτρονικό της κατάστημα, τα πληροφοριακά συστήματα που χρησιμοποιεί και την ανάγκη που εμφανίστηκε για την πρόβλεψη πωλήσεων.

Στο πέμπτο κεφάλαιο θα διερευνήσουμε τα δεδομένα της εταιρείας, περιγράφοντας τις μεταβλητές, το είδος των αρχείων και το μέγεθός τους, έπειτα θα πραγματοποιηθεί μία ανάλυση και οπτικοποίηση των δεδομένων. Τέλος θα γίνει επεξεργασία των

δεδομένων και θα διερευνηθούν οι τάσεις και τα μοτίβα τους και θα αναλυθούν οι συσχετίσεις μεταξύ των μεταβλητών και οι χρονοσειρές.

Στο έκτο κεφάλαιο θα παρουσιαστούν δύο μοντέλα για την πρόβλεψη των πωλήσεων της ROSSMANN. Το πρώτο μοντέλο χρησιμοποιεί μία μέθοδο ανάλυσης χρονοσειρών και ονομάζεται Prophet, ενώ το δεύτερο ονομάζεται XGBoost και είναι μία μέθοδος παλινδρόμησης.

Τέλος, ολοκληρώνονται τα κεφάλαια, με το έβδομο κεφάλαιο, όπου παραθέτουμε τα συμπεράσματα αυτής της διπλωματικής εργασίας.

2. Βιβλιογραφική Αναφορά

2.1. Εισαγωγή

Το επιχειρηματικό περιβάλλον έχει αλλάξει δραματικά τις τρεις τελευταίες δεκαετίες με την αυξανόμενη παγκοσμιοποίηση, την ευρεία υιοθέτηση της τεχνολογίας των πληροφοριών και την εμφάνιση του ηλεκτρονικού επιχειρείν. Οι παράγοντες που απορρέουν από αυτές τις περιβαλλοντικές αλλαγές, όπως ο χρονικός ανταγωνισμός (Golicic et al. (2002)) και ο πολλαπλασιασμός των προϊόντων (Bayus and Putsis (1999); Parker (2002)), έχουν άμεσο αντίκτυπο στις πρακτικές και τις διαδικασίες πρόβλεψης (Moon et al. (2003)), καθιστώντας έτσι σημαντικό να εξεταστεί το πώς έχουν αλλάξει οι πρακτικές διαχείρισης των προβλέψεων από τις μελέτες που πραγματοποιήθηκαν στη δεκαετία του 1980 (Mentzer and Cox (1984)) και 1990s (Mentzer and Kahn (1995)) έως σήμερα (McCarthy (2006)).

Η πρόβλεψη των πωλήσεων διαδραματίζει εξέχοντα ρόλο στον επιχειρηματικό σχεδιασμό και την επιχειρηματική στρατηγική. Η αξία και η σημασία της εκ των προτέρων πληροφόρησης αποτελεί τον ακρογωνιαίο λίθο της προγραμματικής δραστηριότητας και ένας σωστά καθορισμένος στόχος πρόβλεψης μπορεί να καθοδηγήσει αποτελεσματικότερα τις πωλήσεις. Μια πρόβλεψη εξαρτάται συνήθως από πολλούς παράγοντες όπως το χαρακτηριστικό γνώρισμα του προϊόντος, ο περιορισμός της εφοδιαστικής αλυσίδας, η ζήτηση της αγοράς, το μερίδιο αγοράς, η στρατηγική προώθησης, ο ανταγωνισμός, η μακροοικονομική κατάσταση και άλλα. Ωστόσο, τα περισσότερα από αυτά τα δεδομένα είναι δύσκολο ή και αδύνατο να συγκεντρωθούν (Xu and Sharma (2017)).

2.2. Διαχείριση των προβλέψεων πωλήσεων

Η διαχείριση των προβλέψεων πωλήσεων αφορά τη διαχείριση τους εντός ενός οργανισμού. Παρόλο που συνήθως ονομάζεται πρόβλεψη πωλήσεων, ουσιαστικά αυτό που πραγματικά προσπαθούμε να προβλέψουμε είναι η ζήτηση. Δηλαδή θέλουμε να γνωρίζουμε τι απαιτήσεις θα έχουν οι πελάτες μας, ώστε να μπορούμε να προγραμματίσουμε την επίτευξη πωλήσεων σε αυτό το επίπεδο ή κοντά σε αυτό.

Η πρόβλεψη των πωλήσεων περιλαμβάνει την ορθή χρήση διαφόρων ποιοτικών και ποσοτικών τεχνικών στο πλαίσιο των εταιρικών πληροφοριακών συστημάτων, για την κάλυψη των αναγκών που έχουν οι χρήστες των προβλέψεων πωλήσεων και τη

διαχείριση αυτών των διεργασιών. Για να διαχειριστούμε αυτές τις πολυδιάστατες πτυχές, πρέπει να κατανοήσουμε την κάθε μία τεχνική και τις δομές διαχείρισης στις οποίες εκτελούνται οι προβλέψεις πωλήσεων. Ωστόσο, προτού προχωρήσουμε περαιτέρω, θα πρέπει να κατανοήσουμε ακριβώς τι εννοούμε με τον όρο πρόβλεψη πωλήσεων και τη σύγχυση που υπάρχει με τον προγραμματισμό.

Μια πρόβλεψη πωλήσεων ορίζεται, ως η μελλοντική προβολή της αναμενόμενης ζήτησης, λαμβάνοντας υπ'όψιν ένα δεδομένο σύνολο περιβαλλοντικών συνθηκών. Αυτό πρέπει να διακρίνεται από τα επιχειρησιακά σχέδια, τα οποία θα ορίσουμε ως ένα σύνολο συγκεκριμένων δράσεων διαχείρισης που πρέπει να ληφθούν για την επίτευξη της πρόβλεψης των πωλήσεων. Παραδείγματα επιχειρησιακών σχεδίων είναι τα σχέδια παραγωγής, τα σχέδια προμηθειών και τα σχέδια διανομής. Τόσο η πρόβλεψη πωλήσεων όσο και τα επιχειρησιακά σχέδια θα πρέπει να διακριθούν από τον στόχο πωλήσεων, τον οποίο θα ορίσουμε ως επίτευγμα πωλήσεων που δημιουργήθηκε για να δώσει κίνητρα στο προσωπικό πωλήσεων και μάρκετινγκ.

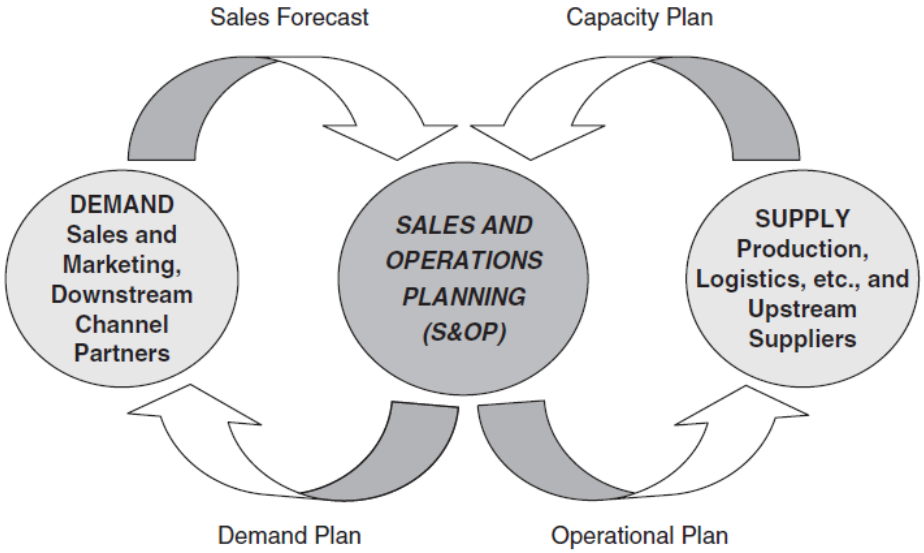
Ο ορισμός της πρόβλεψης πωλήσεων δεν προσδιορίζει την τεχνική (ποσοτική ή ποιοτική), δεν προσδιορίζει ποιος αναπτύσσει την πρόβλεψη στην εταιρεία, ούτε περιλαμβάνει σχέδια διαχείρισης. Ο λόγος για αυτό είναι ότι πολλές εταιρείες συγχέουν τις λειτουργίες της πρόβλεψης, του σχεδιασμού και του καθορισμού στόχων. Τα επιχειρησιακά σχέδια για το επίπεδο των πωλήσεων που πρέπει να επιτευχθούν πρέπει να βασίζονται στην πρόβλεψη της ζήτησης, αλλά οι δύο λειτουργίες διαχείρισης θα πρέπει να διατηρούνται ξεχωριστά. Ομοίως, ο καθορισμός στόχων πρέπει να γίνει με ρεαλιστική εκτίμηση της αναμενόμενης μελλοντικής ζήτησης και αυτή η εκτίμηση προέρχεται από τις προβλέψεις πωλήσεων. Με άλλα λόγια, οι λειτουργίες του σχεδιασμού και του καθορισμού στόχων πρέπει να ενημερώνονται από τις προβλέψεις της ζήτησης, αλλά δεν πρέπει να συγχέονται με τις προβλέψεις πωλήσεων.

Αυτοί οι ορισμοί υπονοούν διαφορετικά μέτρα απόδοσης για τις προβλέψεις πωλήσεων από αυτά των επιχειρησιακών σχεδίων. Επειδή ο σκοπός της πρόβλεψης των πωλήσεων είναι να γίνουν προβολές της ζήτησης δεδομένου ενός συνόλου συγκεκριμένων περιβαλλοντικών παραδοχών, ένα από τα βασικά μέτρα της απόδοσης των προβλέψεων πωλήσεων είναι η ακρίβεια της πρόβλεψης και μία από τις βασικές μεθόδους για να εξηγηθούν οι αποκλίσεις στην ακρίβεια είναι το πώς το περιβάλλον διαφέρει από το καθορισμένο. Αυτή η εξήγηση δεν αποσκοπεί στην δικαιολόγηση της ανακριβής πρόβλεψης, αντίθετα, πρόκειται να μας βοηθήσει να κατανοήσουμε το επιχειρηματικό περιβάλλον και να μπορέσουμε να προβλέψουμε με μεγαλύτερη ακρίβεια στο μέλλον.

Αντίθετα, ο στόχος των επιχειρησιακών σχεδίων δεν είναι η ακρίβεια, αλλά η αποτελεσματική και αποδοτική κάλυψη της προβλεπόμενης ζήτησης. Επιπλέον, ενώ οι προβλέψεις πρέπει να είναι ακριβείς, οι στόχοι πρέπει να πληρούνται ή να υπερβαίνονται. Ένα λάθος που έκαναν πολλές εταιρείες είναι να μπερδέψουν τις προβλέψεις πωλήσεων, όπου ο στόχος είναι η ακρίβεια, με τον στόχο των πωλήσεων, όπου ο στόχος είναι τουλάχιστον να καλύψουν και, ενδεχομένως, να υπερβούν τον στόχο ή την ποσόστωση. Με άλλα λόγια, οι εταιρείες δεν πρέπει ποτέ να συγχέουν τις προβλέψεις με την στρατηγική κινήτρων της επιχείρησης (Mentzer and Moon (2005)).

2.3. Ο ρόλος της πρόβλεψης πωλήσεων στον σχεδιασμό πωλήσεων και λειτουργίας (S&OP)

Σε πολλές εταιρείες, η πρόβλεψη των πωλήσεων αποτελεί αναπόσπαστο μέρος μιας κρίσιμης διαδικασίας για την αντιστοίχιση της ζήτησης και της προσφοράς, η οποία αναφέρεται μερικές φορές ως Σχεδιασμός Πωλήσεων και Λειτουργιών (Sales and Operations Planning - S&OP). Το σχήμα 1 προσφέρει μια απλοποιημένη εικόνα του τρόπου με τον οποίο η πρόβλεψη των πωλήσεων συμβάλλει στη διαδικασία S&OP. Όπως φαίνεται στο Σχήμα 1, μια επιχείρηση μπορεί να θεωρηθεί ότι αποτελείται από δύο βασικές λειτουργίες: μια λειτουργία ζήτησης και μια λειτουργία παροχής. Η ζήτηση είναι ευθύνη των πωλήσεων και του μάρκετινγκ. Σε πολλές εταιρείες, η οργάνωση των πωλήσεων είναι υπεύθυνη για τη δημιουργία και τη διατήρηση της ζήτησης από πολλούς τελικούς πελάτες λιανικής ή από συνεργάτες χονδρικής ή λιανικής πώλησης. Το μάρκετινγκ είναι συνήθως υπεύθυνο για τη δημιουργία και τη διατήρηση της ζήτησης από τους τελικούς καταναλωτές. Το τμήμα ανεφοδιασμού αποτελεί ευθύνη πολλών λειτουργιών, όπως η κατασκευή, ο εφοδιασμός ή η διανομή, οι ανθρώπινοι πόροι και η χρηματοδότηση. Είναι επίσης υπεύθυνο για τους προμηθευτές, οι οποίοι πρέπει να παρέχουν πρώτες ύλες, εξαρτήματα και συσκευασίες. Η διαδικασία S&OP εμπεριέχει ένα σύμπλεγμα πληροφοριών, όπου οι πληροφορίες μπορούν να ρέουν μεταξύ της πλευράς ζήτησης και της πλευράς προσφοράς μιας επιχείρησης.



Σχήμα 1: Σύμπλεγμα S&OP

Όπως φαίνεται στο Σχήμα 1, η κρίσιμη συμβολή στη διαδικασία S&OP είναι η πρόβλεψη πωλήσεων, η οποία, όπως ορίζεται παραπάνω, προβλέπει την μελλοντική προβολή της αναμενόμενης ζήτησης. Η πρόβλεψη των πωλήσεων θα πρέπει να προέρχεται από την πλευρά της ζήτησης της επιχείρησης, διότι η πλευρά της ζήτησης της επιχείρησης (δηλ. πωλήσεις και μαρκετινγκ) είναι υπεύθυνη για τη δημιουργία ζήτησης και πρέπει να έχει την καλύτερη άποψη όσον αφορά τη μελλοντική ζήτηση. Εκτός από τις προβλέψεις πωλήσεων, που προέρχονται από την πλευρά της ζήτησης της εταιρείας, μια άλλη κρίσιμη συμβολή στη διαδικασία S&OP είναι το σχέδιο ανεφοδιαστικής ικανότητας. Το σχέδιο ανεφοδιαστικής ικανότητας είναι μια πρόβλεψη στο μέλλον σχετικά με τις δυνατότητες εφοδιασμού, δεδομένης μιας σειράς περιβαλλοντικών παραδοχών. Αυτή η εισροή παρέχεται από την πλευρά της προσφοράς της επιχείρησης και καταγράφει τόσο μακροπρόθεσμες όσο και βραχυπρόθεσμες δυνατότητες εφοδιασμού. Η διαδικασία που συμβαίνει μέσα στη διαδικασία S&OP, είναι η αντιστοίχιση των μελλοντικών προβολών της ζήτησης (δηλ. της πρόβλεψης των πωλήσεων) με μελλοντικές προβλέψεις εφοδιασμού (δηλαδή το σχέδιο ανεφοδιαστικής ικανότητας).

Από τη διαδικασία S&OP έρχονται δύο κρίσιμα σχέδια, το επιχειρησιακό σχέδιο και το σχέδιο ζήτησης. Όπως αναφέρθηκε παραπάνω, το επιχειρησιακό σχέδιο αποτελείται από σχέδια κατασκευής, σχέδια προμηθειών, σχέδια διανομής και σχέδια ανθρώπινου δυναμικού. Αυτά τα διάφορα επιχειρησιακά σχέδια μπορεί να έχουν βραχυπρόθεσμο χαρακτήρα, όπως ένα μηνιαίο χρονοδιάγραμμα παραγωγής. Μπορούν να έχουν και μακροπρόθεσμο χαρακτήρα, όπως οι εκτεταμένες συμβάσεις για πρώτες ύλες ή ακόμη και σχέδια για επέκταση της παραγωγικής ικανότητας. Το άλλο κρίσιμο

σχέδιο που προκύπτει από τη διαδικασία S&OP είναι το σχέδιο ζήτησης, το οποίο περιλαμβάνει τα σχέδια πωλήσεων και μάρκετινγκ σχετικά με το τι πρέπει να πωληθεί και να διατεθεί στο εμπόριο και πότε, λόγω των δυνατοτήτων προσφοράς της επιχείρησης. Όπως αναφέρθηκε παραπάνω, τα σχέδια ζήτησης ενδέχεται να περιλαμβάνουν την καταστολή της ζήτησης προϊόντων ή υπηρεσιών που έχουν περιορισμένη παραγωγική ικανότητα ή τη μετατόπιση της ζήτησης από προϊόντα χαμηλού περιθωρίου κέρδους σε προϊόντα υψηλού περιθωρίου.

Είναι σημαντικό, να κατανοήσουμε τον κρίσιμο ρόλο που διαδραματίζουν οι προβλέψεις πωλήσεων στις συνολικές δραστηριότητες σχεδιασμού της επιχείρησης. Χωρίς ακριβείς και αξιόπιστες εκτιμήσεις της μελλοντικής ζήτησης, είναι αδύνατο για τους οργανισμούς να διαχειρίζονται αποτελεσματικά τις αλυσίδες εφοδιασμού τους (Mentzer and Moon (2005)).

2.4. Η χρησιμότητα της πρόβλεψης πωλήσεων

Αν μπορούμε απλά να ορίσουμε έναν στόχο πωλήσεων και να περιμένουμε το μάρκετινγκ και τις πωλήσεις να τον ξεπεράσουν, γιατί ακόμη χρειαζόμαστε την πρόβλεψη πωλήσεων; Αυτή είναι μια ερώτηση που πραγματοποιούν πολλοί διευθυντές και συχνά την απαντούν λανθασμένα (δηλαδή ότι δεν χρειαζόμαστε πρόβλεψη), για να διαψευσθούν στο μέλλον.

Η σωστή απάντηση είναι ότι κάθε φορά που αναπτύσσουμε ένα σχέδιο οποιουδήποτε είδους, κάνουμε πρώτα μια πρόβλεψη. Αυτό ισχύει για τους ανθρώπους στην καθημερινότητα, καθώς και για τις κερδοσκοπικές και μη κερδοσκοπικές εταιρείες, τις κυβερνητικές οργανώσεις και στην πραγματικότητα για κάθε οντότητα που κάνει ένα σχέδιο. Μπορεί να είναι τόσο απλό, όσο ο σχεδιασμός του τι θα φορέσουμε αύριο. Όταν αποφασίζουμε να διαλέξουμε ένα ζεστό παντελόνι και ένα πουλόβερ για την επόμενη μέρα, προβλέπουμε ότι ο καιρός θα είναι ψυχρός. Αν προσθέσουμε μια ομπρέλα στο σύνολό μας, προβλέπουμε βροχή. Η ετοιμασία της ενδυμασίας, βασίστηκε στην πρόβλεψη, είτε το σκεφτήκαμε συνειδητά, είτε όχι.

Αυτό δεν είναι πολύ διαφορετικό από μια εταιρεία που κάνει οικονομικά σχέδια με βάση τις αναμενόμενες πωλήσεις και το κόστος της κάλυψης αυτών των πωλήσεων. Το μυστικό είναι να μην πιαστεί κανείς στην παγίδα της "ακούσιας πρόβλεψης πωλήσεων." Η ακούσια πρόβλεψη πωλήσεων γίνεται όταν είμαστε τόσο πεποισμένοι να αναπτύξουμε ένα σχέδιο, που απλά υποθέτουμε ποιες θα είναι οι πωλήσεις, αντί να

δώσουμε βάση σε συγκεντρωμένη σκέψη και ανάλυση των συνθηκών της αγοράς που θα είναι απαραίτητες για τη δημιουργία αυτού του επιπέδου πωλήσεων.

Ένα σημαντικό παράδειγμα μιας τέτοιας ακούσιας πρόβλεψης προήλθε από έναν κατασκευαστή της βιομηχανίας τροφίμων. Ο ιδιοκτήτης της εταιρείας είχε ένα σχέδιο πωλήσεων το οποίο απαιτούσε αύξηση πωλήσεων κατά 5% για το επόμενο έτος. Ωστόσο, ο κλάδος αυτός στη χώρα δεν αυξανόταν και κάθε απόπειρα απόκτησης του μεριδίου αγοράς από τον ανταγωνισμό μπορούσε να καλυφθεί μόνο μέσα από μεγαλύτερες διαφημιστικές δαπάνες, αλλά χωρίς μετατόπιση του μεριδίου αγοράς. Ενημερώθηκε ο ιδιοκτήτης από τους ερευνητές ότι δεν μπορούν να αυξηθούν οι πωλήσεις, τη στιγμή που δεν αλλάζουν το μέγεθος της βιομηχανίας και το μερίδιο αγοράς. Δεν χρειάζεται κανείς μαθηματικές γνώσεις για να καταλάβει ότι αυτό δεν πρόκειται να λειτουργήσει. Η απάντηση του ιδιοκτήτη ήταν ότι η διοίκηση θα έπρεπε απλώς να παρακινήσει όλους να εργαστούν σκληρότερα για να επιτύχουν το (μαθηματικά αδύνατο) σχέδιο. Φυσικά, είναι προφανές τι συνέβη. Κανένα ποσό κινήτρων δεν μπορεί να ξεπεράσει μια αδύνατη κατάσταση και το σχέδιο πωλήσεων φυσικά δεν επιτεύχθηκε. Δεν επιτεύχθηκε επειδή βασίστηκε σε μια ακούσια και άγνωστη πρόβλεψη. Αυτό είναι επίσης ένα κλασικό παράδειγμα διαχείρισης που συγχέει την πρόβλεψη, τον προγραμματισμό και τον καθορισμό στόχων. Στην περίπτωση αυτή, καμία λογική πρόβλεψη δεν θα προέβλεπε αύξηση των πωλήσεων κατά 5%.

Ας δούμε ένα ακόμη παράδειγμα. Ένας μεγάλος διανομέας τροφίμων σε εστιατόρια αναπτύσσει ένα περίτεχνο ετήσιο σχέδιο κέρδους. Εκατοντάδες εργατοωρών εμπλέκονται στην ανάπτυξη αυτού του σχεδίου, και τα στελέχη ενημερώνουν ότι χρειάζονται κέρδη για να αναπτυχθούν το επόμενο έτος κατά 6%. Έπειτα εξετάζουν πόσες πρέπει να είναι οι πωλήσεις για να επιτευχθεί αυτός ο στόχος. Παρατηρήστε ότι ο όρος "στόχος" ειπώθηκε σε αυτό το απόσπασμα. Αυτά τα στελέχη θα έπρεπε να ξεκινήσουν από την διερεύνηση των συνθηκών της αγοράς και του περιβάλλοντος που αντιμετωπίζει η εταιρεία, κατά τον ορίζοντα σχεδιασμού και ποια επίπεδα πωλήσεων θα μπορούσαν να αναμένονται με βάση αυτές τις συνθήκες. Το σχέδιο, στη συνέχεια, καθίσταται καθοριστικό των προσπαθειών μάρκετινγκ και πωλήσεων που είναι απαραίτητες για την επίτευξη και υπέρβαση αυτών των προβλέψεων σε επίπεδο απαραίτητο για την επίτευξη του σχεδίου κέρδους. Το σχέδιο δεν μπορεί να οδηγήσει την πρόβλεψη, πρέπει να υπάρχει άλλος τρόπος.

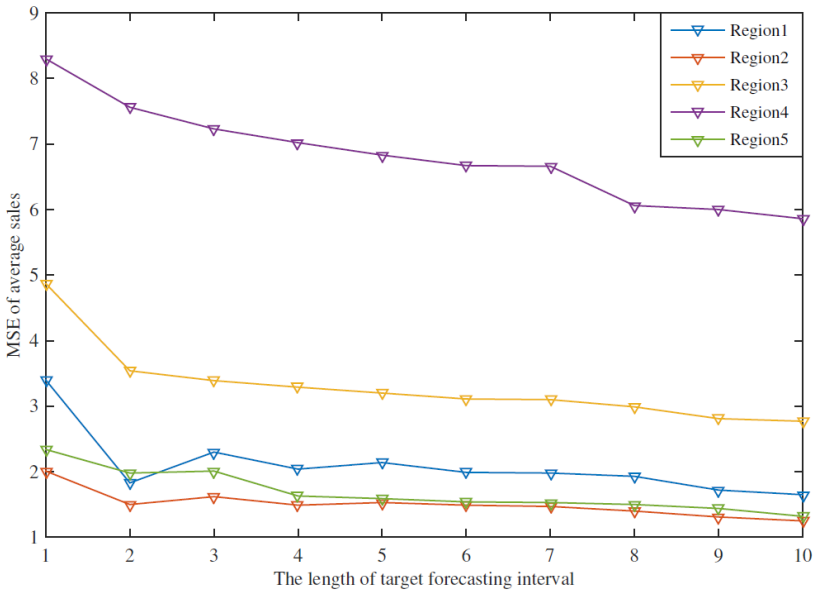
Έτσι, ένας από τους στόχους μας σε αυτό το κεφάλαιο είναι να βοηθήσουμε τους μάνατζερ να δουν τη σημασία των προβλέψεων πωλήσεων ως εισροή στα σχέδιά τους

και να κατανοήσουν πώς αυτές οι προβλέψεις πωλήσεων μπορούν και πρέπει να αναπτυχθούν. Για να γίνει αυτό, ως πρώτο βήμα θα πρέπει να αναλύσουμε τις ανάγκες των προβλέψεων πωλήσεων των πρωταρχικών διευθυντικών λειτουργιών εντός ενός οργανισμού. Με άλλα λόγια, πρέπει να γνωρίζουμε τι πρέπει να κάνει το μάρκετινγκ, οι πωλήσεις, το λογιστήριο, η παραγωγή και το τμήμα εφοδιασμού από την πρόβλεψη των πωλήσεων έως τη συμβολή της στα σχέδιά τους; Για να απαντήσουμε σε αυτήν την ερώτηση, θα καθορίσουμε πρώτα τις συναφείς έννοιες του επιπέδου προβλέψεων πωλήσεων, του χρονικού ορίζοντα, του χρονικού διαστήματος και της μορφής, κυρίως επειδή οι διαφορετικές λειτουργίες διαχείρισης απαιτούν διαφορετικά επίπεδα, ορίζοντες, διαστήματα και μορφές προβλέψεων πωλήσεων.

Το επίπεδο πρόβλεψης των πωλήσεων είναι το επίκεντρο της εταιρικής ιεραρχίας όπου απαιτείται η πρόβλεψη. Μια εταιρική πρόβλεψη, για παράδειγμα, είναι μια πρόβλεψη των συνολικών πωλήσεων για την εταιρεία. Ο χρονικός ορίζοντας των προβλέψεων πωλήσεων συμπίπτει γενικά με το χρονικό πλαίσιο του σχεδίου για το οποίο αναπτύχθηκε. Εάν, για παράδειγμα, συνεχίσουμε το παράδειγμα που μόλις δόθηκε, ένα εταιρικό σχέδιο μπορεί να είναι για τα επόμενα δύο χρόνια και, επομένως, χρειαζόμαστε μια πρόβλεψη πωλήσεων για αυτόν τον χρονικό ορίζοντα δύο ετών. Το χρονικό διάστημα πρόβλεψης πωλήσεων συμπίπτει γενικά με το πόσο συχνά ενημερώνεται το σχέδιο. Εάν το διετές σχέδιο εταιρικών πωλήσεων πρέπει να ενημερώνεται κάθε τρεις μήνες (όχι ένα ασυνήθιστο σενάριο), μπορούμε να πούμε ότι το επίπεδο είναι εταιρικό, ο ορίζοντας είναι δύο χρόνια και το χρονικό διάστημα είναι τριμηνιαίο. Η μορφή πρόβλεψης πωλήσεων είναι αυτό που πρέπει να προβλεφθεί ή να προγραμματιστεί. Ορισμένες λειτουργίες πρέπει να γνωρίζουν ποιες φυσικές μονάδες πρόκειται να παραχθούν ή και να αποσταλούν, ενώ άλλες λειτουργίες πρέπει να γνωρίζουν τα ισοδύναμα σε χρηματικά ποσά αυτών των μονάδων και άλλες λειτουργίες πρέπει να σχεδιάζονται βάσει συνολικών κιλών ή κυβικού όγκου. Αυτά αποτελούν τις μορφές που μπορεί να λάβει μια πρόβλεψη πωλήσεων (Mentzer and Moon (2005)).

Σε μία μελέτη περίπτωσης για την πρόβλεψη πωλήσεων στην ιστοσελίδα ηλεκτρονικού εμπορίου CaiNiao.com μέσω των δεδομένων που παρείχε η πλατφόρμα Alibaba, με τη χρήση του μοντέλου Συνελικτικού Νευρωνικού Δικτύου (Convolutional Neural Network), διαπιστώθηκε ότι όσο μεγαλύτερο είναι το χρονικό διάστημα πρόβλεψης και το πλαίσιο δεδομένων, τόσο πιο εύκολη και έγκυρη πρόβλεψη έχουμε. Αλλά πάντα υπάρχουν περιορισμοί στη χρονική διάρκεια και στο μέγεθος των δεδομένων.

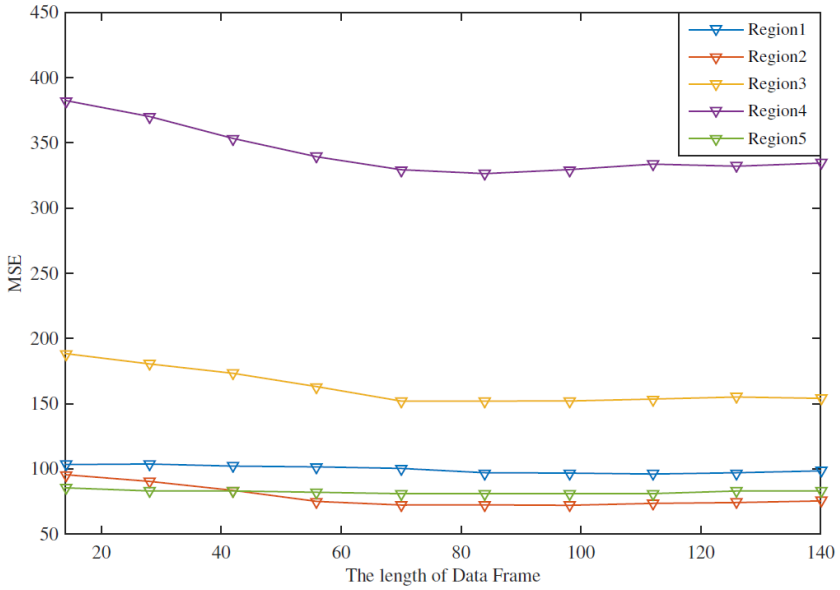
Το σχήμα 2 μας βοηθά να αναλύσουμε τη σχέση μεταξύ της διάρκειας του χρονικού διαστήματος πρόβλεψης και της δυσκολίας πρόβλεψης, που μεταφράζεται μέσα από το μέσο τετραγωνικό σφάλμα (Mean Square Error – MSE) του μέσου όρου πωλήσεων της κάθε περιοχής. Μπορούμε να δούμε ότι όσο μεγαλύτερο είναι το χρονικό διάστημα πρόβλεψης, τόσο πιο εύκολη είναι η πρόβλεψη των πωλήσεων, διότι και στις πέντε περιοχές μετρήσεων (Region 1–5), καθώς αυξάνεται το χρονικό διάστημα πρόβλεψης υπάρχει μείωση του μέσου τετραγωνικού σφάλματος (MSE). Ο κύριος λόγος είναι ότι οι συνολικές πωλήσεις σε ένα μεγάλο χρονικό διάστημα πρόβλεψης, είναι πιο σταθερές από αυτές σε ένα σύντομο χρονικό διάστημα πρόβλεψης. Ωστόσο, η πρόβλεψη με χρήση ενός σύντομου χρονικού διαστήματος επιτρέπει πιο ευέλικτες επιχειρηματικές αποφάσεις. Έτσι, υπάρχει μια αλληλοσύγκρουση μεταξύ της πρακτικής ευελιξίας και της ακρίβειας των προβλέψεων σε πραγματικές εφαρμογές.



Σχήμα 2: Διάγραμμα MSE του μέσου όρου πωλήσεων με τη διάρκεια του χρονικού διαστήματος της πρόβλεψης

Το μέγεθος του πλαισίου δεδομένων είναι μια άλλη κρίσιμη υπερ-παράμετρος στο μοντέλο μας. Αντιπροσωπεύει πόσα ιστορικά δεδομένα χρησιμοποιήσαμε ως είσοδο του μοντέλου μας. Όπως μπορεί να φανεί από το Σχήμα 3, εάν το πλαίσιο δεδομένων είναι πολύ μικρό, οι πληροφορίες που περιέχονται σε αυτό και για τις πέντε περιοχές μετρήσεων είναι ανεπαρκείς, δηλαδή μεγάλο μέσο τετραγωνικό σφάλμα (MSE). Από την άλλη πλευρά, το μεγάλο πλαίσιο δεδομένων μπορεί να περιέχει πάρα πολλές άχρηστες πληροφορίες, οι οποίες συγχέουν τη μηχανική μάθηση και για το λόγο αυτό

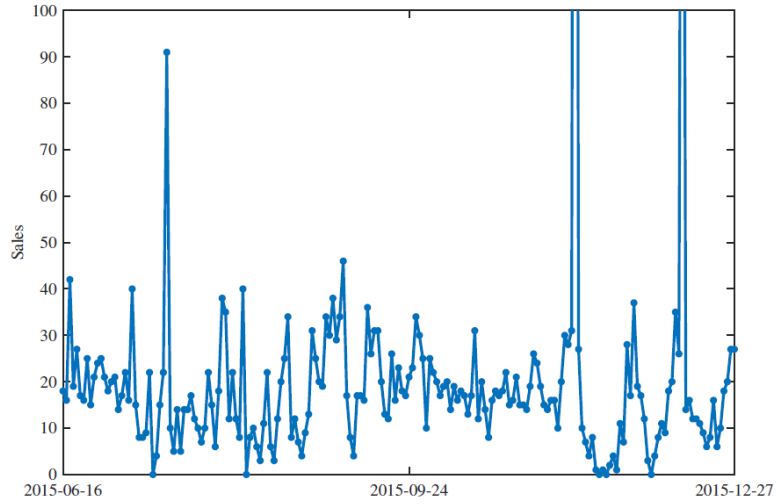
εμφανίζεται μία μικρή αύξηση του μέσου τετραγωνικού σφάλματος (Σχήμα 3). Επιπλέον, το μεγαλύτερο πλαίσιο δεδομένων σημαίνει περισσότερη κατανάλωση υπολογιστικών πόρων.



Σχήμα 3: Διάγραμμα MSE με το μέγεθος πλαισίου των δεδομένων

Το δυναμικό και σύνθετο επιχειρηματικό περιβάλλον στο ηλεκτρονικό εμπόριο δημιουργεί μεγάλες προκλήσεις για τη λήψη αποφάσεων στις επιχειρήσεις. Πολλές ευφυείς τεχνολογίες, όπως οι προβλέψεις πωλήσεων, αναπτύσσονται για να ξεπεράσουν αυτές τις προκλήσεις. Οι προβλέψεις πωλήσεων είναι χρήσιμες για τη διαχείριση του εργατικού δυναμικού, των ταμειακών ροών και των πόρων, έτσι ώστε να βελτιστοποιήσουμε την αλυσίδα εφοδιασμού των παραγωγών.

Η αξία των προβλέψεων πωλήσεων εξαρτάται από την ακρίβειά τους. Οι ανακριβείς προβλέψεις ενδέχεται να οδηγήσουν σε ελλιπή ή μεγάλα αποθέματα, επηρεάζοντας την αποτελεσματικότητα των αποφάσεων στο ηλεκτρονικό εμπόριο. Οι παραδοσιακές τεχνικές πρόβλεψης πωλήσεων βασίζονται σε ανάλυση χρονοσειρών, οι οποίες βασίζονται μόνο στα ιστορικά δεδομένα πωλήσεων. Αυτές οι μέθοδοι μπορούν να χειριστούν καλά τα εμπορεύματα με σταθερές ή εποχιακές τάσεις πωλήσεων (Keller and Gaciu (2012)). Ωστόσο, τα εμπορεύματα στο ηλεκτρονικό εμπόριο είναι πολύ πιο ακανόνιστα στις τάσεις των πωλήσεών τους (ένα παράδειγμα φαίνεται στο σχήμα 4) και η ακρίβεια των προβλέψεων που επιτυγχάνονται με αυτές τις παραδοσιακές μεθόδους είναι πολύ μικρή (Beheshti-Kashi et al. (2015)).



Σχήμα 4: Ιστορικό πωλήσεων ενός τυπικού εμπορεύματος από το σύνολο δεδομένων του CaiNiao.com

Ευτυχώς, είναι διαθέσιμος ένας τεράστιος όγκος δεδομένων στο ηλεκτρονικό εμπόριο και είναι δυνατή η εκμετάλλευση αυτών των δεδομένων για τη βελτίωση της ακρίβειας των προβλέψεων. Εκτός από τα ιστορικά δεδομένα πωλήσεων, μπορούμε να συλλέξουμε πολλά άλλα στοιχεία καταγραφής για online προϊόντα για μεγάλο χρονικό διάστημα, όπως προβολή σελίδας (PV), προβολή σελίδας από αναζήτηση (SPV), προβολή χρηστών (UV), προβολή χρηστών από την αναζήτηση (SUV), την τιμή πώλησης (PAY) και τον όγκο των εμπορευμάτων χονδρικής (GMV) κλπ. Με τη χρήση μεθόδων μάθησης υπό επίβλεψη, όπως είναι τα μοντέλα παλινδρόμησης, αυτές οι πληροφορίες μπορούν να ενσωματωθούν στο μοντέλο πρόβλεψης πωλήσεων και να επιτευχθεί μεγαλύτερη ακρίβεια πρόβλεψης. Το πρώτο βήμα των συμβατικών μεθόδων μηχανικής μάθησης είναι γενικά η μηχανική των χαρακτηριστικών, όπου τα αποτελεσματικά χαρακτηριστικά εξάγονται με το χέρι από τα διαθέσιμα δεδομένα χρησιμοποιώντας την ειδική γνώση του τομέα (Domingos (2012)). Η ποιότητα και η ποσότητα των χαρακτηριστικών μπορεί να επηρεάσει σημαντικά την ακρίβεια του τελικού μοντέλου πρόβλεψης. Ωστόσο, η εξεύρεση αποτελεσματικών χαρακτηριστικών είναι μια δύσκολη και χρονοβόρα διαδικασία. Επιπλέον, αυτά τα χαρακτηριστικά γενικά εξάγονται κατά περίπτωση για συγκεκριμένα εμπορικά σενάρια και τα μοντέλα αυτά είναι δύσκολο να επαναχρησιμοποιηθούν όταν μεταβάλλονται τα δεδομένα ή οι απαιτήσεις. Για παράδειγμα, αφού συγκεντρωθούν περισσότερα δεδομένα για ηλεκτρονικά προϊόντα, θα πρέπει να γίνει και πάλι η μηχανική των χαρακτηριστικών για να ενσωματωθούν οι πληροφορίες που περιέχονται στα νέα δεδομένα στο μοντέλο πρόβλεψης πωλήσεων.

Η εκμάθηση χαρακτηριστικών μπορεί να αποτρέψει την ανάγκη χειροκίνητης μηχανικής χαρακτηριστικών (Bengio, Courville and Vincent (2013)). Μέσω της μάθησης χαρακτηριστικών, τα λειτουργικά χαρακτηριστικά μπορούν να εκπαιδευτούν αυτόματα από τα δεδομένα πρώτης εισόδου και στη συνέχεια να χρησιμοποιηθούν σε συγκεκριμένες εργασίες μηχανικής μάθησης. Το βαθύτερο νευρωνικό δίκτυο είναι μία από τις πιο δημοφιλείς μεθόδους μάθησης χαρακτηριστικών γνωρισμάτων. Είναι εμπνευσμένο από το νευρικό σύστημα, όπου οι κόμβοι λειτουργούν ως νευρώνες και οι άκρες λειτουργούν ως σύναψη. Ένα νευρωνικό δίκτυο χαρακτηρίζει μια συνάρτηση από τη σχέση μεταξύ του επιπέδου εισόδου και του επιπέδου εξόδου, η οποία παραμετροποιείται από τους συντελεστές βαρύτητας που σχετίζονται με τις άκρες. Τα χαρακτηριστικά μαθαίνονται στα κρυφά επίπεδα και στη συνέχεια χρησιμοποιούνται για ταξινόμηση ή παλινδρόμηση στο επίπεδο εξόδου. Υπάρχουν πολλές εργασίες που χρησιμοποιούν τα βαθύτερα νευρωνικά δίκτυα για να μάθουν τα χαρακτηριστικά από τα αδόμητα δεδομένα, όπως από την εικόνα (Krizhevsky, Sutskever and Hinton (2012)), τον ήχο (Graves, Mohamed and Hinton (2013)) και το κείμενο (Blunsom, Grefenstette and Kalchbrenner (2014)).

Σε αυτή τη μελέτη προτείνουν μια καινοτόμο προσέγγιση για να εξάγουν τα αποτελεσματικά χαρακτηριστικά αυτόματα από τα δομημένα δεδομένα χρησιμοποιώντας το Συνελικτικό Νευρωνικό Δίκτυο (CNN), το οποίο είναι μία από τις πιο δημοφιλείς αρχιτεκτονικές βαθύτερων νευρωνικών δικτύων. Πρώτον, μετασχηματίζουμε τα δεδομένα καταγραφής του εμπορεύματος σε ένα σχεδιασμένο πλαίσιο δεδομένων. Στη συνέχεια, εφαρμόζουν το Συνελικτικό Νευρωνικό Δίκτυο σε αυτό το πλαίσιο δεδομένων, όπου τα λειτουργικά χαρακτηριστικά θα εξάγονται στα κρυφά επίπεδα και στη συνέχεια χρησιμοποιούνται για τις προβλέψεις πωλήσεων στο επίπεδο εξόδου. Αυτή η προσέγγιση λαμβάνει τα ακατέργαστα ημερολογιακά δεδομένα των εμπορευμάτων και είναι εύκολο αυτά να ενσωματωθούν στα νέα διαθέσιμα δεδομένα στο μοντέλο πρόβλεψης των πωλήσεων με ελάχιστη ανθρώπινη παρέμβαση. Επιπλέον, η τεχνική της φθίνουσας βαρύτητας του δείγματος (Weight Decay - WD) και η τεχνική της εκμάθησης μεταφοράς (Transfer Learning - TL) χρησιμοποιούνται για να βελτιώσουν περαιτέρω την ακρίβεια των προβλέψεων. Έπειτα δοκιμάζουν την προσέγγισή τους σε ένα μεγάλο σύνολο ρεαλιστικών δεδομένων από το CaiNiao.com και τα εμπειρικά αποτελέσματα επικυρώνουν την αποτελεσματικότητα αυτής της μεθόδου.

Συγκεκριμένα σε αυτό το παράδειγμα αρχικά χρησιμοποιήθηκε η μέθοδος ARIMA, η οποία είναι μία κλασσική μέθοδος ανάλυσης χρονοσειρών που χρησιμοποιεί τα

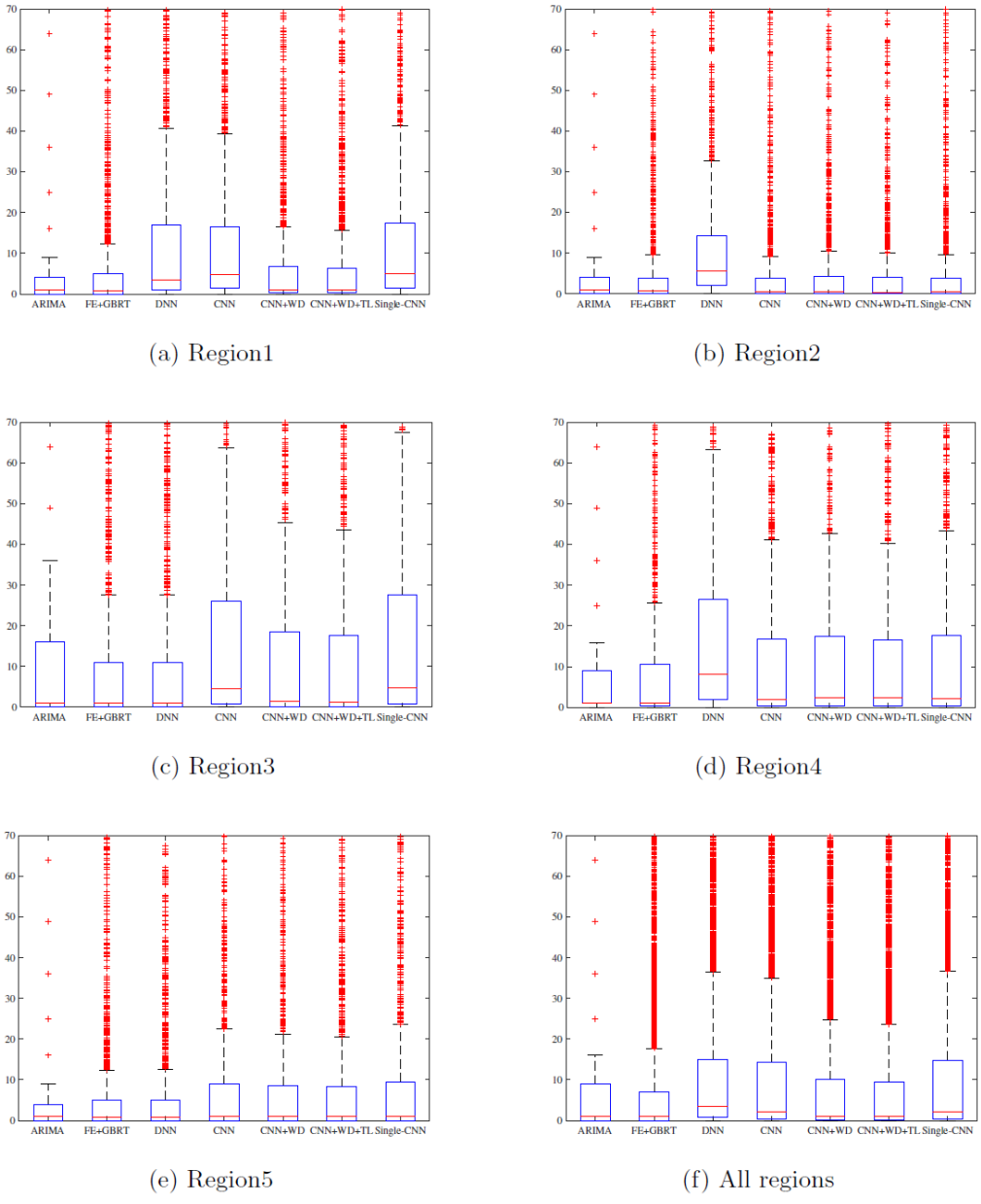
δεδομένα πωλήσεων του παρελθόντος ως είσοδο και προβλέπει τις πωλήσεις της επόμενης ακριβώς στιγμής. Στη συνέχεια εφαρμόστηκε η Βαθμιδωτή Ενίσχυση με Δέντρο Παλινδρόμησης (FE + Gradient Boosting Regression Tree), μια μέθοδος που προσθέτει διαδοχικά τα δεδομένα εκπαίδευσης και τους αποδίδει ένα συντελεστή βαρύτητας. Στη συγκεκριμένη περίπτωση τις παρελθοντικές προβολές από τους χρήστες (UV) σε συνδυασμό με τις μειώσεις των τιμών. Ωστόσο, αντί να αναθέσει διαφορετικά βάρη στους ταξινομητές μετά από κάθε επανάληψη, η μέθοδος αυτή προσαρμόζει το νέο μοντέλο στα νέα υπολείμματα της προηγούμενης πρόβλεψης με στόχο την ενίσχυση των δέντρων αποφάσεων. Ο στόχος είναι να βρεθεί η βέλτιστη διαδρομή με το μικρότερο σφάλμα. Η επόμενη μέθοδος που χρησιμοποιήθηκε ήταν το Βαθύτερο Νευρωνικό Δίκτυο (Deep Neural Network - DNN), το οποίο είναι η απλούστερη δομή των νευρωνικών δικτύων, το οποίο χρησιμοποιεί πολλαπλά επίπεδα μεταξύ των επιπέδων εισόδου και εξόδου. Το Βαθύτερο Νευρωνικό Δίκτυο βρίσκει τη σωστή μαθηματική μέθοδο, είτε είναι γραμμική είτε μη γραμμική σχέση, ώστε να μετατρέψει τη είσοδο σε έξοδο. Το δίκτυο κινείται μεταξύ των επιπέδων υπολογίζοντας την πιθανότητα της κάθε εξόδου. Τέλος εφαρμόστηκαν διάφορες παραλλαγές των Συνελικτικών Νευρωνικών Δικτύων, τα οποία είναι τμήμα των Βαθύτερων Νευρωνικών Δικτύων με τη διαφοροποίηση ότι χρησιμοποιούν μία παραλλαγή πολλαπλών επιπέδων που έχουν σχεδιαστεί με τέτοιο τρόπο, ώστε να προσομοιώνουν την ικανότητα του εγκεφάλου να αναγνωρίζει και να διακρίνει, με αποτέλεσμα να επιλέγουν τη σωστή ακολουθία πράξεων και να απαιτούν την ελάχιστη προεπεξεργασία.

Τα λεπτομερή πειραματικά αποτελέσματα, δηλαδή το μέσο τετραγωνικό σφάλμα (MSE) της κάθε μεθόδου για την κάθε περιοχή παρουσιάζονται στο Σχήμα 5 και η περίληψη αυτών των αποτελεσμάτων παρουσιάζεται στον Πίνακα 1.

Η κλασική μέθοδος εκμάθησης μηχανών (FE + GBRT) λαμβάνει υπ’όψιν περισσότερες πληροφορίες από ότι η μέθοδος ανάλυσης χρονοσειρών (ARIMA) και κατά συνέπεια επιτυγχάνει καλύτερες επιδόσεις. Η απλούστερη αρχιτεκτονική νευρωνικών δικτύων (DNN) εξάγει τα χαρακτηριστικά αυτόματα. Μερικές φορές αποκτά ακόμα πιο χρήσιμη αναπαράσταση χαρακτηριστικών από ότι η μηχανική των χαρακτηριστικών που γίνεται από άνθρωπο, η οποία DNN όπως φαίνεται και στο πείραμα υπερέχει της μεθόδου FE + GBRT σε ορισμένες περιπτώσεις.

Η αρχιτεκτονική του Συνελικτικού Νευρωνικού Δικτύου (CNN) μπορεί να χρησιμοποιήσει πλήρως την εγγενή δομή στα ακατέργαστα δεδομένα για να εξαγάγει πιο αποτελεσματικά χαρακτηριστικά και να επιτύχει σημαντική βελτίωση στην

απόδοση. Η τεχνική της φθίνουσας βαρύτητας του δείγματος (WD) και η τεχνική της εκμάθησης μεταφοράς (TL) είναι εξαιρετικά αποτελεσματικές. Αυξάνουν περαιτέρω την απόδοση και τα τελικά αποτελέσματα είναι πολύ ανταγωνιστικά. Επιπλέον, όπως φαίνεται από τον πίνακα 1 και το σχήμα 5, οι προβλέψεις που προκύπτουν από αυτές τις δύο μεθόδους είναι πιο ισχυρές, διότι η διάμεσος (50% των αποτελεσμάτων) των θηκογραμμάτων (box-plots) στο σχήμα 5 αντιστοιχούν στο μικρότερο σφάλμα. Είναι ενδιαφέρον να διερευνηθεί κατά πόσον είναι δυνατόν να εκπαιδευτεί ένα μοντέλο για την πρόβλεψη των πωλήσεων σε όλες τις περιοχές. Έτσι, εκπαιδεύουμε ένα ενοποιημένο μοντέλο χρησιμοποιώντας όλα τα δείγματα εκπαίδευσης στην D και προβλέπουμε τις πωλήσεις για κάθε περιοχή r αντίστοιχα. Τα αποτελέσματα είναι ελπιδοφόρα αλλά λιγότερο ανταγωνιστικά από τη χρήση μεμονωμένων μοντέλων.



(a) Region1 (b) Region2

(c) Region3 (d) Region4

(e) Region5 (f) All regions

Σχήμα 5: Θηκογράμματα (box-plots) των MSE για κάθε περιοχή

Πίνακας 1: Οι τιμές των MSE όλων των μεθόδων για κάθε περιοχή.

Region	1	2	3	4	5	Avrage
ARIMA	104.37	96.68	190.50	397.08	87.18	175.16
FE+GBRT	97.36	83.90	187.06	329.81	82.21	156.07
DNN	97.50	73.55	181.67	347.50	82.17	156.48
CNN	96.98	72.22	151.96	326.39	80.91	145.69
CNN+WD	89.01	56.14	142.27	301.79	75.09	131.86
CNN+WD+TL	**84.30**	**53.40**	**134.92**	**287.31**	**71.19**	**126.22**
Single-CNN	101.92	75.70	159.48	343.54	85.04	153.22

Συμπερασματικά, σε αυτή τη μελέτη παρουσιάστηκε μια νέα προσέγγιση για να παίρνουμε αποτελεσματικά χαρακτηριστικά αυτόματα από τα δομημένα δεδομένα χρησιμοποιώντας το Συνελικτικό Νευρωνικό Δίκτυο. Μπορεί να αποφευχθεί η ανάγκη χειροκίνητης μηχανικής χαρακτηριστικών, η οποία είναι συνήθως δύσκολη, χρονοβόρα και απαιτεί εξειδικευμένες γνώσεις. Χρησιμοποιήθηκε η προτεινόμενη προσέγγιση για την πρόβλεψη των πωλήσεων λαμβάνοντας τα ακατέργαστα ημερολογιακά δεδομένα και πληροφοριακά γνωρίσματα των εμπορευμάτων ως εισροές. Πρώτον, μετασχηματίστηκαν τα δεδομένα καταγραφής και τα πληροφοριακά γνωρίσματα των εμπορευμάτων, που είναι σε δομημένο τύπο, σε ένα σχεδιασμένο πλαίσιο δεδομένων. Στη συνέχεια, εφαρμόστηκε το Συνελικτικό Νευρωνικό Δίκτυο στο πλαίσιο δεδομένων, όπου τα αποτελεσματικά χαρακτηριστικά θα εξάγονται στα κρυφά επίπεδα και στη συνέχεια θα χρησιμοποιούνται για τις προβλέψεις των πωλήσεων. Δοκιμάζουμε την προσέγγισή μας σε ένα πραγματικό σύνολο δεδομένων από το CaiNiao.com και αυτή επιδεικνύει πολύ καλή απόδοση. Επιπλέον, η τεχνική της φθίνουσας βαρύτητας του δείγματος (WD) και η τεχνική της εκμάθησης μεταφοράς (TL) χρησιμοποιούνται για να βελτιώσουν περαιτέρω την ακρίβεια της πρόγνωσης, οι οποίες αποδείχθηκαν ιδιαίτερα αποτελεσματικές στα πειράματα (Zhao and Wang (2017)).

2.4.1. Οι ανάγκες πρόβλεψης πωλήσεων για το τμήμα Μάρκετινγκ

Τον τομέα μάρκετινγκ συνήθως τον απασχολεί η επιτυχία μεμονωμένων προϊόντων και σειρών προϊόντων που προσφέρει η εταιρεία στους πελάτες της. Αυτή η ανησυχία εκδηλώνεται σε ετήσια σχέδια (ενημερωμένα μηνιαία ή τριμηνιαία) των προσπαθειών μάρκετινγκ για νέα και υπάρχοντα προϊόντα. Τα σχέδια μάρκετινγκ, με τη σειρά τους, περιλαμβάνουν συνήθως τις προβαλλόμενες αλλαγές προϊόντων, τις προωθητικές ενέργειες, την επιλογή καναλιών διανομής και την τιμολόγηση. Για να αναπτύξει αυτά τα σχέδια, οι πωλήσεις χρειάζονται προβλέψεις πωλήσεων που λαμβάνουν υπόψη τις

διάφορες αυτές προσπάθειες και τις πωλήσεις (συνήθως σε χρηματικά ποσά) σε επίπεδο προϊόντος και γραμμής προϊόντων για ετήσιο χρονικό ορίζοντα και με μηνιαία ή τριμηνιαία διαστήματα (Mentzer and Moon (2005)).

2.4.2. Οι ανάγκες πρόβλεψης πωλήσεων για το τμήμα πωλήσεων

Το τμήμα πωλήσεων, ως λειτουργία μάνατζμεντ, συνήθως το απασχολεί ο καθορισμός στόχων για τους μεμονωμένους πωλητές και την παρακίνηση αυτών των πωλητών να υπερβούν αυτούς τους στόχους. Οι τομείς των πωλητών μπορούν να οριστούν με πολλούς τρόπους (γεωγραφικά, ανά βιομηχανία, ανά πελάτη, ανά προϊόν και ούτω καθεξής) και αυτός ο ορισμός βοηθά στον καθορισμό του επιπέδου προβλέψεων πωλήσεων για μια συγκεκριμένη λειτουργία πωλήσεων.

Ο ορίζοντας και το διάστημα καθορίζονται σε μεγάλο βαθμό από το χρονικό πλαίσιο του σχεδίου αποζημίωσης. Εάν, για παράδειγμα, ορισμένοι πωλητές λαμβάνουν τις προμήθειές τους βάσει τριμηνιαίων πωλήσεων και ο μάνατζερ πωλήσεων πρέπει να προγραμματίσει τα επόμενα τέσσερα τρίμηνα, ο ορίζοντας θα είναι ένα έτος και το διάστημα θα είναι τριμηνιαίο. Τουλάχιστον, οι λειτουργίες μάνατζμεντ των πωλήσεων των περισσότερων εταιρειών χρειάζονται προβλέψεις πωλήσεων (σε χρηματικά ποσά) σε επίπεδο τομέα, με τυπικούς ορίζοντες ενός ή δύο ετών και μηνιαία ή τριμηνιαία (Mentzer and Moon (2005)).

2.4.3. Οι ανάγκες πρόβλεψης πωλήσεων για το τμήμα χρηματοοικονομικών

Μεταξύ των άλλων αρμοδιοτήτων, το τμήμα χρηματοοικονομικών (με εισροή από τη λογιστική λειτουργία) επιβαρύνεται με το έργο της προβολής των επιπέδων κόστους και κέρδους και των κεφαλαιακών αναγκών, με βάση συγκεκριμένες προβλέψεις πωλήσεων. Αυτά τα "σχέδια κερδών" είναι συνήθως ετήσια χρονικά διαστήματα και μπορούν να επεκταθούν από ένα έως πέντε έτη. Παρόλο που οι πωλήσεις μεμονωμένων προϊόντων αποτελούν συμβολή σε αυτήν τη διαδικασία σχεδιασμού (επειδή το κόστος διαφορετικών προϊόντων μπορεί να ποικίλλει), η ανησυχία για το σχέδιο κέρδους είναι συνήθως σε εταιρικό ή τμηματικό επίπεδο. Έτσι, οι ανάγκες πρόβλεψης των πωλήσεων στο τμήμα χρηματοοικονομικών, είναι συνήθως πωλήσεις σε χρηματικά ποσά σε εταιρικό επίπεδο, σε τμήμα, σε επίπεδο γραμμής προϊόντος. Ενώ ο ορίζοντας είναι συνήθως ένα έως πέντε χρόνια και το διάστημα είναι τριμηνιαία ή μηνιαία (ανάλογα με το πόσο συχνά ενημερώνεται το σχέδιο) (Mentzer and Moon (2005)).

2.4.4. Οι ανάγκες πρόβλεψης πωλήσεων για το τμήμα παραγωγής και προμηθειών

Το τμήμα παραγωγής και προμηθειών το αφορούν δύο πολύ διαφορετικές προβλέψεις μονάδων, μία μακροπρόθεσμη και μία βραχυπρόθεσμη. Η μακροπρόθεσμη πρόβλεψη χρησιμοποιείται για τον προγραμματισμό της ανάπτυξης προμηθευτών, εγκαταστάσεων και εξοπλισμού, η οποία μπορεί να διαρκέσει αρκετά χρόνια. Επειδή αυτά τα μακροπρόθεσμα σχέδια εξαρτώνται από το μίγμα των πωλήσεων των προϊόντων που πρόκειται να κατασκευαστούν στο εργοστάσιο, η πρόβλεψη πρέπει να είναι σε επίπεδο επιμέρους προϊόντος (συχνά στην ορολογία πρόβλεψης που αναφέρεται ως μονάδες αποθεματοποίησης). Ο ορίζοντας εξαρτάται από το χρόνο που χρειάζεται για να φέρουμε σε λειτουργία νέους προμηθευτές, εγκαταστάσεις και εξοπλισμό και επομένως μπορεί να κυμανθεί από ένα έως τρία χρόνια. Το διάστημα για την ενημέρωση αυτών των προβλέψεων είναι συνήθως τριμηνιαία.

Η πρόβλεψη βραχυπρόθεσμης παραγωγής ή προμήθειας βασίζεται στις ανάγκες του χρονοδιαγράμματος παραγωγής, το οποίο μπορεί να κυμαίνεται από ένα έως έξι μήνες (ανάλογα με τον κύκλο της παραγγελίας πρώτων υλών) και χρειάζεται μια συγκεκριμένη λεπτομέρεια για το ποια προϊόντα παράγονται. Έτσι, αυτή η βραχυπρόθεσμη πρόβλεψη πωλήσεων για το τμήμα παραγωγής και προμηθειών βρίσκεται στο επίπεδο της αποθήκης πρώτων υλών, έχει έναν ορίζοντα σπάνια μεγαλύτερο από έξι μήνες και έχει διαστήματα που κυμαίνονται από ημερήσια έως μηνιαία (Mentzer and Moon (2005)).

Ένα χαρακτηριστικό παράδειγμα στις ανάγκες πρόβλεψης για το τμήμα παραγωγής και προμηθειών, είναι η μελέτη περίπτωσης της πρόγνωσης πωλήσεων άρτου της εταιρείας Favorita. Θα παρουσιάσουμε τη συμμετοχή που κέρδισε τη δεύτερη θέση στον διαγωνισμό προγνωστικών πωλήσεων ψωμιού της εταιρείας Favorita (Kaggle Favorita Grocery (2017)) που φιλοξένησε η διαδικτυακή πλατφόρμα διαγωνισμών ανάλυσης μεγάλων βάσεων δεδομένων, Kaggle.

Τα αρτοποιία είναι ένα από τα είδη επιχειρήσεων που βασίζονται υπερβολικά με την πρόβλεψη αγορών και πωλήσεων. Εάν προβλέψουμε λίγο περισσότερα, τότε τα αρτοποιία παραμένουν γεμάτα με πολλά και φθαρμένα αγαθά. Εάν προβλέψουμε λιγότερα, τότε τα δημοφιλή προϊόντα ξεπουλάνε γρήγορα και οι πελάτες εξαφανίζονται και πηγαίνουν σε άλλο κατάστημα. Το πρόβλημα γίνεται πιο περίπλοκο, καθώς οι έμποροι λιανικής προσθέτουν νέες τοποθεσίες με μοναδικές ανάγκες, νέα προϊόντα, εποχιακές γεύσεις και απρόβλεπτο μάρκετινγκ προϊόντων. Η κοινότητα της Kaggle είχε την πρόκληση να κατασκευάσει ένα μοντέλο που προβλέπει ακριβέστερες πωλήσεις

προϊόντων. Με άλλα λόγια, ο στόχος είναι η δημιουργία ενός μοντέλου με την υψηλότερη ακρίβεια για τις προβλέψεις πωλήσεων, με δυνατότητα χρήσης του μοντέλου στην παραγωγή.

Οι διαγωνισμοί χρονοσειρών που προσφέρονται από τον Kaggle και άλλους οργανισμούς έχουν γίνει δημοφιλείς στη μηχανική μάθηση (machine learning). Με την ύπαρξη κοινών κανόνων συμμετοχής, καθώς και με τα σύνολα δεδομένων εκμάθησης και δοκιμών που μοιράζονται όλοι οι διαγωνιζόμενοι, αυτοί οι διαγωνισμοί μπορούν να συμβάλουν στην προώθηση αυτής της σύγχρονης τεχνολογίας της μηχανικής μάθησης σε ποικίλους τομείς εφαρμογών. Οι χρονοσειρές θεωρούνται μία από τις λιγότερο γνωστές δεξιότητες στον χώρο της ανάλυσης. Η εποχικότητα, οι τάσεις και οι κύκλοι που υπάρχουν στα δεδομένα, είναι δύσκολο να εντοπιστούν και να προβλεφθούν με ακρίβεια λόγω των μη γραμμικών τάσεων και του θορύβου που παρουσιάζονται στη σειρά. Η σημαντική αύξηση της δημοτικότητας των νευρωνικών δικτύων (neural networks) έχει δώσει μια ριζικά διαφορετική κατανόηση του τρόπου με τον οποίο θα μπορούσε να γίνει η πρόβλεψη. Οι πρόοδοι στα μηχανήματα υπολογιστών έχουν καταστήσει δυνατή την αντιμετώπιση προβλημάτων με νευρωνικά δίκτυα σε εύλογο χρονικό διάστημα. Τώρα που είναι μια εφικτή λύση, η βαθύτερη εκμάθηση έχει θέσει πολλά νέα αρχεία για την ακριβή ταξινόμηση σε σύνολα δεδομένων αναφοράς τα τελευταία χρόνια (Ching et al. (2018); Shvets et al. (2018)).

Σύμφωνα με τη φόρμα διαγωνισμών Kaggle, τα δεδομένα χωρίζονται σε δύο τύπους – δεδομένα, ης εκμάθησης και της δοκιμής. Τα δεδομένα της εκμάθησης αντιπροσωπεύουν δεδομένα για την εκπαίδευση του μοντέλου, ενώ τα δεδομένα δοκιμής χωρίζονται σε τμήματα και χρησιμοποιούνται για την αξιολόγηση της ακρίβειας των μοντέλων σε δημόσιους και ιδιωτικούς πίνακες βαθμολογίας. Στον διαγωνισμό, η εταιρεία Favorita μας δίνει 125.497.040 παρατηρήσεις ως δεδομένα εκμάθησης και 3.370.464 ως δεδομένα δοκιμής. Τα σύνολα δεδομένων αποτελούνται από πωλήσεις κατά ημερομηνία, αριθμό καταστήματος, κωδικό είδους και πληροφορίες προώθησης. Εκτός αυτού, δόθηκαν πληροφορίες για τις χρηματικές συναλλαγές, τις τιμές του πετρελαίου, πληροφορίες για τις αποθήκες και τις ημέρες αργιών.

Το καθορισμένο πρόβλημα πρόβλεψης έχει τις ακόλουθες προκλήσεις:

1. Θορυβώδη δεδομένα: ενώ οι διοργανωτές προσπάθησαν όσο το δυνατόν καλύτερα να προετοιμάσουν τα στοιχεία για τους συμμετέχοντες και συγκέντρωσαν ένα μεγάλο όγκο δεδομένων, τα προβλήματα με τις θορυβώδεις καταχωρήσεις υπήρχαν. Μερικά από τα στοιχεία (τιμές πετρελαίου, αργίες, συναλλαγές) δεν συσχετίστηκαν με τον στόχο και δεν χρησιμοποιήθηκαν στο μέλλον.

2. Αόρατα δεδομένα: παρατηρήθηκαν αόρατα δεδομένα στα δεδομένα εκμάθησης. Αυτό σημαίνει ότι το μοντέλο συμπεριφέρεται απρόβλεπτα στα αόρατα δεδομένα καταστημάτων - αγαθών. Ο λόγος για αυτό, είναι ότι το σύνολο των δεδομένων εκμάθησης, δεν περιλαμβάνει εγγραφές για μηδενικές πωλήσεις. Ωστόσο, το σύνολο των δεδομένων δοκιμών, περιλαμβάνει όλους τους συνδυασμούς καταστημάτων - αγαθών, ανεξάρτητα από το αν το στοιχείο αυτό είχε παρατηρηθεί στο παρελθόν σε κάποιο κατάστημα. Τέλος, αυτοί οι συνδυασμοί αντικαταστάθηκαν από μηδενικά με την παραδοχή ότι οι αόρατοι συνδυασμοί ήταν απλά δεδομένα μηδενικών πωλήσεων.

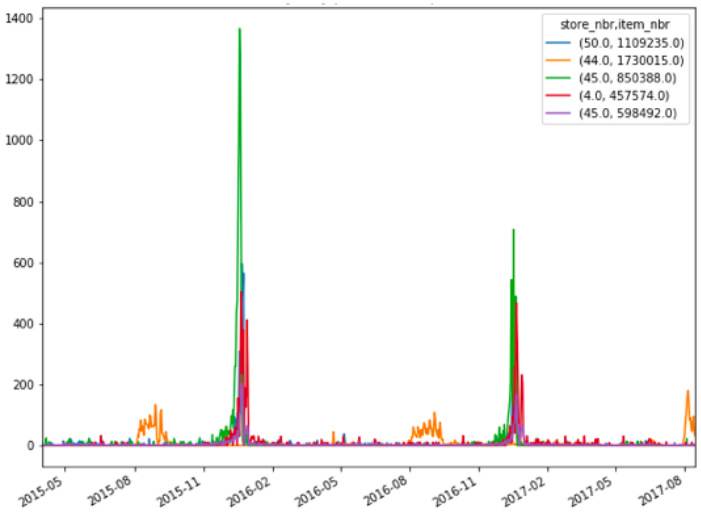
3. Ακρίβεια: δεδομένου ότι το πείραμα αυτό διεξήχθη στο πλαίσιο του διαγωνισμού, χρησιμοποιήθηκαν όλες οι δυνατότητες για την αύξηση της ακρίβειας των προβλέψεων.

Ως γενική κατεύθυνση χρησιμοποιήθηκε το πρωτότυπο μοντέλο συνελικτικού νευρωνικού δικτύου WaveNet (WaveNet CNN) (Vasquez (2017)), με κάποιες επεκτάσεις και τροποποιήσεις (Gulcehre et al. (2014)).

Τα πρόσφατα επιτεύγματα των τεχνικών βαθύτερης μάθησης ώθησαν τους ερευνητές να διερευνήσουν μεθόδους και τεχνικές όπως το WaveNet σε τομείς πρόβλεψης χρονοσειρών (Oosterlee, Borovykh and Bohte (2017)). Το WaveNet είναι ένα γενετικό μοντέλο. Αυτό σημαίνει ότι το μοντέλο μπορεί να παράγει τις ακολουθίες δεδομένων πραγματικής αποτίμησης χρησιμοποιώντας ορισμένες εισόδους υπό όρους.

Η επικύρωση διοργανώθηκε βήμα προς βήμα. Κρατήσαμε τις τελευταίες 16 ημέρες των δεδομένων της εκμάθησης και το χρησιμοποιήσαμε για επαλήθευση. Χρησιμοποιήθηκαν μετατοπισμένα στοιχεία πωλήσεων και περίοδοι προώθησης προϊόντων, για την καλύτερη καταγραφή τριμηνιαίων και ετήσιων προτύπων.

Το μοντέλο μας άρχισε να προβλέπει μετά από 5000 επαναλήψεις μικρών παρτίδων και παρήγαγε προβλέψεις κάθε 2000 επαναλήψεις. Μετά από ορισμένες επαναλήψεις, όπως μπορείτε να δείτε στο Σχήμα 6, ένας μέσος όρος πέντε μοντέλων, έχουν καλύτερη απόδοση από ένα μοντέλο. Για να αυξηθεί η ακρίβεια περισσότερο, χρησιμοποιήθηκε ένας εκθετικός κινητός μέσος με έναν παράγοντα εξομάλυνσης (smooth factor) ο οποίος υπολογίστηκε με την τοπική εγκάρσια επικύρωση (local cross-validation).

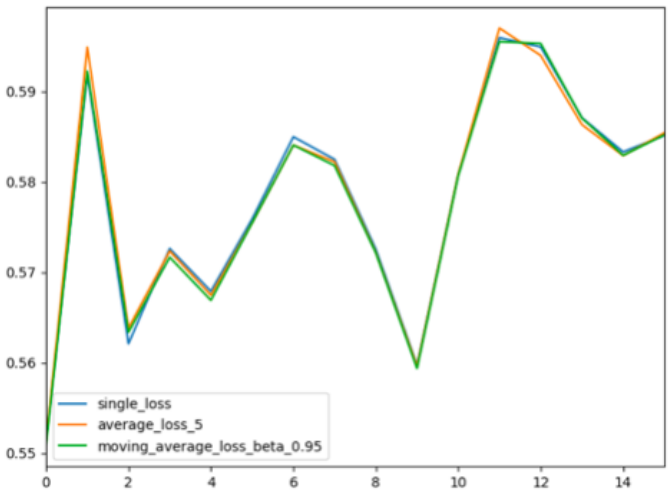


Σχήμα 6: Τάσεις και κύκλοι

Όσον αφορά τα χαρακτηριστικά, το μοντέλο έχει την ικανότητα να πιάσει τα πρότυπα των δεδομένων της χρονοσειράς, οπότε δεν χρησιμοποιήθηκαν πολλά χαρακτηριστικά σε αυτό. Ορισμένα από αυτά είναι πωλήσεις μονάδων και πληροφορίες προώθησης, οι πωλήσεις μονάδων μετατοπίστηκαν στο παρελθόν και οι πληροφορίες προώθησης μετατοπίστηκαν στο μέλλον και στο παρελθόν.

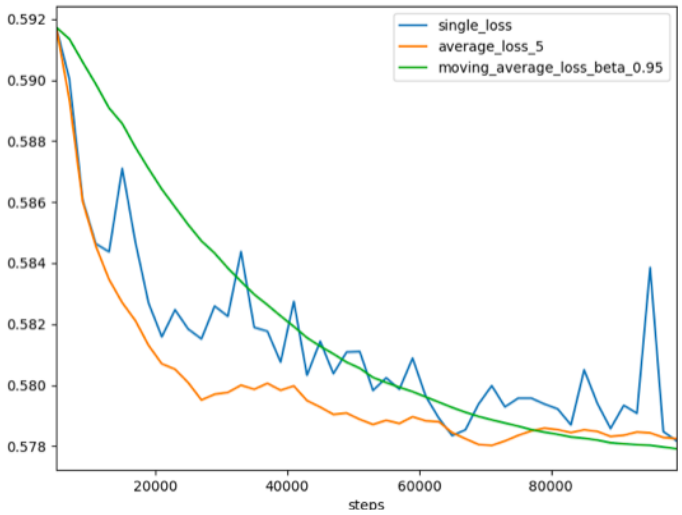
Συμπερασματικά, το προφανές πρόβλημα που αντιμετωπίζει κάθε επιχείρηση είναι ότι οι αγορές είναι απρόβλεπτες. Οποιαδήποτε πρόβλεψη πωλήσεων, όσο αυστηρή και να είναι η ανάλυση των συνθηκών, μπορεί να είναι λάθος. Εάν οι συνθήκες της αγοράς παραμείνουν σχετικά αμετάβλητες, μια αξιόπιστη μέθοδος πρόβλεψης χρησιμοποιεί ιστορικά δεδομένα (Markgraf (2017)).

Η εμπειρία δείχνει ότι τα συνελικτικά νευρωνικά δίκτυα είναι εξαιρετικά στο χειρισμό των ιστορικών δεδομένων και στην αλίευση της εποχικότητας, των τάσεων, των κύκλων και των ακανόνιστων συστατικών, όπως φαίνεται στο σχήμα 7.



Σχήμα 7: Καθημερινή απώλεια

Παραπάνω έγινε περιγραφή μιας προσέγγισης της χρήσης του CNN WaveNet, μιας αλληλουχίας στην αρχιτεκτονική αλληλουχίας, από την άποψη της πρόβλεψης των πωλήσεων, η οποία έδειξε ότι είναι μια πολύ αποτελεσματική μέθοδος (σχήμα 8) για την επίλυση προβλημάτων πρόβλεψης με χρονοσειρές. Εκτός αυτού, περιγράψαμε κάποιες άλλες σύγχρονες τεχνικές που θα μπορούσαν επίσης να χρησιμοποιηθούν ως λύσεις (Kechyn et al. (2018)).



Σχήμα 8: Κινητός μέσος συγκριτικά με προβλέψεις αυτόνομων μοντέλων

2.4.5. Οι ανάγκες πρόβλεψης πωλήσεων για το τμήμα εφοδιασμού

Επειδή είναι ευθύνη του τμήματος εφοδιασμού να μεταφέρει τα προϊόντα που δημιουργεί η παραγωγή στις συγκεκριμένες τοποθεσίες όπου θα απαιτηθούν, το τμήμα χρειάζεται προβλέψεις πωλήσεων στα προϊόντα αποθέματος σύμφωνα με το επίπεδο πρόβλεψης ανά τοποθεσία. Οι ορίζοντες για αυτές τις προβλέψεις είναι επίσης διπλοί: ένας για το μακροπρόθεσμο σχέδιο και ένας για το βραχυπρόθεσμο σχέδιο. Το μακροπρόθεσμο σχέδιο είναι απαραίτητο για την ανάπτυξη των εγκαταστάσεων αποθήκευσης σε διάφορες τοποθεσίες (κατά συνέπεια, προβλέψεις σε μονάδες και κυβικό όγκο) και του εξοπλισμού μεταφοράς για τη μετακίνηση των προϊόντων μεταξύ αυτών των εγκαταστάσεων (επομένως, προβλέψεις σε βάρος). Και πάλι, ο ορίζοντας καθορίζεται από το χρόνο που χρειάζεται για να φέρουμε αυτές τις εγκαταστάσεις σε ετοιμότητα. Μια μεγάλη χημική εταιρεία, για παράδειγμα, χρειάζεται έναν ορίζοντα προγραμματισμού 18 μηνών για τη σύναψη συμβάσεων για την κατασκευή νέων σιδηροδρομικών οχημάτων για να μετακινήσει τα διάφορα προϊόντα της. Έτσι, το μακροπρόθεσμο σχέδιο εφοδιαστικής έχει ως αποτέλεσμα μια πρόβλεψη με ορίζοντα 18 μηνών.

Σε όλες τις εταιρείες, αυτοί οι μακροπρόθεσμοι ορίζοντες μπορεί να κυμαίνονται από μηνιαία, για μισθωμένες εγκαταστάσεις ή συμβατική μεταφορά, έως αρκετά χρόνια για προσαρμοσμένες εγκαταστάσεις ή εξοπλισμό μεταφοράς κατασκευασμένο ειδικά για την εταιρεία. Επειδή και οι δύο χρησιμοποιούνται συχνά, το διάστημα είναι συνήθως μηνιαίο. Επειδή τα σχέδια επηρεάζονται συγκεκριμένα από αυτό το είδος που πρέπει να μετακινηθεί και ποιος είναι ο προορισμός του, το επίπεδο είναι προϊόν κατά τοποθεσία.

Τα βραχυπρόθεσμα σχέδια εφοδιαστικής αφορούν συγκεκριμένες αποφάσεις σχετικά με τα προϊόντα προς μετακίνηση (εκφρασμένα σε μονάδες, όγκο και βάρος) σε ποιες τοποθεσίες και πότε. Έτσι, η πρόβλεψη των πωλήσεων έχει έναν ορίζοντα που καθορίζεται από τον χρόνο κύκλου παραγγελίας από το εργοστάσιο στην εγκατάσταση και επομένως μπορεί να είναι εξαιρετικά βραχυπρόθεσμες (συχνά μηνιαίες, εβδομαδιαίες ή σε μερικές ακραίες περιπτώσεις καθημερινές προβλέψεις). Τα χρονικά διαστήματα για την ενημέρωση αυτών των προβλέψεων είναι επίσης συνήθως στο μηνιαίο ή εβδομαδιαίο (και μερικές φορές ακόμη και καθημερινό) επίπεδο (Mentzer and Moon (2005)).

2.5. Τα εργαλεία διαχείρισης της πρόβλεψης πωλήσεων

Όπως όλες οι σύγχρονες λειτουργίες μάνατζμεντ πρέπει να χρησιμοποιούν τις τελευταίες τεχνολογίες για να διεκπερεώνουν τις εργασίες τους, όπως τα διαθέσιμα πληροφοριακά συστήματα, τις πιο σύγχρονες διαδικασίες και προσεγγίσεις λειτουργιών μάνατζμεντ, και τις μεθόδους μέτρησης και επιβράβευσης των επιδόσεων, έτσι πρέπει να γίνεται και από τη διαχείριση πρόβλεψης πωλήσεων (Mentzer and Moon (2005)). Θα εξετάσουμε εν συντομία καθέναν από αυτούς τους τομείς, παρουσιάζοντας και κάποιες μελέτες περιπτώσεων (case studies).

2.5.1. Τεχνικές προβλέψεων πωλήσεων

Υπάρχει πληθώρα τεχνικών πρόβλεψης, οι οποίες διατίθενται στον διαχειριστή πρόβλεψης πωλήσεων. Στην πραγματικότητα, φαίνεται συχνά ότι υπάρχουν πάρα πολλές τεχνικές και ότι η απόφαση επιλογής μπορεί να συνάδει με την υπερφόρτωση πληροφοριών (σε μία μέτρηση που έγινε το 2005, υπήρχαν πάνω από 70 διαφορετικές τεχνικές χρονοσειρών και μόνο). Ένα τέτοιο σενάριο συχνά αναγκάζει τους υπεύθυνους λήψης αποφάσεων να εγκαταλείψουν κάθε ελπίδα κατανόησης του πλήρους πεδίου των

τεχνικών και να χρησιμοποιούν με συνέπεια μόνο μία ή δύο με τις οποίες είναι εξοικειωμένοι, ανεξάρτητα από το αν οι τεχνικές αυτές είναι κατάλληλες για την περίπτωση ή όχι.

Ευτυχώς, αυτό το σενάριο μπορεί να απλοποιηθεί σημαντικά. Για να κατανοήσουμε τη διαδικασία επιλογής τεχνικών πρόβλεψης πωλήσεων, ο διαχειριστής πρόβλεψης πωλήσεων πρέπει να κατανοήσει τα χαρακτηριστικά ενός σχετικά μικρού συνόλου ομάδων τεχνικών και να καταλάβει σε ποιες καταστάσεις κάθε ομάδα τεχνικών λειτουργεί καλύτερα. Μόλις επιλεγεί η ομάδα τεχνικών, η επιλογή της συγκεκριμένης τεχνικής που χρησιμοποιείται είναι μια πιο απλή απόφαση. Μια απόφαση που μπορεί να επηρεαστεί από μια μεγάλη έρευνα που έχει εξετάσει ποιες τεχνικές χρησιμοποιούνται συχνότερα και πότε λειτουργούν καλύτερα (Mentzer and Kahn (1995)).

Οι κοινές κατηγορίες για τις τεχνικές πρόβλεψης πωλήσεων βασίζονται στο κατά πόσον η τεχνική χρησιμοποιεί υποκειμενική ή στατιστική ανάλυση, ενδογενή δεδομένα (που σημαίνει ότι χρησιμοποιεί μόνο το ιστορικό των πωλήσεων και όχι άλλους παράγοντες που μπορούν να εξηγούν διακυμάνσεις στις πωλήσεις) ή εξωγενή (που σημαίνει τη χρήση άλλων δεδομένων, όπως αλλαγές τιμών ή τακτικές προώθησης ή οικονομικά μέτρα, που μπορούν να εξηγήσουν διακυμάνσεις στις πωλήσεις). Έπειτα αναλύονται αυτά τα δεδομένα, είτε απευθείας από τον άνθρωπο που θα πραγματοποιήσει την πρόβλεψη, είτε απλά εισάγονται σε έναν αλγόριθμο που εκτελεί τον υπολογισμό της πρόβλεψης. Αυτά τα χαρακτηριστικά των τεχνικών πρόβλεψης οδηγούν σε τρεις ευρείες κατηγορίες τεχνικών πρόβλεψης των πωλήσεων:

- τις χρονοσειρές, τόσο τεχνικές σταθερού μοντέλου (fixed-model) όσο και ανοιχτού μοντέλου (open-model),
- την παλινδρόμηση (regression), επίσης αποκαλούμενη συσχέτιση (correlation) και λανθασμένα αποκαλούμενη αιτιώδης τεχνική και
- την κρίσιμη (judgemental), η οποία ονομάζεται επίσης ποιοτική ή υποκειμενική τεχνική (Mentzer and Moon (2005)).

Σύμφωνα με μία μελέτη περίπτωσης, χρειάστηκε να αναθεωρήσουν την τεχνική που χρησιμοποιούσαν στην πρόβλεψη πωλήσεων ενός συγκεκριμένου προϊόντος σε ένα κατάστημα λιανικής πώλησης μεγάλης διανομής, λόγω της ύπαρξης εξωγενών δεδομένων. Για πολλά χρόνια για τον σκοπό αυτό χρησιμοποιήθηκαν στατιστικές μέθοδοι όπως η ARIMA και η Exponential Smoothing. Ωστόσο, αυτές οι στατιστικές μέθοδοι θα μπορούσαν να αποτύχουν σε περίπτωση παρουσίας μεγάλων ακανόνιστων

πωλήσεων, όπως συμβαίνει για παράδειγμα στην περίπτωση προώθησης (promotions), επειδή δεν είναι κατάλληλες για να μοντελοποιήσουν τις μη γραμμικές συμπεριφορές της διαδικασίας πωλήσεων. Τα τελευταία χρόνια χρησιμοποιούνται νέες μέθοδοι βασισμένες στη μηχανική μάθηση για την πραγματοποίηση προβλέψεων. Η προκαταρκτική έρευνα δείχνει ότι οι μέθοδοι που βασίζονται στη Μηχανή Φορέα Υποστήριξης (Support Vector Machine - SVM) είναι πιο ελπιδοφόρες από άλλες μεθόδους μηχανικής μάθησης για την εξεταζόμενη περίπτωση. Η μελέτη αυτή αξιολόγησε την εφαρμογή του SVM στην πρόβλεψη των πωλήσεων υπό το καθεστώς προώθησης, σύγκρινε τη SVM με άλλες στατιστικές μεθόδους και αντιμετώπισε δύο πραγματικές μελέτες περίπτωσης.

Πιο συγκεκριμένα, στο παρελθόν, οι διαχειριστές αυτών των καταστημάτων χρησιμοποιούσαν την εμπειρία τους για να προβλέψουν τις ημερήσιες πωλήσεις και να αποφασίσουν για τις ποσότητες εφοδιασμού. Τα τελευταία χρόνια, με την ανάπτυξη της λήψης αποφάσεων με τη βοήθεια υπολογιστή, η χρήση των μαθηματικών μεθόδων έχει μεταδοθεί ευρέως. Στη δεκαετία του '70 και του '80 οι βασικές μέθοδοι που χρησιμοποιήθηκαν ήταν οι στατιστικές μέθοδοι που βασίστηκαν σε αυτορρυθμιζόμενα μοντέλα χρονοσειρών, όπως η μέθοδος ARIMA, η μέθοδος Box-Jenkins και η εκθετική μέθοδος εξομάλυνσης του Winter (Makridakis et al. (1998)). Αυτές οι μέθοδοι πραγματοποιούν την πρόβλεψη επεξεργάζοντας ως δεδομένα εισόδου, δείγματα από τις ίδιες χρονοσειρές που κάποιος θέλει να προβλέψει. Αν θεωρήσουμε την πρόβλεψη ως την έξοδο της διαδικασίας, μπορούμε να πούμε ότι για αυτές τις μεθόδους η είσοδος και η έξοδος αφορούν τις ίδιες χρονοσειρές.

Στο τέλος της δεκαετίας του '90, αναπτύχθηκε επίσης ένα μαθηματικό μοντέλο διαφορετικό από το Τεχνητό Νευρωνικό Δίκτυο (Artificial Neural Network – ANN) για την ταξινόμηση και την πρόβλεψη, με το όνομα Μηχανή Φορέα Υποστήριξης (SVM) (Cristianini and Shawe-Taylor (2000)). Ο τρόπος ανάλυσης στην SVM προέρχεται από τη Στατιστική Μαθησιακή Θεωρία, ενώ ο αλγόριθμος για την εκπαίδευσή της προέρχεται από τη θεωρία δυαδικότητας του Μαθηματικού Προγραμματισμού. Επίσης, η SVM εκτελεί την πρόβλεψη χρησιμοποιώντας δείγματα των χρονοσειρών που επιθυμεί να προβλέψει, καθώς και δείγματα άλλων γνωρισμάτων.

Τα πολυεπίπεδα ANN, RBF ANN και SVM είναι εργαλεία για μεθόδους μηχανικής μάθησης, που βασίζονται σε μια διαδικασία κατάρτισης που, χρησιμοποιώντας παρωχημένα σύνολα δεδομένων εισόδου και εξόδου, επιτρέπει την πρόβλεψη εξόδων που αντιστοιχούν σε σύνολα δεδομένων εισόδου που δεν χρησιμοποιήθηκαν για την εκπαίδευση του αλγορίθμου. Σε όλες τις περιπτώσεις η διαδικασία εκπαίδευσης

εκτελείται με την επίλυση προβλημάτων μαθηματικής βελτιστοποίησης. Με αυτόν τον τρόπο η μέθοδος μηχανικής μάθησης παρέχει ένα υποκατάστατο μοντέλο ενός σύνθετου άγνωστου φαινομένου.

Σε αυτή τη μελέτη, το περίπλοκο φαινόμενο ανησυχίας είναι πως ο όγκος των πωλήσεων ενός συγκεκριμένου προϊόντος εξαρτάται από διαφορετικά κατάλληλα χαρακτηριστικά εισόδου και ειδικότερα από ένα μη φυσιολογικό χαρακτηριστικό εισόδου, δηλαδή την εμφάνιση των πωλήσεων.

Υπάρχουν πολλές εργασίες στη βιβλιογραφία που ασχολούνται με αυτά τα θέματα. Μία από τις πρώτες εργασίες χρονολογείται στη δεκαετία του '90 (Ansuj et al. (1996)), η οποία έδειξε την υπεροχή της ANN σε σύγκριση με τη μέθοδο ARIMA στην πρόβλεψη των πωλήσεων. Στους Alon et al. (2001) γίνονται αρκετές συγκρίσεις μεταξύ μεθόδων μηχανικής μάθησης και στατιστικών μεθόδων, που δείχνουν από εμπειρικά αποτελέσματα ότι οι μέθοδοι μηχανικής μάθησης έχουν ένα μεγάλο πλεονέκτημα απέναντι στις στατιστικές μεθόδους, ειδικά σε περιόδους ασταθών οικονομικών συνθηκών.

Τα SVM αναπτύχθηκαν στο πλαίσιο της Στατιστικής Μαθησιακής Θεωρίας, αρχικά για την επίλυση των προβλημάτων ταξινόμησης (Burges (1998); Cristianini and Shawe-Taylor (2000)). Τα αποτελέσματα που ελήφθησαν στο πλαίσιο αυτό, δείχνουν ότι η SVM έχει πολύ καλές προοπτικές για τη χρήση της σε άλλους τομείς, υποκινώντας έτσι τη χρήση της SVM στις προβλέψεις.

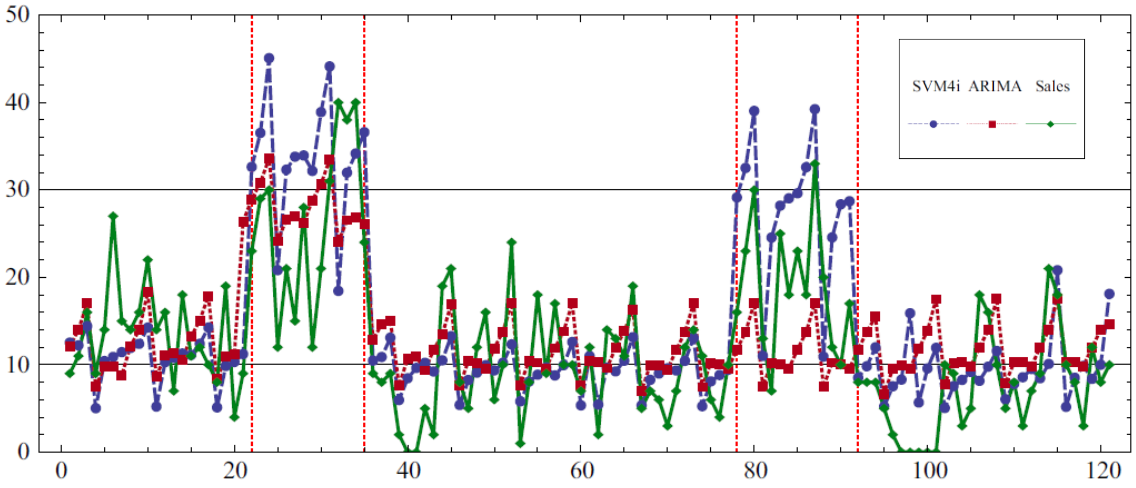
Σε αυτή την ενότητα περιγράφουμε τον τρόπο με τον οποίο χρησιμοποιήθηκε η SVM για την πρόβλεψη των πωλήσεων. Στην εφαρμογή με τα καταστήματα λιανικής πώλησης χρησιμοποιήθηκαν δύο χρονοσειρές εισροών - εκροών που παρέχονται από την ACT Operations Research, που προέρχονται από τις πωλήσεις δύο διαφορετικών καταστημάτων της ίδιας αλυσίδας μεγάλης διανομής. Τα δύο καταστήματα χαρακτηρίζονται από διαφορετικούς όγκους πωλήσεων.

Όσον αφορά την παραγωγή y, μας ενδιαφέρει η καθημερινή πώληση ενός συγκεκριμένου είδους ζυμαρικών ενός δημοφιλούς εμπορικού σήματος, το οποίο μετράται από τον αριθμό των ειδών που πωλούνται. όσον αφορά το διάνυσμα εισόδου x, λαμβάνονται υπ’όψιν εξωγενείς παράγοντες. Συγκεκριμένα, μας ενδιαφέρει να καταγραφούν οι επιδράσεις των πολιτικών προώθησης στις πωλήσεις. Τα δείγματα εισροών - εκροών που χρησιμοποιούνται για την εκπαίδευση και τις δοκιμές καλύπτουν πέντε έτη 2007-2011. Συγκεκριμένα, τα έτη 2007-2010 χρησιμοποιήθηκαν μόνο για την εκπαίδευση και την επαλήθευση, το έτος 2011 για εκπαίδευση, επαλήθευση και δοκιμές με προσέγγιση συρόμενων παραθύρων, όπως περιγράφεται παρακάτω. Οι πρώτες

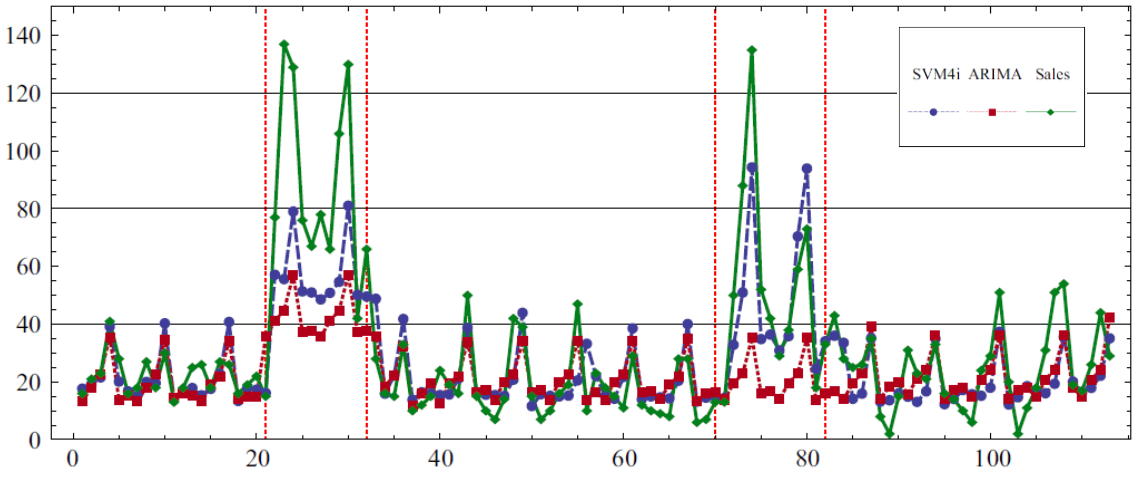
χρονολογικές σειρές λαμβάνονται από το κατάστημα λιανικής πώλησης 1, αυτό με τον μικρότερο όγκο πωλήσεων. Η δεύτερη σειρά προέρχεται από το κατάστημα 2, με τον μεγαλύτερο όγκο πωλήσεων.

Είναι προφανές και όπως φαίνεται σαφώς και στα παρακάτω διαγράμματα (πράσινη γραμμή), ότι υπάρχει αύξηση των πωλήσεων κατά τη διάρκεια των περιόδων προώθησης (τμήματα κόκκινων γραμμών), οι οποίες ήταν 5 κατά τη διάρκεια του έτους. Στο διάγραμμά μας φαίνονται οι δύο από τις πέντε.

Στα παρακάτω σχήματα (σχήμα 9 και 10) φαίνονται τα αποτελέσματα που προέκυψαν από την πρόβλεψη των πωλήσεων κατά τη διάρκεια του 2011 και κάνουμε μια σύγκριση με τις προβλέψεις που παρέχονται από τις παραδοσιακές στατιστικές μεθόδους.



Σχήμα 9: Πραγματικές και προβλεπόμενες πωλήσεις στο κατάστημα # 1 από τον Σεπτέμβριο έως τον Δεκέμβριο του 2011



Σχήμα 10: Πραγματικές και προβλεπόμενες πωλήσεις στο κατάστημα # 2 από τον Σεπτέμβριο έως τον Δεκέμβριο του 2011

Έχουμε περιγράψει πώς μπορεί να εφαρμοστεί η SVM στην πρόβλεψη των πωλήσεων. Η εφαρμογή αφορούσε τις καθημερινές πωλήσεις ενός είδους ζυμαρικών σε δύο καταστήματα λιανικής πώλησης, παρουσία περιόδων προώθησης. Έχουμε επισημάνει τη σημασία μιας κατάλληλης επιλογής των χαρακτηριστικών εισόδου για τη μηχανή μάθησης και της δυνατότητας εξομάλυνσης των διαστασιολογικών ανομοιομορφιών χρησιμοποιώντας την πρόβλεψη ενός χαρακτηριστικού που αντιστοιχεί στα ημερολογιακά γνωρίσματα. Από τα υπολογιστικά αποτελέσματα δείξαμε ότι το SVM παρέχει ένα πολύτιμο εργαλείο για την πρόβλεψη των πωλήσεων, που είναι οι τιμές του μέσου τετραγωνικού σφάλματος (MSE) της πρόβλεψης, οι οποίες είναι μικρότερες από αυτές που παράγονται εφαρμόζοντας τις στατιστικές μεθόδους. Συμπερασματικά, ισχυριζόμαστε ότι οποιοσδήποτε διευθυντής πωλήσεων θα μπορούσε να επωφεληθεί από τη διεύρυνση της κατηγορίας των μεθόδων που χρησιμοποιούνται για την πρόβλεψη των πωλήσεων, ώστε να συμπεριλάβει, μαζί με τις πιο παραδοσιακές στατιστικές μεθόδους και τη SVM που εξετάστηκε σε αυτή την εργασία (Di Pillo et al. (2016)).

Μία ακόμη μελέτη, η οποία ασχολήθηκε με τη χρήση των εξωτερικών παραγόντων στον προγραμματισμό των μοντέλων πρόβλεψης, είναι αυτή της πρόβλεψης πωλήσεων της CPU της Intel Corporation, μιας πολυεθνικής βιομηχανίας ημιαγωγών.

Η μελέτη που παρουσιάζεται εστιάστηκε σε βραχυπρόθεσμες προβλέψεις. Η πρόβλεψη έγινε για τακτικούς λόγους που περιλάμβαναν τον προγραμματισμό της παραγωγής, τη ρύθμιση του στόχου των πωλήσεων, τις βραχυπρόθεσμες απαιτήσεις σε μετρητά και τις προσαρμογές που έπρεπε να γίνουν για τις διακυμάνσεις των πωλήσεων.

Συγκεντρώθηκε η διαθέσιμη πηγή δεδομένων και η απαίτηση πρόβλεψης και συνδυάστηκαν με τη βέλτιστη μεθοδολογία. Σε αυτή τη μελέτη δόθηκαν εβδομαδιαία στοιχεία των πωλήσεων από το 2012. Τα στοιχεία αυτά αντιπροσωπεύουν μια χρονικά ολοκληρωμένη εβδομαδιαία συνολική πώληση της CPU της Intel ανά γραμμή δραστηριότητας. Επιπλέον, δόθηκε η ιστορική μέση τιμή πώλησης, αλλά και τα ιστορικά δεδομένα της κράτησης, τα οποία έδειξαν μια απεικόνιση των κρατήσεων στο παρελθόν. Επίσης ενσωματώθηκαν καινοτομικά στην ποσοτική μοντελοποίηση, η προηγούμενη πώληση, η μελλοντική κράτηση, οι συναλλαγματικές ισοτιμίες, η πρόβλεψη για το ακαθάριστο εγχώριο προϊόν (ΑΕΠ), η εποχικότητα, όπως επίσης και άλλοι δείκτες.

Σε αυτή την μελέτη, έγινε έρευνα και σύγκριση της απόδοσης διαφορετικών τεχνικών στην εβδομαδιαία πρόβλεψη πώλησης CPU της Intel. Δεν αποτελεί έκπληξη

το γεγονός ότι κάθε τύπος μοντέλων έχει τα ιδιαίτερα πλεονεκτήματα και τα μειονεκτήματά του σε σύγκριση με άλλες προσεγγίσεις. Συγκεκριμένα, εξετάστηκαν οι αλγόριθμοι: Extreme Gradient Boosting (XGBoost), Random Forest, Boosting Tree, συναρμολογημένη γραμμική παλινδρόμηση και συναρμολογημένα αυτοτεργασιακά ολοκληρωμένα κινητά μέσα (ensembled autotegrreive integrated moving average models). Τα μοντέλα με μικρότερα σφάλματα επικύρωσης επιλέχθηκαν για να διαμορφώσουν το μοντέλο του συνόλου. Για την καλύτερη καταγραφή των ξεχωριστών χαρακτηριστικών, τα μοντέλα πρόβλεψης εφαρμόστηκαν σε ηγετικό χρόνο και γραμμές επιχειρησιακού επιπέδου. Η διαδικασία επικύρωσης των κινούμενων παραθύρων επιλέγει αυτόματα τα μοντέλα που αντιπροσωπεύουν πιστά την τρέχουσα κατάσταση της αγοράς. Ο εβδομαδιαίος ρυθμός πρόβλεψης επέτρεψε στο μοντέλο να ανταποκριθεί αποτελεσματικά στις διακυμάνσεις της αγοράς. Επίσης αναπτύχθηκε γενικής μεταβλητής σημασίας ανάλυση για να ενισχύσουμε την ερμηνεία του μοντέλου. Αντί να πραγματοποιηθεί υπόθεση σταθερής κατανομής, αυτή η μη παραμετρική ανάλυση μεταβλητής σημασίας μετατόπισης, παρείχε ένα γενικό πλαίσιο σε διάφορες μεθόδους για την αξιολόγηση της μεταβλητής σημασίας. Αυτό το πλαίσιο μεταβλητής σημασίας μπορεί να επεκταθεί περαιτέρω στο πρόβλημα ταξινόμησης τροποποιώντας το μέσο απόλυτο ποσοστό σφάλματος (MAPE) σε λάθος ταξινόμησης. Αυτό το αποτέλεσμα της πρόβλεψης, τώρα βοηθά να διαμορφωθεί μέρος των εισροών που παρέχονται δημόσια και στους επενδυτές, ως κατευθυντήρια γραμμή για το επόμενο τρίμηνο κατά την τριμηνιαία απόδοση της Intel.

Η πρωτοποριακή τεχνική βοήθησε να μειωθούν τα υποκειμενικά σφάλματα και τα σφάλματα μέτρησης, ειδικά στους οικονομικούς δείκτες. Το ευαίσθητο χρονικά πλαίσιο κατάρτισης, επικύρωσης και δοκιμής επέτρεψε στο μοντέλο να αντικατοπτρίζει αποτελεσματικά τις αλλαγές στο περιβάλλον, ενώ με το να αξιολογείται και να αναθεωρείται συνεχώς, διατηρούσε την αξιοπιστία του. Η σύγκριση μεταξύ αυτών των μοντέλων έδειξε μικτά αποτελέσματα. Αυτό υποδηλώνει ότι δεν υπάρχει γενικό συμπέρασμα για το ποια είναι η καλύτερη τεχνική πρόβλεψης για να χρησιμοποιηθεί, αλλά είναι αλήθεια ότι μια πρόβλεψη πρέπει να εξετάζει πολλαπλά μοντέλα, λαμβάνοντας υπ’όψιν τους διαθέσιμους πόρους δεδομένων, τα χαρακτηριστικά της επιχείρησης και την απαίτηση της πρόβλεψης. Ο μεταβλητής σημασίας αλγόριθμος ενίσχυσε την ερμηνεία των μοντέλων του συνόλου και επέτρεψε στους ενδιαφερόμενους να εξετάσουν και να δώσουν έγκαιρα την ανατροφοδότηση. Αυτές οι θετικές πτυχές του μοντέλου μας οδήγησαν στην πρόσφατη υιοθέτηση του (Xu and Sharma (2017)).

2.5.2. Διαχειριστική διαδικασία και προσέγγιση της πρόβλεψης πωλήσεων

Η διαχείριση των προβλέψεων πωλήσεων αφορά τον τρόπο με τον οποίο οργανώνουμε και διεξάγουμε αποδοτικά και αποτελεσματικά τις επιχειρηματικές δραστηριότητες ανάπτυξης και χρησιμοποιούμε τις προβλέψεις πωλήσεων.

Μετά από πολλές έρευνες των Mentzer J.T. και Moon M.A. (2005) σε εκατοντάδες εταιρείες διαπιστώθηκε ότι οι επιχειρήσεις συνήθως οργανώνουν τη λειτουργία της πρόβλεψης πωλήσεων με έναν από τους παρακάτω τέσσερις τρόπους: την *ανεξάρτητη προσέγγιση*, τη *συγκεντρωμένη προσέγγιση*, τη *διαπραγματευόμενη προσέγγιση* και τη *συναίνεση* ως προς τη διαχείριση των προβλέψεων πωλήσεων.

Επιπλέον, διαπιστώσαμε επίσης ότι η αποδοτικότητα και η αποτελεσματικότητα της πρόβλεψης πωλήσεων μιας εταιρείας εξαρτάται από το βαθμό λειτουργικής ολοκλήρωσης που υπάρχει στην εταιρεία. Τα στοιχεία της λειτουργικής ολοκλήρωσης ορίζονται από την επικοινωνία, τον συντονισμός και την συνεργασία. Η επικοινωνία είναι ο γραπτός λόγος, ο προφορικός και η ηλεκτρονική πληροφορία που μοιράζεται μεταξύ των λειτουργικών περιοχών. Ο συντονισμός είναι η επίσημη δομή και απαιτούνται συναντήσεις μεταξύ δύο ή περισσοτέρων λειτουργικών περιοχών. Η συνεργασία είναι ο προσανατολισμός μεταξύ των λειτουργικών περιοχών και μεταξύ της εταιρείας και των βασικών πελατών της, προς τον κοινό καθορισμό στόχων (σε αυτή την περίπτωση, οι κοινοί στόχοι απόδοσης της πρόβλεψης πωλήσεων). Οι τέσσερις διευθυντικές προσεγγίσεις, μαζί με το βαθμό λειτουργικής ολοκλήρωσης που εμπεριέχει κάθε προσέγγιση, θα αναλυθούν παρακάτω σχετικά με τις συνέπειες αυτών των δύο εννοιών στην επιτυχία της πρόβλεψης πωλήσεων μιας εταιρείας.

Οι εταιρείες που χρησιμοποιούν την *ανεξάρτητη προσέγγιση* για την πρόβλεψη των πωλήσεων τείνουν να είναι αρκετά αφελείς στην προσέγγισή τους για την οργάνωση της λειτουργίας πρόβλεψης πωλήσεων. Κάθε λειτουργικό τμήμα της εταιρείας αναπτύσσει μια πρόβλεψη πωλήσεων προσαρμοσμένη στις δικιές της συγκεκριμένες απαιτήσεις. Το πρόβλημα με αυτή την προσέγγιση δεν είναι απαραίτητα ότι κάθε τμήμα αναπτύσσει μια πρόβλεψη με τη μορφή που ταιριάζει στις ιδιαίτερες απαιτήσεις του, αλλά η έλλειψη λειτουργικής ολοκλήρωσης που χαρακτηρίζει αυτή την προσέγγιση.

Η *συγκεντρωμένη* μορφή της πρόβλεψης πωλήσεων αναθέτει την ευθύνη πρόβλεψης σε ένα τμήμα, π.χ. εφοδιασμού ή μάρκετινγκ. Αυτή η διοικητική προσέγγιση καλύπτει τουλάχιστον μερικώς τις πτυχές επικοινωνίας και συντονισμού της λειτουργικής ολοκλήρωσης με μεγαλύτερη αποτελεσματικότητα από την ανεξάρτητη διαχειριστική προσέγγιση. Οι προφορικές, οι γραπτές και ενίοτε οι ηλεκτρονικές επικοινωνίες λαμβάνουν χώρα μεταξύ των διαφόρων χρηστών των

προβλέψεων πωλήσεων που αναπτύσσονται από το υπεύθυνο τμήμα και οι επικοινωνίες αυτές παρέχουν πληροφορίες που μπορούν να ενσωματωθούν στην επίσημη πρόβλεψη. Ωστόσο, αυτή η διαχειριστική προσέγγιση δεν εξετάζει τη πτυχή συνεργασίας της λειτουργικής ολοκλήρωσης, όπως αποδεικνύεται από το γεγονός ότι οι προβλέψεις που αναπτύχθηκαν από το υπεύθυνο τμήμα είναι πολύ προσανατολισμένες στις απαιτήσεις πρόβλεψης και προγραμματισμού αυτού του τμήματος.

Μια εταιρεία που χρησιμοποιεί μια *διαπραγματευμένη προσέγγιση* για τη διαχείριση της διαδικασίας πρόβλεψης των πωλήσεών της, αναπτύσσει τις προβλέψεις πωλήσεων σε κάθε λειτουργικό τμήμα και κατόπιν συγκεντρώνει εκπροσώπους από κάθε τμήμα πριν την περίοδο πρόβλεψης για να διαπραγματευτεί μια επίσημη πρόβλεψη πωλήσεων για κάθε επίπεδο και κάθε κατεύθυνση. Όσον αφορά τη λειτουργική ολοκλήρωση, η προσέγγιση με διαπραγμάτευση ξεπερνά ορισμένα από τα προβλήματα μεροληψίας της συγκεντρωμένης προσέγγισης, ενθαρρύνοντας την επικοινωνία και, ειδικότερα, τον συντονισμό μεταξύ των τμημάτων. Ωστόσο, επειδή κάθε τμήμα αρχικά αναπτύσσει τις δικές του προβλέψεις πωλήσεων για να τις παρουσιάσει στη διαδικασία διαπραγμάτευσης, δεν υπάρχει πραγματική συνεργασία όσον αφορά τη διαδικασία πρόβλεψης, δηλαδή η ανάπτυξη των προβλέψεων πωλήσεων δεν καθοδηγείται από κοινούς στόχους και πληροφορίες αλλά από τους ξεχωριστούς στόχους, τις πληροφορίες και τις απαιτήσεις κάθε επιμέρους τμήματος.

Στη *συναινετική* μορφή του οργανισμού πρόβλεψης πωλήσεων, μια επιτροπή αποτελούμενη από εκπροσώπους του κάθε τμήματος, καθώς και ένα μέλος που ορίζεται ως υπεύθυνο της επιτροπής πρόβλεψης, είναι υπεύθυνοι για την ανάπτυξη προβλέψεων πωλήσεων σύμφωνα με τις ανάγκες όλων των τμημάτων. Μια πραγματική προσέγγιση πρόβλεψης συναίνεσης ενσωματώνει υψηλά επίπεδα πρόβλεψης (της επικοινωνίας, του συντονισμού και της συνεργασίας) ζητώντας από την επιτροπή πρόβλεψης να αναπτύξει μια κοινή πρόβλεψη (η οποία βασίζεται όχι στις μεμονωμένες προβλέψεις διαφόρων τμημάτων, αλλά στην πληροφόρηση από κάθε τμήμα να αναπτύξει μια κοινή πρόβλεψη). Αυτός ο βαθμός λειτουργικής ολοκλήρωσης, μπορεί να βοηθήσει στην υπέρβαση των προκατειλημμένων προβλέψεων που παράγονται από την εστίαση στις μεμονωμένες απαιτήσεις του κάθε τμήματος στην συγκεντρωτική μορφή της πρόβλεψης πωλήσεων. Οι εταιρείες που προτίθενται να ακολουθήσουν αυτή τη διαχειριστική προσέγγιση θα πρέπει να κατανοήσουν ότι είναι ιδιαιτέρως απαιτητική από πλευράς πόρων, τόσο από άποψη χρόνου όσο και προσωπικού. Ωστόσο, εάν μια επιχείρηση διαθέτει τους πόρους για να εφαρμόσει την απαραίτητη λειτουργική

ολοκλήρωση, η συναινετική μορφή οργάνωσης μπορεί να οδηγήσει σε υψηλότερης ποιότητας προβλέψεις πωλήσεων (Mentzer and Moon (2005)).

2.5.3. Μετρήσεις απόδοσης της πρόβλεψης πωλήσεων

Στην ερώτηση, ποια είναι η μέτρηση απόδοσης των προβλέψεων πωλήσεων, η προφανής απάντηση είναι η ακρίβεια. Υπάρχουν όμως πολλά περισσότερα για τις προβλέψεις πωλήσεων από ότι η ακρίβεια. Για παράδειγμα, στην περίπτωση που ένα προϊόν είναι φθηνό και κατέχει μονοπωλιακή θέση στην αγορά του (δηλαδή οι πελάτες δεν μπορούν να το προμηθευτούν από αλλού), ή το κόστος της διατήρησης μεγάλων αποθεμάτων αυτού του προϊόντος είναι χαμηλό ή η πιθανότητα να χάσει τους πελάτες λόγω προσωρινής μη διαθεσιμότητας είναι επίσης χαμηλή, γιατί πρέπει να δαπανήσουμε πολλά χρήματα για την ακριβή πρόβλεψη πωλήσεων αυτού του προϊόντος, όταν δεν υπάρχουν αρνητικές επιπτώσεις στις ανακριβείς προβλέψεις; Αν και αυτό το παράδειγμα είναι μάλλον ακραίο για να είναι αληθινό, παραμένει το γεγονός ότι η επιθυμητή ακρίβεια στην πρόβλεψη των πωλήσεων πρέπει να σταθμιστεί σε σχέση με το κόστος της αλυσίδας εφοδιασμού, τη δυναμικότητα ετήσιων εσόδων και τις επιπτώσεις από την ανακρίβεια όσον αφορά την ικανοποίηση των πελατών (Mentzer and Moon (2005)).

Όπως αναφέρεται σε μία άλλη μελέτη περίπτωσης που πραγματοποιήθηκε για τη Παγκόσμια Ερευνητική Ομάδα Παραγωγής (Global Manufacturing Research Group - GMRG), κατά τη διάρκεια των τελευταίων δεκαετιών, πολλοί ερευνητές επικέντρωσαν τις προσπάθειές τους στην ανάπτυξη νέων μεθόδων και τεχνικών πρόβλεψης με στόχο τη βελτίωση της ακρίβειας της πρόβλεψης (Wright et al. (1986); Armstrong (2001)). Επί χρόνια στο πεδίο των προβλέψεων, ο κύριος στόχος των ερευνητών και των επαγγελματιών ήταν να παρέχουν στις εταιρείες εξελιγμένες ποσοτικές προσεγγίσεις ικανές να μειώσουν τα σφάλματα των προβλέψεων. Αυτό είχε ως αποτέλεσμα να υπάρξει μία βιαστική ανάπτυξη στις νέες μεθόδους πρόβλεψης.

Ωστόσο αρκετές μελέτες βασισμένες σε έρευνες σχετικά με την πραγματική υιοθέτηση ποσοτικών τεχνικών εντός των επιχειρήσεων, αποκαλύπτουν ότι οι ποιοτικές μέθοδοι παραμένουν οι περισσότερο χρησιμοποιούμενες προσεγγίσεις για την επεξεργασία των προβλέψεων. Η ακρίβεια αυτών των προβλέψεων δεν είναι πάντα η υψηλότερη όταν χρησιμοποιούνται σύνθετες τεχνικές και όχι απλές (Mentzer and Cox (1984); Dalrymple (1987); Sanders and Manrodt (1994); Sanders (1997); Lawrence et al. (2000)).

Αυτό το ενδιαφέρον αποτέλεσμα υποδηλώνει ότι οι τεχνικές πρόβλεψης από μόνες τους δεν βελτιώνουν απαραίτητα την ακρίβεια. Οι διευθυντές θα πρέπει επίσης να λάβουν υπ'όψιν και άλλα θέματα που σχετίζονται με τη διαχείριση των διαδικασιών πρόβλεψης (Mentzer and Cox (1984)). Ενστερνίζοντας αυτή την προοπτική, πολλές μελέτες εξετάζουν τις μεταβλητές πρόβλεψης, οι οποίες μπορούν να χρησιμοποιηθούν για τον σχεδιασμό και τη βελτίωση της διαδικασίας πρόβλεψης (Mentzer and Kahn (1997); Mentzer and Bienstock (1998); Chaman (1999); Chase (1999)). Για παράδειγμα, οι συγγραφείς συμφωνούν ότι η συμμετοχή διαφόρων λειτουργικών περιοχών και η διαλειτουργική ολοκλήρωση, είναι κύριες κατά την επεξεργασία πρόβλεψης πωλήσεων (Mentzer and Bienstock (1998); Helms et al. (2000)). Οι εταιρείες θα μπορούσαν να αποφασίσουν να βελτιώσουν τα πληροφοριακά τους συστήματα προκειμένου να διευκολύνουν τη μεταφορά πληροφοριών από μία περιοχή σε άλλη, έτσι ώστε διαφορετικές πηγές πληροφοριών να μπορούν να χρησιμοποιηθούν για την εκπόνηση μίας πρόβλεψης ή ένα κοινό σχέδιο πρόβλεψης να μπορεί να μοιραστεί μέσα στην εταιρεία και να χρησιμοποιηθεί σε διαφορετικές διαδικασίες λήψης αποφάσεων (Mentzer and Kahn (1997); Mentzer and Bienstock (1998); Moon and Mentzer (1998)). Επιπλέον, οι εταιρείες θα μπορούσαν να αναδιοργανώσουν τη διαδικασία πρόβλεψης, προκειμένου να επιτευχθεί μία συντονισμένη διαδικασία προβλέψεων (συναινετική μορφή), αποφεύγοντας με αυτόν τον τρόπο το κάθε τμήμα να αναπτύσσει και να χρησιμοποιεί τη δικιά του (ανεξάρτητη προσέγγιση) (Mentzer and Bienstock (1998); Mentzer et al. (1999); Chaman (2001)).

Η απόφαση για το πώς θα βελτιώσουμε τη διαδικασία πρόβλεψης είναι μία πολύπλοκη εργασία για τις επιχειρήσεις, καθώς η διαδικασία αυτή μπορεί να βελτιωθεί με διαφορετικούς τρόπους και ενεργώντας με διαφορετικούς παράγοντες, που σχετίζονται όχι μόνο με τις τεχνικές που μπορούν να υιοθετηθούν για την εκτέλεση μίας πρόβλεψης, αλλά και με την οργάνωση του συνόλου της διαδικασίας. Ωστόσο, οι επιχειρήσεις έχουν συνήθως περιορισμένους πόρους και δεν μπορούν να αλλάξουν τα πάντα με τη μία. Επομένως, χρειάζονται καθοδήγηση σχετικά με τον τρόπο καθορισμού των προτεραιοτήτων για τη βελτίωση της διαδικασίας πρόβλεψης. Δεδομένου ότι ο σκοπός των εταιρειών είναι να αυξήσουν την απόδοσή τους χάρη σε μια πιο αποτελεσματική διαδικασία πρόβλεψης, η επιχείρηση πρέπει να κατανοήσει ποιες μεταβλητές της διαδικασίας πρόβλεψης σχετίζονται με την απόδοση και εάν μία από αυτές τις μεταβλητές είναι πιο σημαντική από τις άλλες όταν χρησιμοποιηθούν συνδυαστικά. Παρά τη σπουδαιότητα αυτού του ζητήματος, τα συνδυαστικά μοντέλα που βασίζονται στο συστηματικό έλεγχο της σχέσης μεταξύ συγκεκριμένων

μεταβλητών της διαδικασίας πρόβλεψης και των επιδόσεων των επιχειρήσεων, είναι λίγα και παρόμοια.

Επιπλέον, αν μπορούν να δημιουργηθούν και άλλες νέες μεταβλητές, είναι ζωτικής σημασίας για τους διαχειριστές να κατανοήσουν την ύπαρξη «αλληλεπιδραστικών αποτελεσμάτων» μεταξύ των διαφόρων μεταβλητών πρόβλεψης. Η αλληλεπίδραση μπορεί να οριστεί ως το ποσοστό της επίδρασης μιας μεταβλητής σε μια άλλη προς τις τιμές μιας ή περισσοτέρων άλλων μεταβλητών (Gove (1986)). Αυτό σημαίνει ότι ο αντίκτυπος μίας μεταβλητής πρόβλεψης μπορεί να επηρεαστεί από το βαθμό μιας άλλης μεταβλητής πρόβλεψης που θα χρησιμοποιηθεί. Η αλληλεπίδρασή τους, για παράδειγμα, θα μπορούσε να καθορίσει ένα επιπλέον θετικό συνδυαστικό αποτέλεσμα στην απόδοση. Έτσι, οι επιχειρήσεις θα πρέπει να μοχλεύουν ταυτόχρονα και τις δύο αυτές μεταβλητές και να ενθαρρύνουν την αλληλεπίδρασή τους, αντί να επενδύουν και να ενεργούν μόνο στη μία μεταβλητή. Η επιστημονική έρευνα είναι σπάνια σχετικά με τις πιθανές αλληλεπιδράσεις μεταξύ των μεταβλητών πρόβλεψης.

Για την κάλυψη αυτών των κενών σε αυτή την έρευνα διερευνήθηκε:

- ποιες είναι οι σχετικές προγνωστικές μεταβλητές που πρέπει να λαμβάνονται υπ'όψιν όταν αποφασίζουμε πώς να επανασχεδιάζουμε τη διαδικασία πρόβλεψης για να βελτιώσουμε τις επιδόσεις των εταιρειών σε ένα πολυπαραγοντικό περιβάλλον και
- αν κάποιες μεταβλητές πρόβλεψης μπορούν να αλληλοεπιδρούν και να επηρεάσουν την απόδοση των εταιρειών με ένα συνδυαστικό αποτέλεσμα.

Η προς παρουσίαση μελέτη αντιμετωπίζει τα ερευνητικά ερωτήματα μέσω διερευνητικής προσέγγισης. Εξετάζεται η επίδραση ορισμένων μεταβλητών πρόβλεψης και οι αλληλεπιδράσεις τους σε ένα πολυπαραγοντικό περιβάλλον, συμπεριλαμβάνοντας ρητά τους όρους αλληλεπίδρασης στα υπό μελέτη μοντέλα παλινδρόμησης που εκφράζονται ως πολλαπλασιασμός ανεξάρτητων μεταβλητών (Jaccard and Turrisi (2003)). Τα αποτελέσματα των μοντέλων με τις αλληλεπιδράσεις παρέχουν πληρέστερες πληροφορίες σχετικά με την επίδραση των μεταβλητών πρόβλεψης στην απόδοση της επιχείρησης, σε σύγκριση με αυτών που πραγματοποιούν μία μονομερή ανάλυση. Ωστόσο, καθώς αυτή η ανάλυση μπορεί να βοηθήσει στην ερμηνεία της πολυπαραγοντικής ανάλυσης με αλληλεπίδραση, θα αναλυθεί μια σειρά προτάσεων σχετικά με την επίδραση των μεταβλητών πρόβλεψης και τις αλληλεπιδράσεις τους, καθώς και μια σειρά μονοπαραγωγικών υποθέσεων.

Σε αυτή τη μελέτη ελήφθησαν υπ'όψιν διάφοροι τύποι απόδοσης και συγκεκριμένα η πρόβλεψη αποδόσεων της ακρίβειας, κόστους και παράδοσης. Στην πραγματικότητα,

μερικές πρόσφατες μελέτες έχουν δείξει ότι η ακρίβεια της πρόβλεψης δεν είναι η μόνη σχετική απόδοση που πρέπει να μελετηθεί κατά την εξέταση του τρόπου διαχείρισης της διαδικασίας πρόβλεψης (Smaros (2007); Danese and Kalchschmidt (2008)). Μια αποτελεσματική διαχείριση των διαδικασιών πρόβλεψης όχι μόνο μπορεί να συμβάλει στη βελτίωση της ακρίβειας των προβλέψεων και κατά συνέπεια στις αποδόσεις παράδοσης και κόστους, οι οποίες συσχετίζονται συνήθως με το σφάλμα πρόβλεψης, αλλά μπορεί επίσης να έχει άμεσο αντίκτυπο στις αποδόσεις κόστους και παράδοσης.

Στη συγκεκριμένη έρευνα τα δεδομένα για τη δοκιμή των υποθέσεων συλλέχθηκαν από την Παγκόσμια Ερευνητική Ομάδα Παραγωγής (GMRG). Η GMRG συλλέγει πληροφορίες σχετικά με τις πρακτικές κατασκευής που εφαρμόζονται και τις επιδόσεις που επιτελούνται από εταιρείες που δραστηριοποιούνται σε διάφορους τομείς και χώρες σε όλο τον κόσμο.

Τα δεδομένα συλλέχθηκαν κατά τη διάρκεια του χρονικού πλαισίου 2005-2007 από μια διεθνή ομάδα ερευνητών που εργάζονται σε διάφορα πανεπιστήμια σε όλο τον κόσμο. Στο πλαίσιο της ερευνητικής ομάδας, για κάθε χώρα ορίστηκε μία ομάδα ερευνητών και ένας υπεύθυνος για τη συλλογή των δεδομένων. Κάθε ομάδα έπρεπε να διαχειριστεί τα ερωτηματολόγια που έπρεπε να σταλούν στις εταιρείες της εκάστοτε χώρας, είτε με μεμονωμένες επισκέψεις, είτε με αλληλογραφία. Επιπλέον, έπρεπε να βοηθήσουν εκείνες τις εταιρείες που ήρθαν σε επαφή και ενδιαφέρονται να συμμετάσχουν στο ερευνητικό έργο και να διασφαλίσουν ότι οι πληροφορίες που συγκεντρώθηκαν ήταν πλήρεις και ορθές. Τα στοιχεία που χρησιμοποιήθηκαν σε αυτή την έρευνα αποτελούν ένα υποσύνολο της συνολικής έρευνας GMRG.

Συνολικά, 343 επιχειρήσεις επέστρεψαν τα ερωτηματολόγια. Όλες οι εταιρείες ανήκουν στη βιομηχανία κατασκευής και συναρμολόγησης.

Σε αυτό το άρθρο αναλύθηκε ο τρόπος με τον οποίο μπορούν να συμβάλουν στη βελτίωση των επιδόσεων των εταιρειών (π.χ. πρόβλεψη της ακρίβειας, των επιδόσεων κόστους και παράδοσης), σε διαφορετικές συνθήκες πρόβλεψης, δηλαδή σε τεχνικές που υιοθετούνται, σε συνδυασμό για την κατάρτιση μιας πρόβλεψης, . Μια ιδιαίτερα καινοτόμος ιδέα αυτής της μελέτης προέρχεται από την εξέταση του τρόπου αλληλεπίδρασης αυτών των μεταβλητών πρόβλεψης, επηρεάζοντας έτσι τις επιδόσεις των εταιρειών μέσω ενός συνδυαστικού αποτελέσματος.

Η έρευνα αυτή συνέβαλλε στην κατανόηση του αντίκτυπου της πρόβλεψης στις επιδόσεις των εταιρειών με διάφορους τρόπους. Πρώτα από όλα, τα αποτελέσματα της έρευνας βοηθούν στον εντοπισμό ενός καλά καθορισμένου συνόλου προγνωστικών μεταβλητών που πρέπει να τροποποιηθούν, σύμφωνα με την απόδοση του τομέα που

προτίθενται να βελτιώσουν οι εταιρείες. Αρκετές μελέτες σχετικά με το CF υποστηρίζουν ότι η ταυτόχρονη δράση σε όλες τις μεταβλητές πρόβλεψης είναι θεμελιώδης προκειμένου να βελτιωθεί η απόδοση των εταιρειών (Lapide (1999); Helms et al. (2000); Aviv (2001); McCarthy and Golicic (2002)). Σύμφωνα με αυτή την υπόθεση, τα αποτελέσματα δείχνουν ότι, όταν οι εταιρείες προτίθενται να βελτιώσουν τις επιδόσεις κόστους και παράδοσης, θα πρέπει να αφιερώσουν την προσοχή τους σε όλα τα διαφορετικά στοιχεία που χαρακτηρίζουν τον τρόπο διεξαγωγής των προβλέψεων.

Ωστόσο, μια περαιτέρω συμβολή αυτής της έρευνας, η οποία έχει σημαντικές συνέπειες για τους μάνατζερ, ήταν ότι πρότεινε στις εταιρείες τον τρόπο με τον οποίον μπορούν να εξισορροπήσουν τις επενδύσεις τους στην πρόβλεψη, εξετάζοντας πότε είναι συμφέρων να ενεργούν σε δύο ή περισσότερες μεταβλητές ταυτόχρονα, από το να επενδύουν και ενεργούν μόνο σε μία. Αυτό είναι δυνατό χάρη στην ανάλυση των συνδυαστικών αποτελεσμάτων που μπορούν να υπάρχουν μεταξύ των μεταβλητών πρόβλεψης. Για παράδειγμα, αυτή η μελέτη υποστηρίζει τη σημασία της σωστής διαχείρισης πληροφοριών για τη βελτίωση της απόδοσης της επιχείρησης, αλλά αποκαλύπτει επίσης την ύπαρξη συνδυαστικών επιδράσεων μεταξύ της χρήσης διαφορετικών πηγών πληροφοριών και άλλων. Έτσι, οι εταιρείες δεν θα πρέπει να εστιάζουν τις προσπάθειές τους και τις επενδύσεις τους μόνο στη βελτίωση των πληροφοριών που συλλέγονται.

Τέλος, υπογραμμίζεται η ύπαρξη πολύπλοκων συμβιβασμών, οι οποίες πρέπει να ληφθούν υπόψη από τους υπεύθυνους σχεδιασμού κατά τη λήψη αποφάσεων σχετικά με τον τρόπο επανασχεδιασμού της διαδικασίας πρόβλεψης. Εάν αφενός, ο σχεδιασμός μιας επίσημης και δομημένης διαδικασίας βελτιώνει την απόδοση ορισμένων τομέων (π.χ. απόδοση κόστους), αφετέρου, μπορεί να οδηγήσει σε μια διαδικασία λιγότερο αποδοτική, που μπορεί να εμποδίσει την επίτευξη καλής συνολικής απόδοσης (Danese and Kalchschmidt (2010)).

2.6. Ερωτήματα κατά τη διαχείριση της πρόβλεψης πωλήσεων

Κατά το σχεδιασμό μίας πρόβλεψης πωλήσεων δημιουργούνται κάποια ερωτήματα, τα οποία για να απαντηθούν απαιτούν μεγάλη εμπειρία από τους ανθρώπους της εταιρείας και οι απαντήσεις αφορούν ξεχωριστά την κάθε εταιρεία. Οι απαντήσεις σε αυτές τις ερωτήσεις θα σας δείξουν πως πρέπει να λειτουργήσει διαδικασία πρόβλεψης πωλήσεων, ώστε να βοηθήσει αποδοτικά και αποτελεσματικά

την εταιρεία σας να διεξάγει την επιχειρηματική δραστηριότητα ανάπτυξης και χρήσης προβλέψεων πωλήσεων. Οι ερωτήσεις είναι οι εξής:

1. Η πελατειακή βάση είναι στενή ή ευρεία;
2. Χαρακτηριστικά των δεδομένων (αποστολές, πωλήσεις, ζήτηση, ηλικία, λεπτομέρεια, εξωτερικά δεδομένα, ποιότητα);
3. Αριθμός προβλέψεων (ορίζοντες και διαστήματα, προϊόντα, κανάλια, τοποθεσίες);
4. Αριθμός νέων προϊόντων;
5. Γεωγραφικές διαφορές;
6. Εποχικότητα;
7. Εκλέπτυνση του προσωπικού (συστήματα και πρόβλεψη) και συστήματα;
8. Προϋπολογισμός προβλέψεων πωλήσεων;
9. Ακρίβεια που απαιτείται;

Για λόγους συντομίας θα αναλύσουμε μόνο την ερώτηση που αφορά τον αριθμό των νέων προϊόντων και θα παρουσιάσουμε μία μελέτη πρόβλεψης πωλήσεων ενός νέου προϊόντος πριν αυτό κυκλοφορήσει στην αγορά (Mentzer and Moon (2005)).

2.6.1. Νέα προϊόντα

Ο αριθμός των νέων προϊόντων που εισάγονται σε έναν δεδομένο ορίζοντα προγραμματισμού επηρεάζει τον τρόπο με τον οποίο θα προβλέψουμε τις πωλήσεις. Διαπιστώθηκε ότι η πρόβλεψη πραγματικά νέων προϊόντων αναφέρεται από πολλές εταιρείες ως ένα από τα πιο δύσκολα προβλήματα πρόβλεψης που αντιμετωπίζουν. Στην καλύτερη περίπτωση, η πρόβλεψη νέων προϊόντων είναι ένα άλμα στο μέλλον με ελάχιστες ή καθόλου ιστορικές πληροφορίες για να μας οδηγήσουν σε μία πορεία. Για την πρόβλεψη νέων προϊόντων μπορεί να χρειαστεί οι υπάλληλοι της πρόβλεψης πωλήσεων να αφιερώσουν πολλές εργατοώρες, μπορεί να βλάψει την αξιοπιστία της ομάδας πρόβλεψης λόγω μικρής ακρίβειας των προβλέψεων του νέου προϊόντος και επίσης μπορεί να μειώσει το ηθικό της ομάδας πρόβλεψης. Είναι, ωστόσο, μια απαραίτητη λειτουργία σε ένα ανταγωνιστικό περιβάλλον και πρέπει να ενισχυθεί με διάφορες διαδικασίες και τη συμπαράσταση της διοίκησης (Mentzer and Moon (2005)).

Ένα τόσο δύσκολο πρόβλημα είχαν να αντιμετωπίσουν οι επιστήμονες για λογαριασμό μίας τούρκικης εταιρείας που χρειαζόταν μία πρόβλεψη πωλήσεων για ένα νέο προϊόν θερμαντήρα νερού που ήθελε να παράξει.

Το μοντέλο που χρησιμοποιήθηκε είναι το Bass Diffusion, το οποίο είναι ένα κοινό μοντέλο διάχυσης που χρησιμοποιείται για την πρόβλεψη των αρχικών πωλήσεων ενός νέου μακροχρόνιας διάρκειας προϊόντος (ή μίας καινοτομίας νέων προϊόντων). Ο σκοπός αυτού του μοντέλου είναι να δείξει πώς ένα νέο προϊόν (ή μία καινοτομία νέων προϊόντων) υιοθετεί τις πρώτες πωλήσεις του στην κοινωνία με τη βοήθεια μιας μαθηματικής φόρμας. Σύμφωνα με τον Frank Bass, υπάρχουν δύο τύποι πελατών για να πραγματοποιήσουν αυτή την υιοθέτηση. Ο ένας τύπος είναι οι πρωτοπόροι ενώ οι άλλοι μιμητές. Στη μελέτη αυτή, πραγματοποιήθηκαν οι εκτιμήσεις των παραμέτρων χρησιμοποιώντας τα δεδομένα πωλήσεων ενός θερμαντήρα νερού που ελήφθησαν από την TUIK. Στόχος μας είναι να μάθουμε πόσοι πελάτες θα αγοράσουν αυτούς τους θερμαντήρες νερού, πόσοι πελάτες θα υιοθετήσουν αυτό το προϊόν ως καινοτόμοι και πόσοι ως μιμητές, πότε το προϊόν αυτό θα φτάσει στο υψηλότερο ποσοστό πωλήσεων και εάν τα στοιχεία πωλήσεων των θερμαντήρων νερού που παράγονται στην Τουρκία αντικατοπτρίζουν τη διαδικασία διάχυσης.

Στο μοντέλο Bass Diffusion χρησιμοποιείται η μέθοδος ελαχίστων τετραγώνων (OLS), η οποία χρησιμοποιείται συχνά στην οικονομετρία, για την εύρεση των εκτιμήσεων των παραμέτρων. Οι εκτιμήσεις των παραμέτρων πραγματοποιούνται με την τοποθέτηση των συντελεστών που βρέθηκαν με τη βοήθεια της εξίσωσης από τη μέθοδο OLS στις εξισώσεις παραμέτρων που βρέθηκαν από το κλασσικό μοντέλο Bass Diffusion του Frank Bass. Μετά τις ληφθείσες παραμέτρους, εκτιμάται ο χρόνος μέγιστων πωλήσεων και ο μέγιστος αριθμός των πωλήσεων, που βασίζονται και πάλι στο μοντέλο Bass Diffusion. Κατά συνέπεια, το προϊόν θα απαντηθεί από πόσους καταναλωτές ρωτάνε όταν είναι να αγοράσουν.

Μετά την αντικατάσταση των συντελεστών στην εκτιμώμενη εξίσωση με τις εξισώσεις παραμέτρων, εκτιμάται ότι οι θερμαντήρες νερού θα αγοραστούν από συνολικά 9.920.000 άτομα. Ενώ ο αριθμός των πελατών που θα αγοράσουν αυτό το προϊόν ως πρωτοπόροι εκτιμάται ότι ανέρχεται σε 496.000 άτομα, ο αριθμός των ανθρώπων που θα το αγοράσουν ως μιμητές υπολογίστηκε σε 2.380.800 άτομα. Επιπλέον, ο μέγιστος αριθμός πωλήσεων ήταν 860.256 άτομα και ο μέγιστος χρόνος πωλήσεων υπολογίστηκε να είναι περίπου στα 6 χρόνια. Με άλλα λόγια, 6 χρόνια μετά τη διάθεση του προϊόντος θερμαντήρα νερού στην αγορά, θα φτάσει το υψηλότερο ποσοστό πωλήσεων 860.256 τεμάχια. Τέλος, προβλέπεται ότι τα προβλεπόμενα γραφήματα των δεδομένων των πωλήσεων θερμαντήρα νερού αντανακλούν τις διαδικασίες διάχυσης στα γραφήματα της βιβλιογραφία.

Το μοντέλο Bass Diffusion, είναι ένα μοντέλο διάχυσης το οποίο χρησιμοποιείται στο μάρκετινγκ. Το μοντέλο αυτό επικεντρώνεται σε δύο τύπους πελατών: πρωτοπόρους και μιμητές. Αυτό δείχνει πώς οι αρχικές πωλήσεις ενός νέου προϊόντος εξαπλώνονται μέσα στην κοινότητα. Σε αυτό το μοντέλο, υπάρχουν τρεις βασικές παράμετροι που πρέπει να βρεθούν. Η μία αντιπροσωπεύει τους δυνητικούς αγοραστές (m), η άλλη αντιπροσωπεύει τους πρωτοπόρους πελάτες (ρ), ενώ η τελευταία αντιπροσωπεύει τους πελάτες μιμητές (q). Αυτές οι παράμετροι ελήφθησαν από τον Frank Bass, μια εξίσωση που εκτιμήθηκε με τη μέθοδο ελαχίστων τετραγώνων, η οποία χρησιμοποιείται σε πολλούς τομείς. Με τη βοήθεια των παραγόμενων παραμέτρων, προέβλεψε πόσες πωλήσεις θα έχει το νέο προϊόν στο μέλλον (ο αριθμός των μέγιστων πωλήσεων) και πότε θα πραγματοποιηθούν αυτές οι πωλήσεις (μέγιστος χρόνος πωλήσεων). Στη μελέτη μας, την οποία πραγματοποιήσαμε στο τέλος, οι παράμετροι του μοντέλου διάχυσης μπάσων m, ρ και q που πρέπει να εκτιμηθούν είναι περίπου 9.920.000, 0.05 και 0.24, αντίστοιχα. Με άλλα λόγια, ο συνολικός αριθμός δυνητικών αγοραστών που μπορούν να αγοράσουν προϊόντα θερμαντήρων νερού είναι περίπου 9.920.000 άτομα. Μεταξύ του συνολικού αριθμού των αγοραστών, ο αριθμός των πελατών που θα αγοράσει αυτό το προϊόν ως πρωτοπόρο άτομο εκτιμάται ότι είναι 496.000 άτομα, ενώ ο αριθμός των μιμητών υπολογίζεται ότι είναι 2.380.800 άτομα. Ο χρόνος που θα πραγματοποιηθούν οι μέγιστες πωλήσεις εκτιμάται ότι είναι περίπου 6 έτη και ο μέγιστος αριθμός πωλήσεων είναι περίπου 860.256 άτομα. Προβλέπεται ότι οι πωλήσεις του θερμαντήρα νερού θα φτάσουν τις 860.256 στα 6 χρόνια μετά την κυκλοφορία του στην αγορά, ενώ αυτές θα μειώνονται μετά από αυτό το σημείο. Ο πραγματικός αριθμός πωλήσεων για αυτό το προϊόν ήταν τελικά 948.426 άτομα και ο χρόνος των μέγιστων πωλήσεων ήταν στα 5 χρόνια. Διαπιστώνουμε ότι το μοντέλο Bass Diffusion προσφέρει μια καλή προοπτική αναζήτησης για πωλήσεις σε εταιρείες που ασχολούνται με την ανάπτυξη νέων προϊόντων και καινοτομίες νέων προϊόντων (Sevuktekin, Yilmaz and Kara (2018)).

2.7. Πρόβλεψη πωλήσεων και προγραμματισμός – Μία επαναληπτική διαδικασία

Αναπόσπαστο μέρος οποιασδήποτε διαδικασίας πρόβλεψης πωλήσεων είναι η εφαρμογή της επαναληπτικής διαδικασίας πρόβλεψης πωλήσεων και ο προγραμματισμός. Πολλές εταιρείες χρησιμοποιούν το επιχειρηματικό σχέδιο για να πραγματοποιήσουν τις προβλέψεις των πωλήσεων. Αυτή είναι μια αφελής προσέγγιση,

επειδή η πρόβλεψη θα πρέπει να καθοδηγείται από τις πραγματικές συνθήκες της αγοράς και όχι από τις οικονομικές ανάγκες της εταιρείας. Οι πιο εξελιγμένες εταιρείες αναπτύσσουν τις προβλέψεις πωλήσεων ανεξάρτητα από το επιχειρηματικό σχέδιο, αλλά όταν η πρόβλεψη αποκλίνει από το σχέδιο, η πρόβλεψη μεταποιείται έτσι ώστε «να ταιριάζει» με αυτό.

Στην πραγματικότητα, οι εταιρείες που είναι αποτελεσματικές στην πρόβλεψη πωλήσεων και τον επιχειρηματικό προγραμματισμό αρχίζουν με τη διαδικασία της πρόβλεψης πωλήσεων. Ένας ορισμός για την πρόβλεψη πωλήσεων, είναι ο εξής: μια προβολή στο μέλλον της αναμενόμενης ζήτησης, λαμβάνοντας υπόψη ένα καθορισμένο σύνολο περιβαλλοντικών συνθηκών. Δεδομένων των αναμενόμενων οικονομικών και ανταγωνιστικών συνθηκών και των αρχικών σχεδίων μάρκετινγκ, πωλήσεων, παραγωγής, εφοδιασμού, προβάλλουμε τη μελλοντική αναμενόμενη ζήτηση. Υπό αυτή τη βάση, το επιχειρηματικό σχέδιο μπορεί να αναπτυχθεί. Όταν το επιχειρηματικό σχέδιο που προκύπτει δεν ανταποκρίνεται στις οικονομικές ανάγκες και στόχους της εταιρείας, επαναλαμβάνουμε την πρόβλεψη πωλήσεων και εξετάζουμε ποιες πρόσθετες προσπάθειες στο μάρκετινγκ ή και στις πωλήσεις μπορούν να πραγματοποιηθούν ώστε να αυξηθεί η προβλεπόμενη ζήτηση και ποιες επιπλέον προσπάθειες μπορούν να καταβληθούν από την παραγωγή και τον εφοδιασμό για την αύξηση της παραγωγικής ικανότητας στο απαραίτητο επίπεδο για την υλοποίηση του επιχειρηματικού σχεδίου. Η επαναληπτική διαδικασία από την πρόβλεψη πωλήσεων, στο επιχειρηματικό σχέδιο, πίσω πάλι στην πρόβλεψη πωλήσεων, ξανά στο επιχειρηματικό σχέδιο και ούτω καθεξής, είναι αυτή που εξασφαλίζει ένα επιχειρηματικό σχέδιο που βασίζεται στην οικονομική πραγματικότητα και την πραγματικότητα της αγοράς που αντιμετωπίζει η εταιρεία (Mentzer and Moon (2005)).

Η πρόβλεψη πωλήσεων της εταιρείας Rossmann που θα παρουσιάσουμε σε αυτή την εργασία, λαμβάνει υπ’όψιν το σύνολο των περιβαλλοντικών συνθηκών που έχουν τα καταστήματά της, στις οποίες συμπεριλαμβάνονται οι ενέργειες προώθησης, ο ανταγωνισμός, οι μαθητικές και οι εθνικές αργίες.

3. Μεθοδολογία

Αρχικά θα πραγματοποιήσουμε μία γενική περιγραφή της εταιρείας ROSSMANN, της δεύτερης μεγαλύτερης αλυσίδας φαρμακείων στη Γερμανία, που θα αφορά το είδος προϊόντων που εμπορεύεται, τις χώρες που δραστηριοποιείται και τις τακτικές που χρησιμοποιεί για την ανάπτυξή της. Επίσης θα αναλύσουμε τις ανάγκες της, τις αποφάσεις που λαμβάνει και πως θα μπορούσαν να χρησιμοποιηθούν τα δεδομένα από την ανάλυση πρόβλεψης με τη χρήση μηχανικής μάθησης για τη βελτίωση της λειτουργίας της και των αποφάσεων της.

Στη συνέχεια θα μελετήσουμε τον διαγωνισμό που διοργάνωσε η ROSSMANN, στην πλατφόρμα Kaggle, στον οποίο ζητούσε την πρόβλεψη των ημερήσιων πωλήσεων για έξι συνεχείς εβδομάδες. Θα εξετάσουμε τι είδους δεδομένα και μεταβλητές μας δίνονται, ώστε να πραγματοποιήσουμε την πρόβλεψη και θα εξηγήσουμε την κάθε μεταβλητή και τα αρχεία που μας παρέχονται.

Σε αυτόν τον διαγωνισμό έγιναν 3003 υποβολές πρόβλεψης (prediction) για το έπαθλο, των οποίων οι λύσεις δεν ήταν ελεύθερες στο κοινό. Νικητής του διαγωνισμού ήταν ο Gert, του οποίου το σκορ ήταν 0.10021 και με αυτό κατέκτησε τη πρώτη θέση και το μεγαλύτερο χρηματικό έπαθλο. Αντίθετα δόθηκαν στη δημοσιότητα πολλά kernels με διάφορες λύσεις, εκ των οποίων 75 ήταν σε γλώσσα προγραμματισμού R και 68 σε γλώσσα Python και είχαν τουλάχιστον μία ψήφο κοινού. Υπήρχαν και εκατοντάδες άλλα kernels, τα οποία δεν είχαν καμία ψήφο κοινού, οπότε δεν δόθηκε ιδιαίτερη σημασία. Διαβάσαμε αυτές τις λύσεις και καταλήξαμε ότι αυτές που παρουσιάζουν το μεγαλύτερο επιστημονικό ενδιαφέρον και είναι άξιες προς περαιτέρω ανάλυση είναι οι δύο με τις περισσότερες ψήφους κοινού.

Η μία από αυτές τις λύσεις ήταν του Christian Thiele σε γλώσσα προγραμματισμού R και ως αντικείμενο είχε τη διερεύνηση των δεδομένων, χρησιμοποιώντας πίνακες και διάφορες οπτικοποιήσεις. Η δεύτερη ήταν της Elena Petrova, η οποία χρησιμοποίησε την γλώσσα Python και αρχικά παρουσίασε μία μικρή διερεύνηση και επεξεργασία των δεδομένων και έπειτα πραγματοποίησε την πρόβλεψη πωλήσεων με δύο μεθόδους. Η πρώτη μέθοδος ήταν η Prophet, η οποία είναι μία μέθοδος ανάλυσης χρονοσειρών, ενώ η δεύτερη ήταν η XGBoost, η οποία είναι μέθοδος παλινδρόμησης.

Ακολούθως, θα εκτελέσουμε στην πλατφόρμα Kaggle τη λύση του Christian Thiele και θα επεξηγήσουμε τα αποτελέσματα που εξήγαγε. Μελετώντας τη λύση του, με τις αντίστοιχες οπτικοποιήσεις, θα αντιληφθούμε διάφορα στοιχεία για τα δεδομένα του

προβλήματος, όπως τις τάσεις και τις εποχικότητες που ακολουθούν, ώστε να κατανοήσουμε το γενικό επιχειρηματικό πρόβλημα. Κατόπιν, θα δείξουμε την διερεύνηση και επεξεργασία των δεδομένων που έκανε στη δική της λύση η Elena Petrova, ώστε να μπορέσει να τα εισάγει στα μοντέλα πρόβλεψης. Επίσης θα προβάλλουμε από τη λύση της, τη δημιουργία νέων μεταβλητών, τη συσχετίσεις μεταξύ τους και την ανάλυση των χρονοσειρών που ακολουθούν τα δεδομένα. Τέλος, θα εκτελέσουμε στο Kaggle και θα αναπτύξουμε τις δύο μεθόδους που χρησιμοποίησε η Elena Petrova για την πρόβλεψη των ημερήσιων πωλήσεων ενός καταστήματος της ROSSMANN για έξι εβδομάδες, δηλαδή την μέθοδο ανάλυσης χρονοσειρών Prophet και τη μέθοδο παλινδρόμησης XGBoost. Οι μέθοδοι αυτοί θα συγκριθούν και θα επισημανθούν τα κυριότερα πλεονεκτήματα και τα μειονεκτήματά τους.

4. Ανάλυση της μελέτης περίπτωσης: ROSSMANN

4.1. Εισαγωγή

Ήταν το 1972, όταν ο Dirk Rossmann άνοιξε το πρώτο του κατάστημα στο Ανόβερο. Σήμερα η ROSSMANN είναι η δεύτερη μεγαλύτερη αλυσίδα φαρμακείων στην Ομοσπονδιακή Δημοκρατία της Γερμανίας και κατατάσσεται μεταξύ των δέκα σημαντικότερων καταστημάτων λιανικής πώλησης τροφίμων της χώρας. Ενώ κατατάσσεται στην 111η θέση μεταξύ των 250 μεγαλύτερων αλυσίδων λιανικής πώλησης παγκοσμίως. Η έδρα της εταιρείας βρίσκεται πλέον στο Burgwedel κοντά στο Ανόβερο και ο παγκόσμιος όμιλος AS Watson κατέχει το 40% της εταιρείας.

Το μέσο εμβαδόν των υψηλής ποιότητας φαρμακείων της ROSSMANN, είναι 570 τετραγωνικά μέτρα. Τα χρώματα λευκού και κόκκινου που υπάρχουν στο λογότυπο της εταιρείας, κυριαρχούν επίσης τόσο στους εξωτερικούς, όσο και στους εσωτερικούς χώρους. Το έμβλημά της, ο κένταυρος, εμφανίζεται εντός του γράμματος "O". Αυτό το θρυλικό πλάσμα από την ελληνική μυθολογία, συμβολίζει το όνομα του ιδρυτή και ιδιοκτήτη Dirk Rossmann (Rossmann GmbH (n.d.)).

Σχήμα 11: Τα γραφεία της ROSSMANN

4.2. Η επιτυχία της εταιρείας

Από το 2015, έχει εισαχθεί μια νέα εικόνα καταστήματος, με ζεστά χρώματα, νέες σειρές ποικιλιών, ελκυστική καθοδήγηση πελατών και άφθονο έμμεσο φως για μια πιο συναισθηματική προσέγγιση στον πελάτη.

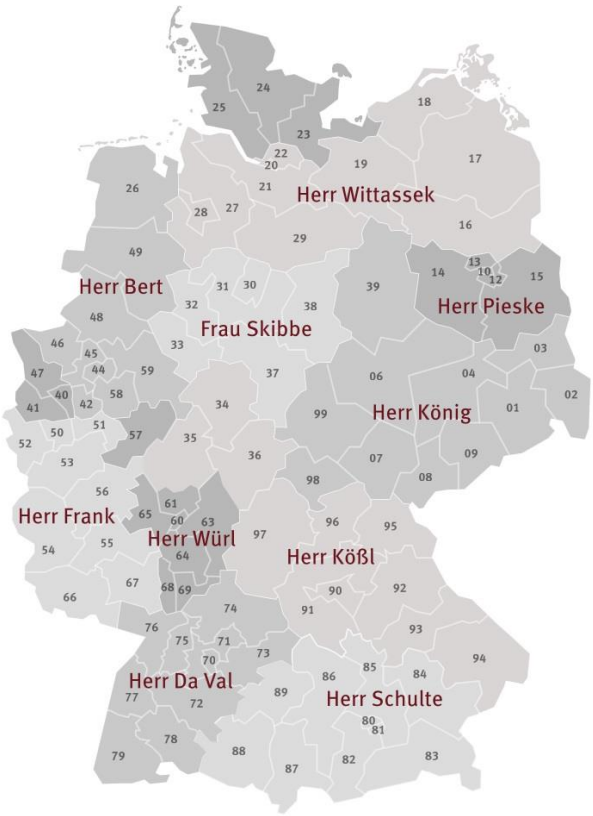
Το 2017, ο όμιλος ROSSMANN είχε 3.790 σημεία πώλησης στη Γερμανία, την Πολωνία, την Ουγγαρία, την Τσεχία, την Αλβανία και την Τουρκία με πωλήσεις ύψους 9 δισεκατομμυρίων ευρώ.

Οι πωλήσεις και η επέκταση των φαρμακείων ROSSMANN παρουσιάζουν σταθερά υψηλή δυναμική εδώ και χρόνια. Με πωλήσεις 9 δισεκατομμύρια ευρώ, η ROSSMANN πέτυχε αύξηση του ομίλου κατά 6,8% το 2017 (προηγούμενο έτος: 8,4 δισ. ευρώ). Οι πωλήσεις στη Γερμανία αυξήθηκαν κατά 4,5% και έφτασαν στα 6,4 δισεκατομμύρια ευρώ (προηγούμενο έτος: 6,12 δισ. ευρώ) μέσω των 2110 φαρμακείων.

Οι πωλήσεις των εταιρειών του ομίλου που βρίσκονται στην Πολωνία, την Ουγγαρία, την Τσεχική Δημοκρατία, την Αλβανία και την Τουρκία αυξήθηκαν κατά 12,9% και άγγιξαν τα 2,6 δισεκατομμύρια ευρώ (προηγούμενο έτος: 2,3 δισ. ευρώ). Έτσι, οι ξένες εταιρείες συνέβαλαν σχεδόν στο 29% των πωλήσεων του Ομίλου. Η ROSSMANN είχε σε λειτουργία εκτός της χώρας της Γερμανίας, συνολικά 1.680 φαρμακεία στις αρχές του 2018.

Μέσα στο 2018 διατηρήθηκε ο υψηλός ρυθμός επέκτασης του ομίλου, ο όγκος των επενδύσεων του οποίου, ανέρχεται σε 210 εκατομμύρια ευρώ. Το σχέδιο ήταν να ανοίξει 230 νέα υποκαταστήματα, εκ των οποίων τα 105 από αυτά να είναι στη Γερμανία.

Ο όμιλος στις αρχές του 2018, προσέφερε εργασία σε 32.000 εργαζόμενους στη Γερμανία, ενώ σε διεθνές επίπεδο, η ROSSMANN ήταν εργοδότης για 54.500 υπαλλήλους. Με τα νέα υποκαταστήματα στη Γερμανία και στο εξωτερικό, φυσικά θα δημιουργούνταν νέες θέσεις εργασίας με πλήρη και μερική απασχόληση (Rossmann GmbH (n.d.)).



Σχήμα 12: Τοποθεσίες καταστημάτων στη Γερμανία

4.3. Τα προϊόντα

Η ROSSMANN είναι πρωτοπόρος εταιρεία στην πώληση φαρμάκων, διότι διαθέτει μία τεράστια ποικιλία προϊόντων, η οποία είναι ιδιαίτερα προσανατολισμένη στον πελάτη. Τα μεγαλύτερα καταστήματα έχουν μία γκάμα προϊόντων που περιλαμβάνει περίπου 21.400 κωδικούς. Το φάσμα των προϊόντων ποικίλλει ανάλογα με το μέγεθος του καταστήματος και την τοποθεσία του σημείου πώλησης.

Εκτός από την κανονική γκάμα φαρμάκων με επίκεντρο τη φροντίδα του δέρματος και του σώματος, τα ποτά, τα βρεφικά είδη, τα απορρυπαντικά, τα καθαριστικά και τα προϊόντα περιποίησης μαλλιών, η εταιρεία έχει προσθέσει στα καταστήματά της επιλεγμένες νέες ιδέες όπως, η φωτογραφική υπηρεσία της ROSSMANN, μια ολοκληρωμένη σειρά με βιολογικά τρόφιμα και κρασιά, ένα τμήμα αρωμάτων με περίπου 400 γνωστά αρώματα, παιχνίδια και χαρτικά, καθώς και εκτεταμένα είδη οικιακών ειδών, τα οποία ολοκληρώνουν το εύρος των προϊόντων ανάλογα με το μέγεθος του υποκαταστήματος. Με αυτές τις νέες προσθήκες προϊόντων και υπηρεσιών, η εταιρεία προσπαθεί να καλύψει πλήρως τις καθημερινές ανάγκες των πελατών της (Rossmann GmbH (n.d.)).

Σχήμα 13: Εσωτερική εικόνα καταστήματος

Σχήμα 14: Φωτογραφική υπηρεσία της ROSSMANN

4.4. Προϊόντα ιδιωτικής ετικέτας

Ο ίδιος ο Dirk Rossmann είχε την ιδέα να αναπτύξει τα δικά του προϊόντα, ειδικά προσαρμοσμένα στις ανάγκες της πελατείας. Όλα ξεκίνησαν το 1997 με τέσσερα προϊόντα ιδιωτικής ετικέτας: το Babydream (προϊόντα για βρέφη), το Facelle (γυναικεία προϊόντα υγιεινής), το Sunozon (αντηλιακό) και το Winston (τροφή για κατοικίδια). Τα προϊόντα αυτά, έδωσαν το έναυσμα για μία επιτυχημένη πορεία.

Σήμερα, η γκάμα έχει αυξηθεί και περιλαμβάνει 29 εμπορικά σήματα της ROSSMANN, από τα enerBiO (βιολογικά τρόφιμα) και τα domol (απορρυπαντικά και καθαριστικά) έως τα altapharma (προϊόντα υγιεινής) και τα ISANA (προϊόντα περιποίησης προσώπου και σώματος), καλύπτοντας όλες τις ανάγκες της καθημερινής ζωής (Rossmann GmbH (n.d.)).

Σχήμα 15: Προϊόντα ιδιωτικής ετικέτας

4.5. Ηλεκτρονικό κατάστημα

Η Rossmann Online GmbH ιδρύθηκε το 1999 και σήμερα απασχολεί περισσότερους από 100 εργαζομένους στον τομέα του εφοδιασμού και της διοίκησης. Για περισσότερα από 10 χρόνια, η Rossmann Online GmbH έχοντας μια ισχυρή γκάμα προϊόντων με περισσότερα από 100.000 είδη, κατείχε ηγετική θέση στο διαδίκτυο όσον αφορά τα προϊόντα φαρμακείων, ομορφιάς, υγείας, οικιακών, τεχνολογίας, ψυχαγωγίας και πολλών άλλων ειδών (Linkedin (n.d.)).

Επιπρόσθετα, η αλυσίδα καταστημάτων Drogerie της Γερμανίας, αποφάσισε τη λειτουργία διαδικτυακού καταστήματος σε συνεργασία με την Amazon. Με αυτή την υπηρεσία, οι πελάτες της Amazon-Prime από το Βερολίνο, μπορούν να έχουν στη διάθεσή τους περίπου 5.000 είδη προϊόντων Rossmann μέσα σε μόνο μια ώρα.

Σύμφωνα με ανακοίνωση της Amazon, τα προϊόντα αυτά μπορεί να είναι είδη οικιακής χρήσης, καλλυντικά, βρεφικά προϊόντα και τρόφιμα.

Η λειτουργία του ηλεκτρονικού καταστήματος γίνεται μέσω της πλατφόρμας της Amazon. Οι παραγγελίες μπορούν να γίνουν από Δευτέρα μέχρι και Σάββατο και ώρες 08:00 έως 00:00. Η ελάχιστη παραγγελία προϊόντων Rossmann είναι στα 20 €, όπως ισχύει και για τις τρέχουσες προσφορές Prime-Now της Amazon. Τα προϊόντα συλλέγονται αμέσως μετά την καταχώρηση της επιθυμητής παραγγελίας και αποστέλλονται στον πελάτη εντός μιας ώρας ή σε προκαθορισμένο χρόνο. Εάν η ανταπόκριση των καταναλωτών είναι ικανοποιητική, δεν αποκλείεται στο μέλλον να επεκταθεί η συγκεκριμένη υπηρεσία και στην πόλη του Μονάχου (Rossmann GmbH (n.d.)).

4.6. Πληροφοριακά συστήματα

Το κλειδί της επιτυχίας για τη Dirk Rossmann GmbH, τη δεύτερη μεγαλύτερη αλυσίδα φαρμακείων στη Γερμανία, είναι η συσχέτιση της γκάμας προϊόντων με τη ζήτηση των καταναλωτών. Για να επιτευχθεί αυτό, ακόμη και σε ώρες αιχμής, η εταιρεία αναβάθμισε την υποδομή των πληροφοριακών συστημάτων για να εξασφαλίσει σταθερά γρήγορους χρόνους απόκρισης. Έχοντας άμεσα ενημερωμένη τη βάση δεδομένων με στοιχεία για τις πωλήσεις των καταστημάτων, οι διαχειριστές μπορούν να λάβουν άμεσα αποφάσεις που ταιριάζουν απόλυτα στους πελάτες της Rossmann.

Για να παραμείνουν ανταγωνιστικοί, η κατανόηση των αναγκών και των προτιμήσεων των καταναλωτών είναι ζωτικής σημασίας. Χωρίς την έγκαιρη πρόσβαση στα δεδομένα πωλήσεων, οι διευθυντές στα φαρμακεία της ROSSMANN, δεν μπορούσαν να προγραμματίσουν αποτελεσματικά το απόθεμά τους και να διατηρήσουν όλο το φάσμα των προϊόντων τους σύμφωνα με τις απαιτήσεις των καταναλωτών, γεγονός που θα μπορούσε να οδηγήσει σε απώλεια πελατείας.

Ο Heike Köhler, επικεφαλής των Κέντρων Δεδομένων της Dirk Rossmann GmbH, εξηγεί ότι ήταν σημαντικό για αυτούς να είναι σε θέση να επεξεργάζονται τα βασικά δεδομένα πωλήσεων αποτελεσματικά, ώστε οι διαχειριστές να λαμβάνουν ενημερωμένες αποφάσεις αγοράς για όλο και μεγαλύτερο αριθμό καταστημάτων. Ενώ αναπτύσσονταν η ROSSMANN γρήγορα, η κακή απόδοση των αναλύσεων επιβράδυνε την επιχειρηματική τους ανταπόκριση και κατέστη σαφές ότι χρειαζόντουσαν την βελτίωση της ταχύτητας και της ικανότητας των δεικτών τους, ώστε να αυξηθεί η παραγωγικότητα και να συνεχίσει η εταιρεία να μεγαλώνει.

Η διαδικασία δημιουργίας αντιγράφων ασφαλείας δεδομένων συχνά υπερφορτωνόταν, πράγμα που σημαίνει ότι τα συστήματα επιβράδυναν τους υπαλλήλους κατά τις εργάσιμες ώρες. Ενώ, οι πωλήσεις, η εφοδιαστική και τα άλλα τμήματα απαιτούσαν την ταχεία διαθεσιμότητα των δυνατοτήτων ανάλυσης ανά πάσα στιγμή.

Αυτό είχε ως αποτέλεσμα η ROSSMANN να χρησιμοποιήσει ένα ERP σύστημα για τη διαχείριση της επιχείρησης, το οποίο εκτελεί το λειτουργικό σύστημα IBM i στους διακομιστές IBM® Power Systems ™ και αποθηκεύει τις συναλλαγές σε ξεχωριστή αποθήκη δεδομένων για εις βάθος αναλύσεις και αναφορές.

Χρησιμοποιώντας την τεχνολογία IBM FlashCopy, επιταχύναν τα αντίγραφα ασφαλείας από 15 ώρες σε μόλις πέντε δευτερόλεπτα, το οποίο είναι περισσότερο από

99,9% ταχύτερο. Τα γρήγορα αντίγραφα ασφαλείας επιτρέπουν την ταχύτερη ανάκτηση και ελαχιστοποιούν τον επιχειρηματικό κίνδυνο.

Χάρη στην υψηλής ταχύτητας επεξεργασία δεδομένων και τις συντομότερες αποθηκεύσεις αντιγράφων ασφαλείας, οι διαχειριστές της ROSSMANN μπορούν τώρα να λαμβάνουν γρήγορα και με ακρίβεια τις αποφάσεις τους. Σε αυτό συμβάλει η ενημερωμένη εικόνα για τα βασικά στοιχεία των πωλήσεων και των αποθεμάτων και ακολούθως για τις προτιμήσεις των καταναλωτών και τις τάσεις της αγοράς ανά πάσα χρονική στιγμή, ακόμη και κατά τις περιόδους αιχμής.

Ο Michael Franke εξηγεί ότι η ROSSMANN εκμεταλλεύεται την υπηρεσία Capacity on Demand που προσφέρει η IBM, η οποία επιτρέπει να ενεργοποιούν τους πρόσθετους πόρους πληροφορικής και μνήμης, κάθε φορά που αυτοί είναι απαραίτητοι, χωρίς διακοπή της λειτουργίας των συστημάτων. Η λύση αυτή παρέχει 50% ταχύτερη πρόσβαση στις αναλύσεις και μικρότερους χρόνους φόρτωσης. Οι διαχειριστές των πωλήσεων και των αγορών μπορούν τώρα να αναλύσουν τις μεγάλες βάσεις δεδομένων και να αποκτήσουν γρηγορότερα νέες προβλέψεις, λαμβάνοντας επιτόπου αποφάσεις που θα οδηγήσουν στη μεγιστοποίηση των πωλήσεων.

Ο Heike Kohler καταλήγει στο συμπέρασμα ότι η λειτουργία της αποθήκης σε DB2 με Linux στο IBM® Power Systems ™ έχει αυξήσει την αποδοτικότητα της πληροφορικής, την παραγωγικότητα του management και μείωσε το κόστος. Οι διαχειριστές της ROSSMANN λαμβάνουν πλέον τις αποφάσεις τους, βάσει άμεσα ενημερωμένων στοιχείων πωλήσεων λιανικής, διασφαλίζοντας ότι η γκάμα των διαθέσιμων προϊόντων ταιριάζει με τις προτιμήσεις των καταναλωτών και μεγιστοποιώντας τα έσοδα από τις πωλήσεις, επιδιώκοντας την ανάπτυξη της εταιρείας (IBM (2015)).

4.7. Η ανάγκη πρόβλεψης πωλήσεων

Η ανάγκη για την πρόβλεψη πωλήσεων για τη ROSSMANN είναι ζωτικής σημασίας και αυτό αποτυπώνεται, όπως είδαμε προηγουμένως, στις μεγάλες επενδύσεις που πραγματοποιεί η εταιρεία στα πληροφοριακά συστήματα. Με την πρόβλεψη πωλήσεων, της δίνεται η δυνατότητα να διαχειρίζεται καλύτερα την εφοδιαστική αλυσίδα, αλλά και να βελτιστοποιεί τον προγραμματισμό των ωραρίων εργασίας του υπαλληλικού προσωπικού, βοηθώντας τους διευθυντές των καταστημάτων να είναι περισσότερο επικεντρωμένοι σε σημαντικότερα ζητήματα, όπως είναι οι πελάτες. Έτσι κατέστη αναγκαίο να πραγματοποιήσει έναν διαγωνισμό μέσα από την πλατφόρμα του

Kaggle, όπου οι διαγωνιζόμενοι, χρησιμοποιώντας πραγματικά δεδομένα της ROSSMANN, θα έπρεπε να παράγουν μία καθημερινή πρόβλεψη πωλήσεων για έξι συνεχόμενες εβδομάδες. Στο επόμενο κεφάλαιο θα περιγράψουμε τα δεδομένα αυτά και θα παρουσιάσουμε μία οπτικοποίησή τους.

5. Διερεύνηση των δεδομένων

Η Rossmann επιθυμούσε να προβλέψει τις καθημερινές πωλήσεις των καταστημάτων της εντός της Γερμανίας, για διάστημα έως έξι εβδομάδων. Οι πωλήσεις των καταστημάτων επηρεάζονται από πολλούς παράγοντες, όπως οι προσφορές, ο ανταγωνισμός, οι σχολικές και κρατικές διακοπές, η εποχικότητα και η τοποθεσία. Με χιλιάδες μεμονωμένους διευθυντές που προβλέπουν πωλήσεις βάσει των ιδιαίτερων περιστάσεων τους, η ακρίβεια των αποτελεσμάτων μπορεί να είναι αρκετά διαφορετική.

Αυτό είχε ως αποτέλεσμα η εταιρεία να διοργανώσει έναν διαγωνισμό στην πλατφόρμα του Kaggle, όπου προέτρεπε την πρόβλεψη έξι εβδομάδων ημερήσιων πωλήσεων για 1115 καταστήματα που βρίσκονται σε όλη τη Γερμανία.

Οι υποβολές των διαγωνιζόμενων αξιολογήθηκαν με το Root Mean Square Percentage Error (RMSPE). Το RMSPE υπολογίζεται ως:

$$\mathrm{RMSPE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right)^2},$$

όπου y_i υποδηλώνει τις πωλήσεις ενός μόνο καταστήματος σε μία μόνο ημέρα και το $\hat{y}_i$ υποδηλώνει την αντίστοιχη πρόβλεψη. Οποιαδήποτε ημέρα και αποθήκευση με μηδενικές (0) πωλήσεις αγνοείται κατά τη βαθμολόγηση.

Ο διαγωνιζόμενος με την καλύτερη πρόβλεψη θα κέρδιζε το πρώτο βραβείο, το οποίο ήταν $ 15000, ο δεύτερος θα έπαιρνε $ 10000, ενώ η τρίτη θέση κέρδιζε $ 5000. Επίσης η ROSSMANN ενδιαφερόταν να προσλάβει τους καλύτερους του διαγωνισμού, οπότε όποιοι επιθυμούσαν δήλωναν τα στοιχεία τους για μία μελλοντική συνεργασία με την εταιρεία (Kaggle Rossmann (2015)).

5.1. Πλατφόρμα Kaggle

Το Kaggle είναι μια διαδικτυακή κοινότητα επιστημόνων που αναλύουν δεδομένα και ασχολούνται με τη μηχανική μάθηση. Δίνει τη δυνατότητα στους χρήστες να βρίσκουν και να δημοσιεύουν σύνολα δεδομένων, να διερευνούν και να δημιουργούν μοντέλα σε ένα διαδικτυακό περιβάλλον Data Science. Όπου το Data Science, είναι ένα διεπιστημονικό πεδίο, το οποίο έχει ως αντικείμενο την εξαγωγή σημαντικών πληροφοριών από αδόμητα ή δομημένα δεδομένα. Η πλατφόρμα του Kaggle δίνει επίσης την ευκαιρία στους χρήστες της, να συνεργάζονται είτε με άλλους επιστήμονες δεδομένων, είτε με προγραμματιστές μηχανικής μάθησης και να συμμετέχουν σε

διαγωνισμούς για την επίλυση προβλημάτων που απαιτούν γνώσεις της επιστήμης ανάλυσης δεδομένων. Στις 8 Μαρτίου 2017 ανακοινώθηκε η εξαγορά του Kaggle από την Google.

Το Kaggle αρχικά ξεκίνησε προσφέροντας μόνο διαγωνισμούς με αντικείμενο τη μηχανική μάθηση, ενώ τώρα προσφέρει επιπλέον σύνολα δεδομένων σε κοινή χρήση, διαδικτυακές (cloud) βιβλιοθήκες εργαλείων που χρησιμοποιούνται στην επιστήμη των δεδομένων και κάποια σύντομα μαθήματα εκπαίδευσης για διάφορους τομείς της επιστήμης. Μία από τις παροχές του Kaggle είναι και τα kernels, τα οποία είναι αποσπάσματα κώδικα και ανήκουν στις διαδικτυακές (cloud) βιβλιοθήκες εργαλείων που αφορούν τη μηχανική μάθηση και γενικά την επιστήμη δεδομένων. Τα kernels επιτρέπουν τους επιστήμονες να ανταλλάσσουν κώδικες και αναλύσεις σε γλώσσες προγραμματισμού Python και R. Έχουν μοιραστεί περισσότερα από 150.000 kernels τα οποία καλύπτουν ένα μεγάλο εύρος πεδίου, από ανάλυση συναισθημάτων έως ανίχνευση αντικειμένων (Wikipedia (n.d)).

5.2. Περιγραφή των δεδομένων

Δόθηκαν ιστορικά στοιχεία πωλήσεων για 1.115 καταστήματα Rossmann. Ο στόχος γίνει πρόβλεψη της στήλης "Πωλήσεις" για το σετ δοκιμών. Να σημειωθεί ότι ορισμένα καταστήματα στο σύνολο δεδομένων έκλεισαν προσωρινά για ανακαίνιση. Τα αρχεία που υπήρχαν ήταν τα εξής:

train.csv: ιστορικά δεδομένα συμπεριλαμβανομένων των πωλήσεων

test.csv: ιστορικά δεδομένα εκτός των πωλήσεων

sample_submission.csv: ένα δείγμα αρχείο υποβολής με τη σωστή μορφή

store.csv: συμπληρωματικές πληροφορίες σχετικά με τα καταστήματα

Παρακάτω παρουσιάζουμε ένα μικρό δείγμα με τις πρώτες πέντε εγγραφές αυτών των αρχείων.

Πίνακας 2: Δείγμα του αρχείου train.csv

Store,"DayOfWeek","Date","Sales","Customers","Open","Promo","StateHoliday","SchoolHoliday"
1,5,2015-07-31,5263,555,1,1,"0","1"
2,5,2015-07-31,6064,625,1,1,"0","1"
3,5,2015-07-31,8314,821,1,1,"0","1"
4,5,2015-07-31,13995,1498,1,1,"0","1"
5,5,2015-07-31,4822,559,1,1,"0","1"

Πίνακας 3: Δείγμα του αρχείου test.csv

Id,"Store","DayOfWeek","Date","Open","Promo","StateHoliday","SchoolHoliday"
1,1,4,2015-09-17,1,1,"0","0"
2,3,4,2015-09-17,1,1,"0","0"
3,7,4,2015-09-17,1,1,"0","0"
4,8,4,2015-09-17,1,1,"0","0"
5,9,4,2015-09-17,1,1,"0","0"

Πίνακας 4: Δείγμα του αρχείου sample_submission.csv

Id,"Sales"
1,
2,
3,
4,
5,

Πίνακας 5: Δείγμα του αρχείου store.csv

Store,"StoreType","Assortment","CompetitionDistance","CompetitionOpenSinceMonth", "CompetitionOpenSinceYear","Promo2","Promo2SinceWeek","Promo2SinceYear","PromoInterval"
1,"c","a",1270,9,2008,0,,,""
2,"a","a",570,11,2007,1,13,2010,"Jan,Apr,Jul,Oct"
3,"a","a",14130,12,2006,1,14,2011,"Jan,Apr,Jul,Oct"
4,"c","c",620,9,2009,0,,,""
5,"a","a",29910,4,2015,0,,,""

Τα περισσότερα από τα πεδία είναι αυτονόητα, παρακάτω είναι περιγραφές για εκείνα που δεν είναι.

Id: ένα αναγνωριστικό που αντιπροσωπεύει μια διπλή ιδιότητα (Κατάστημα, Ημερομηνία) μέσα στο σετ δοκιμών.

Store: ένα μοναδικό Id για κάθε κατάστημα.

Sales: ο κύκλος εργασιών για οποιαδήποτε δεδομένη ημέρα (είναι αυτό που στη συνέχεια θα προβλεφθεί).

Customers: ο αριθμός των πελατών σε μια δεδομένη ημέρα.

Open: μια ένδειξη για το αν το κατάστημα ήταν ανοιχτό: 0 = κλειστό, 1 = ανοιχτό.

StateHoliday: δηλώνει κρατικές αργίες. Κανονικά όλα τα καταστήματα, με λίγες εξαιρέσεις, είναι κλειστά κατά τις κρατικές αργίες. Σημειώστε ότι όλα τα σχολεία είναι κλειστά κατά τις αργίες και τα σαββατοκύριακα. a = δημόσιες αργίες, b = διακοπές για το Πάσχα, c = Χριστούγεννα, 0 = κανένα.

SchoolHoliday: δηλώνει εάν το (Κατάστημα, Ημερομηνία) επηρεάστηκε από το κλείσιμο των δημόσιων σχολείων.

StoreType: διαφοροποιείται μεταξύ 4 διαφορετικών τύπων καταστημάτων: a, b, c, d.

Assortment: περιγράφει το επίπεδο του εύρους προϊόντων: a = βασικό, b = με επιπλέον προϊόντα, c = εκτεταμένο.

CompetitionDistance: απόσταση σε μέτρα από το πλησιέστερο κατάστημα ανταγωνιστών.

CompetitionOpenSince [Month/Year]: δίνει τον κατά προσέγγιση χρόνο και μήνα από τη στιγμή που άνοιξε ο πλησιέστερος ανταγωνιστής.

Promo: δηλώνει εάν ένα κατάστημα εκτελεί μία προωθητική ενέργεια εκείνη την ημέρα.

Promo2: Το Promo2 είναι η συνεχής και διαδοχική προωθητική ενέργεια για μερικά καταστήματα: 0 = το κατάστημα δεν συμμετέχει, 1 = το κατάστημα συμμετέχει.

Promo2Since [Year/Week]: περιγράφει το έτος και την εβδομάδα ημερολογίου, όταν το κατάστημα άρχισε να συμμετέχει στο Promo2.

PromoInterval: περιγράφει τα διαδοχικά διαστήματα που ξεκινάει το Promo2, αναφέροντας τους μήνες που ξεκίνησε η προώθηση, π.χ. "Feb, May, Aug, Nov" σημαίνει ότι κάθε γύρος ξεκινάει τον Φεβρουάριο, τον Μάιο, τον Αύγουστο, τον Νοέμβριο κάθε έτους για το κατάστημα (Kaggle Rossmann (2015)).

5.3. Ανάλυση και οπτικοποίηση των δεδομένων

Στον διαγωνισμό που αναρτήθηκε στο Kaggle για την παρούσα πρόβλεψη, κάποιοι διαγωνιζόμενοι, ανέλυσαν διεξοδικά τα δεδομένα (exploratory analysis) για να εξετάσουν λεπτομερώς και να κατανοήσουν καλύτερα το πρόβλημα. Μία αξιόλογη και λεπτομερή ανάλυση είναι και αυτή του Christian Thiele στις 8 Οκτωβρίου 2015, η οποία ήταν πρώτη στις ψήφους του κοινού, με περισσότερες από τριακόσιους ψήφους (Thie1e (2015)). Ο αλγόριθμος που χρησιμοποιήθηκε για αυτή την ανάλυση από τον Christian Thiele, είναι σε γλώσσα προγραμματισμού R και βρίσκεται στο Παράρτημα Α. Δεν αναλύθηκε περαιτέρω στην παρούσα εργασία, διότι ως γλώσσα προγραμματισμού προς εκμάθηση και εξοικείωση είχαμε την Python.

Ξεκινώντας, εκτελώντας τον αλγόριθμό του, πραγματοποιεί μία καταμέτρηση των δεδομένων που υπάρχουν στα αρχεία test.csv και train.csv και διαπιστώνει ότι το test.csv αποτελείται από 41088 σειρές, ενώ το train.csv από 1017209 σειρές. Στο test.csv το public leaderboard βασίζεται στο 39% των δεδομένων (16024 σειρές) και το

private leaderboard στο 61% (25064 σειρές). Όλα τα καταστήματα που είναι στα test δεδομένα, ανήκουν και στα train δεδομένα. Όμως 259 καταστήματα των train δεδομένων δεν εμπεριέχονται στα test δεδομένα. Συνοπτικά τα στοιχεία παρουσιάζονται στον παρακάτω πίνακα.

Πίνακας 6: Μέγεθος των αρχείων test και train

Αρχείο	Σειρές	Leaderboard		Αριθμός Καταστημάτων
		Public	Private	
test.csv	41088	16024 (39%)	25064 (61%)	856
train.csv	1017209	1017209 (100%)	0	1115

Στην αναζήτηση για ελλιπή δεδομένα, βρέθηκε ότι το κατάστημα 622 έχει 11 μη καταχωρημένες τιμές στη στήλη ‘Open’, αλλά οι υπόλοιπες είναι εντάξει. Όπως οριζόταν και από τους όρους του διαγωνισμού, οι μη καταχωρημένες τιμές θα πάρουν την τιμή ‘1’, δηλαδή ότι το κατάστημα είναι ανοιχτό. Επίσης, όπως παρατηρούμε και στον πίνακα 3, από τα δεδομένα του test λείπει ολόκληρη η στήλη ‘Customers’, αφού αυτά γνωστοποιούνται εκ των υστέρων.

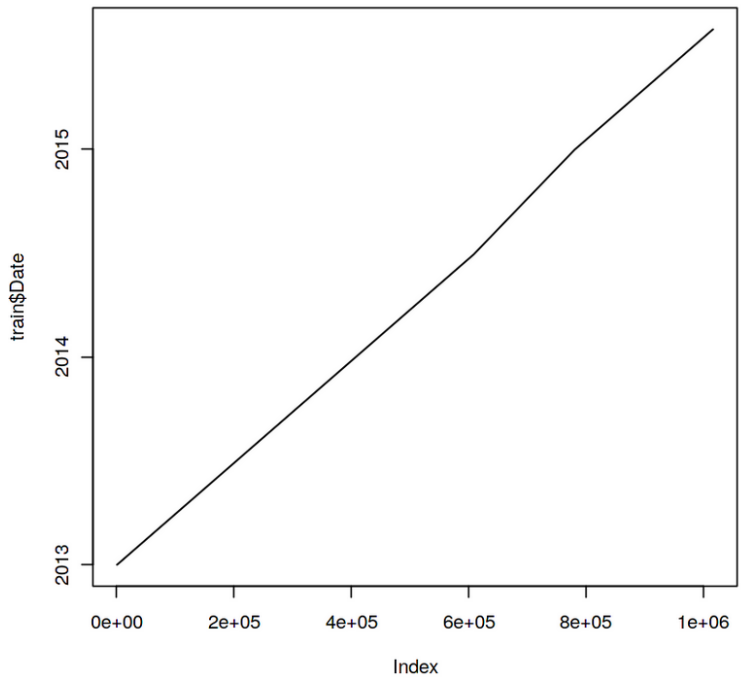
Ένα ακόμη σημείο άξιο αναφοράς είναι, ότι κατά τη διάρκεια της περιόδου που δίνεται στο test, δεν υπάρχουν Πασχαλινές ή Χριστουγεννιάτικες αργίες. Πάραυτα, είναι ενδιαφέρον ότι κατά τη διάρκεια ενός πλήρους έτους, οι σχολικές αργίες αντιστοιχούν σε ένα αρκετά μεγάλο ποσοστό της τάξεως του 44%, ενώ στα δεδομένα train το ποσοστό αυτό ανέρχεται μόλις στο 18%.

Παρακάτω παρουσιάζονται κάποια ποσοστά των τιμών μετά από μία ανάλυση των δεδομένων:

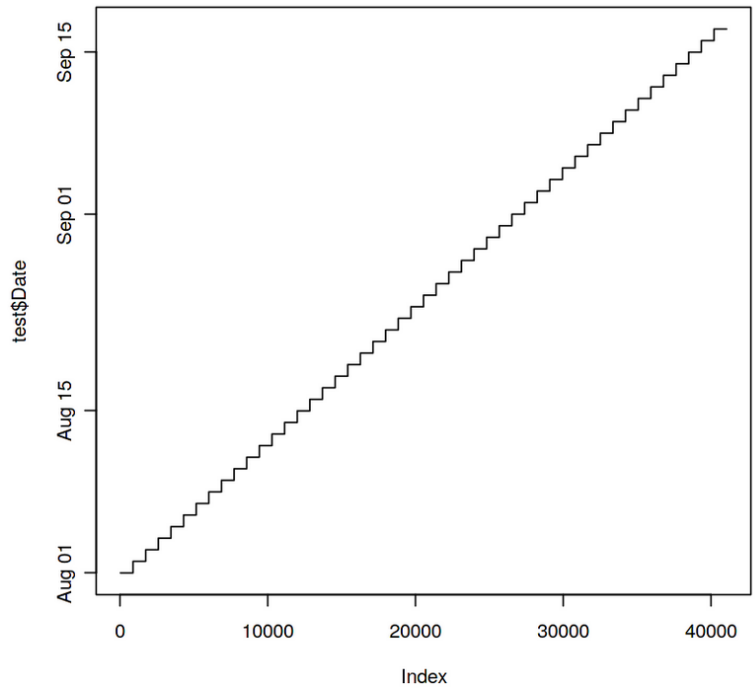
- Ποσοστό ανοιχτών καταστημάτων (train.csv): 83.01 %
- Ποσοστό κλειστών καταστημάτων (train.csv): 16.99 %
- Ποσοστό ανοιχτών καταστημάτων (test.csv): 85.44 %
- Ποσοστό κλειστών καταστημάτων (test.csv): 14.56 %
- Ποσοστό καταστημάτων με προωθητική ενέργεια (train.csv): 38.15 %
- Ποσοστό καταστημάτων χωρίς προωθητική ενέργεια (train.csv): 61.85 %
- Ποσοστό καταστημάτων με προωθητική ενέργεια (test.csv): 39.58 %
- Ποσοστό καταστημάτων χωρίς προωθητική ενέργεια (test.csv): 60.42 %
- Ποσοστό τύπου αργίας (train.csv): 96.95 % καμία, 1.99 % γενική αργία, 0,66 % Πασχαλινή αργία και 0.40 % Χριστουγεννιάτικη αργία

- Ποσοστό τύπου αργίας (test.csv): 0.44 % Χριστουγεννιάτικη ή Πασχαλινή αργία, 99.56 % οποιαδήποτε άλλη αργία
- Ποσοστό σχολικών αργιών (train.csv): 82.14 % δίχως σχολική αργία, 17.86 % με σχολική αργία
- Ποσοστό σχολικών αργιών (test.csv): 55.65 % δίχως σχολική αργία, 44.35 % με σχολική αργία

Από τα δύο επόμενα σχήματα διαπιστώνουμε ότι δεν υπήρχε κάποια ασυνέχεια στα δεδομένα του train και του test αρχείου. Η περίοδος των δεδομένων του αρχείου train είναι από 01/01/2013 έως 31/07/2015, ενώ η περίοδος του test είναι από 01/08/2015 έως 17/09/2015, άρα η πρόβλεψη αφορά 48 ημέρες.

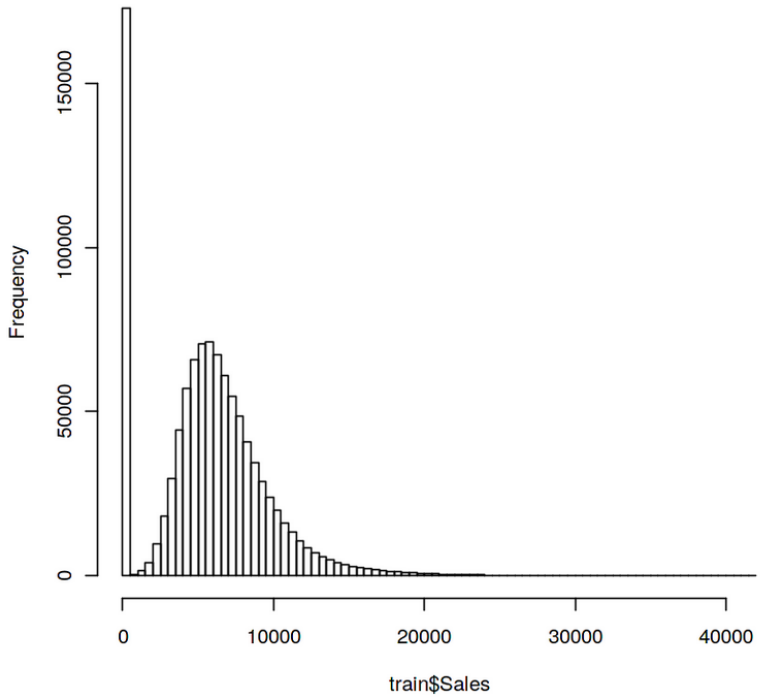


Σχήμα 16: Έλεγχος συνέχειας των δεδομένων του αρχείου train

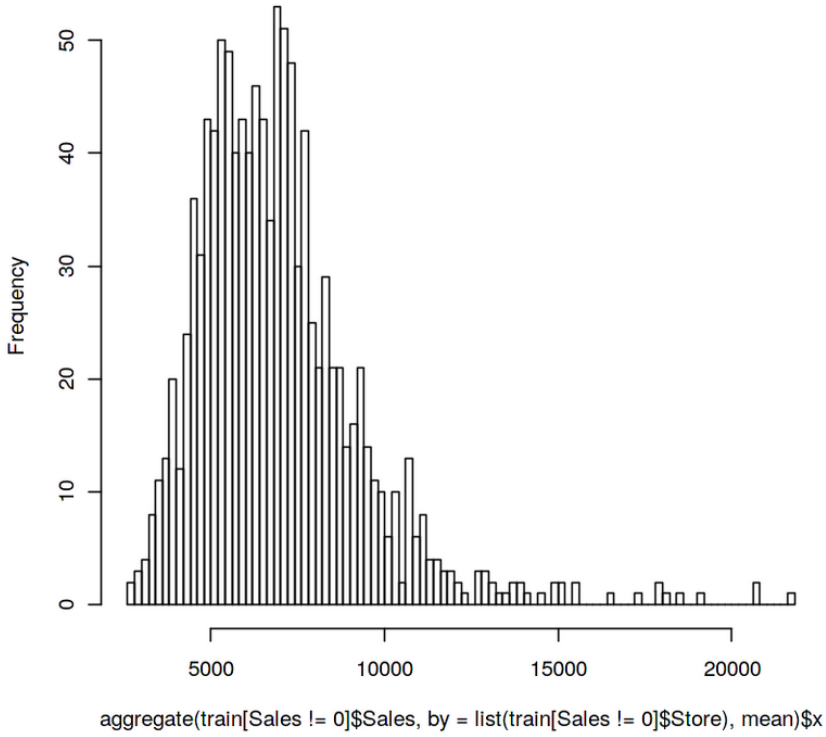


Σχήμα 17: Έλεγχος συνέχειας των δεδομένων του αρχείου test

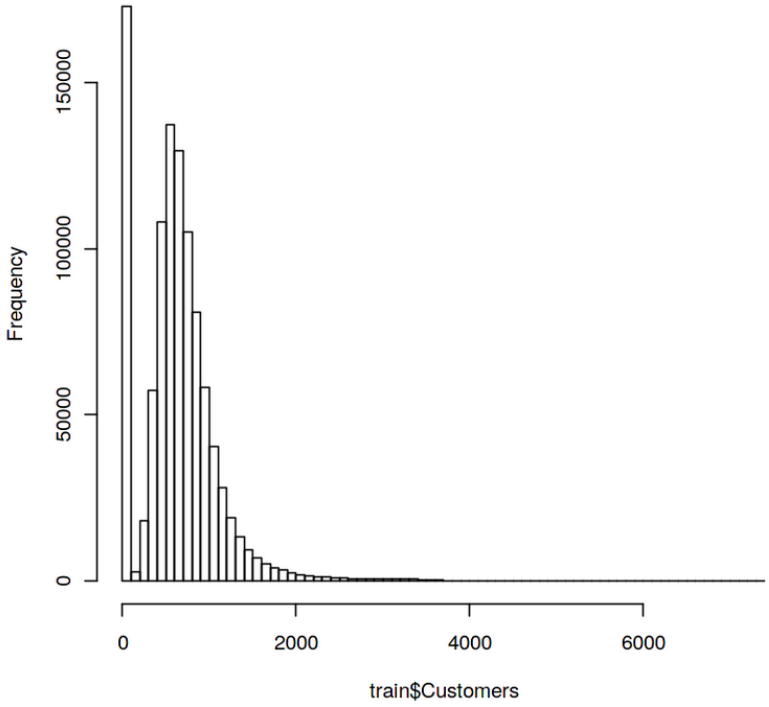
Στη συνέχεια παρουσιάζονται οπτικά τα δεδομένα ανά στήλη του train αρχείου σε ιστογράμματα.



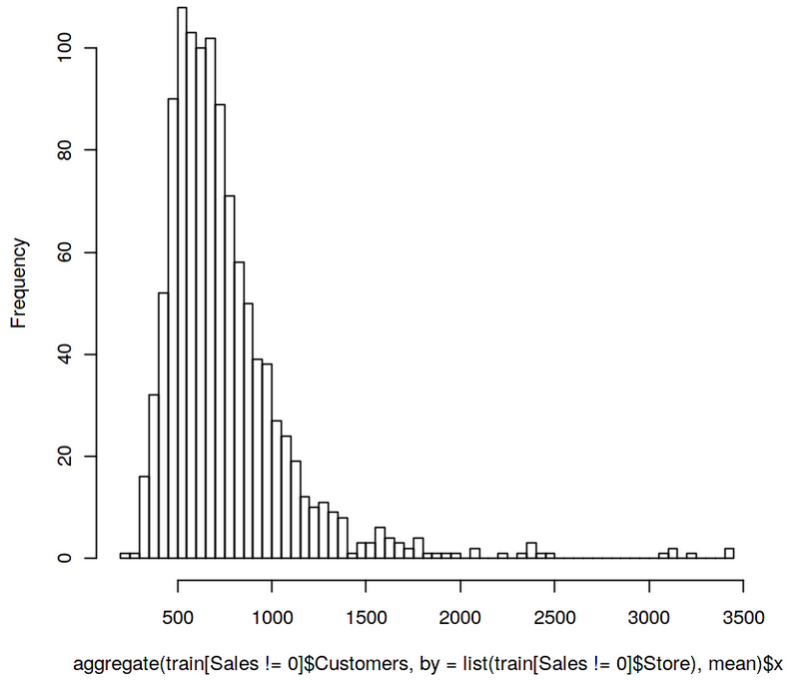
Σχήμα 18: Ιστόγραμμα συχνότητας ύψους πωλήσεων (train.csv)



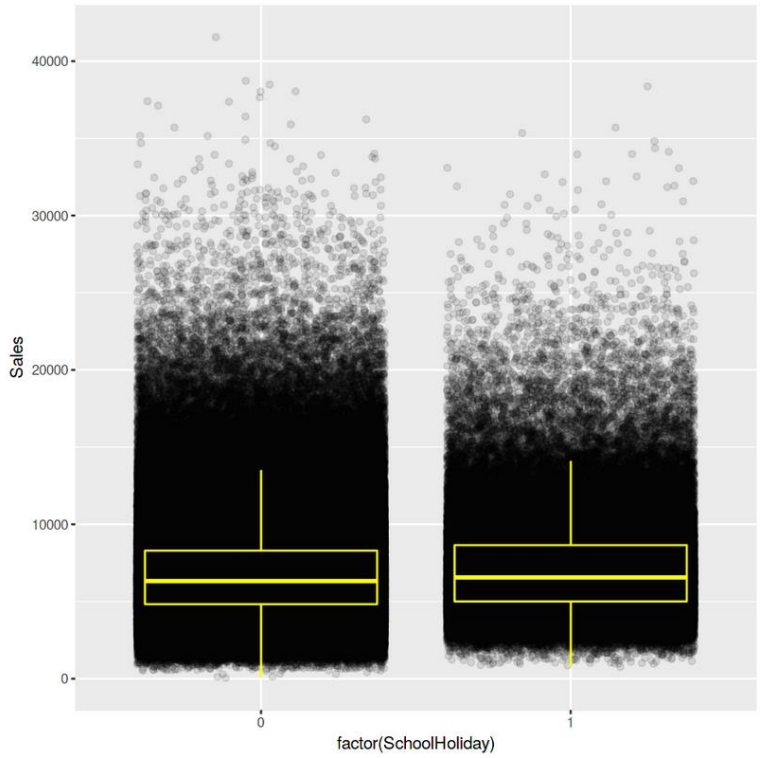
Σχήμα 19: Ιστόγραμμα συχνότητας μέσου όρου πωλήσεων ανά κατάστημα όταν δεν ήταν κλειστό (train.csv)



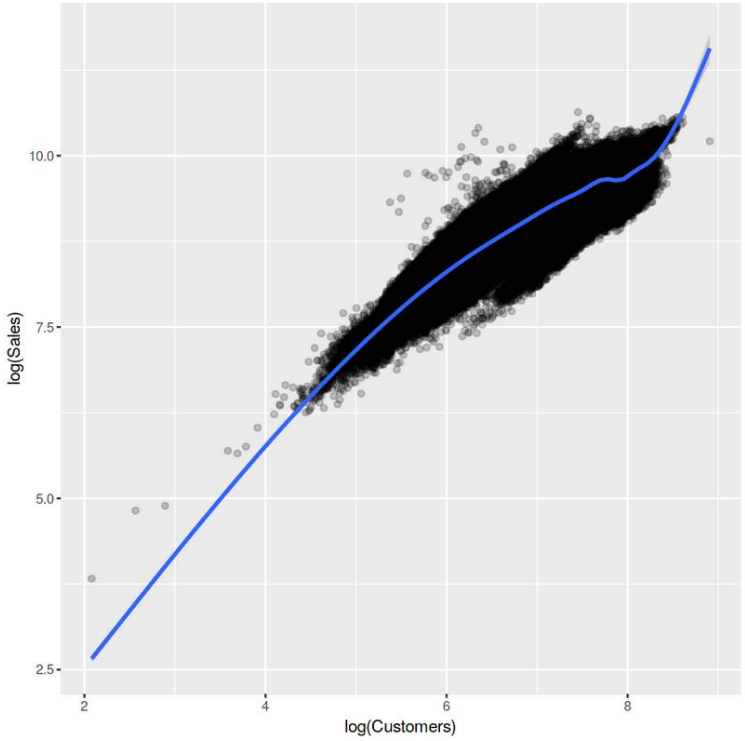
Σχήμα 20: Ιστόγραμμα συχνότητας αριθμού πελατών (train.csv)



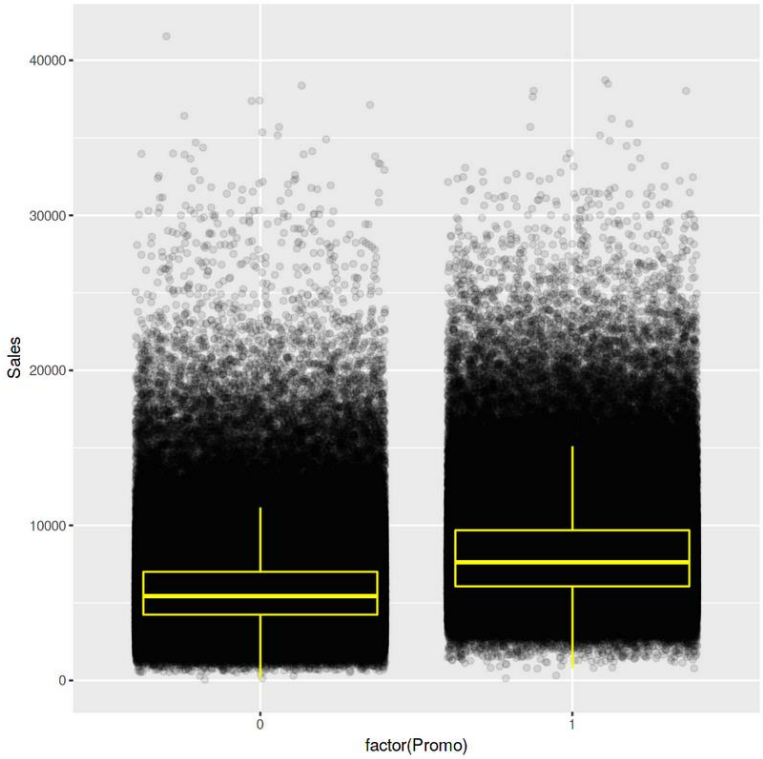
Σχήμα 21: Ιστόγραμμα συχνότητας μέσου όρου πελατών ανά κατάστημα όταν δεν ήταν κλειστό (train.csv)



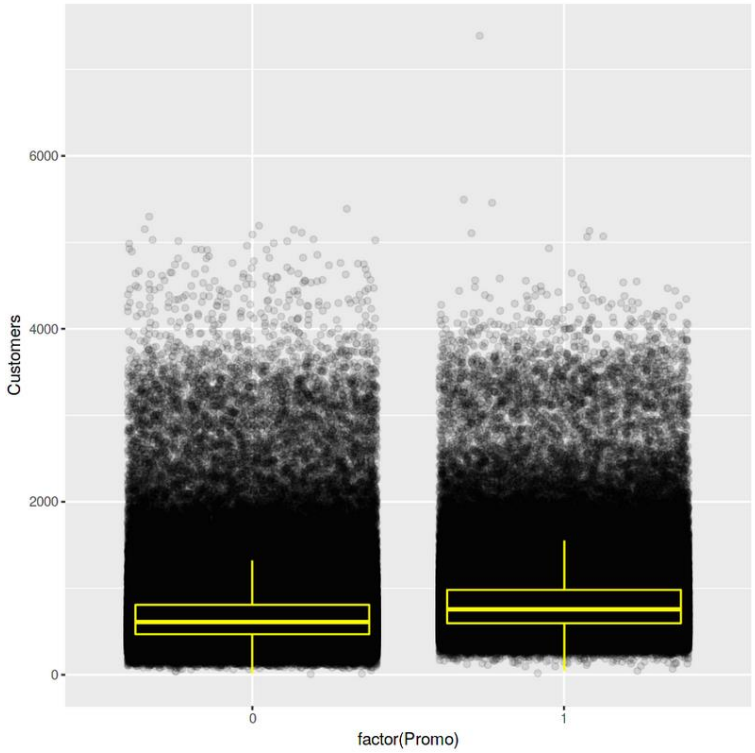
Σχήμα 22: Θηκόγραμμα – επιρροή των πωλήσεων σε σχέση με τις σχολικές αργίες (train.csv)



Σχήμα 23: Λογαριθμικές πωλήσεις συγκριτικά με λογαριθμικό αριθμό πελατών (train.csv)



Σχήμα 24: Θηκόγραμμα – επιρροή των πωλήσεων σε σχέση με τις προωθητικές ενέργειες (train.csv)



Σχήμα 25: Θηκόγραμμα – επιρροή των πελατών σε σχέση με τις προωθητικές ενέργειες (train.csv)

Να σημειωθεί ότι στα θηκογράμματα δεν έχουν συμπεριληφθεί οι ημέρες με μηδενικές πωλήσεις και μηδενικούς πελάτες, διότι θα τα επηρέαζε λανθασμένα.

Επίσης, να επισημάνουμε ότι οι πωλήσεις είναι στενά συσχετισμένες με τον αριθμό των πελατών, αλλά παρατηρούμε ότι το θηκόγραμμα των πωλήσεων επηρεάζεται περισσότερο από τις προωθητικές ενέργειες, σε σχέση με αυτό των πελατών. Αυτό συμβαίνει γιατί οι προωθητικές ενέργειες δεν προσελκύουν κυρίως περισσότερους πελάτες, αλλά προτρέπουν τους πελάτες να ξοδέψουν περισσότερα χρήματα. Η μέση ποσότητα που ξοδεύει ο κάθε πελάτης είναι ένα ευρώ περισσότερο.

Πραγματοποιήθηκε επίσης ένας έλεγχος για την ορθότητα των δεδομένων συγκριτικά με τις πωλήσεις. Όπως φαίνεται και στο σχήμα 26 διαπιστώθηκε ότι υπήρχαν προωθητικές ενέργειες ενώ το αντίστοιχο κατάστημα ήταν κλειστό (μηδενικές πωλήσεις), δηλαδή Promo = 11205 ενώ Sales = 0 και ότι η περίοδος προώθησης καλύπτει περίπου το 45% όλου του χρονικού διαστήματος που εξετάζεται, το οποίο υπολογίστηκε την ημέρες που υπήρχε προωθητική ενέργεια προς τις συνολικές ημέρες ενώ οι πωλήσεις δεν ήταν μηδενικές, δηλαδή από το Promo / (No promo + Promo) = 376875 / 844338 = 0.4464 ή 45%. Στο σχήμα 26 βλέπουμε ότι ορθά δεν υπήρχαν πωλήσεις όταν τα καταστήματα ήταν κλειστά, αλλά παράλληλα εμφανίζονται κάποιες μηδενικές πωλήσεις ενώ τα αντίστοιχα καταστήματα ήταν ανοιχτά και μερικά είχαν και

πελάτες. Κάποιες από αυτές τις μηδενικές πωλήσεις εμφανίζονται σε δύο συνεχόμενες ημέρες, ενώ όπως αναφέρθηκε και προηγουμένως σε κάποια από τα καταστήματα με μηδενικές πωλήσεις υπάρχει παράλληλα και προωθητική ενέργεια (πίνακας 7). Όλες αυτές οι παρατηρήσεις προφανώς είναι εσφαλμένες.

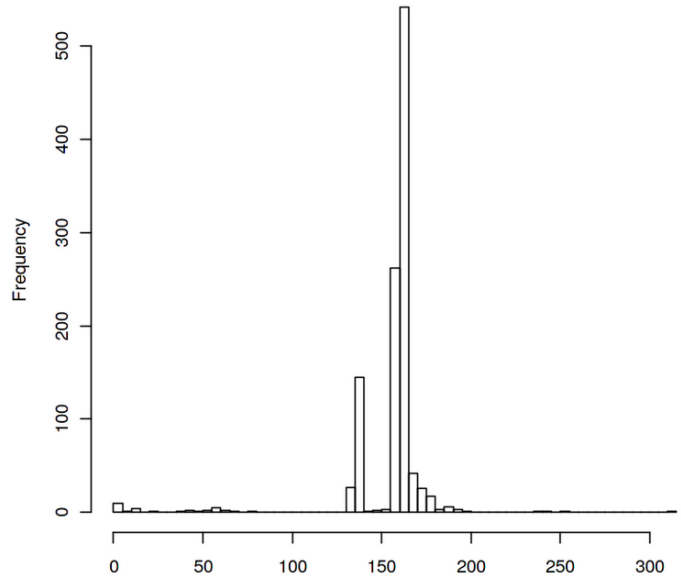
	No promo	Promo
Sales = 0	161666	11205
Sales > 0	467463	376875

	Sales = 0	Sales > 0
Closed	172817	0
Opened	54	844338

Σχήμα 26: Αποτελέσματα προωθητικής ενέργειας - πωλήσεων και πωλήσεων – λειτουργίας καταστήματος (train.csv)

Πίνακας 7: Μηδενικές πωλήσεις σε ανοιχτά καταστήματα

Store	DayOfWeek	Date	Sales	Customers	Open	Promo	StateHoliday	SchoolHoliday
762	4	2013-01-17	0	0	1	0	0	0
232	4	2013-01-24	0	0	1	1	0	0
339	3	2013-01-30	0	0	1	0	0	0
339	4	2013-01-31	0	0	1	0	0	0
259	4	2013-02-07	0	0	1	1	0	0
353	6	2013-03-16	0	0	1	0	0	0
948	4	2013-04-25	0	5	1	1	0	0
589	1	2013-04-29	0	0	1	1	0	0
364	2	2013-05-07	0	0	1	0	0	0
364	3	2013-05-08	0	0	1	0	0	0
681	5	2013-05-10	0	0	1	0	0	0
700	3	2013-06-05	0	0	1	1	0	0
665	5	2013-06-28	0	0	1	0	0	0
665	6	2013-06-29	0	0	1	0	0	0
1039	2	2013-07-09	0	0	1	0	0	0
1039	3	2013-07-10	0	0	1	0	0	0
927	4	2013-08-08	0	0	1	0	0	1
391	3	2013-08-28	0	0	1	1	0	1
663	1	2013-09-02	0	0	1	0	0	1
983	5	2014-01-17	0	0	1	0	0	0
983	6	2014-01-18	0	0	1	0	0	0
623	5	2014-01-24	0	0	1	1	0	0
623	6	2014-01-25	0	0	1	0	0	0
25	3	2014-02-12	0	0	1	0	0	0
25	4	2014-02-13	0	0	1	0	0	0
327	3	2014-03-12	0	0	1	0	0	0
986	2	2014-03-18	0	0	1	1	0	0
850	6	2014-03-29	0	0	1	0	0	0
661	5	2014-04-04	0	0	1	1	0	0
1100	2	2014-04-29	0	3	1	1	0	0
1100	3	2014-04-30	0	0	1	1	0	0
1017	3	2014-06-04	0	0	1	1	0	0
1017	4	2014-06-05	0	0	1	1	0	0
57	2	2014-07-01	0	0	1	1	0	0
925	4	2014-07-03	0	0	1	1	0	0
102	6	2014-07-12	0	0	1	0	0	0
882	3	2014-07-23	0	0	1	0	0	1
887	3	2014-07-23	0	0	1	0	0	0
102	4	2014-07-24	0	0	1	0	0	1
238	4	2014-07-24	0	0	1	0	0	1
303	4	2014-07-24	0	0	1	0	0	1
387	4	2014-07-24	0	0	1	0	0	1
28	2	2014-09-02	0	0	1	1	0	1
28	3	2014-09-03	0	0	1	1	0	1
28	4	2014-09-04	0	0	1	1	0	0
548	5	2014-09-05	0	0	1	1	0	1
835	3	2014-09-10	0	0	1	0	0	0
227	4	2014-09-11	0	0	1	0	0	0
835	4	2014-09-11	0	0	1	0	0	0
357	1	2014-09-22	0	0	1	0	0	0
708	3	2014-10-01	0	0	1	1	0	0
699	4	2015-02-05	0	0	1	1	0	0
674	4	2015-03-26	0	0	1	0	0	0
971	5	2015-05-15	0	0	1	0	0	1



Σχήμα 27: Ιστόγραμμα κατανομής μηδενικών πωλήσεων ανά κατάστημα συνολικά (train.csv)

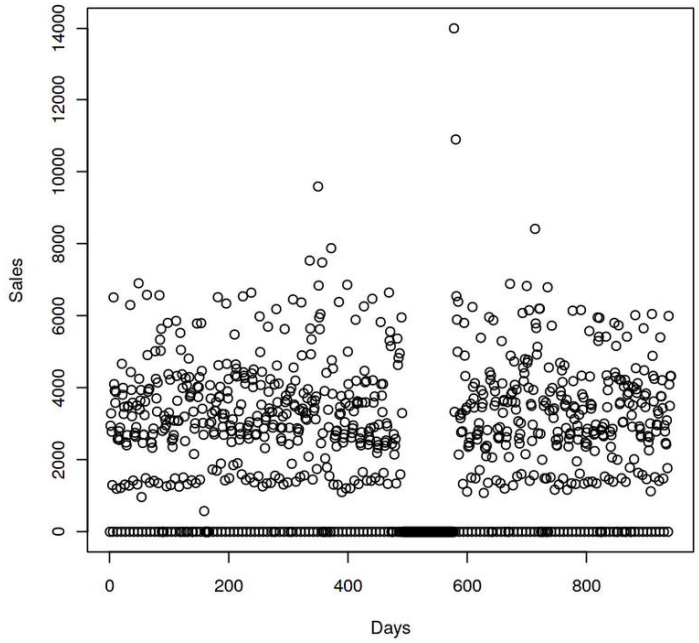
Στο προηγούμενο ιστόγραμμα (σχήμα 27), βλέπουμε στον κάθετο άξονα τον αριθμό των καταστημάτων, ενώ στον οριζόντιο τον αριθμό ημερών που ήταν κλειστά τα καταστήματα. Συνεπώς παρατηρούμε ότι τα καταστήματα έχουν διαφορετικό αριθμό μηδενικών πωλήσεων, δηλαδή ημερών που παρέμειναν κλειστά. Αλλά γενικά υπάρχει μία υψηλή συγκέντρωση των δεδομένων στις περίπου 160 ημέρες ανά κατάστημα (συνολικά 950 από τα 1115 καταστήματα).

Τα δέκα καταστήματα με τον μεγαλύτερο αριθμό μηδενικών πωλήσεων είναι τα εξής:

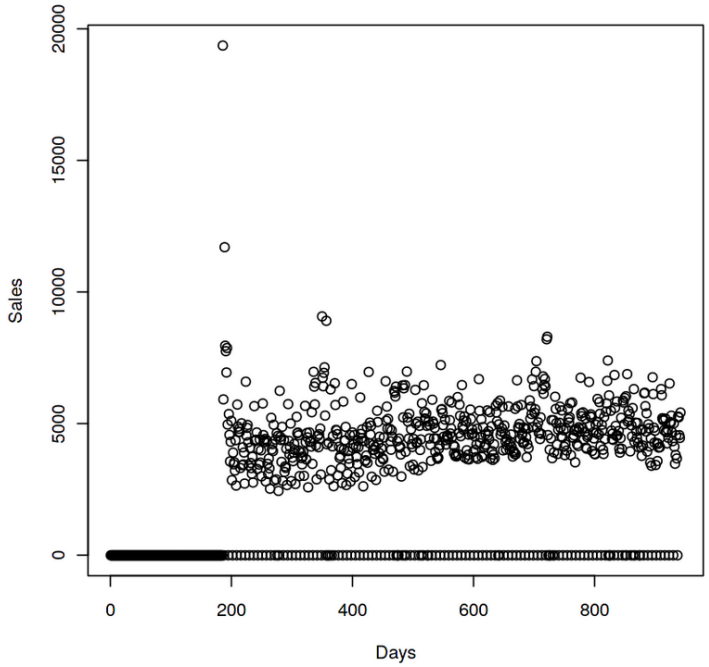
Πίνακας 8: Καταστήματα με τον μεγαλύτερο αριθμό μηδενικών πωλήσεων

Id Καταστήματος	Αριθμός Μηδενικών Πωλήσεων
105	188
339	188
837	191
25	192
560	195
674	197
972	240
349	242
708	255
103	311

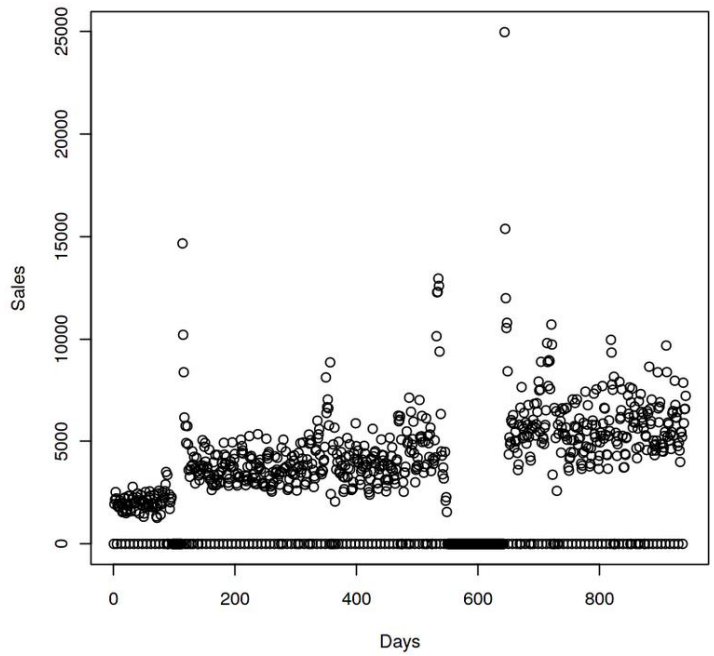
Στη συνέχεια παραθέτονται κάποια διαγράμματα διασποράς των πωλήσεων συγκεκριμένων καταστημάτων που εμφάνισαν μηδενικές πωλήσεις ενώ ήταν ανοιχτά. Στον κάθετο άξονα είναι τα ποσά των πωλήσεων και στον οριζόντιο είναι οι ημέρες του δείγματος, ξεκινώντας από το 1 που αντιστοιχεί στην πρώτη ημέρα (01/01/2013) και καταλήγει στο 941 που αντιστοιχεί στην τελευταία ημέρα (31/07/2015). Παρατηρώντας τα σημεία που αντιστοιχούν στις ημερήσιες πωλήσεις, βλέπουμε ότι υπάρχουν υψηλές εκτινάξεις των πωλήσεων λίγο πριν αυτά κλείσουν και αμέσως μόλις ανοίξουν. Σε κάποια καταστήματα αυτή η περίοδος εμφανίζεται περισσότερες από μία φορές (σχήμα 30).



Σχήμα 28: Κατάστημα 972 - Διασπορά πωλήσεων ανά ημέρα (train.csv)

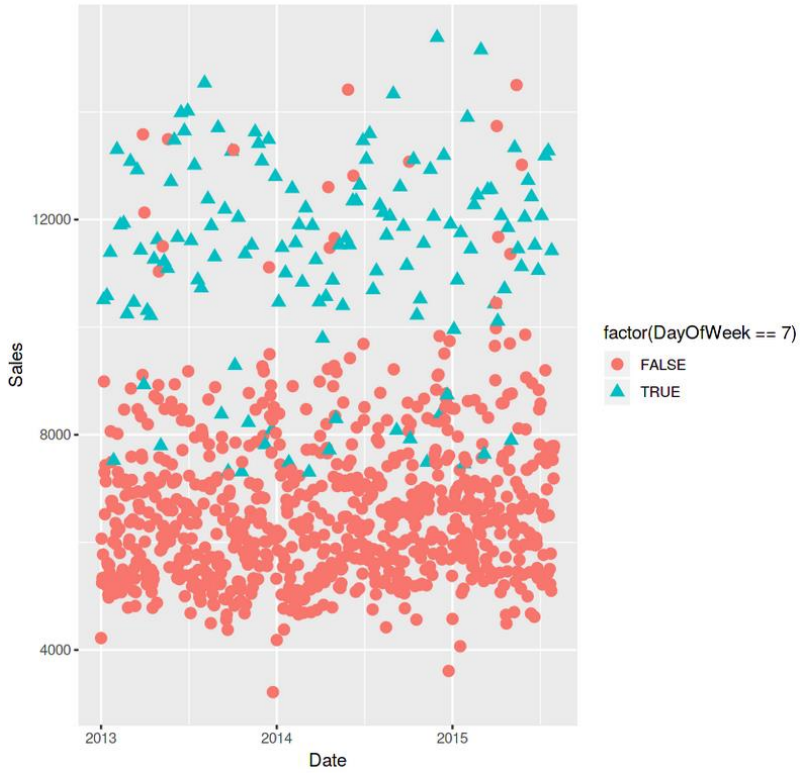


Σχήμα 29: Κατάστημα 103 - Διασπορά πωλήσεων ανά ημέρα (train.csv)

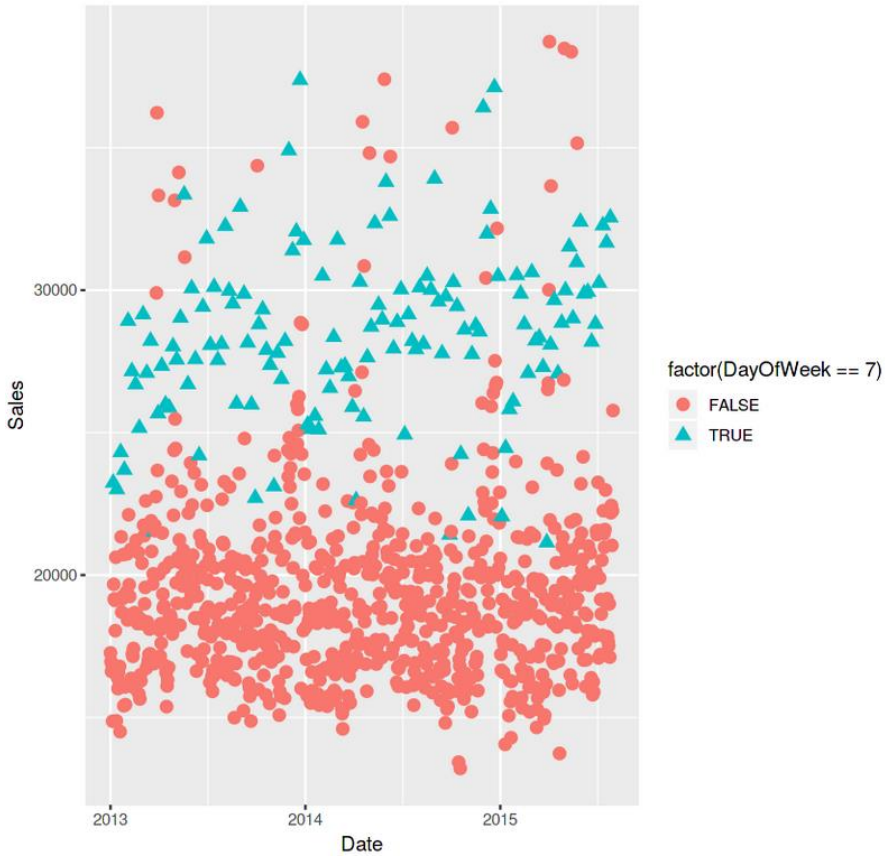


Σχήμα 30: Κατάστημα 708 - Διασπορά πωλήσεων ανά ημέρα (train.csv)

Υπάρχουν επίσης καταστήματα χωρίς καθόλου μηδενικές πωλήσεις στα δεδομένα τους. Αυτά αποτελούν εξαίρεση διότι είναι πάντοτε ανοιχτά, όλες τις Κυριακές και όλες τις αργίες. Στα παρακάτω δύο σχήματα (σχήματα 31 και 32) βλέπουμε τη διασπορά των πωλήσεων στα καταστήματα νούμερο 85 και 262. Με κόκκινους κύκλους βλέπουμε τις ημερήσιες πωλήσεις του καταστήματος, όταν αυτές γίνονται από Δευτέρα έως Σάββατο, ενώ με μπλε τρίγωνα εμφανίζονται οι ημερήσιες πωλήσεις μόνο των Κυριακών. Παρατηρούμε ότι οι πωλήσεις αυτών των καταστημάτων τις Κυριακές είναι ιδιαίτερα υψηλές.

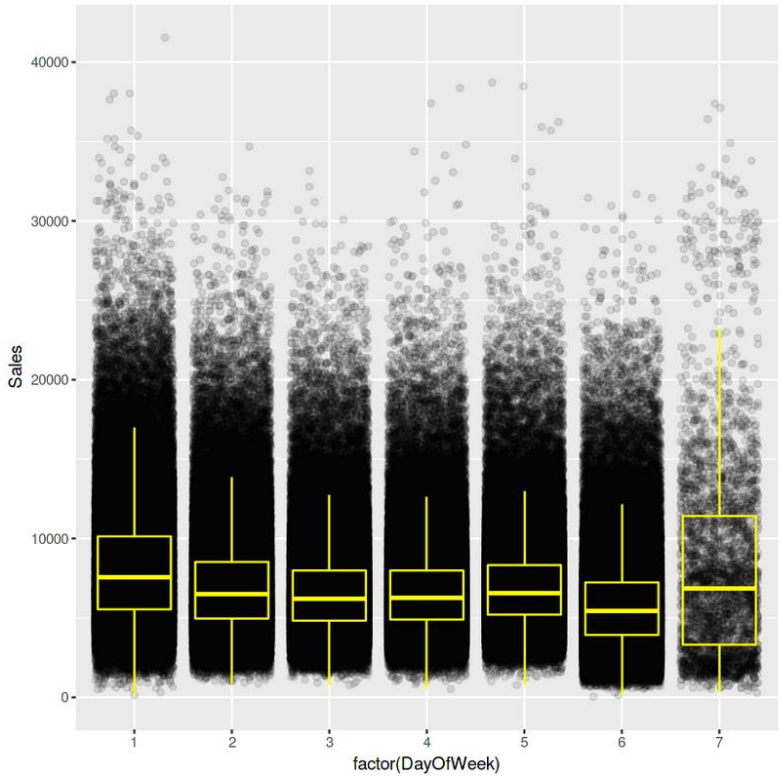


Σχήμα 31: Κατάστημα 85 - Σύγκριση των Κυριακάτικων πωλήσεων. [TRUE εάν είναι Κυριακή] (train.csv)

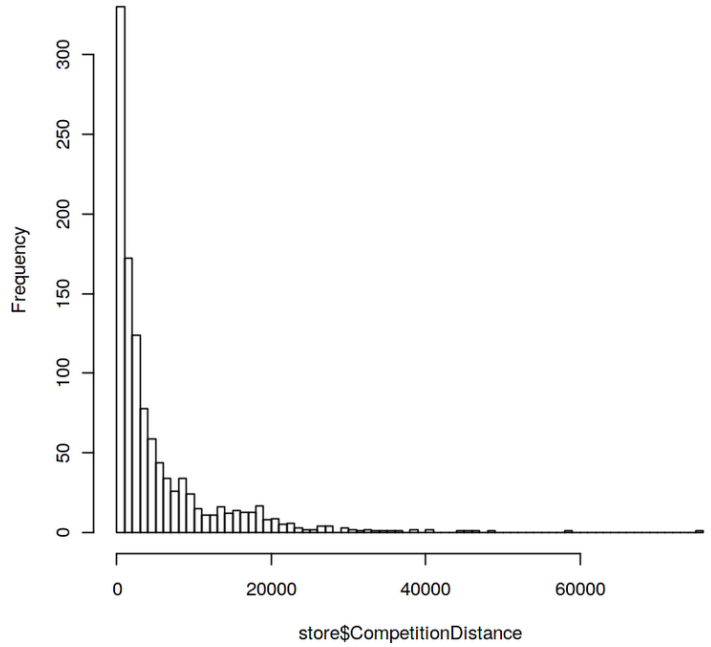


Σχήμα 32: Κατάστημα 262 - Σύγκριση των Κυριακάτικων πωλήσεων. [TRUE εάν είναι Κυριακή] (train.csv)

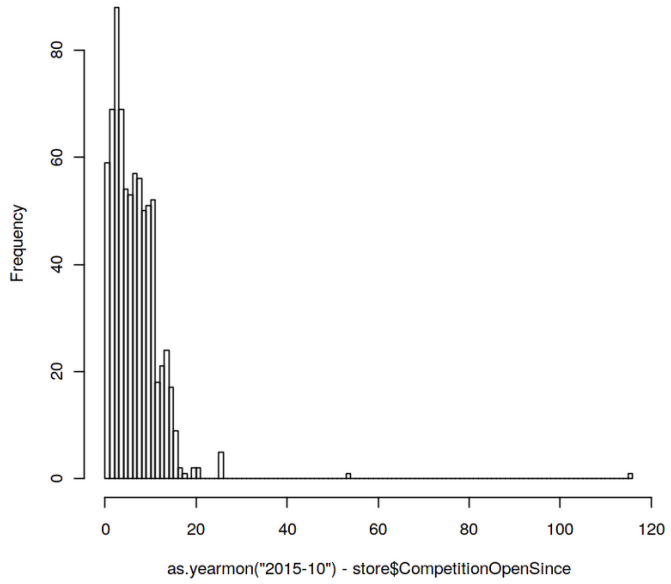
Οι υψηλές πωλήσεις τις Κυριακές δεν ισχύουν γενικά. Υπάρχει μεγάλη διασπορά του ύψους των πωλήσεων τις Κυριακές, από χαμηλές τιμές έως πολύ υψηλές. Αυτό επηρεάζει τον μέσο όρο, ο οποίος δεν είναι ιδιαίτερα υψηλός σε σχέση με τις άλλες ημέρες (σχήμα 33). Υψηλές πωλήσεις αυτή την ημέρα παρουσιάζουν καταστήματα που είναι ευρέως γνωστά ότι είναι ανοιχτά συνέχεια, όπως τα δύο που είδαμε παραπάνω.



Σχήμα 33: Θηκογράμματα - Σύγκριση των μέσων ημερήσιων πωλήσεων (train.csv)



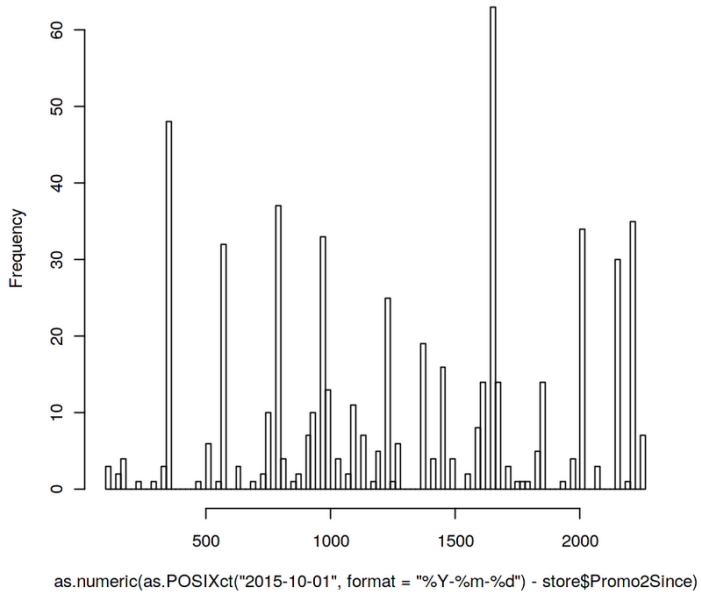
Σχήμα 34: Ιστόγραμμα συχνότητας απόστασης καταστήματος ανταγωνιστή (train.csv)



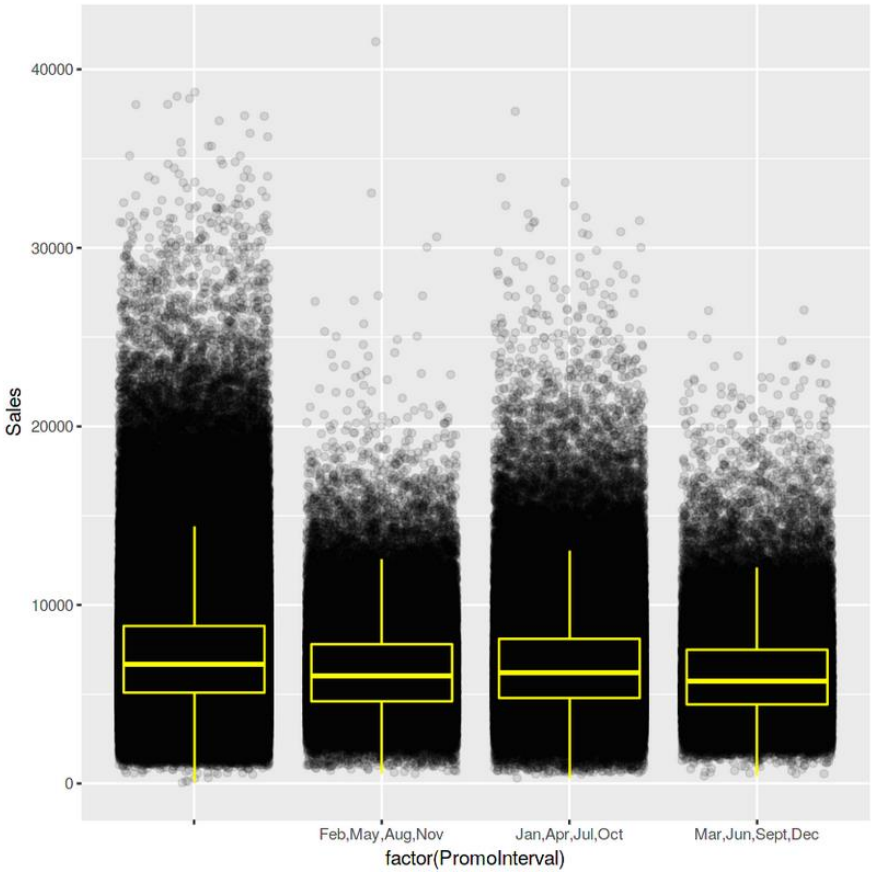
Σχήμα 35: Ιστόγραμμα συχνότητας ετών από τότε που άνοιξε κατάστημα ανταγωνιστή – Σημείο αναφοράς Οκτώβριος 2015 (train.csv)

Από τα παραπάνω ιστογράμματα βγαίνει το πόρισμα ότι τα καταστήματα των ανταγωνιστών άνοιξαν σε πολύ κοντινή απόσταση από αυτά της ROSSMANN (μικρότερη των χιλίων μέτρων) και αυτό συνέβη κατά το πλείστων τα τελευταία δεκαπέντε έτη (από τον Οκτώβριο 2015).

Στη συνέχεια παρουσιάζεται μία ανάλυση των δεδομένων που αφορούν τις περιόδους προώθησης. Στα επόμενα σχήματα βλέπουμε τη συχνότητα των διαστημάτων εφαρμογής τους καθώς και την επιρροή τους στις πωλήσεις.



Σχήμα 36: Ιστόγραμμα συχνότητας ημερών από τότε που ξεκίνησε το Promo2 – Σημείο αναφοράς 01/10/2015 (train.csv)



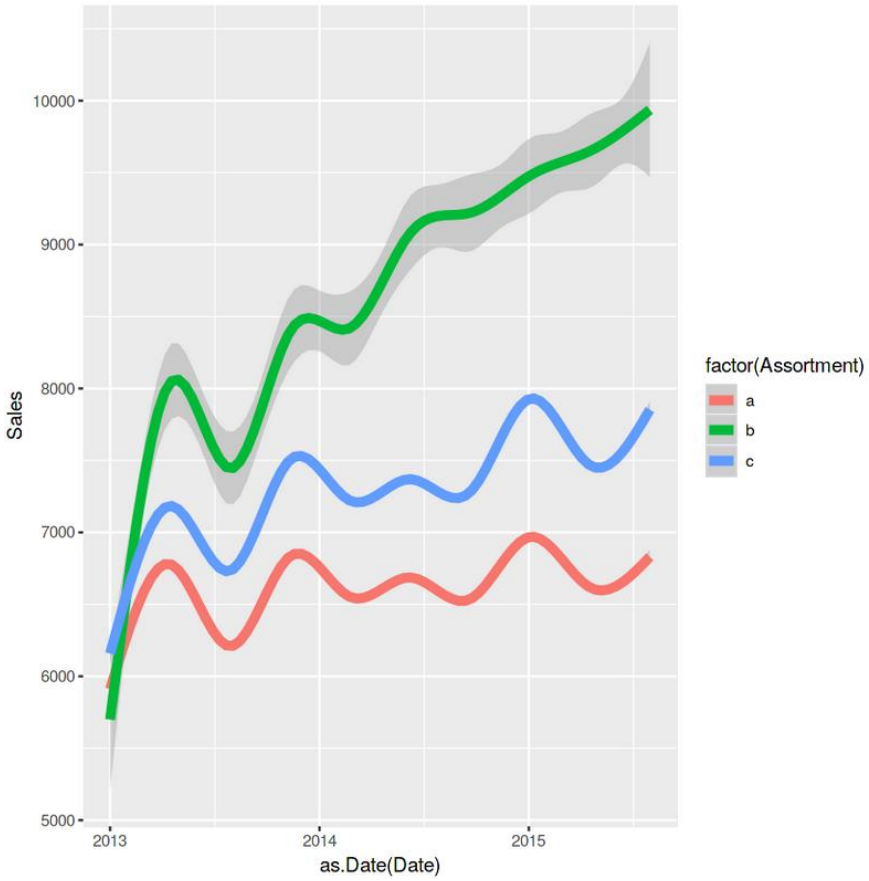
Σχήμα 37: Ιστόγραμμα συχνότητας ημερών από τότε που ξεκίνησε το Promo2 – Σημείο αναφοράς 01/10/2015 (train.csv)

Στο σχήμα 37 παρατηρούμε ότι τα καταστήματα που συμμετέχουν σε προωθητικές ενέργειες έχουν την τάση να έχουν χαμηλότερες πωλήσεις. Αυτό δεν σημαίνει απαραίτητα ότι οι ενέργειες αυτές δεν βοηθούν ή είναι αντιπαραγωγικές. Απλά συνήθως είναι μέτρα που λαμβάνονται κυρίως σε καταστήματα που έχουν χαμηλές πωλήσεις εξ ’αρχής.

Παρακάτω, στον πίνακα 9, σημειώνεται η συσχέτιση μεταξύ των τεσσάρων τύπων καταστημάτων και της ποικιλίας των προϊόντων που αυτά διαθέτουν. Αυτή η συσχέτιση προκύπτει από απλή κατανομή των δοθέντων δεδομένων. Ενώ ακολούθως οπτικοποιούμε σε διαγράμματα την σχέση της ποικιλίας προϊόντων με το ύψος των πωλήσεων και τον αριθμό πελατών. Από τα διαγράμματα συμπεραίνουμε ότι οι διαφορετικές ποικιλίες προϊόντων, συνεπάγονται με διαφορετικά επίπεδα πωλήσεων και τα καταστήματα με την κατηγορία ποικιλίας ‘b’ ακολουθεί διαφορετική τάση στην αγορά, το οποίο είναι λογικό εάν αναλογιστούμε ότι αυτά είναι μόνο 9 στον αριθμό (πίνακας 9). Συγκεκριμένα στο σχήμα 38 βλέπουμε ότι από την 01/01/2013 έως και περίπου το πρώτο τρίμηνο του 2014 οι πωλήσεις των καταστημάτων ανεξαρτήτου της ποικιλίας που έχουν, ακολουθούν την ίδια τάση, με σχετικά αυξημένες πωλήσεις των καταστημάτων που έχουν την κατηγορία ποικιλίας ‘b’. Από εκείνο το σημείο και έπειτα τα καταστήματα με την ποικιλία ‘b’ ξεκινούν μία ανοδική πορεία πωλήσεων, ενώ τα καταστήματα με τις ποικιλίες ‘a’ και ‘c’ συνεχίζουν μία παράλληλη μεταξύ τους πιο σταθερή πορεία.

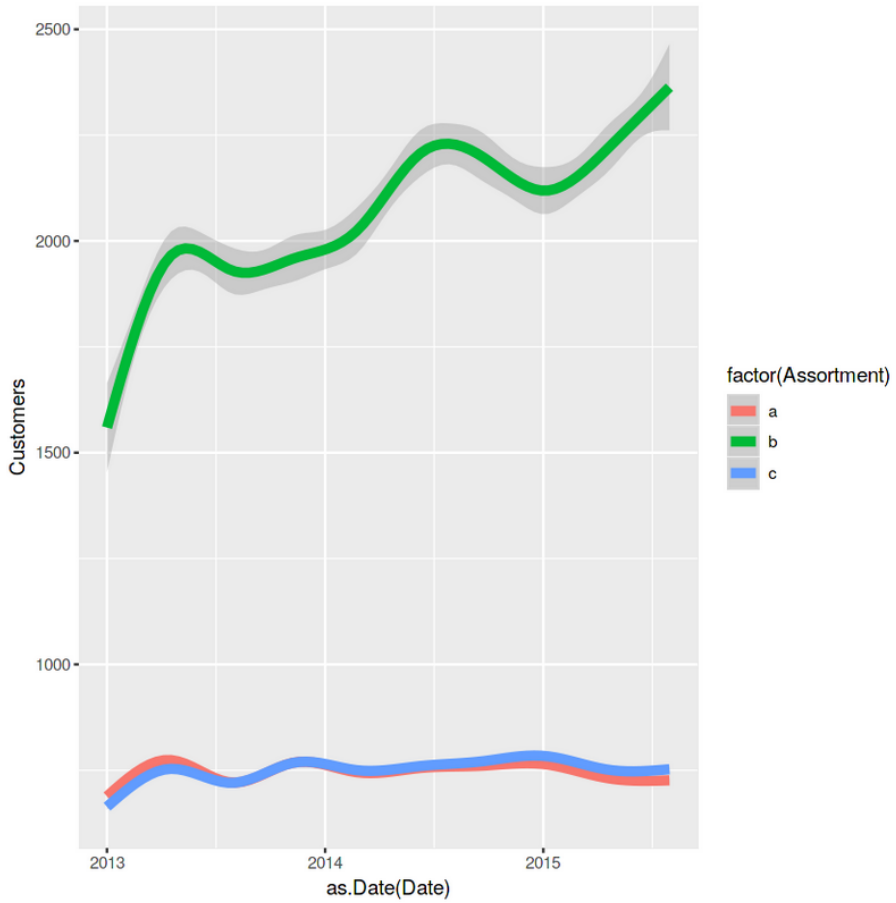
Πίνακας 9: Συσχέτιση μεταξύ των τεσσάρων τύπων καταστημάτων και της ποικιλίας προϊόντων που διαθέτουν

		Store Type				
		a	b	c	d	**Total**
Assortment	a	381	7	77	128	**593**
	b	0	9	0	0	**9**
	c	221	1	71	220	**513**
	Total	**602**	**17**	**148**	**348**	**1115**



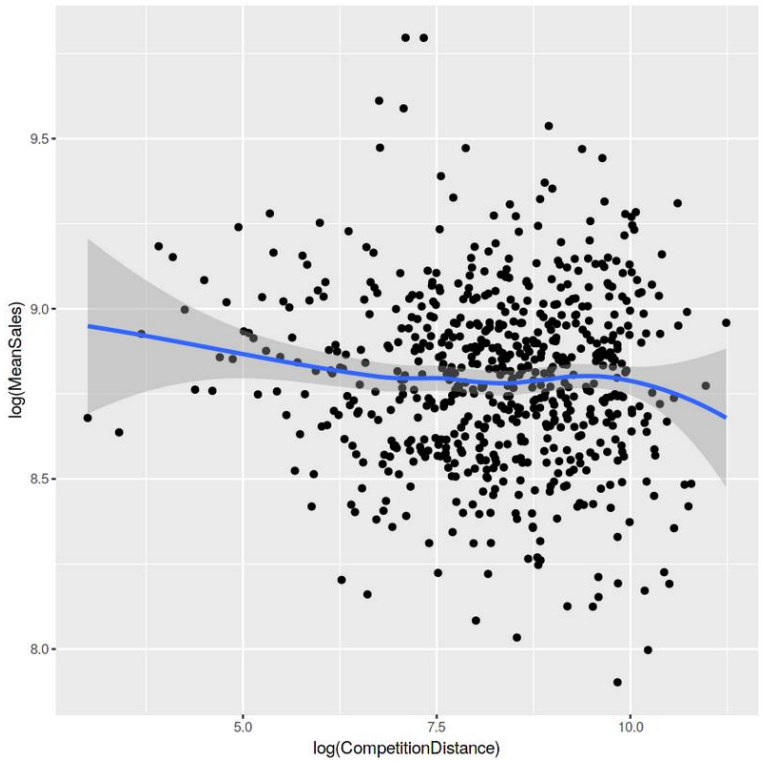
Σχήμα 38: Πωλήσεις συγκριτικά με την ποικιλία προϊόντων των καταστημάτων (train.csv)

Σε τελείως διαφορετική κατάσταση όπως παρατηρούμε και στο σχήμα 39, βρίσκεται ο αριθμός των πελατών που επισκέπτονται τα καταστήματα ανάλογα με την κατηγορία ποικιλίας των προϊόντων. Ενώ στα καταστήματα των κατηγοριών ποικιλίας ‘a’ και ‘c’ ο αριθμός πελατών συμβαδίζει και είναι περίπου της τάξης των 500 πελατών ημερησίως, σε αυτά της ποικιλίας ‘b’ ο αριθμός τους κυμαίνεται περίπου από 1500 έως 2300 πελάτες ημερησίως, έχοντας μια διαρκώς ανοδική πορεία.



Σχήμα 39: Πελάτες συγκριτικά με την ποικιλία προϊόντων των καταστημάτων (train.csv)

Η επιρροή των πωλήσεων από την απόσταση των γειτονικών ανταγωνιστικών καταστημάτων ερευνάται στη συνέχεια. Μετά τη δημιουργία λογαριθμικού διαγράμματος μέσης τιμής πωλήσεων σε σύγκριση με την απόσταση από κατάστημα ανταγωνιστή (σχήμα 40), εκτιμάται ότι η σχέση τους δεν είναι ξεκάθαρη. Υπάρχει όμως ένα ελαφρώς αυξημένο ποσό πωλήσεων όταν η απόσταση αυτή είναι μικρή. Το οποίο πιθανόν να συμβαίνει διότι τα καταστήματα που έχουν πολύ κοντά κάποιον ανταγωνιστή, βρίσκονται σε κάποιο πυκνοκατοικημένο αστικό κέντρο, οπότε οι πωλήσεις λόγω του πληθυσμού είναι ούτως ή άλλως αυξημένες. Άρα οι πωλήσεις δεν επηρεάζονται εμφανώς από την απόσταση των ανταγωνιστών, αλλά πιο λογικό είναι να επηρεάζονται από τα χαρακτηριστικά της περιοχής που βρίσκονται.

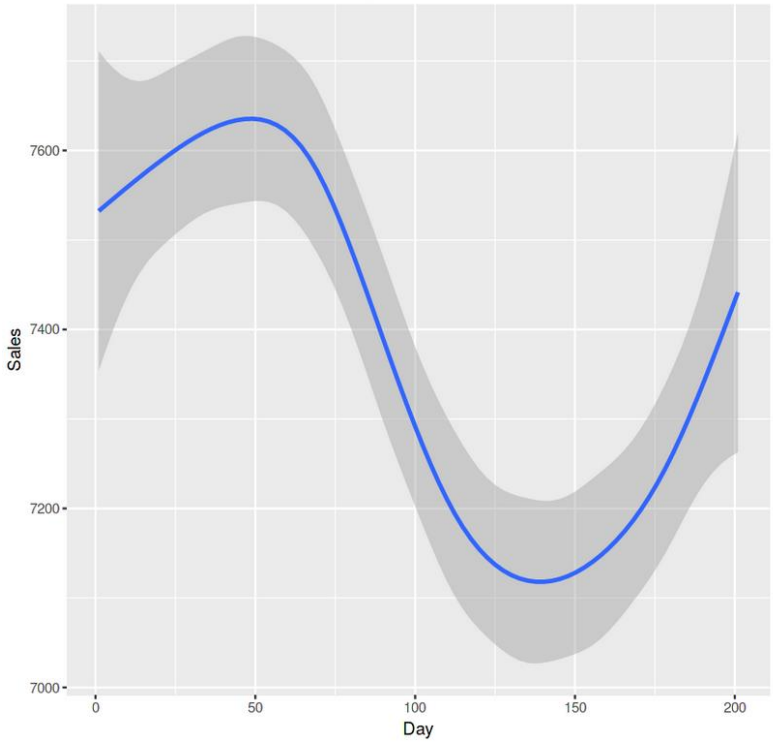


Σχήμα 40: Λογαριθμικός μέσος όρος πωλήσεων συγκριτικά με την λογαριθμική απόσταση των ανταγωνιστών (train.csv)

Το τι συμβαίνει εάν κάποιος ανταγωνιστής ανοίξει κατάστημα στην περιοχή που υπάρχει ένα της ROSSMANN, θα το εξετάσουμε από τα δεδομένα των καταστημάτων που αρχικά έχουν την τιμή 'NA' ως 'CompetitorDistance' (δηλαδή δεν έχουν κάποιον ανταγωνιστή κοντά) και αργότερα εμφανίζεται ένας αριθμός στα αντίστοιχα κελιά της στήλης.

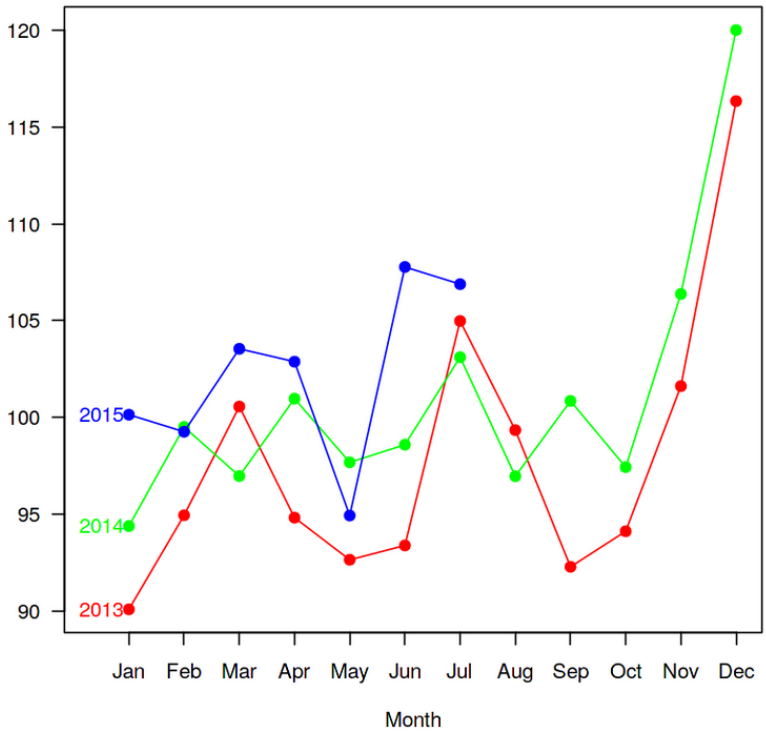
Επειδή από τα δεδομένα είναι γνωστός μόνο ο μήνας του ανοίγματος του ανταγωνιστή, θα χρησιμοποιηθεί ένα αρκετά μεγάλο χρονικό διάστημα για να δούμε το αποτέλεσμα. Οπότε στο γράφημα θα προηγηθούν εκατό ημέρες πωλήσεων πριν το άνοιγμα του ανταγωνιστή. Την εκατοστή ημέρα ανοίγει ο ανταγωνιστής. Κατά τη διάρκεια του διαθέσιμου χρονικού διαστήματος, 147 καταστήματα δεν είχαν αρχικά και απέκτησαν στην πορεία έναν ανταγωνιστή στην περιοχή τους.

Στο σχήμα 41 εμφανίζεται μία πτώση των πωλήσεων μετά την ημέρα 100, όταν και ανοίγει ο ανταγωνιστής. Βέβαια υπήρχε από πριν μία πτωτική τάση στις πωλήσεις, αλλά δεν μπορεί να αρνηθεί κάποιος και την επιρροή του ανταγωνιστή στη συνέχεια.



Σχήμα 41: Επιρροή των πωλήσεων όταν δεν προϋπήρχε και ανοίγει (Day = 100) κατάστημα ανταγωνιστή στην περιοχή (train.csv)

Τέλος παρουσιάζονται οπτικά οι μέσες μηνιαίες πωλήσεις ανά κατάστημα τη χρονική περίοδο των δεδομένων που μας παρέχονται στο αρχείο train.csv (Thie1e (2015)).



Σχήμα 42: Μέσες μηνιαίες πωλήσεις ανά κατάστημα (train.csv)

5.4. Επεξεργασία των δεδομένων και διερεύνηση τάσεων και μοτίβων

Παρακάτω θα παρουσιάσουμε μία διαφορετική διερεύνηση και επεξεργασία των δεδομένων, από άλλη οπτική γωνία, εμβαθύνοντας σε άλλες πτυχές των δεδομένων. Αυτή η διερεύνηση πραγματοποιήθηκε από την Elena Petrova για τον ίδιο διαγωνισμό στην πλατφόρμα του Kaggle (Petrova (2016)). Η γλώσσα προγραμματισμού που χρησιμοποιήθηκε είναι η Python και ο αλγόριθμος επισυνάπτεται στο Παράρτημα Β. Στο επόμενο κεφάλαιο, το μοντέλο πρόβλεψης που θα αναπτύξουμε, θα βασιστεί σε αυτή την επεξεργασία και ανάλυση των δεδομένων.

Αρχικά θα γίνει μία σύντομη επεξεργασία των δεδομένων, ώστε να τα μετατρέψουμε σε μορφή κατάλληλη για τη δημιουργία του μοντέλου πρόβλεψης. Έπειτα θα παρουσιαστούν οι τάσεις και τα μοτίβα των δεδομένων, συμβάλλοντας στην καλύτερη κατανόηση των συσχετίσεων για την περαιτέρω ανάλυση. Τέλος θα γίνει μία ανάλυση των χρονοσειρών των δεδομένων.

Ως πρώτο βήμα στην εκτέλεση του αλγορίθμου, είναι η φόρτωση των έτοιμων αλγορίθμων που χρησιμοποιούνται για την επεξεργασία και οπτικοποίηση των δεδομένων στο Kernel. Η γλώσσα προγραμματισμού που χρησιμοποιήθηκε, όπως προαναφέρθηκε είναι η Python. Παρακάτω δίνονται τμήματα του κώδικα μαζί με τα αποτελέσματά τους:

```
import warnings
warnings.filterwarnings("ignore")

# loading packages
# basic + dates
import numpy as np
import pandas as pd
from pandas import datetime

# data visualization
import matplotlib.pyplot as plt
import seaborn as sns # advanced vizs
%matplotlib inline

# statistics
from statsmodels.distributions.empirical_distribution import ECDF

# time series analysis
from statsmodels.tsa.seasonal import seasonal_decompose
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf

# prophet by Facebook
from fbprophet import Prophet
```

```
# importing train data to learn
train = pd.read_csv("../input/train.csv",
                    parse_dates = True, low_memory = False, index_col = 'Date')

# additional store data
store = pd.read_csv("../input/store.csv",
                    low_memory = False)
# time series as indexes
train.index
```

```
DatetimeIndex(['2015-07-31', '2015-07-31', '2015-07-31', '2015-07-31',
               '2015-07-31', '2015-07-31', '2015-07-31', '2015-07-31',
               '2015-07-31', '2015-07-31',
               ...
               '2013-01-01', '2013-01-01', '2013-01-01', '2013-01-01',
               '2013-01-01', '2013-01-01', '2013-01-01', '2013-01-01',
               '2013-01-01', '2013-01-01'],
              dtype='datetime64[ns]', name='Date', length=1017209, freq=None)
```

Στη συνέχεια θα διαχειριστούμε τις ελλιπείς καταχωρήσεις (missing values), τις οποίες διαπιστώσαμε και προηγουμένως και θα δημιουργήσουμε νέες μεταβλητές για περαιτέρω ανάλυση. Τα δεδομένα του train.csv είναι δεδομένα χρονοσειρών, οπότε θα είναι χρήσιμο να εξάγουμε και να τα αναλύσουμε ανά ημερομηνία. Επίσης υπάρχουν δύο συσχετιζόμενες μεταβλητές οι οποίες μπορούν να συνδυαστούν σε μία καινούργια. Αυτές είναι οι πωλήσεις και ο αριθμός των πελατών, από τις οποίες δημιουργήθηκαν οι πωλήσεις ανά πελάτη.

```
# data extraction
train['Year'] = train.index.year
train['Month'] = train.index.month
train['Day'] = train.index.day
train['WeekOfYear'] = train.index.weekofyear

# adding new variable
train['SalePerCustomer'] = train['Sales']/train['Customers']
train['SalePerCustomer'].describe()
```

```
count    844340.000000
mean          9.493619
std           2.197494
min           0.000000
25%           7.895563
50%           9.250000
75%          10.899729
max          64.957854
Name: SalePerCustomer, dtype: float64
```

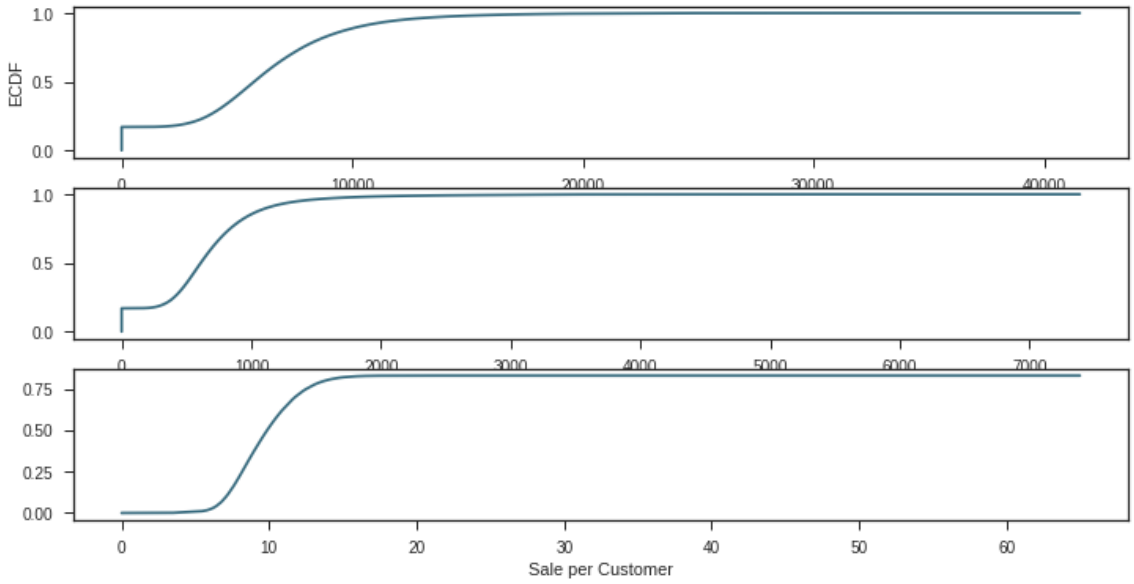
Από τα παραπάνω συμπεραίνουμε ότι ο κάθε πελάτης ξοδεύει 9.50 € ανά ημέρα, ενώ υπάρχουν και ημέρες που δεν ξοδεύουν τίποτα. Για να αποκτήσουμε μία πρώτη εικόνα για την συνέχεια των δεδομένων της νέας μεταβλητής που δημιουργήσαμε, θα εξάγουμε το διάγραμμα εμπειρικής συσσωρευτικής λειτουργίας κατανομής (Empirical Cumulative Distribution Function - ECDF).

```
sns.set(style = "ticks")# to format into seaborn
c = '#386B7F' # basic color for plots
plt.figure(figsize = (12, 6))

plt.subplot(311)
cdf = ECDF(train['Sales'])
plt.plot(cdf.x, cdf.y, label = "statmodels", color = c);
plt.xlabel('Sales'); plt.ylabel('ECDF');

# plot second ECDF
plt.subplot(312)
cdf = ECDF(train['Customers'])
plt.plot(cdf.x, cdf.y, label = "statmodels", color = c);
plt.xlabel('Customers');

# plot second ECDF
plt.subplot(313)
cdf = ECDF(train['SalePerCustomer'])
plt.plot(cdf.x, cdf.y, label = "statmodels", color = c);
plt.xlabel('Sale per Customer');
```



Σχήμα 43: ECDF των πωλήσεων ανά πελάτη

Παρατηρούμε ότι το 20% των δεδομένων αντιστοιχεί σε μηδενικές πωλήσεις ανά πελάτη και σχεδόν το 80% των καθημερινών πωλήσεων ήταν λιγότερο από 1000 €.

Όπως διερευνήσαμε και στη προηγούμενη ενότητα, οι μηδενικές πωλήσεις δεν αφορούν μόνο τα κλειστά καταστήματα, αλλά βρέθηκαν και κάποιες περιπτώσεις όπου ενώ ήταν ανοιχτά τα καταστήματα, δεν σημείωσαν πωλήσεις (54 καταχωρήσεις). Αυτές

οι 54 καταχωρήσεις μαζί με τις 172817 καταχωρήσεις όπου ήταν κλειστά τα καταστήματα, δεν θα συμπεριληφθούν στην πρόβλεψη πωλήσεων που θα εκτελεστεί στο επόμενο κεφάλαιο για να αποφύγουμε τυχόν λανθασμένες επιρροές. Οι καταχωρήσεις αυτές αποτελούν περίπου το 10% του συνόλου.

Όσον αφορά το αρχείο store.csv, εξετάζουμε αρχικά εάν υπάρχουν κενές καταχωρήσεις.

```
# missing values?
store.isnull().sum()
```

```
Store                        0
StoreType                    0
Assortment                   0
CompetitionDistance          3
CompetitionOpenSinceMonth  354
CompetitionOpenSinceYear   354
Promo2                       0
Promo2SinceWeek            544
Promo2SinceYear            544
PromoInterval              544
dtype: int64
```

Είναι λίγες οι μεταβλητές με κενές τιμές σε αυτό το αρχείο. Η μεταβλητή CompetitionDistance έχει μόνο τρεις, οι οποίες δεν παρουσιάζουν κάποιο μοτίβο, απλά δεν είναι καταχωρημένες. Σε αυτή την περίπτωση θα τις αντικαταστήσουμε με τη μεσαία τιμή (η οποία είναι μισή από την μέση τιμή).

```
# fill NaN with a median value (skewed distribuion)
store['CompetitionDistance'].fillna(store['CompetitionDistance']
.median(), inplace = True)
```

Συνεχίζοντας με τις μεταβλητές Promo2SinceWeek, CompetitionOpenSinceMonth και CompetitionOpenSinceYear, εφόσον δεν έχουμε πληροφορίες για αυτές, τις κενές καταχωρήσεις θα τις αντικαταστήσουμε με μηδενικά.

```
# no promo = no information about the promo?
_ = store[pd.isnull(store.Promo2SinceWeek)]
_[_.Promo2 != 0].shape
```

```
(0, 10)
```

```
# replace NA's by 0
store.fillna(0, inplace = True)
```

Το επόμενο βήμα είναι να ενώσουμε το αρχείο train.csv με το αρχείο store.csv, ώστε να μπορούμε να το επεξεργαστούμε αποτελεσματικότερα. Αυτό θα πραγματοποιηθεί έχοντας μία στήλη κοινή, ώστε να συσχετιστούν οι σειρές μεταξύ τους. Ως κοινή στήλη χρησιμοποιήθηκε το 'Κατάστημα' (Store). Το νέο αρχείο που δημιουργήθηκε έχει 22 στήλες και 844338 γραμμές (πίνακας 10).

```
print("Joining train set with an additional store information.")

# by specifying inner join we make sure that only those observations
# that are present in both train and store sets are merged together
train_store = pd.merge(train, store, how = 'inner', on = 'Store')

print("In total: ", train_store.shape)
train_store.head()
```

```
Joining train set with an additional store information.
In total:  (844338, 22)
```

Ένα δείγμα των πρώτων πέντε γραμμών του τελικού αποτελέσματος μετά τη συγχώνευση είναι το εξής:

Πίνακας 10: Δείγμα πέντε γραμμών συγχωνευμένου πίνακα των δεδομένων train και store (στήλες 22)

	Store	DayOfWeek	Sales	Customers	Open	Promo	StateHoliday	SchoolHoliday	Year	Month	...	SalePerCustomer
0	1	5	5263	555	1	1	0	1	2015	7	...	9.482883
1	1	4	5020	546	1	1	0	1	2015	7	...	9.194139
2	1	3	4782	523	1	1	0	1	2015	7	...	9.143403
3	1	2	5011	560	1	1	0	1	2015	7	...	8.948214
4	1	1	6102	612	1	1	0	1	2015	7	...	9.970588

StoreType	Assortment	CompetitionDistance	CompetitionOpenSinceMonth	CompetitionOpenSinceYear	Promo2
c	a	1270.0	9.0	2008.0	0
c	a	1270.0	9.0	2008.0	0
c	a	1270.0	9.0	2008.0	0
c	a	1270.0	9.0	2008.0	0
c	a	1270.0	9.0	2008.0	0

Promo2SinceWeek	Promo2SinceYear	PromoInterval
0.0	0.0	0
0.0	0.0	0
0.0	0.0	0
0.0	0.0	0
0.0	0.0	0

Σειρά έχει η εξέταση των διαφορετικών τύπων καταστημάτων (StoreType) και πως διανέμεται η μεταβλητή Sales στον εκάστοτε τύπο.

```
train_store.groupby('StoreType')['Sales'].describe()
```

	count	mean	std	min	25%	50%	75%	max
StoreType								
a	457042.0	6925.697986	3277.351589	46.0	4695.25	6285.0	8406.00	41551.0
b	15560.0	10233.380141	5155.729868	1252.0	6345.75	9130.0	13184.25	38722.0
c	112968.0	6933.126425	2896.958579	133.0	4916.00	6408.0	8349.25	31448.0
d	258768.0	6822.300064	2556.401455	538.0	5050.00	6395.0	8123.25	38037.0

Τα καταστήματα τύπου ‘b’ έχουν τον υψηλότερο μέσο πωλήσεων, ωστόσο έχουν τον μικρότερο αριθμό δεδομένων. Οπότε ας εξάγουμε το άθροισμα των πωλήσεων και των πελατών, ώστε να διαπιστώσουμε ποιο είναι το κατάστημα με τον μεγαλύτερο τζίρο και τους περισσότερους πελάτες.

```
train_store.groupby('StoreType')['Customers', 'Sales'].sum()
```

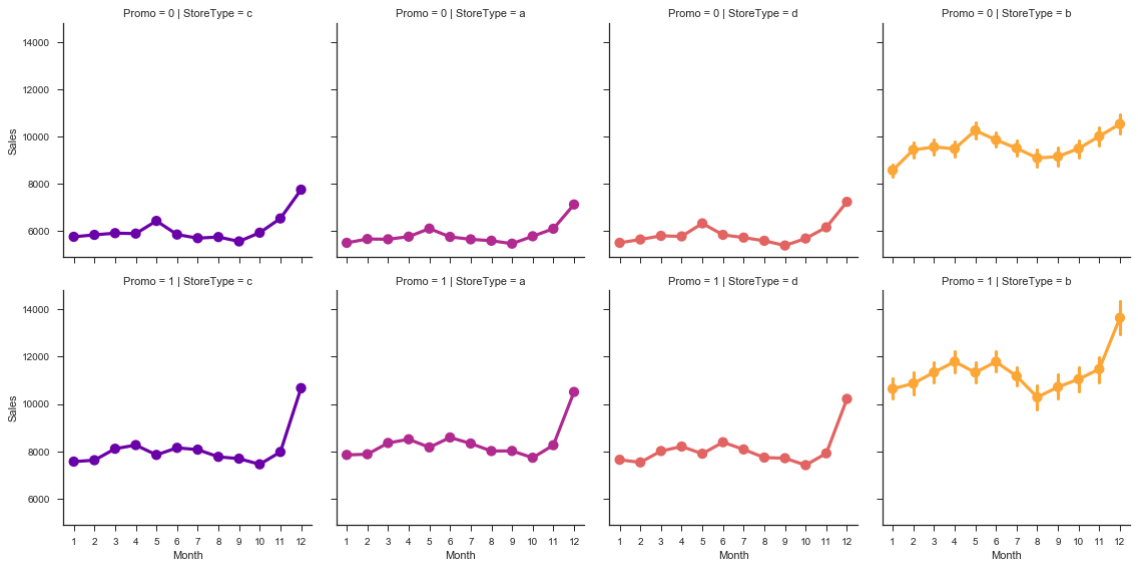
	Customers	Sales
StoreType		
a	363541431	3165334859
b	31465616	159231395
c	92129705	783221426
d	156904995	1765392943

Ολοφάνερο είναι ότι τα καταστήματα τύπου ‘a’ έχουν τον μεγαλύτερο τζίρο και τους περισσότερους πελάτες, ενώ τα καταστήματα τύπου ‘d’ βρίσκονται στη δεύτερη θέση. Για να συμπεριλάβουμε και τον χρόνο στις πωλήσεις, θα χρησιμοποιήσουμε το εργαλείο Seaborn’sfacet grid.

Πρώτα θα παρουσιαστούν διαγράμματα πωλήσεων σε σχέση με τον τύπο καταστήματος, όπου σε κάθε στήλη είναι οι διαφορετικοί τύποι καταστημάτων ενώ στην επάνω σειρά είναι χωρίς προωθητική ενέργεια (Promo = 0) και στην κάτω είναι εντός προωθητικής ενέργειας (Promo = 1).

```
# sales trends
sns.factorplot(data = train_store, x = 'Month', y = "Sales",
               col = 'StoreType', # per store type in cols
               palette = 'plasma',
               hue = 'StoreType',
               row = 'Promo', # per promo in the store in rows
               color = c)
```

```
<seaborn.axisgrid.FacetGrid at 0x7fb34d03cc50>
```

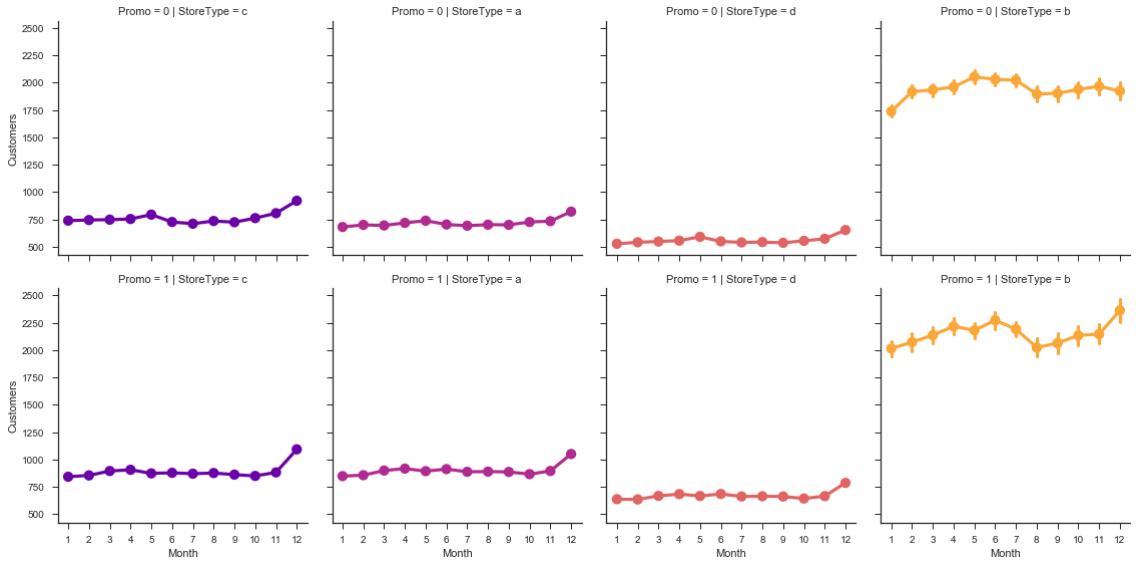


Σχήμα 44: Πωλήσεις ανάλογα με τον τύπο καταστήματος και την παρουσία ή όχι προωθητικής ενέργειας

Στη συνέχεια παρουσιάζονται διαγράμματα αριθμού πελατών σε σχέση με τον τύπο καταστήματος. Εφαρμόζεται η ίδια λογική, όπως προηγουμένως.

```
# sales trends
sns.factorplot(data = train_store, x = 'Month', y = "Customers",
               col = 'StoreType', # per store type in cols
               palette = 'plasma',
               hue = 'StoreType',
               row = 'Promo', # per promo in the store in rows
               color = c)
```

```
<seaborn.axisgrid.FacetGrid at 0x7fb34d0d2d30>
```



Σχήμα 45: Πελάτες ανάλογα με τον τύπο καταστήματος και την παρουσία ή όχι προωθητικής ενέργειας

Παρατηρούμε ότι όλοι οι τύποι ακολουθούν την ίδια τάση, σε διαφορετικά επίπεδα πωλήσεων και αριθμού πελατών, ανάλογα την παρουσία προωθητικής ενέργειας (Promo) ή όχι. Μία μικρή διαφοροποίηση υπάρχει στα καταστήματα τύπου 'b'.

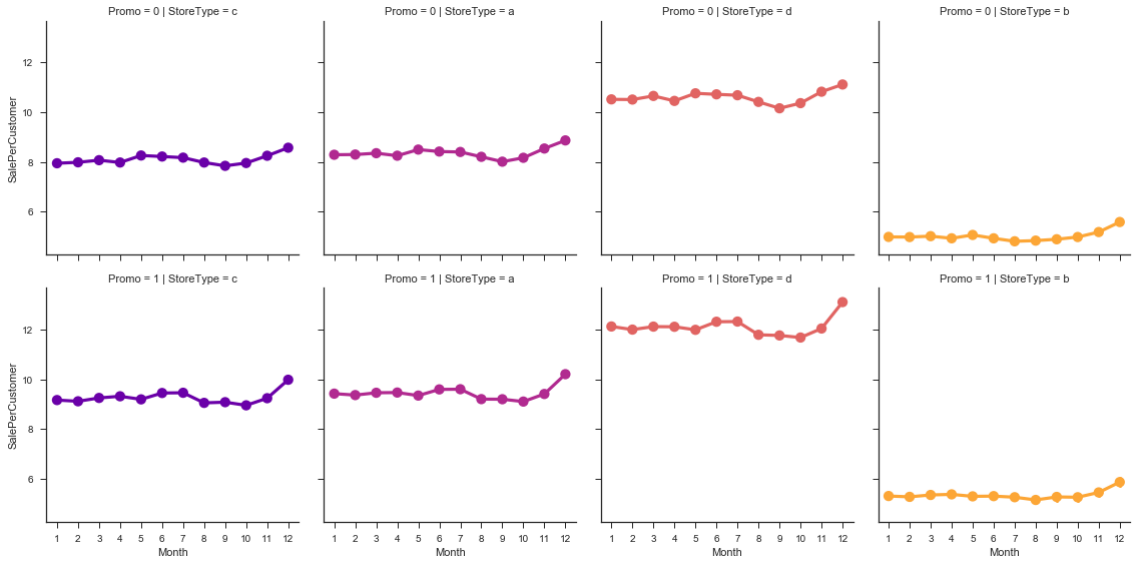
Επίσης είναι αξιοσημείωτο να αναφερθεί ότι πριν τις διακοπές των Χριστουγέννων υπάρχει μία σημαντική αύξηση στις πωλήσεις, αλλά οι εποχικότητες και οι τάσεις θα συζητηθούν αργότερα στην ανάλυση των χρονοσειρών.

Παρακάτω χρησιμοποιούμε τη νέα μεταβλητή που δημιουργήσαμε και εκτελούμε την ίδια διαδικασία με τα διαγράμματα. Διαπιστώνουμε όμως, ότι ενώ τα καταστήματα τύπου Β φαινόντουσαν αυτά με τον μεγαλύτερο τζίρο και την καλύτερη απόδοση, στην πραγματικότητα δεν είναι αλήθεια. Ο τύπος καταστήματος που εμφάνισε τις υψηλότερες πωλήσεις ανά πελάτη είναι ο 'd', με 12 € υπό προωθητική ενέργεια και 10 € χωρίς, ενώ τα καταστήματα τύπου 'a' και 'c' έφτασαν περίπου τα 9 €.

Οι χαμηλές πωλήσεις ανά πελάτη των καταστημάτων τύπου 'b' πιθανόν να οφείλεται στις αγορές χαμηλής αξίας πραγμάτων από τους πελάτες ή μικρής ποσότητας. Επιπλέον, προηγουμένως είδαμε ότι αυτός ο τύπος είχε συνολικά τον μικρότερο τζίρο και τους λιγότερους πελάτες σε όλη την περίοδο που εξετάζουμε.

```
# sale per customer trends
sns.factorplot(data = train_store, x = 'Month', y = "SalePerCustomer",
               col = 'StoreType', # per store type in cols
               palette = 'plasma',
               hue = 'StoreType',
               row = 'Promo', # per promo in the store in rows
               color = c)
```

```
<seaborn.axisgrid.FacetGrid at 0x7fb34cff1f98>
```

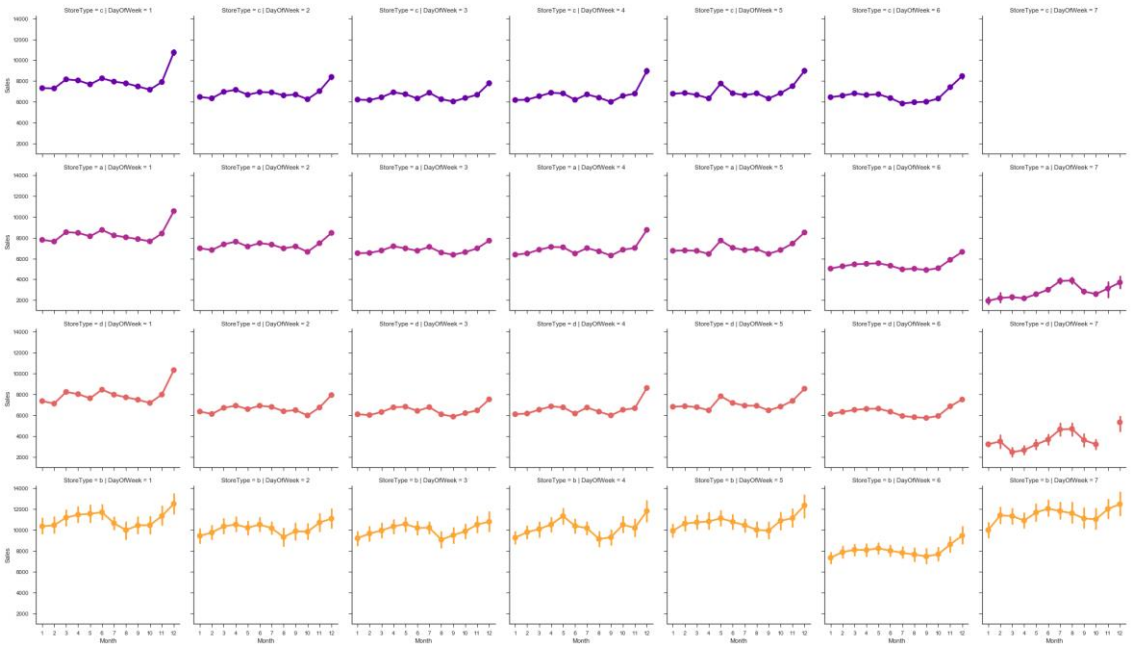


Σχήμα 46: Πωλήσεις ανά πελάτη ανάλογα με τον τύπο καταστήματος και την παρουσία ή όχι προωθητικής ενέργειας

Τα επόμενα διαγράμματα είναι ένας συνδυασμός των πωλήσεων ανά μήνα, αλλά χωρίζονται οριζόντια σύμφωνα με τους τέσσερις τύπους καταστημάτων και κάθετα ανά ημέρα της εβδομάδας.

```
# customers
sns.factorplot(data = train_store, x = 'Month', y = "Sales",
               col = 'DayOfWeek', # per store type in cols
               palette = 'plasma',
               hue = 'StoreType',
               row = 'StoreType', # per store type in rows
               color = c)
```

```
<seaborn.axisgrid.FacetGrid at 0x7fb348748e48>
```



Σχήμα 47: Μηνιαίες πωλήσεις ανάλογα με τον τύπο καταστήματος (οριζόντια) και ανάλογα την ημέρα της εβδομάδας (κάθετα)

Σύμφωνα με τα διαγράμματα, τα καταστήματα τύπου ‘c’ είναι κλειστά τις Κυριακές, ενώ τα υπόλοιπα είναι τον περισσότερο χρόνο ανοιχτά. Ενδιαφέρον παρουσιάζει ότι τα τύπου ‘d’ καταστήματα είναι κλειστά τις Κυριακές από τον Οκτώβριο έως τον Δεκέμβριο. Ωστόσο τα καταστήματα που είναι ανοιχτά τις Κυριακές είναι τα εξής:

```
# stores which are opened on Sundays
train_store[(train_store.Open == 1) & (train_store.DayOfWeek == 7)]['Store']
.unique()
```

```
array([  85,  122,  209,  259,  262,  274,  299,  310,  335,  353,  423,
        433,  453,  494,  512,  524,  530,  562,  578,  676,  682,  732,
        733,  769,  863,  867,  877,  931,  948, 1045, 1081, 1097, 1099])
```

Ολοκληρώνοντας την αρχική ανάλυση των δεδομένων, θα δημιουργήσουμε και κάποιες νέες μεταβλητές που θα περιγράφουν την περίοδο του χρόνου όπου ήταν σε εξέλιξη μία προωθητική ενέργεια και την περίοδο του χρόνου όπου λειτουργούσε ένα ανταγωνιστικό κατάστημα στην περιοχή.

```
# competition open time (in months)
train_store['CompetitionOpen'] = 12 * (train_store.Year - train_store.CompetitionOpenSinceYear)
+ \
        (train_store.Month - train_store.CompetitionOpenSinceMonth)

# Promo open time
train_store['PromoOpen'] = 12 * (train_store.Year - train_store.Promo2SinceYear) + \
        (train_store.WeekOfYear - train_store.Promo2SinceWeek) / 4.0

# replace NA's by 0
train_store.fillna(0, inplace = True)

# average PromoOpen time and CompetitionOpen time per store type
train_store.loc[:, ['StoreType', 'Sales', 'Customers', 'PromoOpen', 'CompetitionOpen']].groupby
('StoreType').mean()
```

	Sales	Customers	PromoOpen	CompetitionOpen
StoreType				
a	6925.697986	795.422370	12918.492198	7115.514452
b	10233.380141	2022.211825	17199.328069	11364.495244
c	6933.126425	815.538073	12158.636107	6745.418694
d	6822.300064	606.353935	10421.916846	9028.526526

Τα καταστήματα τύπου ‘a’ με την περισσότερη πελατεία και τον μεγαλύτερο τζίρο συμπεραίνουμε ότι δεν είναι τα πιο εκτεθειμένα στον ανταγωνισμό. Αντιθέτως ο τύπος ‘b’ είναι αυτός που έχει την μεγαλύτερη συνύπαρξη με τους ανταγωνιστές στην περιοχή του και παράλληλα έχει την μεγαλύτερη χρονική διάρκεια σε προωθητικές ενέργειες (Petrova (2016)).

5.5. Ανάλυση συσχετίσεων

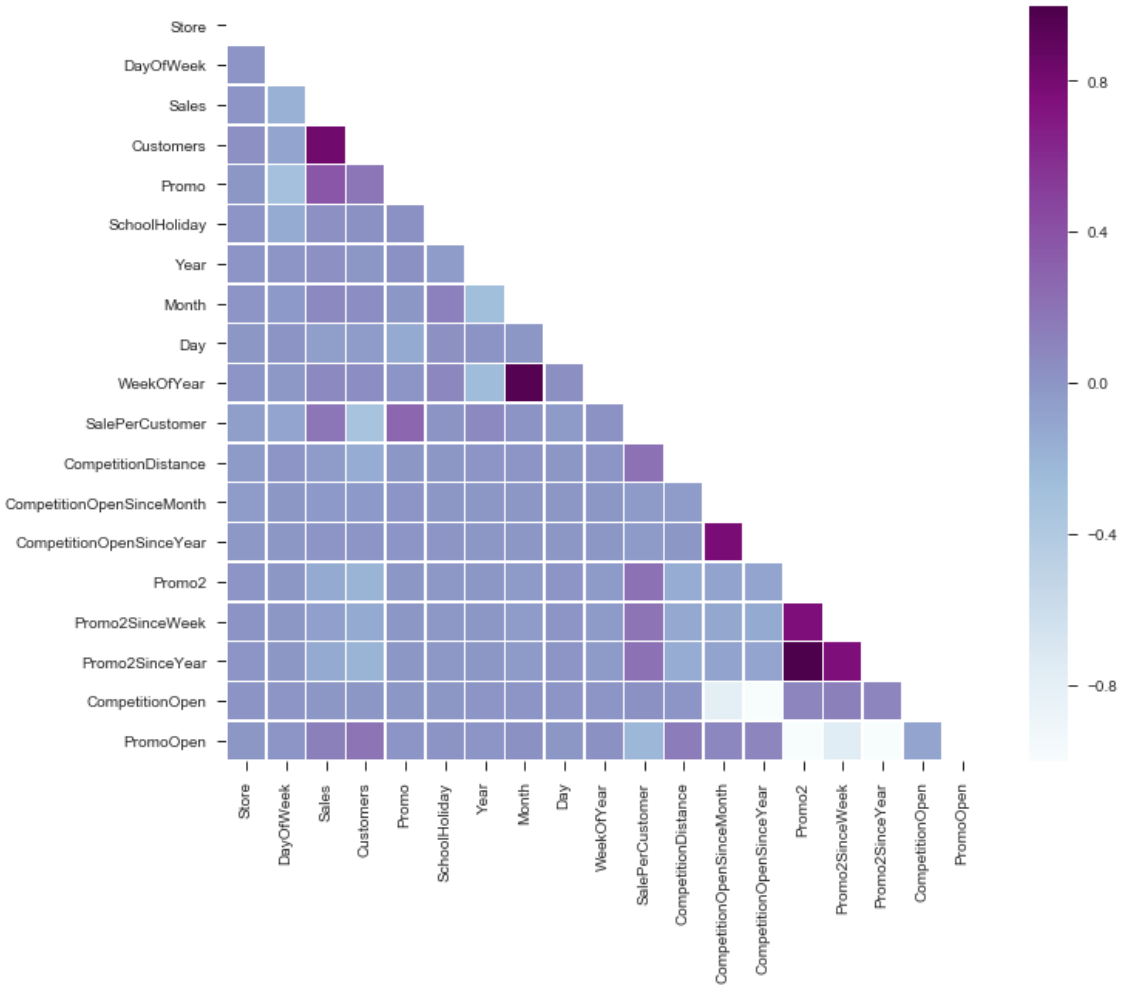
Έχουμε ολοκληρώσει την επεξεργασία και τη δημιουργία νέων μεταβλητών στα δεδομένα, οπότε τώρα μπορούμε να ελέγξουμε την συνολική συσχέτιση εκτυπώνοντας το θερμοδιάγραμμα (heatmap) seaborn.

```
# Compute the correlation matrix
# exclude 'Open' variable
corr_all = train_store.drop('Open', axis = 1).corr()

# Generate a mask for the upper triangle
mask = np.zeros_like(corr_all, dtype = np.bool)
mask[np.triu_indices_from(mask)] = True

# Set up the matplotlib figure
f, ax = plt.subplots(figsize = (11, 9))

# Draw the heatmap with the mask and correct aspect ratio
sns.heatmap(corr_all, mask = mask,
            square = True, linewidths = .5, ax = ax, cmap = "BuPu")
plt.show()
```



Σχήμα 48: Συσχέτιση των μεταβλητών μέσω του θερμοδιαγράμματος (heatmap) seaborn

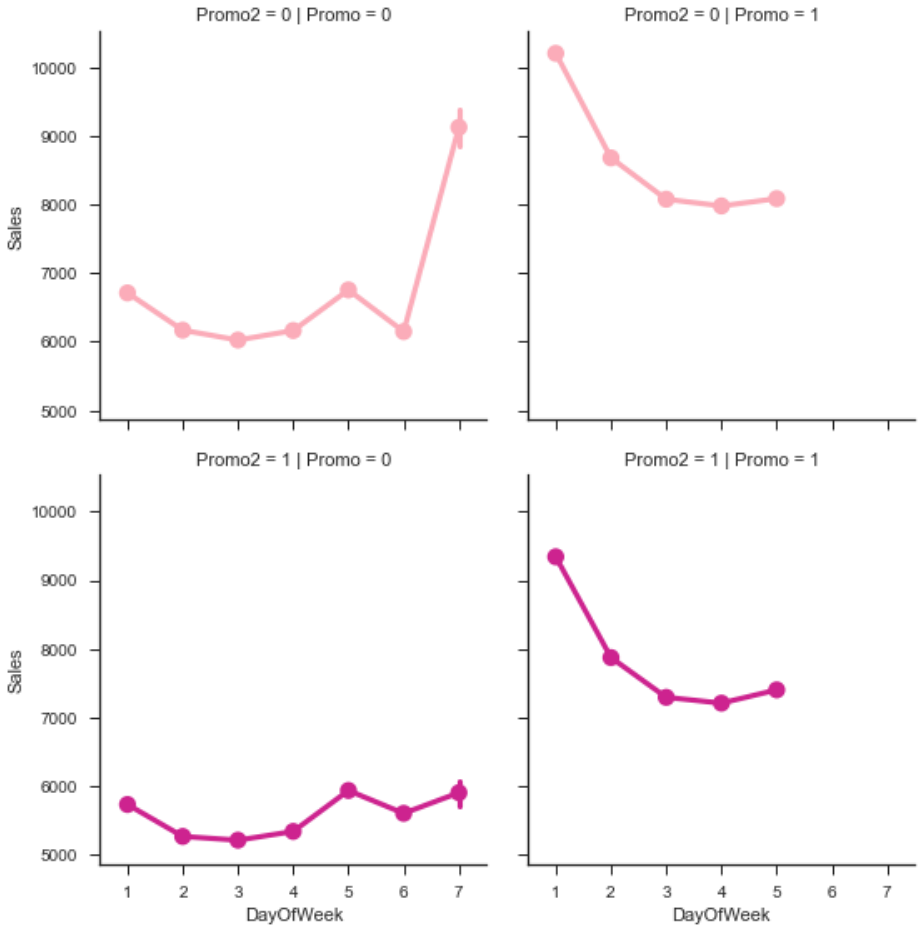
Όπως προαναφέρθηκε υπάρχει σημαντική θετική συσχέτιση μεταξύ του ύψους πωλήσεων και του αριθμού των πελατών σε ένα κατάστημα. Επίσης διακρίνεται θετική συσχέτιση μεταξύ του γεγονότος ότι το κατάστημα είναι σε προωθητική ενέργεια (Promo = 1) και του αριθμού των πελατών.

Ωστόσο, εάν το κατάστημα συνεχίσει μια διαδοχική προωθητική ενέργεια (Promo2 = 1), τότε η ποσότητα των πελατών και των πωλήσεων φαίνεται να παραμένει στάσιμη ή να μειώνεται, το οποίο περιγράφεται από την ελαφρά αρνητική συσχέτιση του θερμοδιαγράμματος. Παρόμοια αρνητική συσχέτιση παρατηρείται και μεταξύ της παρουσίας προωθητικής ενέργειας και της ημέρας της εβδομάδας.

Τέλος θα παρουσιαστούν διαγράμματα με τις πωλήσεις ανά πελάτη σε σχέση με την ημέρα της εβδομάδας και την παρουσία ή όχι προωθητικής ενέργειας.

```
# sale per customer trends
sns.factorplot(data = train_store, x = 'DayOfWeek', y = "Sales",
               col = 'Promo',
               row = 'Promo2',
               hue = 'Promo2',
               palette = 'RdPu')
```

```
<seaborn.axisgrid.FacetGrid at 0x7fb3445326a0>
```



Σχήμα 49: Πωλήσεις ανά πελάτη σε σχέση με την ημέρα της εβδομάδας και την παρουσία ή όχι προωθητικής ενέργειας

Στην περίπτωση καθόλου προωθητικής ενέργειας (Promo = 0 και Promo2 = 0), οι πωλήσεις τείνουν να αυξάνονται σημαντικά την Κυριακή. Ωστόσο πρέπει να σημειωθεί ότι τα καταστήματα τύπου ‘c’ δεν είναι ανοιχτά τις Κυριακές, οπότε τα δεδομένα είναι μόνο από τα καταστήματα τύπου ‘a’, ‘b’ και ‘d’.

Σε αντίθεση, τα καταστήματα που συμμετέχουν σε ενέργειες προώθησης, τείνουν να πραγματοποιούν τις περισσότερες πωλήσεις τη Δευτέρα. Αυτό το γεγονός θα μπορούσε να είναι ένας σημαντικός παράγοντας για τις καμπάνιες μάρκετινγκ της ROSSMANN. Την ίδια τάση ακολουθούν και τα καταστήματα που εμπλέκονται και στις δύο προωθητικές ενέργειες ταυτόχρονα (Promo = 1 και Promo2 = 1).

Το Promo2 από μόνο του δεν δείχνει να συσχετίζεται με κάποια αξιοσημείωτη αλλαγή στις πωλήσεις κατά τη διάρκεια της εβδομάδας. Αυτό αποδεικνύεται και από το ανοιχτό χρώμα του θερμοδιαγράμματος προηγουμένως (Petrova (2016)).

5.6. Συμπεράσματα από την διερεύνηση των δεδομένων

Τα κυριότερα συμπεράσματα που προέκυψαν από την μέχρι τώρα ανάλυση είναι ότι ο τύπος καταστήματος με τις περισσότερες πωλήσεις και με τον μεγαλύτερο αριθμό πελατών είναι ο τύπος ‘a’.

Τα καταστήματα τύπου ‘d’ έχουν τις υψηλότερες πωλήσεις ανά πελάτη, κάτι το οποίο πρέπει να λάβει υπ’όψιν η ROSSMANN και μπορεί να το εκμεταλλευτεί προτείνοντας μεγαλύτερη ποικιλία προϊόντων.

Οι χαμηλές πωλήσεις ανά πελάτη στα καταστήματα τύπου ‘b’ υποδηλώνει μία πιθανότητα οι άνθρωποι να πηγαίνουν εκεί μόνο για χαμηλής αξίας πράγματα. Αν και αυτός ο τύπος καταστημάτων είχε συνολικά τις μικρότερες πωλήσεις ανά πελάτη σε όλη την περίοδο, παρουσιάζει μία αξιόλογη δυναμική.

Οι πελάτες τείνουν να αγοράζουν περισσότερο τις Δευτέρες όταν υπάρχει προωθητική ενέργεια (Promo) και τις Κυριακές όταν δεν υπάρχει καμία (Promo = 0 και Promo2 = 0).

Τέλος το Promo2 από μόνο του δεν φαίνεται να συσχετίζεται με καμία σημαντική αλλαγή στις πωλήσεις (Petrova (2016)).

5.7. Ανάλυση χρονοσειρών ανά τύπο καταστήματος

Αυτό που διαφοροποιεί ένα πρόβλημα χρονοσειρών από ένα συνηθισμένο πρόβλημα παλινδρόμησης είναι η εξάρτηση του από τον χρόνο. Η βασική παραδοχή της γραμμικής παλινδρόμησης ότι οι παρατηρήσεις είναι ανεξάρτητες μεταξύ τους, δεν ισχύει στις χρονοσειρές. Οι χρονοσειρές έχουν κάποια μορφή τάσεων εποχικότητας, με τάσεις αύξησης ή μείωσης, δηλαδή παραλλαγές που αφορούν το συγκεκριμένο χρονικό πλαίσιο. Στη περίπτωση της ROSSMANN βλέπουμε μία τέτοια τάση στις διακοπές των Χριστουγέννων.

Στο παρόν κεφάλαιο θα δημιουργήσουμε μία ανάλυση χρονοσειρών που αφορά τους τύπους καταστημάτων και όχι μεμονωμένα καταστήματα. Το κύριο πλεονέκτημα αυτής της προσέγγισης είναι η απλότητα της παρουσίασης και ο συνδυασμός των διαφορετικών τάσεων και της εποχικότητας των δεδομένων. Επίσης θα αναλύσουμε τα στοιχεία των χρονοσειρών, δηλαδή τις τάσεις, τις εποχικότητες και την αυτοσυσχέτιση (autocorrelation). Στο τέλος της ανάλυσης θα προσπαθήσουμε να κατανοήσουμε τα δεδομένα, ώστε να είμαστε έτοιμοι να χρησιμοποιήσουμε τη μεθοδολογία Prophet για την πρόβλεψη.

5.7.1. Εποχικότητα

Επιλέγουμε τέσσερα καταστήματα, ένα από κάθε τύπο καταστημάτων για την αντιπροσώπευσή τους.

Το κατάστημα νούμερο 2 από τα καταστήματα τύπου ‘a’.

Το κατάστημα νούμερο 85 από τα καταστήματα τύπου ‘b’.

Το κατάστημα νούμερο 1 από τα καταστήματα τύπου ‘c’.

Το κατάστημα νούμερο 13 από τα καταστήματα τύπου ‘d’.

Για την καλύτερη απεικόνιση των τάσεων, μετατρέπουμε τα δεδομένα από ημερήσια σε εβδομαδιαία, χρησιμοποιώντας τη μέθοδο resample.

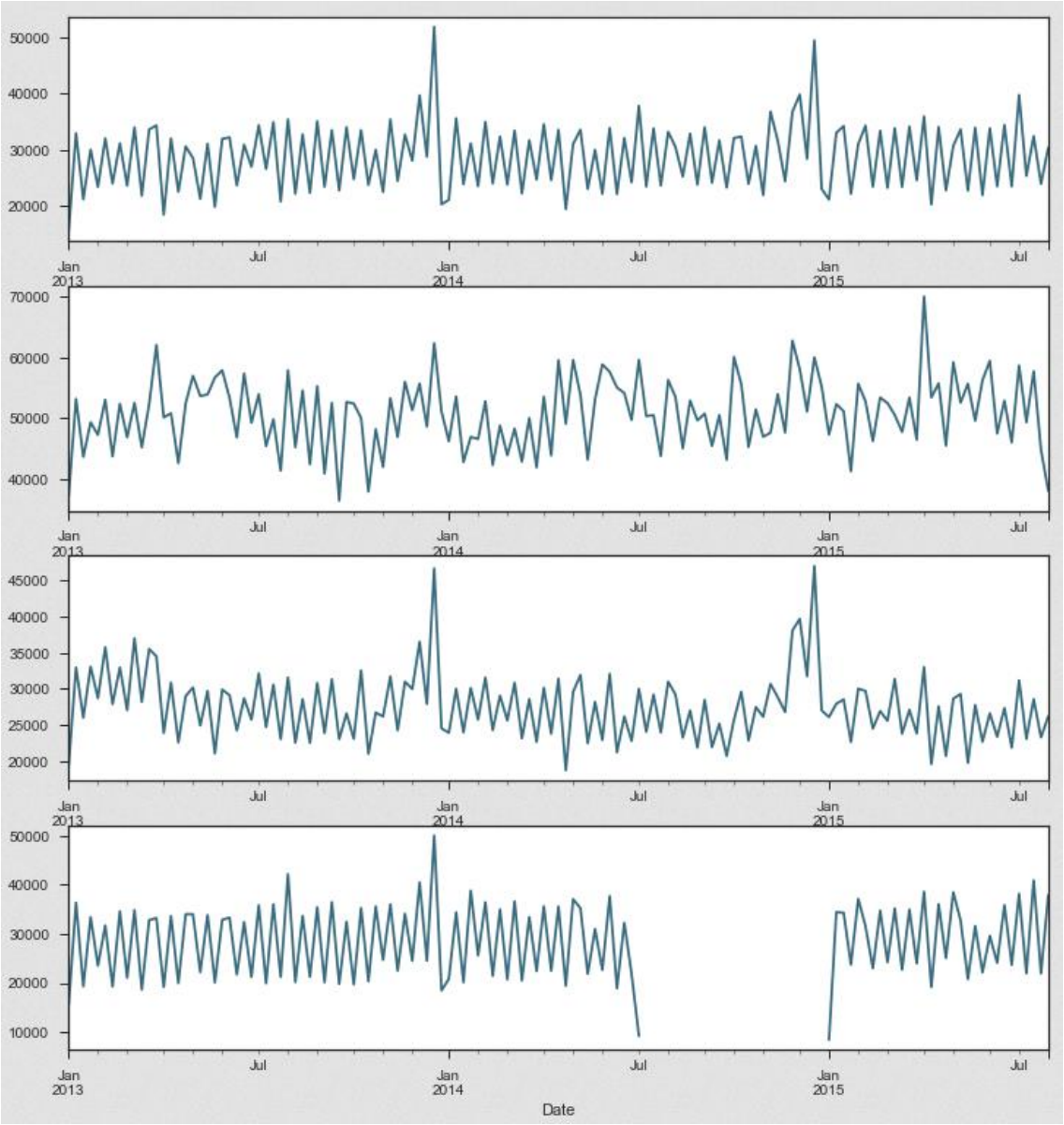
```
# preparation: input should be float type
train['Sales'] = train['Sales'] * 1.0

# store types
sales_a = train[train.Store == 2]['Sales']
sales_b = train[train.Store == 85]['Sales'].sort_index(ascending = True)
# solve the reverse order
sales_c = train[train.Store == 1]['Sales']
sales_d = train[train.Store == 13]['Sales']

f, (ax1, ax2, ax3, ax4) = plt.subplots(4, figsize = (12, 13))

# store types
sales_a.resample('W').sum().plot(color = c, ax = ax1)
sales_b.resample('W').sum().plot(color = c, ax = ax2)
sales_c.resample('W').sum().plot(color = c, ax = ax3)
sales_d.resample('W').sum().plot(color = c, ax = ax4)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fb31e218e48>
```



Σχήμα 50: Απεικόνιση των τάσεων των δεδομένων

Παρατηρούμε ότι οι πωλήσεις των καταστημάτων που ανήκουν στους τύπους ‘a’ και ‘c’ τείνουν να εκτινάσσονται την περίοδο των Χριστουγέννων και να μειώνονται αμέσως μετά από αυτή. Την ίδια τάση φαίνεται να έχει και το κατάστημα νούμερο 13 που ανήκει στον τύπο ‘d’, αλλά δεν υπάρχουν δεδομένα από τον Ιούλιο 2014 έως τον Ιανουάριο 2015, πιθανώς γιατί ήταν κλειστό. Το κατάστημα νούμερο 85 (τύπος ‘b’) δεν δείχνει να ακολουθεί κάποια συγκεκριμένη τάση εποχικότητας, όπως τα άλλα.

5.7.2. Ετήσια εποχικότητα

Τώρα θα εξετάσουμε την παρουσία εποχικότητας στις σειρές. Θα χρησιμοποιήσουμε την μέθοδο seasonal_decompose( ), η οποία διαχωρίζει τα δεδομένα

των χρονοσειρών σε εποχικότητες, τάσεις και υπολειπόμενες συμπεριφορές ή θόρυβο, δίνοντάς μας τη δυνατότητα καλύτερης επίγνωσης των μοτίβων που ακολουθούν.

```
f, (ax1, ax2, ax3, ax4) = plt.subplots(4, figsize = (12, 13))

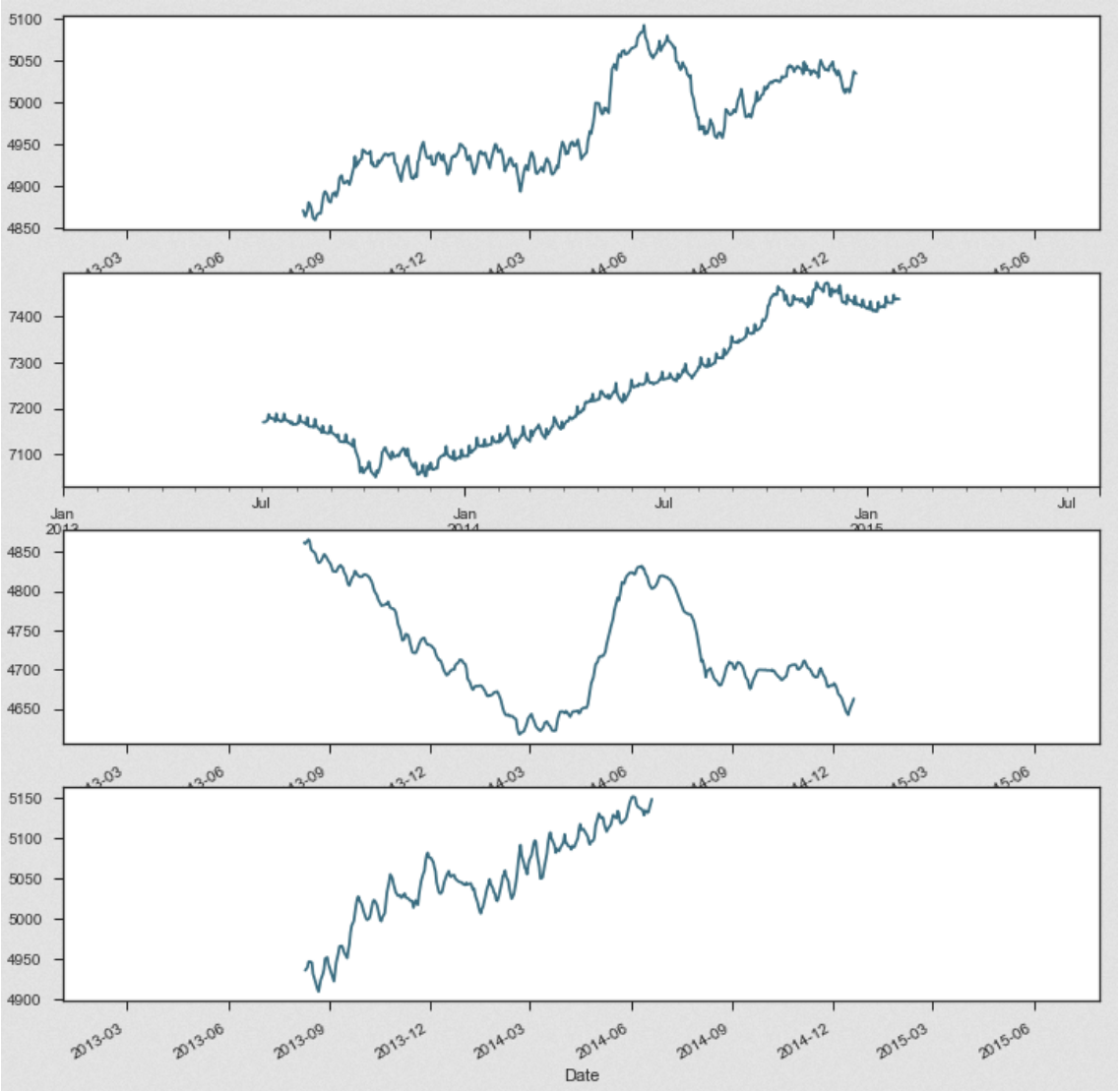
# monthly
decomposition_a = seasonal_decompose(sales_a, model = 'additive',
freq = 365)
decomposition_a.trend.plot(color = c, ax = ax1)

decomposition_b = seasonal_decompose(sales_b, model = 'additive',
freq = 365)
decomposition_b.trend.plot(color = c, ax = ax2)

decomposition_c = seasonal_decompose(sales_c, model = 'additive',
freq = 365)
decomposition_c.trend.plot(color = c, ax = ax3)

decomposition_d = seasonal_decompose(sales_d, model = 'additive',
freq = 365)
decomposition_d.trend.plot(color = c, ax = ax4)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fb31de4e3c8>
```



Σχήμα 51: Απεικόνιση της εποχικότητας των χρονοσειρών χρησιμοποιώντας τη μέθοδο seasonal_decompose

Συνολικά οι πωλήσεις όλων των καταστημάτων φαίνονται να αυξάνονται, εκτός από το κατάστημα που ανήκει στον τύπο ‘c’ (τρίτο από πάνω), στο οποίο μειώνονται. Το κατάστημα που ανήκει στον τύπο ‘a’, παρόλο που συνολικά από τα δεδομένα ο τύπος του έχει τις περισσότερες πωλήσεις, δείχνει να ακολουθεί την φθίνουσα τροχιά που έχει ήδη διαγράψει το κατάστημα τύπου ‘c’.

5.7.3. Αυτοσυσχέτιση

Σε αυτή την ενότητα θα επεξηγήσουμε τα διαγράμματα της λειτουργία αυτοσυσχέτισης (Autocorrelation Function – ACF) και της λειτουργίας μερικής αυτοσυσχέτισης (Partial Autocorrelation Function – PACF).

Η ACF είναι ένα μέτρο συσχέτισης μεταξύ των χρονοσειρών και μίας μεταχρονισμένης εκδοχής (lagged version) των ιδίων. Για παράδειγμα εάν εφαρμόσουμε lag = 5, η ACF θα συγκρίνει τις σειρές στην χρονική στιγμή $t_1 \ldots t_n$ με τις χρονικές στιγμές $t_{1-5} \ldots t_{n-5}$ (τα t_{1-5} και t_n είναι τα τελικά σημεία).

Η PACF από την άλλη μεριά, μετράει τη συσχέτιση μεταξύ των χρονοσειρών και μίας μεταχρονισμένης εκδοχής (lagged version) των ιδίων, αλλά αφού εξαλειφθούν οι παραλλαγές που εξηγούνται από τις παρεμβάσεις των συγκρίσεων. Για παράδειγμα εάν εφαρμόσουμε lag = 5, θα ελέγξει τη συσχέτιση, αλλά αφού αφαιρέσει τις επιδράσεις που ήδη εξηγούνται από τις χρονικές περιόδους 1 έως 4.

```
# figure for subplots
plt.figure(figsize = (12, 8))

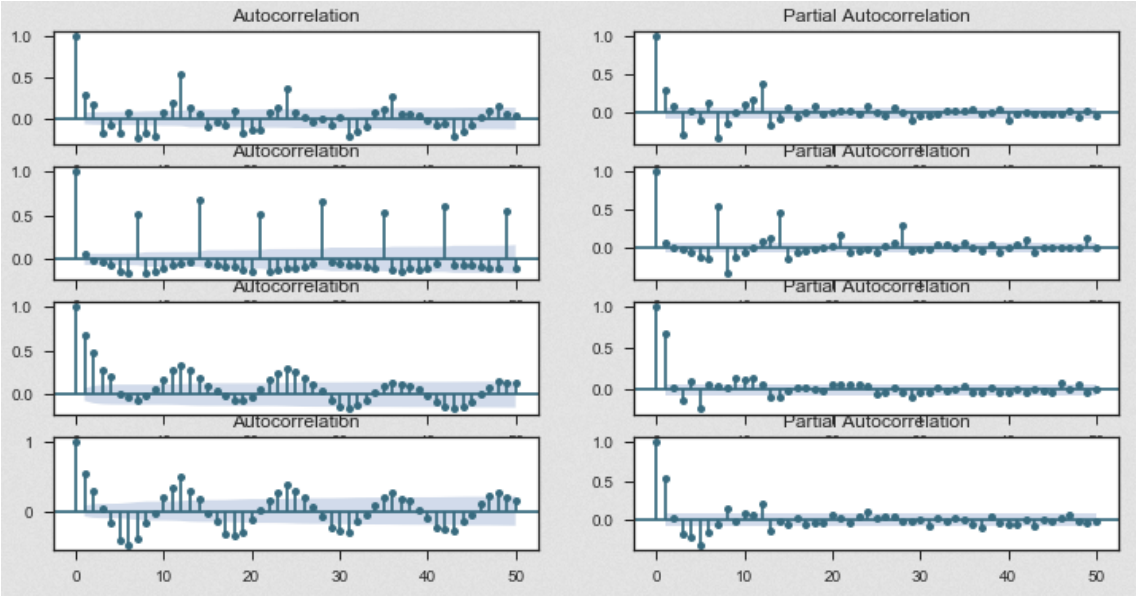
# acf and pacf for A
plt.subplot(421); plot_acf(sales_a, lags = 50, ax = plt.gca(), color = c)
plt.subplot(422); plot_pacf(sales_a, lags = 50, ax = plt.gca(), color = c)

# acf and pacf for B
plt.subplot(423); plot_acf(sales_b, lags = 50, ax = plt.gca(), color = c)
plt.subplot(424); plot_pacf(sales_b, lags = 50, ax = plt.gca(), color = c)

# acf and pacf for C
plt.subplot(425); plot_acf(sales_c, lags = 50, ax = plt.gca(), color = c)
plt.subplot(426); plot_pacf(sales_c, lags = 50, ax = plt.gca(), color = c)

# acf and pacf for D
plt.subplot(427); plot_acf(sales_d, lags = 50, ax = plt.gca(), color = c)
plt.subplot(428); plot_pacf(sales_d, lags = 50, ax = plt.gca(), color = c)

plt.show()
```



Σχήμα 52: Συσχέτιση των χρονοσειρών με τις ίδιες μεταχρονισμένες κατά x χρονικές περιόδους

Κάθε οριζόντιο ζευγάρι από τα παραπάνω διαγράμματα αντιστοιχεί σε έναν τύπο καταστήματος από τον ‘a’ έως τον ‘d’. Γενικά αυτά τα διαγράμματα απεικονίζουν τη συσχέτιση μεταξύ των χρονοσειρών και των ίδιων μεταχρονισμένων κατά x χρονικές περιόδους.

Υπάρχουν δύο κοινά πράγματα για κάθε οριζόντιο ζευγάρι, η μη τυχαία κατανομή των χρονοσειρών και η υψηλή καθυστέρηση lag = 1.

Οι τύποι καταστήματος ‘a’ και ‘b’ δείχνουν εποχικότητα σε συγκεκριμένα χρονικά διαστήματα. Ο τύπος ‘a’ έχει κάθε δωδέκατο χρονικό διάστημα θετικές εκτινάξεις, δηλαδή έχει στο 12ο, στο 24ο, στο 36ο χρονικό διάστημα κ.ο.κ., ενώ ο τύπος ‘b’ έχει μία εβδομαδιαία τάση με θετικές εκτινάξεις στο 7ο, στο 14ο, στο 21ο και στο 28ο χρονικό διάστημα.

Οι τύποι ‘c’ και ‘d’ παρουσιάζουν μία περίπλοκη συσχέτιση, φαίνεται πως κάθε παρατήρηση είναι συσχετισμένη μόνο με τη γειτονική της (Petrova (2016)).

6. Μοντέλο πρόβλεψης

Σε αυτό το κεφάλαιο θα παρουσιαστεί ο τρόπος με τον οποίο κατασκευάστηκε το μοντέλο πρόβλεψης πωλήσεων, του οποίο το kernel ήταν δεύτερο σε ψήφους κοινού, με 160 ψήφους. Το kernel αυτό το δημιούργησε η Elena Petrova το 2017 σε γλώσσα προγραμματισμού Python και είναι αναρτημένο στην διαδικτυακή πλατφόρμα του Kaggle και μπορείτε να το βρείτε και στο Παράρτημα Β. Η Elena Petrova, αρχικά πραγματοποίησε μία πρόβλεψη των πωλήσεων με τη μέθοδο Prophet, η οποία μέθοδος παρουσιάστηκε πρόσφατα από τη Facebook.

Η μέθοδος αυτή επιλέχτηκε διότι έχει πολύ καλά χαρακτηριστικά για την μοντελοποίηση των αργιών, των οποίων τα δεδομένα χρησιμοποιούνται και στην περίπτωση της ROSSMANN. Στη συνέχεια πέρα από την Prophet, πραγματοποίησε και μία πρόβλεψη των πωλήσεων με τη μέθοδο Extreme Gradient Boosting (XGBoost). Τέλος θα παρουσιαστούν τα πλεονεκτήματα και τα μειονεκτήματα της Seasonal ARIMA και της Prophet (Petrova (2017)).

6.1. Εισαγωγή στη μέθοδο Prophet και XGBoost

6.1.1. Μέθοδος Prophet

Δεν είναι δυνατή η επίλυση όλων των προβλημάτων πρόβλεψης με την ίδια μέθοδο. Υπάρχουν διάφορες μέθοδοι ανάλογα με την φύση του προβλήματος. Η Prophet είναι μία βελτιστοποιημένη μέθοδος για επιχειρηματικές προβλέψεις, οι οποίες συνήθως έχουν κάποιο από τα παρακάτω χαρακτηριστικά:

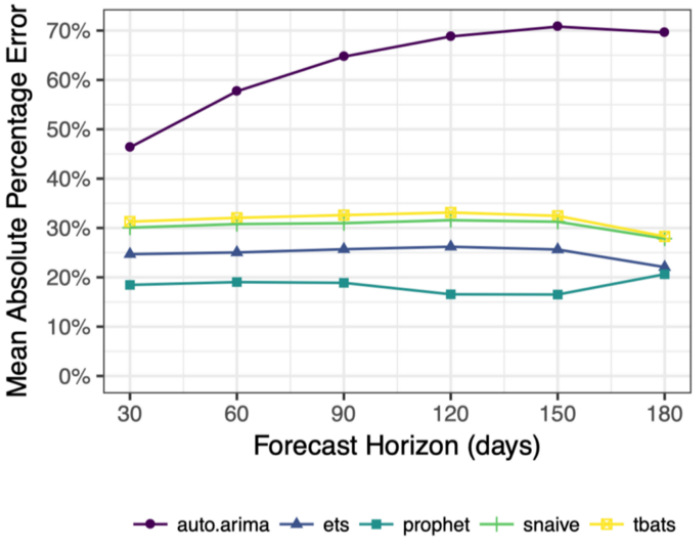
- Ωριαίες, καθημερινές ή εβδομαδιαίες παρατηρήσεις με τουλάχιστον μερικούς μήνες (κατά προτίμηση ένα χρόνο) ιστορίας.
- Εποχικότητα της ανθρώπινης κλίμακας, όπως ποια ημέρα της εβδομάδας είναι και ποια ημερομηνία του έτους είναι.
- Αργίες, που συμβαίνουν σε ακανόνιστα διαστήματα, τα οποία όμως είναι γνωστά εκ των προτέρων (πχ. Χριστούγεννα, Πάσχα).
- Έναν λογικό αριθμό ελλιπών παρατηρήσεων ή εσφαλμένων καταχωρήσεων.
- Ιστορικές μεταβολές τάσεων, για παράδειγμα εξαιτίας λανσαρίσματος νέων προϊόντων.
- Τάσεις, οι οποίες δεν είναι γραμμικές καμπύλες ανάπτυξης, αλλά η τάση αγγίζει ένα όριο (φτάνει σε κορεσμό) (Taylor and Letham (2017)).

Ουσιαστικά, η Prophet είναι μία διαδικασία αθροιστικής παλινδρόμησης με τέσσερα βασικά στοιχεία:

$$y(t) = g(t) + s(t) + h(t) + \epsilon$$

- $g(t)$: Μοντελοποιεί την τάση, το οποίο περιγράφει μία μακροπρόθεσμη αύξηση ή μείωση των δεδομένων. Το Prophet ενσωματώνει δύο μοντέλα τάσεων, ένα μοντέλο λογικής ανάπτυξης (κορεσμού) και ένα μερικώς γραμμικό μοντέλο, ανάλογα με το είδος του προβλήματος πρόβλεψης. Το Prophet εντοπίζει αυτόματα τις αλλαγές στις τάσεις, επιλέγοντας τα σημεία αλλαγής από τα δεδομένα.
- $s(t)$: Μοντελοποιεί την εποχικότητα με τη χρήση σειρών Fourier, το οποίο περιγράφει πως τα δεδομένα επηρεάζονται από εποχιακούς παράγοντες, όπως η χρονική στιγμή στο έτος.
- $h(t)$: Μοντελοποιεί τις επιπτώσεις των διακοπών ή των μεγάλων γεγονότων που επηρεάζουν τις χρονοσειρές των επιχειρήσεων (πχ. Χριστούγεννα, Black Friday, τελικός Champions league, κλπ). Η λίστα των σημαντικών αργιών ορίζεται χειροκίνητα από τον χρήστη.
- ϵ: Αντιπροσωπεύει τον όρο του μη αναστρέψιμου σφάλματος.

Η μέθοδος Prophet αξιολογεί αυτόματα τις επιδόσεις πρόβλεψης και προειδοποιεί όπου απαιτείται χειροκίνητη παρέμβαση. Ένας από τους ευκολότερους τρόπους αξιολόγησης είναι να ορίσουμε ένα επίπεδο με μερικές απλές μεθόδους πρόβλεψης (π.χ. εποχιακή αφελής τάση, μέσος όρος δείγματος, κλπ.). Είναι χρήσιμο να συγκρίνετε τις απλές και προηγμένες μεθόδους πρόβλεψης για να είστε σε θέση να προσδιορίσετε εάν μπορεί να επιτευχθεί πρόσθετη απόδοση χρησιμοποιώντας ένα πιο πολύπλοκο μοντέλο. Μερικές φορές, ίσως είναι καλύτερο να χρησιμοποιήσετε απλώς μία απλοϊκή μέθοδο.



Σχήμα 53: Η Prophet έχει χαμηλότερο σφάλμα πρόβλεψης από τις άλλες μεθόδους

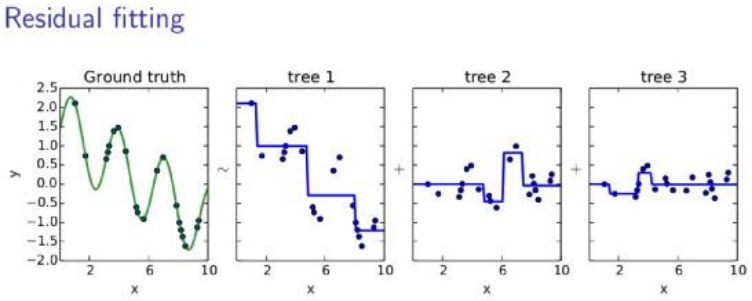
Από την άλλη, αυτή η μέθοδος αξιολογεί την απόδοση της χρησιμοποιώντας μια διαδικασία που ονομάζεται προσομοίωση ιστορικών προβλέψεων (SHFs). Τα SHF δουλεύουν με την παραγωγή 'K' προγνώσεων σε διάφορα χρονικά σημεία της ιστορίας, τα οποία στη συνέχεια ταιριάζουν στο μοντέλο του αναμενόμενου σφάλματος σε διαφορετικούς ορίζοντες πρόβλεψης (Liu (2018)).

6.1.2. Μέθοδος XGBoost

Ο Tianqi Chen τον Μάρτιο του 2014, πραγματοποιώντας μία έρευνα για το Deep Machine Learning Community (DMLC) Group, δημιούργησε το XGBoost, το οποίο σημαίνει Ακραία Βαθμιδωτή Ενίσχυση (Extreme Gradient Boosting) και βασίζεται στη μέθοδο πρόβλεψης Βαθμιδωτής Ενίσχυσης (Gradient Boosting).

Το Gradient Boosting είναι μια μέθοδος που προσθέτει διαδοχικά τα δεδομένα εκπαίδευσης και τους αποδίδει ένα συντελεστή βαρύτητας. Ωστόσο, αντί να αναθέσει διαφορετικά βάρη στους ταξινομητές μετά από κάθε επανάληψη, η μέθοδος αυτή προσαρμόζει το νέο μοντέλο στα νέα υπολείμματα της προηγούμενης πρόβλεψης και στη συνέχεια ελαχιστοποιεί την απώλεια όταν προσθέτει την τελευταία κάθε φορά πρόβλεψη. Ο στόχος είναι να βρεθούν οι βέλτιστες παράμετροι που έχουν τη μεγαλύτερη μείωση στη λειτουργία απώλειας. Αυτός είναι ο τρόπος με τον οποίο αυτή η μέθοδος επιχειρεί να ελαχιστοποιήσει τα σφάλματα. Έτσι, στο τέλος ενημερώνεται το μοντέλο χρησιμοποιώντας βαθμιδωτή (gradient) μείωση, δικαιολογώντας το όνομα Gradient Boosting (Βαθμιδωτή Ενίσχυση). Το άθροισμα των προγνωστικών μας γίνεται ολοένα και ισχυρότερο μετά από κάθε βήμα. Αυτή η διαδικασία επαναλαμβάνεται

μέχρι να κατασκευαστεί μία τελική πρόβλεψη. Το παρακάτω διάγραμμα ‘Ground Truth’ απεικονίζει ένα σύνολο δεδομένων, με μια γραμμή που τρέχει μέσα από κάθε ένα από τα σημεία.



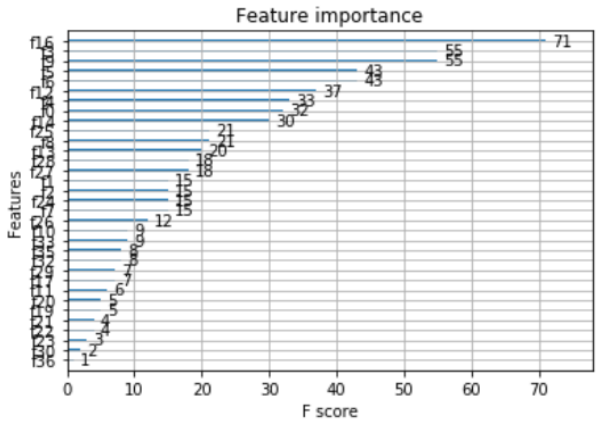
Σχήμα 54: Ο κύκλος μάθησης στο Gradient Boosting

Το ‘tree 1’ είναι η καλύτερη εφαρμογή των δεδομένων. Το ‘tree 2’ είναι μια καμπύλη που σχεδιάζει τα σφάλματα από την όδευση του ‘tree 1’. Αυτά τα σφάλματα βασίζονται στον τρόπο με τον οποίο το ‘tree 1’ παρερμήνευσε το αρχικό διάγραμμα (το Ground truth σε αυτή την περίπτωση). Τέλος, το ‘tree 3’ είναι ένας συνδυασμός του ‘tree 1’ και του ‘tree 2’. Αυτός είναι ο κύκλος μίας μάθησης στο Gradient Boosting. Συνδυάζοντας τις επαναλήψεις μάθησης, το τελικό μας μοντέλο μπορεί να αντιληφθεί ένα μεγάλο μέρος του σφάλματος από το αρχικό μοντέλο και το μειώνει με την πάροδο του χρόνου. Αυτή η μέθοδος υποστηρίζει τόσο προβλήματα παλινδρόμησης όσο και ταξινόμησης. (Lutins (2017)).

Το XGBoost συγκεκριμένα, υλοποιεί αυτόν τον αλγόριθμο για την ενίσχυση των δέντρων αποφάσεων με έναν πρόσθετο σημαντικό όρο ρύθμισης (το Penalty, το οποίο δικαιολογεί και την έννοια του Boosting) στην Objective λειτουργία (Loss Function + Penalty).

```
xgb.plot_importance(my_model)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x23a5fb65400>
```



Σχήμα 55: Διάγραμμα βαρύτητας των μεταβλητών (plot_importance)

Ένα σπουδαίο χαρακτηριστικό που προσφέρεται από το XGBoost είναι το διάγραμμα plot_importance που μας παρέχει μια γραφική παράσταση των χαρακτηριστικών του μοντέλου μας και τη βαρύτητα της σημασίας τους. Αυτή είναι μια μεγάλη προσθήκη για να αναλύσουμε το μοντέλο μας και να αξιολογήσουμε τα χαρακτηριστικά του. Μπορούμε να αναγνωρίσουμε ποιες μεταβλητές έχουν σημαντικό ρόλο στην πρόβλεψή μας και ποιες όχι, δίνοντας μας τη δυνατότητα είτε να τις χρησιμοποιήσουμε στο μοντέλο μας, είτε να τις αγνοήσουμε (εάν είναι ασήμαντες). Ένα μειονέκτημα της μεθόδου XGBoost είναι ότι ως δεδομένα εισόδου δέχεται μόνο αριθμούς, δεν μπορεί να αντιληφθεί οποιασδήποτε άλλης φύσης δεδομένα, όπως πχ ημερομηνίες, γράμματα ή ολόκληρες λέξεις (Raghu (2018)).

6.2. Ανάλυση χρονοσειρών και πρόβλεψη πωλήσεων με τη μέθοδο Prophet

Σε αυτή την ενότητα θα πραγματοποιήσουμε την πρόβλεψη πωλήσεων για τις επόμενες έξι εβδομάδες του πρώτου καταστήματος. Η μέθοδος πρόβλεψης για δεδομένα χρονοσειρών που θα χρησιμοποιηθεί, είναι η Prophet, η οποία πρόσφατα, όπως προαναφέρθηκε, δημοσιεύθηκε από την ομάδα Core Data Science της Facebook. Η μέθοδος βασίζεται σε ένα πρόσθετο μοντέλο, όπου μη γραμμικές τάσεις έχουν προσαρμοστεί σε ετήσιες και εβδομαδιαίες εποχικότητες με την προσθήκη των αργιών. Δίνει τη δυνατότητα να εκτελεστούν αυτοματοποιημένες προβλέψεις σε γλώσσα προγραμματισμού Python 3, οι οποίες έχουν ήδη εφαρμοστεί σε γλώσσα R.

```
# importing data
df = pd.read_csv("../input/train.csv",
                 low_memory = False)

# remove closed stores and those with no sales
df = df[(df["Open"] != 0) & (df['Sales'] != 0)]

# sales for the store number 1 (StoreType C)
sales = df[df.Store == 1].loc[:, ['Date', 'Sales']]

# reverse to the order: from 2013 to 2015
sales = sales.sort_index(ascending = False)

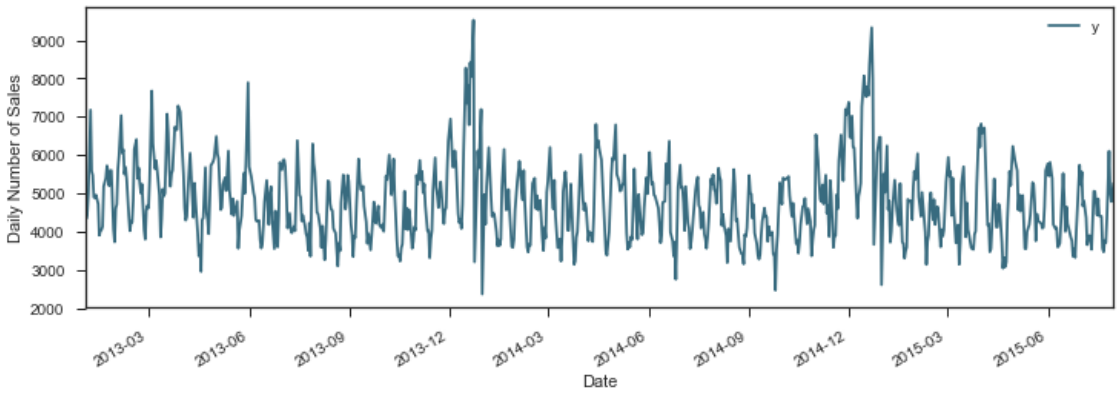
# to datetime64
sales['Date'] = pd.DatetimeIndex(sales['Date'])
sales.dtypes
```

```
Date     datetime64[ns]
Sales             int64
dtype: object
```

```
# from the prophet documentation every variables should have specific names
sales = sales.rename(columns = {'Date': 'ds',
                                'Sales': 'y'})
sales.head()
```

	ds	y
1014980	2013-01-02	5530
1013865	2013-01-03	4327
1012750	2013-01-04	4486
1011635	2013-01-05	4997
1009405	2013-01-07	7176

```
# plot daily sales
ax = sales.set_index('ds').plot(figsize = (12, 4), color = c)
ax.set_ylabel('Daily Number of Sales')
ax.set_xlabel('Date')
plt.show()
```



Σχήμα 56: Διάγραμμα ημερησίων πωλήσεων για όλη την περίοδο δεδομένων

Η μέθοδος Prophet μας επιτρέπει να μοντελοποιήσουμε τις αργίες που υπάρχουν στα δεδομένα μας, οπότε σε αυτό το σημείο θα εισάγουμε τα αντίστοιχα δεδομένα μας που αφορούν τις αργίες. Η μεταβλητή StateHoliday των δεδομένων της ROSSMANN υποδηλώνει τις εθνικές αργίες κατά τις οποίες όλα τα καταστήματα φυσιολογικά είναι κλειστά. Επίσης υπάρχουν και οι σχολικές αργίες στα δεδομένα μας κατά τις οποίες κάποια συγκεκριμένα καταστήματα είναι επίσης κλειστά. Δημιουργούμε έναν πίνακα που να αντιστοιχεί τις ημερομηνίες με τις αργίες.

```
# create holidays dataframe
state_dates = df[(df.StateHoliday == 'a') | (df.StateHoliday == 'b') & (d
f.StateHoliday == 'c')].loc[:, 'Date'].values
school_dates = df[df.SchoolHoliday == 1].loc[:, 'Date'].values

state = pd.DataFrame({'holiday': 'state_holiday',
                      'ds': pd.to_datetime(state_dates)})
school = pd.DataFrame({'holiday': 'school_holiday',
                       'ds': pd.to_datetime(school_dates)})

holidays = pd.concat((state, school))
holidays.head()
```

	ds	holiday
0	2015-06-04	state_holiday
1	2015-06-04	state_holiday
2	2015-06-04	state_holiday
3	2015-06-04	state_holiday
4	2015-06-04	state_holiday

Πριν την εκτέλεση, παραμετροποιούμε το μοντέλο της Prophet και αλλάζουμε το διάστημα αβεβαιότητας από 80% σε 95%. Εκτυπώνουμε την πρώτη εβδομάδα της πρόβλεψης.

```
# set the uncertainty interval to 95% (the Prophet default is 80%)
my_model = Prophet(interval_width = 0.95,
                   holidays = holidays)
my_model.fit(sales)

# dataframe that extends into future 6 weeks
future_dates = my_model.make_future_dataframe(periods = 6*7)

print("First week to forecast.")
future_dates.tail(7)
```

```
First week to forecast.
```

	ds
816	2015-09-05
817	2015-09-06
818	2015-09-07
819	2015-09-08
820	2015-09-09
821	2015-09-10
822	2015-09-11

```
# predictions
forecast = my_model.predict(future_dates)

# preditions for last week
forecast[['ds', 'yhat', 'yhat_lower', 'yhat_upper']].tail(7)
```

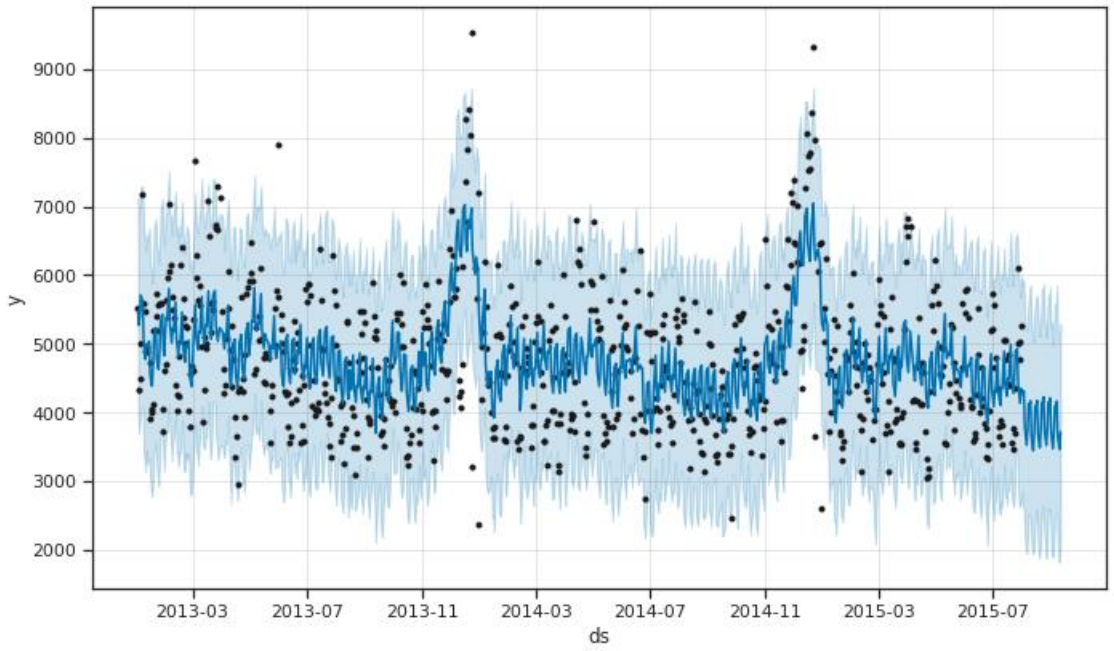
	ds	yhat	yhat_lower	yhat_upper
816	2015-09-05	4092.315126	2516.799472	5590.759650
817	2015-09-06	4087.136609	2522.210456	5550.350687
818	2015-09-07	4161.901669	2532.280427	5843.295785
819	2015-09-08	3664.283500	2064.019465	5210.161131
820	2015-09-09	3551.082173	1838.992757	5081.631360
821	2015-09-10	3462.584725	1814.992113	5052.909537
822	2015-09-11	3717.252800	2130.510109	5291.021113

Το αντικείμενο πρόβλεψης εδώ είναι ένα νέο πλαίσιο δεδομένων που περιλαμβάνει μια στήλη με την πρόβλεψη yhat, καθώς και δύο στήλες που ορίζουν τα διαστήματα αβεβαιότητας (yhat_lower και yhat_upper).

```
fc = forecast[['ds', 'yhat']].rename(columns = {'Date': 'ds', 'Forecast':
'yhat'})
```

Στο παρακάτω διάγραμμα της Prophet (σχήμα 57), εκτυπώνονται οι παρατηρούμενες τιμές των χρονοσειρών (μαύρες κουκκίδες), οι τιμές πρόβλεψης (μπλε γραμμή) και τα διαστήματα αβεβαιότητας της πρόβλεψής μας (γαλάζια περιοχή).

```
# visualizing predicions
my_model.plot(forecast);
```

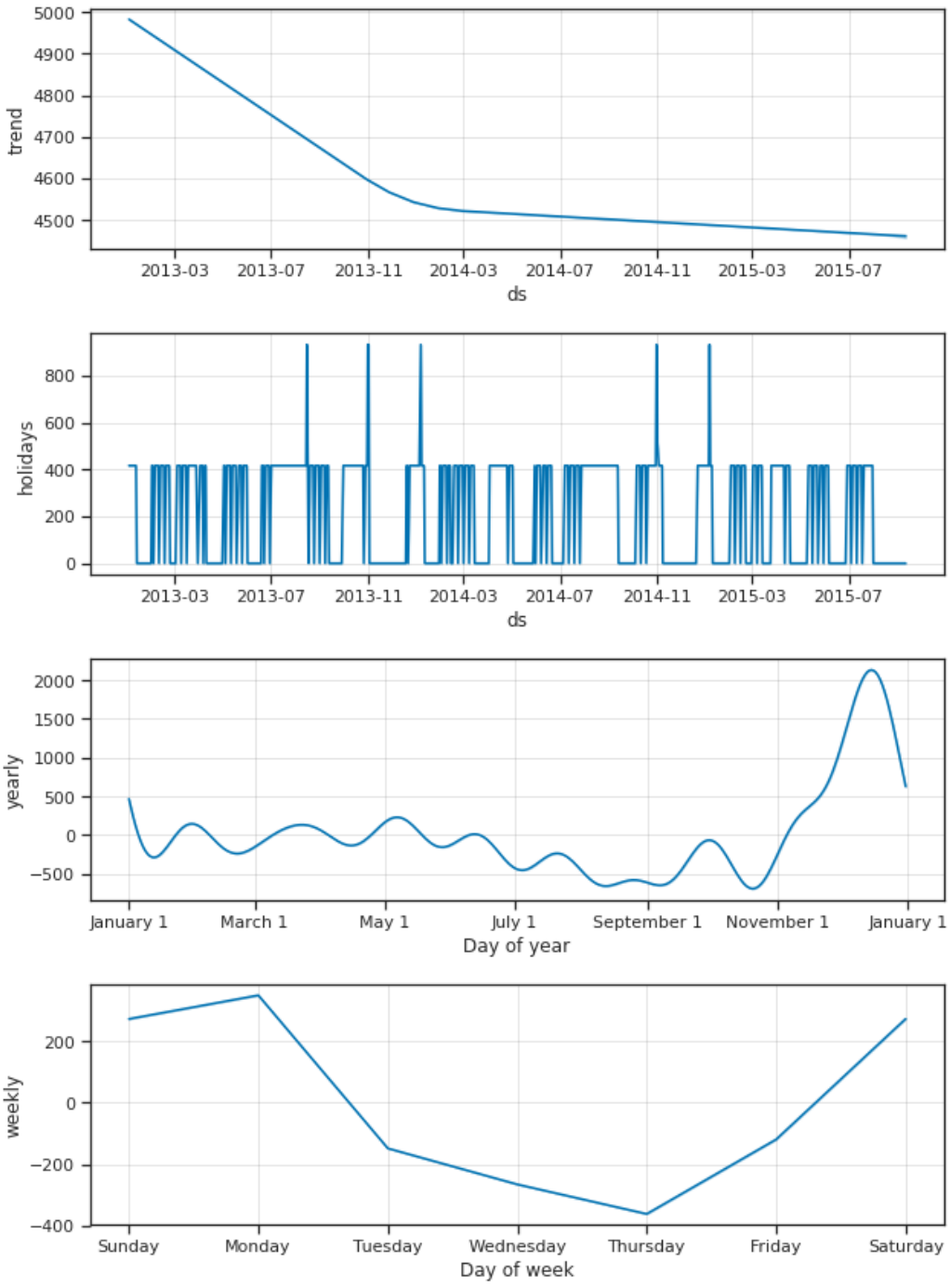


Σχήμα 57: Διάγραμμα πρόβλεψης πωλήσεων της Prophet

Όπως παρατηρούμε η μέθοδος Prophet αποτυπώνει τις τάσεις και τις περισσότερες φορές προβλέπει τις μελλοντικές τιμές σωστά.

Ένα ακόμη ισχυρό χαρακτηριστικό αυτής της μεθόδου είναι η ικανότητά της να μας επιστρέφει τα στοιχεία της πρόβλεψής μας. Αυτό μας αποκαλύπτει πως τα ημερήσια, εβδομαδιαία και ετήσια μοτίβα των χρονοσειρών προστιθέμενων των αργιών, συμβάλλουν συνολικά στις τιμές πρόβλεψης.

```
my_model.plot_components(forecast);
```



Σχήμα 58: Μοτίβα χρονοσειρών

Το πρώτο διάγραμμα στο σχήμα 58 δείχνει ότι οι μηνιαίες πωλήσεις του καταστήματος νούμερο 1 έχουν μειωθεί γραμμικά στον χρόνο. Στο δεύτερο διάγραμμα εμφανίζονται τα διαστήματα των αργιών που έχουν οριστεί στο μοντέλο. Το τρίτο διάγραμμα παρουσιάζει την περίοδο των Χριστουγέννων ως την περίοδο με την περισσότερη κίνηση, ενώ στο τέταρτο διάγραμμα φαίνεται ότι ο εβδομαδιαίος όγκος των πωλήσεων της περασμένης εβδομάδας έχει αντίκτυπο στις πωλήσεις της Δευτέρας της επομένης, εμφανίζοντας τις υψηλότερες πωλήσεις.

6.2.1. Σύνοψη της πρόβλεψης με χρήση χρονοσειρών

Στο κεφάλαιο αυτό προσαρμόσαμε το μοντέλο πρόβλεψης χρησιμοποιώντας τη νέα διαδικασία Prophet από τη Facebook. Θα παραθέσουμε τα κύρια πλεονεκτήματα και μειονεκτήματα της πρόβλεψης με τη χρήση χρονοσειρών.

Πλεονεκτήματα:

- Ισχυρό εργαλείο για την πρόβλεψη χρονοσειρών, καθώς εξηγεί τις εξαρτήσεις του χρόνου, τις εποχικότητες και τις αργίες (στο Prophet εισάγονται χειροκίνητα).
- Υλοποιείται εύκολα με το auto.arima( ) της γλώσσας προγραμματισμού R, το οποίο πακέτο πρόβλεψης εκτελεί μία σύνθετη αναζήτηση πλέγματος και έναν εξελιγμένο αλγόριθμο στο παρασκήνιο.

Μειονεκτήματα:

- Δεν εντοπίζει αλληλεπιδράσεις μεταξύ εξωτερικών χαρακτηριστικών, οι οποίες μπορούν να βελτιώσουν την ακρίβεια του μοντέλου πρόβλεψης. Στην περίπτωση της ROSSMANN οι μεταβλητές αυτές είναι το Promo και το CompetitionOpen.
- Αν και το Prophet προσφέρει μία αυτοματοποιημένη λύση για την ARIMA, αυτό το μοντέλο βρίσκεται ακόμη υπό εξέλιξη και δεν είναι απόλυτα σταθερό.
- Η χρήση εποχιακού μοντέλου ARIMA απαιτεί τέσσερα έως πέντε ολόκληρα έτη στα δεδομένα, το οποίο είναι το μεγαλύτερο μειονέκτημα για νέες επιχειρήσεις.
- Το εποχιακό μοντέλο της ARIMA της Python έχει επτά υπερπαραμετρικά στοιχεία, τα οποία μπορούν να συντονιστούν μόνο χειροκίνητα, επηρεάζοντας σημαντικά την ταχύτητα της διαδικασίας πρόβλεψης (Petrova (2016)).

6.3. Εναλλακτική προσέγγιση: Παλινδρόμηση XGBoost

Το XGBoost είναι μία εφαρμογή των βαθμιδωτά ενισχυμένων δέντρων αποφάσεων (Gradient Boosted Decision Trees) σχεδιασμένο για τη βελτίωση της ταχύτητας και της απόδοσης. Καταλληλότερο όνομα θα ήταν η κανονικοποιημένη βαθμιδωτή ενίσχυση (Regularized Gradient Boosting), καθώς χρησιμοποιεί μία πιο κανονικοποιημένη τυποποίηση του μοντέλου για τον έλεγχο της υπερφόρτωσης.

Τα πλεονεκτήματα αυτού του αλγόριθμου είναι τα εξής:

- Το XGBoost έχει ένα αυτοματοποιημένο σύστημα διαχείρισης των κενών τιμών. Συγκεκριμένα χρησιμοποιεί μία προεπιλεγμένη διαδρομή μάθησης για τις τιμές που λείπουν. Επιλέγει την καλύτερη διαδρομή, η οποία ελαχιστοποιεί την απώλεια δεδομένων.
- Πραγματοποιεί μία ανάλυση αλληλεπιδραστικών χαρακτηριστικών (μέχρι στιγμής εφαρμοσμένη μόνο σε R), δηλαδή σχεδιάζει τα δέντρα αποφάσεων με τις διακλαδώσεις και τα πλαίσια αποφάσεων.
- Εκτελεί μία ανάλυση της βαρύτητας των χαρακτηριστικών, όπου παρουσιάζει σε ένα διαβαθμισμένο διάγραμμα με μπάρες τις πιο σημαντικές μεταβλητές.

Όπως είδαμε και στην προηγούμενη ενότητα, τα δεδομένα μας δεν είναι καθόλου τυχαία, αλλά ακολουθούν μία αυστηρά εποχιακή τάση. Επομένως, πριν τη χρήση του μοντέλου, πρέπει να εξομαλύνουμε τις μεταβλητές που θα χρησιμοποιηθούν για την πρόβλεψη πωλήσεων. Το τυπικό βήμα της πρώτης επεξεργασίας είναι ο μετασχηματισμός των δεδομένων σε ερωτήματα. Μόλις πραγματοποιήσουμε την πρόβλεψη, θα αντιστρέψουμε τη διαδικασία μετασχηματισμών.

Παρακάτω θα παρουσιάσουμε μία σύντομη εκτέλεση της μεθόδου.

```
# to predict to
test = pd.read_csv("~/Documents/projects/rossmann/test.csv",
                   parse_dates = True, low_memory = False, index_col = 'Date')
test.head()
```

	Id	Store	DayOfWeek	Open	Promo	StateHoliday	SchoolHoliday
Date							
2015-09-17	1	1	4	1.0	1	0	0
2015-09-17	2	3	4	1.0	1	0	0
2015-09-17	3	7	4	1.0	1	0	0
2015-09-17	4	8	4	1.0	1	0	0
2015-09-17	5	9	4	1.0	1	0	0

Η Id μεταβλητή αντιπροσωπεύει μία διπλή σημασία (Κατάστημα, Ημερομηνία) στο αρχείο test.csv.

```
# test: missing values?
test.isnull().sum()
```

```
Id                0
Store             0
DayOfWeek         0
Open             11
Promo             0
StateHoliday      0
SchoolHoliday     0
dtype: int64
```

	Id	Store	DayOfWeek	Open	Promo	StateHoliday	SchoolHoliday
Date							
2015-09-17	480	622	4	NaN	1	0	0
2015-09-16	1336	622	3	NaN	1	0	0
2015-09-15	2192	622	2	NaN	1	0	0
2015-09-14	3048	622	1	NaN	1	0	0
2015-09-12	4760	622	6	NaN	0	0	0
2015-09-11	5616	622	5	NaN	0	0	0
2015-09-10	6472	622	4	NaN	0	0	0
2015-09-09	7328	622	3	NaN	0	0	0
2015-09-08	8184	622	2	NaN	0	0	0
2015-09-07	9040	622	1	NaN	0	0	0
2015-09-05	10752	622	6	NaN	0	0	0

Παρόλο που οι τιμές του Open είναι κενές, υποθέτουμε ότι τα καταστήματα σε αυτές τις ημερομηνίες ήταν ανοιχτά, οπότε αντικαθιστούμε τις κενές τιμές με τον αριθμό 1.

```
# replace NA's in Open variable by 1
test.fillna(1, inplace = True)
```

6.3.1. Κωδικοποίηση δεδομένων

Η μέθοδος XGBoost δεν αντιλαμβάνεται τίποτα άλλο, παρά αριθμούς. Οπότε το πρώτο βήμα πριν την δημιουργία του μοντέλου, είναι η μετατροπή ορισμένων μεταβλητών, σε αριθμητικές τιμές, όπως επίσης και η μετατροπή των ημερομηνιών σε απλούς αριθμούς.

```
# data extraction
test['Year'] = test.index.year
test['Month'] = test.index.month
test['Day'] = test.index.day
test['WeekOfYear'] = test.index.weekofyear

# to numerical
mappings = {'0':0, 'a':1, 'b':2, 'c':3, 'd':4}
test.StateHoliday.replace(mappings, inplace = True)

train_store.Assortment.replace(mappings, inplace = True)
train_store.StoreType.replace(mappings, inplace = True)
train_store.StateHoliday.replace(mappings, inplace = True)
train_store.drop('PromoInterval', axis = 1, inplace = True)

store.StoreType.replace(mappings, inplace = True)
store.Assortment.replace(mappings, inplace = True)
store.drop('PromoInterval', axis = 1, inplace = True)
```

Επιστρέφουμε στον ενοποιημένο πίνακα δεδομένων train_store μετά την μετατροπή όλων των τιμών σε αριθμούς (θα εκτυπώσουμε μόνο τις πρώτες 5 γραμμές x 22 στήλες).

```
# take a look on the train and store again
train_store.head()
```

Πίνακας 11: Δείγμα πέντε γραμμών συγχωνευμένου πίνακα των δεδομένων train και store (στήλες 22)

	Store	DayOfWeek	Sales	Customers	Open	Promo	StateHoliday	SchoolHoliday	Year	Month	...	StoreType
0	1	5	5263	555	1	1	0	1	2015	7	...	3
1	1	4	5020	546	1	1	0	1	2015	7	...	3
2	1	3	4782	523	1	1	0	1	2015	7	...	3
3	1	2	5011	560	1	1	0	1	2015	7	...	3
4	1	1	6102	612	1	1	0	1	2015	7	...	3

Assortment	CompetitionDistance	CompetitionOpenSinceMonth	CompetitionOpenSinceYear	Promo2
1	1270.0	9.0	2008.0	0
1	1270.0	9.0	2008.0	0
1	1270.0	9.0	2008.0	0
1	1270.0	9.0	2008.0	0
1	1270.0	9.0	2008.0	0

Promo2SinceWeek	Promo2SinceYear	CompetitionOpen	PromoOpen
0.0	0.0	82.0	24187.75
0.0	0.0	82.0	24187.75
0.0	0.0	82.0	24187.75
0.0	0.0	82.0	24187.75
0.0	0.0	82.0	24187.75

Παρακάτω θα συγχωνεύσουμε επίσης τα δεδομένα του αρχείου test με αυτά του store.

```
print("Joining test set with an additional store information.")
test_store = pd.merge(test, store, how = 'inner', on = 'Store')

test_store['CompetitionOpen'] = 12 * (test_store.Year - test_store.CompetitionOpenSinceYear) + (test_st
ore.Month - test_store.CompetitionOpenSinceMonth)
test_store['PromoOpen'] = 12 * (test_store.Year - test_store.Promo2SinceYear) + (test_store.WeekOfYear
- test_store.Promo2SinceWeek) / 4.0

print("In total: ", test_store.shape)
test_store.head()
```

Ο πίνακας που προκύπτει από αυτή τη συγχώνευση είναι διαστάσεων 41088 σειρών και 21 στηλών. Παρακάτω φαίνονται οι πρώτες 5 σειρές.

Πίνακας 12: Δείγμα πέντε γραμμών συγχωνευμένου πίνακα των δεδομένων test και store (στήλες 21)

	Id	Store	DayOfWeek	Open	Promo	StateHoliday	SchoolHoliday	Year	Month	Day	...	StoreType	Assortment
0	1	1	4	1.0	1	0	0	2015	9	17	...	3	1
1	857	1	3	1.0	1	0	0	2015	9	16	...	3	1
2	1713	1	2	1.0	1	0	0	2015	9	15	...	3	1
3	2569	1	1	1.0	1	0	0	2015	9	14	...	3	1
4	3425	1	7	0.0	0	0	0	2015	9	13	...	3	1

CompetitionDistance	CompetitionOpenSinceMonth	CompetitionOpenSinceYear	Promo2	Promo2SinceWeek
1270.0	9.0	2008.0	0	0.0
1270.0	9.0	2008.0	0	0.0
1270.0	9.0	2008.0	0	0.0
1270.0	9.0	2008.0	0	0.0
1270.0	9.0	2008.0	0	0.0

Promo2SinceYear	CompetitionOpen	PromoOpen
0.0	84.0	24189.50
0.0	84.0	24189.50
0.0	84.0	24189.50
0.0	84.0	24189.50
0.0	84.0	24189.25

6.3.2. Εκπαίδευση μοντέλου

Η προσέγγιση που θα γίνει για την εκπαίδευση του μοντέλου είναι η εξής:

1. Θα διαχωρίσουμε τα δεδομένα train σε δεδομένα εκπαίδευσης και δεδομένα δοκιμής, με τα οποία θα αξιολογήσουμε το μοντέλο.

2. Θα θέσουμε τον ρυθμό εκμάθησης (eta) σε μία σχετικά υψηλή τιμή (π.χ. 0.05 ~ 0.1) και το num_round σε 300 ~ 500.
3. Θα χρησιμοποιήσουμε την αναζήτηση πλέγματος (grid), ώστε να βρούμε τον βέλτιστο συνδυασμό επιπρόσθετων παραμέτρων.
4. Θα μειώσουμε το eta μέχρι να φτάσουμε το ιδανικό.
5. Θα χρησιμοποιήσουμε τα δεδομένα δοκιμής, ως λίστα παρακολούθησης για να επανακατασκευάσουμε το μοντέλο με τις βέλτιστες παραμέτρους.

```
# split into training and evaluation sets
# excluding Sales and Id columns
predictors = [x for x in train_store.columns if x not in ['Customers', 'Sales', 'SalePerCustomer']]
y = np.log(train_store.Sales) # log transformation of Sales
X = train_store

# split the data into train/test set
X_train, X_test, y_train, y_test = train_test_split(X, y,
                                                    test_size = 0.3, # 30% for the evaluation set
                                                    random_state = 42)
```

```
# predictors
X.columns
```

```
Index(['Store', 'DayOfWeek', 'Sales', 'Customers', 'Open', 'Promo',
       'StateHoliday', 'SchoolHoliday', 'Year', 'Month', 'Day', 'WeekOfYear',
       'SalePerCustomer', 'StoreType', 'Assortment', 'CompetitionDistance',
       'CompetitionOpenSinceMonth', 'CompetitionOpenSinceYear', 'Promo2',
       'Promo2SinceWeek', 'Promo2SinceYear', 'CompetitionOpen', 'PromoOpen'],
      dtype='object')
```

```
# evaluation metric: rmspe
# Root Mean Square Percentage Error
# code chunk shared at Kaggle

def rmspe(y, yhat):
    return np.sqrt(np.mean((yhat / y-1) ** 2))

def rmspe_xg(yhat, y):
    y = np.expm1(y.get_label())
    yhat = np.expm1(yhat)
    return "rmspe", rmspe(y, yhat)
```

Πριν τον συντονισμό των παραμέτρων θα επεξηγηθεί ο ρόλος τους:

- *eta*: Ο ρυθμός εκμάθησης που χρησιμοποιείται για την ενημέρωση της βαρύτητας των μεταβλητών. Μία χαμηλή τιμή σε αυτό, σημαίνει βραδύτερη εκπαίδευση αλλά καλύτερη σύγκλιση.
- *num_round*: Ο συνολικός αριθμός των επαναλήψεων.
- *subsample*: Η αναλογία των δεδομένων εκπαίδευσης που χρησιμοποιούνται σε κάθε επανάληψη. Θα πρέπει να διαμορφώνεται σε ένα εύρος από 30% έως 80% του συνόλου των δεδομένων κατάρτισης και να συγκρίνεται με μία τιμή του 100%, χωρίς δειγματοληψία.

- *colsample_bytree*: Η αναλογία των χαρακτηριστικών που χρησιμοποιούνται σε κάθε επανάληψη, η προεπιλεγμένη τιμή είναι το 1.
- *max_depth*: Το μέγιστο βάθος κάθε δέντρου. Εάν δεν περιορίσουμε το μέγιστο βάθος, το gradient boosting θα υπερφορτωθεί.
- *early_stopping_rounds*: Εάν δεν υπάρχει βελτίωση της βαθμολογία επικύρωσης για τον δεδομένο αριθμό επαναλήψεων, ο αλγόριθμος θα σταματήσει νωρίτερα για να καταπολεμήσει την υπερφόρτωση.

```
# base parameters
params = {
    'booster': 'gbtree',
    'objective': 'reg:linear', # regression task
    'subsample': 0.8, # 80% of data to grow trees and prevent overfitting
    'colsample_bytree': 0.85, # 85% of features used
    'eta': 0.1,
    'max_depth': 10,
    'seed': 42} # for reproducible results
```

```
# XGB with xgboost library
dtrain = xgb.DMatrix(X_train[predictors], y_train)
dtest = xgb.DMatrix(X_test[predictors], y_test)

watchlist = [(dtrain, 'train'), (dtest, 'test')]

xgb_model = xgb.train(params, dtrain, 300, evals = watchlist,
                      early_stopping_rounds = 50, feval = rmspe_xg, verbose_eval = True)
```

Last five rows:

```
[295]    train-rmspe:0.106959    test-rmspe:0.111575
[296]    train-rmspe:0.106855    test-rmspe:0.111498
[297]    train-rmspe:0.106467    test-rmspe:0.111439
[298]    train-rmspe:0.106348    test-rmspe:0.111331
[299]    train-rmspe:0.105759    test-rmspe:0.111298
```

Ουσιαστικά, θέλουμε τη μικρότερη αξία. Το μοντέλο με υπερπαραμετρικές βάσεις δίνει καλύτερα αποτελέσματα στα train set, υποδεικνύοντας το ζήτημα της υπερφόρτωσης.

6.3.3. Αναζήτηση πλέγματος από το sklearn

Η μέθοδος Scikit είναι διάσημη για την GridSearchCV και την RandomizedSearchCV. Μεταξύ αυτών των δύο, τις περισσότερες φορές προτιμάται η RandomizedSearchCV γιατί είναι ταχύτερη έκδοση από την GridSearchCV.

Ως είσοδο, η RandomnizedSearchCV παίρνει μόνο το sklearn wrap του XGboost, οπότε αντί να χρησιμοποιήσουμε την πρώτη έκδοση ενός μοντέλου, χτίζουμε το ανάλογο μοντέλο στο sklearn με το XGBRegressor.

```
# XGB with sklearn wrapper
# the same parameters as for xgboost model
params_sk = {'max_depth': 10,
             'n_estimators': 300, # the same as num_rounds in xgboost
             'objective': 'reg:linear',
             'subsample': 0.8,
             'colsample_bytree': 0.85,
             'learning_rate': 0.1,
             'seed': 42}

skrg = XGBRegressor(**params_sk)

skrg.fit(X_train, y_train)
```

```
XGBRegressor(base_score=0.5, colsample_bylevel=1, colsample_bytree=0.85,
       gamma=0, learning_rate=0.1, max_delta_step=0, max_depth=10,
       min_child_weight=1, missing=None, n_estimators=300, nthread=-1,
       objective='reg:linear', reg_alpha=0, reg_lambda=1,
       scale_pos_weight=1, seed=42, silent=True, subsample=0.8)
```

Θα καθορίσουμε την παράμετρο ρύθμισης reg_alpha που θα μειώσει την πολυπλοκότητα του μοντέλου και θα βελτιώσει την απόδοση, καθώς και την παράμετρο gamma που αντιπροσωπεύει την ελάχιστη μείωση απώλειας που απαιτείται για να γίνει split. Επίσης θα ορίσουμε το max_depth που χρησιμοποιείται για τον έλεγχο της υπερφόρτωσης.

```
import scipy.stats as st

params_grid = {
    'learning_rate': st.uniform(0.01, 0.3),
    'max_depth': list(range(10, 20, 2)),
    'gamma': st.uniform(0, 10),
    'reg_alpha': st.expon(0, 50)}

search_sk = RandomizedSearchCV(skrg, params_grid, cv = 5) # 5 fold cross validation
search_sk.fit(X_train, y_train)

# best parameters
print(search_sk.best_params_); print(search_sk.best_score_)
```

```
{'gamma': 0.80198330585415034, 'learning_rate': 0.044338624448041611, 'max_depth': 16, 'reg_alpha': 23.0
08226565535971}
0.999596090945
```

```
# with new parameters
params_new = {
    'booster': 'gbtree',
    'objective': 'reg:linear',
    'subsample': 0.8,
    'colsample_bytree': 0.85,
    'eta': 0.044338624448041611,
    'max_depth': 16,
    'gamma': 0.80198330585415034,
    'reg_alpha': 23.008226565535971,
    'seed': 42}

model_final = xgb.train(params_new, dtrain, 300, evals = watchlist,
                        early_stopping_rounds = 50, feval = rmspe_xg, verbose_eval = True)
```

Οι τελευταίες πέντε σειρές είναι οι εξής:

```
[295]    train-rmspe:0.213295    test-rmspe:0.147425
[296]    train-rmspe:0.213214    test-rmspe:0.147367
[297]    train-rmspe:0.213061    test-rmspe:0.147199
[298]    train-rmspe:0.213084    test-rmspe:0.14701
[299]    train-rmspe:0.212913    test-rmspe:0.146808
```

Επιλύσαμε ένα πρόβλημα με υπερφόρτωση, αλλά λόγω της μείωσης του ρυθμού εκμάθησης (eta), πήραμε λίγο χειρότερη συνολική βαθμολογία στο test set (~0.11 έως ~0.14).

```
yhat = model_final.predict(xgb.DMatrix(X_test[predictors]))
error = rmspe(X_test.Sales.values, np.exp(yhat))

print('First validation yelds RMSPE: {:.6f}'.format(error))
```

```
First validation yelds RMSPE: 0.146761
```

Παρόλο που είχαμε μια λίγο υψηλότερη τιμή RMSPE, πρέπει να θυμόμαστε ότι ο αντίστοιχος eta για την πρώτη έκδοση ήταν 0.1, ο οποίος θεωρείται γενικά υψηλός. Είναι αξιοσημείωτο ότι πήραμε σχεδόν το ίδιο αποτέλεσμα, αλλά με δύο φορές χαμηλότερο eta (0.1 έως 0.04).

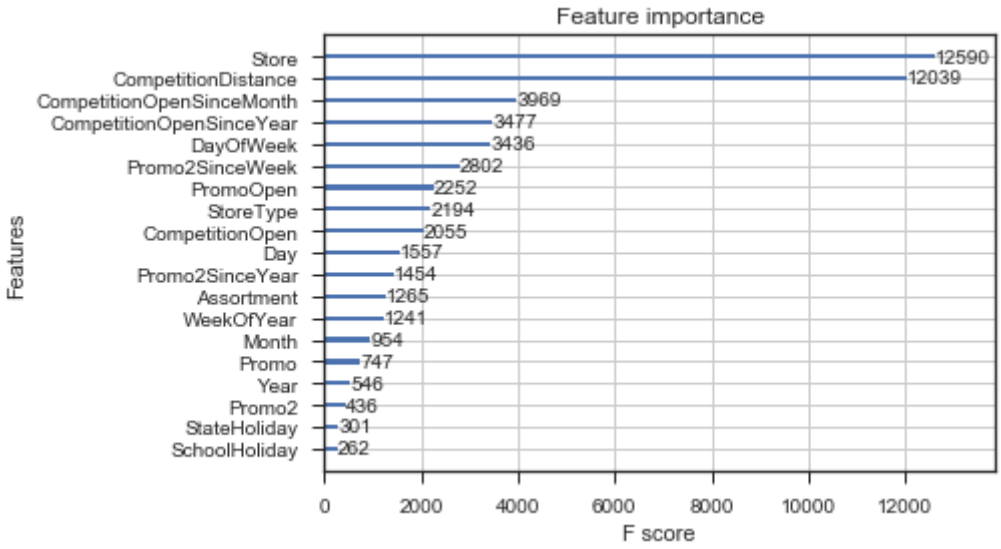
Τα επόμενα βήματα για τη βελτίωση του μοντέλου θα συμπεριλαμβάνουν περαιτέρω μείωση του eta και συντονισμό των αντίστοιχων gamma και max_depth.

6.3.4. Κατανόηση μοντέλου

Οι βαρύτητες σημαντικότητας των χαρακτηριστικών, μας βοηθούν να κατανοήσουμε ποιες από τις μεταβλητές συνέβαλαν περισσότερο στο τελικό αποτέλεσμα.

```
xgb.plot_importance(model_final)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x117a35748>
```



Σχήμα 59: Διάγραμμα βαρύτητας των μεταβλητών της ROSSMANN

Οι πιο σημαντικές μεταβλητές που κυριαρχούν είναι το Store και το CompetitionDistance. Μετά ακολουθούν οι μεταβλητές CompetitionOpenSinceMonth, CompetitionOpenSinceYear, DayOfWeek, Promo2SinceWeek και την πρώτη επτάδα κλείνει η PromoOpen.

6.3.5. Πρόβλεψη στα κρυφά δεδομένα

Παρακάτω παρουσιάζεται η εφαρμογή της τελικής πρόβλεψης πωλήσεων στα κρυφά δεδομένα του test.csv αρχείου. Θα εμφανίσουμε την πρόβλεψη πωλήσεων για τα πρώτα πέντε καταστήματα στις 17/9/2015 (Petrova (2017)).

```
# predictions to unseen data
unseen = xgb.DMatrix(test_store[predictors])
test_p = model_final.predict(unseen)

forecasts = pd.DataFrame({'Id': test['Id'],
                          'Sales': np.exp(test_p)})
# forecasts
forecasts.head()
```

	Id	Sales
Date		
2015-09-17	1	4428.240723
2015-09-17	2	4577.308105
2015-09-17	3	4895.477051
2015-09-17	4	5571.973145
2015-09-17	5	4961.958496

6.4. Συμπεράσματα για τις μεθόδους πρόβλεψης

Αρχικά χρησιμοποιήθηκε ένα νέο μοντέλο πρόβλεψης, το Prophet, το οποίο είναι ακόμα υπό ανάπτυξη, αλλά έχει τα πάντα για την προηγμένη μοντελοποίηση, καθώς μπορεί να υπολογίζει τα σημεία αλλαγής στις τάσεις και να συμπεριλαμβάνει τις αργίες στα δεδομένα. Παρόλα αυτά το πιο εξελιγμένο μέχρι στιγμής εργαλείο για την ανάλυση των χρονοσειρών παραμένει το Auto-ARIMA από το πακέτο πρόβλεψης της γλώσσας R.

Ωστόσο, ένα σημαντικό άλμα στην απόδοση της πρόβλεψης μπορεί να πραγματοποιηθεί με το μοντέλο XGBoost και τη χρήση της XGBoost library, εάν επιτευχθεί μία αύξηση του αριθμού και του εύρους των υπερπαραμέτρων που θα χρησιμοποιηθούν. Λόγω του όγκου των παρατηρήσεων (περίπου 800000) και με τα μέσα που χρησιμοποιήθηκαν, η πιο λεπτομερής αναζήτηση πλέγματος (grid) θα χρειαζόταν περίπου δύο με τρεις ημέρες για να μοντελοποιηθεί, οπότε αφέθηκε ως μία πρόταση για μελλοντική έρευνα.

Μια άλλη μέθοδος που δεν αναλύθηκε εδώ, είναι το μοντέλο παλινδρόμησης Stacking, το οποίο λειτουργεί καλά για σύνολα δεδομένων μικρού ή μεσαίου μεγέθους. Κατά αυτό το μοντέλο, θα έπρεπε να συνδυάσουμε αρχικά τα XGBoost, RandomForest, NN και SVM για παλινδρόμηση και στη συνέχεια να τα συνθέτουμε όλα μαζί για την κατασκευή του τελικού μοντέλου.

Όσον αφορά τον διαγωνισμό της Kaggle, έγιναν δύο υποβολές από την Elena Petrova στο leaderboard, μία πρόβλεψη από τη βάση (forecast from the base) και ένα tuned model.

```
# first
# 0.66419
test_base = xgb_model.predict(unseen)

forecasts_base = pd.DataFrame({'Id': test['Id'],
                               'Sales': np.exp(test_base)})
forecasts_base.to_csv("xgboost_2_submission.csv", index = False)
```

```
# final
# 0.60553
forecasts.to_csv("xgboost_submission.csv", index = False)
```

Submission and Description	Private Score	Public Score
xgboost_2_submission.csv a few seconds ago by Elena Petrova add submission details	0.68262	0.66419
xgboost_submission.csv 17 minutes ago by Elena Petrova	0.62590	0.60553

Το αποτέλεσμα με τέσσερις συντονισμένες παραμέτρους βελτιώθηκε από 0.66 σε 0.60 στο Public Board και από 0.62 σε 0.68 στο Private Board (το test set το οποίο έχει δεσμευτεί από το ίδιο το Kaggle). Αυτά τα αποτελέσματα παραμένουν σχετικά χαμηλά, αλλά είναι μία καλή αρχή για περαιτέρω βελτίωση (Petrova (2017)).

7. Συμπεράσματα

Στην παρούσα εργασία κατανοήσαμε την όλη λειτουργία του Data Science και ποια είναι η χρήση του σε επιχειρηματικά προβλήματα. Επίσης κατανοήσαμε, ποιος είναι ο ρόλος της διερεύνησης των δεδομένων, της μηχανικής μάθησης και των μεθόδων πρόβλεψης, χρησιμοποιώντας μεγάλες βάσεις δεδομένων που προέρχονται από επιχειρήσεις, ώστε να βελτιώσουμε τη διαδικασία λήψης αποφάσεων αυτών των επιχειρήσεων.

Επικεντρώνοντας την προσοχή μας σε μία μόνο μελέτη περίπτωσης, αυτή της ROSSMANN, και αναλύοντάς την διεξοδικά, εστιάσαμε πρακτικά στη χρήση των τεχνολογιών του Data Science και όχι θεωρητικά, οπότε αποκτήσαμε μία εμπεριστατωμένη άποψη για αυτές τις τεχνολογίες. Αντιληφθήκαμε επίσης ότι υπάρχουν διάφορες μέθοδοι για τις προβλέψεις, ανάλογα με το πρόβλημα που θέλουμε να αντιμετωπίσουμε, οι οποίες πλέον είναι πολύ εξελιγμένες. Ενώ συνειδητοποιήσαμε ότι για την εκτέλεσή τους δεν απαιτούνται ισχυρά συστήματα υπολογιστών, διότι μπορούν να εκτελεστούν σε cloud πλατφόρμες.

Επίσης μετά την ανάλυση αυτής της μελέτης περίπτωσης, διαπιστώσαμε ότι αυτοί οι αλγόριθμοι έχουν την δυνατότητα να εξάγουν αξιόλογα αποτελέσματα, πολύ κοντά στα πραγματικά και με τις κατάλληλες τροποποιήσεις μπορούν να χρησιμοποιηθούν και από άλλες μεγάλες επιχειρήσεις, που έχουν συλλέξει τον κατάλληλο όγκο δεδομένων και επιθυμούν παρόμοιες προβλέψεις.

Τέλος, μέσω της εκπόνησης της εργασίας, εξοικειωθήκαμε με τις βασικές έννοιες των μεθόδων πρόβλεψης, όπως είναι τα train και test δεδομένα, η εποχικότητα, οι τάσεις των δεδομένων, η ανάλυση χρονοσειρών, η μέθοδος XGBoost, το RMSPE, το Kaggle και οι cloud πλατφόρμες.

Έχοντας εκτελέσει σε αυτή την εργασία την πρόβλεψη με μία μέθοδο από την κατηγορία των χρονοσειρών και με μία από την κατηγορία της παλινδρόμησης, μία πρόταση για μελλοντική έρευνα θα ήταν η εκτέλεση του ίδιου προβλήματος, με τη χρήση μεθόδου πρόβλεψης με νευρωνικά δίκτυα.

Βιβλιογραφία

Agrawal, D. and Schorling, C. (1997), "Market share forecasting: an empirical comparison of artificial neural networks and multinomial logit model", Journal of Retailing 72 (4), 383–407.

Alon, I., Qi, M. and Sadowski, R.J. (2001), "Forecasting aggregate retail sales: a comparison of artificial neural networks and traditional methods", Journal of Retailing and Consumer Services 8, 147–156.

Ansuj, A.P., Camargo, M.E., Radharamanan, R. and Petry, D.G. (1996), "Sales forecasting using time series and neural networks", Computers & Industrial Engineering – Journal 31, 421–424.

Armstrong, J.S. (2001), Principles of Forecasting: A Handbook for Researchers and Practitioners, Kluwer, Boston.

Aviv, Y. (2001), "The effect of collaborative forecasting on supply chain performance", Management Science 47 (10), 1326–1343.

Barksdale, H.C. and Hilliard, J.E. (1975), "A cross-spectral analysis of retail inventories and sales", The Journal of Business 48 (3), 365–382.

Bayus, B.L. and Putsis, W.P. (1999), "Product proliferation: an empirical analysis of product line determinants and market outcomes", Marketing Science 18 (2), 137–155.

Beheshti-Kashi, S., Karimi, H.R., Thoben, K.D., Lutjen, M. and Teucke, M. (2015), "A survey on retail sales forecasting and prediction in fashion markets", Systems Science & Control Engineering 3 (1), 154-161.

Bengio, Y., Courville, A. and Vincent, P. (2013), "Representation learning: A review and new perspectives", IEEE transactions on pattern analysis and machine intelligence 35 (8), 1798-1828.

Blunsom, P., Grefenstette, E., Kalchbrenner, N., et al. (2014). A convolutional neural network for modelling sentences. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, Maryland, USA.

Burges, C.J.C. (1998), "A tutorial on support vector machines for pattern recognition", Data Mining and Knowledge Discovery 2, 121–167.

Chaman, L.J. (1999), "Ten commandments of selling forecasts to forecast users", Journal of Business Forecasting Methods & Systems 17 (4), 2–5.

Chaman, L.J. (2001), "Forecasting practices in corporate America", Journal of Business Forecasting Methods & Systems 20 (2), 2–4.

Chase, C. (1999), “Sales forecasting at the dawn of the new millennium?”, Journal of Business Forecasting Methods & Systems 18 (3), 2–5.

Chen, F.L. and Ou, T.Y. (2011), “Sales forecasting system based on Gray extreme learning machine with Taguchi method in retail industry”, Expert Systems with Applications 38 (3), 1336–1345.

Ching, T., Himmelstein, D.S., Beaulieu-Jones, B.K., Kalinin, A.A., Do, B.T., Way, G.P., Ferrero, E., Agapow, P.M., Zietz, M., Hoffman, M.M., et al. (2018), “Opportunities and obstacles for deep learning in biology and medicine”, bioRxiv, 142760.

Cristianini, N. and Shawe-Taylor, J. (2000), An introduction to support vector machines and other kernel-based learning methods, Cambridge University Press.

Curtis, A.B., Lundholm, R.J. and McVay, S.E. (2014), “Forecasting sales: a model and some evidence from the retail industry”, Contemporary Accounting Research 31 (2), 581–608.

Dalrymple, D.J. (1987), “Sales forecasting practices: results of a United States survey”, International Journal of Forecasting 3, 379–392.

Danese, P. and Kalchschmidt M. (2008). The impact of forecasting on performances: is accuracy the only matter? In: Proceedings of the XV Working Seminar on Production Economics, Innsbruck, Austria, 153–166.

Danese, P. and Kalchschmidt, M. (2010), “The impact of forecasting on companies’ performance: Analysis in a multivariate setting”, International Journal of Production Economics 133, 458-469.

Di Pillo, G., Latorre, V., Lucidi, S. and Procacci, E. (2016), “An application of support vector machines to sales forecasting under promotions”, 4OR - A Quarterly Journal of Operations Research 14 (3), 309-325.

Domingos, P. (2012), “A few useful things to know about machine learning”, Communications of the ACM 55 (10), 78-87.

Golicic, S.L., Davis, D.F., McCarthy, T.M. and Mentzer J.T. (2002), “The impact of e-commerce on supply chain relationships”, International Journal of Physical Distribution and Logistics Management 23 (10), 851–871.

Gove, P.B. (1986), Webster’s Third New International Dictionary of the English Language Unabridged, Merriam-Webster, Inc., Springfield, M.A.

Graves, A., Mohamed, A.R. and Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In: 2013 IEEE international conference on acoustics, speech and signal processing, 6645-6649, Vancouver, BC, Canada.

Gulcehre, C., Bougares, F., Schwenk, H., Cho, K., Merrienboer, B. and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 1724–1734, Doha, Qatar.

Helms, M.M., Ettkin, L.P. and Chapman, S. (2000), "Supply chain forecasting: collabora- tive forecasting supports supply chain management", Business Process Management Journal 6 (5), 392–407.

IBM, "Gaining a competitive advantage by having key business figures available at all times". Available at: https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=POC03218WWEN (2015).

Jaccard, J. and Turrisi, R. (2003), Interaction Effects in Multiple Regression, Second Edition Sage Publications, Thousand Oaks, CA.

Kaggle Favorita Grocery, "Corporación Favorita Grocery Sales Forecasting", Available at: https://www.kaggle.com/c/favorita-grocery-sales-forecasting (2017).

Kaggle Rossmann, "Rossmann Store Sales". Available at: https://www.kaggle.com/c/rossmann-store-sales (2015).

Kechyn, G., Yu, L., Zang, Y. and Kechyn, S. (2018), "Sales forecasting using WaveNet within the framework of the Kaggle competition", arXiv:1803.04037.

Keller, G. and Gaciu, N. (2012), Managerial statistics, South-Western Cengage Learning.

Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012). Imagenet classi_cation with deep convolutional neural networks. In: Advances in neural information processing systems, 1097-1105, Lake Tahoe, Nevada, USA.

Lapide, L. (1999), "How about collaborative forecasting?", Journal of Business Forecasting Methods & Systems 18 (3), 24–25.

Lawrence, M., O'Connor, M. and Edmundson, B. (2000), "A field study of sales forecasting accuracy and processes", European Journal of Operational Research 122, 151–160.

Linkedin, "Rossmann Online GmbH". Available at: https://www.linkedin.com/company/rossmann-online-gmbh/about/

Liu, S., "Forecasting with Prophet". Available at: https://towardsdatascience.com/forecasting-with-prophet-d50bbfe95f91 (2018).

Lutins, E., " Boosting in Machine Learning and the Implementation of XGBoost in Python". Available at: https://towardsdatascience.com/boosting-in-machine-learning-and-the-implementation-of-xgboost-in-python-fb5365e9f2a0 (2017).

Makridakis, S., Wheelwright, S.C. and Hyndman, R.J. (1998), Forecasting: methods and applications, John Wiley & Sons Inc.

Markgraf, B., "The risks of sales forecasting". Available at: https://smallbusiness.chron.com/risks-sales-forecasting-47339.html (2017).

McCarthy, T.M., Davis, D.F., Golicic, S.L. and Mentzer, J.T. (2006), "The Evolution of Sales Forecasting Management: A 20-Year Longitudinal Study of Forecasting Practices", Journal of Forecasting 25, 303-324.

McCarthy, T.M. and Golicic, S.L. (2002), "Implementing collaborative forecasting to improve supply chain performance", International Journal of Physical Distribution & Logistics Management 32 (6), 431–454.

Mentzer, J.T. and Bienstock, C.C. (1998), Sales Forecasting Management: understanding the techniques, systems and management of the sales forecasting process. SAGE Publications, Incorporated.

Mentzer, J.T., Bienstock, C. and Kahn, K. (1999), "Benchmarking sales forecasting management", Business Horizons 42 (3), 48–56.

Mentzer, J.T. and Cox, J.E. Jr. (1984), "Familiarity, application, and performance of sales forecasting techniques", Journal of Forecasting 3 (1), 27–36.

Mentzer, J.T. and Kahn, K.B. (1995), "Forecasting technique familiarity, satisfaction, usage, and application", Journal of Forecasting 14, 465–476.

Mentzer, J.T. and Kahn, K.B. (1997), "Sales forecasting systems in corporate America", Journal of Business Forecasting Methods & Systems 16 (1), 6–12.

Mentzer, J.T. and Moon, M.A. (2005), Sales Forecasting Management. A Demand Management Approach, Sage Puublications, Incorporated.

Moon, M.A. and Mentzer, J.T. (1998), "Seven keys to better forecasting", Business Horizons 41 (5), 44–52.

Moon, M.A., Mentzer, J.T. and Smith, C.D. (2003), "Conducting a sales forecasting audit", International Journal of Forecasting 19, 5–25.

Oosterlee, C.W., Borovykh, A. and Bohte, S. (2017), "Conditional time series forecasting with convolutional neural networks", arXiv:1703.04691.

Parker, K. (2002), "Events happen, but demand is always", Manufacturing Business Technology 20 (2), 40–43.

Petrova, E., "Github: Rossmann Store Sales". Available at: https://github.com/elena-petrova/rossmann_TSA_forecasts/blob/master/Rossmann_Sales.ipynb (2017).

Petrova, E., “Time Series Analysis and Forecasting with Prophet”. Available at: https://www.kaggle.com/elenapetrova/time-series-analysis-and-forecasts-with-prophet (2016).

Raghu Raj Rai, “XGBoost: The Excalibur for Everyone”. Available at: https://towardsdatascience.com/xgboost-the-excalibur-for-everyone-8009bd015f1e (2018).

Rossmann GmbH. Available at: https://www.rossmann.de/unternehmen/ueber-uns.html

Sanders, N.R. (1997), “The status of forecasting in manufacturing firms”, Production and Inventory Management Journal, Second Quarter, 32–36.

Sanders, N.R. and Manrodt, K.B. (1994), “Forecasting practices in US corporations: survey results”, Interfaces 24 (2), 92–100.

Sevuktekin, M., Yilmaz, T. and Kara, M. (2018), “Sales Forecast of New Product With Bass Diffusion Model”, International Journal of Economic and Administrative Studies, 399-414.

Shvets, A., Rakhlin, A., Kalinin, A.A. and Iglovikov, V. (2018), “Automatic instrument segmentation in robot-assisted surgery using deep learning”, arXiv preprint arXiv:1803.01207.

Smaros, J. (2007), “Forecasting collaboration in the European grocery sector: observations from a case study”, Journal of Operations Management 25, 702–716.

Taylor, S.J. and Letham, B., “Prophet: forecasting at scale”. Available at: https://research.fb.com/prophet-forecasting-at-scale/ (2017).

Thiele Christian. Available at: https://www.kaggle.com/thie1e/exploratory-analysis-rossmann (2015).

Vasquez, “Web Traffic Forecasting”. Available at: https://github.com/sjvasquez/web-traffic-forecasting (2017).

Wikipedia Foundation Inc, “Kaggle”. Available at: https://en.wikipedia.org/wiki/Kaggle

Wright, D.J., Capon, G., Page, R., Quiroga, J., Taseen, A.A. and Tomasini, F. (1986), “Evaluation of forecasting methods for decision support”, International Journal of forecasting 2 (2), 139–153.

Xu, Q. and Sharma, V. (2017). Ensemble Sales Forecasting Study in Semiconductor Industry. In: Advances in Data Mining. Applications and Theoretical Aspects, part of the 17th Industrial Conference, ICDM 2017, New York, USA.

Zhao, K. and Wang, C. (2017), “Sales Forecast in E-commerce using Convolutional Neural Network”, arXiv:1708.07946.

Παράρτημα Α: Διερεύνηση των δεδομένων (κώδικας σε γλώσσα R)

https://www.kaggle.com/kokovidis/rossmann-r-kernel

```
---
title: "Rossmann Exploratory Analysis"
author: "Christian Thiele"
date: "8. Oktober 2015"
output: html_document
---

This is an exploratory analysis of the Rossmann Store Sales data which
can be
found [here](https://www.kaggle.com/c/rossmann-store-sales). The data
isn't huge
but the speedup using data.table is noticeable. It is nice to have
unmasked data which allows for some interpretations.

Read in the data:

```{r}
library(data.table)
library(zoo)
library(forecast)
library(ggplot2)
test <- fread("../input/test.csv")
train <- fread("../input/train.csv")
store <- fread("../input/store.csv")
```

Let's have a first look at the data:

```{r}
str(train)
str(test)
str(store)
# head(train); tail(train)
# head(test); tail(test)
train[, Date := as.Date(Date)]
test[, Date := as.Date(Date)]
store
train <- train[order(Date)]
test <- test[order(Date)]
summary(train)
summary(test)
test[is.na(test$Open), ] # Only store 622
test$Open[test$Store == 622]
```

The test set has just 41088 rows while the train set has 1017209 rows.
The public leaderboard is based on 39% of the data (16024 rows) and
the private leaderboard
is based on 61% of the data (25064 rows). Store 622 has 11 missing
values in the
`Open` columns, but not all of the data in that column of that store
is missing.
```

As was pointed out [here](https://www.kaggle.com/c/rossmann-store-sales/forums/t/16835/open-is-blank-in-test-file-for-store-622) it should probably be imputed as 1.

Additionally, the whole `Customers` column is missing from the test data (since
that data is only known ex post).

```{r}
test[is.na(test)] <- 1
```

During the test period there are no Easter or Christmas holidays but interestingly
during a rather large portion of the time (44%) there are school holidays while that
is the case for only 18% of the train data:

```{r}
# Unique values per column
train[, lapply(.SD, function(x) length(unique(x)))]
test[, lapply(.SD, function(x) length(unique(x)))]
# All test stores are also in the train data
sum(unique(test$Store) %in% unique(train$Store))
# 259 train stores are not in the test data
sum(!(unique(train$Store) %in% unique(test$Store)))
table(train$Open) / nrow(train) # Percent Open Train
table(test$Open) / nrow(test) # Percent Open Test
table(train$Promo) / nrow(train) # Percent of the time promo in train
table(test$Promo) / nrow(test) # Percent of the time promo in test
table(train$StateHoliday) / nrow(train) # Percent of the time holiday in train
table(test$StateHoliday) / nrow(test) # no b and c = no easter holiday and no christmas
table(train$SchoolHoliday) / nrow(train) # Percent of the time school holiday in train
table(test$SchoolHoliday) / nrow(test) # Percent of the time school holiday in test
```

There are no obvious breaks in the data.
The test period ranges from 2015-08-01 to 2015-09-17, so the task is to predict 48 days.
The train period ranges from 2013-01-01 to 2015-07-31.

```{r}
plot(train$Date, type = "l")
plot(test$Date, type = "l")
# As expected all 856 stores to be predicted daily
all(table(test$Date) == 856)
```

Let's look at the columns that are unique to the train set:

```{r}
hist(train$Sales, 100)
hist(aggregate(train[Sales != 0]$Sales,
               by = list(train[Sales != 0]$Store), mean)$x, 100,
     main = "Mean sales per store when store was not closed")

hist(train$Customers, 100)
hist(aggregate(train[Sales != 0]$Customers,
               by = list(train[Sales != 0]$Store), mean)$x, 100,
     main = "Mean customers per store when store was not closed")
```

```
ggplot(train[Sales != 0], aes(x = factor(SchoolHoliday), y = Sales)) +
    geom_jitter(alpha = 0.1) +
    geom_boxplot(color = "yellow", outlier.colour = NA, fill = NA)
ggplot(train[train$Sales != 0 & train$Customers != 0],
       aes(x = log(Customers), y = log(Sales))) +
    geom_point(alpha = 0.2) + geom_smooth()
ggplot(train[train$Sales != 0 & train$Customers != 0],
       aes(x = factor(Promo), y = Sales)) +
    geom_jitter(alpha = 0.1) +
    geom_boxplot(color = "yellow", outlier.colour = NA, fill = NA)
ggplot(train[train$Sales != 0 & train$Customers != 0],
       aes(x = factor(Promo), y = Customers)) +
    geom_jitter(alpha = 0.1) +
    geom_boxplot(color = "yellow", outlier.colour = NA, fill = NA)
```

Note: I chose to not plot that data including days with 0 sales
or customers because that would have biased the boxplots.

Sales is as expected strongly correlated with the number of customers.
It looks like the Boxplots of customers overlap a little more than the
boxplots
of sales. This would mean that the promos are not mainly attracting
more customers
but make customers spend more. The mean amount spent per customer is
about one
Euro higher:

```{r}
with(train[train$Sales != 0 & train$Promo == 0], mean(Sales /
Customers))
with(train[train$Sales != 0 & train$Promo == 1], mean(Sales /
Customers))
```

There are sometimes promos while the respective store is closed and
there are promos
45% of the time:

```{r}
table(ifelse(train$Sales != 0, "Sales > 0", "Sales = 0"),
      ifelse(train$Promo, "Promo", "No promo"))
```

At least there are no sales when the stores are closed but there are
some stores
that, according to the data, made no sales although they were opened
even if
they had some customers. These observations *may* be errors in the
data / outliers:

```{r}
table(ifelse(train$Open == 1, "Opened", "Closed"),
      ifelse(train$Sales > 0, "Sales > 0", "Sales = 0"))
# That tends to happen on consecutive days. Some stores even had
customers
# (who bought nothing?)
train[Open == 1 & Sales == 0]
```

The stores have different amounts of days with zero sales.
There are spikes in the sales before the stores close and after the
reopen:

```{r}
zerosPerStore <- sort(tapply(train$Sales, list(train$Store),
function(x) sum(x == 0)))
hist(zerosPerStore,100)
# Stores with the most zeros in their sales:
tail(zerosPerStore, 10)
# Some stores were closed for some time, some of those were closed
multiple times
plot(train[Store == 972, Sales], ylab = "Sales", xlab = "Days", main =
"Store 972")
plot(train[Store == 103, Sales], ylab = "Sales", xlab = "Days", main =
"Store 103")
plot(train[Store == 708, Sales], ylab = "Sales", xlab = "Days", main =
"Store 708")
```

There are also stores that have *no* zeros in their sales. These are
the exception
since they are opened also on sundays / holidays. The sales of those
stores
on sundays are particularly high:

```{r}
ggplot(train[Store == 85],
       aes(x = Date, y = Sales,
           color = factor(DayOfWeek == 7), shape = factor(DayOfWeek ==
7))) +
    geom_point(size = 3) + ggtitle("Sales of store 85 (True if
sunday)")
ggplot(train[Store == 262],
       aes(x = Date, y = Sales,
           color = factor(DayOfWeek == 7), shape = factor(DayOfWeek ==
7))) +
    geom_point(size = 3) + ggtitle("Sales of store 262 (True if
sunday)")
```

That is not true in general. The variability of sales on sundays is
quite high
while the median is not:

```{r}
ggplot(train[Sales != 0],
       aes(x = factor(DayOfWeek), y = Sales)) +
    geom_jitter(alpha = 0.1) +
    geom_boxplot(color = "yellow", outlier.colour = NA, fill = NA)

```

**The `store` file contains information about the stores that can be
linked to
`train` and `test` via the store ID.**

```{r}
summary(store)
table(store$StoreType)
table(store$Assortment)
# There is a connection between store type and type of assortment
table(data.frame(Assortment = store$Assortment, StoreType =
store$StoreType))
hist(store$CompetitionDistance, 100)
# Convert the CompetitionOpenSince... variables to one Date variable
store$CompetitionOpenSince <-
as.yearmon(paste(store$CompetitionOpenSinceYear,
```

```
store$CompetitionOpenSinceMonth, sep = "-"))
# One competitor opened 1900
hist(as.yearmon("2015-10") - store$CompetitionOpenSince, 100,
     main = "Years since opening of nearest competition")
# Convert the Promo2Since... variables to one Date variable
# Assume that the promo starts on the first day of the week
store$Promo2Since <- as.POSIXct(paste(store$Promo2SinceYear,
                                      store$Promo2SinceWeek, 1, sep = "-
"),
                                format = "%Y-%U-%u")
hist(as.numeric(as.POSIXct("2015-10-01", format = "%Y-%m-%d") -
store$Promo2Since),
     100, main = "Days since start of promo2")
table(store$PromoInterval)
```

The stores with promos tend to make lower sales. This does not
necessary mean
that the promos don't help or are counterproductive. They are possibly
measures
that are taken mainly by stores with low sales in the first place:

```{r}
# Merge store and train
train_store <- merge(train, store, by = "Store")
ggplot(train_store[Sales != 0], aes(x = factor(PromoInterval), y =
Sales)) +
    geom_jitter(alpha = 0.1) +
    geom_boxplot(color = "yellow", outlier.colour = NA, fill = NA)
```

The different store types and assortment types imply different overall
levels of sales and seem to
be exhibiting different trends:

```{r}
ggplot(train_store[Sales != 0],
       aes(x = as.Date(Date), y = Sales, color = factor(StoreType))) +
    geom_smooth(size = 2)
ggplot(train_store[Customers != 0],
       aes(x = as.Date(Date), y = Customers, color =
factor(StoreType))) +
    geom_smooth(size = 2)
ggplot(train_store[Sales != 0],
       aes(x = as.Date(Date), y = Sales, color = factor(Assortment)))
+
    geom_smooth(size = 2)
ggplot(train_store[Sales != 0],
       aes(x = as.Date(Date), y = Customers, color =
factor(Assortment))) +
    geom_smooth(size = 2)
```

The effect of the distance to the next competitor is a little
counterintuitive.
Lower distance to the next competitor implies (slightly, possibly not
significantly)
higher sales. This may occur
(my assumption) because stores with a low distance to the next
competitor are
located in inner cities or crowded regions with higher sales in
general. Maybe

the effects of being in a good / bad region and having a competitor / not
having a competitor cancel out:

```{r}
salesByDist <- aggregate(train_store[Sales != 0 &
!is.na(CompetitionDistance)]$Sales,
              by = list(train_store[Sales != 0 &
!is.na(CompetitionDistance)]$CompetitionDistance), mean)
colnames(salesByDist) <- c("CompetitionDistance", "MeanSales")
ggplot(salesByDist, aes(x = log(CompetitionDistance), y =
log(MeanSales))) +
    geom_point() + geom_smooth()
```

A missing value for `CompetitionDistance` doesn't necessarily mean that there is
no competiton. Maybe that data was just not collected, yet. There is no obvious
connection between sales and having `NA` as `CompetitionDistance`:

```{r}
ggplot(train_store[Sales != 0],
       aes(x = factor(!is.na(CompetitionOpenSinceYear)), y = Sales)) +
    geom_jitter(alpha = 0.1) +
    geom_boxplot(color = "yellow", outlier.colour = NA, fill = NA) +
    ggtitle("Any competition?")
```

So what happens if a competitor opens? In order to assess this effect we fetch
data from all stores that first have `NA` as `CompetitorDistance` and later some
value. **Only the month, not the date, of the opening of the competitor is known
so we need a rather large window to see the effect (100 days)**. 147 stores
had a competitor move into their area during the available time span. The
competition leaves a 'dent' in the sales which looks a little different
depending on the chosen `timespan` so I wouldn't like to argue about statistical
significance based on this plot alone. It's informative to look at anyway:

```{r}
# Sales before and after competition opens
train_store$DateYearmon <- as.yearmon(train_store$Date)
train_store <- train_store[order(Date)]
timespan <- 100 # Days to collect before and after Opening of
competition
beforeAndAfterComp <- function(s) {
    x <- train_store[Store == s]
    daysWithComp <- x$CompetitionOpenSince >= x$DateYearmon
    if (any(!daysWithComp)) {
        compOpening <- head(which(!daysWithComp), 1) - 1
        if (compOpening > timespan & compOpening < (nrow(x) -
timespan)) {
           x <- x[(compOpening - timespan):(compOpening + timespan), ]
            x$Day <- 1:nrow(x)
            return(x)
        }
    }
```

```
}
temp <-
lapply(unique(train_store[!is.na(CompetitionOpenSince)]$Store),
beforeAndAfterComp)
temp <- do.call(rbind, temp)
# 147 stores first had no competition but at least 100 days before the
end
# of the data set
length(unique(temp$Store))
ggplot(temp[Sales != 0], aes(x = Day, y = Sales)) +
    geom_smooth() +
    ggtitle(paste("Competition opening around day", timespan))
```


The seasonplot is adapted from
[spsrini](https://www.kaggle.com/spsrini/rossmann-store-
sales/seasonplot-month/files)
(edit: Replace sum and show sales in relation to mean daily sales of a
store which
better accounts for missing data / closed stores):

```{r, results='hide'}
temp <- train
temp$year <- format(temp$Date, "%Y")
temp$month <- format(temp$Date, "%m")
temp[, StoreMean := mean(Sales), by = Store]
temp <- temp[, .(MonthlySalesMean = mean(Sales / (StoreMean)) * 100),
             by = .(year, month)]
temp <- as.data.frame(temp)
```

```{r}
SalesTS <- ts(temp$MonthlySalesMean, start=2013, frequency=12)
col = rainbow(3)
seasonplot(SalesTS, col=col, year.labels.left = TRUE, pch=19, las=1)
```

Παράρτημα Β: Μοντέλο πρόβλεψης πωλήσεων (κώδικας σε γλώσσα Python)

https://www.kaggle.com/kokovidis/rossman-predict

```
import warnings
warnings.filterwarnings("ignore")

# loading packages
# basic + dates
import numpy as np
import pandas as pd
from pandas import datetime

# data visualization
import matplotlib.pyplot as plt
import seaborn as sns # advanced vizs
%matplotlib inline

# statistics
from statsmodels.distributions.empirical_distribution import ECDF

# time series analysis
from statsmodels.tsa.seasonal import seasonal_decompose
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf

# prophet by Facebook
from fbprophet import Prophet

# machine learning: XGB
import xgboost as xgb
from sklearn.model_selection import train_test_split, GridSearchCV,
RandomizedSearchCV
from xgboost.sklearn import XGBRegressor # wrapper

# importing train data to learn
train = pd.read_csv("../input/train.csv",
                    parse_dates = True, low_memory = False, index_col
= 'Date')

# additional store data
store = pd.read_csv("../input/store.csv",
                    low_memory = False)
# time series as indexes
train.index

# first glance at the train set: head and tail
print("In total: ", train.shape)
train.head(5).append(train.tail(5))

# data extraction
train['Year'] = train.index.year
train['Month'] = train.index.month
train['Day'] = train.index.day
train['WeekOfYear'] = train.index.weekofyear

# adding new variable
train['SalePerCustomer'] = train['Sales']/train['Customers']
train['SalePerCustomer'].describe()
```

```
sns.set(style = "ticks")# to format into seaborn
c = '#386B7F' # basic color for plots
plt.figure(figsize = (12, 6))

plt.subplot(311)
cdf = ECDF(train['Sales'])
plt.plot(cdf.x, cdf.y, label = "statmodels", color = c);
plt.xlabel('Sales'); plt.ylabel('ECDF');

# plot second ECDF
plt.subplot(312)
cdf = ECDF(train['Customers'])
plt.plot(cdf.x, cdf.y, label = "statmodels", color = c);
plt.xlabel('Customers');

# plot second ECDF
plt.subplot(313)
cdf = ECDF(train['SalePerCustomer'])
plt.plot(cdf.x, cdf.y, label = "statmodels", color = c);
plt.xlabel('Sale per Customer');

# closed stores
train[(train.Open == 0) & (train.Sales == 0)].head()

# opened stores with zero sales
zero_sales = train[(train.Open != 0) & (train.Sales == 0)]
print("In total: ", zero_sales.shape)
zero_sales.head(5)

print("Closed stores and days which didn't have any sales won't be
counted into the forecasts.")
train = train[(train["Open"] != 0) & (train['Sales'] != 0)]

print("In total: ", train.shape)

# additional information about the stores
store.head()

# missing values?
store.isnull().sum()

# missing values in CompetitionDistance
store[pd.isnull(store.CompetitionDistance)]

# fill NaN with a median value (skewed distribuion)
store['CompetitionDistance'].fillna(store['CompetitionDistance'].media
n(), inplace = True)

# no promo = no information about the promo?
_ = store[pd.isnull(store.Promo2SinceWeek)]
_[_.Promo2 != 0].shape

# replace NA's by 0
store.fillna(0, inplace = True)

print("Joining train set with an additional store information.")

# by specifying inner join we make sure that only those observations
# that are present in both train and store sets are merged together
train_store = pd.merge(train, store, how = 'inner', on = 'Store')

print("In total: ", train_store.shape)
train_store.head()
```

```
train_store.groupby('StoreType')['Sales'].describe()

train_store.groupby('StoreType')['Customers', 'Sales'].sum()

# sales trends
sns.factorplot(data = train_store, x = 'Month', y = "Sales",
               col = 'StoreType', # per store type in cols
               palette = 'plasma',
               hue = 'StoreType',
               row = 'Promo', # per promo in the store in rows
               color = c)

# sales trends
sns.factorplot(data = train_store, x = 'Month', y = "Customers",
               col = 'StoreType', # per store type in cols
               palette = 'plasma',
               hue = 'StoreType',
               row = 'Promo', # per promo in the store in rows
               color = c)

# sale per customer trends
sns.factorplot(data = train_store, x = 'Month', y = "SalePerCustomer",
               col = 'StoreType', # per store type in cols
               palette = 'plasma',
               hue = 'StoreType',
               row = 'Promo', # per promo in the store in rows
               color = c)

# customers
sns.factorplot(data = train_store, x = 'Month', y = "Sales",
               col = 'DayOfWeek', # per store type in cols
               palette = 'plasma',
               hue = 'StoreType',
               row = 'StoreType', # per store type in rows
               color = c)

# stores which are opened on Sundays
train_store[(train_store.Open == 1) & (train_store.DayOfWeek ==
7)]['Store'].unique()

# competition open time (in months)
train_store['CompetitionOpen'] = 12 * (train_store.Year -
train_store.CompetitionOpenSinceYear) + \
        (train_store.Month - train_store.CompetitionOpenSinceMonth)

# Promo open time
train_store['PromoOpen'] = 12 * (train_store.Year -
train_store.Promo2SinceYear) + \
        (train_store.WeekOfYear - train_store.Promo2SinceWeek) / 4.0

# replace NA's by 0
train_store.fillna(0, inplace = True)

# average PromoOpen time and CompetitionOpen time per store type
train_store.loc[:, ['StoreType', 'Sales', 'Customers', 'PromoOpen',
'CompetitionOpen']].groupby('StoreType').mean()

# Compute the correlation matrix
# exclude 'Open' variable
corr_all = train_store.drop('Open', axis = 1).corr()

# Generate a mask for the upper triangle
mask = np.zeros_like(corr_all, dtype = np.bool)
```

```
mask[np.triu_indices_from(mask)] = True

# Set up the matplotlib figure
f, ax = plt.subplots(figsize = (11, 9))

# Draw the heatmap with the mask and correct aspect ratio
sns.heatmap(corr_all, mask = mask,
            square = True, linewidths = .5, ax = ax, cmap = "BuPu")
plt.show()

# sale per customer trends
sns.factorplot(data = train_store, x = 'DayOfWeek', y = "Sales",
               col = 'Promo',
               row = 'Promo2',
               hue = 'Promo2',
               palette = 'RdPu')

# preparation: input should be float type
train['Sales'] = train['Sales'] * 1.0

# store types
sales_a = train[train.Store == 2]['Sales']
sales_b = train[train.Store == 85]['Sales'].sort_index(ascending =
True) # solve the reverse order
sales_c = train[train.Store == 1]['Sales']
sales_d = train[train.Store == 13]['Sales']

f, (ax1, ax2, ax3, ax4) = plt.subplots(4, figsize = (12, 13))

# store types
sales_a.resample('W').sum().plot(color = c, ax = ax1)
sales_b.resample('W').sum().plot(color = c, ax = ax2)
sales_c.resample('W').sum().plot(color = c, ax = ax3)
sales_d.resample('W').sum().plot(color = c, ax = ax4)

f, (ax1, ax2, ax3, ax4) = plt.subplots(4, figsize = (12, 13))

# monthly
decomposition_a = seasonal_decompose(sales_a, model = 'additive', freq
= 365)
decomposition_a.trend.plot(color = c, ax = ax1)

decomposition_b = seasonal_decompose(sales_b, model = 'additive', freq
= 365)
decomposition_b.trend.plot(color = c, ax = ax2)

decomposition_c = seasonal_decompose(sales_c, model = 'additive', freq
= 365)
decomposition_c.trend.plot(color = c, ax = ax3)

decomposition_d = seasonal_decompose(sales_d, model = 'additive', freq
= 365)
decomposition_d.trend.plot(color = c, ax = ax4)

# figure for subplots
plt.figure(figsize = (12, 8))

# acf and pacf for A
plt.subplot(421); plot_acf(sales_a, lags = 50, ax = plt.gca(), color =
c)
plt.subplot(422); plot_pacf(sales_a, lags = 50, ax = plt.gca(), color
= c)

# acf and pacf for B
```

```
plt.subplot(423); plot_acf(sales_b, lags = 50, ax = plt.gca(), color =
c)
plt.subplot(424); plot_pacf(sales_b, lags = 50, ax = plt.gca(), color
= c)

# acf and pacf for C
plt.subplot(425); plot_acf(sales_c, lags = 50, ax = plt.gca(), color =
c)
plt.subplot(426); plot_pacf(sales_c, lags = 50, ax = plt.gca(), color
= c)

# acf and pacf for D
plt.subplot(427); plot_acf(sales_d, lags = 50, ax = plt.gca(), color =
c)
plt.subplot(428); plot_pacf(sales_d, lags = 50, ax = plt.gca(), color
= c)

plt.show()

# importing data
df = pd.read_csv("../input/train.csv",
                    low_memory = False)

# remove closed stores and those with no sales
df = df[(df["Open"] != 0) & (df['Sales'] != 0)]

# sales for the store number 1 (StoreType C)
sales = df[df.Store == 1].loc[:, ['Date', 'Sales']]

# reverse to the order: from 2013 to 2015
sales = sales.sort_index(ascending = False)

# to datetime64
sales['Date'] = pd.DatetimeIndex(sales['Date'])
sales.dtypes

# from the prophet documentation every variables should have specific
names
sales = sales.rename(columns = {'Date': 'ds',
                                'Sales': 'y'})
sales.head()

# plot daily sales
ax = sales.set_index('ds').plot(figsize = (12, 4), color = c)
ax.set_ylabel('Daily Number of Sales')
ax.set_xlabel('Date')
plt.show()

# create holidays dataframe
state_dates = df[(df.StateHoliday == 'a') | (df.StateHoliday == 'b') &
(df.StateHoliday == 'c')].loc[:, 'Date'].values
school_dates = df[df.SchoolHoliday == 1].loc[:, 'Date'].values

state = pd.DataFrame({'holiday': 'state_holiday',
                      'ds': pd.to_datetime(state_dates)})
school = pd.DataFrame({'holiday': 'school_holiday',
                      'ds': pd.to_datetime(school_dates)})

holidays = pd.concat((state, school))
holidays.head()

# set the uncertainty interval to 95% (the Prophet default is 80%)
my_model = Prophet(interval_width = 0.95,
                   holidays = holidays)
```

```
my_model.fit(sales)

# dataframe that extends into future 6 weeks
future_dates = my_model.make_future_dataframe(periods = 6*7)

print("First week to forecast.")
future_dates.tail(7)

# predictions
forecast = my_model.predict(future_dates)

# preditions for last week
forecast[['ds', 'yhat', 'yhat_lower', 'yhat_upper']].tail(7)

fc = forecast[['ds', 'yhat']].rename(columns = {'Date': 'ds',
'Forecast': 'yhat'})

# visualizing predicions
my_model.plot(forecast);

my_model.plot_components(forecast);

# to predict to
test = pd.read_csv("../input/test.csv",
                   parse_dates = True, low_memory = False, index_col
= 'Date')
test.head()

# test: missing values?
test.isnull().sum()

test[pd.isnull(test.Open)]

# replace NA's in Open variable by 1
test.fillna(1, inplace = True)

# data extraction
test['Year'] = test.index.year
test['Month'] = test.index.month
test['Day'] = test.index.day
test['WeekOfYear'] = test.index.weekofyear

# to numerical
mappings = {'0':0, 'a':1, 'b':2, 'c':3, 'd':4}
test.StateHoliday.replace(mappings, inplace = True)

train_store.Assortment.replace(mappings, inplace = True)
train_store.StoreType.replace(mappings, inplace = True)
train_store.StateHoliday.replace(mappings, inplace = True)
train_store.drop('PromoInterval', axis = 1, inplace = True)

store.StoreType.replace(mappings, inplace = True)
store.Assortment.replace(mappings, inplace = True)
store.drop('PromoInterval', axis = 1, inplace = True)

# take a look on the train and store again
train_store.head()

print("Joining test set with an additional store information.")
test_store = pd.merge(test, store, how = 'inner', on = 'Store')

test_store['CompetitionOpen'] = 12 * (test_store.Year -
test_store.CompetitionOpenSinceYear) + (test_store.Month -
test_store.CompetitionOpenSinceMonth)
```

```
test_store['PromoOpen'] = 12 * (test_store.Year -
test_store.Promo2SinceYear) + (test_store.WeekOfYear -
test_store.Promo2SinceWeek) / 4.0

print("In total: ", test_store.shape)
test_store.head()

# split into training and evaluation sets
# excluding Sales and Id columns
predictors = [x for x in train_store.columns if x not in ['Customers',
'Sales', 'SalePerCustomer']]
y = np.log(train_store.Sales) # log transformation of Sales
X = train_store

# split the data into train/test set
X_train, X_test, y_train, y_test = train_test_split(X, y,
                                                    test_size = 0.3, #
30% for the evaluation set
                                                    random_state = 42)

# predictors
X.columns

# evaluation metric: rmspe
# Root Mean Square Percentage Error
# code chunk shared at Kaggle

def rmspe(y, yhat):
    return np.sqrt(np.mean((yhat / y-1) ** 2))

def rmspe_xg(yhat, y):
    y = np.expm1(y.get_label())
    yhat = np.expm1(yhat)
    return "rmspe", rmspe(y, yhat)

# base parameters
params = {
    'booster': 'gbtree',
    'objective': 'reg:linear', # regression task
    'subsample': 0.8, # 80% of data to grow trees and prevent
overfitting
    'colsample_bytree': 0.85, # 85% of features used
    'eta': 0.1,
    'max_depth': 10,
    'seed': 42} # for reproducible results

# XGB with xgboost library
dtrain = xgb.DMatrix(X_train[predictors], y_train)
dtest = xgb.DMatrix(X_test[predictors], y_test)

watchlist = [(dtrain, 'train'), (dtest, 'test')]

xgb_model = xgb.train(params, dtrain, 300, evals = watchlist,
                      early_stopping_rounds = 50, feval = rmspe_xg,
verbose_eval = True)

# XGB with sklearn wrapper
# the same parameters as for xgboost model
params_sk = {'max_depth': 10,
            'n_estimators': 300, # the same as num_rounds in xgboost
            'objective': 'reg:linear',
            'subsample': 0.8,
            'colsample_bytree': 0.85,
            'learning_rate': 0.1,
```

```
            'seed': 42}

skrg = XGBRegressor(**params_sk)

skrg.fit(X_train, y_train)

import scipy.stats as st

params_grid = {
    'learning_rate': st.uniform(0.01, 0.3),
    'max_depth': list(range(10, 20, 2)),
    'gamma': st.uniform(0, 10),
    'reg_alpha': st.expon(0, 50)}

search_sk = RandomizedSearchCV(skrg, params_grid, cv = 5) # 5 fold
cross validation
search_sk.fit(X_train, y_train)

# best parameters
print(search_sk.best_params_); print(search_sk.best_score_)

# with new parameters
params_new = {
    'booster': 'gbtree',
    'objective': 'reg:linear',
    'subsample': 0.8,
    'colsample_bytree': 0.85,
    'eta': 0.044338624448041611,
    'max_depth': 16,
    'gamma': 0.80198330585415034,
    'reg_alpha': 23.008226565535971,
    'seed': 42}

model_final = xgb.train(params_new, dtrain, 300, evals = watchlist,
                        early_stopping_rounds = 50, feval = rmspe_xg,
verbose_eval = True)

yhat = model_final.predict(xgb.DMatrix(X_test[predictors]))
error = rmspe(X_test.Sales.values, np.exp(yhat))

print('First validation yelds RMSPE: {:.6f}'.format(error))

xgb.plot_importance(model_final)

# predictions to unseen data
unseen = xgb.DMatrix(test_store[predictors])
test_p = model_final.predict(unseen)

forecasts = pd.DataFrame({'Id': test['Id'],
                          'Sales': np.exp(test_p)})
# forecasts
forecasts.head()

# first
# 0.66419
test_base = xgb_model.predict(unseen)

forecasts_base = pd.DataFrame({'Id': test['Id'],
                          'Sales': np.exp(test_base)})
forecasts_base.to_csv("xgboost_2_submission.csv", index = False)

# final
# 0.60553
forecasts.to_csv("xgboost_submission.csv", index = False)
```