ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ

Χρήση Τεχνολογιών Μεγάλων Δεδομένων στο Ηλεκτρονικό Επιχειρείν

ΖΑΜΙΧΟΣ ΠΑΝΤΕΛΗΣ
ΚΩΝΣΤΑΝΤΙΝΟΣ

Υποβλήθηκε ως απαιτούμενο για την απόκτηση του μεταπτυχιακού διπλώματος στην Αναλυτική των Επιχειρήσεων και Επιστήμη των Δεδομένων

Τμήμα Οργάνωσης και Διοίκησης Επιχειρήσεων

Επιβλέπων Καθηγητής: Ταραμπάνης Κωνσταντίνος

Σεπτέμβριος 2022

Στην οικογένειά μου,

Ευχαριστίες

Θα ήθελα αρχικά να ευχαριστήσω τους καθηγητές μου κύριο Ταραμπάνη Κωνσταντίνο και κύριο Καλαμπόκη Ευάγγελο για την ευκαιρία που μου δώσανε να ασχοληθώ με το συγκεκριμένο θέμα, καθώς και για την καθοδήγηση και τις συμβουλές τους, στην εκπόνηση της διπλωματικής μου εργασίας. Επίσης θα ήθελα να ευχαριστήσω τους φίλους μου και τη γυναίκα μου Αλεξάνδρα, για τη στήριξη και τη συνεχή ενθάρρυνση καθ' όλη τη διάρκεια των σπουδών μου, καθώς και κατά τη διαδικασία της έρευνας και της συγγραφής αυτής της διατριβής, η ολοκλήρωση της οποίας θα ήταν αδύνατη χωρίς αυτούς. Σας ευχαριστώ.

Περίληψη

Τα τελευταία χρόνια, καθημερινά παράγονται τεράστιες ποσότητες δεδομένων είτε από οργανισμούς είτε από επιχειρήσεις. Τα δεδομένα αυτά, χάρη στην ραγδαία ανάπτυξη της τεχνολογίας, μπορούν με ευχέρεια να αποθηκευτούν και να επεξεργαστούν. Στόχος είναι η παραγωγή πληροφορίας, υψηλότερης προστιθέμενης αξίας, αξιοποιώντας τις τεράστιες ποσότητες δεδομένων που είναι στη διάθεση των οργανισμών και των επιχειρήσεων. Η ανάπτυξη της βαθιάς μάθησης (deep learning) ως η κινητήρια δύναμη της μηχανικής μάθησης (machine learning) έφερε επανάσταση στον τομέα της πληροφορικής και της ανάλυσης των δεδομένων. Ένα από τα προβλήματα που απασχολούν τον κλάδο της ανάλυσης δεδομένων, όσο και των επιχειρήσεων είναι η πρόβλεψη των μελλοντικών πωλήσεων. Μέσω της συλλογής ιστορικών δεδομένων, τα οποία σχετίζονται με προηγούμενες πωλήσεις και με τη χρήση των κατάλληλων αλγορίθμων μπορούμε να προβλέψουμε τις μελλοντικές πωλήσεις. Πληθώρα μμοντέλων Βαθιάς Μάθησης έχουν προταθεί στη βιβλιογραφία για την επίλυση των προβλημάτων πρόβλεψης πωλήσεων.

Πρόσφατα, η βαθιά μάθηση έχει αναδιαρθρώσει το δυνητικό μέλλον της πρόβλεψης πωλήσεων, επιτρέποντας στα μοντέλα να κωδικοποιούν πολλαπλές χρονοσειρές σε ένα μόνο μοντέλο, με τη χρήση όχι μόνο αριθμητικών αλλά και κατηγορικών μεταβλητών.

Στην τρέχουσα διπλωματική εργασία, αντιμετωπίζουμε το πρόβλημα της πρόβλεψης πωλήσεων, σε καταστήματα λιανικού εμπορίου. Η πρόβλεψη πραγματοποιήθηκε για τα καταστήματα της αμερικάνικης πολυεθνικής εταιρίας Walmart, για τις επόμενες είκοσι οκτώ (28) ημέρες. Χρησιμοποιήθηκε η μέθοδος μηχανικής μάθησης LightGBM για την πραγματοποίηση της πρόβλεψης, ανά πολιτεία, κατάστημα, κατηγορία και τμήμα, λαμβάνοντας επίσης υπόψη και επεξηγηματικές μεταβλητές, όπως η τιμή των προϊόντων, προσφορές, ημέρα της εβδομάδας και ειδικές εκδηλώσεις, όπως για παράδειγμα γιορτές, καθώς επίσης και επιπλέον μεταβλητών που δημιουργήθηκαν για την βελτιστοποίηση των μοντέλων και των προβλέψεων. Πειραματικά μοντέλα παρουσιάζονται για τις παραπάνω μεθόδους.

Ως μέτρηση αξιολόγησης και σύγκρισης των μεθόδων, χρησιμοποιήθηκε η ρίζα του μέσου τετραγωνικού σφάλματος (RMSE).

Η παρακάτω έρευνα μπορεί να χρησιμοποιηθεί, από ιδιωτικές επιχειρήσεις καθώς και από οργανισμούς, για τη λήψη αποφάσεων όπως και καλύτερου προγραμματισμού των εταιρειών. Μέσω της συγκεκριμένης έρευνας μπορεί να επιτευχθεί μείωση του λειτουργικού κόστους εργασίας και αύξηση του προσδοκώμενου κέρδους των επιχειρήσεων.

Λέξεις Κλειδιά: Πρόβλεψη Πωλήσεων, Πρόβλεψη Χρονοσειρών, Μηχανική Μάθηση, LightGBM

Abstract

Over the last years, huge amounts of data are generated every day either by organizations or businesses. This data, thanks to the rapid development of technology, can easily be stored and processed. The goal is, to provide information, with higher added value, utilizing the huge amounts of data,that are available to organizations and businesses. The rise of deep learning as the driving force behind machine learning has revolutionized the field of computer science and data analysis. One of the problems that concern the field of data analysis, as well as business, is the prediction of future sales. Through the collection of historical data related to past sales, and using appropriate algorithms, we can predict future sales. Numerous Deep Learning models have been proposed in the literature, in order to solve sales forecasting projects.

Recently, deep learning has restructured the potential future of sales forecasting by allowing models to encode multiple time series into a single model, using not only numerical but also categorical variables.

In the current thesis, we deal with the problem of forecasting sales in retail stores. The forecast was held for the stores of the American multinational company Walmart, for the next twenty-eight (28) days. The LightGBM machine learning method was used in order to make the prediction possible, per state, store, category and department, also taking account explanatory variables such as product price, promotions, day of the week and special events such as holidays, as well as additional variables created to optimize models and predictions. Experimental models demonstrate the above methods.

Root mean square error (RMSE) was used as a metric to evaluate and compare the methods.

The following research can be used, by private businesses as well as organizations, for decision making, as well as better planning of companies. Through this specific research, a reduction in operational labor costs and an increase in the expected profit of businesses can be achieved.

Keywords: Sales Forecasting, Time Series Forecasting, Machine Learning, LightGBM

Περιεχόμενα

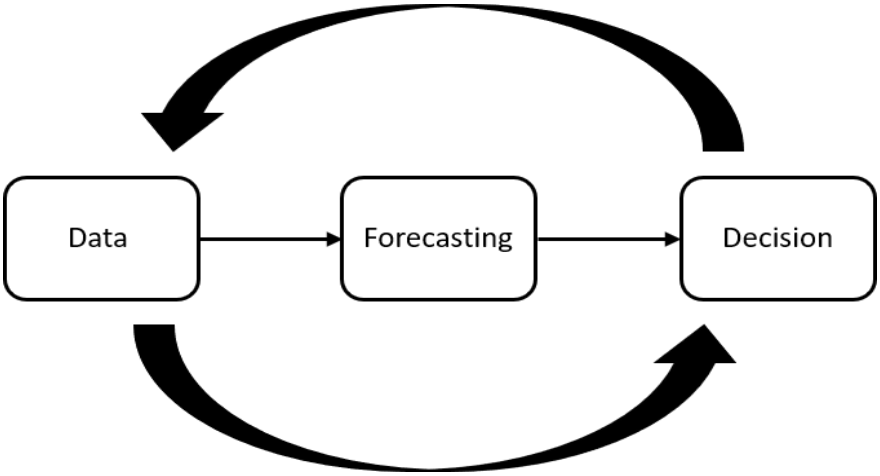
Λίστα Εικόνων

Λίστα Πινάκων

1. Εισαγωγή

1.1. Περιγραφή του προβλήματος

Η πρόβλεψη είναι ίσως η πιο κοινή εφαρμογή της μηχανικής μάθησης στον πραγματικό κόσμο. Οι επιχειρήσεις προβλέπουν τη ζήτηση προϊόντων, οι κυβερνήσεις προβλέπουν οικονομική και πληθυσμιακή ανάπτυξη, οι μετεωρολόγοι προβλέπουν τον καιρό. Η κατανόηση του τι πρόκειται να συμβεί, είναι επιτακτική τόσο για την επιστήμη, όσο και για τη βιομηχανία και τις κυβερνήσεις, με τη χρήση της μηχανικής μάθησης να εφαρμόζεται όλο και περισσότερο σε αυτούς τους τομείς, προκειμένου να αντιμετωπιστεί η ανάγκη της πρόβλεψης. Όπως φαίνεται στην Εικόνα 1, η πρόβλεψη βασίζεται στη συλλογή δεδομένων. Τα διαθέσιμα δεδομένα χρησιμοποιούνται για την πρόβλεψη, η οποία βοηθά στην λήψη αποφάσεων. Η διαδικασία αυτή είναι ένας ατέρμων κύκλος καθώς οι νέες αποφάσεις αποτελούν τα νέα δεδομένα τα οποία θα χρησιμοποιηθούν για μελλοντικές προβλέψεις.



Εικόνα 1: Η διαδικασία λήψης αποφάσεων με την βοήθεια της πρόβλεψης βασισμένη στα διαθέσιμα δεδομένα. [30]

Η πρόβλεψη χρονοσειρών είναι ένα ευρύ πεδίο με μακρά ιστορία. Παλαιότερα, οι εταιρείες συνήθιζαν να παράγουν προϊόντα χωρίς να λάβουν υπόψη τον αριθμό των πωλήσεων, καθώς και τη ζήτηση. Για να καθορίσει οποιοσδήποτε κατασκευαστής αν θα προβεί σε αύξηση ή μείωση της παραγωγής προϊόντων, απαιτούνται στοιχεία και δεδομένα σχετικά με τη ζήτηση των προϊόντων στην αγορά. Οι εταιρείες μπορεί να

αντιμετωπίσουν ζημιές, εάν δεν λάβουν υπόψη αυτά τα δεδομένα, στον αγώνα τους για επιβίωση, ανάπτυξη και δημιουργία κερδών, σε μια ανταγωνιστική αγορά.

Τα τελευταία χρόνια, σε ένα ιδιαίτερα υψηλά ανταγωνιστικό περιβάλλον, και σε ένα συνεχώς μεταβαλλόμενο καταναλωτικό τοπίο, όλο και περισσότερες επιχειρήσεις προσπαθούν να προβλέψουν τις μελλοντικές πωλήσεις των προϊόντων και των υπηρεσιών τους. Είναι ένα κρίσιμο βήμα για τον προγραμματισμό της επιχείρησης, τόσο στην ακριβή και έγκαιρη κατασκευή - δημιουργία προϊόντων ή υπηρεσιών, όσο και στη διανομή και τη λιανική πώληση των αγαθών**.** Οι βραχυπρόθεσμες προβλέψεις βοηθούν κυρίως στον προγραμματισμό παραγωγής και τη διαχείριση αποθεμάτων, ενώ οι μακροπρόθεσμες προβλέψεις εστιάζουν στην επιχειρηματική ανάπτυξη και τη λήψη αποφάσεων. (Doganis et al. 2006)

Η πρόβλεψη πωλήσεων είναι η έκφραση των αναμενόμενων εσόδων από τις πωλήσεις. Υπολογίζει τι πρόκειται να πουλήσει η εταιρεία μέσα σε μια συγκεκριμένη χρονική περίοδο, όπως τον επόμενο μήνα τρίμηνο ή έτος. Οι προβλέψεις πωλήσεων διαφέρουν ως προς το πού και πώς λαμβάνουν τα δεδομένα τους, για παράδειγμα, μπορεί να βασίζονται στη διαίσθηση των αντιπροσώπων πωλήσεων ή ακόμη και στην τεχνητή νοημοσύνη (AI). Ανεξάρτητα όμως από το πως γίνεται η πρόβλεψη, απαντώνται δυο βασικά ερωτήματα.

- **Πόσο:** Κάθε ευκαιρία πώλησης έχει το δικό της προβλεπόμενο ποσό που θα φέρει σε κάθε επιχείρηση. Μέσω της πρόβλεψης πωλήσεων, οι αναλυτές προσπαθούν να βρουν το ποσό αυτό, λαμβάνοντας υπόψη όλα όσα γνωρίζουν, σε συνδυασμό με τα αποτελέσματα των αναλύσεων.
- **Πότε:** Οι προβλέψεις πωλήσεων υποδεικνύουν ένα μήνα, τρίμηνο ή έτος, κατά το οποίο η ομάδα πωλήσεων αναμένει ό,τι τα έσοδα θα αυξηθούν ή θα μειωθούν αντίστοιχα.

Η αφθονία των δεδομένων πωλήσεων και των συσχετιζόμενων πληροφοριών, μπορούν να χρησιμοποιηθούν μέσω διαφόρων τεχνικών Μηχανικής Μάθησης, για την ανάπτυξη προβλεπτικών μοντέλων. Είναι μια τεχνική η οποία δεν επηρεάζεται από την όποια διαίσθηση των διευθυντών πωλήσεων, είναι ευέλικτη, πράγμα που σημαίνει ότι είναι εύκολα προσαρμόσιμη στις οποιεσδήποτε αλλαγές των νέων δεδομένων, καθώς και υπολογίζει με μεγαλύτερη ακρίβεια μια πρόβλεψη ακόμα και από έναν ειδικό. Για παράδειγμα τις προηγούμενες δεκαετίες όταν οι εταιρίες

συνήθιζαν να παράγουν προϊόντα χωρίς να λάβουν υπόψη τον αριθμό των πωλήσεων και της ζήτησης, συχνά ερχόταν αντιμέτωπες με αρκετά προβλήματα, από τη στιγμή που δεν ήξεραν πόση ποσότητα δυνητικά θα πουλήσουν, ώστε ο κατασκευαστής να αυξήσει ή αντίστοιχα να μειώσει τον αριθμό των παραγόμενων μονάδων.

Υπάρχουν διάφορες τεχνικές μέσω των οποίων οι εταιρίες έχουν επικεντρωθεί στην πρόβλεψη των πωλήσεων, διάφορα στατιστικά μοντέλα για χρονοσειρές, γραμμική παλινδρόμηση, feature engineering, τυχαία δάση και αρκετές ακόμη.

Οι χρονοσειρές περιέχουν σημεία δεδομένων που αποθηκεύονται σε μια καθορισμένη περίοδο και χρησιμοποιούνται για να προβλεφθεί το μέλλον. Είναι δηλαδή μια συλλογή σημείων που κατανεμημένα διαδοχικά και ομοιόμορφα σε μια περίοδο. Στις εφαρμογές πρόβλεψης, οι παρατηρήσεις συνήθως καταγράφονται με τακτική συχνότητα, όπως για παράδειγμα ημερήσια ή μηνιαία. Κύρια χαρακτηριστικά των χρονοσειρών που μελετώνται είναι οι τάσεις (trends), η εποχικότητα (seasonality), η κυκλικότητα (cyclicity).

1.2. Αντικείμενο και Στόχοι της Μελέτης

Οι εποχές που διανύουμε χαρακτηρίζονται από παγκόσμια οικονομική αβεβαιότητα. Οι επιχειρηματίες ενδιαφέρονται για τη μεγιστοποίηση των κερδών τους, καθώς και για την εκτίμηση - πρόβλεψη τους, τόσο μακροπρόθεσμα όσο και βραχυπρόθεσμα. Πολλοί από αυτούς εξακολουθούν να παλεύουν με τον τρόπο πρόβλεψης των εσόδων για το επόμενο έτος, το οποίο είναι συχνά το σημείο εκκίνησης των ετήσιων προϋπολογισμών των επιχειρήσεων. Τώρα περισσότερο από ποτέ, οι επιχειρήσεις βασίζονται στην ικανότητά των ομάδων πωλήσεων, να προβλέπουν, για να σχεδιάσουν ολόκληρη τη στρατηγική ανάπτυξής τους.

Θα μπορούσαμε να πούμε ότι η πρόβλεψη πωλήσεων είναι ένας συνδυασμός επιστήμης και τέχνης. Οι υπεύθυνοι λήψης αποφάσεων βασίζονται σε αυτές τις προβλέψεις για να σχεδιάσουν την επέκταση της επιχείρησης και να καθορίσουν τρόπους και τεχνικές ανάπτυξης της εταιρείας. Επομένως, η πρόβλεψη πωλήσεων επηρεάζει με πολλούς τρόπους, όλους τους τομείς σε μια επιχείρηση.

Ωστόσο, για τις περισσότερες εταιρείες μια ακριβή πρόβλεψη πωλήσεων εξακολουθεί να αποτελεί μια σημαντική πρόκληση. Χάρη σε ανακριβείς μεθόδους

πρόβλεψης που βασίζονται στη διαίσθηση, οι εταιρείες καταλήγουν να έχουν κακή διορατικότητα στις προβλεπόμενες πωλήσεις.

Η πρόβλεψη πωλήσεων είναι ιδιαίτερα σημαντική, για παράδειγμα στην αγορά τροφίμων. Λόγω της περιορισμένης διάρκειας ζωής αρκετών αγαθών, εσφαλμένη πρόβλεψη οδηγεί σε απώλεια κερδών τόσο σε περιπτώσεις έλλειψης όσο και σε καταστάσεις πλεονάσματος. Περισσότερες από τις αναμενόμενες παραγγελίες - πωλήσεις οδηγούν σε έλλειψη προϊόντων και οι λιγότερες παραγγελίες - πωλήσεις οδηγούν σε έλλειψη ευκαιρίας δημιουργίας κέρδους. Επιπλέον, η ζήτηση των καταναλωτών παρουσιάζει διαρκώς διακυμάνσεις βασή παραγόντων όπως η τιμή, οι προσφορές, οι αλλαγές στις προτιμήσεις των καταναλωτών ή οι αλλαγές καιρικές σε καιρικές συνθήκες (Van der Vorst et al. 1998). Επομένως η επιχείρηση, για να παραμένει ανταγωνιστική στην αγορά, χρειάζεται συνεχώς να έχει κατάλληλα επίπεδα αποθέματος στην ποσότητα των διαθέσιμων προϊόντων ή υπηρεσιών της.

Ανακεφαλαιώνοντας, η πρόβλεψη πωλήσεων είναι σημαντική γιατί, βοηθά κάθε επιχείρηση να πάρει καλύτερες επιχειρηματικές αποφάσεις, βοηθά στον συνολικό επιχειρηματικό σχεδιασμό, τον προϋπολογισμό και τη διαχείριση κινδύνων. Επιτρέπει στις εταιρείες να κατανέμουν αποτελεσματικά τους πόρους για μελλοντική ανάπτυξη και να διαχειρίζονται τις ταμειακές ροές τους. Βοηθά τις ομάδες πωλήσεων να επιτύχουν τους στόχους τους, εντοπίζοντας σημάδια έγκαιρης προειδοποίησης στη γραμμή πωλήσεων ώστε να προβούν σε διορθώσεις πριν να είναι πολύ αργά. Βοηθά τις επιχειρήσεις να εκτιμήσουν με ακρίβεια το κόστος και τα έσοδά τους, βάσει των οποίων είναι σε θέση να προβλέψουν τη βραχυπρόθεσμη και μακροπρόθεσμη απόδοσή τους.

1.3. Συνεισφορά

Στη συγκεκριμένη διπλωματική εργασία εστιάζουμε στο πρόβλημα της πρόβλεψης πωλήσεων για τα καταστήματα λιανικής πωλήσεις, το οποίο αντιμετωπίζεται σαν ένα πρόβλημα πρόβλεψης χρονοσειρών με μεθόδους μηχανικής μάθησης. Αρχικά γίνεται μία εισαγωγή στις έννοιες την των Μεγάλων Δεδομένων, και της Μηχανικής Μάθησης, ενώ στην συνέχεια αναλύονται σύνηθες τεχνικές εκπαίδευσης και αξιολόγησης που συναντώνται στη βιβλιογραφία για την πρόβλεψη χρονοσειρών. Επίσης, γίνεται αναφορά σε κάποιες από τις πιο αποτελεσματικές μεθόδους για την πρόβλεψη χρονοσειρών. Επιπρόσθετα σε αυτή τη διπλωματική γίνεται αξιολόγηση

και σύγκριση των μεθόδων που αναλύονται με βάση δεδομένα του πραγματικού κόσμου τα οποία παρέχονται από την αμερικάνικη πολυεθνική εταιρία Walmart. Τα πειράματα που παρουσιάζονται περιέχουν ανάλυση τόσο για την εύρεση των βέλτιστων παραμέτρων των αλγορίθμων που οδηγούν σε προβλέψεις με μικρό σφάλμα, όσο και ανάλυση για την εύρεση των συνδυασμό των διαθέσιμων μεταβλητών οι οποίες επηρεάζουν την πρόβλεψη για το συγκεκριμένο πρόβλημα.

1.4. Δομή μελέτης

Το υπόλοιπο της εργασίας ακολουθεί την εξής δομή. Αρχικά, στο Κεφάλαιο 2 δίνεται μια εισαγωγή σε βασικές έννοιες όπως τα Μεγάλα Δεδομένα, η εξόρυξη δεδομένων, η Μηχανική Μάθηση. Επίσης, αναλύονται οι κύριες κατηγορίες της Μηχανικής Μάθησης και των Συνδυαστικών Μοντέλων μάθησης. Στη συνέχεια, γίνεται ο ορισμός των χρονοσειρών, καθώς και η περιγραφή των βασικών τους χαρακτηριστικών. Το κεφάλαιο κλείνει με την περιγραφή βασικών τεχνικών αξιολόγησης για προβλήματα Μηχανικής Μάθησης και πρόβλεψης χρονοσειρών. Στο Κεφάλαιο 3, γίνεται η βιβλιογραφική ανασκόπηση με μερικές από τις σημαντικότερες τεχνικές για την πρόβλεψη χρονοσειρών και λεπτομερής περιγραφή αυτών. Στη συνέχεια, στο Κεφάλαιο 4 παρουσιάζονται τα πειράματα και τα αποτελέσματα τους που έγιναν για την αξιολόγηση των προαναφερθέντων μεθόδων, μαζί με τα εργαλεία υλοποίησης των πειραμάτων και μίας λεπτομερής ανάλυσης του συνόλου δεδομένων που χρησιμοποιήθηκαν για τα πειράματα. Τέλος, στο Κεφάλαιο 5 γίνεται η σύνοψη της εργασία και παρουσιάζονται τα συμπεράσματα που προέκυψαν από αυτή.

2. Θεωρητικό Υπόβαθρο

2.1. Μεγάλα Δεδομένα

Όπως αναφέρθηκε στο Κεφάλαιο 1, τα δεδομένα παίζουν πολύ σημαντικό ρόλο όχι μόνο στη λήψη αποφάσεων, αλλά αποτελούν και την κύρια είσοδο για τα μοντέλα πρόβλεψης. Τα τελευταία χρόνια με τη ραγδαία ανάπτυξη της τεχνολογίας και του διαδικτύου τα δεδομένα υπάρχουν σε μεγάλες ποσότητες, ενώ συχνά αναφέρονται σαν Μεγάλα Δεδομένα (Big Data).

Τα δεδομένα έχουν χαρακτηριστεί ως το «Νέο πετρέλαιο» [1], ωστόσο η χρήση τους απαιτεί την κατάλληλη επεξεργασία, όπως και με το πετρέλαιο «Το πετρέλαιο πρέπει να μετατραπεί σε αέριο, πλαστικό, χημικά κ.λπ. να δημιουργήσει μια πολύτιμη οντότητα που οδηγεί σε κερδοφόρα δραστηριότητα. Επομένως, τα δεδομένα πρέπει να επεξεργαστούν και να αναλυθούν για να έχουν αξία».

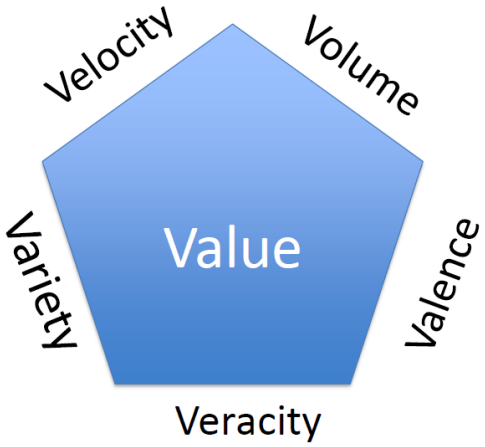
Το πεδίο των Μεγάλων Δεδομένων είναι ένα πεδίο που αντιμετωπίζει τρόπους ανάλυσης, συστηματικής εξαγωγής πληροφοριών από, σύνολα δεδομένων, τα οποία είναι πολύ μεγάλα ή πολύπλοκα για να αντιμετωπιστούν από τα παραδοσιακά λογισμικά εφαρμογών επεξεργασίας δεδομένων.

Τα κύρια χαρακτηριστικά των Μεγάλων Δεδομένων είναι τα γνωστά ως 4 V's (πλέον 6) και είναι τα ακόλουθα, όπως φαίνονται και από την Εικόνα 2:

- **Όγκος (Volume):** αναφέρεται στο τεράστιο μέγεθος των συνόλων δεδομένων που πρέπει να υποβληθούν σε επεξεργασία. Ο όγκος των δεδομένων που διαχειρίζονται οι εταιρείες εκτοξεύτηκε στα ύψη γύρω στο 2012, όταν άρχισαν να συλλέγουν περισσότερα από τρία εκατομμύρια στοιχεία για κάθε δεδομένα.
- **Ταχύτητα (Velocity)**: είναι η συχνότητα των εισερχόμενων δεδομένων που πρέπει να υποβληθούν σε επεξεργασία.
- **Αλήθεια (Veracity):** αναφέρεται στην αξιοπιστία των δεδομένων.
- **Ποικιλία (Variety)**: Τα μεγάλα δεδομένα προέρχονται από μια μεγάλη ποικιλία πηγών. Μια εταιρεία μπορεί να λάβει δεδομένα από πολλές διαφορετικές πηγές: από εσωτερικές συσκευές έως τεχνολογία GPS smartphone ή τι λένε οι άνθρωποι στα κοινωνικά δίκτυα.

- **Αξία (Value)**: η αξία που μπορούμε να λάβουμε από τα δεδομένα και πώς τα μεγάλα δεδομένα οδηγούν σε καλύτερα αποτελέσματα.

- **Valence**: Μέτρο συνδεσιμότητας. Αυτό αναφέρεται στο πώς τα μεγάλα δεδομένα μπορούν να συνδεθούν μεταξύ τους, δημιουργώντας συνδέσεις μεταξύ διαφορετικών κατά τα άλλα συνόλων δεδομένων.



Εικόνα 2: Χαρακτηριστικά των μεγάλων δεδομένων

Τα μεγάλα δεδομένα επηρεάζουν πολλούς τομείς της μοντέρνας οικονομίας, και είναι η βασική είσοδος των μοντέλων μηχανικής μάθησης

2.2. Εξόρυξη Δεδομένων

Η εξόρυξη δεδομένων περιγράφεται ως μια διαδικασία εξαγωγής χρήσιμών πληροφοριών από μια μεγαλύτερη συλλογή ακατέργαστων δεδομένων με τη χρήση των αρχών της στατιστικής, της τεχνητής νοημοσύνης, μηχανικής μάθησης και μεθόδους αναγνώρισης προτύπων[2][3]. Στόχος της εξόρυξης δεδομένων είναι η πληροφορία που θα εξαχθεί και τα πρότυπα που θα προκύψουν να έχουν δομή κατανοητή προς τον άνθρωπο έτσι ώστε να τον βοηθήσουν να πάρει τις κατάλληλες αποφάσεις [4] Η Εξόρυξη Δεδομένων θεωρείται μια συστηματική και επαναληπτική διαδικασία ανακάλυψης γνώσης, στην οποία οι αυτοματοποιημένες μέθοδοι αναγνώρισης προτύπων συνδυάζονται με τις ειδικές γνώσεις του αναλυτή. Αυτή η διαδικασία ονομάζεται διαδικασία Ανακάλυψης Γνώσης σε Βάσεις Δεδομένων - Knowledge Discovery in Databases (KDD).[5]

2.3. Μηχανική Μάθηση

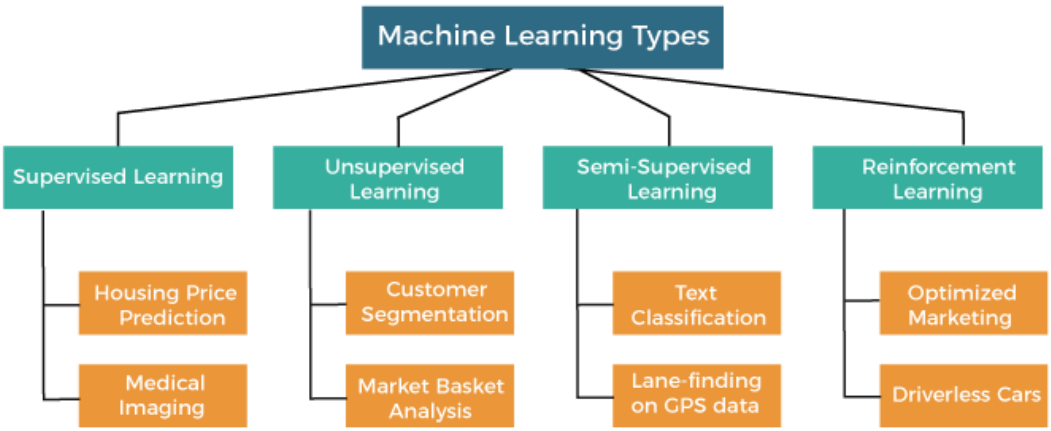
Ο όρος Μηχανική Μάθηση επινοήθηκε από τον Αμερικανό επιστήμονα υπολογιστών Arthur Samuel το 1959 και ορίζεται ως το πεδίο μελέτης που επιτρέπει στους υπολογιστές να μαθαίνουν χωρίς να εχουν ρητά προγραμματιστεί .

Οι αλγόριθμοι μηχανικής μάθησης χρησιμοποιούν υπολογιστικές μεθόδους για να ''μάθουν'' και να εξάγουν πληροφορίες απευθείας από δεδομένα, χωρίς να βασίζονται σε μια προκαθορισμένη εξίσωση ως μοντέλο. Οι αλγόριθμοι προσαρμοστικά βελτιώνουν την απόδοσή τους καθώς αυξάνεται ο αριθμός των δειγμάτων που είναι διαθέσιμα για εκμάθηση. Η βαθιά μάθηση είναι μια εξειδικευμένη μορφή μηχανικής μάθησης.[31] Γενικά, είναι μια μέθοδος που μπορεί να διαχειριστεί διάφορες εργασίες, αναλύοντας και διερευνώντας δεδομένα[6].

Παραδείγματα εφαρμογών μηχανικής εκμάθησης αποτελούν η ανίχνευση ανεπιθύμητων μηνυμάτων ηλεκτρονικού ταχυδρομείου, ηλεκτρονικές απάτες, προβλέψεις αξίας μετοχών, έξυπνοι προσωπικοί βοηθοί, προτάσεις νέων προϊόντων, αυτοοδηγούμενα οχήματα, ανάλυση συναισθημάτων κ.λπ.

Βασιζόμενοι στις μεθόδους και στον τρόπο μάθησης, η μηχανική μάθηση μπορει να χωριστεί σε τέσσερις κυρίως τύπους, οι οποίοι είναι:

1. Supervised Learning - Εποπτευόμενη Μάθηση
2. Unsupervised Learning - Μη εποπτευόμενη Μάθηση
3. Semi-Supervised Learning - Ημι-Εποπτευόμενη Μάθηση
4. Reinforcement Learning - Ενισχυτική Μάθηση



Εικόνα 3: Κατηγορίες μηχανικής μάθησης [32]

1. Εποπτευόμενη Μάθηση

Η Εποπτευόμενη Μηχανική Μάθηση βασίζεται στην εποπτεία, ο αλγόριθμος διδάσκεται με παραδείγματα. Ο χειριστής παρέχει στον αλγόριθμο μηχανικής μάθησης ένα γνωστό σύνολο δεδομένων εισόδου και γνωστές αποκρίσεις - τα δεδομένα εξόδου, ο αλγόριθμος εκπαιδεύει ένα μοντέλο, προσδιορίζοντας μοτίβα, για να δημιουργήσει λογικές προβλέψεις για απόκριση σε νέα δεδομένα. Τα δεδομένα με ετικέτα προσδιορίζουν ότι ορισμένες από τις εισόδους έχουν ήδη αντιστοιχιστεί στην έξοδο. Πρώτα, εκπαιδεύουμε το μηχάνημα με την είσοδο και την αντίστοιχη έξοδο και, στη συνέχεια, ζητάμε από το μηχάνημα να προβλέψει την έξοδο χρησιμοποιώντας το σύνολο δεδομένων δοκιμής, αυτή η διαδικασία συνεχίζεται έως ότου ο αλγόριθμος επιτύχει υψηλό επίπεδο ακρίβειας/απόδοσης. Κατηγορίες προβλημάτων που χρησιμοποιούν εποπτευόμενη μάθηση, είναι ταξινόμηση (classification) και παλινδρόμηση (regression) και πρόβλεψη (forecasting).

Στα προβλήματα ταξινόμησης, το πρόγραμμα μηχανικής μάθησης προσπαθεί να εξάγει ένα συμπέρασμα από τις παρατηρούμενες τιμές και να προσδιορίσει σε ποια κατηγορία ανήκουν οι νέες παρατηρήσεις. Για παράδειγμα, όταν θέλουμε να χαρακτηρίσουμε μηνύματα ηλεκτρονικού ταχυδρομείου ως "ανεπιθύμητα" ή "μη ανεπιθύμητα", ο αλγόριθμος εξετάζει τα υπάρχοντα δεδομένα παρατήρησης και να φιλτράρει τα νέα μηνύματα αναλόγως.

Στα προβλήματα παλινδρόμησης, ο αλγόριθμος μηχανικής εκμάθησης προσπαθεί να εκτιμήσει και να κατανοήσει τις σχέσεις μεταξύ των μεταβλητών. Η ανάλυση παλινδρόμησης εστιάζει σε μια εξαρτημένη μεταβλητή και σε μια σειρά από άλλες μεταβαλλόμενες μεταβλητές, καθιστώντας την ιδιαίτερα χρήσιμη για την πρόβλεψη. Όπως για παράδειγμα την τιμή ενός σπιτιού, βασιζόμενος στα τετραγωνικά του σπιτιού, στο πόσο κοντά βρίσκεται σε ένα σχολείο και άλλους παράγοντες.

Στα προβλήματα πρόβλεψης, ο αλγόριθμος προσπαθεί να προβλέψει το μέλλον βασισμένος σε δεδομένα του παρελθόντος και του παρόντος. Χρησιμοποιείται συνήθως για την ανάλυση των τάσεων.

2. Μη Εποπτευόμενη Μάθηση

Στην μη εποπτευόμενη μάθηση, ο αλγόριθμος μηχανικής μάθησης μελετά δεδομένα για να προσδιορίσει μοτίβα. Δεν υπάρχει σωστή απάντηση, ούτε καθοδήγηση από κάποιον ανθρώπινο παράγοντα. Αντ'αυτού ο αλγόριθμος καθορίζει τους συσχετισμούς και τις σχέσεις αναλύοντας τα διαθέσιμα δεδομένα. Σε μια διαδικασία εκμάθησης χωρίς επίβλεψη, ο αλγόριθμος μηχανικής μάθησης προσπαθεί να ερμηνεύσει μεγάλα σύνολα δεδομένων και να αντιμετωπίσει αυτά τα δεδομένα ανάλογα. Ο αλγόριθμος επιχειρεί να οργανώσει αυτά τα δεδομένα με κάποιο τρόπο ώστε να περιγράψει τη δομή του. Αυτό μπορεί να σημαίνει ομαδοποίηση των δεδομένων σε συσχετίσεις ή την τακτοποίησή τους με βάση κάποια κοινά χαρακτηριστικά. Καθώς αξιολογεί περισσότερα δεδομένα, η ικανότητά του να λαμβάνει αποφάσεις για αυτά, σταδιακά βελτιώνεται και γίνεται πιο εκλεπτυσμένη. Η μη εποπτευόμενη μάθηση, περιλαμβάνει την ομαδοποίηση (clustering) και τη μείωση διαστάσεων (dimension reduction).

Η ομαδοποίηση περιλαμβάνει την ομαδοποίηση συνόλων παρόμοιων δεδομένων (με βάση καθορισμένα κριτήρια). Είναι χρήσιμο για την τμηματοποίηση δεδομένων σε πολλές ομάδες και την εκτέλεση αναλύσεων σε κάθε σύνολο δεδομένων για την εύρεση μοτίβων.

Η μείωση της διαστάσεων μειώνει τον αριθμό των μεταβλητών που εξετάζονται, για την εύρεση των ακριβών πληροφοριών που απαιτούνται.

Κάποια παραδείγματα μη εποπτευόμενης μάθησης είναι: ανάλυση καλαθιού αγοράς (market basket analysis), ώστε να προτείνει προϊόντα με βάσει προηγούμενες παραγγελίες των πελατών, ομαδοποίηση πελατών, αναγνωρίζοντας και δημιουργώντας ομάδες με κοινά χαρακτηριστικά, σημασιολογική ομαδοποίηση (semantic clustering), ομαδοποιώντας σημασιολογικά ταυτόσημα ερωτήματα αναζήτησης - λέξεις, φράσεις και προτάσεις - σε ομάδες με βάση το νόημα.

3. Ημι-Εποπτευόμενη Μάθηση

Η ημι-εποπτευόμενη μάθηση είναι παρόμοια με την εποπτευόμενη μάθηση, αλλά αντ' αυτού χρησιμοποιεί τόσο δεδομένα με ετικέτα όσο και χωρίς ετικέτα δεδομένα, Συνήθως τα δεδομένα με ετικέτα είναι αρκετά λιγότερα σε σχέση με τα δεδομένα χωρίς ετικέτα. Τα δεδομένα με ετικέτα περιέχουν ουσιαστικά πληροφορίες

σημαντικές, έτσι ώστε ο αλγόριθμος να μπορεί να κατανοήσει τα δεδομένα, ενώ τα δεδομένα χωρίς ετικέτα στερούνται πληροφορίας. Ωστόσο, η προσθήκη ετικετών είναι συχνά μια δύσκολη, δαπανηρή ή ακόμη και χρονοβόρα διαδικασία, καθώς απαιτείται έμπειρο και εξειδικευμένο προσωπικό. Αντίθετα, τα δεδομένα χωρίς ετικέτα μπορεί να είναι σχετικά εύκολο να συλλεχθούν, αλλά υπάρχουν λίγοι τρόποι χρήσης τους. Η ημι-εποπτευόμενη μάθηση αντιμετωπίζει αυτό το πρόβλημα χρησιμοποιώντας μεγάλο όγκο δεδομένων χωρίς ετικέτα, μαζί με δεδομένα με ετικέτα, για τη δημιουργία καλύτερων ταξινομητών. Επειδή η ημι-εποπτευόμενη μάθηση απαιτεί λιγότερη ανθρώπινη προσπάθεια και δίνει μεγαλύτερη ακρίβεια, είναι πολύ ενδιαφέρον τόσο στη θεωρία όσο και στην πράξη.[7] Ένα συνηθισμένο παράδειγμα εφαρμογής ημι-εποπτευόμενης μάθησης είναι ένας ταξινομητής εγγράφων κειμένου. Σε αυτή την περίπτωση ημι-εποπτευόμενη μάθηση είναι ιδανική, επειδή θα ήταν σχεδόν αδύνατο να βρεθεί μεγάλος αριθμός εγγράφων κειμένου με ετικέτα.

4. **Ενισχυτική Μάθηση**

Η ενισχυτική μάθηση εστιάζει σε διεργασίες μάθησης, όπου ένας αλγόριθμος μηχανικής μάθησης εφοδιάζεται με ένα σύνολο ενεργειών, παραμέτρων και τελικών τιμών. Καθορίζοντας τους κανόνες, ο αλγόριθμος μηχανικής εκμάθησης προσπαθεί στη συνέχεια να εξερευνήσει διαφορετικές επιλογές και δυνατότητες, παρακολουθώντας και αξιολογώντας κάθε αποτέλεσμα για να προσδιορίσει ποιο είναι το βέλτιστο, μεγιστοποιώντας την ανταμοιβή ή ελαχιστοποιώντας το ρίσκο. Η ενισχυτική μάθηση διδάσκει τον αλγόριθμο μέσω των δοκιμών και σφαλμάτων. Μαθαίνει από προηγούμενες εμπειρίες και αρχίζει να προσαρμόζει την προσέγγισή του, για να επιτύχει το καλύτερο δυνατό αποτέλεσμα. Παραδείγματα στα οποία χρησιμοποιείται ενισχυτική μάθηση είναι τα αυτοδηγούμενα οχήματα, η ρομποτική, ακόμα και το ψηφιακό μάρκετινγκ.

2.4. Συνδιασμός Μοντέλων (Ensemble Techniques)

Ένα μοντέλο πρόβλεψης που προκύπτει από μια διαδικασία Μηχανικής Μάθησης δεν είναι πάντα τέλειο. Η απόδοση του κυμαίνεται ανάλογα με τον αριθμό, την ποιότητα των δεδομένων και την καταλληλότητα του αλγορίθμου που

χρησιμοποιήθηκε σε σχέση με τα δεδομένα. Ένας τρόπος για να βελτιώσουμε την απόδοση του είναι να συνδυάσουμε τις αποφάσεις πολλών διαφορετικών μοντέλων πρόβλεψης. Οι πιο διαδεδομένες μέθοδοι είναι:

A) Bagging, B) Boosting, C) Stacking (για μοντέλα ταξινόμησης και παλινδρόμησης)

D) Error-Correcting Output Codes (για μοντέλα ταξινόμησης με πάνω από 2 κατηγορίες)

Οι μέθοδοι Bagging, Boosting και Error-Correcting Output Codes συνδυάζουν ομογενή μοντέλα. Μοντέλα, που προκύπτουν από την εκπαίδευση ίδιων αλγορίθμων σε διαφορετικά σύνολα δεδομένων ή από μοντέλα που προκύπτουν με διαφορετική παραμετροποίηση του ίδιου αλγορίθμου. Στη μέθοδο Stacking συνδυάζονται και ετερογενή μοντέλα. Οι παραπάνω μέθοδοι σχεδόν πάντα αυξάνουν την απόδοση πρόβλεψης σε σχέση με αυτήν που θα είχε ένα και μόνο μοντέλο. Ωστόσο, το συνδυαστικό μοντέλο είναι πολύπλοκο και είναι δύσκολο να αναλυθούν οι παράγοντες που συνεισφέρουν στην τελική συνδυασμένη απόφαση.

2.4.1. Bagging (Bootstrap Aggregating)

Η μέθοδος Bagging εφαρμόζεται σε αλγορίθμους μάθησης οι οποίοι είναι ασταθείς, παρουσιάζουν δηλαδή μεγάλη διακύμανση (variance). Μικρές αλλαγές στα δεδομένα εκπαίδευσης, προκαλούν αλλαγές στο μοντέλο με αποτέλεσμα να βγάζει διαφορετικές αποφάσεις για ορισμένες περιπτώσεις. Η μέθοδος Bagging είναι σχεδιασμένη για να βελτιώνει τη σταθερότητα και την ακρίβεια των αλγορίθμων μηχανικής μάθησης που χρησιμοποιούνται για ταξινόμηση και παλινδρόμηση. Μειώνει τη διακύμανση και βοηθά στο να αποφευχθεί το overfitting. Συνήθως εφαρμόζεται σε μεθόδους δέντρων αποφάσεων. Είναι μια ειδική περίπτωση του μοντέλου με προσέγγιση μέσου όρου.

Αρχικά δημιουργεί (τυχαία) πολλά διαφορετικά σύνολα δεδομένων από το αρχικό μέσω "Δειγματοληψίας με Αντικατάσταση" (sampling with replacement). Το νέο σύνολο δεδομένων έχει ίδιο αριθμό δεδομένων με το αρχικό, αλλά κάποια δεδομένα έχουν επαναληφθεί ενώ κάποια δεν έχουν συμπεριληφθεί καθόλου. Στη συνέχεια εφαρμόζεται ο ασταθής αλγόριθμος μάθησης σε όλα τα νέα σύνολα δεδομένων και παράγονται αντίστοιχα μοντέλα πρόβλεψης. Για τη διαδικασία πρόβλεψης λαμβάνονται οι αποφάσεις όλων των μοντέλων. Η τελική τιμή είναι είτε η κατηγορία που συγκεντρώνει τις περισσότερες αποφάσεις μοντέλων (ψήφους) είτε ο μέσος

όρος των αριθμητικών προβλέψεων των διαφορετικών μοντέλων. Για παράδειγμα, ο αλγόριθμος τυχαίων δασών (random forest) συνδυάζει τυχαία δέντρα αποφάσεων (random trees) με bagging για να επιτευχθεί υψηλή ακρίβεια στην ταξινόμηση.

2.4.2. Boosting

Στην μηχανική μάθηση, ο μέθοδος boosting είναι ένας μετα-αλγόριθμος που μετατρέπει αδύναμους αλγορίθμους σε ισχυρούς ελαττώνοντας την προκατάληψη (bias) και τη διακύμανση (variance). Στηρίζεται στην ανάθεση βαρών (θετικών αριθμών) στα δεδομένα, έτσι ώστε δεδομένα με μεγαλύτερο βάρος να παίζουν σημαντικότερο ρόλο στη μάθηση. Αν ο αλγόριθμος μπορεί να χειριστεί βάρη στα δεδομένα, τότε δεν υπάρχει πρόβλημα στην εφαρμογή του Boosting. Αν ο αλγόριθμος δεν μπορεί να χειριστεί βάρη, τότε μετατρέπουμε ένα σύνολο δεδομένων με βάρη σε ένα χωρίς βάρη μέσω "Δειγματοληψίας με Αντικατάσταση". Η πιθανότητα επιλογής ενός δεδομένου κατά τη δειγματοληψία είναι ανάλογη του βάρους του. Έτσι δεδομένα με μεγαλύτερο βάρος εμφανίζονται περισσότερες φορές, ενώ δεδομένα με μικρότερο βάρος μπορεί να μην εμφανιστούν ακόμη και καθόλου. Τα βάρη αλλάζουν τον τρόπο υπολογισμού της απόδοσης ενός αλγορίθμου. Χωρίς βάρη, το ποσοστό λάθους είναι ο αριθμός των δεδομένων ελέγχου που ταξινομούνται λάθος προς το συνολικό αριθμό των δεδομένων ελέγχου. Με βάρη, είναι το άθροισμα των βαρών των δεδομένων ελέγχου που ταξινομούνται λάθος προς το συνολικό άθροισμα των βαρών όλων των δεδομένων ελέγχου.

Η μέθοδος Boosting είναι μια επαναληπτική διαδικασία. Τα διαφορετικά μοντέλα κατασκευάζονται το ένα μετά το άλλο. Η απόδοση του προηγούμενου μοντέλου επηρεάζει την κατασκευή του επόμενου. Συγκεκριμένα το Boosting προσπαθεί να κατασκευάσει το επόμενο μοντέλο έτσι ώστε να μην κάνει τα ίδια λάθη με αυτά που έκανε το προηγούμενο και αυτό επιτυγχάνεται χρησιμοποιώντας τα βάρη. Η διαδικάσία που ακολουθείται είναι η εξής. Το πρώτο μοντέλο παράγεται από το αρχικό σύνολο δεδομένων. Αν ο αλγόριθμος χειρίζεται βάρη τότε θέτουμε σε όλα τα δεδομένα του αρχικού συνόλου ίσα βάρη. Έπειτα τα δεδομένα ταξινομούνται από το μοντέλο. Αν η απόφαση του μοντέλου για κάποιο δεδομένο είναι λάθος τότε το βάρος του αυξάνεται ενώ αν είναι σωστή μειώνεται. Η διαδικασία επαναλαμβάνεται για τη μάθηση του επόμενου μοντέλου. Υπενθυμίζεται ότι αν ο αλγόριθμος δεν χειρίζεται βάρη, θα πρέπει το σύνολο δεδομένων εκπαίδευσης να προκύψει από το

προηγούμενο με "Δειγματοληψία με Αντικατάσταση". Με τη διαδικασία του Boosting, παράγονται επαναληπτικά μοντέλα πρόβλεψης που επικεντρώνονται στα δύσκολα δεδομένα τα οποία δεν ταξινομούνται σωστά από το αρχικό μοντέλο.

Για δεδομένα που ταξινομούνται λάθος το βάρος παραμένει όσο ήταν αρχικά, ενώ για αυτά που ταξινομούνται σωστά το βάρος μειώνεται αντιστρόφως ανάλογα με το ποσοστό λαθών e του ταξινομητή στα δεδομένα.

$$weight = weght \frac{e}{1-e}$$

Στη συνέχεια τα βάρη ομογενοποιούνται έτσι ώστε το άθροισμα τους να παραμείνει όσο και πριν. Κάθε βάρος διαιρείται με το άθροισμα των νέων βαρών και πολλαπλασιάζεται με το άθροισμα των παλιών βαρών. Έτσι αυτόματα αυξάνεται το βάρος των δεδομένων που ταξινομούνται λάθος και μειώνεται αυτών που ταξινομούνται σωστά. Η διαδικασία επαναλαμβάνεται μέχρι το ποσοστό λάθους του τρέχοντος μοντέλου γίνει είτε 0 είτε μεγαλύτερο ή ίσο του 0.5. Στη δεύτερη περίπτωση, το τελευταίο αυτό μοντέλο διαγράφεται. Για τη ταξινόμηση άγνωστων δεδομένων, συνδυάζονται οι αποφάσεις όλων των μοντέλων μέσω ψηφοφορίας με βάρη. Το βάρος της απόφασης κάθε μοντέλου είναι αντίστοιχο του ποσοστού λαθών e στα δεδομένα από τα οποία εκπαιδεύτηκε:

$$weight = -log \frac{e}{1-e}$$

Η ποιο κοινή εφαρμογή της μεθόδου boosting είναι ο αλγόριθμος AdaBoost, παρόλο που κάποιοι καινούριοι αλγόριθμοι έχουν πετύχει καλύτερα αποτελέσματα. Διατυπώθηκε από τους Yoav Freund και Robert Schapire το 1995, οι οποίοι κέρδισαν το βραβείο Gödel το 2003 για το έργο τους.[8] Ο AdaBoost (με δέντρα αποφάσεων ως αδύναμοι μαθητές) αναφέρεται συχνά ως ο καλύτερος out-of-the-box ταξινομητής. Είναι αρκετά δημοφιλής και ίσως ο πιο ιστορικά σημαντικός αλγόριθμός, που μπορούσε να προσαρμοστεί στους αδύναμους μαθητές (weak learners). Ωστόσο υπάρχουν πολλοί πιο πρόσφατοι αλγόριθμοι όπως LPBoost, TotalBoost, BrownBoost, XGBoost, MadaBoost, LogitBoost LightGBM και αρκετοί ακόμα.

Οι διαφορές που εντοπίζονται στους αλγορίθμους AdaBoost, GBoost και XGBoost, σχετίζονται με τη συνάρτηση κόστους που χρησιμοποιεί ο καθένας από αυτούς. Παρατηρήθηκε θεωρητικά ότι ο AdaBoost μπορεί να γενικευτεί σε έναν αλγόριθμο (Gradient Boost - GBoost) που θα μπορεί να χρησιμοποιεί διαφορετικές συναρτήσεις

κόστους για τη βελτιστοποίηση των μοντέλων που συνδυάζονται. Η συνάρτηση κόστους του AdaBoost είναι αυτή που διορθώνει τα επιμέρους σφάλματα του κάθε μοντέλου (με προσθετικό τρόπο). Ο GBoost ελαχιστοποιεί οποιαδήποτε συνάρτηση κόστους μέσω gradient descent, το μειονέκτημα όμως είναι ότι στηρίζεται σε πολλές υπερ-παραμέτρους. .Ο XGBoost στηρίζεται στον GBoost, καθώς η λειτουργικότητά του είναι παρόμοια, αλλά με σημαντικές βελτιώσεις. Κύριες επέκτασεις είναι η εφαρμογή μεθόδου ομαλοποίησης (regularization) καθώς και η υποστήριξη παράλληλης επεξεργασίας, διαχείριση ελλιπών τιμών, κλάδεμα δέντρου, ενσωματωμένη μέθοδο cross-validation και άλλες.

2.4.3. Stacking

Σε αντίθεση με τις μεθόδους Bagging και Boosting, η μέθοδος Stacking χρησιμοποιείται συνήθως για τον συνδυασμό μοντέλων πρόβλεψης που προέκυψαν από διαφορετικούς αλγορίθμους μάθησης. Το Stacking στηρίζεται στη παραγωγή ενός μετά-μοντέλου (μοντέλο επιπέδου 1) με δεδομένα εκπαίδευσης τις αποφάσεις ενός συνόλου μοντέλων (μοντέλα επιπέδου 0). Το μετά-μοντέλο αυτό θα προσπαθήσει να μάθει ποιά μοντέλα είναι τα πιο αξιόπιστα από αυτά που συμμετέχουν, έτσι ώστε να δίνει τη σωστή απόφαση κάθε φορά. Η μέθοδος Stacking αποδίδει συνήθως καλύτερα από οποιονδήποτε μεμονωμένο αλγόριθμο. Δεν χρησιμοποιείται τόσο ευρέως όσο οι bagging και boosting μέθοδοι, ωστόσο έχει χρησιμοποιηθεί με επιτυχία και σε προβλήματα εποπτευόμενης μάθησης (παλινδρόμησης, ταξινόμησης) και σε προβλήματα χωρίς επίβλεψη (εκτίμηση πυκνότητας). Πρόσφατα στο διαγωνισμό του Netflix οι δύο κορυφαίες υλοποιήσεις, χρησιμοποίησαν blending, που θεωρείται μια μέθοδος stacking.

Συνδιασμός Μεθόδων (Ensemble methods) στην Python

<u>Bagging</u>

Python (sklearn): sklearn.ensemble.BaggingClassifier(base_estimator, n_estimators, max_samples, bootstrap={True, False}, bootstrap_features={True, False}, oob_score={True, False})

<u>Stacking</u>

Python (sklearn): sklearn.ensemble.StackingClassifier(estimators, final_estimator, cv (cross-validation splitting strategy))

AdaBoost classifier

Python (sklearn): sklearn.ensemble.AdaBoostClassifier(base_estimator, n_estimators, learning_rate, algorithm={SAMME, SAMME.R})

Gboost ensemble classifier

Python (sklearn): sklearn.ensemble.GradientBoostingClassifier(loss={deviance, exponential}, learning_rate, n_estimators, subsample, criterion={MAE, MSE, Friedman MSE}, min_samples_split, min_samples_leaf, max_depth, max_features)

XGBoost classifier

Python (XGBoost): xgboost.XGBClassifier(eta (learning rate), booster = {dart, gblinear, gbtree}, objective={reg:logistic, binary:logistic, multi:softmax, …}, eval_metric={RMSE, MAE, logloss, error, …})

2.4.4. Error-Correcting Output Codes (ECOC)

Η “ECOC” είναι μια μέθοδος για τη μετατροπή ενός προβλήματος ταξινόμησης με παραπάνω από δύο κατηγορίες (multiclass problem) σε πολλά προβλήματα με δύο κατηγορίες. Αυτό καθιστά επιλύσιμα τα multiclass problems από αλγορίθμους μάθησης που χειρίζονται μόνο δύο κατηγορίες (π.χ. Support Vector Machines). Χρησιμοποιείται και για το συνδυασμό πολλαπλών μοντέλων αλγορίθμων που χειρίζονται multiclass προβλήματα, γιατί αυξάνει την απόδοση πρόβλεψης.

2.5. Χρονοσειρές και Χαρακτηριστικά Χρονοσειρών

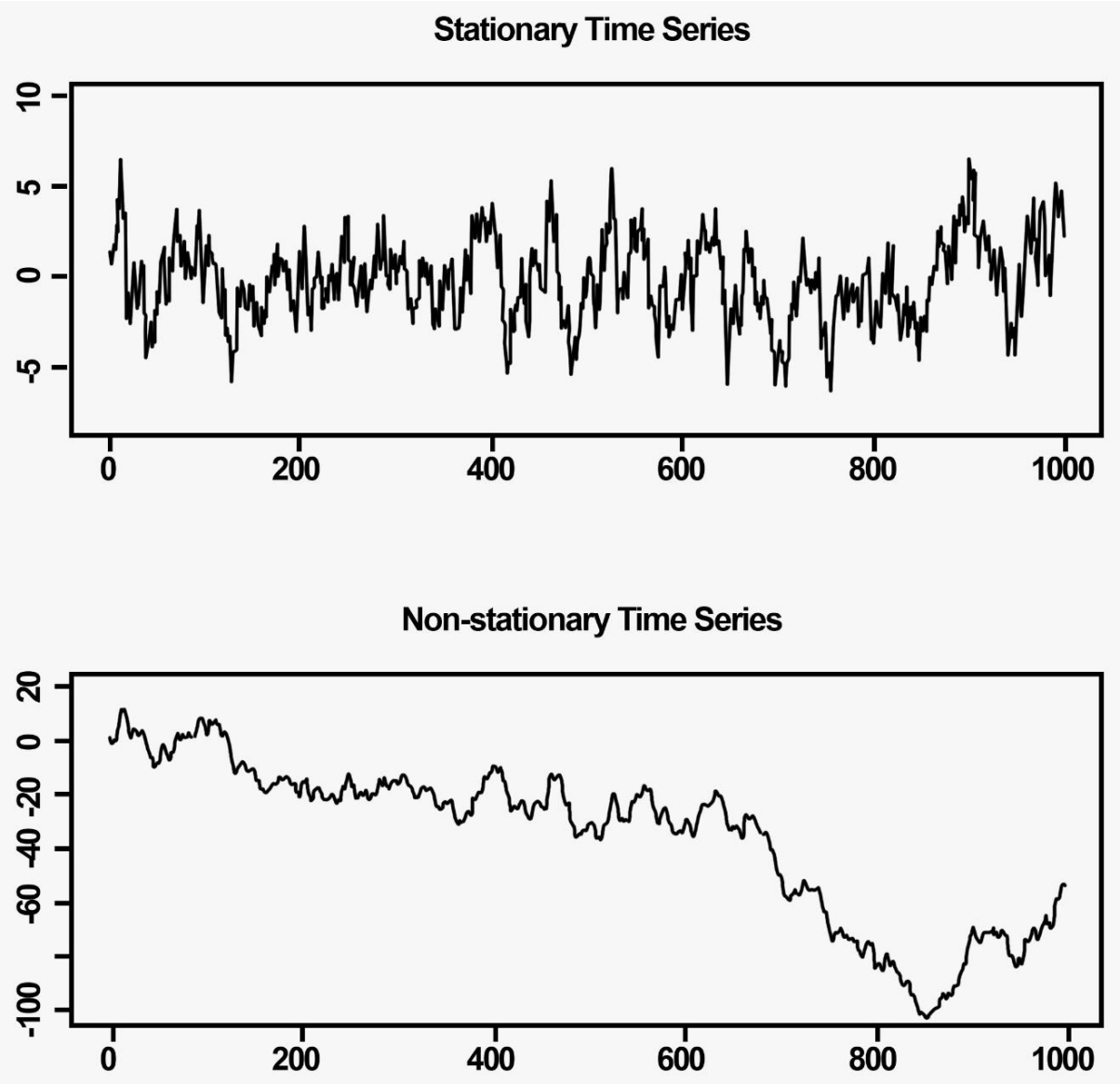
Χρονοσειρά είναι μια ακολουθία παρατηρήσεων που καταγράφονται με την πάροδο του χρόνου. Στις εφαρμογές πρόβλεψης, οι παρατηρήσεις συνήθως καταγράφονται με τακτική συχνότητα, όπως μηνιαία, τριμηνιαία ή ετήσια, σε μερικές περιπτώσεις εβδομαδιαία, ημερησίως ή ωριαία (μελέτη οδικής κυκλοφορίας, τηλεφωνική κίνηση). Η ανάλυση χρονοσειρών αποτελείται από μεθόδους που μας βοηθούν να τις κατανοήσουμε σε βάθος, ώστε να είμαστε σε θέση να κάνουμε προβλέψεις.[9]

Τα τέσσερα χαρακτηριστικά, καθένα από τα οποία εκφράζει μια ιδιαίτερη πτυχή της κίνησης των τιμών της χρονοσειράς, είναι:

- Σταθερότητα

Μια χρονοσειρά χαρακτηρίζεται από σταθερότητα, όταν η πιθανολογική συμπεριφορά κάθε συλλογής τιμών είναι πανομοιότυπη με εκείνη του συνόλου με

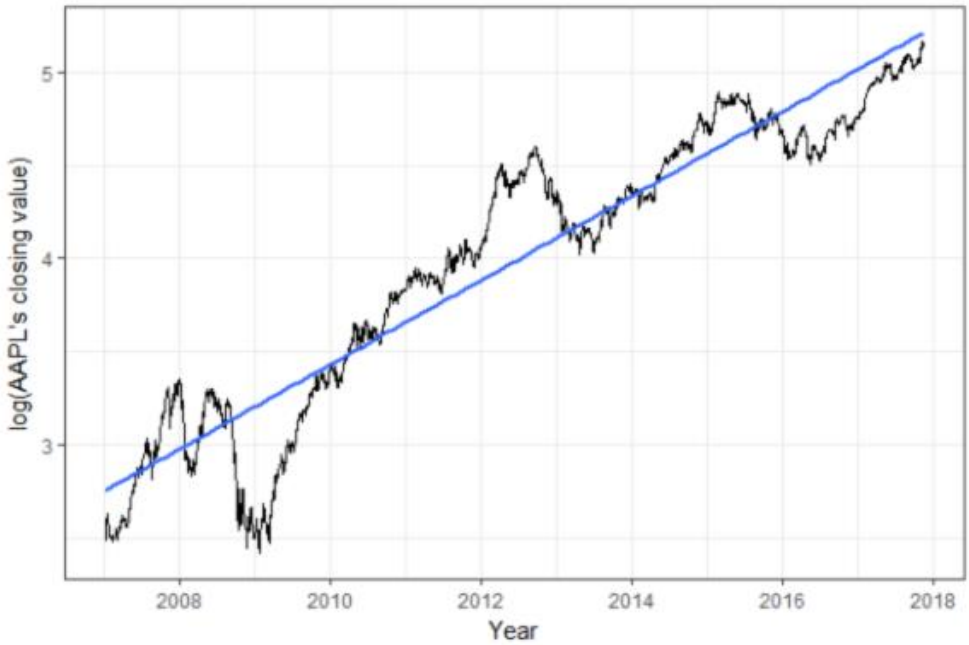
μετατόπιση χρόνου. Δεν σημαίνει ότι η σειρά δεν αλλάζει με την πάροδο του χρόνου, απλά ο τρόπος που αλλάζει, δεν αλλάζει τη ίδια τη χρονοσειρά με την πάροδο του χρόνου. Το αλγεβρικό ισοδύναμο είναι επομένως μια γραμμική συνάρτηση, ίσως, και όχι μια σταθερή, η τιμή μιας γραμμικής συνάρτησης αλλάζει καθώς αυξάνεται το $\boldsymbol{x}$, αλλά ο τρόπος που αλλάζει παραμένει σταθερός — έχει σταθερή κλίση, μια τιμή που καταγράφει αυτόν τον ρυθμό αλλαγής. Στα παρακάτω γραφήματα φαίνεται η διαφορά μίας χρονοσειράς με σταθερότητα και μια η οποία δεν παρουσιάζει.



Εικόνα 4: Stationary and Non-stationary Time Series [33]

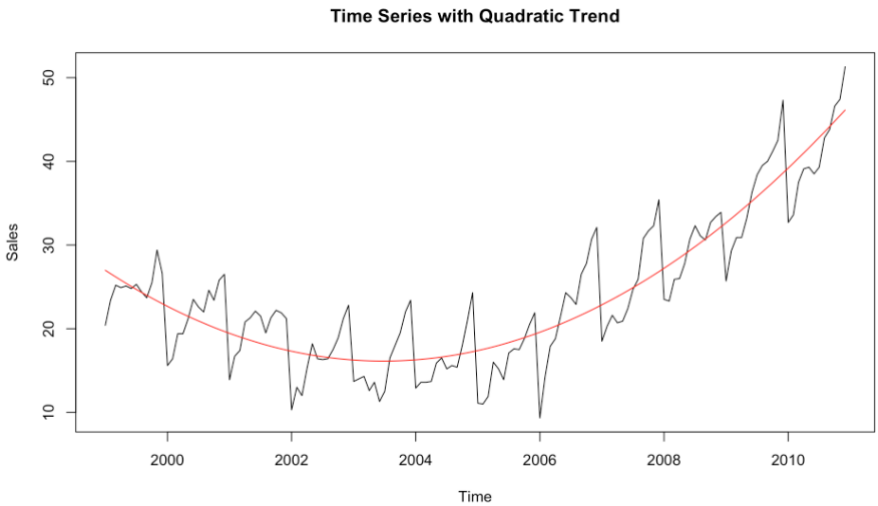
- <u>Τάση</u>

Η τάση περιγράφει την κίνηση κατά μήκος μιας περιόδου. Υπάρχουν αρκετοί τύποι τάσεων που μπορεί να συναντήσουμε στην πράξη. Η πιο δημοφιλής είναι μια γραμμική τάση όπου τα δεδομένα κυμαίνονται γύρω από μια γραμμή. Επιπλέον, μια τετραγωνική τάση, ή μια εκθετική τάση μπορεί να εμφανιστεί σε μια χρονοσειρά όπου η αλλαγή των χρονικών βημάτων αναφέρεται σε μια ταχύτερη αύξηση ή πτώση των παρατηρούμενων τιμών.



Εικόνα 5: Γραμμική Τάση σε Χρονοσειρές

Στο παραπάνω γράφημα φαίνεται μια αύξουσα γραμμική τάση για την τιμή της μετοχής της Apple για το χρονικό διάστημα 2007 - 2018.

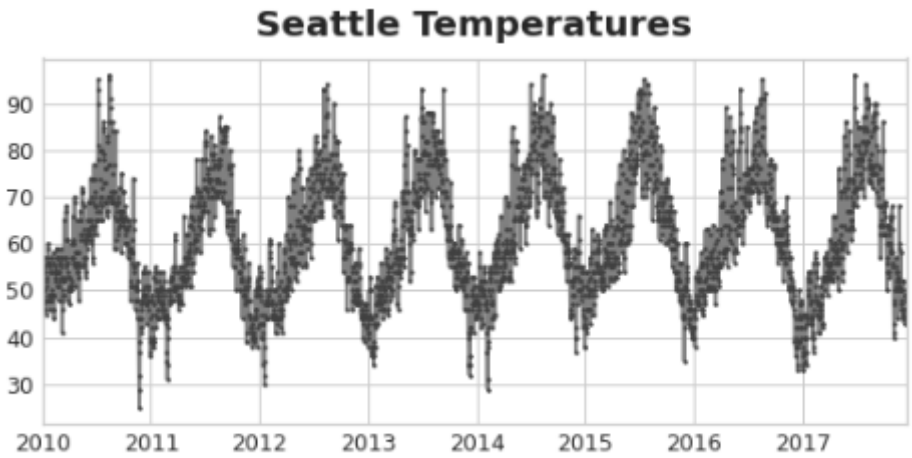


Εικόνα 6: Τετραγωνική Τάση σε Χρονοσειρές

Στο παραπάνω γράφημα φαίνεται η τετραγωνική τάση στις πωλήσεις για το χρονικό διάστημα 1999 – 2011 [34]

- Εποχικότητα

Αντιπροσωπεύει τις εποχικές αλλαγές, καταγράφει ένα μοτίβο μέσα σε ένα διάστημα (συνήθως λιγότερο από ένα χρόνο). Για παράδειγμα οι πωλήσεις μαγιό το καλοκαίρι αυξάνονται και το χειμώνα μειώνονται, αυτό επαναλαμβάνεται για χρόνια. Το παράδειγμα αυτό δείχνει τη διακύμανση των πωλήσεων μεταξύ των πραγματικών ετήσιων εποχών. Θα μπορούσε επίσης να χαρακτηριστεί από μια μικρότερη ή μεγαλύτερη χρονική περίοδο, όπως για παράδειγμα μια εβδομάδα, ένα μήνα, ένα εξάμηνο. Με άλλα λόγια, κάθε φορά που η συμπεριφορά μιας χρονοσειράς επηρεάζεται με περιοδικό τρόπο από το ημερολόγιο, την ονομάζουμε εποχιακή.



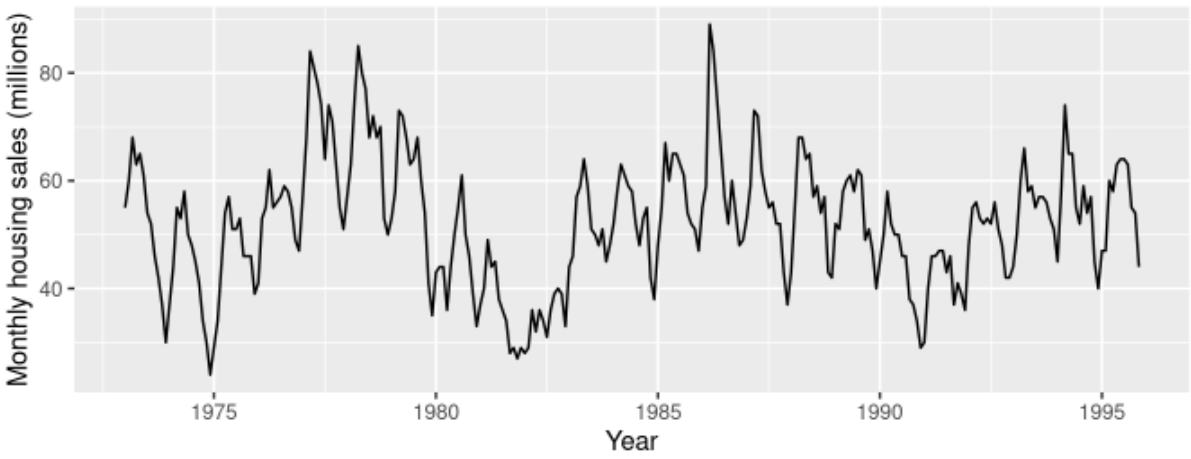
Εικόνα 7:Κυκλικότητα σε Χρονοσειρές [35]

Στο παραπάνω γράφημα φαίνονται οι αλλαγές στη θερμοκρασία για την πόλη του Seattle. Οι αυξομειώσεις της θερμοκρασίας παρατηρούνται μέσα στη χρονική περίοδο ενός έτους.

- Κυκλικότητα

Είναι παρόμοια με την εποχικότητα, αντιστοιχεί όμως σε διακυμάνσεις, αυξήσεις ή μειώσεις, που δεν είναι σταθερής περιόδου, όπως για παράδειγμα η ηφαιστειακή δραστηριότητα. Η διάρκεια αυτών των διακυμάνσεων είναι συνήθως τουλάχιστον 2 χρόνια. Για παράδειγμα οι οικονομικοί κύκλοι που συνήθως διαρκούν αρκετά χρόνια, αλλά η διάρκεια του τρέχοντος κύκλου είναι άγνωστη εκ των προτέρων. Η διαφορά

με την εποχικότητα, είναι ότι η εποχικότητα έχει πάντα το ίδιο μήκος - περίοδο.[10][36]



Εικόνα 8: Εποχικότητα και Κυκλικότητα σε Χρονοσειρές

Στο παραπάνω γράφημα, όπου παρουσιάζονται οι μηνιαίες πωλήσεις νέων μονοκατοικιών που πωλήθηκαν στις ΗΠΑ την περίοδο 1973-1995, φαίνεται η διαφορά της εποχικότητας και τη κυκλικότητας. Υπάρχει έντονη εποχικότητα μέσα σε κάθε χρόνο, καθώς και έντονη κυκλική συμπεριφορά με περίοδο περίπου 6 – 10 ετών.

Μια χρονοσειρά είναι μια ακολουθία σημείων δεδομένων που ταξινομούνται χρονικά. Το γεγονός ότι τα δεδομένα χρονοσειρών είναι ταξινομημένα χρονικά, τους δίνει ένα πλεονέκτημα, λόγω του ότι συχνά εμφανίζουν σειριακή εξάρτηση. Η σειριακή εξάρτηση εμφανίζεται όταν η τιμή ενός σημείου δεδομένων κάθε φορά εξαρτάται στατιστικά από ένα άλλο σημείο δεδομένων σε κάποια άλλη χρονική στιγμή. Ωστόσο, το χαρακτηριστικό αυτό των χρονοσειρών, παραβιάζει μία από τις θεμελιώδεις παραδοχές πολλών στατιστικών αναλύσεων ότι τα δεδομένα είναι στατιστικά ανεξάρτητα.[11]

Επίσης ένα χαρακτηριστικό που συχνά εμφανίζεται στις χρονοσειρές είναι η αυτοσυσχέτιση (autocorrelation) του συνόλου δεδομένων. Η συσχέτιση, αρχικά, μετρά την ισχύ της γραμμικής σχέσης μεταξύ δύο ακολουθιών. Όσο πιο κοντά είναι η συσχέτιση στο 1, τόσο ισχυρότερη είναι η θετική γραμμική σχέση, όσο πιο κοντά είναι στο -1, τόσο ισχυρότερη είναι η αρνητική γραμμική σχέση, και όσο πιο κοντά είναι στο 0, τόσο πιο αδύναμη είναι η γραμμική σχέση. Ο όρος αυτοσυσχέτιση αντίστοιχα, αναφέρεται στον βαθμό ομοιότητας μεταξύ μιας δεδομένης χρονοσειράς,

αλλά με μια εκδοχή του εαυτού της με κάποια καθυστέρηση (lag), σε σχέση ορισμένων διαδοχικών χρονικών διαστημάτων. Με άλλα λόγια, η αυτοσυσχέτιση μετράει τη σχέση μεταξύ της τιμής μιας μεταβλητής και τυχόν προηγούμενων τιμών, των ιστορικών δεδομένων αυτής της μεταβλητής. Σχεδιάζει τη μια σειρά πάνω στην άλλη και καθορίζει τον βαθμό ομοιότητας μεταξύ των δύο. Όπως και με τη συσχέτιση, οι τιμές της αυτοσυσχέτησης κυμαίνονται στο ίδιο εύρος τιμών, από -1 έως και 1. Οι τιμές κοντά στο 1 δηλώνουν ότι υπάρχει ισχυρή θετική αυτοσυσχέτιση, στο -1 ότι υπάρχει ισχυρή αρνητική, και στο 0 ότι δεν υπάρχει. Ουσιαστικά για τον υπολογισμό της αυτοσυσχέτισης ακολουθούμε την ίδια ακριβώς διαδικασία που θα ακολουθούσαμε κατά τον υπολογισμό της συσχέτισης μεταξύ δύο διαφορετικών συνόλων τιμών χρονοσειρών. Η κύρια διαφορά είναι ότι η αυτοσυσχέτιση χρησιμοποιεί την ίδια χρονική σειρά δύο φορές, μια φορά με τις αρχικές της τιμές και έπειτα μια φορά με μερικές διαφορετικές χρονικές περιόδους. Η αυτοσυσχέτιση είναι μια ιδανική μέθοδος για την αποκάλυψη τάσεων και μοτίβων σε δεδομένα χρονοσειρών που διαφορετικά θα έμεναν άγνωστα. Μπορούμε να τη χρησιμοποιήσουμε για να βοηθήσουμε στον εντοπισμό εποχικότητας (seasonality), τάσης (trend) στα δεδομένα χρονοσειρών μας.

2.6. Τεχνικές αξιολόγησης

2.6.1. Cross-Validation

Η μέθοδος Cross-Validation είναι μια στατιστική τεχνική, με την οποία προσπαθούμε να εκτιμήσουμε την ικανότητα γενίκευσης ενός μοντέλου μηχανικής μάθησης. Προσπαθούμε να προβλέψουμε την απόδοσή του, βασισμένοι σε κριτήρια αξιολόγησης. Μπορούμε να χρησιμοποιήσουμε οποιοδήποτε κατάλληλο κριτήριο για την αξιολόγηση, για παράδειγμα μπορούμε να χρησιμοποιήσουμε το κριτήριο ακρίβειας εάν έχουμε πρόβλημα ταξινόμησης ή να χρησιμοποιήσουμε το μέσο τετραγωνικό σφάλμα (Mean Squared Error) εάν αντιμετωπίζουμε πρόβλημα παλινδρόμησης.

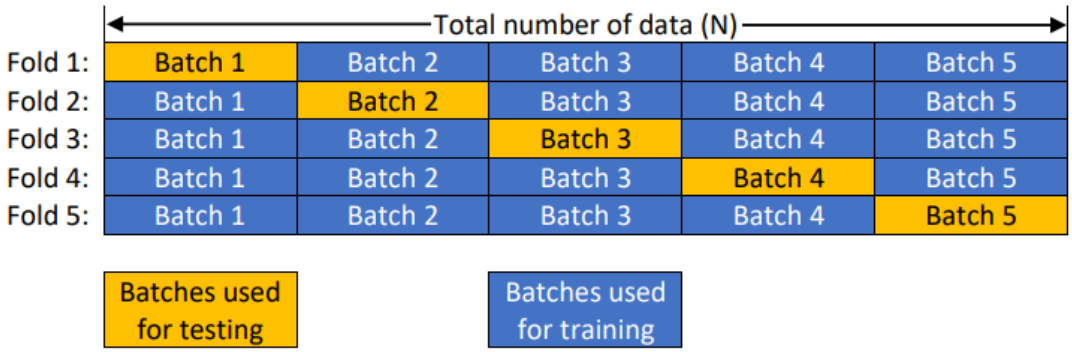
Η βασική προσέγγιση της μεθόδου είναι ο τυχαίος διαχωρισμός των δεδομένων σε δύο μέρη: το train set και το test (ή valuation) set. Το μοντέλο εκπαιδεύεται στο train set και η απόδοσή του κρίνεται στο test (ή valuation) set. Ένα τέτοιο πείραμα

ονομάζεται «fold». Δεδομένου ότι ένα μόνο fold μπορεί να μην παράγει στατιστικά ασφαλή αποτελέσματα, συνήθως εκτελούμε K folds και λαμβάνουμε τη μέση απόδοση, για να εκτιμήσουμε την αναμενόμενη ικανότητα γενίκευσης του μοντέλου. Ο αριθμός K μπορεί να πάρει οποιαδήποτε θετική ακέραια τιμή, π.χ. 5, 10, 100 κ.λπ. Όσο μεγαλύτερη είναι η τιμή του K τόσο ασφαλέστερη είναι η στατιστική μας εκτίμηση.

Σε όλες τις μεθόδους Cross-Validation τα δεδομένα χωρίζονται τυχαία, έτσι ώστε τα σετ εκπαίδευσης σε διαφορετικά folds να μην είναι τα ίδια. Μερικές από τις πιο κοινές μεθόδους Cross-validation είναι:

2.6.2. K-fold Cross-validation

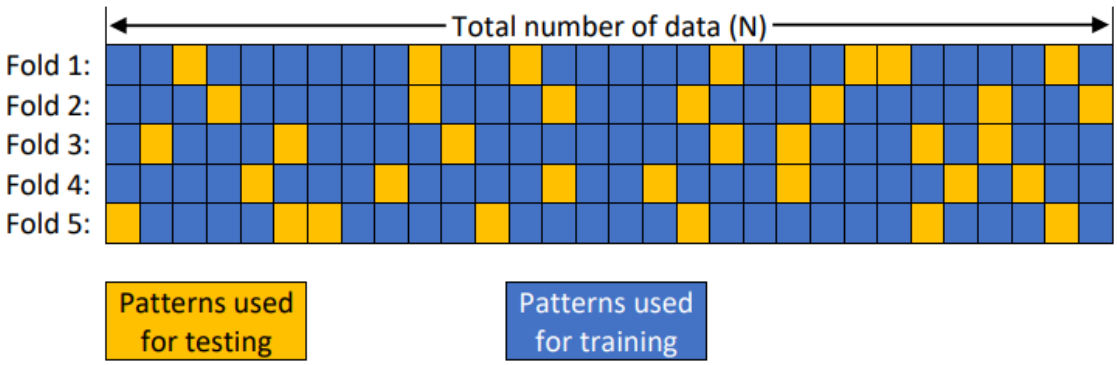
Σε αυτή την προσέγγιση τα δεδομένα χωρίζονται σε K μη τεμνόμενες ομάδες που ονομάζονται παρτίδες και εκτελούνται K-folds. Στο fold i εκπαιδεύουμε τον αλγόριθμο χρησιμοποιώντας όλες τις παρτίδες εκτός από την i-η που χρησιμοποιείται για δοκιμή. Στο παρακάτω σχήμα βλέπουμε πώς χρησιμοποιούνται οι παρτίδες σε κάθε fold.



Εικόνα 9: K-fold Cross-validation

2.6.3. Τυχαία Διάσπαση - Random split

Σε αυτή την προσέγγιση επιλέγουμε τα δείγματα δοκιμής τυχαία χωρίς αντικατάσταση. Είναι πιθανό ότι ένα τμήμα του συνόλου δεδομένων μπορεί να χρησιμοποιηθεί στο σετ δοκιμής (test set) πολλαπλών folds. Φυσικά, είναι επίσης πιθανό να μην εμφανιστεί κάποιο τμήμα του συνόλου δεδομένων στο σετ δοκιμης σε κανένα fold. Το παρακάτω σχήμα απεικονίζει αυτήν την τυχαία διαίρεση.

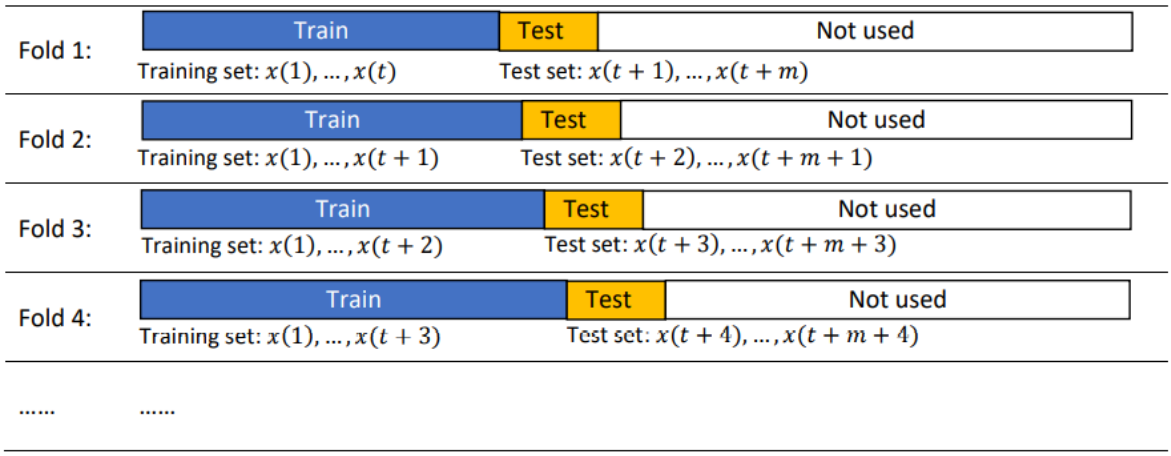


Εικόνα 10: Τυχαία Διάσπαση - Random split

Το πλεονέκτημα αυτής της τεχνικής είναι το γεγονός ότι μπορούμε να επιλέξουμε τον αριθμό των folds ανεξάρτητα από τον αριθμό των τμημάτων στο σύνολο δεδομένων. Πρέπει ωστόσο να επιλέξουμε το ποσοστό των τμημάτων που θα χρησιμοποιηθούν για τη δοκιμή σε κάθε πτυχή. Συνήθως, το σύνολο δεδομένων χωρίζεται 70/30 ή 80/20 (εκπαίδευση έναντι δοκιμής).

2.6.4. Walk Forward validation

Όταν προσπαθούμε να εκτιμήσουμε την ικανότητα γενίκευσης ενός μοντέλου που προβλέπει μια χρονοσειρά, δεν έχει νόημα να το εκπαιδεύσουμε χρησιμοποιώντας δεδομένα που εμφανίζονται χρονολογικά μετέπειτα χρονικά από τα δεδομένα δοκιμής, επειδή δεν έχει νόημα να χρησιμοποιούμε τις τιμές από το μέλλον για να προβλέψουμε τιμές στο παρελθόν. Στις χρονοσειρές υπάρχει μια χρονική εξάρτηση μεταξύ των παρατηρήσεων και πρέπει να διατηρήσουμε αυτή τη σχέση κατά τη διάρκεια των δοκιμών. Η μέθοδος που μπορεί να χρησιμοποιηθεί για cross-validation σε χρονοσειρές είναι cross-validation με κυλιόμενη βάση. Κάθε fold θα χρησιμοποιεί μια συγκεκριμένη χρονική περίοδο για δοκιμή, για παράδειγμα t + 1, … ,t + m. Αυτό σημαίνει ότι η περίοδος εκπαίδευσης θα είναι 1, … ,t και όλα τα δεδομένα μετά από χρόνο t + m θα αγνοηθούν. Στο επόμενο fold τα δεδομένα εκπαίδευσης θα περιλαμβάνουν ένα ακόμη σημείο x(t + 1), το χρονικό παράθυρο για τα δεδομένα δοκιμής θα μετατοπιστεί κατά ένα: t + 2, … ,t + m + 1 και τα δεδομένα μετά το t + m + 1 δεν θα χρησιμοποιηθούν. Στο ακόλουθο γράφημα παρουσιάζεται ο τρόπος με τον οποίο δημιουργούνται τα splits για cross-validation σε χρονοσειρές.[12][13][14][37][38]



Εικόνα 11: Walk-Forward validation

2.7. Μετρικές αξιολόγησης

Συνήθως σε προβλήματα πρόβλεψης χρονοσειρών χρησιμοποιούνται οι παρακάτω μετρικές σφάλματος για την αξιολόγηση των μοντέλων και τη σύγκριση των αποτελεσμάτων πραγματικού χρόνου με τα αποτελέσματα της πρόβλεψης. Οι μετρικές σφάλματος χρησιμοποιούνται για τη βελτιστοποίηση και την εύρεση καλύτερων παραμέτρων στα προβλεπτικά μοντέλα, είτε για την εύρεση ακρίβειας πρόβλεψης που έγινε με τη χρήση χρονοσειρών.

MSE (mean_squared_error) μπορεί να χρησιμοποιηθεί ως συνάρτηση απώλειας.

$$MSE = \frac{\sum(y_i - y_p)^2}{n}$$

Το MSE μετράται σε μονάδες που είναι το τετράγωνο της μεταβλητής στόχου. Λόγω της διατύπωσής του, το MSE, ακριβώς όπως η συνάρτηση τετραγωνικής απώλειας από την οποία προκύπτει, τιμωρεί αποτελεσματικά τα μεγαλύτερα σφάλματα πιο αυστηρά. Είναι πιο αυστηρό θα μπορούσαμε να πούμε σε μεγαλύτερα σφάλματα, σε σχέση με το RMSE.

RMSE (Root Mean Square Error) - (Σφάλμα μέσου τετραγώνου ρίζας)

$$RMSE = \sqrt{MSE}$$

Το RMSE από τον τύπο του, είναι η τετραφωνική ρίζα του MSE (Mean Squared Error), επομένως μετράται στις ίδιες μονάδες με τη μεταβλητή στόχο.

Υπάρχει και η δυνατότητα να χρησιμοποιηθούν επιπλέον μετρικές σφάλματος, οι οποίες όμως δεν χρησιμοποιήθηκαν για την αξιολόγηση των μοντέλων της συγκεκριμένης έρευνας. Για παράδειγμα:

r2_score Για την αξιολόγηση της απόδοσης του μοντέλου.

Το **r2_score** αξιολογεί την απόδοση ενός μοντέλου υπολογίζοντας το ποσοστό της διαφοράς μεταξύ της διακύμανσης γύρω από τον μέσο όρο των δεδομένων και της διακύμανσης γύρω από το προσαρμοσμένο μοντέλο στα δεδομένα που είναι εύκολο να υπολογιστεί και δίνει μια καλή ιδέα για την απόδοση του μοντέλου.
Αν το **r2_score** είναι 0.60, σημαίνει ότι το προσαρμοσμένο μοντέλο έχει 60% μικρότερη διακύμανση από τη μέση διακύμανση.

MAE (Mean Absolute Error) Εκφράζει τη μέση απόλυτη διαφορά, μεταξύ της πραγματικής και της προβλέψιμης τιμής.

$$MAE = \frac{|y_i - y_p|}{n}$$

$\boldsymbol{y_i}$ είναι η πραγματική τιμή, $\boldsymbol{y_p}$ είναι η προβλέψιμη τιμή και n ο αριθμός των παρατηρήσεων.

MAPE (Mean Absolute Percentage Error) Για να αξιολογήσουμε τις προβλέψεις μας.
Εκφράζει το μέσο απόλυτο ποσοστό της διαφοράς της πραγματικής με την προβλέψιμη τιμή.

$$MAPE = \frac{1}{n}\sum \frac{|y_i - y_p|}{|y_i|}$$

Σε αντίθεση με άλλες μετρήσεις αξιολόγησης όπως το MAE και το RMSE, το MAPE είναι ανεξάρτητο από κλίμακα, δηλαδή δεν εξαρτάται από το εύρος τιμών.

Weighted Mean Absolute Percentage Error (WMAPE) - (Σταθμισμένο μέσο απόλυτο ποσοστό σφάλματος)

$$WMAPE = \frac{\sum|y_i - y_p|}{\sum|y_i|}$$

Η μέτρηση του σταθμισμένου μέσου απόλυτου ποσοστού σφάλματος WMAPE χρησιμοποιείται για τις περιπτώσεις όπου η προτεραιότητα κάθε id προϊόντος λαμβάνεται υπόψη μαζί με τον αριθμό των πωλήσεων που πραγματοποιήθηκαν. Οι σταθμισμένοι συντελεστές υπολογίζονται διαιρώντας τη διαφορά μεταξύ των τιμών πρόβλεψης και των πραγματικών τιμών πωλήσεων με τη μέση τιμή τους.

3. Βιβλιογραφική Ανασκόπηση

Λόγω της σημασίας της πρόβλεψης σε διάφορους τομείς, πληθώρα ερευνών έχουν αναπτυχθεί με κεντρικό άξονα την πρόβλεψη πωλήσεων. Η πρόβλεψη πωλήσεων τροφίμων είναι μια εργασία πρόβλεψης χρονοσειρών. Κλασικές στατιστικές τεχνικές, όπως ο αυτοπαλινδρομικός κινητός μέσος όρος (ARMA) και ο αυτοπαλινδρομικός ολοκληρωμένος κινητός μέσος όρος (ARIMA) μπορούν να χρησιμοποιηθούν για την αντιμετώπιση τέτοιων προβλημάτων. Ωστόσο, μια προσέγγιση μηχανικής μάθησης για την αντιμετώπιση μιας πρόβλεψης χρονοσειρών είναι συχνά πιο ισχυρή και ευέλικτη. Επιτρέπει τη χρήση σύγχρονων εποπτευόμενων αλγορίθμων μηχανικής μάθησης, όπως μηχανές υποστήριξης διανυσμάτων για παλινδρόμηση και μοντέλα δέντρων αποφάσεων (Landwehr et al. 2005) [15] (Δέντρα αποφάσεων με συναρτήσεις γραμμικής παλινδρόμησης στα φύλλα). Επίσης είναι ευέλικτη, γιατί συμπεριλαμβάνει πρόσθετες χρήσιμες μεταβλητές εισόδου, εκτός της δεδομένης χρονοσειράς.

Στην περίπτωση των τροφίμων και ιδιαίτερα των προϊόντων μικρής διάρκειας ζωής, όπως το γάλα, απαιτούνται ημερήσιες προβλέψεις πωλήσεων [16] και των φρέσκων τροφίμων, τα οποία παράγονται καθημερινά [18]. Για προϊόντα μεγαλύτερης διάρκειας ζωής, οι εβδομαδιαίες προβλέψεις πωλήσεων είναι επαρκείς για τη διαχείριση αποθεμάτων [19]. Ανεξάρτητα από τη διάρκεια ζωής των προϊόντων, οι τριμηνιαίες προβλέψεις μπορούν να βοηθήσουν τους διαχειριστές να λάβουν αποφάσεις επιχειρηματικής ανάπτυξης στους τομείς της χρηματοδότησης, του σχεδιασμού της υποδομής και του μάρκετινγκ [16]

Επίσης, τα εξωτερικά χαρακτηριστικά που χρησιμοποιούνται για την πρόβλεψη πωλήσεων μπορεί να είναι δεδομένα που σχετίζονται με τον καιρό [17], οικονομικοί δείκτες και χαρακτηριστικά που σχετίζονται με ημερομηνίες, όπως διακοπές ή εκδηλώσεις που προκαλούν μαζική κατανάλωση (π.χ. το Super Bowl στις ΗΠΑ).

Τα εσωτερικά χαρακτηριστικά, που χρησιμοποιούνται για την πρόβλεψη, [16] είναι οι πωλήσεις των προηγούμενων 6 ημερών, οι πωλήσεις της αντίστοιχης ημέρας του προηγούμενου έτους (ίδια ημέρα της εβδομάδας, περίπου ίδια ημερομηνία), οι πωλήσεις των προηγούμενων 6 ημερών εκείνης της ημέρας και η εκατοστιαία μεταβολή στις πωλήσεις μεταξύ του τρέχοντος έτους και του προηγούμενου έτους. Τα 5 χαρακτηριστικά που βρέθηκαν πιο χρήσιμα εμπειρικά ήταν οι φετινές πωλήσεις

με lag 1 (προηγούμενη ημέρα) και lag 6 (την ίδια μέρα της προηγούμενης εβδομάδας, καθώς τα καταστήματα ήταν ανοιχτά 6 ημέρες την εβδομάδα) και οι αντίστοιχες πωλήσεις του προηγούμενου έτους με lag 3, 5 και 6. Τα ακόλουθα εσωτερικά και εξωτερικά χαρακτηριστικά εξετάστηκαν στο (Žliobaite et al. 2009)[20]: Πωλήσεις προϊόντων, κινητός μέσος όρος πωλήσεων προϊόντων, περσινός κινητός μέσος όρος πωλήσεων προϊόντων, αθροιστικές πωλήσεις όλων των προϊόντων, περσινές σωρευτικές πωλήσεις όλων των προϊόντων, προωθήσεις προϊόντων, ημερολόγιο, θρησκεία και σχολικές διακοπές ως δυαδικά χαρακτηριστικά, εποχές όπως 4 δυαδικά χαρακτηριστικά, μήνες ως 13 δυαδικά χαρακτηριστικά και θερμοκρασία. Οι πωλήσεις για τα τελευταία έξι χρονικά σημεία (εβδομάδες σε εκείνη την περίπτωση) των σχετικών προϊόντων χρησιμοποιήθηκαν ως χαρακτηριστικά στο Meulstee and Pechenizkiy (2008) [21]. Επίσης, σχετικά προϊόντα θεωρήθηκαν αυτά που ανήκουν στο ίδιο σύμπλεγμα μετά από μια αθροιστική ιεραρχική διαδικασία ομαδοποίησης. Πειράματα με μια ποικιλία μετρήσεων απόστασης, έδειξαν ότι η δυναμική στρέβλωση χρόνου και η μεγαλύτερη κοινή υποακολουθία οδηγούν σε ικανοποιητικά αποτελέσματα.

Μια απλή τεχνική πρόβλεψης είναι ο κινητός μέσος όρος (MA). Η πρόβλεψη για μια μεταβλητή στόχο είναι ο μέσος όρος της τιμής της κατά τις τελευταίες n παρατηρήσεις. Μια προσέγγιση υβριδικής υπολογιστικής νοημοσύνης χρησιμοποιείται στο Doganis et al. (2006) [16]. Ένα δίκτυο συνάρτησης ακτινικής βάσης (RBF) εκπαιδεύεται με γρήγορο, χωρίς επίβλεψη τρόπο, χρησιμοποιώντας τον αλγόριθμο ασαφών μέσων.

Μια προσέγγιση συνόλου χρησιμοποίησαν οι Meulstee and Pechenizkiy (2008) [21]. Το σύνολο αποτελούνταν από 24 μοντέλα που παράγονται με τη χρήση 8 αλγορίθμων μάθησης (2 decision tree learners, 2 rule learners, 2 lazy learners, a support vector machine και logistic regression) με 3 διαφορετικά χρονικά παράθυρα (13, 26 και 52 χρονικά σημεία). Το σύνολο συνδυάστηκε χρησιμοποιώντας δυναμική ολοκλήρωση [22]

Η βαθιά μάθηση εφαρμόστηκε τα τελευταία χρόνια για την πρόβλεψη των πωλήσεων ενός ιαπωνικού σούπερ μάρκετ [17]. Συγκεκριμένα, η προσέγγιση περιελάμβανε έναν αυτόματο κωδικοποιητή στοιβαγμένης αποθορυβοποίησης για τη δημιουργία χαρακτηριστικών υψηλού επιπέδου, οι οποίες στη συνέχεια τροφοδοτήθηκαν σε ένα δίκτυο βραχυπρόθεσμης μνήμης για την πρόβλεψη μελλοντικών πωλήσεων.

3.1. Πρόβλεψη Χρονοσειρών

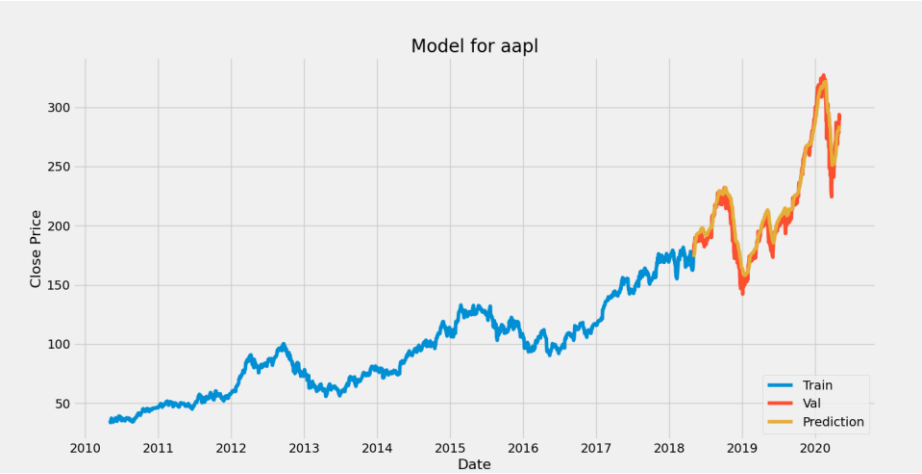
Η πρόβλεψη χρονοσειρών είναι η διαδικασία ανάλυσης ιστορικών δεδομένων με τη χρήση στατιστικής και μηχανικής μάθησης, για την δημιουργία προβλέψεων ώστε να είμαστε σε θέση να λάβουμε στρατηγικές αποφάσεις. Όσο πιο ολοκληρωμένα είναι τα δεδομένα που έχουμε στη διάθεση μας, τόσο πιο ακριβείς μπορούν να γίνουν οι προβλέψεις. Η πρόβλεψη σειρών χρησιμοποιείται συχνά σε συνδυασμό με την ανάλυση χρονοσειρών. Η ανάλυση χρονοσειρών περιλαμβάνει την ανάπτυξη μοντέλων για την κατανόηση των δεδομένων. Η ανάλυση μπορεί να παρέχει το «γιατί» πίσω από τα αποτελέσματα που βλέπουμε και στη συνέχεια, η πρόβλεψη κάνει το επόμενο βήμα για το τι πρέπει να γίνει με αυτή τη γνώση.

Η πρόβλεψη μπορεί να έχει εφαρμογή σε διάφορους κλάδους, από πρόβλεψη καιρού, πρόβλεψη κλίματος, μέχρι χρηματοοικονομικές προβλέψεις, προβλέψεις υγειονομικής περίθαλψης, προβλέψεις προϊόντων λιανικής πώλησης, επιχειρηματικές προβλέψεις, προβλέψεις περιβαλλοντικών μελετών, πρόβλεψη κοινωνικών μελετών και πολλά άλλα. Σε οποιονδήποτε τομέα υπάρχουν διαθέσιμα ιστορικά δεδομένα, μπορούν να επεξεργαστούν με μεθόδους ανάλυσης χρονοσειρών και στη συνέχεια να μοντελοποιηθεί και να πραγματοποιηθεί η πρόβλεψη.

Φυσικά, υπάρχουν περιορισμοί όταν προσπαθούμε αντιμετωπίζουμε το απρόβλεπτο και το άγνωστο. Η πρόβλεψη χρονοσειρών δεν είναι αλάνθαστη και δεν είναι κατάλληλη ή χρήσιμη για όλες τις περιπτώσεις. Επειδή στην πραγματικότητα δεν υπάρχει κανόνας για το πότε πρέπει ή όχι να χρησιμοποιούμε την πρόβλεψη, εναπόκειται στους αναλυτές και τις ομάδες ανάλυσης δεδομένων να γνωρίζουν τους περιορισμούς της ανάλυσης και τι μπορούν να υποστηρίξουν τα μοντέλα. Δεν ταιριάζει κάθε μοντέλο σε κάθε σύνολο δεδομένων και δεν απαντάται κάθε ερώτηση. Οι ομάδες ανάλυσης δεδομένων θα πρέπει να χρησιμοποιούν πρόβλεψη χρονοσειρών όταν κατανοούν την επιχειρηματική ερώτηση και έχουν τα κατάλληλα δεδομένα και δυνατότητες πρόβλεψης για να απαντήσουν σε αυτήν την ερώτηση. Μία καλή πρόβλεψη λειτουργεί με καθαρά, ορθά χρονικώς δομιμένα δεδομένα και μπορεί να εντοπίσει τάσεις και μοτίβα στα ιστορικά δεδομένα. Οι αναλυτές μπορούν να εντοπίσουν τη διαφορά μεταξύ τυχαίων διακυμάνσεων ή ακραίων τιμών και μπορούν να διαχωρίσουν τις γνήσιες πληροφορίες από τις εποχιακές διακυμάνσεις. Η ανάλυση χρονοσειρών δείχνει πώς αλλάζουν τα δεδομένα με την πάροδο του

χρόνου και μια καλή πρόβλεψη, μπορεί να προσδιορίσει την κατεύθυνση προς την οποία αλλάζουν τα δεδομένα.

Ακολουθούν κάποια παραδείγματα από διάφορους κλάδους, ώστε να γίνουν πιο συγκεκριμένες οι έννοιες της ανάλυσης και πρόβλεψης χρονοσειρών. Πρόβλεψη της τιμής κλεισίματος μιας μετοχής στο τέλος της ημέρας, πρόβλεψη μονάδων προϊόντων που πωλούνται καθημερινά στα καταστήματα, πρόβλεψη των επιπέδων της ανεργίας μιας χώρας, πρόβλεψη της μέσης τιμής της βενζίνης. Στην παρακάτω εικόνα φαίνεται η πρόβλεψη της τιμής κλεισίματος της αμερικανικής εταιρείας τεχνολογίας Apple Inc. Με μπλε φαίνονται οι τιμές οι οποίες χρησιμοποιήθηκαν για την εκπαίδευση του μοντέλου, με πορτοκαλί οι προβλέψεις του και με κόκκινο οι πραγματικές τιμές, με τις οποίες θα γίνει η εκτίμηση ακρίβειας του μοντέλου. Στο συγκεκριμένο παράδειγμα φαίνεται ξεκάθαρα, ότι η πρόβλεψη είναι αρκετά κοντά με τις πραγματικές τιμές.



Εικόνα 12: Παράδειγμα πρόβλεψης τιμής μετοχής [39]

Ωστόσο γεγονότα που είναι τυχαία δεν μπορούν να προβλεφθούν ποτέ με ακρίβεια, ανεξάρτητα από το πόσα δεδομένα συλλέγουμε. Για παράδειγμα: μπορούμε να παρατηρούμε και να συλλέγουμε δεδομένα κάθε εβδομάδα για κάθε νικητή της λοταρίας, αλλά δεν μπορούμε ποτέ να προβλέψουμε ποιος θα κερδίσει την επόμενη. Εξαρτάται από τα δεδομένα και την ανάλυση δεδομένων για το πότε μπορούμε να χρησιμοποιήσουμε την πρόβλεψη, επειδή η πρόβλεψη εξαρτάται σε μεγάλο βαθμό από διάφορους παράγοντες.

3.1.1. Αλγόριθμοι Μηχανικής Μάθησης

Βασιζόμενοι σε ιστορικά δεδομένα, μπορούμε ως ένα βαθμό να προβλέψουμε τα γεγονότα του μέλλοντος. Για μεγάλο χρονικό χρονικό διάστημα, στατιστικά μοντέλα χρησιμοποιούνταν συνήθως για τη διεξαγωγή προβλέψεων. Με τη συνεχή πρόοδο της επιστήμη των υπολογιστών, είμαστε πλέον σε θέση να χρησιμοποιούμε μοντέλα μηχανικής μάθησης, προκειμένου να πραγματοποιήσουμε αυτές τις προβλέψεις. Οι αλγόριθμοι μηχανικής μάθησης που χρησιμοποιούνται συνήθως είναι για παράδειγμα οι Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines), Τυχαία παλινδρόμηση δασών (Random Forest Regression), Gradient Boosting, XGBoost, LightGBM και Απλή Γραμμική Παλινδρόμηση (Simple Linear Regression).

3.1.1. XGBoost

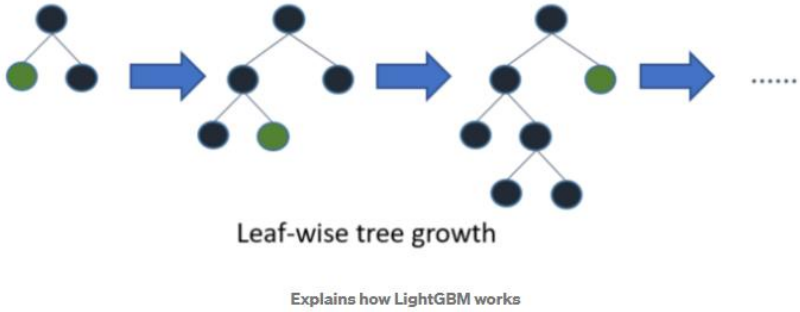
Η μέθοδος eXtreme Gradient Boosting, γνωστή και ως XGBoost είναι ένας αποτελεσματικός αλγόριθμος gradient boosting που χρησιμοποιείται για προβλήματα ταξινόμησης και παλινδρόμησης. Είναι γρήγορος όσο και αποτελεσματικός, έχει καλή απόδοση, αν όχι την καλύτερη, σε ένα ευρύ φάσμα εργασιών πρόβλεψης. Ο XGBoost μπορεί επίσης να χρησιμοποιηθεί για πρόβλεψη χρονοσειρών, αν και πρώτα πρέπει να μετατραπεί το σύνολο δεδομένων χρονοσειρών σε πρόβλημα εποπτευμένης μάθησης. Απαιτεί επίσης τη χρήση μιας εξειδικευμένης τεχνικής για την αξιολόγηση του μοντέλου που ονομάζεται walk-forward validation, καθώς η αξιολόγηση του μοντέλου με τη χρήση k-fold cross validation, θα είχε ως αποτέλεσμα αισιόδοξα μεροληπτικά αποτελέσματα.[40] Ο αλγόριθμος ανήκει στην κατηγορία των Gradient Boosting αλγορίθμων. Η μέθοδος Boosting είναι μια ensemble μέθοδος, η οποία βασίζεται στη βασική ιδέα του συνδύασμου των αποτελεσμάτων πολλών αδύναμων αλγορίθμων, των οποίων το σφάλμα (error rate) είναι ελαφρώς καλύτερο από την τυχαία επιλογή (Freund & Schapire).[24][23] Με την τεχνική του Boosting μειώνεται η διακύμανση όσο και η μεροληψία. Η Gradient Boosting είναι μια τεχνική μηχανικής μάθησης, όπου προσθέτοντας διαδοχικά νέα μοντέλα, τα οποία ειναι decission trees, έχει στόχο να βελτιώσει τα σφάλματα (τη διαφορά μεταξύ των πραγματικών και των προβλεπόμενων τιμών), μέχρις ότου να μην υπάρχει περαιτέρω βελτίωση. Όταν ένα δέντρο αποφάσεων είναι ο αδύναμος μαθητής (weak learner), ο αλγόριθμος που προκύπτει ονομάζεται gradient-boosted trees, και

συνήθως υπερτερεί του random forest. Ο Gradient Boosting μπορεί να χαρακτηριστεί ως άπληστος, και να υπερπροσαρμοστεί γρήγορα ένα σύνολο δεδομένων εκπαίδευσης (overfitting). Μπορεί ωστόσο να επωφεληθεί από μεθόδους κανονικοποίησης που τιμωρούν διάφορα μέρη του αλγορίθμου και γενικά βελτιώνουν την απόδοση του αλγορίθμου μειώνοντας την υπερπροσαρμογή. [25][26][27] Ο XGBoost είναι μια μέθοδος που βασίζεται στα δέντρα, είναι μια προέκταση του gradient boosting. Τα αδύναμα μοντέλα είναι δέντρα παλινδρόμησης τα οποία προστίθενται επανειλημμένα, για να προβλέψουν τα υπολείμματα προηγούμενων δέντρων και τα οποία στη συνέχεια συνδυάζονται με τα προηγούμενα δέντρα για να κάνουν την τελική πρόβλεψη. Σε προβλήματα με μη δομημένα δεδομένα (όπως εικόνες, κειμενο, κλπ), τα νευρωνικά δίκτυα τείνουν να έχουν καλύτερη απόδοση, ενώ σε δομημένα δεδομένα οι αλγόριθμοί που βασίζονται σε δέντρα αποφάσεων θεωρούνται αποδοτικότεροι.

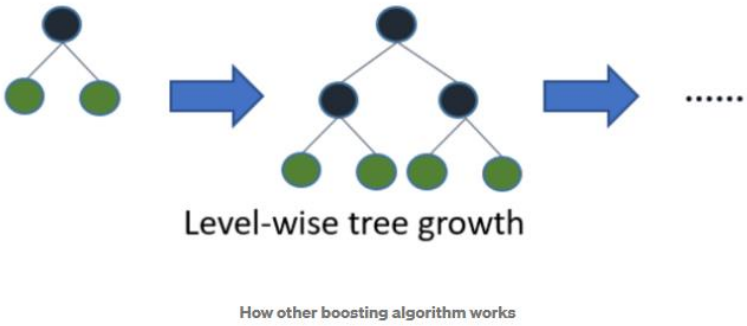
Ο XGBoost ξεκίνησε αρχικά ως ερευνητικό έργο από τον Tianqi Chen [28] ως μέρος της ομάδας Distributed (Deep) Machine Learning Community (DMLC). Είναι μια βιβλιοθήκη λογισμικού ανοιχτού κώδικα, μπορεί να τρέξει σε Windows, Linux και OS X και να χρησιμοποιηθεί σε αρκετές γλώσσες προγραμματισμού όπως, C++, Java, Python (μέσω της βιβλιοθήκης scikit-learn), R (μέσω του πακέτου caret), Julia, Perl και Scala [41]. Τα τελευταία χρόνια εφαρμόζεται συχνά και έχει κερδίσει αρκετούς διαγωνισμούς μηχανικής μάθησης στην πλατφόρμα της Kaggle.

3.1.1. LightGBM

Ο LightGBM είναι ένας σχετικά νέος αλγόριθμος μηχανικής μάθησης, ανήκει στη οικογένεια των boosting αλγορίθμων και χρησιμοποιεί μεθόδους μάθησης βασισμένες στα δέντρα. Το χαρακτηριστικό που τον κάνει να διαφοροποιείται από τους άλλους αλγορίθμους που βασίζονται στα δέντρα, είναι ότι ο LightGBM αναπτύσσει τα δέντρα κατακόρυφα ενώ οι άλλοι αλγόριθμοι αναπτύσσουν τα δέντρα οριζόντια, πράγμα που σημαίνει ότι ο LightGBM αναπτύσσει τα κατά φύλλα (leaf-wise), ενώ οι άλλοι τα αναπτύσσουν σε επίπεδα (level-wise). Θα επιλέξει το φύλλο με μέγιστη απώλεια δέλτα για να αναπτυχθεί. Αναπτύσσοντας το ίδιο φύλλο, ο leaf-wise αλγόριθμος μπορεί να μειώσει περισσότερες απώλειες από έναν level-wise αλγόριθμο. Στις παρακάτω εικόνες φαίνεται η υλοποίηση του LightGBM και των υπόλοιπων boosting αλγορίθμων.



Εικόνα 13: Πως λειτουργεί ο LightGBM



Εικόνα 14: Πως λειτουργούν οι boosting αλγόριθμοι [42]

Δεδομένου ότι το μέγεθος των δεδομένων αυξάνεται μέρα με τη μέρα, γίνεται όλο και δυσκολότερο για τους παραδοσιακούς αλγόριθμους της επιστήμης των δεδομένων να αποδώσουν ταχύτερα και να εξάγουν αποτελέσματα. Το πρόθεμα «Light» στον αλγόριθμο LightGBM χαρακτηρίζει την υψηλή του ταχύτητά εκμάθησης. Είναι σε θέση να διαχειριστεί αρκετά μεγάλο μέγεθος δεδομένων χρησιμοποιώντας μικρότερη μνήμη προκειμένου να εκτελεστεί. Ένας άλλος λόγος για τον οποίο ο LightGBM είναι αρκετά δημοφιλής είναι η υψηλή απόδοση και επειδή εστιάζει στην ακρίβεια των αποτελεσμάτων. Επίσης ο LightGBM μπορεί να υποστηρίξει και παράλληλη, κατανεμημένη μάθηση μέσω GPU, καθώς είναι ικανός να χειρίζεται δεδομένα μεγάλης κλίμακας. Παρόλα αυτά δεν μπορεί να χρησιμοποιηθεί σε όλες τις περιπτώσεις. Δεν συνίσταται η χρήση του σε μικρά σύνολα δεδομένων γιατί είναι

αρκετά εύκολο να προβεί σε υπερπροσαρμογή (overfitting), ειδικά σε μικρά σύνολα. Δεν υπάρχει ωστόσο ένα όριο στον αριθμό των σειρών.

Μερικές από τις βασικές παραμέτρους του LightGBM είναι οι ακόλουθοι:

Παράμετροι ελέγχου

max_depth: Περιγράφει το μέγιστο βάθος του δέντρου. Αυτή η παράμετρος χρησιμοποιείται για τον χειρισμό της υπερπροσαρμογής του μοντέλου. Καθώς παρατηρούμε ότι το μοντέλο τείνει να υπερπροσαρμόζεται μπορούμε να μειώσουμε το max_depth.

min_data_in_leaf: Είναι ο ελάχιστος αριθμός των εγγραφών που μπορεί να έχει ένα φύλλο. Η προεπιλεγμένη τιμή είναι 20. Μπορεί επίσης να χρησιμοποιηθεί για την αντιμετώπιση της υπερπροσαρμογής (overfitting).

feature_fraction: Χρησιμοποιείται όταν ο τύπος του αλγορίθμου boosting είναι τυχαίο δάσος (random forest). Η τιμή 0,8 σημαίνει ότι ο LightGBM θα επιλέξει τυχαία το 80% των παραμέτρων σε κάθε επανάληψη για την κατασκευή δέντρων.

bagging_fraction: καθορίζει το κομμάτι των δεδομένων που θα χρησιμοποιηθεί για κάθε επανάληψη και γενικά χρησιμοποιείται για την επιτάχυνση της εκπαίδευσης και την αποφυγή υπερπροσαρμογής.

early_stopping_round: Αυτή η παράμετρος βοηθάει στο να επιταχύνουμε την ανάλυσή μας. Το μοντέλο θα σταματήσει να εκπαιδεύεται εάν μια μετρική από τα validation data δεν βελτιωθεί στους τελευταίους γύρους. Ως αποτέλεσμα μειώνονται οι υπερβολικές επαναλήψεις.

lambda: αναφέρεται στην εξομάλυνση. Η τυπική τιμή κυμαίνεται από 0 έως 1. Ελέγχει πόσο θέλουμε να τιμωρήσουμε το άθροισμα των βαρών των τετραγωνικών φύλλων. Συνήθως αναφερόμαστε σε αυτήν την παράμετρο ως όρο εξομάλυνσης L2, καθώς εφαρμόζεται η συνάρτηση τετραγώνου.

min_gain_to_split: Αυτή η παράμετρος περιγράφει το ελάχιστο κέρδος όταν πρόκειται να γίνει διαίρεση. Μπορεί να χρησιμοποιηθεί για τον έλεγχο του αριθμού των διαχωρισμών, που θα είναι χρήσιμοι, στο δέντρο.

max_cat_group: Όταν ο αριθμός της κατηγορίας είναι μεγάλος, η εύρεση του σημείου διαχωρισμού σε αυτήν μπορεί εύκολα να γίνει overfitting. Έτσι το LightGBM τα συγχωνεύει σε ομάδες 'max_cat_group' και βρίσκει τα σημεία διαχωρισμού στα όρια της ομάδας. Σαν προεπιλογή ορίζεται το 64.

Βασικές Παράμετροι

task: Καθορίζει την εργασία που θέλουμε να εκτελέσουμε στα δεδομένα. Μπορεί να είναι είτε εκπαίδευση είτε πρόβλεψη.
application: Είναι η πιο σημαντική παράμετρος και καθορίζει την εφαρμογή του μοντέλου, είτε πρόκειται για πρόβλημα παλινδρόμησης είτε για πρόβλημα ταξινόμησης. Το LightGBM σαν προεπιλογή θεωρεί το μοντέλο ως μοντέλο παλινδρόμησης.

regression: για παλινδρόμηση
binary: για δυαδική ταξινόμηση
multiclass: για πρόβλημα ταξινόμησης πολλαπλών κλάσεων

boosting: ορίζει τον τύπο του αλγορίθμου που θέλουμε να εκτελέσουμε, default = gdbt

gbdt: Gradient Boosting Decision Tree
rf: τυχαίο δάσος
dart: Dropouts συναντούν Multiple Additive Regression Trees
goss: Gradient-based One-Side Sampling

num_boost_round: Αριθμός επαναλήψεων ενίσχυσης, συνήθως 100+

learning_rate: Καθορίζει την επίδραση κάθε δέντρου στο τελικό αποτέλεσμα. Το GBM λειτουργεί ξεκινώντας με μια αρχική εκτίμηση η οποία ενημερώνεται

χρησιμοποιώντας την έξοδο κάθε δέντρου. Η παράμετρος εκμάθησης ελέγχει το μέγεθος αυτής της αλλαγής στις εκτιμήσεις. Τυπικές τιμές: 0,1, 0,001, 0,003…

num_leaves: αριθμός φύλλων σε πλήρες δέντρο, προεπιλογή: 31

device: προεπιλογή cpu, μπορεί επίσης να δεχθεί και gpu

Παράμετροι μετρικών

metric: Μια από τις σημαντικές παραμέτρους καθώς καθορίζει το σφάλμα για την κατασκευή μοντέλου. Μερικα σφάλματα για παλινδρόμηση και ταξινόμηση είναι:

mae: μέσο απόλυτο σφάλμα
mse: μέσο τετραγωνικό σφάλμα
binary_logloss: απώλεια για δυαδική ταξινόμηση
multi_logloss: απώλεια για πολλαπλή ταξινόμηση

Παράμετροι IO

max_bin: υποδηλώνει τον μέγιστο αριθμό bin που θα τοποθετηθεί η τιμή του χαρακτηριστικού.

categorical_feature: Δηλώνει το δείκτη των κατηγορικών χαρακτηριστικών. Αν categorical_features=0,1,2 τότε η στήλη 0, η στήλη 1 και η στήλη 2 είναι κατηγορικές μεταβλητές.

ignore_column: οι στήλες που υποδηλώνει ο δείκτης θα αγνοηθούν.

save_binary: Ο καθορισμός της παραμέτρου true θα αποθηκεύσει το σύνολο δεδομένων σε ένα δυαδικό αρχείο, αυτό το δυαδικό αρχείο θα επιταχύνει τον χρόνο ανάγνωσης των δεδομένων την επόμενη φορά.
num_leaves: Η κύρια παράμετρος για τον έλεγχο της πολυπλοκότητας του μοντέλου δέντρων. Στην ιδανική περίπτωση, η τιμή των num_leaves θα πρέπει να

είναι μικρότερη ή ίση με 2^(max_depth). Μεγαλύτερη τιμή από αυτή θα έχει ως αποτέλεσμα την υπερπροσαρμογή.

min_data_in_leaf: Η ανάθεση μιας μεγάλης τιμής μπορεί να αποφύγει την ανάπτυξη του δέντρου σε βάθος, όμως μπορεί να προκαλέσει κακή προσαρμογή (under-fitting). Μπορεί να ρυθμιστεί σε εκατοντάδες ή χιλιάδες για ένα μεγάλο σύνολο δεδομένων.

max_depth: Το max_depth χρησιμοποιείται για να περιορίσουμε το βάθος του δέντρου.

[29][43][44]

4. Πειράματα

4.1. Σύνοψη πειραμάτων

Στο παρών κεφάλαιο περιγράφονται μια σειρά πειραμάτων που έγιναν για τη πρόβλεψη πωλήσεων χρησιμοποιώντας τον LightGBM αλγόριθμο για ένα σύνολο δεδομένων το οποίο περιέχει ιστορικά δεδομένα πωλήσεων της εταιρίας Walmart τα όποια θα αναλυθούν με λεπτομέρεια στην επόμενη ενότητα.

4.2. Δεδομένα

Τα δεδομένα για την υλοποίηση της παρούσας διπλωματικής αντλήθηκαν από την ιστοσελίδα της Kaggle και συγκεκριμένα από τον M5 Forecasting - Accuracy διαγωνισμό, που διοργανώθηκε από το Makridakis Open Forecasting Center (MOFC) του πανεπιστημίου της Λευκωσίας. Στο dataset περιέχονται ιστορικά δεδομένα πωλήσεων από την αμερικάνικη πολυεθνική εταιρία Walmart, για σχεδόν 30.000 διαφορετικά προϊόντα και για 1941 ημέρες, συγκεκριμένα από 29 Ιανουαρίου του 2011 έως και 19 Ιουνίου του 2016 . Τα δεδομένα συλλέχθηκαν από συνολικά δέκα (10) καταστήματα, τριών πολιτειών της Αμερικής, (Καλιφόρνια, Τέξας και Γουισκόνσιν), και περιέχουν πληροφορίες όπως επίπεδο αντικειμένου, σε ποιο τμήμα βρίσκεται το κάθε αντικείμενο, κατηγορίες προϊόντων, λεπτομέρειες καταστήματος. Επιπλέον, περιέχει και επεξηγηματικές μεταβλητές όπως τιμή, προσφορές, ημέρα της εβδομάδας και ειδικές εκδηλώσεις - γιορτές. Αναλυτικότερα:

Στο dataset περιλαμβάνονται τα παρακάτω αρχεία:

1. **calendar.csv** - Περιέχει πληροφορίες για την ημέρα της εβδομάδας, ειδικά γεγονότα για τις ημερομηνίες πώλησης των προϊόντων.
2. **sales_train_validation.csv** - Περιέχει τον όγκο πωλήσεων των προϊόντων ανά ημέρα κατάστημα και κατηγορία προϊόντος [d_1 - d_1913].
3. **sell_prices.csv** - Περιέχει πληροφορίες για την τιμή πώλησης των προϊόντων ανά κατάστημα και ημερομηνία.

4. **sales_train_evaluation.csv** - Περιέχει τον όγκο πωλήσεων των προϊόντων ανα ημέρα κατάστημα, κατηγορία προϊόντος [d_1 - d_1941].

Τα ιστορικά δεδομένα με τα οποία εργαζόμαστε, περιλαμβάνουν τον όγκο πωλήσεων για 3049 μοναδικά προϊόντα, τα οποία πωλούνται σε 3 πολιτείες της Αμερικής, συνολικά σε 10 καταστήματα. Όλα τα καταστήματα έχουν τρεις (3) κατηγορίες προϊόντων, οι οποίες περαιτέρω χωρίζονται σε επτά (7) υποκατηγορίες συνολικά.
Για την πολιτεία της Καλιφόρνιας έχουμε τέσσερα καταστήματα: CA_1, CA_2, CA_3, CA_4.
Για την πολιτεία του Τέξας έχουμε τρία καταστήματα: TX_1, TX_2, TX_3.
Για την πολιτεία του Γουισκόνσιν έχουμε τρία καταστήματα: WI_1, WI_2, WI_3.
Για κάθε ένα κατάστημα τα προϊόντα χωρίζονται σε τρεις (3) κατηγορίες: Hobbies, Foods και Household
Στην κατηγορία Hobbies, έχουμε 2 τμήματα, HOBBIES_1, HOBBIES_2,
στην κατηγορία Foods, 3 τμήματα, FOODS_1, FOODS_2, FOODS_3 και
στην κατηγορία Household, 2 τμήματα, HOUSEHOLD_1, HOUSEHOLD_2,
όπου έπειτα το κάθε ένα προϊόν αριθμείται με αύξουσα σειρά από το 1 έως το 3049.
Λόγω ότι το σύνολο δεδομένων είναι αρκετά μεγάλο, από άποψη ημερομηνιών, είναι σχετικά δύσκολη η εύρεση όλων των υποκείμενων μοτίβων που υπάρχουν στο σύνολο δεδομένων. Για την καλύτερη κατανόηση και αναζήτηση μοτίβων, που διαφορετικά θα μπορούσαν να είναι δύσκολο να προσδιοριστούν, με μια ολοκληρωτική προσέγγιση, προσπαθήσαμε να απαντήσουμε στις παρακάτω ερωτήσεις.

4.3. Εργαλεία υλοποίησης

Για την υλοποίηση της παρούσας έρευνας, την εισαγωγή των δεδομένων, τον καθαρισμό και την επεξεργασία τους, όπως επίσης και για την δημιουργία του προβλεπτικού μοντέλου πωλήσεων για τα καταστήματα λιανικής πώλησης Walmart, χρησιμοποιήθηκε η γλώσσα προγραμματισμού Python, μέσω της διανομής ανοιχτού κώδικα Anaconda. Έγινε χρήση διαφόρων βιβλιοθηκών της Python για την εισαγωγή και την επεξεργασία των δεδομένων. Συγκεκριμένα χρησιμοποιήθηκαν οι

βιβλιοθήκες os για εύκολη διαχείριση του συστήματος, pandas και numpy για την εισαγωγή και διαχείριση των δεδομένων, matplotlib (plyplot) και seaborn για την οπτικοποίηση των δεδομένων και δημιουργία γραφημάτων, joblib από scikit-learn (externals) για την αποθήκευση και ανάκτηση των μοντέλων, pickle για αποθήκευση αρχείων, csv για αποθήκευση των δεδομένων σε αρχεία .csv και mse για την αξιολόγηση των μοντέλων καθώς επίσης και το μοντέλο μηχανικής μάθησης lightgbm.

4.4. Μετρικές Αξιολόγησης

Για να αξιολογηθούν οι προβλέψεις πωλήσεων που έγιναν με χρήση του LightGBM, χρησιμοποιήθηκε η μέτρηση σφάλματος RMSE. Η χρονική σειρά που χρησιμοποιήθηκε σε αυτήν την περίπτωση, έλαβε τον μέσο αριθμό πωλήσεων που πραγματοποιήθηκαν, κατά τη διάρκεια σχεδόν 1900 ημερών, λαμβάνοντας υπόψη την αρχική ιεραρχική διάταξη των id των προϊόντων. Η χρησιμότητα της μέτρησης σφάλματος RMSE θεωρείται αρκετά έγκυρη, δεδομένου ότι, λαμβάνουμε υπόψη τον μέσο αριθμό πωλήσεων που πραγματοποιήθηκαν, ακυρώνοντας έτσι την επίδραση πολλών μηδενικών για πολλά αναγνωριστικά προϊόντων, παράλληλα με τον αριθμό των ημερών.

RMSE (Root Mean Square Error) - (Σφάλμα μέσου τετραγώνου ρίζας)

$$RMSE = \sqrt{MSE}$$

4.5. Ανάλυση Δεδομένων

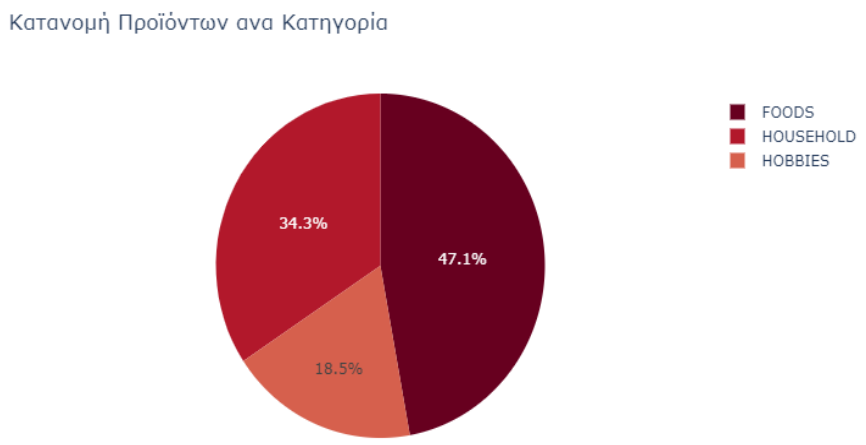
4.5.1. Εισαγωγή

Πέρα από την πρόβλεψη των πωλήσεων των προϊόντων της εταιρείας λιανικού εμπορίου Walmart ανά πολιτεία, κατάστημα, κατηγορία προϊόντος και τμήμα, είναι εξίσου σημαντική μια πρακτική ερμηνεία (μη επιστημονική) των δεδομένων με σκοπό την εύρεση των κύριων χαρακτηριστικών της μορφής των δεδομένων και γνώση του πεδίου προέλευσης τους.

Σε αυτή την ενότητα παρουσιάζονται μέσω διαφόρων γραφημάτων, οπτικοποιήσεις για το διαθέσιμο σύνολο δεδομένων, για μια πρώτη εικόνα και την καλύτερη κατανόηση των δεδομένων.

4.5.2. Γραφική Ανάλυση

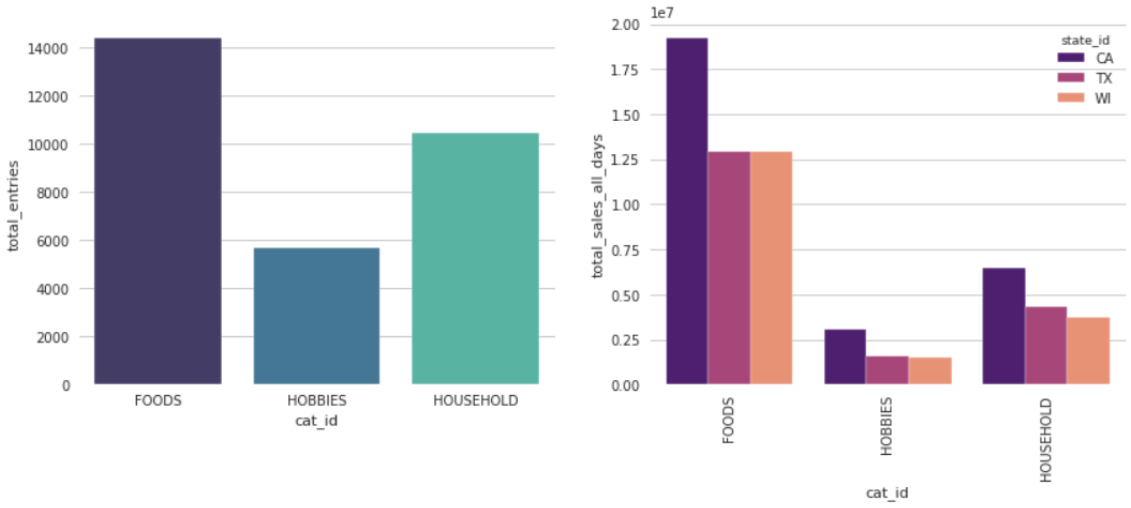
Για την υλοποίηση της παρούσας διπλωματικής λήφθηκαν δεδομένα από 1941 ημέρες, συγκεκριμένα από 29 Ιανουαρίου του 2011 έως και 19 Ιουνίου του 2016.
Παρακάτω μπορούμε να δούμε ορισμένες διάφορες - σχέσεις μεταξύ των καταστημάτων, τα προϊόντα που πωλούνται, τις κατηγορίες που ενδιαφέρουν περισσότερο τους καταναλωτές, καθώς επίσης και την πώληση ειδών ανά συγκεκριμένη τοποθεσία. Στο αρχείο Sellprices περιέχονται πληροφορίες σε επίπεδο καταστημάτων, περίπου 6 εκατομμυρίων καταχωρήσεων, ενώ το train_sales έχει πληροφορίες για περίπου 30.000 διαφορετικά προϊόντα. Επιπλέον, το σύνολο δεδομένων στο calendar_df το οποίο περιέχει ημερολογιακά δεδομένα, καθιστά δυνατή την ανάλυση χρονοσειρών των προϊόντων που πωλούνται.
Τα προϊόντα κατανέμονται σε τρεις κατηγορίες. Τα είδη διατροφής - Foods είναι η κατηγορία με τις περισσότερες πωλήσεις (47,1%), ακολουθούν τα είδη σπιτιού – νοικοκυριού - Household (34,3%), και έπειτα τα είδη για χόμπι - Hobbies (18,5%) .



Εικόνα 15: Κατανομή προϊόντων ανά κατηγορία

Όσον αφορά τον συνολικό αριθμό των πωλήσεων, είναι εμφανές ότι συνολικά στα είδη που πωλήθηκαν, οι περισσότερες πωλήσεις πραγματοποιήθηκαν στην Καλιφόρνια CA (43,6%), ακολουθεί το Τέξας TX (28,8%) και τέλος το Ουισκόνσιν WI με μικρή διαφορά από το Τέξας (27,6%).

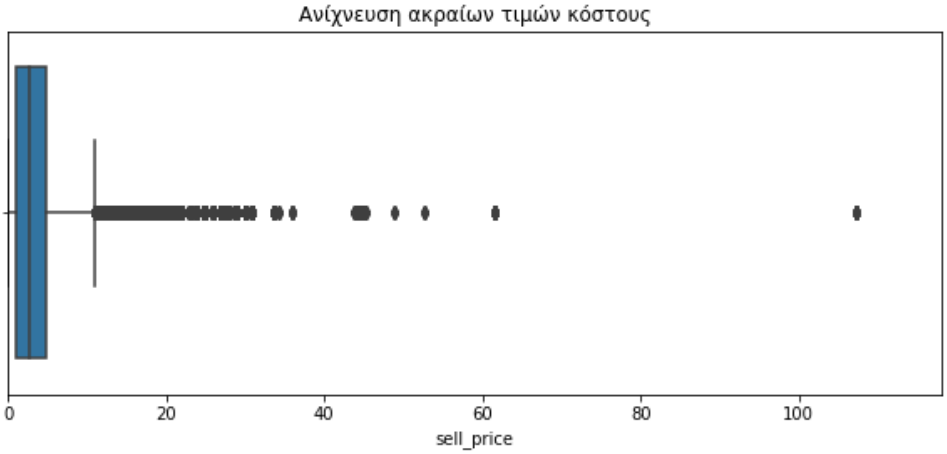
Εικόνα 16: Κατανομή των συνολικών πωλήσεων ανά πολιτεία



Εικόνα 17: Πωλήσεις προϊόντων ανά κατηγορία και ανά πολιτεία

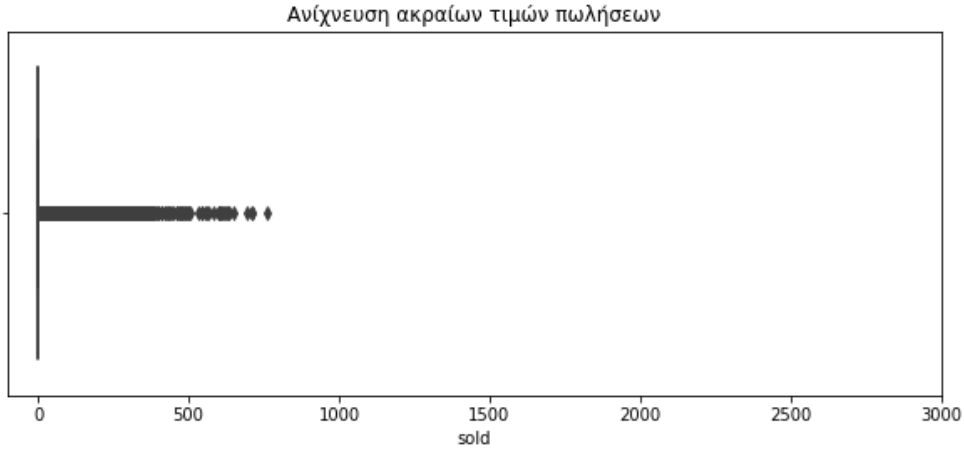
Τα παραπάνω γραφήματα υποδηλώνουν, ότι τα περισσότερα προιόντα που πωλήθηκαν ανήκαν στην κατηγορία FOODS, έπειτα στην κατηγορία HOUSEHOLD και οι λιγότερες πωλήσεις προιόντων ανήκαν στην κατηγορία HOBBIES. Επίσης

παρατηρείται ότι και στις τρεις κατηγορίες προϊόντων (FOODS, HOBBIES και HOUSEHOLD), οι περισσότερες πωλήσεις πραγματοποιηθήκανε στην πολιτεία της Καλιφόρνια, και ακολουθούν το Τέξας και το Ουισκόνσιν με ίδιο αριθμό πωλήσεων για την κατηγορία FOODS, και με μικρή διαφορά στις άλλες δυο κατηγορίες (HOBBIES και HOUSEHOLD) στο Τέξας παρουσιάζονται περισσότερες πωλήσεις σε σχέση με το Ουισκόνσιν.



Εικόνα 18: Εύρος τιμής πώλησης των προϊόντων

Στο παραπάνω γράφημα παρατηρείται η κατανομή των τιμών πώλησης των προϊόντων, όπου σαν ακραίες τιμές θεωρήθηκαν οι τιμές πάνω από 60, οι οποίες δεν συμπεριλήφθηκαν στα πειράματα μας, επειδή θεωρήθηκε ότι εισάγουν περισσότερο θόρυβο παρά χρήσιμη πληροφορία στα μοντέλα.



Εικόνα 19: Μέγιστος αριθμός πωλήσεων ανά ημέρα

Στην παραπάνω εικόνα φαίνεται ο μέγιστος αριθμός πωλήσεων προϊόντων ανά ημέρα. Δεν παρατηρούνται κάποιες ακραίες τιμές, οπότε όλες συμπεριλήφθηκαν στα πειράματά μας.

4.6. Ορισμός πειραμάτων

Η διαδικασία των πειραμάτων ξεκίνησε δημιουργώντας αρχικά ορισμένα βασικά μοντέλα χωρίς να προσθέσουμε επιπλέον στήλες-μεταβλητές στα δεδομένα μας. Τα βασικά μοντέλα είναι απλά μοντέλα χωρίς επιπρόσθετα χαρακτηριστικά, εργαστήκαμε δηλαδή μόνο με τις στήλες που ήδη υπήρχαν στο αρχικό σύνολο δεδομένων. Στόχος ήταν να έχουμε ένα μέτρο σύγκρισης, των μοντέλων χωρίς επιπλέον μεταβλητές, και των μετέπειτα προβλεπτικών μοντέλων με τις ενσωματωμένες προβλεπτικές μεταβλητές που δημιουργήθηκαν. Εδώ αξίζει να σημειώσουμε ότι λόγω περιορισμένης επεξεργαστικής ισχύς, δεν τρέξαμε ένα αρχικό μοντέλο για όλες τις τιμές του συνόλου δεδομένων, οι οποίες ήταν περίπου 59.181.000 γραμμές, αλλά εργαστήκαμε χωρίζοντας το αρχικό σύνολο δεδομένων και δημιουργώντας τέσσερα προβλεπτικά μοντέλα. Συγκεκριμένα ανά πολιτεία (state), ανά κατάστημα (store), ανά κατηγορία προϊόντος (category) και ανά τμήμα (department).
Τα πειράματα υλοποιήθηκαν με τη χρήση του LightGBM, ο οποίος χαρακτηρίζεται τόσο για την υψηλή του ταχύτητά εκμάθησης, όσο για την υψηλή απόδοση του και για την ακρίβεια των αποτελεσμάτων.
Η αρχική προσέγγιση μας ήταν να συγκρίνουμε διαφορετικά μοντέλα μηχανικής μάθησης όπως για παράδειγμα RandomForestRegressor, XGBRegressor όμως λόγω του όγκου δεδομένων και περιορισμένης υπολογιστικής δύναμης δεν ήταν δυνατή η ολοκλήρωση της εκπαίδευσης αυτών των μοντέλων, οπότε έγινε χρήση αποκλειστικά του LGBMRegressor.

Αρχικά, λόγω όγκου δεδομένων και δοκιμής μεταξύ της αρχικής τιμής (n-estimators = 100) και δοκιμών με μεγαλύτερες τιμές, η παράμετρος ορίστηκε σε οι n-estimators = 1000. Οι υπόλοιπες παράμετροι του αλγορίθμου ήταν οι προκαθορισμένες

4.7. Προ-Επεξεργασία Δεδομένων

Στο αρχικό σύνολο δεδομένων, οι ημέρες καθορίζονται σε διαφορετικές στήλες η κάθε μια. Έγινε μια μετατροπή των δεδομένων, έτσι ώστε για κάθε προϊόν σε κάθε κατάστημα σε κάθε πολιτεία, να δημιουργηθεί μια καινούρια σειρά για την κάθε ημέρα, με αποτέλεσμα το αρχικό σύνολο δεδομένων διαστάσεων 30.490 γραμμών x 1947 στηλών, να μετατραπεί σε ένα σύνολο δεδομένων 59.181.090 γραμμών x 8 στηλών. Αυτό είχε σαν αποτέλεσμα το πρόβλημά μας να αντιμετωπιστεί ως ένα πρόβλημα χρονοσειρών.

Έπειτα το νέο σύνολο δεδομένων συγχωνεύτηκε με το σύνολο δεδομένων calendar με βάση την ημέρα, και στη συνέχεια συγχωνεύτηκε με το σύνολο δεδομένων sell_prices με βάση το κατάστημα, προϊόν και wm_yr_wk, με αποτέλεσμα να προκύψει το τελικό σύνολο δεδομένων στο οποίο εργαστήκαμε, με 59.181.090 γραμμές x 22 στήλες.

Επίσης από την Εικόνα 18: Εύρος τιμής πώλησης των προϊόντων θεωρήσαμε σαν ακραίες τιμές, τις τιμές πώλησης προϊόντων οι οποίες είναι μεγαλύτερες του 60, οπότε δεν τις συμπεριλάβαμε στη δημιουργία των μοντέλων μας.

4.7.1. Δημιουργία Προβλεπτικών Μεταβλητών

Προκειμένου να πετύχουμε καλύτερα αποτελέσματα στα προβλεπτικά μοντέλα μας, πέρα των στηλών που υπήρχαν στο σύνολο δεδομένων, δημιουργήσαμε ορισμένες επιπρόσθετες προβλεπτικές μεταβλητές - στήλες. Ο λόγος είναι για να βοηθήσουμε το μοντέλο να δημιουργήσει συσχετίσεις ανάμεσα στις μεταβλητές και να εξάγει καλύτερες και ακριβέστερες προβλέψεις.

<u>Οι στήλες που υπήρχαν στα αρχεία του συνόλου δεδομένων.</u>

Id: αναγνωριστικό των προϊόντων
item_id: το id του προϊόντος
dept_id: σε ποιό τμήμα ανήκει το προϊόν
cat_id: σε ποιά κατηγορία ανήκει το προϊόν (FOODS, HOUSEHOLD, HOBBIES)

store_id: το αναγνωριστικό του καταστήματος και στις τρεις πολιτείες CA_1, CA_2, CA_3, CA_4, TX_1, TX_2, TX_3, WI_1, WI_2 & WI_3
state_id: την πολιτεία (CA, TX, WI)

wm_yr_wk: περιέχει πληροφορίες για την ημερομηνία
weekday: η ημέρα της εβδομάδας (Monday, Tuesday, Wednesday, …. Sunday)
wday: αρίθμηση των ημερών της εβδομάδας, ξεκινώντας από 1 (Saturday), 2 (Sunday), 3 (Monday) ….. 7 (Friday)
month: οι μήνες του έτους σε αύξοντα αριθμό από το 1 (Ιανουάριος) έως το 12 (Δεκέμβριος)
year: έτος
d: η ημέρα από d_1 έως d_1969
event_name_1: κάποια γιορτή ή κάποιο ειδικό γεγονός (πχ Superbowl, Easter κλπ)
event_name_2: κάποια γιορτή ή κάποιο ειδικό γεγονός
event_type_1: το είδος της γιορτής ή του γεγονότος
event_type_2: το είδος της γιορτής ή του γεγονότος

Η ομοσπονδιακή κυβέρνηση των Ηνωμένων Πολιτειών παρέχει ένα επίδομα διατροφικής βοήθειας που ονομάζεται Συμπληρωματικό Πρόγραμμα Βοήθειας Διατροφής - Supplement Nutrition Assistance Program (SNAP). Το πρόγραμμα παρέχει σε οικογένειες και άτομα χαμηλού εισοδήματος, που πληρούν ορισμένες προϋποθέσεις και οικονομικά κριτήρια, μένουν σε μια από τις πολιτείες που συμμετέχουν στο πρόγραμμα, μια κάρτα ηλεκτρονικής μεταφοράς παροχών. Αυτή η κάρτα μπορεί να χρησιμοποιηθεί ως χρεωστική κάρτα για την αγορά προϊόντων διατροφής, σε εξουσιοδοτημένα καταστήματα λιανικής πώλησης τροφίμων. Στις πολιτείες που συμμετέχουν στο πρόγραμμα, τα χρηματικά οφέλη διανέμονται στους δικαιούχους 10 ημέρες του μήνα. [#] https://www.benefits.gov/benefit/361

snap_CA: λαμβάνει τιμές 0 (όχι) ή 1 (ναι), ανάλογα με το αν είναι ημέρα του προγράμματος στην Καλιφόρνια ή όχι
snap_WI: λαμβάνει τιμές 0 (όχι) ή 1 (ναι), ανάλογα με το αν είναι ημέρα του προγράμματος στο Γουισκόνσιν ή όχι
snap_TX: λαμβάνει τιμές 0 (όχι) ή 1 (ναι), ανάλογα με το αν είναι ημέρα του προγράμματος στο Τέξας ή όχι

sell_price: η τιμή στην οποία πωλήθηκαν τα προϊόντα

Οι στήλες που δημιουργήθηκαν.

sold: το πλήθος των προϊόντων που πωλήθηκαν
revenue: μας δίνει τα έσοδα, υπολογίζοντας το γινόμενο του αριθμού των προϊόντων που πωλήθηκαν [sold] επί την τιμή του προϊόντος [sell_price]
rolling_mean: είναι ο κινητός μέσος όρος των προϊόντων που πωλούνται σε διάστημα μίας εβδομάδας
expanding_mean: εδώ υπολογίζουμε το μέσο όρο απλά κάθε φορά προσθέτουμε επιπλέον τιμές
Η διαφορά του expanding με το rolling είναι ότι το μήκος του rolling_mean παραμένει σταθερό, υπολογίζοντας τον μέσο όρο από τις πιο πρόσφατες τιμές και αγνοώντας τις παλαιότερες, ενώ στο expanding_mean κάθε φορά προστίθεται μια νέα τιμή και έπειτα υπολογίζεται ο μέσος όρος.
avg: ο μέσος όρος των πωλήσεων ανά προϊόν, τμήμα, κατηγορία προϊόντος, κατάστημα και πολιτεία
daily_avg: ο μέσος όρος των ημερήσιων πωλήσεων ανά προϊόν, τμήμα, κατηγορία προϊόντος, κατάστημα και πολιτεία
selling_trend: είναι η τάση της πώλησης που παρουσιάζεται στα προϊόντα, παίρνει μια θετική τιμή αν οι ημερήσιες πωλήσεις είναι περισσότερες από τον μέσο όρο όλων των πωλήσεων (d_1 – d_1969), ή αρνητική τιμή αν είναι λιγότερες. Υπολογίζεται ως η διαφορά του daily_avg – avg.
store_avg: ο μέσος όρος πωλήσεων ανά κατάστημα
item_avg: ο μέσος όρος πωλήσεων του κάθε προϊόντος
store_item_avg: ο μέσος όρος πωλήσεων ανά κατάστημα και προϊόν

item_first_sale: πριν πόσες ημέρες έγινε η πρώτη πώληση του προϊόντος
item_last_sale: πριν πόσες ημέρες έγινε η τελευταία πώληση του προϊόντος
item_shop_first_sale: πριν πόσες ημέρες έγινε η πρώτη πώληση του προϊόντος, ανά κατάστημα
item_shop_last_sale: πριν πόσες ημέρες έγινε η τελευταία πώληση του προϊόντος, ανά κατάστημα

dept_avg: ο μέσος όρος πωλήσεων ανά τμήμα
state_avg: ο μέσος όρος πωλήσεων ανά πολιτεία
state_store_avg: ο μέσος όρος πωλήσεων ανά κατάστημα, πολιτεία
cat_avg: ο μέσος όρος πωλήσεων ανά κατηγορία
dept_item_avg: ο μέσος όρος πωλήσεων ανά προϊόν στο κάθε κατάστημα
cat_item_avg: ο μέσος όρος πωλήσεων ανά προϊόν σε κάθε κατηγορία
cat_dept_avg: ο μέσος όρος πωλήσεων ανά προϊόν σε κάθε τμήμα
state_shop_cat_avg: ο μέσος όρος πωλήσεων ανά πολιτεία, κατάστημα και κατηγορία
store_cat_dept_avg: ο μέσος όρος πωλήσεων ανά κατάστημα, κατηγορία και τμήμα

Δημιουργήθηκαν **lag features** για την μεταβλητή **sold** με τιμές [1,2,3,6,13,30] για να πάρουμε τις πωλήσεις αντίστοιχα την προηγούμενη ημέρα, την παραπροηγούμενη, 3 ημέρες πριν, την ίδια ημέρα μία εβδομάδα πριν, την ίδια ημέρα δύο εβδομάδες πριν και την ίδια ημέρα πριν από ένα μήνα αντίστοιχα. Να επισημάνουμε ότι τα lag features είναι ένας κλασικός τρόπος όπου τα προβλήματα χρονοσειρών μετατρέπονται σε προβλήματα εποπτευόμενης μάθησης.
Επίσης από μαθηματικής σκοπιάς, η μέση τιμή αντιπροσωπεύει μια πιθανότητα της μεταβλητής στόχου. Κατά κάποιο τρόπο, η μέση τιμή ενσωματώνει τη μεταβλητή στόχο στην κωδικοποιημένη τιμή της.
Επιπλέον σε επόμενα πειράματα δημιουργήθηκαν οι ακόλουθες μεταβλητές, με τις οποίες το σφάλμα RMSE ανέβηκε, θεωρήθηκε ότι δημιουργήθηκε επιπλέον περιττή και μη σχετική πληροφορία, οπότε δεν λήφθηκαν υπόψη στη διαδικασία των πειραμάτων. Οι μεταβλητές αυτές είναι:

rolling_revenue_mean: είναι ο κινητός μέσος όρος των εσόδων σε διάστημα μίας εβδομάδας
expanding_revenue_mean: εδώ υπολογίζουμε το μέσο όρο των εσόδων, απλά κάθε φορά προσθέτουμε επιπλέον τιμές
monthly_avg: ο μέσος όρος των μηνιαίων πωλήσεων ανά προϊόν, τμήμα, κατηγορία προϊόντος, κατάστημα και πολιτεία
monthly_selling_trend: είναι η μηνιαία τάση της πώλησης που παρουσιάζεται στα προϊόντα, παίρνει μια θετική τιμή αν οι μηνιαίες πωλήσεις είναι περισσότερες από

τον μέσο όρο όλων των πωλήσεων (d_1 – d_1969), ή αρνητική τιμή αν είναι λιγότερες. Υπολογίζεται ως η διαφορά του monthly _avg – avg.

4.8. Αποτελέσματα και ερμηνεία

4.8.1. Εισαγωγή

Τα μοντέλα που παρουσιάζονται παρακάτω αφορούν την πρόβλεψη των πωλήσεων για τις επόμενες 28 ημέρες, των προϊόντων για την αμερικάνικη πολυεθνική εταιρία λιανικού εμπορίου Walmart. Λόγω του αρκετά μεγάλου όγκου δεδομένων, 59.181.090 γραμμές x 22 στήλες, αποφασίστηκε να δημιουργηθούν τέσσερα προβλεπτικά μοντέλα για τις πωλήσεις των προϊόντων στις επόμενες 28 ημέρες, αντί να εργαστούμε με μια ολοκληρωτική προσέγγιση, στο σύνολο δεδομένων. Τα μοντέλα αφορούν προβλέψεις για τις πωλήσεις προϊόντων ανά πολιτεία, ανά κατάστημα, ανά κατηγορία προϊόντος και ανά τμήμα, όπου για κάθε ένα από αυτά πήραμε αποτελέσματα από 3 εκδόσεις. Η πρώτη έκδοση, η οποία μπορεί να χαρακτηριστεί ως βασική (initial), περιλαμβάνει μόνο τις μεταβλητές – στήλες του αρχικού συνόλου δεδομένων. Η δεύτερη έκδοση περιέχει όλες τις επιπλέον (extra) μεταβλητές που δημιουργήσαμε και ενσωματώθηκαν στο αρχικό σύνολο δεδομένων. Και η τρίτη έκδοση περιλαμβάνει τις μεταβλητές, αυτές δηλαδή που συνέβαλαν όπως φαίνεται και στα παρακάτω γραφήματα στην εκπαίδευση του μοντέλου.
Τα προβλεπτικά μοντέλα υλοποιήθηκαν με τη χρήση μηχανικής μάθησης και του αλγορίθμου LightGBM, με την ενσωμάτωση και χρήση των επιπρόσθετων μεταβλητών.

Τα αποτελέσματα αξιολογήθηκαν με τη χρήση της μετρικής σφαλμάτων RMSE (Root Mean Square Error) – Ρίζα του μέσου τετραγωνικού σφάλματος.
Συνολικά δημιουργήθηκαν και αξιολογήθηκαν 69 πειράματα των οποίων τα αποτελέσματα και τα σκορ που λάβαμε ήταν τα ακόλουθα:

4.8.2. Αποτελέσματα ανά πολιτεία (state)

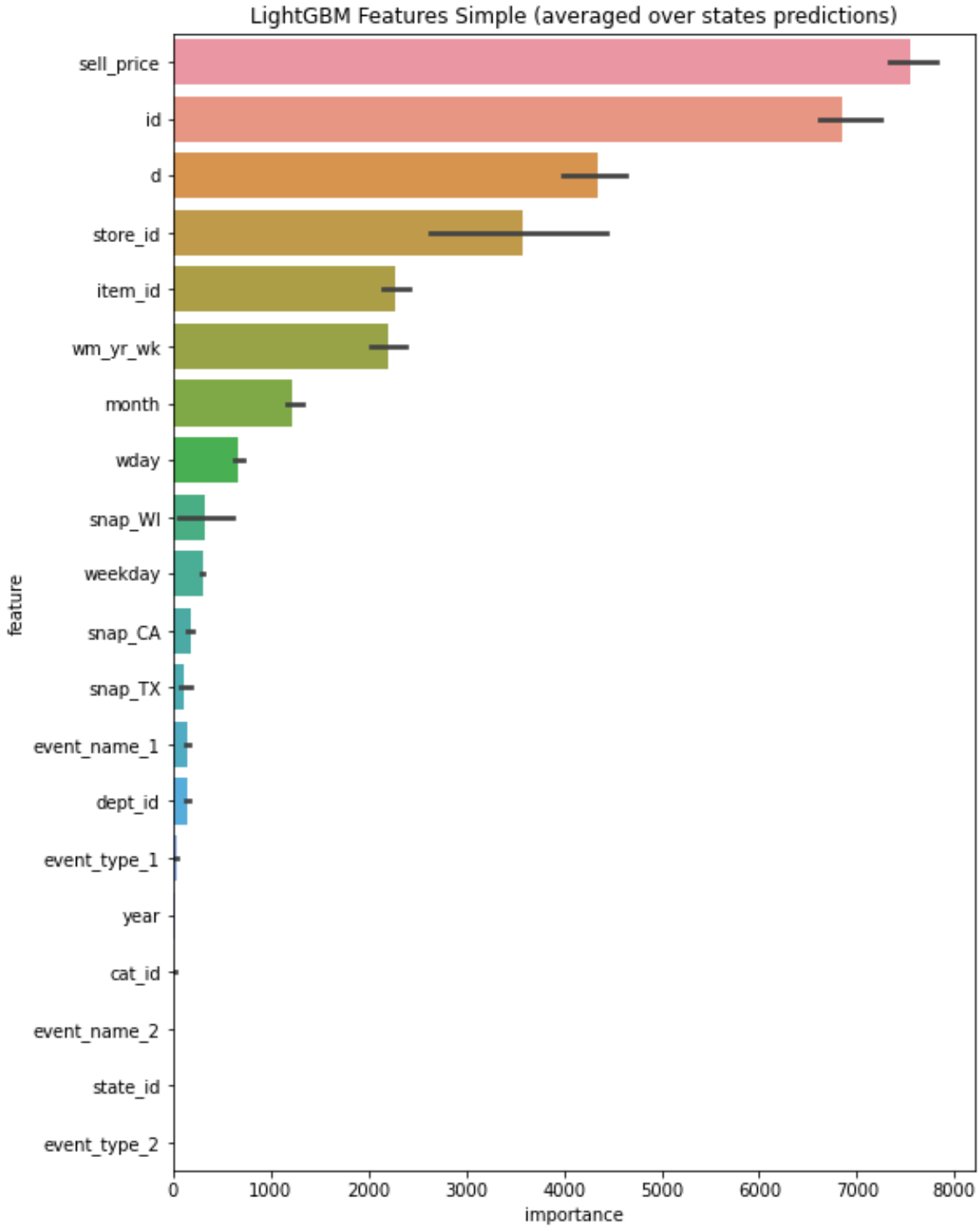
RSME	initial	extra	important
CA	2.45	0.23	0.25
TX	2.18	0.25	0.24
WI	2.18	0.35	0.31

Πίνακας 1: RMSE σκορ ανά πολιτεία

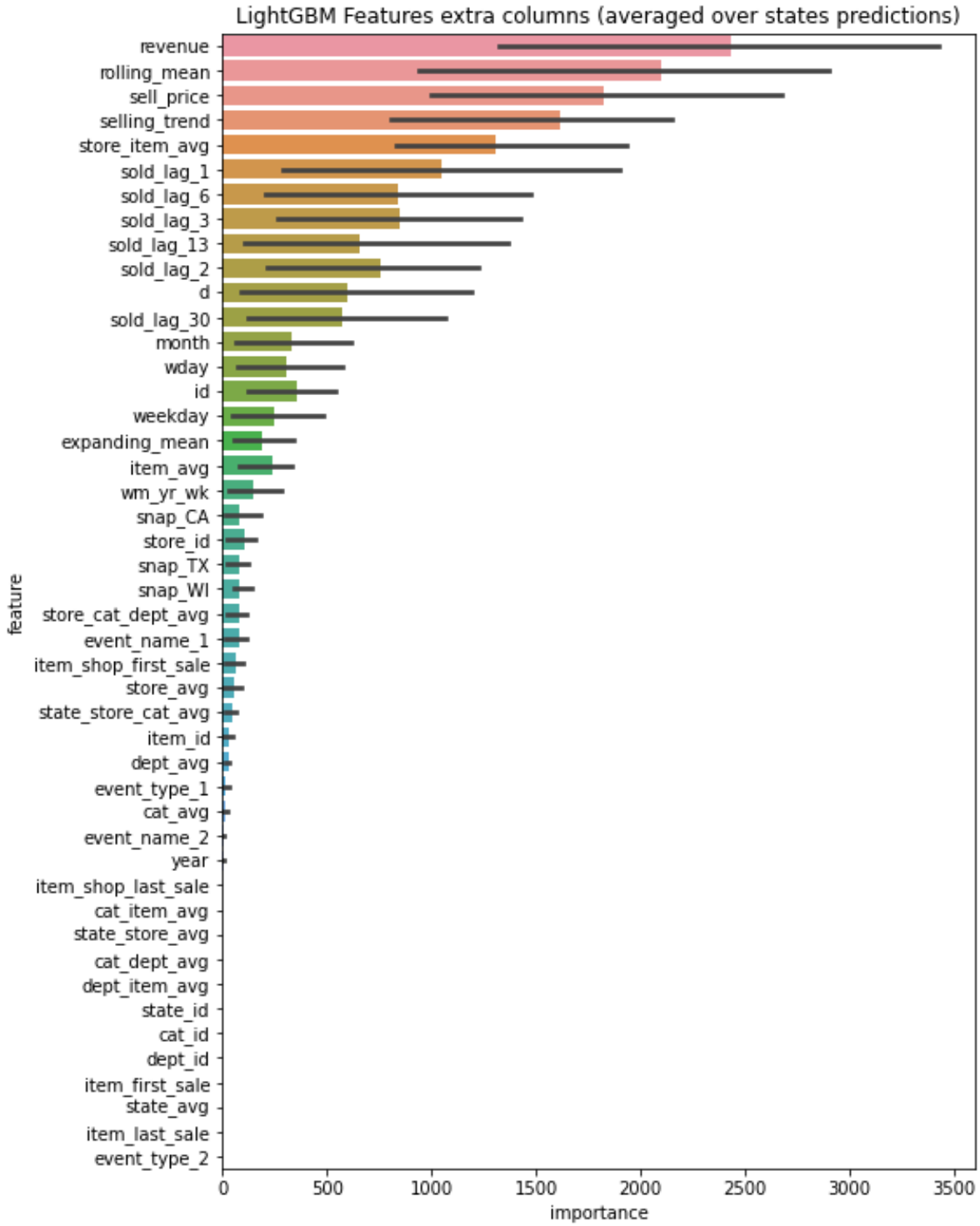
Το αρχικό – βασικό μοντέλο (initial) δεν παρουσιάζει αρκετά ικανοποιητικές τιμές με βάση τη μετρική RSME. Όπως φαίνεται και στον παραπάνω πίνακα για τα καταστήματα της πολιτείας της Καλιφόρνιας CA το σφάλμα παίρνει την τιμή 2.46, για το Τέξας TX 2.18 και για το Γουισκόνσιν WI 2.18 .

Με την εισαγωγή των επιπρόσθετων (extra) μεταβλητών που δημιουργήθηκαν, παρουσιάζονται και αναλύονται στην ενότητα 4.7.1 (Δημιουργία Προβλεπτικών Μεταβλητών), τα σκορ βελτιωθήκανε αρκετά με τις τιμές να διαμορφώνονται ανάλογα, CA 0.23, TX 0.25, WI 0.35 . Οι μεταβλητές που συντελέσανε στη δημιουργία των μοντέλων φαίνονται επίσης και στα παρακάτω γραφήματα. Στη συνέχεια εκτελέστηκε και ένα τρίτο πείραμα (important), λαμβάνοντας υπόψη μόνο τις μεταβλητές που είχαν τη μεγαλύτερη επίδραση στο μοντέλο και αφαιρώντας τις μεταβλητές που είχαν ελάχιστη έως και καθόλου επίδραση. Παρατηρούμε ότι με την εισαγωγή των επιπλέον μεταβλητών, στο δεύτερο πείραμα , το σφάλμα RMSE μειώθηκε αισθητά, οι σημαντικότερες μεταβλητές που συντέλεσαν στη δημιουργία του μοντέλου ήταν οι revenue, rolling_mean, selling_trend, store_item_avg, καθώς και τα lag features. Από της αρχικές στήλες που υπήρχαν στο σύνολο δεδομένων αρκετά μεγάλη σημασία είχε η sell_price, και ακολουθούν σε σημαντικότητα ημερολογιακές μεταβλητές όπως η ημέρα d, ο μήνας month, η ημέρα της εβδομάδας wday, weekday, wm_yr_wk. Οι στήλες snap_CA, snap_WI και snap_TX δεν έχουν και τόσο σημαντικό ρόλο στη δημιουργία του μοντέλου. Στο τρίτο πείραμα (important) αφαιρέθηκαν οι μεταβλητές item_shop_last_sale, cat_item_avg, state_store_avg, cat_dept_avg, state_id,cat_id, dept_id, item_first_sale, state_avg, item_last_sale, event_type_2 και κρατήθηκαν αυτές μόνο που έχουν τη μεγαλύτερη επίδραση λάβαμε ελαφρώς καλύτερο σκορ για τις πολιτείες TX και WI ενώ ελαφρώς χειρότερο για την πολιτεία CA.

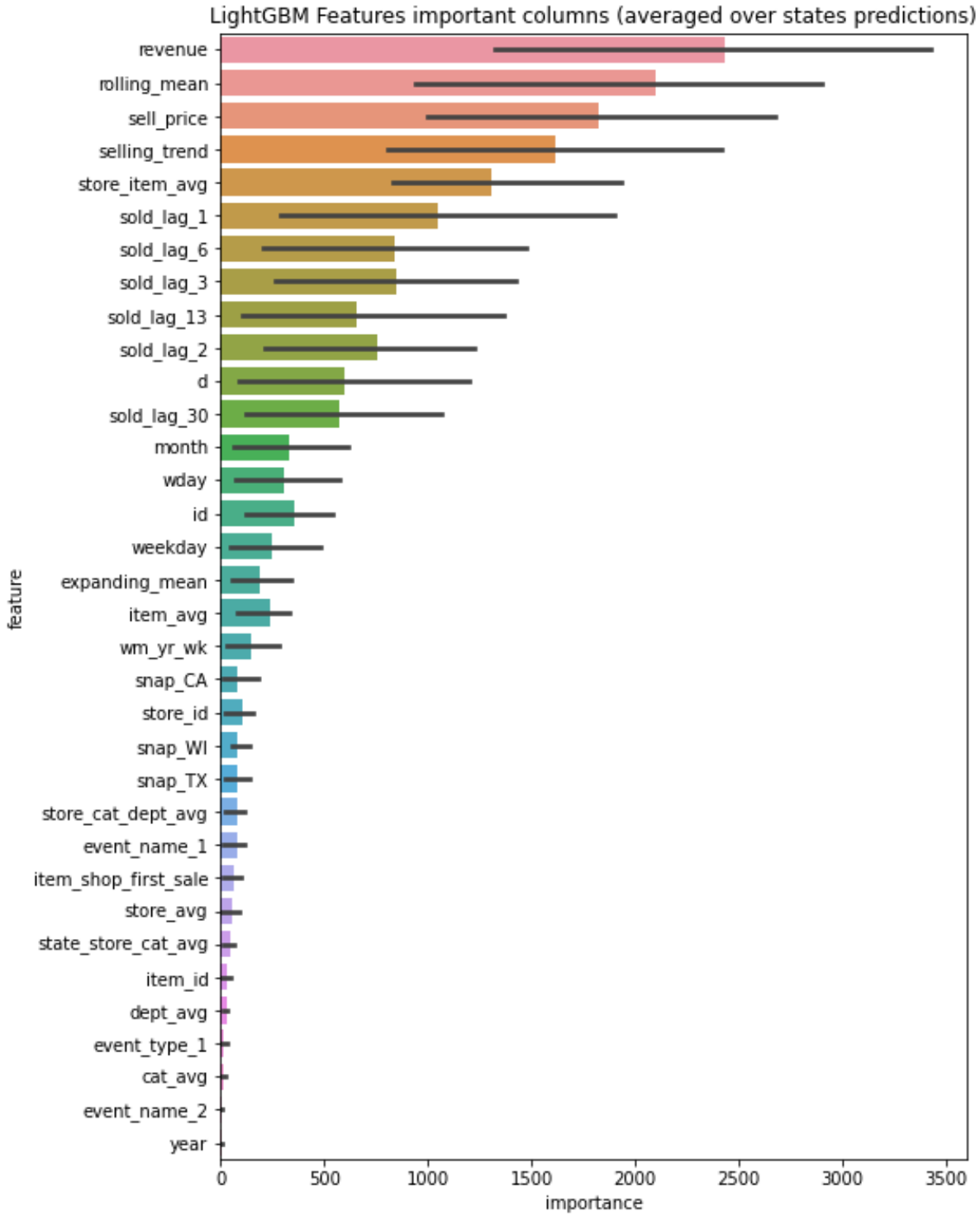
Στα παρακάτω γραφήματα παρουσιάζεται η σημαντικότητα κάθε μεταβλητής του συγκεκριμένου προβλεπτικού μοντέλου ανά πολιτεία.



Εικόνα 20: Η επίδραση μεταβλητών στη δημιουργία των μοντέλων ανά πολιτεία χωρίς την προσθήκη επιπλέον μεταβλητών



Εικόνα 21: Η επίδραση όλων των μεταβλητών στη δημιουργία των μοντέλων ανά πολιτεία



Εικόνα 22: Η επίδραση των σημαντικότερων μεταβλητών στη δημιουργία των μοντέλων ανά πολιτεία

4.8.3. Αποτελέσματα ανά κατάστημα (store)

RMSE	initial	extra	important
CA_1	2.56	0.30	0.32
CA_2	2.47	0.24	0.24
CA_3	3.09	0.34	0.35
CA_4	1.58	0.11	0.14
TX_1	2.08	0.20	0.20
TX_2	2.24	0.24	0.25
TX_3	2.42	0.33	0.33
WI_1	1.97	0.13	0.12
WI_2	3.54	0.39	0.41
WI_3	2.33	0.28	0.28

Πίνακας 2: RMSE σκορ ανά κατάστημα

Όπως και στα πειράματα τα οποία εκτελέστηκαν ανά πολιτεία, έτσι και στα πειράματα ανά κατάστημα το αρχικό – βασικό μοντέλο (initial) δεν παρουσιάζει ικανοποιητικές τιμές με βάση τη μετρική RSME. Σύμφωνα με τον παραπάνω πίνακα για τα καταστήματα της πολιτείας της Καλιφόρνιας CA_1, CA_2, CA_3 και CA_4, τα σφάλματα πήραν αντίστοιχα τις τιμές 2.56, 2.47, 3.09, 1.58 για τα καταστήματα της πολιτείας του Τέξας TX_1, TX_2 και TX_3 τα σφάλματα ήταν αντίστοιχα 2.08, 2.24, 2.42 και για τα καταστήματα της πολιτείας Γουισκόνσιν WI_1, WI_2, WI_3 τα σφάλματα ήταν αντίστοιχα 1.97, 3.54, 2.33 .

Με την εισαγωγή των επιπρόσθετων (extra) μεταβλητών που δημιουργήθηκαν, παρουσιάζονται και αναλύονται στην ενότητα 4.7.1 (Δημιουργία Προβλεπτικών Μεταβλητών), τα σκορ βελτιωθήκανε αρκετά με τις τιμές να διαμορφώνονται ανάλογα CA_1 0.30, CA_2 0.24, CA_3 0.34, CA_4 0.11, TX_1 0.20, TX_2 0.24, TX_3 0.33, WI_1 0.13, WI_2 0.39 και WI_3 0.28, όπου όπως παρατηρείται η εισαγωγή νέων μεταβλητών συνέβαλε αρκετά στη μείωση των σφαλμάτων, επομένως και στη βελτίωση των προβλεπτικών μοντέλων, με το μοντέλο για το κατάστημα CA_4 να παρουσιάζει το χαμηλότερο σκορ 0.11, επομένως και την καλύτερη απόδοση. Οι μεταβλητές που συντελέσανε στη δημιουργία των μοντέλων φαίνονται επίσης και στα παρακάτω γραφήματα. Στη συνέχεια εκτελέστηκε και ένα τρίτο πείραμα (important), κρατώντας μόνο τις μεταβλητές που είχαν τη μεγαλύτερη

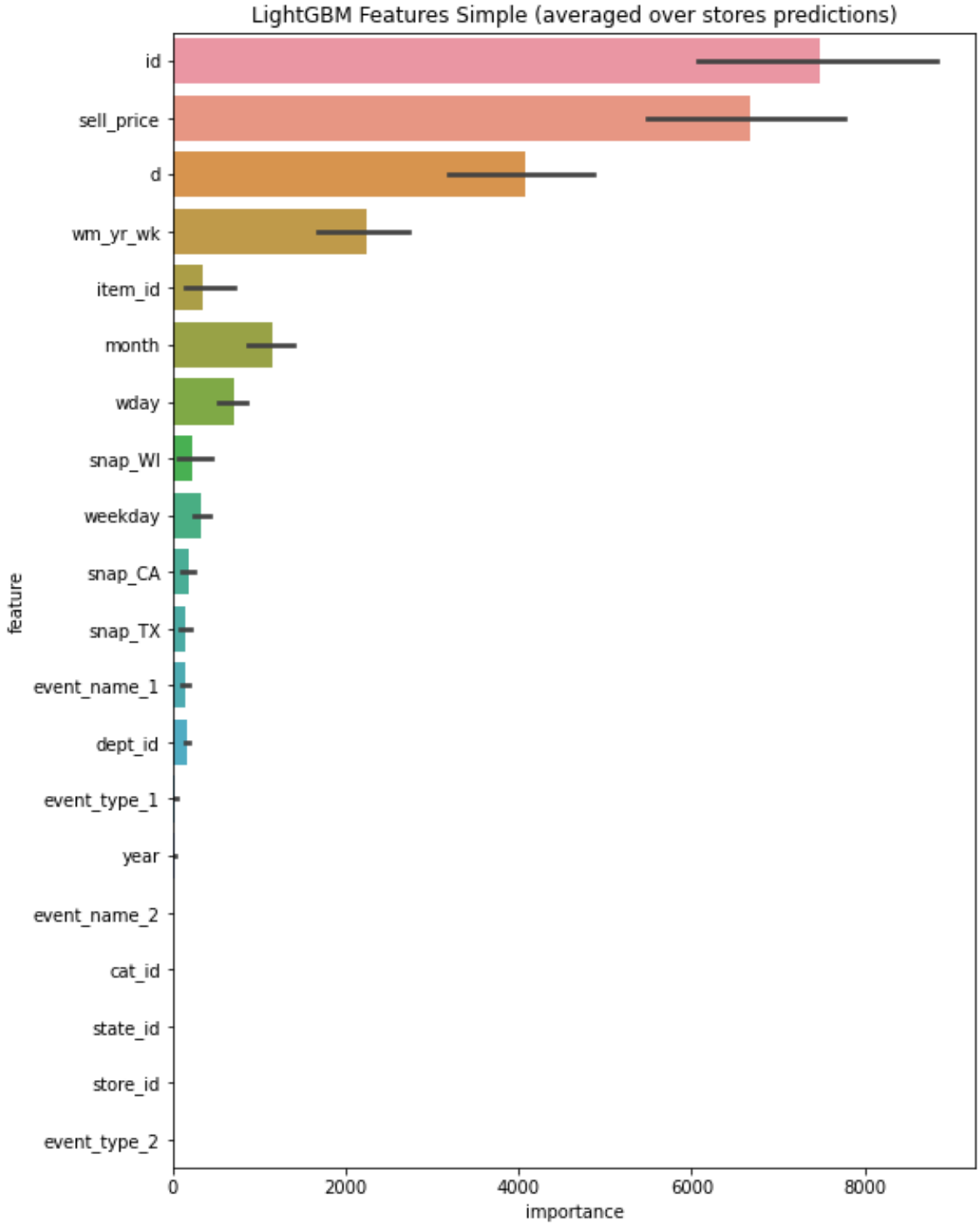
επίδραση στο μοντέλο και αφαιρώντας τις μεταβλητές που είχαν ελάχιστη έως και καθόλου επίδραση. Παρατηρούμε και σε αυτή την περίπτωση ότι με την εισαγωγή των επιπλέον μεταβλητών, στο δεύτερο πείραμα, το σφάλμα RMSE μειώθηκε αισθητά, οι σημαντικότερες μεταβλητές που συντέλεσαν στη δημιουργία του μοντέλου ήταν και εδώ οι revenue, rolling_mean, selling_trend, store_item_avg, καθώς και τα lag features. Από της αρχικές στήλες που υπήρχαν στο σύνολο δεδομένων αρκετά μεγάλη σημασία είχε η sell_price, και ακολουθούν σε σημαντικότητα ημερολογιακές μεταβλητές όπως η ημέρα d, ο μήνας month, η ημέρα της εβδομάδας wday, weekday, wm_yr_wk. Οι στήλες snap_CA, snap_WI και snap_TX δεν έχουν και εδώ τόσο σημαντικό ρόλο στη δημιουργία του μοντέλου. Στο τρίτο πείραμα (important) αφαιρέθηκαν οι μεταβλητές item_shop_last_sale, state_avg ,cat_item_avg, state_store_avg, dept_item_avg, , state_store_cat_avg, state_id, cat_id, store_id, store_avg, cat_dept_avg, state_id,cat_id, dept_id, item_first_sale, , item_last_sale, event_type_2 και κρατήθηκαν αυτές μόνο που έχουν τη μεγαλύτερη επίδραση λάβαμε ελαφρώς καλύτερο σκορ για το κατάστημα WI_1 ενώ για όλα τα υπόλοιπα καταστήματα στα σκορ παραμείνανε ίδια ή και ελαφρώς χειρότερα. Από αυτό συμπεραίνουμε ότι ορισμένες οι μεταβλητές που αφαιρεθήκαν αν και είχαν μικρή ή και καθόλου επίδραση στο μοντέλο, το βελτίωναν έστω και στο ελάχιστο

Στην παρακάτω εικόνα φαίνονται οι τιμές που παίρνει η μετρική της ρίζας του τετραγωνικού μέσου σφάλματος RMSE, και το σφάλμα l2, κατά τη διάρκεια της εκπαίδευσης του μοντέλου. Στην εικόνα παρουσιάζονται μόνο οι τιμές κατά την εκπαίδευση του μοντέλου για το κατάστημα CA_1 της πολιτείας της Καλιφόρνια.

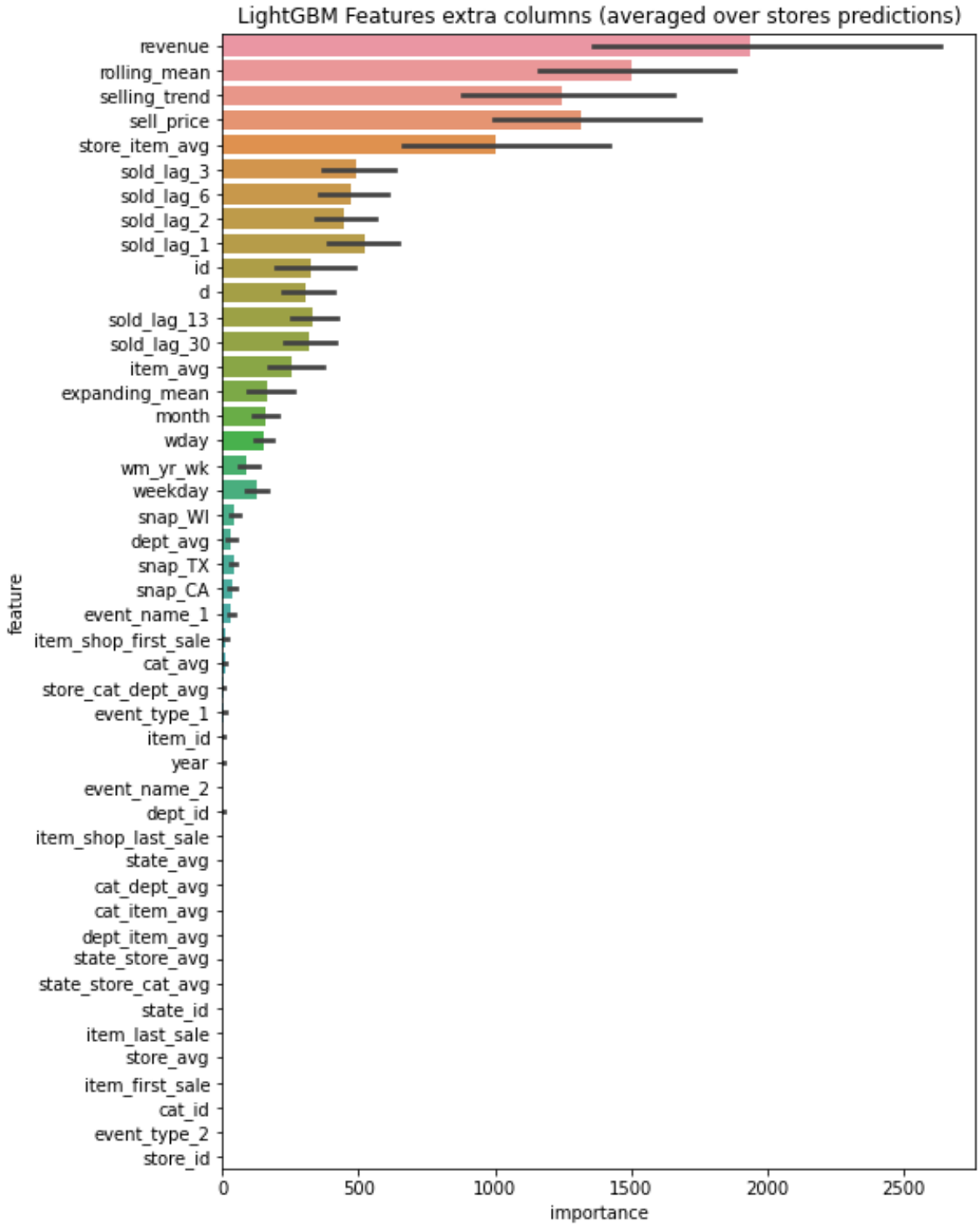
```
*****Prediction for Store: CA_1*****
(5918109, 46)
Training until validation scores don't improve for 20 rounds
[20]    training's rmse: 0.913741       training's l2: 0.834922 valid_1's rmse: 0.724213        valid_1's l2: 0.524485
[40]    training's rmse: 0.59342        training's l2: 0.352148 valid_1's rmse: 0.416191        valid_1's l2: 0.173215
[60]    training's rmse: 0.516879       training's l2: 0.267164 valid_1's rmse: 0.348493        valid_1's l2: 0.121447
[80]    training's rmse: 0.470372       training's l2: 0.221249 valid_1's rmse: 0.331978        valid_1's l2: 0.110209
[100]   training's rmse: 0.439209       training's l2: 0.192905 valid_1's rmse: 0.32155 valid_1's l2: 0.103394
[120]   training's rmse: 0.410073       training's l2: 0.16816  valid_1's rmse: 0.317316        valid_1's l2: 0.10069
[140]   training's rmse: 0.387564       training's l2: 0.150206 valid_1's rmse: 0.316434        valid_1's l2: 0.100131
[160]   training's rmse: 0.371669       training's l2: 0.138138 valid_1's rmse: 0.311924        valid_1's l2: 0.0972964
[180]   training's rmse: 0.354504       training's l2: 0.125673 valid_1's rmse: 0.306018        valid_1's l2: 0.0936472
[200]   training's rmse: 0.340339       training's l2: 0.115831 valid_1's rmse: 0.30467 valid_1's l2: 0.0928238
[220]   training's rmse: 0.325919       training's l2: 0.106223 valid_1's rmse: 0.305729        valid_1's l2: 0.0934702
[240]   training's rmse: 0.314236       training's l2: 0.098744 valid_1's rmse: 0.303924        valid_1's l2: 0.09237
[260]   training's rmse: 0.304749       training's l2: 0.0928719        valid_1's rmse: 0.302709        valid_1's l2: 0.091633
Early stopping, best iteration is:
[254]   training's rmse: 0.307666       training's l2: 0.0946581        valid_1's rmse: 0.301089        valid_1's l2: 0.0906544
[-4.06540312e-03  1.24795470e-03  7.25762902e-04 ...  1.01849289e+00
  9.61070773e-01  4.93740277e+00]
RMSE Test score: 0.31303084999918956
```

Εικόνα 23: Πορεία υπολογισμού σφαλμάτων κατά την εκπαίδευση

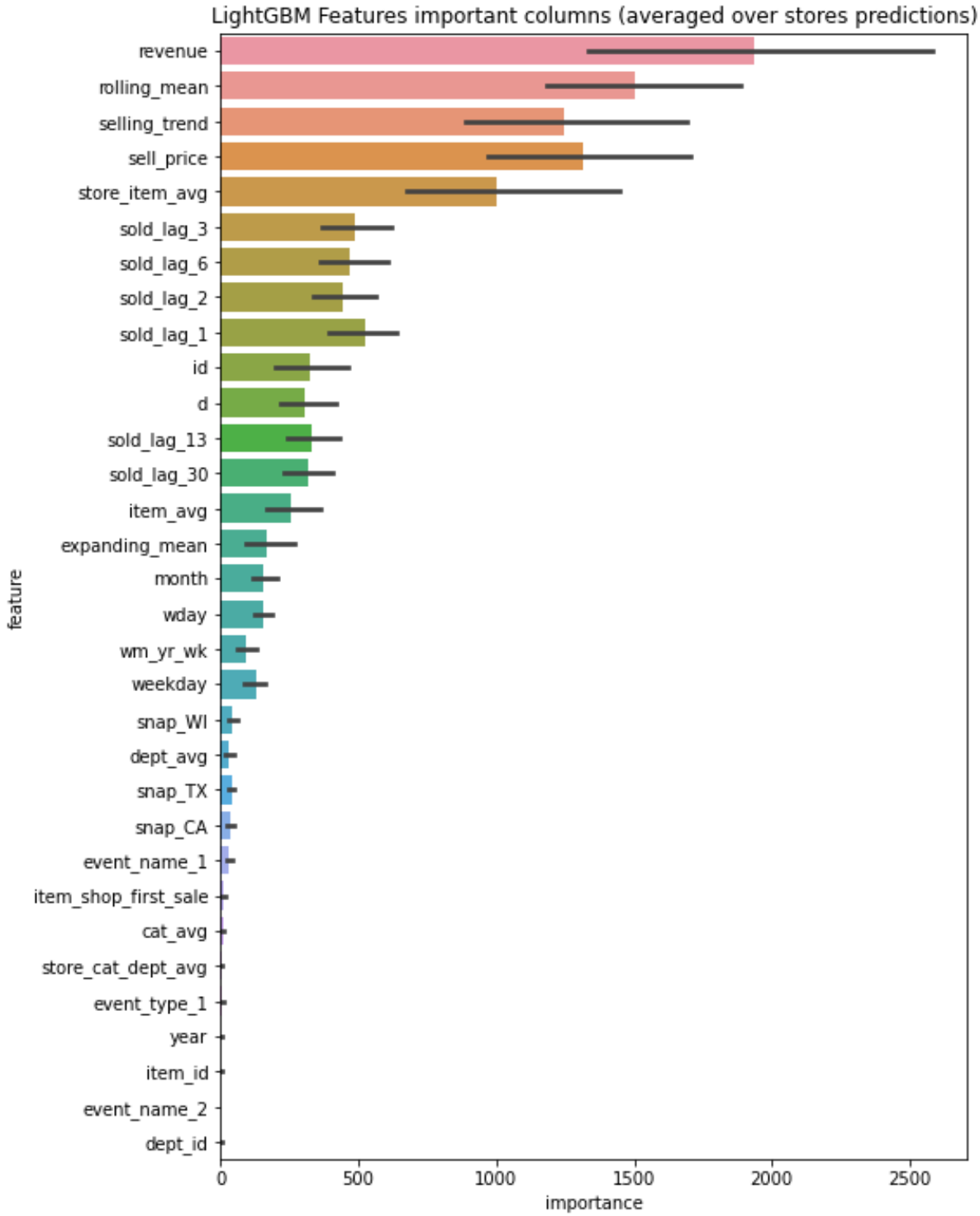
Στα παρακάτω γραφήματα παρουσιάζεται η σημαντικότητα κάθε μεταβλητής του συγκεκριμένου προβλεπτικού μοντέλου ανά κατάστημα.



Εικόνα 24: Η επίδραση μεταβλητών στη δημιουργία των μοντέλων ανά κατάστημα χωρίς την προσθήκη επιπλέον μεταβλητών



Εικόνα 25: Η επίδραση όλων των μεταβλητών στη δημιουργία των μοντέλων ανά κατάστημα



Εικόνα 26: Η επίδραση των σημαντικότερων μεταβλητών στη δημιουργία των μοντέλων ανά κατάστημα

4.8.4. Αποτελέσματα ανά κατηγορία προϊόντος (category)

RMSE	initial	extra	important
HOBBIES	1.65	0.12	0.11
HOUSEHOLD	1.56	0.17	0.17
FOODS	3.08	0.36	0.38

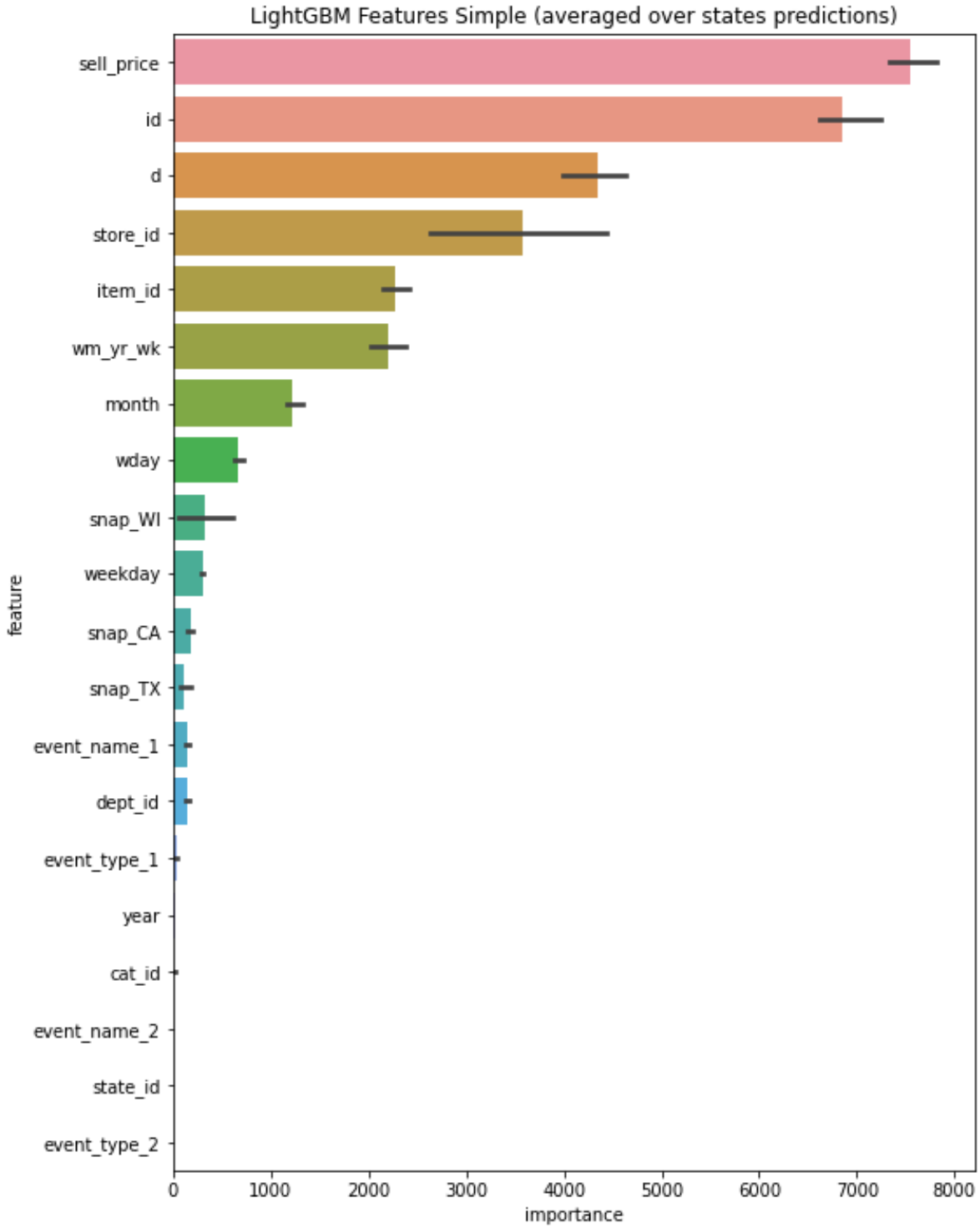
Πίνακας 3: RMSE σκορ ανά κατηγορία προϊόντος

Όπως και στα πειράματα τα οποία εκτελέστηκαν ανά πολιτεία και ανά κατάστημα, έτσι και στα πειράματα ανά κατηγορία προϊόντος το αρχικό – βασικό μοντέλο (initial) δεν παρουσιάζει ικανοποιητικές τιμές με βάση τη μετρική RSME. Σύμφωνα με τον παραπάνω πίνακα για την κατηγορία HOBBIES το σφάλμα ήταν 1.65, για την κατηγορία HOUSEHOLD το σφάλμα ήταν 1.56 και για την κατηγορία FOODS το σφάλμα ήταν 3.08.

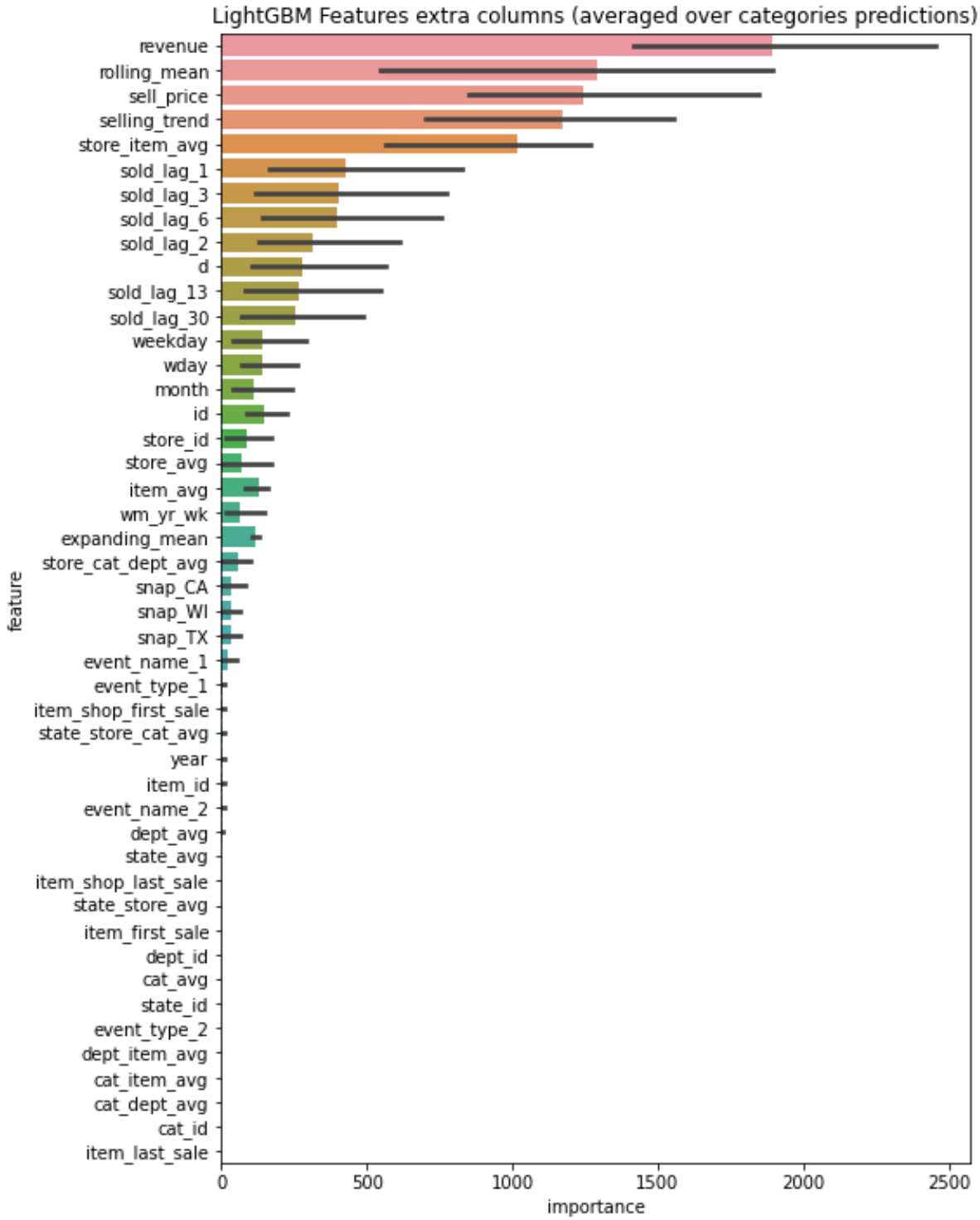
Με την εισαγωγή των επιπρόσθετων (extra) μεταβλητών που δημιουργήθηκαν, παρουσιάζονται και αναλύονται στην ενότητα 4.7.1 (Δημιουργία Προβλεπτικών Μεταβλητών), τα σκορ βελτιωθήκανε και σε αυτή την περίπτωση αρκετά με τις τιμές να διαμορφώνονται ανάλογα HOBBIES 0.21, HOUSEHOLD 0.17 και FOODS 0.36, όπου όπως παρατηρείται η εισαγωγή νέων μεταβλητών συνέβαλε αρκετά στη μείωση των σφαλμάτων, επομένως και στη βελτίωση των προβλεπτικών μοντέλων, με το μοντέλο για την κατηγορία HOBBIES να παρουσιάζει το χαμηλότερο σκορ 0.12, επομένως και την καλύτερη απόδοση. Οι μεταβλητές που συντελέσανε στη δημιουργία των μοντέλων φαίνονται επίσης και στα παρακάτω γραφήματα. Στη συνέχεια εκτελέστηκε και ένα τρίτο πείραμα (important), κρατώντας μόνο τις μεταβλητές που είχαν τη μεγαλύτερη επίδραση στο μοντέλο και αφαιρώντας τις μεταβλητές που είχαν ελάχιστη έως και καθόλου επίδραση. Παρατηρούμε και σε αυτή την περίπτωση, όπως και στις δύο προηγούμενες περιπτώσεις, στα πειράματα ανά πολιτείες και ανά καταστήματα ότι με την εισαγωγή των επιπλέον μεταβλητών, στη δεύτερη κατηγορία πειραμάτων (extra) , το σφάλμα RMSE μειώθηκε αισθητά, οι σημαντικότερες μεταβλητές που συντέλεσαν στη δημιουργία του μοντέλου ήταν και εδώ οι revenue, rolling_mean, selling_trend, store_item_avg, καθώς και τα lag features. Από της αρχικές στήλες που υπήρχαν στο σύνολο δεδομένων αρκετά μεγάλη σημασία είχε η sell_price, και ακολουθούν σε σημαντικότητα ημερολογιακές μεταβλητές όπως η ημέρα d, ο μήνας month, η ημέρα της εβδομάδας wday,

weekday, wm_yr_wk. Οι στήλες snap_CA, snap_WI και snap_TX δεν έχουν και εδώ τόσο σημαντικό ρόλο στη δημιουργία του μοντέλου. Στο τρίτο πείραμα (important) αφαιρέθηκαν οι μεταβλητές item_shop_first_sale, item_shop_last_sale, year, item_id, event_name_2, dept_avg, state_avg ,cat_item_avg, state_store_avg, dept_item_avg, cat_avg, state_store_cat_avg, state_id, cat_id, store_id, store_avg, cat_dept_avg, state_id, cat_id, dept_id, item_first_sale, state_store_avg , item_last_sale, event_type_2 και κρατήθηκαν αυτές μόνο που έχουν τη μεγαλύτερη επίδραση λάβαμε ελαφρώς καλύτερο σκορ για την κατηγορία HOBBIES ενώ για όλα τις υπόλοιπες δυο κατηγορίες HOUSEHOLD και FOODS στα σκορ παραμείνανε ίδια ή και ελαφρώς χειρότερα. Από αυτό συμπεραίνουμε και σε αυτή την περίπτωση ότι οι μεταβλητές που αφαιρεθήκαν αν και είχαν μικρή ή και καθόλου επίδραση στο μοντέλο, το βελτίωναν έστω και στο ελάχιστο.

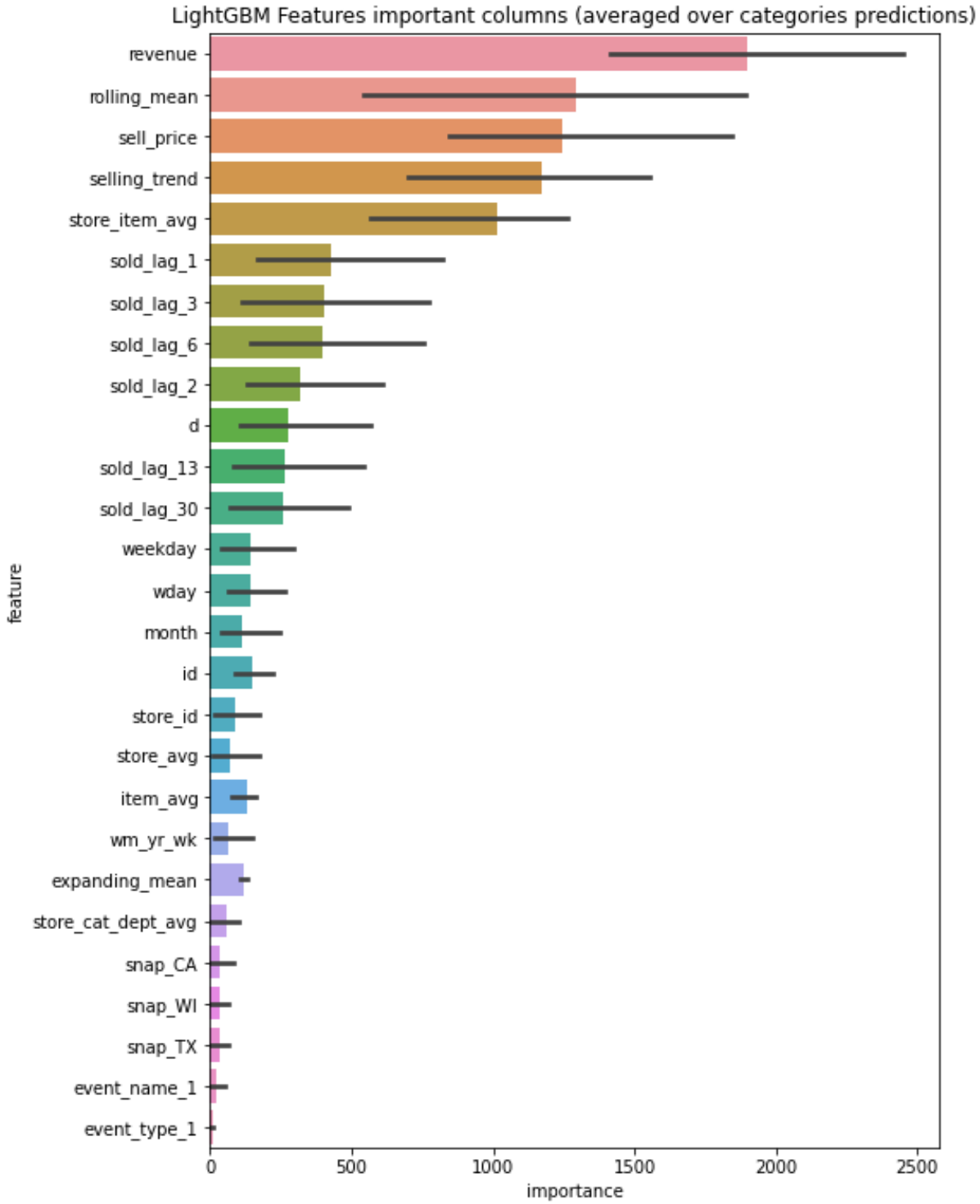
Στα παρακάτω γραφήματα παρουσιάζεται η σημαντικότητα κάθε μεταβλητής του συγκεκριμένου προβλεπτικού μοντέλου ανά κατηγορία προϊόντος.



Εικόνα 27:Η επίδραση μεταβλητών στη δημιουργία των μοντέλων ανά κατηγορία προϊόντων χωρίς την προσθήκη επιπλέον μεταβλητών



Εικόνα 28: Η επίδραση όλων των μεταβλητών στη δημιουργία των μοντέλων ανά κατηγορία



Εικόνα 29: Η επίδραση των σημαντικότερων μεταβλητών στη δημιουργία των μοντέλων ανά κατηγορία

4.8.5. Αποτελέσματα ανά τμήμα (department)

RMSE	initial	extra	important
HOBBIES_1	1.85	0.13	0.13
HOBBIES_2	0.83	0.02	0.02
HOUSEHOLD_1	1.93	0.15	0.15
HOUSEHOLD_2	0.84	0.15	0.15
FOODS_1	2.68	0.32	0.32
FOODS_2	2.03	0.38	0.38
FOODS_3	3.41	0.34	0.34

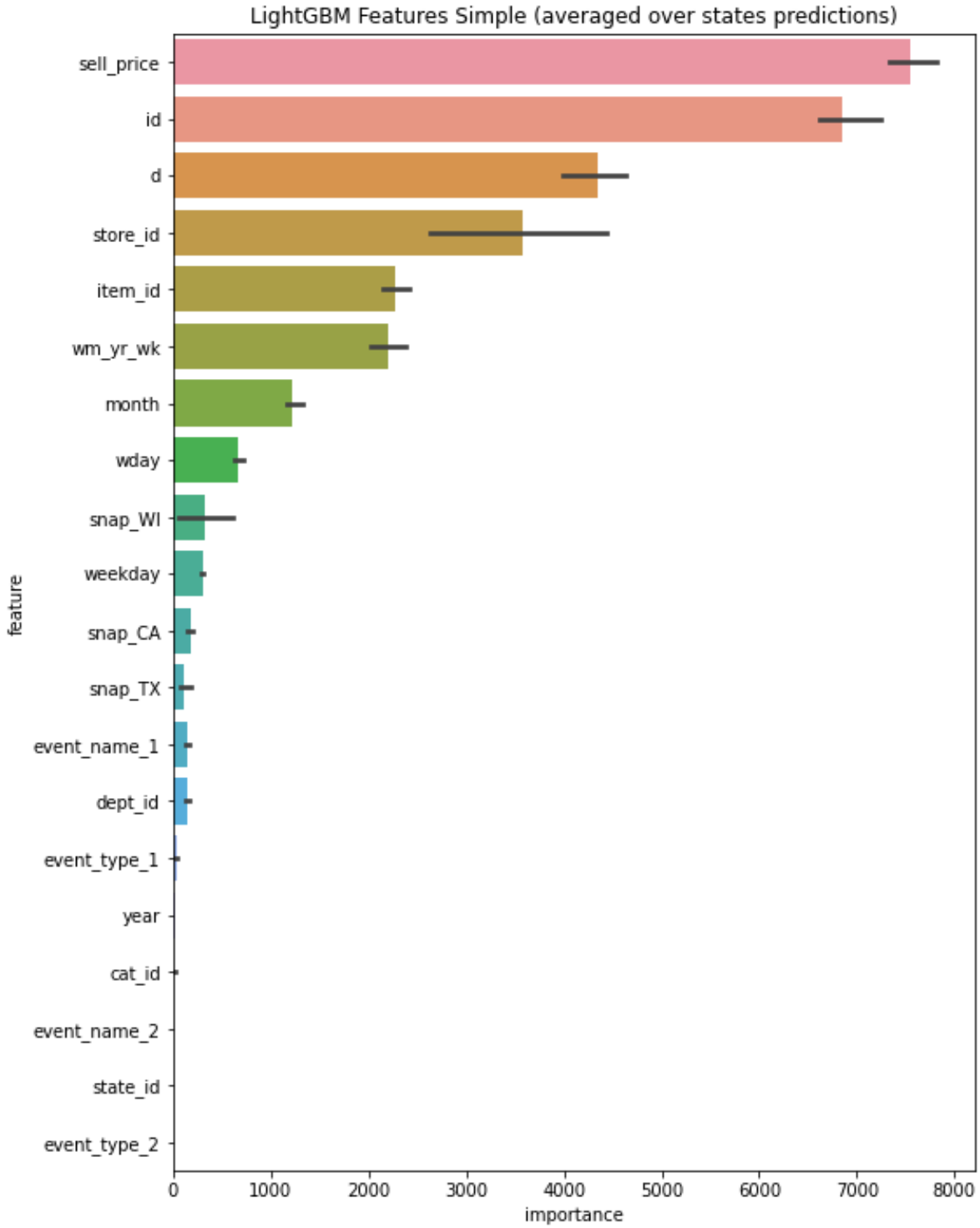
Πίνακας 4: RMSE σκορ ανά τμήμα

Και σε αυτή την περίπτωση όπως και στα πειράματα τα οποία εκτελέστηκαν ανά πολιτεία και ανά κατάστημα και ανά κατηγορία προϊόντος, έτσι και στα πειράματα ανά τμήμα το αρχικό – βασικό μοντέλο (initial) δεν παρουσιάζει αρκετά ικανοποιητικές τιμές με βάση τη μετρική RSME. Σύμφωνα με τον παραπάνω πίνακα για την κατηγορία HOBBIES_1 το σφάλμα ήταν 1.85, HOBBIES_2 το σφάλμα ήταν 0.83 (αρκετά ικανοποιητικό, το οποίο στη συνέχεια βελτιώνεται αρκετά), για την κατηγορία HOUSEHOLD_1 το σφάλμα ήταν 1.93, HOUSEHOLD_2 το σφάλμα ήταν 0.84 (και εδώ αρκετά ικανοποιητικό, το οποίο στη συνέχεια βελτιώνεται αρκετά) και για την κατηγορία FOODS_1 το σφάλμα ήταν 2.68, FOODS_2 το σφάλμα ήταν 2.03, FOODS_3 το σφάλμα ήταν 3.41.

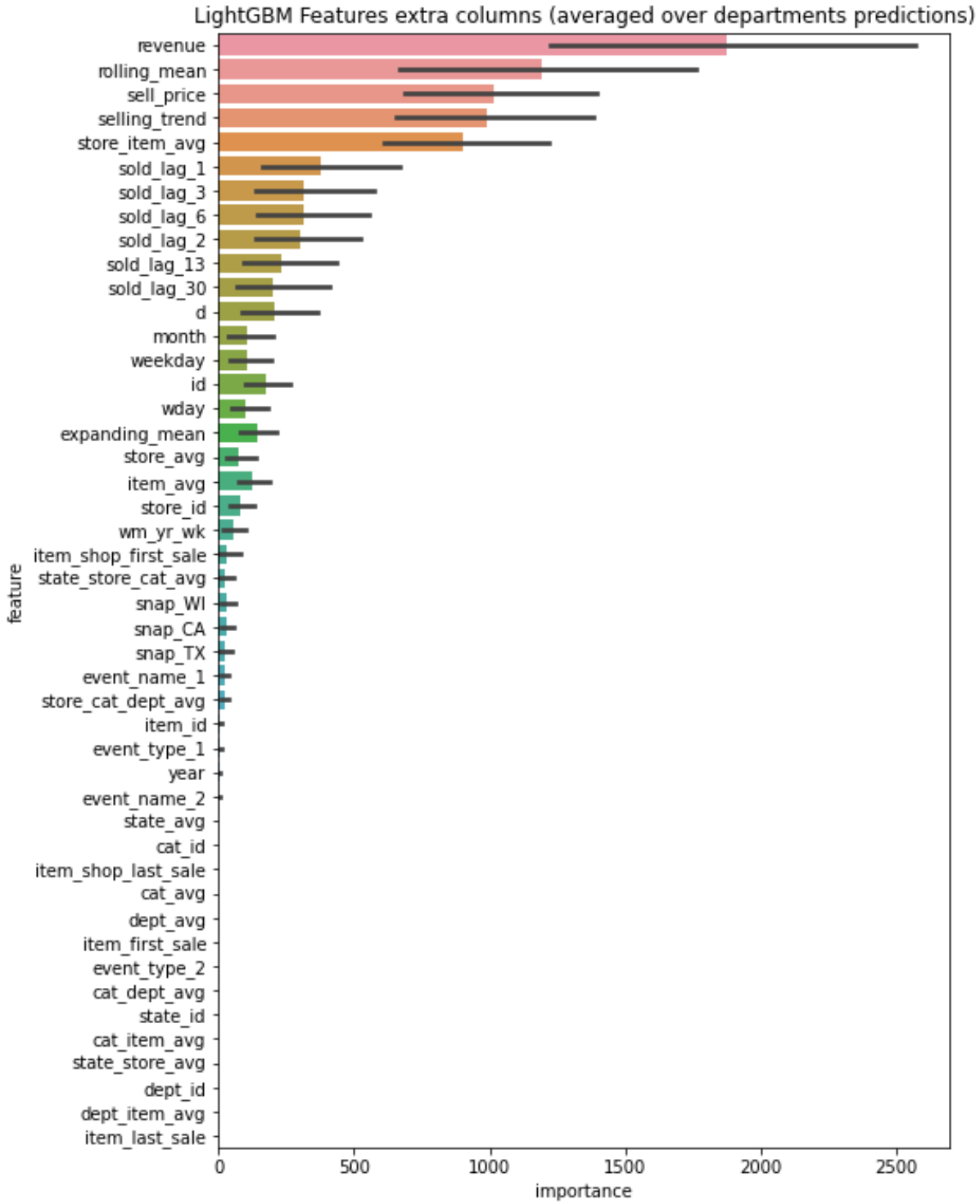
Με την εισαγωγή των επιπρόσθετων (extra) μεταβλητών που δημιουργήθηκαν, παρουσιάζονται και αναλύονται στην ενότητα 4.7.1 (Δημιουργία Προβλεπτικών Μεταβλητών), τα σκορ βελτιωθήκανε και σε αυτή την περίπτωση αρκετά με τις τιμές να διαμορφώνονται ανάλογα HOBBIES_1 0.13, HOBBIES_2 0.02, HOUSEHOLD_1 0.15, HOUSEHOLD_2 0.15 και FOODS_1 0.32, FOODS_1 0.38, FOODS_1 0.34, όπου όπως παρατηρείται η εισαγωγή νέων μεταβλητών συνέβαλε αρκετά στη μείωση των σφαλμάτων, επομένως και στη βελτίωση των προβλεπτικών μοντέλων, με το μοντέλο για την κατηγορία HOBBIES_2 να παρουσιάζει το χαμηλότερο σκορ

0.02, και καλύτερη γενικά απόδοση από όλα τα πειράματα τα οποία εκτελέσαμε. Οι μεταβλητές που συντελέσανε στη δημιουργία των μοντέλων φαίνονται επίσης και στα παρακάτω γραφήματα. Στη συνέχεια εκτελέστηκε και ένα τρίτο πείραμα (important), κρατώντας μόνο τις μεταβλητές που είχαν τη μεγαλύτερη επίδραση στο μοντέλο και αφαιρώντας τις μεταβλητές που είχαν ελάχιστη έως και καθόλου επίδραση. Παρατηρούμε και σε αυτή την περίπτωση, όπως και στις τρεις προηγούμενες περιπτώσεις, στα πειράματα ανά πολιτείες, ανά καταστήματα και ανά κατηγορίες προιόντων ότι με την εισαγωγή των επιπλέον μεταβλητών, στη δεύτερη κατηγορία πειραμάτων (extra), το σφάλμα RMSE μειώθηκε αισθητά, οι σημαντικότερες μεταβλητές που συντέλεσαν στη δημιουργία του μοντέλου ήταν και εδώ οι revenue, rolling_mean, selling_trend, store_item_avg, καθώς και τα lag features. Από της αρχικές στήλες που υπήρχαν στο σύνολο δεδομένων αρκετά μεγάλη σημασία είχε η sell_price, και ακολουθούν σε σημαντικότητα ημερολογιακές μεταβλητές όπως η ημέρα d, ο μήνας month, η ημέρα της εβδομάδας wday, weekday, wm_yr_wk. Οι στήλες snap_CA, snap_WI και snap_TX δεν έχουν και εδώ τόσο σημαντικό ρόλο στη δημιουργία του μοντέλου. Στο τρίτο πείραμα (important) αφαιρέθηκαν οι μεταβλητές item_shop_last_sale, year, item_id, event_name_1, event_name_2, dept_avg, state_avg ,cat_item_avg, state_store_avg, dept_item_avg, cat_avg, state_id, cat_id, cat_dept_avg, state_id, cat_id, dept_id, item_first_sale, state_store_avg , item_last_sale, event_type_2 και κρατήθηκαν αυτές μόνο που έχουν τη μεγαλύτερη επίδραση λάβαμε ακριβώς τι ίδιο σκορ για όλες της στα σκορ παραμείνανε ίδια. Από αυτό συμπεραίνουμε και σε αυτή την περίπτωση ότι οι μεταβλητές που αφαιρεθήκαν αν και είχαν μικρή ή και καθόλου επίδραση στο μοντέλο, δεν επηρέασαν στην αλλαγή της απόδοσης των μοντέλων .

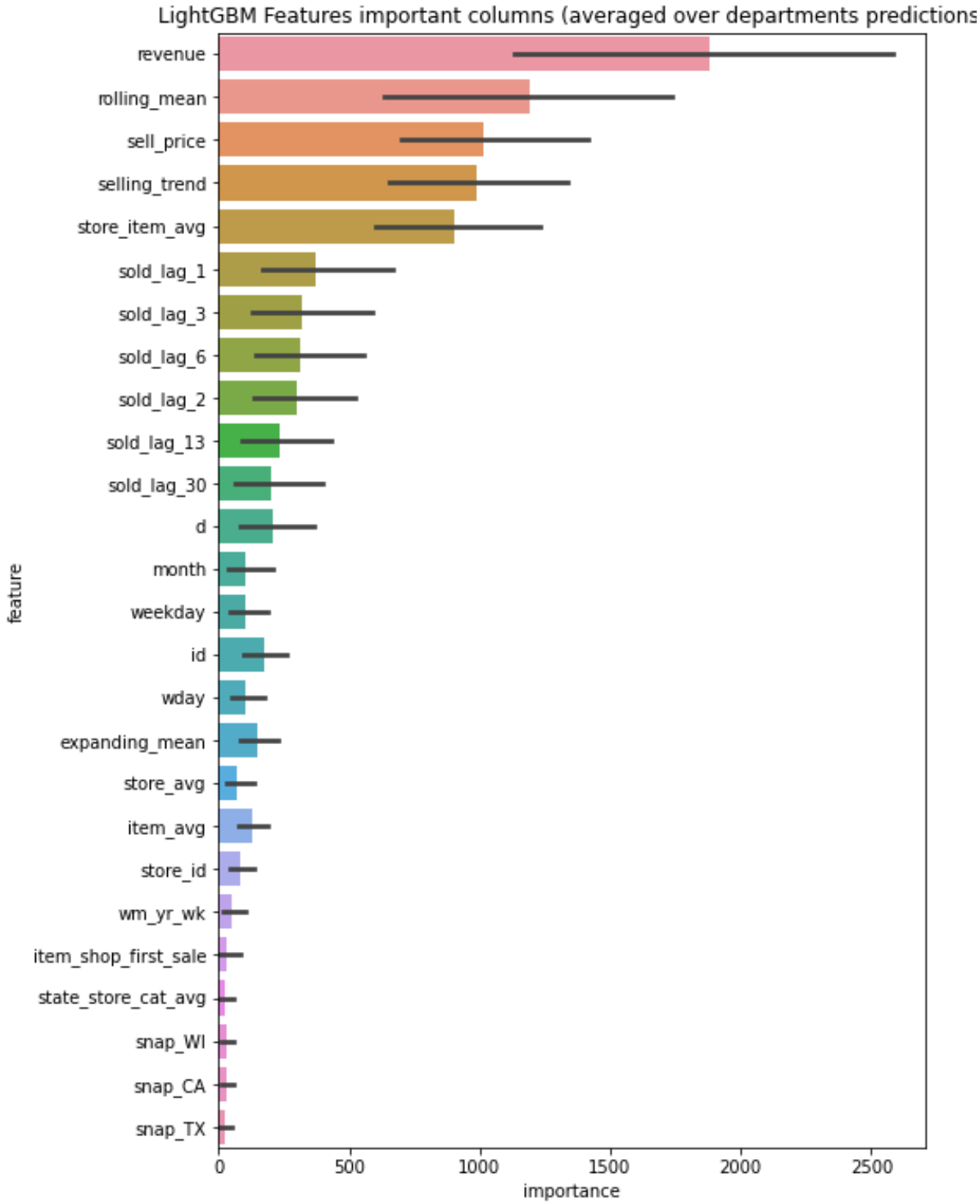
Στα παρακάτω γραφήματα παρουσιάζεται η σημαντικότητα κάθε μεταβλητής του συγκεκριμένου προβλεπτικού μοντέλου ανά κατηγορία προϊόντος.



Εικόνα 30: Η επίδραση μεταβλητών στη δημιουργία των μοντέλων ανά τμήμα χωρίς την προσθήκη επιπλέον μεταβλητών



Εικόνα 31: Η επίδραση όλων των μεταβλητών στη δημιουργία των μοντέλων ανά τμήμα



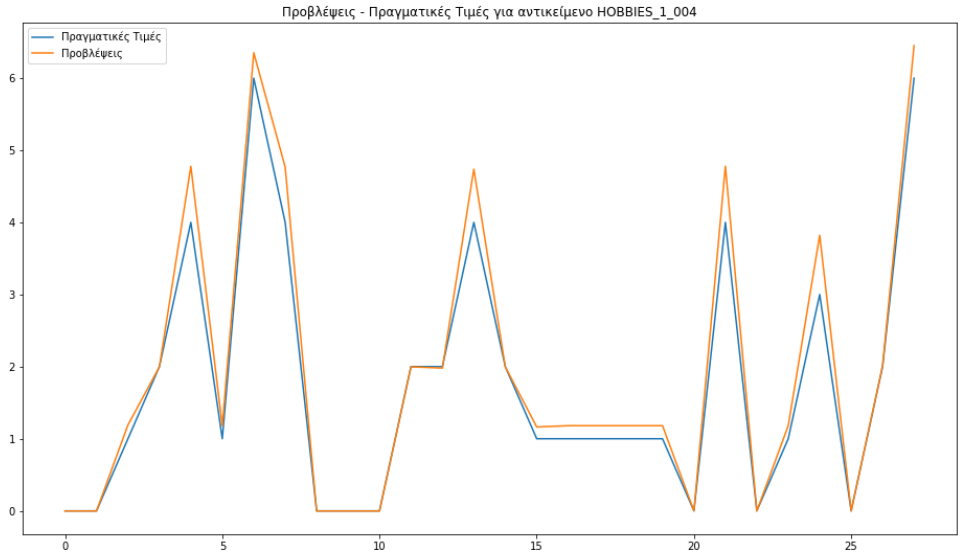
Εικόνα 32: Η επίδραση των σημαντικότερων μεταβλητών στη δημιουργία των μοντέλων ανά τμήμα

4.9. Συμπεράσματα

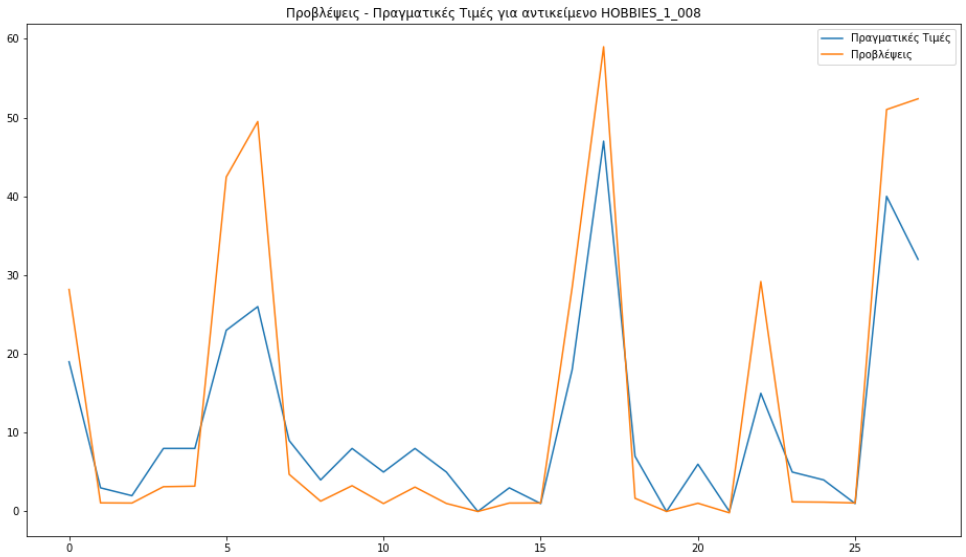
Σύμφωνα με τα πειράματα που εκτελέσαμε, παρατηρείται ότι και στις τέσσερις περιπτώσεις, ανά πολιτεία, ανά κατάστημα, ανά κατηγορία και ανά τμήμα, η χρήση επιπλέον μεταβλητών οδηγεί σε μικρότερο σφάλμα και επομένως σε καλύτερα αποτελέσματα. Ιδιαίτερα σε όλες τις περιπτώσεις όπως φαίνεται και στα παραπάνω γραφήματα, το σημαντικότερο ρόλο στη δημιουργία των μοντέλων έχουν οι μεταβλητές – στήλες revenue, rolling_mean, selling_trend, store_item_avg, και τα lag features με ιδιαίτερη σημασία τις πωλήσεις της προηγούμενης ημέρας, καθώς και της προηγούμενης εβδομάδας. Μικρότερης σημασίας είναι οι μεταβλητές snap_CA, snap_TX και snap_WI οι οποίες αναφέρονται στις ημέρες του μήνα, όπου οι πολίτες και οικογένειες με χαμηλά εισοδηματικά κριτήρια, μπορούν να κάνουν χρήση προπληρωμένων καρτών για την αγορά προϊόντων διατροφής. Επίσης οι ημερολογιακές μεταβλητές καθότι είναι σημαντικές, δεν έχουν όμως και τον μεγαλύτερο ρόλο στη δημιουργία των μοντέλων. Οι μεταβλητές item_shop_last_sale, year, item_id, event_name_1, event_name_2, dept_avg, state_avg ,cat_item_avg, state_store_avg, dept_item_avg, cat_avg, state_id, cat_id, cat_dept_avg, state_id, cat_id, dept_id, item_first_sale, state_store_avg , item_last_sale, event_type_2 όπως παρουσιάζεται και στις προηγούμενες ενότητες είχανε μικρή και έως καθόλου επίδραση στη δημιουργία των προβλεπτικών μοντέλων.

Στα πειράματα τα οποία λάβαμε υπόψη όχι όλες, άλλα μόνο τις μεταβλητές που είχαν ακόμη και ένα μικρό ρόλο στη δημιουργία των μοντέλων, φαίνεται ότι στα σφάλματα δεν μειώθηκαν αισθητά, και ότι ακόμη και σε κάποιες περιπτώσεις υπήρχε μικρή αύξηση των σφαλμάτων.

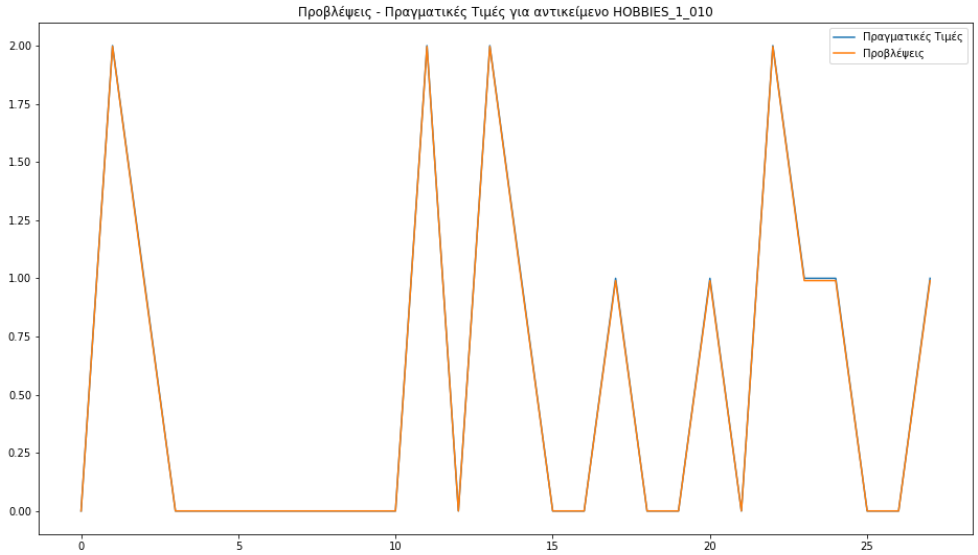
Στα γραφήματα που ακολουθούν παρουσιάζονται οι προβλέψεις και οι πραγματικές τιμές των πωλήσεων για ενδεικτικά πέντε από τα 3049 προϊόντα του συνόλου δεδομένων, για τις 28 ημέρες πρόβλεψης.



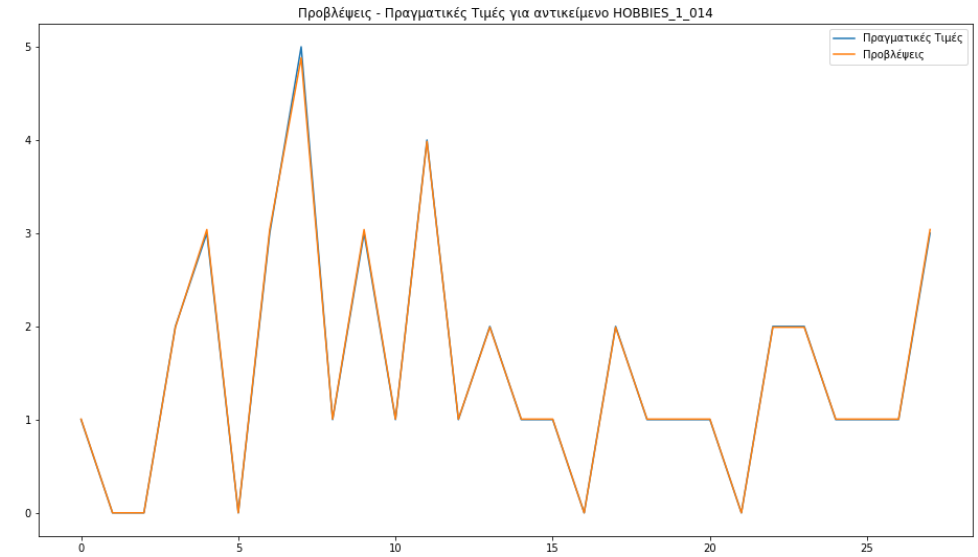
Εικόνα 33: Προβλέψεις - Πραγματικές Τιμές για το προϊόν HOBBIES_1_004



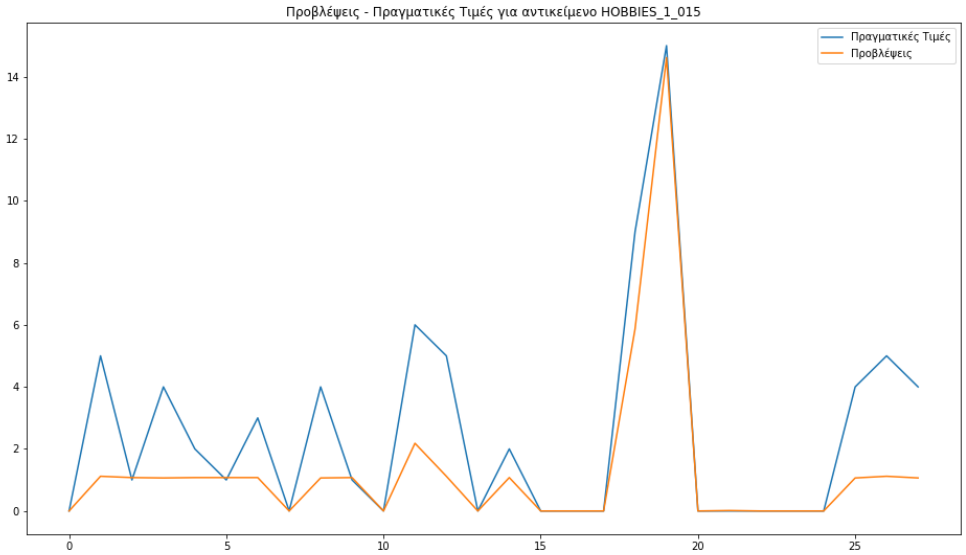
Εικόνα 34: Προβλέψεις - Πραγματικές Τιμές για το προϊόν HOBBIES_1_008



Εικόνα 35: Προβλέψεις - Πραγματικές Τιμές για το προϊόν HOBBIES_1_010



Εικόνα 36: Προβλέψεις - Πραγματικές Τιμές για το προϊόν HOBBIES_1_014



Εικόνα 37: Προβλέψεις - Πραγματικές Τιμές για το προϊόν HOBBIES_1_015

Ο συνολικός αριθμός των πειραμάτων που πραγματοποιήσαμε και παρουσιάσαμε στην συγκεκριμένη διπλωματική, ανέρχεται στα 69. Αναλυτικότερα: 3 πολιτείες επί 3 μοντέλα (initial, extra και important), συνολικά 10 καταστήματα επί 3 μοντέλα (initial, extra και important), 3 κατηγορίες προϊόντων επί 3 μοντέλα (initial, extra και important) και 7 τμήματα επί 3 μοντέλα (initial, extra και important)

Αξίζει να σημειωθεί ότι από τα 69 μοντέλα που δημιουργήθηκαν κατά τη διάρκεια των πειραμάτων για τα παραπάνω γραφήματα επιλέχθηκε το μοντέλο για το κατάστημα 4 της Καλιφόρνιας (CA_4), όπου το η μετρική RMSE (η ρίζα του μέσου τετραγωνικού σφάλματος) μας έδωσε, στο μοντέλο με όλες τις έξτρα μεταβλητές που δημιουργήσαμε, σκορ 0.11, το οποίο θεωρείται εξαιρετικά ικανοποιητικό και για αυτό βλέπουμε τόσο μεγάλη ακρίβεια στις προβλέψεις, όπως παράδειγμα για τα προϊόντα HOBBIES_1_010 και HOBBIES_1_014.

5. Σύνοψη και Μελλοντικές Προτάσεις

5.1. Σύνοψη και συμπεράσματα

Στην παρούσα διπλωματική εργασία, ασχοληθήκαμε με το πρόβλημα της πρόβλεψης πωλήσεων, στα καταστήματα λιανικού εμπορίου και συγκεκριμένα το σύνολο δεδομένων, αφορούσε δεδομένα πωλήσεων της αμερικάνικης πολυεθνικής εταιρείας Walmart. Αντιμετωπίστηκε σαν ένα πρόβλημα πρόβλεψης χρονοσειρών με μεθόδους μηχανικής μάθησης. Αρχικά ορίστηκε το πρόβλημα και έγινε εισαγωγή και βιβλιογραφική αναφορά σε βασικές έννοιες μηχανικής μάθησης, χρονοσειρών και ορισμένων βασικών αλγορίθμων που χρησιμοποιούνται τα τελευταία χρόνια για την αντιμετώπιση και επίλυση της συγκεκριμένης κατηγορίας προβλημάτων. Έπειτα ορίσαμε τα πειράματα, τα οποία υλοποιήθηκαν με τη χρήση του αλγορίθμου μηχανικής μάθησης LightGBM. Στο σύνολο δεδομένων γίνανε οι κατάλληλες τροποποιήσεις ώστε να είναι σε μορφή όπου μπορεί να αντιμετωπιστεί ως πρόβλημα χρονοσειρών. Υλοποιήθηκαν τρείς βασικές κατηγορίες πειραμάτων, ένα αρχικό μοντέλο χωρίς την προσθήκη επιπλέον μεταβλητών (initial), ένα ακόμη μοντέλο με την προσθήκη όλων των μεταβλητών που δημιουργήθηκαν από εμάς (extra), και τέλος ένα χωρίς να λάβουμε υπόψη τις μεταβλητές που έχουν ελάχιστο και έως καθόλου ρόλο στη δημιουργία των μοντέλων (important). Τα πειράματα υλοποιήθηκαν, ανά πολιτεία (state), ανά κατάστημα (store), ανά κατηγορία προϊόντος (category) και ανά τμήμα (department). Συνολικά δημιουργήθηκαν 69 μοντέλα. Δημιουργήθηκαν και χρησιμοποιήθηκαν μεταβλητές όπως το revenue όπου υπολογίζονται τα έσοδα, ως το γινόμενο του αριθμού των προϊόντων που πωλήθηκαν επί την τιμή του προϊόντος, το rolling_mean και expanding_mean, ο κινητός μέσος όρος των προϊόντων που πωλούνται σε διάστημα μίας εβδομάδας και ο μέσος όρος με την προσθήκη επιπλέον τιμών κάθε φορά και το selling_trend όπου φαίνεται η τάση της πώλησης που παρουσιάζεται στα προϊόντα καθώς και αρκετές μεταβλητές που υπολογίζουν μέσους όρους, όπως για παράδειγμα η store_item_avg. Δημιουργήθηκαν επίσης lag features για την προηγούμενη ημέρα, 2 και 3 ημέρες πριν, την προηγούμενη εβδομάδα, 2 εβδομάδες πριν και πριν από ένα μήνα. Σε όλα τα πειράματα παρατηρήσαμε ότι οι σημαντικότερες μεταβλητές που

συντέλεσαν στη δημιουργία των μοντέλων ήταν οι μεταβλητές που δημιουργήσαμε revenue, rolling_mean, selling_trend, store_item_avg, καθώς και τα lag features και ιδιαίτερα οι πωλήσεις της προηγούμενης ημέρας. Από της αρχικές στήλες που υπήρχαν στο σύνολο δεδομένων αρκετά μεγάλη σημασία είχε η sell_price, και ακολούθησαν σε σημαντικότητα ημερολογιακές μεταβλητές όπως η ημέρα d, ο μήνας month, η ημέρα της εβδομάδας wday, weekday, wm_yr_wk. Οι στήλες snap_CA, snap_WI και snap_TX δεν έχαν και τόσο σημαντικό ρόλο στη δημιουργία των μοντέλων. Τέλος φάνηκε ότι οι μεταβλητές item_shop_last_sale, year, item_id, event_name_1, event_name_2, dept_avg, state_avg ,cat_item_avg, state_store_avg, dept_item_avg, cat_avg, state_id, cat_id, cat_dept_avg, state_id, cat_id, dept_id, item_first_sale, state_store_avg , item_last_sale, event_type_2 είχανε μικρή και έως καθόλου επίδραση στη δημιουργία των προβλεπτικών μοντέλων.

5.2. Μελλοντικές προτάσεις

Η παρούσα διπλωματική έχει υλοποιηθεί με τη χρήση του αλγορίθμου μηχανικής μάθησης LightGBM και με τη δημιουργία αρκετών προβλεπτικών μεταβλητών. Λαμβάνοντας υπόψη και αξιοποιώντας τη συγκεκριμένη έρευνα σε συνδυασμό με πληροφορίες που μπορούμε να αποκτήσουμε και από υπόλοιπες έρευνες που έχουν πραγματοποιηθεί πάνω στο συγκεκριμένο επιστημονικό πεδίο, για περεταίρω βελτίωση θα μπορούσαν να προταθούν τα εξής:

Δημιουργία και σύγκριση μοντέλων τα οποία μπορούν να υλοποιηθούν με τη χρήση επιπλέον αλγορίθμων όπως για παράδειγμα XGBoost, ο αλγόριθμος Prophet του Facebook, Random Forest, ακόμα και με τη χρήση Νευρωνικών Δικτύων.

Εύρεση και ενσωμάτωση καιρικών μεταβλητών, για το χρονολογικό εύρος των δεδομένων. Αναζήτηση για ακραία καιρικά φαινόμενα και καταστροφές. Επίσης μπορούν να προστεθούν πληροφορίες σχετικά με την κίνηση και την κυκλοφορία των οχημάτων στις ευρύτερες περιοχές των καταστημάτων, για την ίδια ημέρα. Πληροφορίες σχετικά με τρόπους μετακίνησης προς και από τα καταστήματα, καθώς και τα ωράρια των δρομολογίων στα μέσα μαζικής μεταφοράς.

Πληροφορίες για τον αριθμό των πελατών στα καταστήματα κατά τη διάρκεια της ημέρας. Πληροφορίες σχετικά με ασθένειες, για την ίδια ημέρα και για κάποια προηγούμενη χρονική περίοδο, όπως ιώσεις, γρίπες, εποχιακές γρίπες ακόμα και πανδημίες.

Η παραπάνω προβλεπτική διαδικασία θα μπορούσε να υλοποιηθεί με τη χρήση παράλληλου προγραμματισμού (parallel computing), ή ακόμη και με τη χρήση GPU ή TPU, για βελτίωση της ταχύτητας εκτέλεσης του.

Βιβλιογραφία

1. Arthur, Charles (23 August 2013). "Tech giants may be huge, but nothing matches big data". The Guardian. ISSN 0261-3077. Retrieved 30 April 2019.
2. Rygielski, C., Wang, J. C., & Yen, D. C. (2002). Data mining techniques for customer relationship management. Technology in society, 24(4), 483-502.
3. Gazi, B. (2010). Krzysztof J. Cios, Witold Pedrycz, Roman W. Swiniarski and Lukasz A. Kurgan Data Mining: A Knowledge Discovery Approach. Springer ((2007)). ISBN: 978-0387333335.£ 55.99. 606 pp. Hardcover. The Computer Journal, 53(4), 489-490.
4. Γολέμη, Ε. (2010). Κρυπτογραφία και εξόρυξη δεδομένων (Doctoral dissertation).
5. Krause-Traudes, M., Scheider, S., Rüping, S., & Meßner, H. (2008, May). Spatial data mining for retail sales forecasting. In 11th AGILE International Conference on Geographic Information Science (pp. 1-11).
6. Alpaydin, E. (2020). Introduction to machine learning. MIT press.
7. Zhu, X. J. (2005). Semi-supervised learning literature survey.
8. Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences, 55(1), 119-139.
9. Dodge, Y. (2008). The concise encyclopedia of statistics. Springer Science & Business Media.
10. Shumway, R. H., Stoffer, D. S., & Stoffer, D. S. (2000). Time series analysis and its applications (Vol. 3). New York: springer.
11. Bisgaard, S., & Kulahci, M. (2011). Time series analysis and forecasting by example. John Wiley & Sons.
12. Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. Information Sciences, 191, 192-213.
13. Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: an analysis and review. International journal of forecasting, 16(4), 437-450.
14. Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. BMC bioinformatics, 7(1), 1-8.

15. Landwehr, N., Hall, M., & Frank, E. (2005). Logistic model trees. Machine learning, 59(1), 161-205.
16. Doganis, P., Alexandridis, A., Patrinos, P., & Sarimveis, H. (2006). Time series sales forecasting for short shelf-life food products based on artificial neural networks and evolutionary computing. Journal of Food Engineering, 75(2), 196-204.
17. Liu, X., & Ichise, R. (2017, July). Food sales prediction with meteorological data—a case study of a japanese chain supermarket. In International Conference on Data Mining and Big Data (pp. 93-104). Springer, Cham.
18. Cai, X., Chen, J., Xiao, Y., & Xu, X. (2010). Optimization and coordination of fresh product supply chains with freshness-keeping effort. Production and Operations management, 19(3), 261-278.
19. Tsoumakas, G. (2019). A survey of machine learning techniques for food sales prediction. Artificial Intelligence Review, 52(1), 441-447.
20. Žliobaite, I., & Kuncheva, L. I. (2009, December). Determining the training window for small sample size classification with concept drift. In 2009 IEEE International Conference on Data Mining Workshops (pp. 447-452). IEEE.
21. Meulstee, P., & Pechenizkiy, M. (2008, December). Food sales prediction:" If only it knew what we know". In 2008 IEEE International Conference on Data Mining Workshops (pp. 134-143). IEEE.
22. Tsymbal, A., Pechenizkiy, M., Cunningham, P., & Puuronen, S. (2008). Dynamic integration of classifiers for handling concept drift. Information fusion, 9(1), 56-68.
23. Kearns, M. (1988). Thoughts on hypothesis boosting. Unpublished manuscript, 45, 105.
24. Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences, 55(1), 119-139.
25. Piryonesi, S. M., & El-Diraby, T. E. (2020). Data analytics in asset management: Cost-effective prediction of the pavement condition index. Journal of Infrastructure Systems, 26(1), 04019036.
26. Hastie, T., Tibshirani, R., & Friedman, J. (2009). Boosting and additive trees. In The elements of statistical learning (pp. 337-387). Springer, New York, NY.

27. Madeh Piryonesi, S., & El-Diraby, T. E. (2021). Using machine learning to examine impact of type of performance indicator on flexible pavement deterioration modeling. Journal of Infrastructure Systems, 27(2), 04021005.
28. Chen, T. (2016). Story and lessons behind the evolution of XGBoost.
29. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30.
30. https://towardsdatascience.com/sales-forecasting-from-time-series-to-deep-learning-5d115514bfac
31. How Machine Learning Works https://www.mathworks.com/discovery/machine-learning.html
32. https://www.javatpoint.com/types-of-machine-learning
33. https://towardsdatascience.com/stationarity-in-time-series-analysis-90c94f27322
34. https://www.kaggle.com/code/ryanholbrook/trend
35. https://www.kaggle.com/code/ryanholbrook/seasonality
36. https://robjhyndman.com/hyndsight/cyclicts/
37. https://robjhyndman.com/hyndsight/tscv/
38. https://otexts.com/fpp3/tscv.html
39. https://devpost.com/software/stock-portfolio-allocation
40. https://machinelearningmastery.com/xgboost-for-time-series-forecasting/
41. https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d
42. https://medium.com/@pushkarmandot/https-medium-com-pushkarmandot-what-is-lightgbm-how-to-implement-it-how-to-fine-tune-the-parameters-60347819b7fc
43. https://lightgbm.readthedocs.io/en/latest/Parameters.html
44. https://medium.com/@pushkarmandot/https-medium-com-pushkarmandot-what-is-lightgbm-how-to-implement-it-how-to-fine-tune-the-parameters-60347819b7fc