ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΣΤΗ ΔΙΟΙΚΗΣΗ ΕΠΙΧΕΙΡΗΣΕΩΝ

Διπλωματική Εργασία

ΑΝΑΛΥΤΙΚΗ ΔΕΔΟΜΕΝΩΝ ΣΤΙΣ ΕΠΙΧΕΙΡΗΣΕΙΣ: ΜΕΛΕΤΗ ΠΕΡΙΠΤΩΣΗΣ AIRBNB

του

ΦΩΤΙΟΥ Γ. ΖΗΣΟΠΟΥΛΟΥ

Υποβλήθηκε ως απαιτούμενο για την απόκτηση του μεταπτυχιακού διπλώματος ειδίκευσης στη Διοίκηση Επιχειρήσεων

ΙΑΝΟΥΑΡΙΟΣ 2019

Περίληψη

Η ταξινόμηση και οι συστάσεις αναζήτησης είναι θεμελιώδη προβλήματα κρίσιμου ενδιαφέροντος για τις μεγάλες εταιρείες του διαδικτύου, συμπεριλαμβανομένων των μηχανών αναζήτησης ιστού, ιστοσελίδων δημοσίευσης περιεχομένου και αγορών. Ωστόσο, παρά το γεγονός ότι μοιράζονται μερικά κοινά χαρακτηριστικά, δεν υπάρχει μια κοινή λύση για όλους σε αυτόν τον χώρο. Αντίστοιχα στον τομέα του τουρισμού, τα προβλήματα προβολής και ταξινόμησης αποτελεσμάτων αναζήτησης και οι συστάσεις προς τους χρήστες, έχουν τις δικές τους ιδιαιτερότητες. Ειδικότερα στην Airbnb, μια αγορά βραχυπρόθεσμης ενοικίασης ακινήτων, η σύσταση καταλυμάτων στους χρήστες της είναι μια μοναδική πρόκληση, καθώς πρόκειται για μια αγορά με δύο κατηγορίες χρηστών, τους φιλοξενούμενους και τους οικοδεσπότες. Ακόμη μεγαλύτερη γίνεται αυτή η πρόκληση όταν η εταιρία δεν διαθέτει αρκετές πληροφορίες για κάποιο χρήστη. Αυτή την περίπτωση αντιμετωπίζει η εταιρία κάθε φορά που πρέπει να προτείνει καταλύματα σε νέους χρήστες, οι οποίοι δεν έχουν κάνει ακόμη κράτηση μέσω της πλατφόρμας της. Σε αυτή την εργασία περιγράφεται ο τρόπος που μπορεί να λυθεί ένα τέτοιο πρόβλημα με χρήση της επιστήμης των δεδομένων και της μηχανικής μάθησης. Θα αναλύσουμε συγκεκριμένα μια από τις λύσεις που υποβλήθηκαν σε διαγωνισμό της Airbnb που διεξήχθη μέσα από την πλατφόρμα της Kaggle.

Πίνακας περιεχομένων

Λίστα διαγραμμάτων

Λίστα πινάκων

Λίστα εικόνων

1. Εισαγωγή

1.1. Ορισμός προβλήματος

Η αλματώδης τεχνολογική ανάπτυξη των τελευταίων δεκαετιών δεν θα μπορούσε παρά να έχει επίδραση και στον κλάδο του τουρισμού. Το ερευνητικό ενδιαφέρον στρέφεται γύρω από πολλά θέματα αναφορικά με τη μελέτη των ταξιδιών και του τουρισμού γενικότερα, ωστόσο το μεγαλύτερο ενδιαφέρον προσελκύει η μελέτη του τρόπου λήψης των αποφάσεων. Δεδομένης και της ποικιλόμορφης φύσης του τουριστικού κλάδου, η πρόκληση περιλαμβάνει μια ευρεία γκάμα ξεκινώντας από την επιλογή του προορισμού καθώς και όλων των συναφών υπηρεσιών κατά τη διάρκεια ενός ταξιδιού. Είναι λοιπόν φυσικό, η μελέτη του τρόπου λήψης αυτών των αποφάσεων να προσελκύει το ενδιαφέρον τόσο των ερευνητών σε θεωρητικό επίπεδο όσο και των επαγγελματιών του κλάδου, οι οποίοι αναγνωρίζουν την τεράστια οικονομική επίδραση στην παγκόσμια οικονομία. Η συνεχόμενη ανάπτυξη της τεχνολογίας της πληροφορικής και των εφαρμογών που σχετίζονται με την λήψη αποφάσεων στον τουρισμό, δημιούργησε την ανάγκη για καλύτερη κατανόηση και προσοχή στον τομέα αυτό, καθώς τα μοντέλα αποφάσεων ταξιδιών θα πρέπει να τροποποιούνται συνεχώς και να αντικατοπτρίζουν στη δομή τους τις σύγχρονες απαιτήσεις.

Η αλλαγή αυτή δεν θα πρέπει να θεωρηθεί εύκολη για πολλούς λόγους. Μία από τις πιο δραματικές αλλαγές είναι η μετατροπή της ιεραρχίας πρόσβασης στις πληροφορίες από κάθετη σε οριζόντια δομή. Θεωρητικά, οποιοσδήποτε με έναν ηλεκτρονικό υπολογιστή και πρόσβαση στον Παγκόσμιο Ιστό έχει την ικανότητα να αναζητήσει, να έχει πρόσβαση και να ταξινομήσει τεράστιες ποσότητες πληροφοριών από όλο τον κόσμο. Η δυνατότητα αυτή έχει ήδη κάνει την εμφάνιση της τα προηγούμενα χρόνια ως ένα τεραστίων διαστάσεων φαινόμενο στον τομέα των ταξιδιών, επηρεάζοντας τον τρόπο με τον οποίο οι άνθρωποι συγκεντρώνουν πληροφορίες σχετικά με το που να πάνε και τι να κάνουν, αλλά και τον τρόπο με τον οποίο χρησιμοποιούν τις διάφορες πηγές πληροφοριών. Έτσι λοιπόν, οι περισσότεροι πλέον προτιμούν διαδικτυακές πηγές ως πρώτη επιλογή αναζήτησης, αντικαθιστώντας τον παραδοσιακό τρόπο του παρελθόντος που προϋπέθετε την ανταλλαγή πληροφορίων από στόμα σε στόμα, ενώ προτιμούν και τη διαδικτυακή αγορά προϊόντων και υπηρεσιών τουρισμού (Fesenmeier, Woeber, Werthner, 2006).

Συνοψίζοντας, γίνεται φανερό ότι η καταναλωτική συμπεριφορά έχει διαφοροποιηθεί σημαντικά και θα συνεχίσει να διαφοροποιείται. Ήδη τα περισσότερα μοτίβα που αναγνωρίστηκαν από τους ερευνητές στο παρελθόν έχουν ξεπεραστεί και δεν ενδείκνυνται για σχεδιασμό, διαχείριση, προώθηση και διαμόρφωση πολιτικής. Η αλλαγή αυτή εγείρει την ανάγκη για την διερεύνηση και τη δημιουργία νέων μοντέλων αποφάσεων και συστημάτων συστάσεων (travel recommendation systems).

1.2. Το πλαίσιο και οι στόχοι της έρευνας

Στην παρούσα εργασία, στα πλαίσια της ανάλυσης των τουριστικών συστημάτων συστάσεων της Airbnb, θα παρουσιαστεί η επίλυση ενός προβλήματος μηχανικής μάθησης. Το πρόβλημα προκύπτει από την έλλειψη πληροφοριών για νέους χρήστες της Airbnb και η επίλυσή του δίνει τη δυνατότητα πρόβλεψης της χώρας προορισμού για την πρώτη κράτηση ενός νέου χρήστη. Θα παρουσιαστεί και θα αναλυθεί μια από τις λύσεις που υποβλήθηκαν σε διαγωνισμό της ηλεκτρονικής πλατφόρμας Kaggle. Το μοντέλο δημιουργήθηκε με βάση δεδομένα όπως προσωπικά στοιχεία που εισήγαγαν οι ίδιοι κατά τη δημιουργία του λογαριασμού τους και δεδομένα που συλλέχθηκαν από τον ιστότοπο αναφορικά με το τι έκαναν οι χρήστες κατά τη διάρκεια παραμονής τους σε αυτόν. Στόχος μας είναι μέσα από αυτή τη διαδικασία της ανάλυσης του προβλήματος και της λύσης, να παρουσιάσουμε τον τρόπο με τον οποίο οι σύγχρονες επιχειρήσεις αντιμετωπίζουν τέτοιου είδους προβλήματα.

1.3. Δομή της εργασίας

Στο κεφάλαιο 2, αφού δώσουμε έναν ορισμό για τον τουρισμό και κάνουμε μια ιστορική αναδρομή για τις διαδικτυακές τουριστικές πλατφόρμες, θα παρουσιάσουμε διαθέσιμες πληροφορίες που υπάρχουν στη βιβλιογραφία γύρω από το θέμα των συστημάτων συστάσεων, ειδικά στον τομέα του τουρισμού. Στο κεφάλαιο 3, θα παρουσιάσουμε τη μεθοδολογική προσέγγιση του θέματος της παρούσας εργασίας, αντιστοιχίζοντας τα βήματά της με τα κεφάλαια που ακολουθούν. Στο κεφάλαιο 4, παρουσιάζουμε τη χρήση της επιστήμης των δεδομένων από την Airbnb και επικεντρωνόμαστε στο σύστημα εξατομικευμένων συστάσεων της εταιρείας. Στο

κεφάλαιο 5, περιγράφεται το πρόβλημα που θέλησε να λύσει η Airbnb μέσα από το διαγωνισμό στην Kaggle. Γίνεται η επιλογή της λύσης αφού αναλυθούν πρώτα τα κριτήρια με βάσει τα οποία θα γίνει αυτή η επιλογή. Στη συνέχεια, στο κεφάλαιο 6, θα παρουσιάσουμε τα δεδομένα τα οποία δόθηκαν στους συμμετέχοντες του διαγωνισμού από την ίδια την Airbnb και θα προχωρήσουμε σε μια διερευνητική ανάλυση των μεταβλητών. Στο κεφάλαιο 7, θα παρουσιάσουμε τη λύση που επιλέξαμε, περιγράφοντας και αναλύοντας τα τμήματα κώδικα που χρησιμοποιήθηκαν για την εκπαίδευση του μοντέλου. Στο τελευταίο κεφάλαιο θα κλείσουμε την εργασία μας με τα συμπεράσματα στα οποία καταλήξαμε.

2. Βιβλιογραφική επισκόπηση

2.1. Έννοια του τουρισμού

Εξαιτίας του γεγονότος ότι ο τουρισμός εμφανίζεται με διαφορές μορφές, ο ορισμός του αποτελεί μια αρκετά δύσκολη υπόθεση. Μία από τις πρώτες προσπάθειες ήταν αυτή των Hunziker και Krapf το 1942, οι οποίοι όρισαν τον τουρισμό ως το σύνολο των σχέσεων και φαινομένων που προκύπτουν από την πραγματοποίηση ενός ταξιδιού και τη διαμονή αγνώστων, υπό τον όρο ότι η διαμονή δεν συνεπάγεται τη δημιουργία μόνιμης κατοικίας και δεν συνδέεται με κερδοσκοπική δραστηριότητα. Σε μια πιο σύγχρονη προσέγγιση, ο Παγκόσμιος Οργανισμός Τουρισμού (World Tourism Organization) ορίζει τον τουρισμό ως τις δραστηριότητες των ανθρώπων που ταξιδεύουν και διαμένουν σε προορισμούς διαφορετικούς από αυτούς που αποτελούν το συνηθισμένο περιβάλλον τους και για χρονικό διάστημα που δεν ξεπερνά τον έναν χρόνο, με σκοπό την αναψυχή, την εκπλήρωση επαγγελματικών υποχρεώσεων κ.α.

Στη σύγχρονη μορφή του, ο τουρισμός, αποτελεί ένα οικονομικό και κοινωνικό φαινόμενο, η ανάπτυξη του οποίου, μετά το Β' Παγκόσμιο Πόλεμο προσελκύει το ενδιαφέρον ολοένα και περισσότερων αναπτυγμένων αλλά και αναπτυσσόμενων χωρών. Από το 1997, άλλωστε, αποτελεί πλέον την μεγαλύτερη βιομηχανία του κόσμου (Ηγουμενάκης, Κραβαρίτης, 2004).

2.2. Ιστορική αναδρομή

Η διείσδυση του παγκόσμιου ιστού στην καθημερινότητα μας άλλαξε θεμελιωδώς όχι μόνο την εμπειρία των ταξιδιών, αλλά και ολόκληρη τη βιομηχανία, καθώς εμφανίστηκαν νέα δυναμικά εμπορικά σήματα, ενώ οι υπάρχοντες «παίκτες» αγωνίζονταν να συμβαδίσουν στον ταχύτατο αυτό ρυθμό. Οι καταναλωτές, όλο και πιο ενδυναμωμένοι και συνδεδεμένοι άρχισαν να έχουν μεγαλύτερες απαιτήσεις από τους ταξιδιωτικούς φορείς, με αποτέλεσμα τα τελευταία 25 χρόνια να αλλάξει δραματικά το τοπίο των εξελίξεων στον τουριστικό κλάδο.

Πολύ πριν οι ηλεκτρονικοί υπολογιστές εισβάλουν στη ζωή των περισσότερων νοικοκυριών, το 1989, ο Βρετανός επιστήμονας πληροφορικής Tim Bernerns-Lee έγραψε μια πρόταση για το πως τελικά θα εξελιχθεί ο Παγκόσμιος Ιστός.

Μερικά χρόνια αργότερα, το 1994, ο ιστότοπος Travelweb.com αναδείχθηκε ως η πρώτη λεπτομερής λίστα ξενοδοχείων απ' όλο τον κόσμο, περιλαμβάνοντας λίγο αργότερα και τη δυνατότητα κρατήσεων. Ο διαδικτυακός αυτός ιστότοπος δημιουργήθηκε από μία μικρή σχετικά ομάδα υπό τον John Davis, βασικό παράγοντα στην ιστορία των online ταξιδιών, ο οποίος και δημιούργησε έναν από τους σημαντικότερους προμηθευτές της τεχνολογίας κρατήσεων, τον Πήγασο. Η ιστοσελίδα λειτουργούσε κυρίως ως βιτρίνα, για να καλύπτει τις υπόλοιπες δραστηριότητες της εταιρίας, ωστόσο έγινε αρκετά δημοφιλής στους πρώιμους επισκέπτες του διαδικτύου, λόγω της μοναδικότητας της. Την ίδια χρονιά, ο Αμερικάνος ταξιδιώτης και συγγραφέας Jeff Greenwald, πυροδότησε αυτό που έμελλε να εξελιχθεί τα επόμενα χρόνια σε μια τεράστια βιομηχανία για ονειροπόλους, ταξιδιώτες και αμέτρητους ανθρώπους σε όλο τον κόσμο, το ταξιδιωτικό blogging με τη δημοσίευση ενός άρθρου 1.600 λέξεων για λογαριασμό του πρώτου εμπορικού ιστότοπου ενός τουριστικού γραφείου στο Μεξικό.

Την επόμενη χρονιά, το 1995, η εταιρεία Viator Systems (η σημερινή Viator), ξεκίνησε μια επιχείρηση τεχνολογίας ταξιδιών για την παροχή κρατήσεων για εκδρομές, ένα σύστημα το οποίο συναντάται ακόμα και σήμερα πίσω από τα σημερινά συγγενικά συστήματα κρατήσεων της εταιρίας. Παράλληλα, η εταιρία Lonely Planet, ο κύριος ηγέτης στους γραπτούς ταξιδιωτικούς οδηγούς, αναγνωρίζοντας την ανάγκη της επέκτασης στο διαδίκτυο, πραγματοποίησε μια επιτυχημένη επέκταση στον παγκόσμιο ιστό και ενέπνευσε εκατοντάδες, ίσως και χιλιάδες παρόμοιους εμπορικούς οδηγούς να συνδεθούν με το διαδίκτυο και να προσφέρουν τις συμβουλές τους.

Το 1996 αποτελεί χρονιά σταθμό για τον διαδικτυακό τουρισμό. Η Microsoft, μία από τις πλέον εξελιγμένες εταιρίες της εποχής, συνειδητοποίησε ότι υπήρχε κάτι στην πρώιμη αυτή ορμή και πραγματοποίησε την προσπάθεια της με ένα online ταξιδιωτικό γραφείο, γνωστό ως Expedia. Η διαδικτυακή αυτή τοποθεσία έθεσε στην ουσία τα θεμέλια στον κλάδο των ψυχαγωγικών και εμπορικών ταξιδιών, ακολουθούμενη από αμέτρητους μιμητές. Σήμερα αποτελεί έναν από τους μεγαλύτερους ομίλους παροχής διαδικτυακών υπηρεσιών ταξιδιών, που περιλαμβάνει ιστότοπους όπως οι CarRentals.com, Hotels.com, Trivago, Travelocity κ.α. Τα πρώτα της βήματα

πραγματοποίησε την ίδια χρονιά και η πιο διαδεδομένη μηχανή αναζήτησης του πλανήτη, η Google, με τους παρόχους ταξιδιών να διαθέτουν σταδιακά δισεκατομμύρια για να διαφημιστούν παράλληλα με τα αποτελέσματα αναζήτησης, μια σχέση η οποία συνεχίζεται μέχρι σήμερα με αμείωτο ρυθμό. Στην Ευρώπη, η αεροπορική βιομηχανία έμελλε να αλλάξει για πάντα όταν δύο εταιρίες, η EasyJet και η Ryanair, προώθησαν το μοντέλο των αεροπορικών εταιριών χαμηλού κόστους. Έτσι το «Κάντε κράτηση μέσω του ταξιδιωτικού πρακτορείου σας» άρχισε να αντικαθίσταται από το «Κάντε κράτηση στην ιστοσελίδα μας».

Καθώς τα online ταξιδιωτικά γραφεία απέκτησαν γερές ρίζες στις Ηνωμένες Πολιτείες, στην Ευρώπη το 1998, δύο νέοι, οι Brent Hoberman και Martha Lane Fox οδήγησαν τον αστείρευτο και παιχνιδιάρικο κόσμο του Lastminute.com στον ακόμα καινούργιο κόσμο του διαδικτυακού εμπορίου, ενισχύοντας την ιδέα στο μυαλό των καταναλωτών ότι τα ταξίδια μπορούν να αγοραστούν μέσω διαδικτύου με ευκολία και την τελευταία στιγμή. Εν τέλει η εταιρία εξαγοράστηκε το 2005 από την αμερικάνικη ομολογιακή εταιρία Travelocity.com έναντι του ποσού των 577 εκατομμυρίων λιρών.

Το 1999, η αμερικάνικη ιστοσελίδα FareChase εμφανίστηκε στο προσκήνιο με μία διαφορετική πρόταση στην εμπειρία του online shopping ακολουθούμενη τα επόμενα χρόνια από αμέτρητα (και πιο επιτυχημένα) εμπορικά σήματα. Γνωστή με τον αγγλικό όρο «metasearch» στην αγορά, η ιστοσελίδα έγινε γνωστή ως μια απλή μηχανή αναζήτησης που συλλέγει ναύλους από τοποθεσίες κρατήσεων (πρακτορεία, αεροπορικές εταιρίες, ξενοδοχεία). Αργότερα, το 2004 εξαγοράστηκε από τη Yahoo και έθεσε τις βάσεις για δημοφιλείς ιστοσελίδες όπως οι Sidestep, Kayak, TravelSupermarket και Skyscanner.

Ενδεχομένως λίγοι θα μπορούσαν να προβλέψουν το 2000 ότι μια μικρή επιχείρηση με έδρα στη Μασαχουσέτη, θα εξελισσόταν σε μία από τις δημοφιλέστερες ταξιδιωτικές ιστοσελίδες στον κόσμο. Το TripAdvisor, το πρωτότυπο και μέχρι στιγμής το δημοφιλέστερο και μεγαλύτερο κοινωνικό ταξιδιωτικό site παγκοσμίως, υιοθέτησε μια απίστευτα απλή λογική: οι ταξιδιώτες έγραφαν κριτικές για ξενοδοχεία που είχαν επισκεφθεί, οι συνάδελφοι ταξιδιώτες μπορούσαν να βρουν αυτές τις κριτικές και να αποφασίσουν αν θα προτιμήσουν ένα ξενοδοχείο ή όχι, με βάση αυτό που είχαν διαβάσει. Οι ξενοδόχοι τρομοκρατήθηκαν, αλλά οι ταξιδιώτες σχεδόν λάτρεψαν αυτή τη νέα «σοφία του πλήθους». Σήμερα το TripAdvisor έχει καταγεγραμμένες πάνω από 702

εκατομμύρια κριτικές, προσελκύει μηνιαία 490 εκατομμύρια επισκέπτες, ενώ οι λίστες του περιλαμβάνουν πάνω από 8.000.000 καταλύματα, εστιατόρια κ.α.

Η νέα χιλιετία έφερε νέες προκλήσεις και απαιτήσεις στον κλάδο του τουρισμού, ειδικά μετά και το τρομοκρατικό χτύπημα της 11ης Σεπτεμβρίου, το οποίο προκάλεσε μια ύφεση τόσο στα ταξίδια αναψυχής όσο και στα επαγγελματικά. Από το 2004 και ειδικότερα τα επόμενα χρόνια, η εταιρία Priceline, γνωστή τα προηγούμενα χρόνια για την ιδέα "Name Your Own Price", ξεκίνησε μια ανοδική πορεία, η οποία εκτοξεύτηκε με την εξαγορά του ιστότοπου ActiveHotels και κυρίως του Booking.com. Οι δύο εξαγορές προωθούσαν μια διεθνή και ανανεωμένη εστίαση στις ξενοδοχειακές κρατήσεις και κατάφεραν να δημιουργήσουν μία από τις πλέον αναγνωρίσιμες μάρκες του κλάδου, την Booking.com. Σήμερα ο όμιλος Booking Holdings Inc., όπως μετονομάστηκε τον Φεβρουάριο του 2018, διαθέτει ιστοσελίδες σε πάνω από 40 γλώσσες και 220 χώρες, ενώ οι κρατήσεις για το 2017 ανέρχονται σε πάνω από 81,2 δισεκατομμύρια δολάρια, κατατάσσοντας τον στον μεγαλύτερο ηγέτη στις διαδικτυακές κρατήσεις και συναφείς υπηρεσίες.

Την ίδια χρονική περίοδο λανσαρίστηκε στην αγορά μια από τις πλέον αναγνωρίσιμες πλατφόρμες κοινωνικής δικτύωσης, το Facebook. Το δημιούργημα του Mark Zuckerberg σηματοδότησε την αρχή μιας νέας ιδέας στην διαδικτυακή ταξιδιωτική εμπειρία: την κοινή χρήση. Οι ταξιδιώτες στράφηκαν αρχικά στο Facebook και αργότερα και στα υπόλοιπα μέσα κοινωνικής δικτύωσης (Twitter, FourSquare, Pinterest, Instagram κ.α.) χρησιμοποιώντας τα ως ένα τρόπο να δείξουν στους φίλους τους ποια μέρη είχαν ή σκόπευαν να επισκεφθούν. Με την πάροδο των χρόνων, η αυξανόμενη δημοτικότητα της πλατφόρμας, που σήμερα απαριθμεί πάνω από 2,2 δισεκατομμύρια ενεργούς χρήστες, ανάγκασε όλες τις ταξιδιωτικές εταιρίες να έχουν ενεργή και καθημερινή παρουσία σε τουλάχιστον ένα από αυτά τα μέσα, επιτρέποντας την αλληλεπίδραση με τους καταναλωτές.

Αν και δεν ευθύνεται άμεσα για την εξέλιξη των διαδικτυακών ταξιδιών, η κυκλοφορία του πρώτου iPhone της Apple το 2007 έδωσε τη δυνατότητα στους χρήστες να έχουν πρόσβαση στο διαδίκτυο μέσω των κινητών συσκευών τους και άσκησε μεγάλη επιρροή στην έναρξη χρήσης των ταξιδιωτικών υπηρεσιών σε κινητές συσκευές. Παράλληλα το αντίστοιχο App Store της Apple πυροδότησε τη δημιουργία των εφαρμογών για κινητές συσκευές των ταξιδιωτικών ιστότοπων.

Το 2008 δημιουργήθηκε η AirBedAndBreakfast, ή όπως είναι κατά κόσμον γνωστή η Airbnb. Με έδρα το Σαν Φρανσίσκο, η Airbnb είναι μια αξιόπιστη ηλεκτρονική αγορά με σκοπό την καταγραφή, την ανακάλυψη και την κράτηση καταλυμάτων σε όλο τον κόσμο. Η λειτουργία της είναι παρόμοια με τις σελίδες κράτησης για ξενοδοχεία, όπου ο χρήστης επιλέγει το κατάλυμα και τις ημερομηνίες που επιθυμεί. Η κύρια διαφορά έγκειται στο γεγονός ότι τα καταλύματα διαχειρίζονται από ιδιώτες και όχι από επαγγελματίες. Πριν από την κράτηση, οι χρήστες πρέπει να παρέχουν προσωπικά στοιχεία και πληροφορίες πληρωμής, ενώ οι οικοδεσπότες παρέχουν τιμές και άλλες λεπτομέρειες για την ενοικίαση ή για σχετικές εκδηλώσεις. Η τιμολόγηση καθορίζεται από τους ίδιους τους οικοδεσπότες, με συστάσεις της Airbnb, με αποτέλεσμα οι κρατήσεις να είναι πολλές φορές φθηνότερες από την αντίστοιχη κράτηση σε ένα ξενοδοχείο. Από τις απαρχές της εταιρίας ήταν κοινό μυστικό η μεγάλη χρήση της επιστήμης των δεδομένων για να δημιουργήσει νέες προσφορές και να αξιοποιήσει νέες πρωτοβουλίες μάρκετινγκ. Σε επόμενο κεφάλαιο θα αναλυθεί λεπτομερώς η περίπτωση της Airbnb και συγκεκριμένα η χρήση της επιστήμης των δεδομένων με τη βοήθεια της οποίας η εταιρεία επιλύει επιχειρηματικά προβλήματα.

2.3. Συστήματα συστάσεων

2.3.1 Ορισμός

Η τεράστια ανάπτυξη του παγκοσμίου ιστού και του όγκου των πληροφοριών που είναι πλέον διαθέσιμες, δημιούργησε όπως είναι αναμενόμενο, μία σύγχυση σε πολλούς καταναλωτές, οι οποίοι δεν ήταν εύκολο να αξιολογήσουν σωστά τις επιλογές που είχαν στη διάθεση τους. Οι τεχνικές εξατομίκευσης αναπτύχθηκαν για να δώσουν λύση σε αυτό το πρόβλημα, παρέχοντας προσαρμοσμένες πληροφορίες στους χρήστες, με βάση τις προτιμήσεις τους, τους περιορισμούς ή ακόμα και το γούστο τους (Gao, Liu & Wu, 2010). Οι τεχνικές αυτές είναι ιδιαιτέρως χρήσιμες και εφαρμόσιμες στη χρήση των συστημάτων συστάσεων. Τα συστήματα συστάσεων (recommender systems) είναι εργαλεία και τεχνικές λογισμικού που παρέχουν προτάσεις για στοιχεία που είναι χρήσιμα για ένα χρήστη. Οι προτάσεις μπορεί να σχετίζονται με διάφορες διαδικασίες λήψης αποφάσεων, όπως ποια είδη να αγοράσει ο καταναλωτής, τι μουσική να ακούσει ή ακόμα και τι να

διαβάσει στο διαδίκτυο (Ricci, Rokach & Shapira, 2011). Στον τομέα του τουρισμού, τα συστήματα συστάσεων αποσκοπούν στην αντιστοίχιση των χαρακτηριστικών του τουρισμού και των πόρων αναψυχής με τις ανάγκες των χρηστών (Borras, Moreno, Valls, 2014).

2.3.2 Διάκριση τουριστικών συστημάτων συστάσεων

Τα συστήματα συστάσεων έχουν ταξινομηθεί κλασσικά, ανάλογα με τον τρόπο που αναλύουν τις πληροφορίες του χρήστη και φιλτράρουν τις λίστες των στοιχείων, σε τέσσερις μεγάλες κατηγορίες: σε συστήματα βάση περιεχομένου (content based systems – CB), σε συστήματα συνεργατικού φιλτραρίσματος (collaborative filtering systems– CL), σε δημογραφικά συστήματα (demographic systems – DM) και υβριδικά συστήματα (hybrid approaches) (Burke, 2002).

2.3.2.1 Συστήματα συστάσεων βάση περιεχομένου (Content based systems)

Τα εν λόγω συστήματα (για συντομία CB), υπολογίζουν ένα βαθμό ομοιότητας μεταξύ των χρηστών και αυτών που συνιστώνται σε αυτούς. Η όλη διαδικασία διεξάγεται συγκρίνοντας τα χαρακτηριστικά της εκάστοτε σύστασης σε σχέση με τις προτιμήσεις του χρήστη, με βάση αγορές ή αναζητήσεις που πραγματοποίησε στο παρελθόν. Το αποτέλεσμα της σύγκρισης είναι συνήθως μια συνολική βαθμολογία απόδοσης, η οποία υποδηλώνει το βαθμό του «ταιριάσματος» ανάμεσα στο χρήστη και στην εκάστοτε εναλλακτική. Όσο υψηλότερη είναι αυτή η βαθμολογία τόσο μεγαλύτερη είναι η απόδοση της εναλλακτικής για ένα συγκεκριμένο χρήστη.

Για τη λειτουργία των CB συστημάτων έχουν προταθεί διάφορες προσεγγίσεις στη βιβλιογραφία. Κάποιες από αυτές χρησιμοποιούν απλές συντακτικές συγκρίσεις μεταξύ των χαρακτηριστικών των εξεταζόμενων στοιχείων ή υπηρεσιών και των χαρακτηριστικών των προφίλ των χρηστών, ενώ άλλες έχουν δυνατότητες πρόβλεψης για τη βελτίωση της απόδοσης του συστήματος. Για το σκοπό αυτό έχουν χρησιμοποιηθεί τόσο δίκτυα Bayesian όσο και λογική βασισμένη σε κανόνες (Kabassi, 2010).

Όλες αυτές οι προσεγγίσεις έχουν μια κοινή αδυναμία στη συντακτική φύση τους, η οποία επιτρέπει να ανιχνευθεί μόνο η ομοιότητα μεταξύ αντικειμένων που διαθέτουν τα ίδια χαρακτηριστικά. Ωστόσο, το κύριο πλεονέκτημα των CB συστημάτων είναι ότι βασίζονται μόνο σε γεγονότα που αφορούν το συγκεκριμένο χρήστη και, ως εκ τούτου, είναι αληθινά. Φυσικά σε ορισμένες περιπτώσεις αυτό μπορεί να θεωρηθεί και ως μειονέκτημα εξαιτίας του γεγονότος ότι η συγκεκριμένη μέθοδος μπορεί να οδηγήσει σε υπερβολικά εξειδικευμένες προτάσεις, με τις οποίες ο χρήστης είναι ήδη γνώριμος (Adomavicius,Tuzhilin, 2005). Ένα ακόμα πλεονέκτημα των CB συστημάτων είναι ότι είναι ικανό να παρακολουθήσει τις αλλαγές στις προτιμήσεις του χρήστη με δεδομένο βέβαια ότι τα στοιχεία που εισάγει ο χρήστης αφορούν τον ίδιο και δεν πρόκειται για κάποιον άλλο χρήστη ή αποσκοπούν στην αγορά π.χ. κάποιου δώρου. Τέλος, ένα ακόμα μειονέκτημα των συγκεκριμένων συστημάτων είναι η αλληλεπίδραση με τους νέους χρήστες, λόγω της έλλειψης στοιχείων για αυτούς στο πρώιμο αυτό στάδιο, γνωστό και ως «ψυχρή εκκίνηση» (cold start). Ένα παρόμοιο πρόβλημα είναι αυτό της Airbnb, το οποίο θα αναλυθεί σε επόμενο κεφάλαιο.

2.3.2.2 Συστήματα συνεργατικού φιλτραρίσματος (COLLABORATIVE FILTERING SYSTEMS)

Τα συστήματα συνεργατικού φιλτραρίσματος, για συντομία CL, πραγματοποιούν συστάσεις βασιζόμενα σε ομάδες χρηστών με παρόμοιες προτιμήσεις. Η ομοιότητα μεταξύ των χρηστών υπολογίζεται συγκρίνοντας συνήθως τις αξιολογήσεις που οι ίδιοι δίνουν σε ορισμένα από τα αντικείμενα που τους συστάθηκαν. Όταν το σύστημα εντοπίσει ποιοι είναι οι χρήστες που μοιράζονται παρόμοια ενδιαφέροντα με τον τρέχοντα χρήστη, τότε προτείνει στον τελευταίο τα στοιχεία που άρεσαν στους παραπάνω χρήστες. Σε αυτή την προσέγγιση είναι απαραίτητη κάποια ανατροφοδότηση αναφορικά με τις παρεχόμενες προτάσεις, προκειμένου να διαπιστωθεί ποια στοιχεία ήθελε ή αντιπάθησε ο χρήστης (Borras, Moreno, Valls, 2014) .

Μπορούμε να διακρίνουμε δύο είδη μεθόδων των συστημάτων CL: τις μεθόδους με βάση το χρήστη και τις μεθόδους με βάση τα στοιχεία. Η πρώτη μέθοδος βρίσκει «γείτονες», όπως ονομάζονται, ενός χρήστη-στόχου ταιριάζοντας τις απόψεις του με εκείνες άλλων χρηστών του συστήματος (Kabassi, 2010). Η δεύτερη μέθοδος δημιουργεί

ομάδες βρίσκοντας ομοιότητες στα στοιχεία που προτίμησαν (ή όχι) οι χρήστες στο παρελθόν.

Όπως όλα τα συστήματα, έτσι και τα CL συστήματα διαθέτουν αδύναμα σημεία. Τα τρία που μπορούμε να κατονομάσουμε εδώ είναι αυτά της ακεραιότητας των δεδομένων, του «γκρίζου πρόβατου» και της κλιμάκωσης. Το πρώτο πρόβλημα εμφανίζεται όταν ο αριθμός των αξιολογήσεων των χρηστών είναι μικρός συγκριτικά με τον συνολικό αριθμό των αντικειμένων, έτσι ώστε η πιθανότητα εύρεσης χρηστών που βαθμολογούν τα ίδια στοιχεία να είναι πολύ χαμηλή για να μπορούν να γίνουν σωστές εκτιμήσεις. Το δεύτερο πρόβλημα, το «γκρίζο πρόβατο», αναφέρεται σε ένα χρήστη, ο οποίος διαθέτει ένα προφίλ πολύ διαφορετικό σε σχέση με τους υπόλοιπους χρήστες του συστήματος. Σε αυτή την περίπτωση είναι δύσκολο να βρεθούν οι κατάλληλες συστάσεις, λόγω της έλλειψης στοιχείων για παρόμοιους χρήστες. Τέλος, το τελευταίο πρόβλημα της κλιμάκωσης μπορεί να εμφανιστεί όταν η κοινότητα των χρηστών είναι πολύ μεγάλη (Borras, Moreno, Valls, 2014).

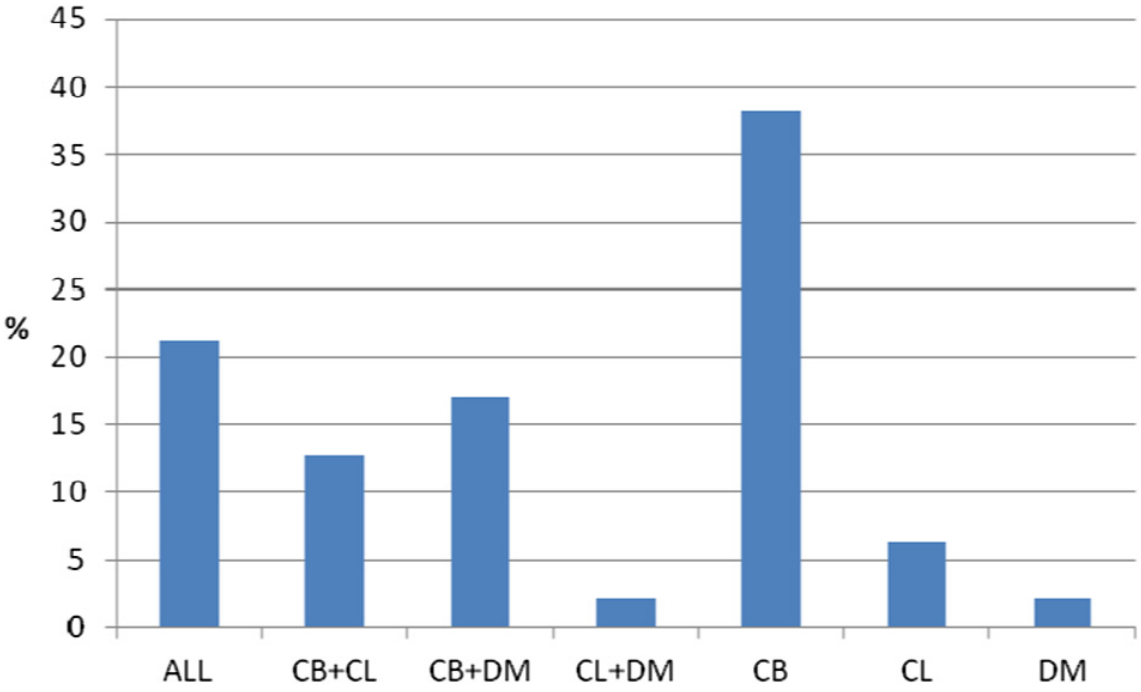
2.3.2.3 Δημογραφικά συστήματα (DEMOGRAPHIC BASED SYSTEMS)

Αυτή η κατηγορία συστημάτων, εν συντομία DM, βασίζεται στα δημογραφικά δεδομένα του χρήστη, όπως η ηλικία, η χώρα προέλευσης, το επίπεδο σπουδών κ.α. Στην περίπτωση αυτή, η σύσταση δεν βασίζεται στα ενδιαφέροντα και στις προτιμήσεις του χρήστη, αλλά στα προσωπικά του χαρακτηριστικά. Οι χρήστες κατηγοριοποιούνται συνήθως σε μια συγκεκριμένη στερεότυπη τάξη, ανάλογα με τα δημογραφικά στοιχεία τους, έτσι ώστε τα μέλη της ίδιας ομάδας να μοιράζονται ένα κοινό δημογραφικό προφίλ. Ένα DM σύστημα διαθέτει εσωτερικές γνώσεις σχετικά με τις τυποποιημένες προτιμήσεις κάθε στερεοτύπου, γεγονός που χρησιμοποιείται για να παρέχει τις συστάσεις στους χρήστες (Borras, Moreno, Valls, 2014).

Σε σύγκριση με τις άλλες προσεγγίσεις, το πλεονέκτημα ενός DM συστήματος είναι ότι δεν απαιτείται ιστορικό βαθμολογίας των χρηστών ή αξιολόγησης των αντικειμένων προς πρόταση. Ως εκ τούτου οι νέοι χρήστες μπορούν να λαμβάνουν συστάσεις, παρακάμπτοντας το πρόβλημα της «ψυχρής εκκίνησης», που αναφέρθηκε παραπάνω (Wang, Chan, & Ngai, 2012).

2.3.2.4 Υβριδικά συστήματα (HYBRID APPROACHES)

Δεδομένου ότι κάθε μία από τις προσεγγίσεις έχει κάποια μειονεκτήματα, ο συνδυασμός διαφορετικών τεχνικών αποτελεί τα τελευταία χρόνια μια ευρέως διαδεδομένη τακτική. Στο παρακάτω διάγραμμα φαίνεται η κατανομή σε ποσοστά της χρήσης των διαφορετικών τύπων των συστημάτων συστάσεων στον κλάδο του τουρισμού.



Διάγραμμα 2.1 Ποσοστά χρήσης κάθε τύπου συστήματος συστάσεων

Τα υβριδικά συστήματα ενσωματώνουν τις τεχνικές αυτές με διαφορετικούς τρόπους. Μπορούμε ενδεικτικά να αναφερθούμε σε τρεις προσεγγίσεις (Borras, Moreno, Valls, 2014):

i. Επιλογή της μεθόδου: το υβριδικό σύστημα ενσωματώνει μεθόδους DM, CB και CL, αλλά μόνο μία από αυτές εφαρμόζεται ανάλογα με την ιδιαιτερότητα κάθε χρήστη. Παραδείγματος χάριν, κατά την πρώτη επαφή με το χρήστη, χρησιμοποιείται μία μέθοδος που βασίζεται σε δημογραφικά δεδομένα. Εν συνεχεία, εάν μπορέσει το σύστημα να ταυτοποιήσει το χρήστη με μία ομάδα, πραγματοποιείται μία σύσταση με βάση ένα CL σύστημα, αλλιώς, σε διαφορετική περίπτωση, εφαρμόζεται διαδικασία βάσει ενός CB συστήματος.

ii. Διαδοχική χρήση: κάθε τεχνική συστάσεων χρησιμοποιείται σε διαφορετικά στάδια της διαδικασίας. Ένα πολύ καλό παράδειγμα αυτής της τεχνικής αποτελεί το σύστημα SPETA, το οποίο χρησιμοποιεί τη γνώση της τρέχουσας θέσης του χρήστη, τις προτιμήσεις καθώς και το ιστορικό των τοποθεσιών, προκειμένου να παρέχει το είδος των υπηρεσιών που οι χρήστες περιμένουν από έναν πραγματικό ξεναγό. Για να το πετύχει αυτό ακολουθεί τέσσερα στάδια: αρχικά χρησιμοποιούνται σχετικές πληροφορίες, όπως η θέση ή η ώρα, για να γίνει η πρώτη σύσταση. Στη συνέχεια, επιτυγχάνεται ένα πιο ακριβές σύνολο αποτελεσμάτων χρησιμοποιώντας τεχνικές φιλτραρίσματος που βασίζονται στη γνώση, υπολογίζοντας τη σημασιολογική ομοιότητα μεταξύ των προτιμήσεων των χρηστών και των τουριστικών υπηρεσιών. Στο τρίτο στάδιο χρησιμοποιούνται CL τεχνικές για να βελτιωθεί το σύνολο των επιλογών, ενώ τέλος χρησιμοποιείται ένα διάνυσμα προτιμήσεων για την τελική επιλογή (Grespo, Chamizo, Mencke, Colomo, Gomez, 2009).

iii. Ολοκληρωμένη χρήση: κατά την εκτέλεση συνδυάζονται οι τεχνικές CB και CL. Ένα παράδειγμα αυτού του συνδυασμού αποτελεί το SigTur (Borras, Flor, Perez, Moreno, Valls, Isern, Orellana, Russo, Clave, 2011). Το SigTur αποτελεί ένα σύστημα συστάσεων το οποίο σχεδιάστηκε από το Τεχνολογικό Πάρκο Τουρισμού και Αναψυχής της πόλης Vila-Seca, στην επαρχία της Ταραγόνας, με τη χρηματοδότηση της Ευρωπαϊκής Περιφερειακής Ανάπτυξης. Στο SigTur, λοιπόν, υπολογίζονται διαφορετικές αξιολογήσεις για να εκτιμηθεί το ενδιαφέρον μιας δραστηριότητας για ένα χρήστη – στόχο. Οι αξιολογήσεις λαμβάνονται με χρήση ομαδοποίησης DM, CL και CB. Στη συνέχεια, οι αξιολογήσεις αυτές συγχωνεύονται για να βρεθεί μια συνολική βαθμολογία ποιότητας για κάθε στοιχείο και να γίνει το φιλτράρισμα των καλύτερων τουριστικών αξιοθέατων. Ένα ακόμα παράδειγμα αποτελεί η πρόταση του υβριδικού συστήματος του Lucas (Lucas, 2013), στο οποίο οι χρήστες ταξινομούνται σε ομάδες χρησιμοποιώντας ταυτόχρονα προσωπικά δημογραφικά στοιχεία (DM), πληροφορίες σχετικά με το περιεχόμενο των αντικειμένων που έχουν επιλεγεί προηγουμένως από το χρήστη (CB) και

τις πληροφορίες άλλων χρηστών (CL). Εν συνεχεία, δημιουργείται αυτόματα μια σειρά από ασαφείς κανόνες, ώστε οι νέοι χρήστες να μπορούν να ταξινομηθούν αυτόματα σε διάφορες ομάδες με διαφορετικούς βαθμούς συμμετοχής. Η λίστα ή τα συνιστώμενα στοιχεία προέρχονται τελικά από μία πρόβλεψη που βασίζεται στις ομάδες, στις οποίες ανήκει ο χρήστης.

Τα τελευταία χρόνια, και συγκεκριμένα μετά το 2012, παρατηρείται γενικότερα μια αυξανόμενη τάση στη χρήση των συστημάτων συνεργατικού φιλτραρίσματος (CL), κυρίως σε υβριδικά συστήματα. Πιο συγκεκριμένα, από το 2008 έως το 2011, μόνο το 25% των συστημάτων συστάσεων χρησιμοποίησε αυτή τη μέθοδο, ενώ από το 2012, το ποσοστό αυξήθηκε σε 75%.

3. Μεθοδολογία

3.1. Εισαγωγή

Σε αυτό το κεφάλαιο θα περιγράψουμε τη μεθοδολογική προσέγγιση που θα ακολουθήσουμε για την ανάλυση του προβλήματος και της λύσης του. Θα χωρίσουμε την προσέγγισή μας σε πέντε βήματα, τα οποία περιγράφονται παρακάτω.

3.2. Τα βήματα της μεθοδολογίας

Η μεθοδολογίας μας αποτελείται από τα εξής βήματα:

- Συλλογή πληροφοριών και στοιχείων από την υπάρχουσα βιβλιογραφία και από ειδησεογραφικές πηγές αναφορικά με τη χρήση της επιστήμης των δεδομένων στην Airbnb. Παρουσίαση του συστήματος συστάσεων που χρησιμοποιεί η εταιρεία.

- Ανάλυση του προβλήματος για το οποίο αναζήτησε λύση η Airbnb μέσω του Kaggle. Παρουσίαση των βασικών στοιχείων του διαγωνισμού. Επισκόπηση της λίστας των λύσεων που υποβλήθηκαν και ορισμός κριτηρίων για την επιλογή μίας εκ των λύσεων.

- Διερευνητική ανάλυση των δεδομένων του προβλήματος. Αφού παρουσιάσουμε τα δεδομένα που παραχώρησε η Airbnb στους συμμετέχοντες του διαγωνισμού, θα περιγράψουμε τις μεταβλητές τους. Στο τμήμα της διερευνητικής ανάλυσης, θα αναλύσουμε τα διαθέσιμα σύνολα δεδομένων και θα συνοψίσουμε τα κύρια χαρακτηριστικά τους, οπτικοποιώντας τα σημαντικότερα από αυτά με διαγράμματα.

- Αναπαραγωγή του κώδικα της λύσης που επιλέξαμε, εκτελώντας τον μέσα από το Jupiter Notebook της Kaggle. Τμηματοποίηση και σχολιασμός του κώδικα. Διαχωρισμός της λύσης στις φάσεις της προεπεξεργασίας των δεδομένων και της δημιουργίας και εκπαίδευσης του μοντέλου.

- Διατυπώνουμε τα συμπεράσματα στα οποία καταλήγουμε. Παραθέτουμε τα προβλήματα, τους περιορισμούς και τις αδυναμίες που αντιμετωπίσαμε.

Κάθε ένα από τα παραπάνω βήματα θα αποτελέσει ένα ξεχωριστό κεφάλαιο στη συνέχεια της παρούσας εργασίας.

4. Η χρήση της επιστήμης δεδομένων στην Airbnb

4.1. Εισαγωγή

Από τα πρώτα της βήματα, η Airbnb δεν έκρυψε την εκτενή χρήση της επιστήμης των δεδομένων για να δημιουργήσει νέες προσφορές προϊόντων, να βελτιώσει τις υπηρεσίες της και να αξιοποιήσει νέους τρόπους μάρκετινγκ. Ο Riley Newman, πρώην επικεφαλής του τμήματος της επιστήμης των δεδομένων στην Airbnb, εξηγεί ότι η εταιρεία βλέπει και εξετάζει τα δεδομένα ως τη φωνή του πελάτη και την επιστήμη των δεδομένων ως την ερμηνεία αυτής της φωνής.

4.2. Σύστημα συστάσεων στην Airbnb

Όταν ένας χρήστης ψάχνει για κατάλυμα που θα τον φιλοξενήσει στην επόμενη απόδρασή του, η Airbnb ταξινομεί και προβάλει τη λίστα των αποτελεσμάτων αναζήτησης, με καθόλου τυχαίο τρόπο. Υπάρχει πληθώρα μεταβλητών που ταξινομούν τις δεκάδες χιλιάδες καταχωρίσεις που είναι διαθέσιμες για μια συγκεκριμένη τοποθεσία. Σε αντίθεση με τα μηχανήματα, είναι αδύνατο για έναν άνθρωπο να αναζητήσει ολόκληρες τις λίστες αποτελεσμάτων και ειδικά για αναποφάσιστους χρήστες, αυτό θα δημιουργούσε μεγάλο πρόβλημα. Γι’ αυτό οι αλγόριθμοι μηχανικής μάθησης της Airbnb αναλαμβάνουν αυτή τη δουλειά για τους χρήστες της, εστιάζοντας στα σωστά δεδομένα και παρέχοντας εξατομικευμένες πληροφορίες, τόσο για τους φιλοξενούμενους όσο και για τους οικοδεσπότες.

Η βελτιστοποίηση των αντιστοιχιών μεταξύ φιλοξενουμένων και οικοδεσποτών θα είναι κρίσιμη για την επιτυχία της Airbnb καθώς αυτή συνεχίζει να αναπτύσσεται. Η ποικιλία σε τύπους καταλυμάτων που έχει η Airbnb θα συνεχίσει να είναι το μεγάλο της πλεονέκτημα, αρκεί να εξασφαλίζει ότι οι επισκέπτες μπορούν εύκολα να βρουν έναν οικοδεσπότη που ικανοποιεί τα κριτήρια τους. Καθώς η Airbnb συνεχίζει να προσθέτει ακόμη περισσότερα στα ήδη 4 εκατομμύρια καταλύματα που έχει στους καταλόγους της, το να εξασφαλίζει ότι τόσο οι φιλοξενούμενοι όσο και οι οικοδεσπότες θα είναι ικανοποιημένοι, θα γίνεται όλο και πιο κρίσιμο. Εάν οι χρήστες μπορούν να βρουν

ακριβώς το κατάλυμα που ψάχνουν, ειδικά αν είναι σε φθηνότερη τιμή, είναι απίθανο να επιστρέψουν στη χρήση ξενοδοχείων.

Η Airbnb μέχρι τώρα έχει καταφέρει να κάνει μια εκπληκτική δουλειά στη βελτιστοποίηση της αντιστοίχισης επισκέπτη-οικοδεσπότη. Σύμφωνα με το ειδησεογραφικό άρθρο «The 5 Machine Learning Use Cases that Optimize Your Airbnb Travel Experience» υπάρχουν 5 σημαντικές περιπτώσεις χρήσης της μηχανικής μάθησης που αναπτύσσονται επί του παρόντος από τους μηχανικούς της επιστήμης δεδομένων της Airbnb. Αυτές οι περιπτώσεις χρήσεις είναι οι ακόλουθες:

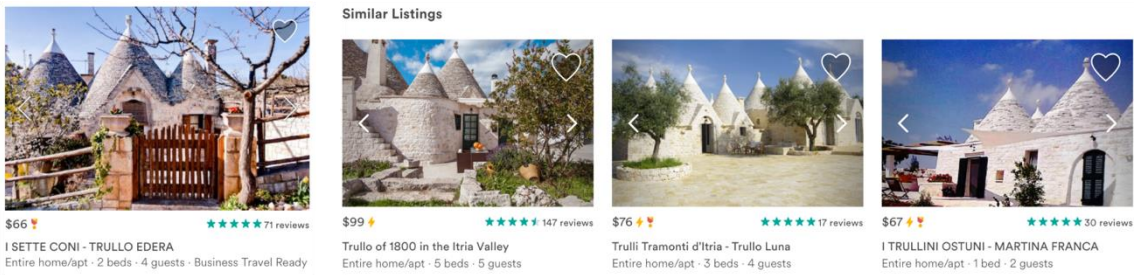
- Εντοπισμός προτιμήσεων των οικοδεσποτών. Ο στόχος αυτού του συστήματος μηχανικής μάθησης είναι να ανακαλύψει τι επηρεάζει τις αποφάσεις των οικοδεσποτών για αποδοχή αιτημάτων στέγασης και πώς η Airbnb θα μπορούσε να αυξήσει τις αποδοχές και τα «ζευγαρώματα» φιλοξενούμενων - οικοδεσποτών στην πλατφόρμα της.

- Παροχή πληροφοριών στους οικοδεσπότες. Ο στόχος αυτού του συστήματος μηχανικής μάθησης είναι να απαντήσει στην κοινή ερώτηση από τους οικοδεσπότες της Airbnb για το πώς μπορούν να επιλέξουν τη σωστή τιμή. Ο καθορισμός μιας τιμής μπορεί να είναι δύσκολος χωρίς αξιόπιστες πληροφορίες σχετικά με άλλες καταχωρήσεις στην περιοχή του οικοδεσπότη, τις τάσεις των ταξιδιών καθώς και το ενδιαφέρον που έχουν οι άνθρωποι για την παροχή προσφερόμενης φιλοξενίας.

- Προβλέψεις τιμών για τα καταλύματα. Στην Airbnb, η πρόβλεψη των τιμών είναι μια περίπτωση χρήσης της μοντελοποίησης της αξίας πελατείας, η οποία καταγράφει την προβαλλόμενη αξία ενός χρήστη για ένα συγκεκριμένο χρονικό ορίζοντα. Σε εταιρείες όπως η Airbnb, η γνώση για την αξία των καταλυμάτων των χρηστών τους, τους επιτρέπουν να κατανείμουν πιο αποτελεσματικά τον προϋπολογισμό τους σε διαφορετικά κανάλια μάρκετινγκ, να υπολογίζουν ακριβέστερα τιμές υποβολής προσφορών για διαδικτυακό μάρκετινγκ βασισμένες σε λέξεις-κλειδιά και να δημιουργούν καλύτερες ομαδοποιήσεις των καταχωρίσεων.

- Καταπολέμηση της οικονομικής απάτης. Η καταπολέμηση της οικονομικής απάτης είναι ένα από τα πιο σημαντικά καθήκοντα της Airbnb για να εξασφαλίσει την εμπιστοσύνη στην πλατφόρμα της. Η εταιρεία κάνει χρήση της μηχανικής μάθησης, με πειραματισμούς και αναλυτικά στοιχεία για τον εντοπισμό και αποκλεισμό των απατεώνων, ελαχιστοποιώντας τις επιπτώσεις στους καλούς χρήστες που είναι η συντριπτική πλειοψηφία των χρηστών της.

- Εξατομικευμένες προτάσεις καταχωρίσεων. Πολύ πρόσφατα, η Airbnb ανέπτυξε μια τεχνική ενσωμάτωσης καταχωρήσεων με σκοπό τη βελτίωση παρόμοιων συστάσεων για καταλύματα και εξατομίκευση της λίστας σε πραγματικό χρόνο κατά την ταξινόμηση των αποτελεσμάτων της αναζήτησης. Οι ενσωματώσεις (embeddings) είναι στοχευμένες απεικονίσεις κατοικιών της Airbnb, η εκμάθηση προβολής των οποίων γίνεται μέσα από παρελθοντικές αναζητήσεις που επιτρέπουν στην εταιρία να μετρήσει τις ομοιότητες μεταξύ των καταλυμάτων. Κωδικοποιούν αποτελεσματικά πολλές παραμέτρους των καταχωρίσεων, όπως η τοποθεσία, η τιμή, ο τύπος του καταλύματος, η αρχιτεκτονική και το στυλ του καταλύματος.



Εικόνα 4.1 Ενσωμάτωση στοχευμένων απεικονίσεων κατοικιών

Οι ενσωματώσεις είναι μια ιδέα που προήλθε από την επεξεργασία της φυσικής γλώσσας για την αναπαράσταση λέξεων. Όπως ακριβώς μπορεί κανείς να εκπαιδεύσει ενσωματώσεις λέξεων αντιμετωπίζοντας μια ακολουθία λέξεων σε μια πρόταση ως πλαίσιο, το ίδιο μπορεί να γίνει και για την εκπαίδευση ενσωματώσεων ενεργειών των χρηστών, αντιμετωπίζοντας την ακολουθία των ενεργειών του χρήστη ως πλαίσιο. Παραδείγματα αποτελούν η εκμάθηση αναπαράστασης αντικειμένων στα οποία έγινε κλικ ή αγορά ή αναζήτηση καθώς και διαφημίσεις στις οποίες έγινε κλικ. Αυτές οι ενσωματώσεις χρησιμοποιούνται στη συνέχεια για ποικίλες συστάσεις στο διαδίκτυο.

Στην Airbnb, εκπαίδευσαν και βελτιστοποίησαν τα μοντέλα στην εκμάθηση των ενσωματώσεων για 4,5 εκατομμύρια ενεργές καταχωρίσεις, χρησιμοποιώντας περισσότερες από 800 εκατομμύρια συνεδρίες που είχαν κλικ αναζήτησης, με αποτέλεσμα την υψηλής ποιότητας παρουσίαση αποτελεσμάτων.

Δεδομένου ότι γίνονται πολλές αναζητήσεις πριν από μια κράτηση, κάνοντας κλικ σε διάφορες καταχωρίσεις ή από την επικοινωνία με διάφορους οικοδεσπότες, χρησιμοποιούνται τα σήματα από τη διάρκεια των συνεδριών αναζήτησης, δηλαδή τα κλικ και η επικοινωνία με οικοδεσπότες, με στόχο την εξατομίκευση των αποτελεσμάτων σε πραγματικό χρόνο. Στόχος είναι να εμφανιστούν στο χρήστη περισσότερες καταχωρίσεις παρόμοιες με αυτές που πιστεύεται ότι του άρεσαν από την έναρξη της αναζήτησης. Παράλληλα χρησιμοποιούνται αρνητικά σήματα όπως η προσπέραση αποτελεσμάτων που βρισκόταν υψηλά στη λίστα, ώστε να εμφανίζονται λιγότερα αποτελέσματα με αυτά που πιστεύεται ότι δεν άρεσαν στο χρήστη.

Για να είναι εφικτός ο υπολογισμός των ομοιοτήτων μεταξύ των καταχωρίσεων με τις οποίες ο επισκέπτης αλληλοεπιδρά και των υποψήφιων καταχωρίσεις προς ταξινόμηση, χρησιμοποιούνται ενσωματώσεις καταχωρίσεων. Αξιοποιούνται οι ομοιότητες για τη δημιουργία εξατομικευμένης λίστας αποτελεσμάτων αναζήτησης και της λίστας συστάσεων παρόμοιων καταχωρίσεων. Τα δύο αυτά μέσα, οδηγούν στο 99% των κρατήσεων της Airbnb. (Mihajlo Grbovic, Haibin Cheng, 2018)

Είναι σαφές ότι το σύστημα συστάσεων της Airbnb είναι εξαιρετικά επιτυχημένο στην παροχή προσαρμοσμένων ταξιδιωτικών εμπειριών για εκατομμύρια χρήστες. Εκτός από αυτές τις χρήσεις, η Airbnb γενικά έχει επενδύσει σε μεγάλο βαθμό στην επιστήμη των δεδομένων. Από την κεντρική βάση γνώσεων, η οποία αρχειοθετεί και μεταφέρει τη γνώση σε ολόκληρο τον οργανισμό, στο Superset που κλιμακώνει την πρόσβαση στα δεδομένα και στις οπτικές πληροφορίες. Από το Dataportal που παρέχει πολύτιμους πόρους και εργαλεία που καλύπτουν ζητήματα της επιστήμης δεδομένων, μέχρι το σύστημα Automated ML, το οποίο αυξάνει κατά πολύ την παραγωγικότητα των ειδικών, η επιστήμη των δεδομένων χρησιμοποιείται εκτενώς σε ατομικό, ομαδικό και οργανωσιακό επίπεδο.

4.3. Συμπεράσματα

Με σταθερή υποδομή δεδομένων, εξελιγμένα εσωτερικά εργαλεία και αξιόπιστη βάση δεδομένων, η Airbnb είναι αναμφισβήτητα μία από τις καλύτερες εταιρίες τεχνολογίας που βασίζονται σε δεδομένα και αξιοποιούν τις νέες αυτές τεχνολογικές τάσεις. Τα παραπάνω, αποτελούν λόγους που συνέβαλαν στην απόφαση να εστιάσουμε στην παρούσα εργασία σε παράδειγμα προβλήματος που αντιμετώπισε η εταιρία Airbnb.

5. Επιλογή λύσης

5.1. Εισαγωγή

Όπως αναφέραμε στο προηγούμενο κεφάλαιο, υπάρχει ένα ευρύ φάσμα χρήσης της επιστήμης των δεδομένων στην Airbnb. Στη συνέχεια θα επικεντρωθούμε σε μια συγκεκριμένη περίπτωση των εξατομικευμένων προτάσεων των καταχωρίσεων. Κάθε χρήστης έχει τις δικές του προσωπικές ανάγκες και επιθυμίες. Όσο περισσότερο χρησιμοποιεί την πλατφόρμα για τις ταξιδιωτικές του εμπειρίες, τόσο καλύτερα τον «μαθαίνει» το σύστημα, πετυχαίνοντας σταδιακά όλο και πιο επιτυχημένες εξατομικευμένες προτάσεις.

Μια ιδιαίτερη περίπτωση αποτελούν νέοι χρήστες, οι οποίοι έγιναν πρόσφατα μέλη της πλατφόρμας και δεν έχουν προβεί ακόμη σε κρατήσεις. Καθώς δεν υπάρχει ιστορικό κρατήσεων, θα πρέπει να γίνει προσπάθεια πρόβλεψης του πρώτου προορισμού αυτών των χρηστών, ώστε να είναι εφικτή και σε αυτή την κατηγορία χρηστών μια στοχευμένη χρήση των εξατομικευμένων συστάσεων. Για την εύρεση της βέλτιστης λύσης που θα λύνει αυτό ακριβώς το πρόβλημα, η Airbnb απευθύνθηκε στην κοινότητα της πλατφόρμας Kaggle, δημιουργώντας ένα διαγωνισμό με αυτό το θέμα.

5.2. Kaggle

Το Kaggle είναι μια online κοινότητα επιστημόνων δεδομένων και ειδικών στη μηχανική μάθηση, το οποίο ανήκει στην Google Inc από το Μάρτιο του 2017. Το Kaggle επιτρέπει στους χρήστες να βρίσκουν και να δημοσιεύουν σύνολα δεδομένων, να εξερευνούν και να δημιουργούν μοντέλα σε ένα web-based περιβάλλον επιστήμης δεδομένων, να συνεργάζονται με άλλους επιστήμονες δεδομένων και μηχανικούς μηχανικής μάθησης, και να συμμετάσχουν σε διαγωνισμούς για την επίλυση προκλήσεων της επιστήμης δεδομένων.

Το Kaggle ξεκίνησε προσφέροντας διαγωνισμούς στη μηχανική μάθηση αλλά πλέον προσφέρει επίσης μια πλατφόρμα δημόσιων δεδομένων, ένα πλαίσιο εργασίας βασισμένο σε cloud για την επιστήμη δεδομένων καθώς και μια σύντομη εκπαίδευση

στην τεχνητή νοημοσύνη. Οι χρήστες έχουν τη δυνατότητα να συμμετέχουν σε διάφορους διαγωνισμούς, αντιμετωπίζοντας προβλήματα ανάλυσης μεγάλου όγκου δεδομένων. Το άτομο ή η ομάδα που επιτυγχάνει το καλύτερο αποτέλεσμα μπορεί να λάβει χρηματικό έπαθλο, ευκαιρία εργασίας για μια εταιρεία ή αναγνώριση της κοινότητας για την προσπάθεια και την καινοτομία.

Οποιοσδήποτε επιθυμεί, μπορεί να εγγραφεί και να προσπαθήσει να λύσει ένα πρόβλημα από μια μεγάλη ποικιλία προκλήσεων. Αυτή η απλή ιδέα προσέλκυσε περισσότερους από 1.000.000 εγγεγραμμένους χρήστες, καθιστώντας την τη μεγαλύτερη επιστημονική κοινότητα στον κόσμο. Οι χρήστες ανήκουν σε όλα τα είδη επαγγελματικών κλάδων, τα οποία εμπλουτίζουν τις λύσεις και επιτρέπουν την επίλυση προβλημάτων διαφορετικής φύσης. Πολλές εταιρείες το χρησιμοποιούν ως εργαστήριο για την επίλυση προβλημάτων που έχουν με τα δεδομένα τους. Η ποικιλία των συμμετεχόντων δίνει στις εταιρείες τη δυνατότητα να αξιολογήσουν τις διάφορες προτεινόμενες λύσεις και να επιλέξουν τον νικητή για το διαγωνισμό τους. Μια τέτοια διάκριση προσφέρει σήμερα ένα ορισμένο κύρος στον κόσμο της ανάλυσης δεδομένων, και όπως προαναφέρθηκε, οι εταιρείες προσφέρουν επίσης ένα χρηματικό έπαθλο που μερικές φορές μπορεί να είναι αρκετά δελεαστικό.

Ο διαγωνισμός που αναλύουμε εδώ από την Airbnb ήταν ένας διαγωνισμός πρόσληψης, όπου οι υποβαλλόμενες λύσεις αξιολογήθηκαν από την εταιρεία και εκείνοι που εντυπωσίασαν περισσότερο με τις απαντήσεις τους, θα εξεταζόταν για συνέντευξη που θα τους έδινε την ευκαιρία να συμμετάσχουν στην ομάδα Data Science και Analytics της Airbnb. Χρησιμοποιήθηκε μια συγκεκριμένη μέτρηση για την αξιολόγηση των λύσεων που συμμετείχαν στο διαγωνισμό. Η ορθότητα κάθε υποβληθείσας λύσης για την πρόβλεψη, υπολογίστηκε με τον αλγόριθμο βαθμολόγησης NDCG5.

Διάγραμμα 5.1 Λειτουργικό διάγραμμα της Kaggle

5.3. Ο διαγωνισμός

Ο διαγωνισμός που παρουσιάζουμε σε αυτή την εργασία δημοσιεύθηκε με το όνομα "Airbnb New User Bookings" και ήταν ενεργός στην πλατφόρμα Kaggle για 4 μήνες, από τις 25 Νοεμβρίου έως τις 11 Φεβρουαρίου 2016. (https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings)

Οι νέοι χρήστες της Airbnb μπορούν να κάνουν κράτηση για να μείνουν σε 34.000+ πόλεις σε 190+ χώρες. Ο στόχος του διαγωνισμού ήταν η δημιουργία ενός μοντέλου που θα προβλέπει τον πρώτο προορισμό ενός νέου χρήστη, από 12 πιθανά αποτελέσματα για τη χώρα προορισμού: Ηνωμένες Πολιτείες, Γαλλία, Καναδά, Μεγάλη Βρετανία, Ισπανία, Ιταλία, Πορτογαλία, Ολλανδία, Γερμανία, Αυστραλία, προορισμός δεν βρέθηκε (NDF).

Με την ακριβή πρόβλεψη του προορισμού ενός νέου χρήστη και κατ' επέκταση την πιθανότερη κράτηση για την πρώτη του ταξιδιωτική εμπειρία μέσω της πλατφόρμας, η Airbnb μπορεί να παρέχει στοχευμένο περιεχόμενο στους χρήστες της, να μειώσει τον μέσο χρόνο μέχρι την ολοκλήρωση της πρώτης κράτησης και να αυξήσει την πιθανότητα ότι οι νέοι χρήστες τελικά θα προβούν σε κράτηση μέσω της πλατφόρμας. Αυτή η πρόβλεψη θα βοηθήσει επίσης την εταιρεία να έχει καλύτερη πρόβλεψη της ζήτησης. Γνωρίζοντας ποιες πόλεις θα έχουν τη μεγαλύτερη ζήτηση και σε ποιες ημερομηνίες, μπορεί να παρέχει τη δυνατότητα εκτίμησης και καθορισμού των τιμών για τους ιδιοκτήτες των ακινήτων.

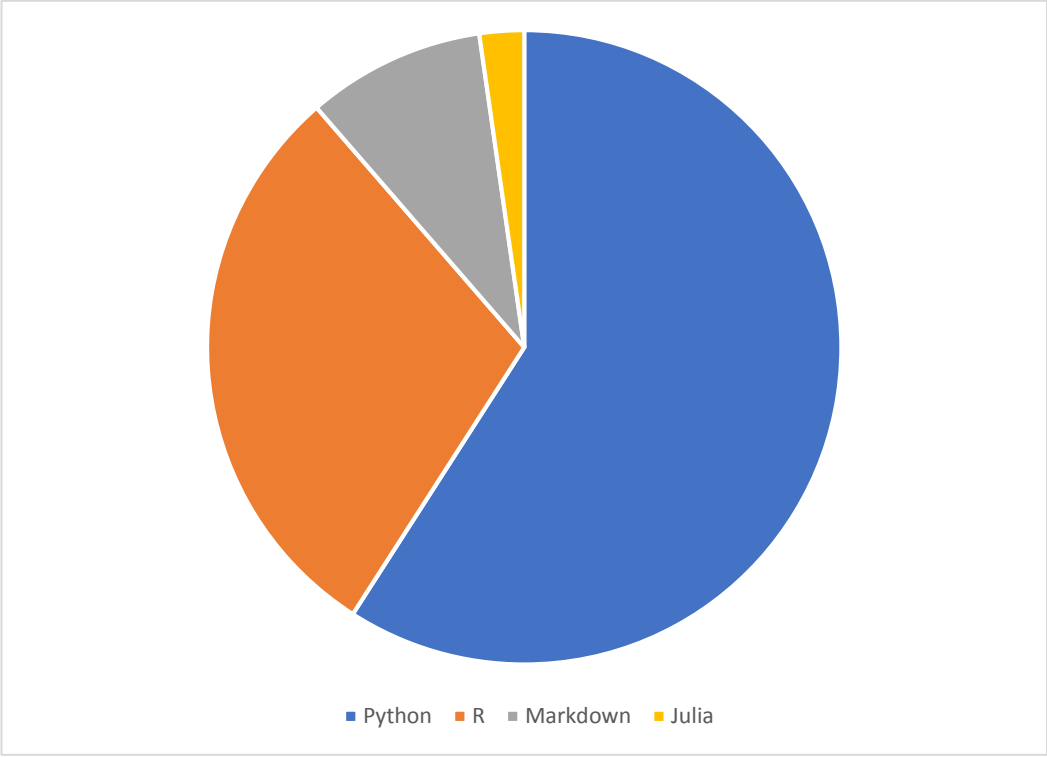
Η συμμετοχή στο διαγωνισμό ήταν μεγάλη συγκεντρώνοντας συνολικά 1471 υποβληθείσες λύσεις. Εμείς επικεντρωθήκαμε στην ανάλυσή μας στις υλοποιήσεις (Kernel) που είχαν τουλάχιστον δύο ψήφους από την κοινότητα της Kaggle, οι οποίες είναι στο σύνολό τους 44. Στη συνέχεια του κεφαλαίου όπου θα αναφερόμαστε στις υλοποιήσεις, θα εννοούμε αυτό το συγκεκριμένο σύνολο.

Παρατηρήσαμε ότι οι περισσότερες από τις υλοποιήσεις έχουν παρόμοια προσέγγιση, χρησιμοποιούν τους ίδιους αλγορίθμους και διαφοροποιούνται σε μικρές λεπτομέρειες, όπως οι ρύθμιση των παραμέτρων. Αυτές οι «λεπτομέρειες» βέβαια, έχουν μεγάλη σημασία για το τελικό αποτέλεσμα, αφού επηρεάζουν σε σημαντικό βαθμό την αποτελεσματικότητα της λύσης. Για την επιλογή της λύσης που θα παρουσιάσουμε αναλυτικά σε επόμενο κεφάλαιο, λήφθηκαν υπόψη διάφορα κριτήρια όπως η γλώσσα

προγραμματισμού που χρησιμοποιήθηκε για τον κώδικα της λύσης, η προσέγγιση μηχανικής μάθησης, ο αλγόριθμος που χρησιμοποιήθηκε, η πολυπλοκότητα και η τεκμηρίωση της λύσης καθώς και η ανταπόκριση από την κοινότητα της Kaggle, μετρήσιμη σε αριθμό ψήφων (votes).

5.4. Γλώσσα προγραμματισμού

Διερευνώντας τις διαθέσιμες υλοποιήσεις, παρατηρήσαμε ότι υπάρχει δεν υπάρχει μεγάλη ποικιλία στην επιλογή της γλώσσα προγραμματισμού που χρησιμοποιείται για τον κώδικα που θα οδηγήσει στη λύση. Συγκεκριμένα από τις 44 λύσεις, στις 26 χρησιμοποιήθηκε η Python, στις 13 η R, στις 4 η Markdown και σε μια η Julia. Όπως φαίνεται και στο διάγραμμα 5.2, είναι ξεκάθαρο ότι η προτιμώμενη γλώσσα προγραμματισμού σε αυτό το διαγωνισμό, είναι η Python. Ως εκ τούτου, η λύση που θα παρουσιάσουμε στη συνέχεια, θα είναι μια από αυτές όπου ο κώδικας έχει γραφτεί σε γλώσσα Python.



Διάγραμμα 5.2 Προτίμηση γλώσσας προγραμματισμού

Η Python εμφανίστηκε τις τελευταίες δεκαετίες ως εργαλείο πρώτης κατηγορίας για επιστημονικές υπολογιστικές εργασίες, συμπεριλαμβανομένης της ανάλυσης και απεικόνισης μεγάλων συνόλων δεδομένων. Η ίδια η γλώσσα δεν σχεδιάστηκε ειδικά για την ανάλυση δεδομένων ή την επιστημονική υπολογιστική. Η χρησιμότητα της Python για την επιστήμη των δεδομένων προέρχεται κυρίως από το μεγάλο και ενεργό οικοσύστημα πακέτων τρίτων, με σημαντικότερες τις βιβλιοθήκες NumPy και Pandas.

Το όνομα NumPy σημαίνει "Numeric Python" ή "Αριθμητική Python". Πρόκειται για μια βιβλιοθήκη ανοιχτού κώδικα της Python, η οποία παρέχει γρήγορους μαθηματικούς υπολογισμούς σε πίνακες και μήτρες. Η NumPy παρέχει τις βασικές πολυδιάστατες λειτουργίες υπολογιστικής προσανατολισμένες σε πίνακες, σχεδιασμένες για μαθηματικές λειτουργίες υψηλού επιπέδου και επιστημονικούς υπολογισμούς. Το κύριο αντικείμενο της NumPy είναι η ομοιογενής πολυδιάστατοι πίνακες, ένας πίνακας με στοιχεία του ίδιου τύπου, δηλ. χαρακτήρες (ομοιογενείς) και συνήθως ακέραιοι.

Παρόμοια με τη NumPy, η Pandas είναι μία από τις πιο ευρέως χρησιμοποιούμενες βιβλιοθήκες της Python στην επιστήμη των δεδομένων. Παρέχει δομές υψηλής ανάλυσης, εύχρηστες δομές και εργαλεία ανάλυσης δεδομένων. Σε αντίθεση με την βιβλιοθήκη NumPy που παρέχει αντικείμενα για πολυδιάστατους πίνακες, η Pandas παρέχει πίνακα 2 διαστάσεων που ονομάζεται Dataframe. Είναι σαν ένα υπολογιστικό φύλλο με ονόματα στηλών και ετικέτες γραμμών. Ως εκ τούτου, με τους πίνακες δύο διαστάσεων, η Pandas είναι ικανή να παρέχει πολλές πρόσθετες λειτουργίες, όπως δημιουργία πινάκων pivot, υπολογισμό στηλών με βάση άλλες στήλες και γραφική αναπαράσταση αποτελεσμάτων. (Fabio Nelli, 2015)

5.5. Εποπτευόμενη μηχανική μάθηση

Η προσέγγιση της μηχανικής μάθησης θα είναι η εποπτευόμενη μηχανική μάθηση καθώς το σύνολο δεδομένων που παρέχεται από την Airbnb για το διαγωνισμό, έχει σχεδιαστεί για αυτή τη μέθοδο. Έχουμε πρόσβαση σε ένα σύνολο δεδομένων εκπαίδευσης (training set) με τα σωστά αποτελέσματα και ένα σύνολο δεδομένων δοκιμών (test set) χωρίς αποτελέσματα.

Η εποπτευόμενη μηχανική μάθηση είναι μια τεχνική όπου ο υπολογιστής χρησιμοποιεί δεδομένα εκπαίδευσης με γνωστά αποτελέσματα, για να δημιουργήσει ένα μοντέλο το οποίο θα μπορεί να εφαρμόσει σε δεδομένα όπου το αποτέλεσμα είναι άγνωστο, ώστε να προβλέψει το αποτέλεσμα αυτών των δεδομένων. Το μοντέλο που παράγεται από αυτή τη διαδικασία μάθησης μπορεί να χωριστεί σε δύο κατηγορίες: ταξινόμησης και παλινδρόμησης. Ένα μοντέλο τύπου ταξινόμησης στοχεύει να βάλει μια ετικέτα σε ένα σύνολο δεδομένων, όπως η πρόβλεψη εάν ένα μήνυμα ηλεκτρονικού ταχυδρομείου είναι spam ή όχι. Τα μοντέλα παλινδρόμησης χρησιμοποιούνται για την πρόβλεψη συνεχών μετρήσεων, αυτό μπορεί να χρησιμοποιηθεί για την πρόβλεψη της αλλαγής του καιρού με βάση τα τρέχοντα δεδομένα που υπάρχουν για τον καιρό.

Η ροή εργασίας όταν εργαζόμαστε με εποπτευόμενη μηχανική μάθηση είναι η εξής (Kotsiantis, 2007):

- Ανάλυση δεδομένων: Τα δεδομένα πρέπει να αναλυθούν ώστε να βοηθήσουμε τον υπολογιστή να τα κατανοήσει όσο το δυνατόν καλύτερα. Κατά τη διάρκεια αυτού του βήματος θα πρέπει να μορφοποιήσουμε τα δεδομένα σε μια πιο κατανοητή μορφή και να αφαιρέσουμε περιττές καταχωρήσεις.

- Απόφαση αλγορίθμου: Η απόφαση με ποιον αλγόριθμο θα δουλέψουμε εξαρτάται από το ποιοι είναι οι περιορισμοί του συστήματός μας. Ορισμένοι αλγόριθμοι χρησιμοποιούν πολύ λιγότερη μνήμη, αλλά μπορεί να χρειάζονται περισσότερο χρόνο για να παράγουν ένα μοντέλο από κάποιους άλλους. Άλλοι αλγόριθμοι παράγουν καλύτερα μοντέλα για προβλέψεις αλλά η συλλογιστική τους μπορεί να είναι πολύ πολύπλοκη για να την καταλάβουμε, το οποίο καθιστά την ρύθμιση παραμέτρων πολύ πιο δύσκολη.

- Επικύρωση μοντέλου: Όταν ο επιλεγμένος αλγόριθμος έχει δημιουργήσει ένα μοντέλο πρέπει να επικυρώσουμε την ακρίβεια του μοντέλου. Υπάρχουν αρκετές προσεγγίσεις γι' αυτό, και μία από αυτές είναι διασταυρωμένη επικύρωση. Η διασταυρωμένη επικύρωση είναι μια μέθοδος όπου διαιρούμε το σύνολο δεδομένων εκπαίδευσης σε Ν αριθμό υποσυνόλων, αφαιρούμε τα αποτελέσματα από ένα υποσύνολο και το χρησιμοποιούμε ως σύνολο δεδομένων δοκιμών. Εκπαιδεύουμε το μοντέλο μας με τα υπόλοιπα δεδομένα. Αυτή η διαδικασία επαναλαμβάνεται στη συνέχεια για κάθε υποσύνολο ώστε να προκύψει μια γενική βαθμολογία αποδοτικότητας του μοντέλου (Arlot & Celisse, 2010).

- Προσαρμογή παραμέτρων: Εάν η ακρίβεια του μοντέλου θεωρηθεί πολύ χαμηλή, προσαρμόζουμε τις παραμέτρους των αλγορίθμων και χρησιμοποιούμε αυτές τις νέες παραμέτρους για την εκπαίδευση ενός νέου μοντέλου. Αυτό μπορεί να γίνει π.χ. με την αλλαγή του μεγέθους του δέντρου εάν επιλέξαμε να εργαστούμε με έναν αλγόριθμο δέντρων αποφάσεων. Αν επιλέξουμε να ακολουθήσουμε μια πιο αυτοματοποιημένη προσέγγιση θα χρησιμοποιήσουμε αναζήτηση πλέγματος κατά την κατασκευή του μοντέλου, θα φορτώσουμε την αναζήτηση πλέγματος με ένα σύνολο παραμέτρων σε διαφορετικά εύρη και ο έλεγχος αναζήτησης πλέγματος θα ελέγξει ποιος συνδυασμός παραμέτρων παράγει το πιο ακριβές αποτέλεσμα.

- Εφαρμογή του μοντέλου: Αφού έχουμε ένα ακριβές μοντέλο, μπορούμε τώρα να το χρησιμοποιήσουμε για να προβλέψουμε τα αποτελέσματα των νέων δεδομένων.

5.6. Ο αλγόριθμος XGBoost

Ο αλγόριθμος που πρόκειται να χρησιμοποιηθεί, πρέπει να είναι γρήγορος και να μην καταναλώνει πολλή μνήμη, αφού έχουμε πολύ μεγάλο όγκο δεδομένων για επεξεργασία. Πρέπει επίσης να λάβουμε υπόψη το γεγονός ότι θέλουμε πολυταξινόμηση που σημαίνει ότι θέλουμε να μάθουμε ουσιαστικά πόσο πιθανό είναι ο χρήστης να θέλει να ταξιδέψει σε όλους τους προορισμούς και όχι μόνο ποιος προορισμός έχει την υψηλότερη πιθανότητα. Ένας καλός αλγόριθμος γι' αυτό είναι τα δέντρα αποφάσεων με ενισχυμένη κλίση, ο οποίος χρησιμοποιεί αρκετά μικρότερα δέντρα αποφάσεων με βαθμολογημένα αποτελέσματα, όπου κάθε αποτέλεσμα παίρνει διαφορετική βαθμολογία όταν διατρέχει το δέντρο. Ο συγκεκριμένος αλγόριθμος που χρησιμοποιήθηκε στις περισσότερες λύσεις είναι ο XGBoost. Αυτός ο αλγόριθμος χρησιμοποιείται και στην υλοποίηση που θα παρουσιάσουμε στη συνέχεια. Η σημασία του ονόματος είναι η «εξαιρετικά ενισχυμένη κλίση» (extreme gradient boost), και προτιμάται σε σχέση με άλλες υλοποιήσεις ενισχυμένων δέντρων λόγω της ανωτερότητάς του στην ταχύτητα.

Ο αλγόριθμος αυτός δημιουργήθηκε αρχικά από τον Tianqi Chen από το Πανεπιστήμιο της Ουάσιγκτον και είναι μια εφαρμογή των ενισχυμένων με κλίση δέντρων αποφάσεων. Ένας τυποποιημένος αλγόριθμος δέντρων αποφάσεων, δημιουργεί ένα δέντρο βασισμένο σε όλα τα χαρακτηριστικά, για να προσπαθήσει να προβλέψει το αποτέλεσμα. Αυτό το επιτυγχάνει μέσα από το συνεχή διαχωρισμό των δεδομένων σε δύο ή περισσότερες ομάδες και οδηγώντας αυτές τις ομάδες σε περισσότερους διαχωρισμούς σε κάθε κλάδο. Έτσι για παράδειγμα, θα μπορούσε να υπάρχει διαχωρισμός σε αρσενικό, θηλυκό ή άγνωστο ως ο πρώτος κόμβος στο δέντρο και κάθε απόφαση θα διακλαδωθεί σε νέους κόμβους απόφασης ή σε έναν κόμβο φύλλων ο οποίος είναι τελικά ο προορισμός των χρηστών. Το φύλλο στο οποίο καταλήγουν δίνει την πληροφορία στον ταξινομητή, ποιος προορισμός είναι πιο πιθανό να επιλεγεί. Το μειονέκτημα των κανονικών δέντρων αποφάσεων είναι το μέγεθος που απαιτείται για την πλήρη εξήγηση των δεδομένων. Για παράδειγμα κάθε φύλο χρειάζεται τον δικό του κόμβο απόφασης που περιγράφει την ηλικία του για να είναι πλήρως σε θέση να δημιουργήσει το προφίλ του χρήστη.

Ένα ενισχυμένο δέντρο χρησιμοποιεί την ίδια λογική ενός δέντρου με κόμβους απόφασης και τελικούς κόμβους φύλλων, αλλά η διαφορά είναι το μέγεθος. Ένα ενισχυμένο δέντρο χρησιμοποιεί μικρότερα δέντρα που εξηγούν μόνο ένα κομμάτι των

χαρακτηριστικών σε κάθε δέντρο. Στη συνέχεια δημιουργούνται αρκετά από αυτά τα δέντρα και κάθε κόμβος φύλλου περιέχει μια βαθμολογία. Οι βαθμολογίες από όλα τα υποδέντρα αθροίζονται για να επιτευχθεί μια τελική βαθμολογία. Το μεταβαλλόμενο μέρος του αλγορίθμου προέρχεται από τη δημιουργία κάθε νέου δέντρου. Ο αλγόριθμος δημιουργεί ένα νέο δέντρο που αναλύει τα προηγούμενα δένδρα και δημιουργεί στη συνέχεια ένα νέο δέντρο που επιχειρεί να διορθώσει τα σφάλματα στα προηγούμενα.

Τα σύνολα δεδομένων γίνονται όλο και μεγαλύτερα και δημιουργούν πρόβλημα για το υπάρχον hardware, όπου για παράδειγμα η μνήμη είναι πολύ περιορισμένη. Απαιτείται λοιπόν καλός χειρισμός της μνήμης από τους αλγόριθμους μηχανικής μάθησης ώστε να μην εμφανίζονται πάρα πολλά από τα κοινά προβλήματα εξάντλησης της μνήμης «out-of-memory». Ο XGBoost επικεντρώνεται στην υπολογιστική ταχύτητα και την απόδοση του μοντέλου. Ο αλγόριθμος σχεδιάστηκε με στόχο την αποδοτικότερη χρήση των πόρων μνήμης, αλλά και την ταχύτητα. Ένας από τους στόχους κατά το σχεδιασμό ήταν να γίνει η καλύτερη χρήση των διαθέσιμων πόρων κατά την εκπαίδευση ενός μοντέλου. Ορισμένες βασικές λειτουργίες εφαρμογής του αλγορίθμου περιλαμβάνουν:

- Αυτόματο χειρισμό δεδομένων που λείπουν
- Δομή block για να υποστηρίζει τον παραλληλισμό της δομής των δένδρων
- Συνεχής εκπαίδευση με νέα δεδομένα ώστε να είναι εφικτή η περαιτέρω ενίσχυση σε ένα ήδη προσαρμοσμένο μοντέλο

Η υλοποίηση του μοντέλου υποστηρίζει επίσης τις δυνατότητες της βιβλιοθήκης scikit-learn. Ο XGBoost κυριαρχεί γενικά σε δομημένα ή πινακοειδή σύνολα δεδομένων για προβλήματα πρότυπης μοντελοποίησης ταξινόμησης και παλινδρόμησης. Έχει επιλεγεί από πολλούς νικητές διαγωνισμών στο Kaggle. (Praxitelis - Nikolaos Kouroupetroglou, 2017)

5.7. Συμπεράσματα

Αναλύοντας τις διαθέσιμες υλοποιήσεις, παρατηρήσαμε ότι κάποιες από αυτές επικεντρώνονται ιδιαίτερα στην ανάλυση και επεξεργασία των δεδομένων χωρίς να παραθέτουν όμως στη συνέχεια τον αλγόριθμο που έχουν χρησιμοποιήσει. Άλλες πάλι παρουσιάζουν τη χρήση του ταξινομητή και δημιουργία του μοντέλου, υστερούν όμως σε πρότερη ανάλυση των δεδομένων, και έτσι δεν είναι εμφανές από που προκύπτουν οι παράμετροι που χρησιμοποιούν. Λαμβάνοντας υπόψη το γεγονός αυτό, σε συνδυασμό με όλα τα παραπάνω κριτήρια επιλογής που παρουσιάσαμε σε αυτό το κεφάλαιο, καταλήξαμε στην επιλογή δύο υλοποιήσεων. Η μια ανήκει στον διαγωνιζόμενο με το όνομα David Gasquez, ο οποίος περιγράφει ένα τμήμα της λύσης του, αυτό της ανάλυσης των δεδομένων, το οποίο είναι το πιο δημοφιλής Kernel του συγκεκριμένου διαγωνισμού, με 120 ψήφους. Η δεύτερη υλοποίηση είναι του διαγωνιζόμενου με το όνομα Sandro. Σε αυτή την υλοποίηση έχουμε το τμήμα του κώδικα που αφορά τη χρήση του ταξινομητή και τη δημιουργία του τελικού αρχείου με τις προβλέψεις, όπως ζητούσε ο διαγωνισμός. Η ολοκληρωμένη λύση του συγκεκριμένου διαγωνιζόμενου, κατατάχθηκε βάση της τελική βαθμολογίας στις 3 καλύτερες του διαγωνισμού. Στη λύση που θα παρουσιάσουμε σε επόμενο κεφάλαιο, θα συνδυάσουμε αυτές τις δύο υλοποιήσεις.
(https://www.kaggle.com/davidgasquez/user-data-exploration)
(https://www.kaggle.com/svpons/script-0-8655/data)

6. Ανάλυση δεδομένων και εξομοίωση των ευρημάτων

6.1. Εισαγωγή

Η διερευνητική ανάλυση δεδομένων είναι μια προσέγγιση που αναλύει σύνολα δεδομένων για να συνοψίσει τα κύρια χαρακτηριστικά τους, συχνά με μεθόδους οπτικοποίησης. Μπορεί να γίνει με τη χρήση κάποιου στατιστικού μοντέλου, δεν είναι όμως απαραίτητο. Πρωτίστως η διερευνητική ανάλυση δεδομένων είναι για να δούμε τι πληροφορίες μπορούμε να λάβουμε από τα δεδομένα μας πέρα από την επίσημη μοντελοποίηση ή για δοκιμές υποθέσεων.

Αρχικά, τα δεδομένα της μελέτης μας θα οργανωθούν μέσω κανονικοποίησης των μεταβλητών ώστε να μπορέσουμε στη συνέχεια να εφαρμόσουμε στατιστικές τεχνικές από της οποίες θα λάβουμε τη μέση τιμή, τη διακύμανση και άλλα χρήσιμα στατιστικά στοιχεία που καθορίζουν το σύνολο προς μελέτη. Θα ακολουθήσει μια οπτικοποίηση των υφιστάμενων σχέσεων. Τέλος, θα μελετήσουμε τα "ειδικά" μοτίβα, όπως τα υπερμεγέθη, προσπαθώντας να εξηγήσουμε την ύπαρξή τους καθώς και να αποφασίσουμε αν είναι χρήσιμα για το μοντέλο μας.

Συνοπτικά, μπορούμε να απαριθμήσουμε τους σημαντικότερους στόχους της ανάλυσης των δεδομένων μας ως εξής:

- Κατανόηση των δεδομένων
- Διερεύνηση των δομών τους
- Στόχευση μεταβλητών με τη μεγαλύτερη σημασία
- Εντοπισμός ανωμαλιών
- Έλεγχος υποθέσεων
- Προσδιορισμός πιθανών αλλαγών και πώς θα επηρεάσουν το πρόβλημα

6.2. Αρχικά δεδομένα

Πριν εξετάσουμε λεπτομερώς τα δεδομένα, θα παρουσιάσουμε τα σύνολα των δεδομένων όπως δόθηκαν στο κοινό από την Airbnb για το διαγωνισμό στην πλατφόρμας της Kaggle. Τα δεδομένα αυτά αποτελούν τη βάση για την εκπαίδευση και αξιολόγηση του μοντέλου πρόβλεψης.

Υπάρχουν διαθέσιμα 5 σύνολα δεδομένων σε μορφή csv:

1. train_users: Αυτό το σύνολο δεδομένων περιέχει 213.451 δεδομένα χρηστών από την Airbnb. Όλα τα δεδομένα είναι ανώνυμα, πράγμα που σημαίνει ότι όλα τα δεδομένα που χρησιμοποιούνται για την αναγνώριση αντικαθίστανται από ένα πεδίο ταυτότητας (ID). Τα διάφορα στοιχεία των δεδομένων σε αυτό το πεδίο, περιέχουν δεδομένα που συλλέχθηκαν από την πρώτη καταγεγραμμένη επίσκεψη του χρήστη και τα δεδομένα που εισήχθησαν κατά τη διαδικασία της εγγραφής του. Για το σετ αυτό που θα χρησιμοποιηθεί για την εκπαίδευση, έχουμε επίσης την πληροφορία για τη χώρα στην οποία ο χρήστης επίλεξε τελικά να ταξιδέψει.

2. test_users: Το τεστ σετ των χρηστών, περιέχει 62.096 χρήστες για τους οποίους θα πρέπει να γίνει η πρόβλεψη. Έχει την ίδια μορφή με το train_users, με εξαίρεση τα πεδία date_first_booking και country_destination, τα οποία φυσικά δεν υπάρχουν.

3. sessions: Περιέχει εγγραφές των ενεργειών του χρήστη στον ιστότοπο. Κάθε χρήστης αναγνωρίζεται με ένα πεδίο user_id που αντιστοιχεί στο πεδίο id των συνόλων δεδομένων χρήστη. Κάθε εγγραφή σε αυτό το σύνολο δεδομένων περιέχει δεδομένα σχετικά με την ενέργεια που εκτελέστηκε, τον τύπο της συσκευής από την οποία εκτελέστηκε καθώς και το χρόνο σε δευτερόλεπτα από την τελευταία ενέργεια που αναφέρθηκε. Περιέχει μόνο δεδομένα για χρήστες από το 2014.

4. countries: Σύνοψη των χωρών που είναι πιθανοί προορισμοί, καθώς και δεδομένα για την τοποθεσία τους και την κύρια γλώσσα τους

5. age_gender_bkts: Αυτό το σύνολο δεδομένων ομαδοποιεί τους χρήστες στο σετ εκπαίδευσης, σε ηλικιακές ομάδες με διαφορά 5 ετών και παρουσιάζει πληροφορίες σχετικά με το φύλο και τον προορισμό χώρας που αποφασίζει κάθε ομάδα.

Το πρώτο πρόβλημα λοιπόν με το οποίο βρισκόμαστε αντιμέτωποι, είναι η ποικιλία των πηγών από τις οποίες αντλούμε τις πληροφορίες. Για να μπορέσει να εφαρμοστεί μια τεχνική μηχανικής μάθησης, θα πρέπει να οργανώσουμε τα δεδομένα αυτά με τέτοιο τρόπο έτσι ώστε κάθε σύνολο πληροφοριών να αναπαρίσταται ως μία γραμμή σε έναν πίνακα. Η Airbnb προσφέρει αυτά τα δεδομένα στους συμμετέχοντες του διαγωνισμού με αυτό τον τρόπο, ώστε ο καθένας να αποφασίσει μόνος του πώς θα χρησιμοποιήσει το κάθε αρχείο.

Τα πιο σημαντικά αρχεία είναι αυτά που μας δείχνουν δεδομένα των χρηστών και των συνεδριών. Κάθε χρήστης έχει έναν αριθμό Ν ενεργειών στο αρχείο sessions.csv και αυτό σημαίνει ότι δεν μπορούμε να τα χρησιμοποιήσουμε ως έχουν, αλλά χρειάζονται επιπλέον επεξεργασία ώστε να μπορέσουμε να επωφεληθούμε στο μέγιστο βαθμό από αυτές τις πληροφορίες.

Έχουμε λοιπόν μια σχέση ένα προς πολλά μεταξύ του αρχείου χρηστών και εκείνο των sessions, όπου το πρωτεύον κλειδί και στα δύο είναι το ID των χρηστών. Το αρχείο sessions έχει μέγεθος 55MB και περιέχει μέχρι 5000 συμβάντα ανά χρήστη, άρα το ενδεχόμενο να έχουμε όλα αυτά τα δεδομένα συγκεντρωτικά σε μια στήλη, απορρίπτεται. Θα δούμε σε επόμενη ενότητα πως θα προσεγγίσουμε αυτό το πρόβλημα, αφού πρώτα κάνουμε μια περιγραφή των μεταβλητών.

6.3. Περιγραφή μεταβλητών

Στα αρχεία χρηστών (train_users και test_users) έχουμε τις εξής μεταβλητές:

id: Το ID του χρήστη. Είναι μοναδικό για κάθε χρήστη
date_account_created: Η ημερομηνία δημιουργίας του λογαριασμού στην Airbnb
timestamp_first_active: Χρονική σφραγίδα της πρώτης δραστηριότητας του χρήστη. Αυτή μπορεί να είναι χρονικά πριν τη δημιουργία του λογαριασμού, καθώς οι χρήστες μπορούν να κάνουν αναζητήσεις χωρίς να είναι εγγεγραμμένοι.
date_first_booking: Η ημερομηνία κατά την οποία ο χρήστης έκανε την πρώτη του κράτηση.
gender: Το φύλο του χρήστη
age: Η ηλικία του χρήστη
signup_method: Η μέθοδος που χρησιμοποιήθηκε για την εγγραφή
signup_flow: Η σελίδα από την οποία έγινε η εγγραφή
language: Η επιλεγμένη γλώσσα του χρήστη
affiliate_channel: Συσχετισμένος τύπος μάρκετινγκ
affiliate_provider: Συσχετισμένη εταιρεία μάρκετινγκ
first_affiliate_tracked: Από ποια εταιρεία πρωτοήρθε ο χρήστης
signup_app: Εφαρμογή μέσω της οποίας ήταν συνδεδεμένος
first_device_type: Συσκευή από την οποία ήταν συνδεδεμένος
first_browser: Το πρόγραμμα περιήγησης που χρησιμοποιήθηκε για την πρόσβαση διαδικτυακά
country_destination: Χώρα προορισμού του χρήστη

Στο αρχείο sessions, έχουμε τις ακόλουθες μεταβλητές:

user_id: Αναγνωριστικό χρήστη. Υπάρχουν επαναλαμβανόμενα αναγνωριστικά μέσα σε αυτό το αρχείο δεδομένου ότι υπάρχουν πολλά γεγονότα για ένα χρήστη
action: Ενέργεια
action_type: Τύπος ενέργειας
action_detail: Λεπτομέρειες της ενέργειας.
device_type: Τύπος συσκευής
secs_elapsed: Δευτερόλεπτα που έχουν περάσει. Η Airnbnb δεν διευκρινίζει

περισσότερα για αυτή τη μεταβλητή, επομένως θα πρέπει να κάνουμε και να επαληθεύσουμε κάποιες υποθέσεις.

Στη συνέχεια θα συνεχήσουμε με την διερευνητική ανάλυση ώστε να αναλυθούν και να κατανοηθούν λεπτομέρειες των δεδομένων μας.

6.4. Διερευνητική ανάλυση μεταβλητών

Πρώτα από όλα, θα πρέπει να προβούμε σε μια γενική σάρωση των δεδομένων:

- να παρατηρήσουμε τα δεδομένα των χρηστών
- να ελέγξουμε αν υπάρχουν σφάλματα στα δεδομένα
- να δούμε αν συμπεριφέρονται τα δεδομένα μας με κάποιο συγκεκριμένο τρόπο
- να αναγνωρίσουμε αν πρέπει να διορθώσουμε ή να διαγράψουμε κάποια από τα δεδομένα για να τα κάνουμε πιο ρεαλιστικά.

Αρχικά να αναφέρουμε ότι έχουμε 213.451 διαφορετικούς χρήστες εκπαίδευσης και 62.096 χρήστες για τα τεστ . Έχουμε λοιπόν συνολικά 275.547 χρήστες, ένας ικανός αριθμός ώστε να βρούμε διαφορετικές συμπεριφορές και διαφορετικούς τύπους χρηστών. Θα ενώσουμε τα δεδομένα αυτά, ώστε να μπορέσουμε να εργαστούμε με όλα τα δεδομένα για την ανάλυση των μεταβλητών.

	affiliate_channel	affiliate_provider	age	country_destination	date_account_created	date_first_booking	first_affiliate_tracked
0	direct	direct	NaN	NDF	2010-06-28	NaN	untracked
1	seo	google	38.0	NDF	2011-05-25	NaN	untracked
2	direct	direct	56.0	US	2010-09-28	2010-08-02	untracked
3	direct	direct	42.0	other	2011-12-05	2012-09-08	untracked
4	direct	direct	41.0	US	2010-09-14	2010-02-18	untracked

first_browser	first_device_type	gender	language	signup_app	signup_flow	signup_method	timestamp_first_active
Chrome	Mac Desktop	-unknown-	en	Web	0	facebook	20090319043255
Chrome	Mac Desktop	MALE	en	Web	0	facebook	20090523174809
IE	Windows Desktop	FEMALE	en	Web	3	basic	20090609231247
Firefox	Mac Desktop	FEMALE	en	Web	0	facebook	20091031060129
Chrome	Mac Desktop	-unknown-	en	Web	0	basic	20091208061105

Εικόνα 6.1 Τμήμα από το συγκεντρωτικό πίνακας χρηστών

Τα δεδομένα φαίνεται να είναι σε μορφή που μπορούν να χρησιμοποιηθούν, επομένως το επόμενο σημαντικό βήμα είναι να εντοπίσουμε τα δεδομένα που λείπουν. Συνήθως τα δεδομένα που λείπουν σημειώνονται ως NaN, αλλά αν κοιτάξουμε την εικόνα 6.1, μπορούμε να δούμε στη στήλη του φύλου κάποιες τιμές να είναι -unknown-. Θα πρέπει πρώτα να μετατρέψουμε αυτές τις τιμές σε NaN. Μια επισκόπηση του ποσοστού των δεδομένων που λείπουν για κάθε στήλη, παρουσιάζεται στον πίνακα 6.1.

Πίνακας 6.1 Ποσοστό NaN για κάθε χαρακτηριστικό

date_first_booking	67,8%
first_affiliate_tracked	2,2%
first_browser	16,1%
gender	46,5%
age	42,4%

Για το χαρακτηριστικό date_first_booking, σχεδόν το 68% των τιμών είναι NaN, καθώς αυτή η τιμή δεν υπάρχει στους χρήστες του συνόλου τεστ, και έτσι δεν θα το χρειαστούμε στο κομμάτι της μοντελοποίησης. Αντίθετα, τα υψηλά ποσοστά NaN θα οδηγήσουν σε χαμηλότερες επιδόσεις των ταξινομητών που θα δημιουργήσουμε.

Ας αναλύσουμε λοιπόν τις μεταβλητές του φύλου και της ηλικίας πιο αναλυτικά. Για το φύλο μόνο δύο τιμές θα ήταν αναμενόμενες, αρσενικό και θηλυκό. Οι αριθμοί εμφάνισης τις κάθε τιμής παρουσιάζονται στον πίνακα 6.2. Το 46% των χρηστών δεν έχει καταχωρήσει φύλο και το 0,1% έχει επιλέξει ως τιμή άλλο φύλο.

Πίνακας 6.2 Αριθμός επαναλήψεων των τιμών για τη μεταβλητή φύλο

-unknown-	129.480
FEMALE	77.524
MALE	68.209
OTHER	334

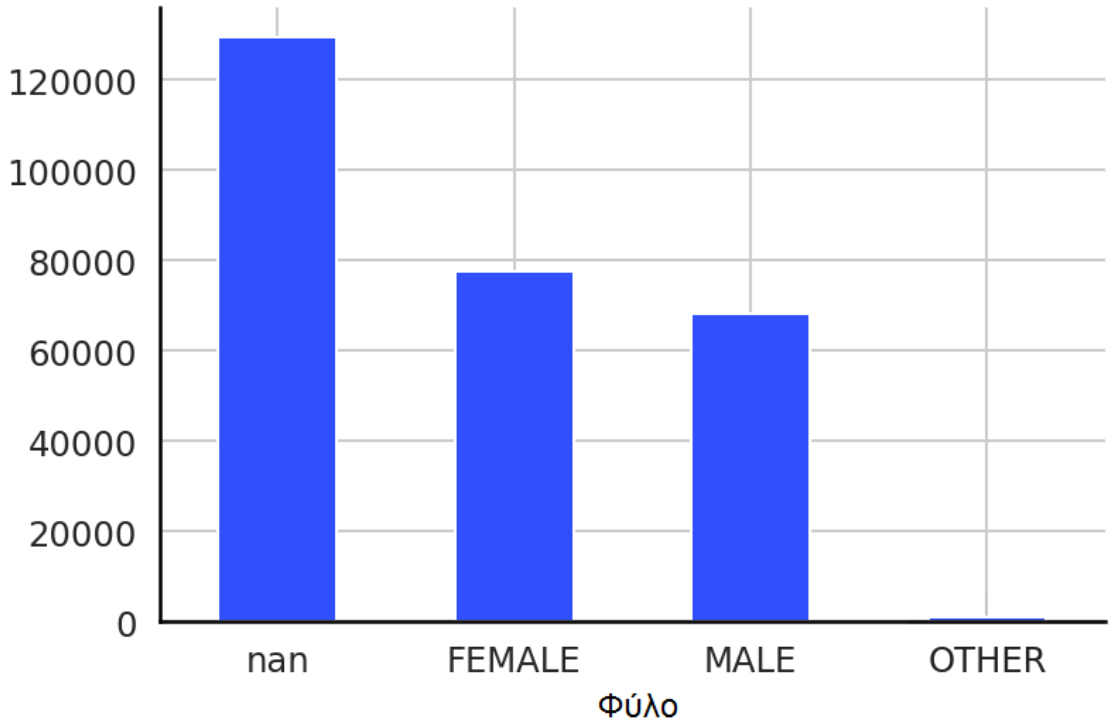
Για τις ηλικίες, θα αναμέναμε τιμές στο διάστημα [1, 100], παρατηρώντας όμως τον πίνακα 6.3, μπορούμε να δούμε ότι ο μέσος όρος είναι τα 47 χρόνια. Είναι προφανές ότι αντιμετωπίζουμε κάποιο πρόβλημα βλέποντας την τυπική απόκλιση. Συμπεραίνουμε πως δεν είναι όλα τα δεδομένα μας σωστά, και αυτό γίνεται ξεκάθαρο από τη μέγιστη τιμή που είναι το 2014. Μπορούμε να υποθέσουμε ότι αναφέρεται σε ένα συγκεκριμένο έτος, δηλαδή σε μια ημερομηνία και όχι στην πραγματική ηλικία του χρήστη. Υπάρχει λοιπόν μια ασυνέπεια στην ηλικία ορισμένων χρηστών. Θα μπορούσε να οφείλεται στο γεγονός ότι το πεδίο εισαγωγής της ηλικίας δεν καθαρίστηκε ή δεν περιορίστηκε σωστά ή υπήρξαν κάποια λάθη στην επεξεργασία των δεδομένων.

Πίνακας 6.3 Στατιστική σύνοψη της μεταβλητής ηλικίας

mean	47,145310
std	142,629468
min	1,000000
25%	28,000000
50%	33,000000
75%	42,000000
max	2.014,000000

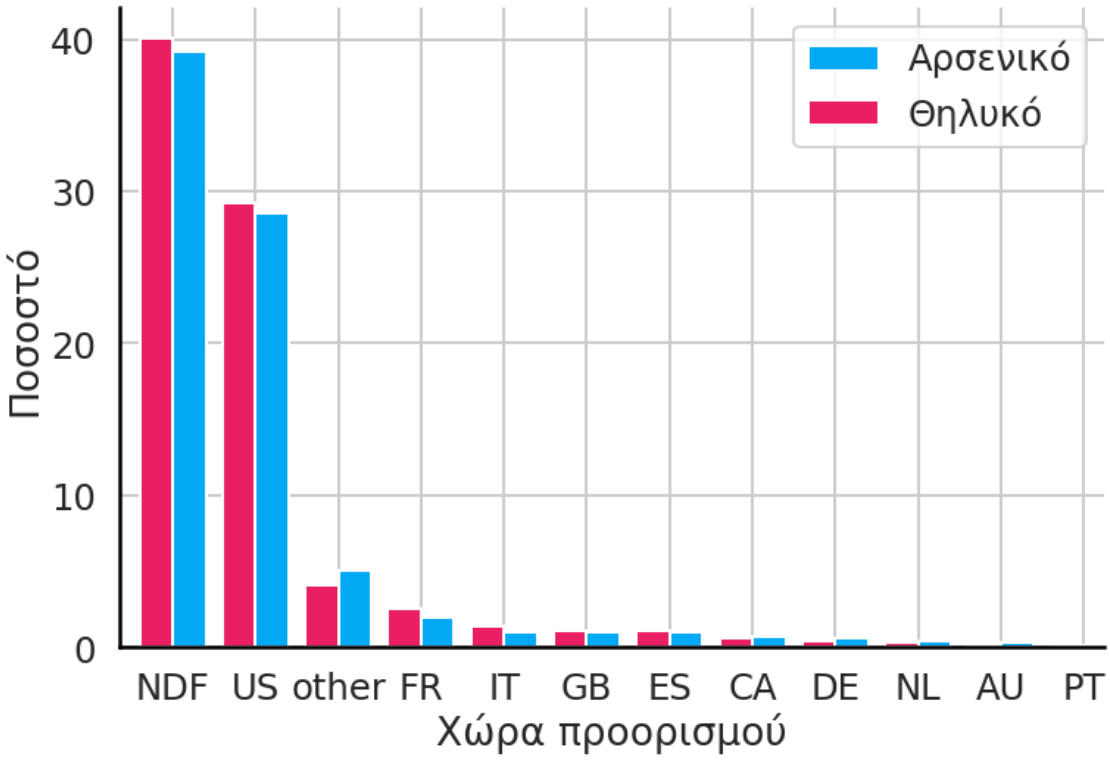
Υπάρχουν 830 χρήστες ηλικίας άνω των 100 ετών και 188 κάτω των 18. Για τιμές που είναι μεταξύ 1940 και 2014 μπορούμε να υπολογίσουμε την ηλικία κάνοντας μια απλή αφαίρεση. Στα υπόλοιπα δείγματα, απορρίπτουμε τις τιμές. Ειδικότερα τιμές μεγαλύτερες από 100 και μικρότερες από 18, θα μετατραπούν σε NaN.

Θα συνεχίσουμε με μια σειρά γραφικών αναπαραστάσεων για να προσπαθήσουμε να κατανοήσουμε καλύτερα τα δεδομένα και να δούμε αν υπάρχουν κάποια μοτίβα σε αυτά. Η απεικόνιση των μεταβλητών καθιστά ευκολότερο τον εντοπισμό των αποκλίσεων και των σφαλμάτων, μελετώντας τα γραφήματα.



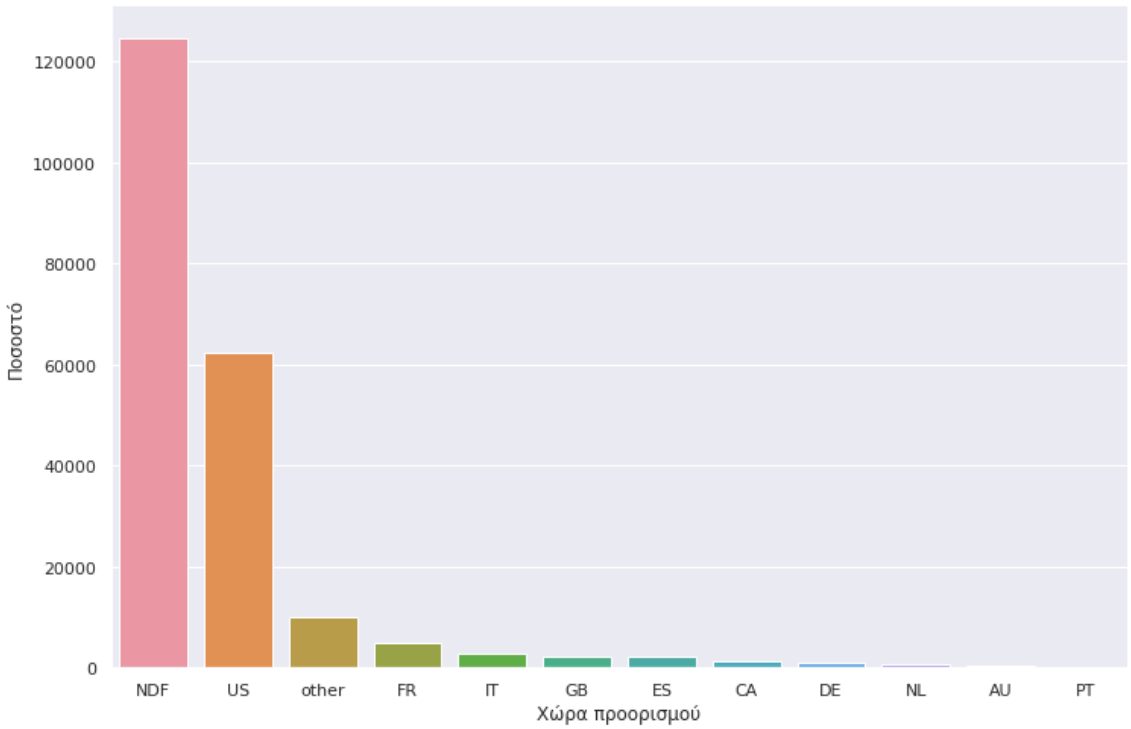
Διάγραμμα 6.1 Ιστόγραμμα τιμών για τη μεταβλητή φύλο

Όπως παρατηρήσαμε και προηγουμένως, έχουμε μεγάλο αριθμό χρηστών, οι οποίοι δεν έχουν ορίσει το φύλο τους. Το διάγραμμα 6.1. μας βοηθάει για να οπτικοποιήσουμε αυτό τον αριθμό, αλλά μας δείχνει επίσης ότι υπάρχει μια μικρή διαφορά στο φύλο των χρηστών. Να σημειώσουμε εδώ ότι αυτό μπορεί να μην είναι η πραγματική κατανομή, καθώς μπορεί να υπάρχουν περισσότεροι άντρες που έχουν εγγραφεί και απλώς δεν καταχώρησαν την πληροφορία αυτή στο σύστημα.



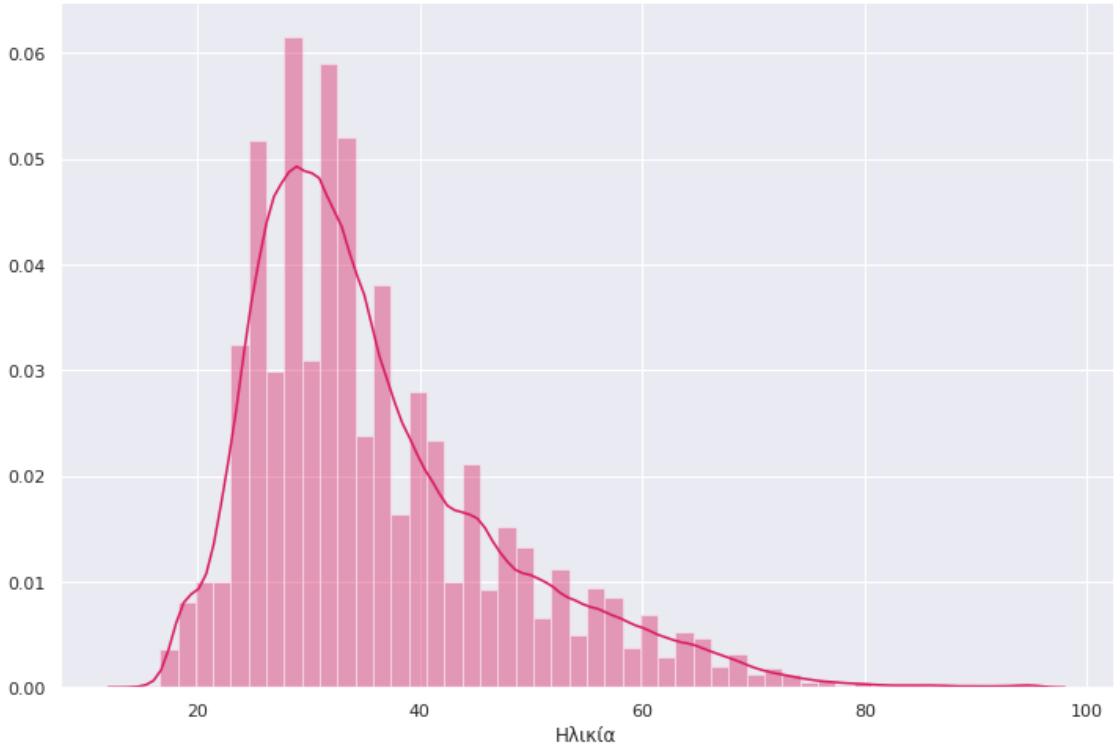
Διάγραμμα 6.2 Ιστόγραμμα προτίμησης χώρας προορισμού μεταξύ των δύο φύλων

Όπως βλέπουμε στο διάγραμμα 6.2, δεν υπάρχουν μεγάλες διαφορές στις προτιμήσεις της χώρας ανάλογα με το φύλο. Ενδιαφέρον παρουσιάζει βέβαια η πληροφορία της σχετικής συχνότητας των χωρών προορισμού, η οποία φαίνεται πιο ξεκάθαρα στο διάγραμμα 6.3. Διακρίνοντας τους αριθμούς των χρηστών ανά χώρα, μπορούμε εύκολα να καταλήξουμε στο συμπέρασμα ότι αν ένας χρήστης προβεί σε κράτηση μέσω Airbnb, το πιθανότερο είναι να ταξιδέψει εντός ΗΠΑ. Θα πρέπει να σημειώσουμε βέβαια ότι το 45% των χρηστών δεν έκαναν καμία κράτηση. Αυτές οι δύο επιλογές δημιουργούν μια ανισορροπία στο πρόβλημα.



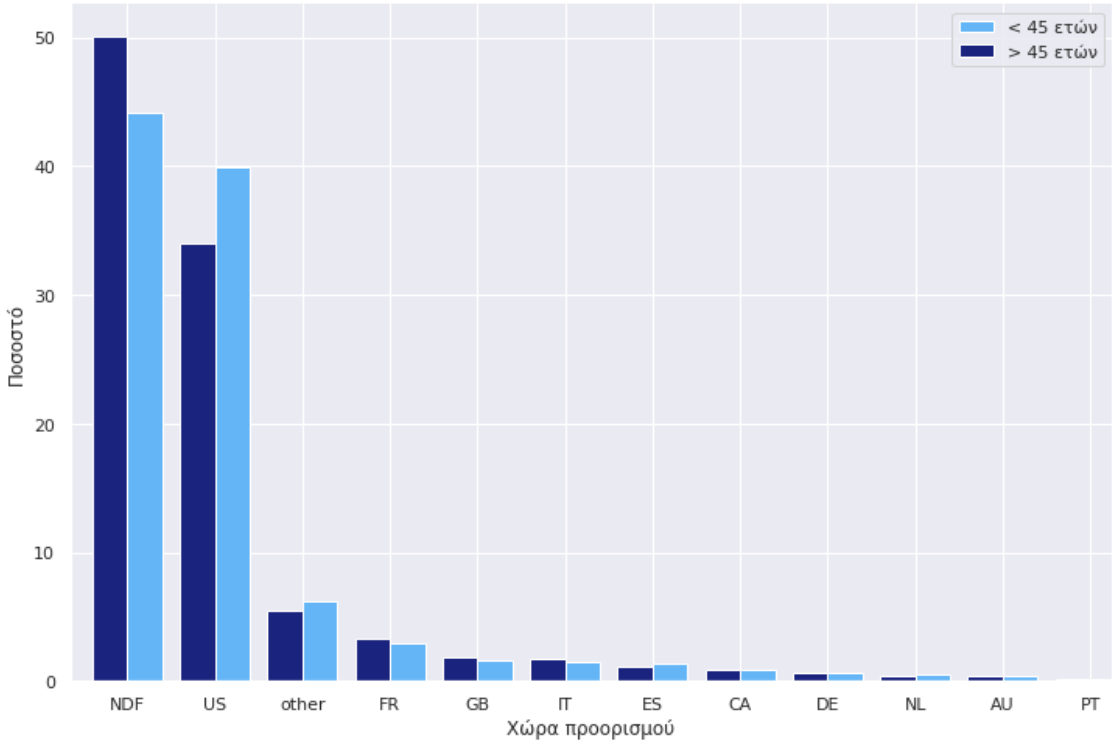
Διάγραμμα 6.3 Σχετικής συχνότητας των χωρών προορισμού

Αφού πήραμε μια ιδέα για το πώς κατανέμονται οι προτιμήσεις των δύο φύλων και τον αριθμό των χρηστών που κάνουν κράτηση σε κάθε χώρα, θα διερευνήσουμε τη μεταβλητή ηλικία. Αρχικά, μπορούμε να δούμε μια κατανομή των ηλικιών των χρηστών στο διάγραμμα 6.4.



Διάγραμμα 6.4 Κατανομή της ηλικίας των χρηστών.

Όπως αναμενόταν, η συνήθης ηλικία των ταξιδιωτών είναι μεταξύ 25 και 40 ετών. Θα κάνουμε έναν έλεγχο να δούμε αν οι ηλικιωμένοι ταξιδεύουν με διαφορετικό τρόπο από τους νεότερους. Ας επιλέξουμε αυθαίρετα την ηλικία των 45 ώστε να χωρίσουμε τους χρήστες σε δύο ομάδες.

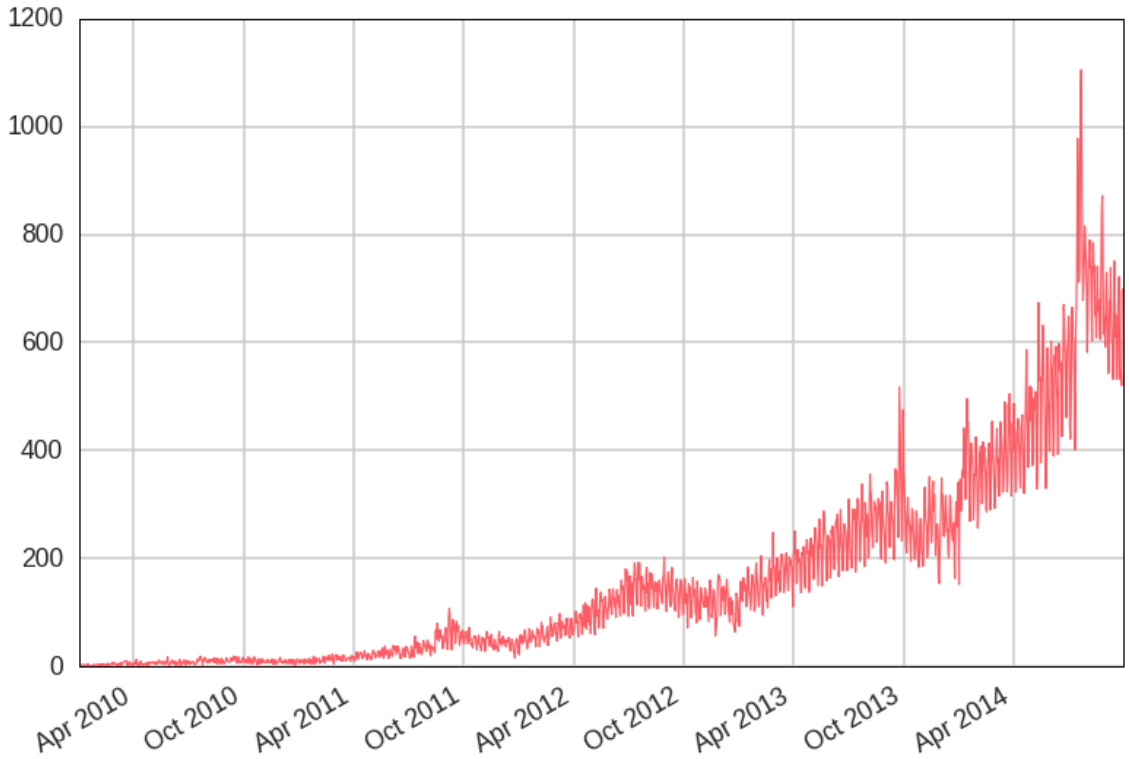


Διάγραμμα 6.5 Προτίμηση χώρας αναλογικά με την ηλικία

Μελετώντας αυτό το διάγραμμα, μπορούμε να παρατηρήσουμε με μεγαλύτερη σαφήνεια τη συμπεριφορά των χρηστών ανάλογα με την ηλικία τους. Μπορούμε να δούμε πώς οι νέοι που βρίσκονται σε ηλικία κάτω των 45 χρόνων, τείνουν να παραμένουν στις Ηνωμένες Πολιτείες, και οι ηλικιωμένοι τείνουν να επισκέπτονται συχνότερα άλλες χώρες. Πρέπει να σημειωθεί όμως ότι έχουμε μόνο το 58% των ηλικιών των χρηστών.

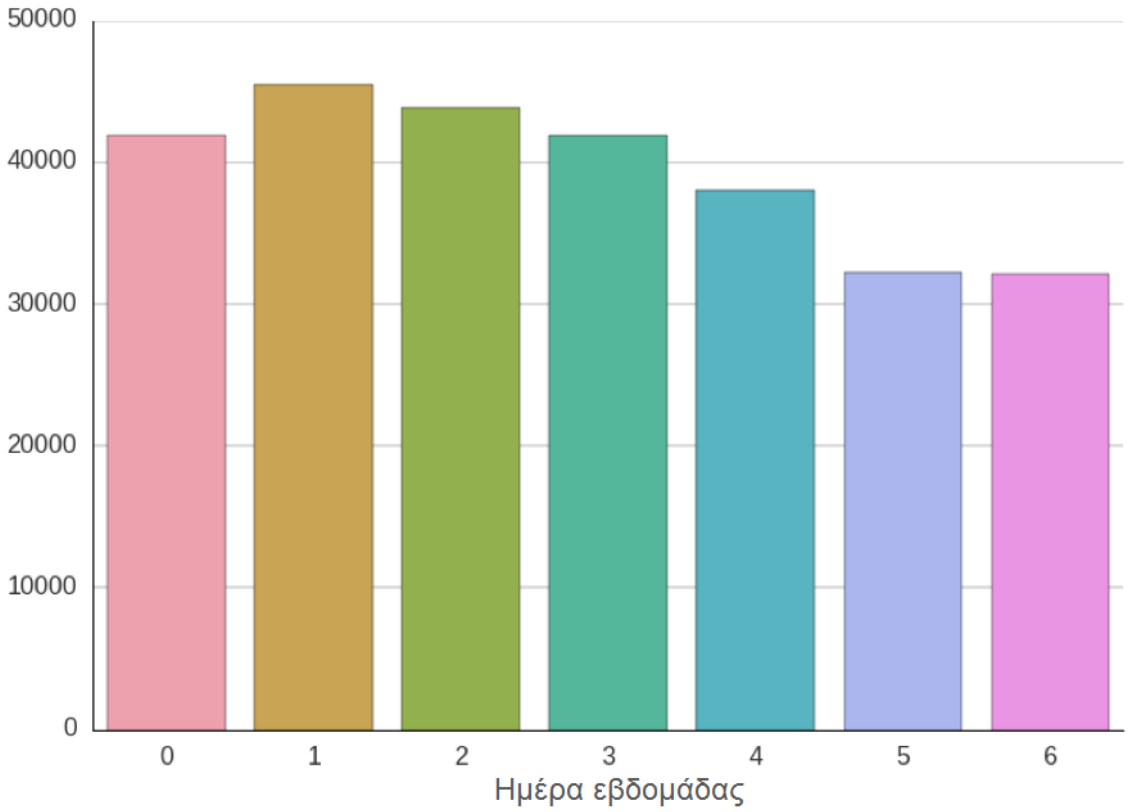
Μια ακόμη σημαντική παράμετρος στην επιλογή χώρας προορισμού, είναι η μητρική γλώσσα του χρήστη. Από τα δεδομένα μας βλέπουμε ότι το 96% των χρηστών έχει την αγγλική γλώσσα ως μητρική. Το τόσο υψηλό ποσοστό ίσως είναι μια εξήγηση γιατί πολλοί χρήστες αποφασίζουν να μείνουν στις ΗΠΑ. Ο αντίλογος βέβαια σε αυτό είναι ότι αν η γλώσσα είναι τόσο σημαντικό κριτήριο, γιατί να μην ταξιδέψουν στο Ηνωμένο Βασίλειο. Πρέπει να έχουμε κατά νου ότι υπάρχουν πολλοί παράγοντες που δεν συνυπολογίζουμε εδώ, οπότε το να κάνουμε τέτοιες υποθέσεις ή προβλέψεις δεν είναι σκόπιμο.

Για να αναλύσουμε τις ημερομηνίες και τις χρονοσφραγίδες των χρηστών θα πρέπει να παρατηρήσουμε στο διάγραμμα 6.6., τον αριθμό των λογαριασμών που δημιουργήθηκαν ανά χρονική περίοδο.



Διάγραμμα 6.6 Αριθμός νέων χρηστών

Είναι αξιοσημείωτη η ανάπτυξη της Airbnb σε αυτό το χρονικό διάστημα. Όπως μπορούμε να δούμε, ο αριθμός των χρηστών δεν έχει σταματήσει να αυξάνεται και μάλιστα με σημαντικό ρυθμό. Παρόλο που είναι μια σχεδόν εκθετική καμπύλη, βλέπουμε πως είναι πολύ απότομη και έχει αρκετές κορυφές σε πολλές από τις ημερομηνίες. Το μέγεθος αυτών των κορυφών μεγαλώνει όσο μεγαλώνει ο αριθμός των χρηστών. Κάνοντας διάφορες υποθέσεις, αποφασίσαμε να κάνουμε μια ανάλυση των ημερών της εβδομάδας για να δούμε πόσοι χρήστες κάνουν εγγραφή σε κάθε μία από αυτές. Φυσικά γίνονται καθημερινά πολλές εγγραφές όμως υπάρχουν διαφορές στην κατανομή, όπως φαίνεται στο γράφημα 6.7. Με τον αριθμό 0 απεικονίζεται η Δευτέρα και με τον αριθμό 6 η Κυριακή. Τα τοπικά ελάχιστα λοιπόν ήταν συνήθως Κυριακές, όπου ενδεχομένως οι άνθρωπο χρησιμοποιούν λιγότερο το διαδίκτυο, και ο μεγαλύτερος αριθμός εγγραφών χρηστών παρατηρείται τις Τρίτες.



Διάγραμμα 6.7 Εγγραφές ανά ημέρα εβδομάδας

Εκτός από την ανάλυση των μεταβλητών των αρχείων users που παρουσιάσαμε μέχρι τώρα, θα προχωρήσουμε σε ανάλυση και του αρχείου sessions. Σε αυτό το αρχείο έχουμε 135.484 ID χρηστών. Αν συγκρίνουμε τον αριθμό αυτό με τους 275.547 χρήστες που διαθέτουμε συνολικά, αποδεικνύεται ότι έχουμε μόνο το 50% των συνεδριών των χρηστών. Θα διερευνήσουμε λοιπόν τις ενέργειες των χρηστών για τις οποίες υπάρχουν εγγραφές.

Ξεκινάμε από τον τύπο της ενέργειας που μπορεί να έχει μία από τα παρακάτω τιμές: nan, click, data, view, submit, message_post, -unknown-, booking_request, partner_callback, booking_response και modify.

Το πρώτο πράγμα που παρατηρούμε είναι ότι υπάρχουν δύο τρόποι ορισμού των μηδενικών τιμών, null ή unknown. Κάνοντας αντικατάσταση και κανονικοποίηση των δεδομένων θα δούμε ποιοι είναι οι συνηθέστεροι τύποι ενεργειών ανάλογα με το πόσες φορές επαναλαμβάνονται στο αρχείο sessions:

Πίνακας 6.4 Αριθμός επαναλήψεων ενεργειών στο αρχείο συνεδρίας.

view	3.560.902
data	2.103.770
click	1.996.183
submit	623.357
message_post	87.103
partner_callback	19.132
booking_request	18.773
modify	1.139
booking_response	4

Από τα ονόματα δεν μπορούμε να διακρίνουμε τη ακριβή σημασία της κάθε ενέργειας, εκτός ίσως από τα booking_request και booking_response. Αυτά βέβαια εμφανίζονται σε σχετικά χαμηλή συχνότητα και έτσι δεν αποτελούν αντικείμενο περαιτέρω ερευνών. Υπάρχουν πολλές πιθανές ενέργειες, 360 για την ακρίβεια, και η δημιουργία οποιουδήποτε γραφήματος ή πίνακα θα μπορούσε να οδηγήσει σε σύγχυση. Στον πίνακα 6.5. βλέπουμε των αριθμό τον μοναδικών τιμών κάθε παραμέτρου.

Πίνακας 6.5 Αριθμός μοναδικών τιμών κάθε παραμέτρου

action	360
action_type	11
action_detail	156
device_type	14
secs_elapsed	337.662

Σε πολλές εγγραφές του πίνακα sessions, έχουμε μηδενικές τιμές για τις υπάρχουσες παραμέτρους. Μια σύνοψη των αριθμών των μηδενικών τιμών φαίνεται στον πίνακα 6.6. Όπως βλέπουμε, όπου δεν υπάρχει τιμή για την μεταβλητή action type, δεν υπάρχει ούτε για τη μεταβλητή action_detail. Δεν ισχύει το ίδιο για την ίδια την ενέργεια και έτσι βλέπουμε ότι η μεταβλητή action έχει πολύ λιγότερες μηδενικές τιμές.

Πίνακας 6.6 Αριθμός μηδενικών τιμών κάθε παραμέτρου

action	79.626
action_type	1.126.204
action_detail	1.126.204
device_type	0
secs_elapsed	136.031

Ενδιαφέρον παρουσιάζει και η πληροφορία για τον τύπο της συσκευής από την οποία είχαν οι χρήστες πρόσβαση στην Airbnb, πληροφορία που αντικατοπτρίζεται στον πίνακα 6.7. Το πεδίο αυτό μπορεί να μας πει πολλά για τον χρήστη, καθώς οι συσκευές που χρησιμοποιούμε σήμερα είναι σε μεγάλο βαθμό προσωπικές προτιμήσεις έτσι ακριβώς και ο τρόπος που θέλουμε να περάσουμε τις διακοπές μας. Εδώ όμως επίσης παρατηρούμε ότι τα δεδομένα δεν είναι απόλυτα ορθά καθώς έχουμε επίσης ένα μεγάλο αριθμό ενεργειών που έγιναν από αγνώστου τύπου συσκευή.

Πίνακας 6.7 Καταμέτρηση συσκευών από τις ενέργειες χρηστών

Mac Desktop	3.594.286
Windows Desktop	2.658.539
iPhone	2.105.031
Android Phone	839.637
iPad Tablet	683.414
AndroidApp unknown Phone/Tablet	273.652
-unknown-	211.279
Tablet	139.886
Linux Desktop	28.373
Chromebook	22.348
iPodtouch	8.198
Windows Phone	2.047
Blackberry	979
Opera Phone	68

6.5. Περίληψη

Παρουσιάσαμε σε αυτό το κεφάλαιο τα δεδομένα που δόθηκαν από την Airbnb στους συμμετέχοντες στο διαγωνισμό της Kaggle. Αναλύσαμε τις μεταβλητές των διαθέσιμων συνόλων δεδομένων και εντοπίσαμε κάποιες ελλείψεις και αδυναμίες. Έχοντας τώρα μια ιδέα για τη μορφή και την κατάσταση των δεδομένων, θα προχωρήσουμε στη φάση επεξεργασίας των δεδομένων και στην εφαρμογή του μοντέλου.

7. Παρουσίαση και ανάλυση της λύσης

7.1. Εισαγωγή

Σε αυτό το κεφάλαιο θα παρουσιάσουμε τον συνδυασμό των λύσεων από το διαγωνισμό της Kaggle, ώστε να καταλήξουμε σε ένα μοντέλο η εφαρμογή του οποίου να μπορεί να απαντά με ακρίβεια στο ερώτημα «Σε ποια χώρα θα επιλέξει να ταξιδέψει ένα νέος χρήστης της Airbnb». Το πρώτο μέρος της λύσης αποτελεί η προεπεξεργασία των δεδομένων και είναι το κομμάτι στο οποίο διαφοροποιούνται οι λύσεις των περισσότερων διαγωνιζόμενων. Αφού εξηγήσουμε και παρουσιάσουμε την προεπεξεργασία, θα εκπαιδεύσουμε το μοντέλο μας με τη χρήση του αλγορίθμου δημιουργώντας στο τέλος το αρχείο υποβολής με τον πίνακα των αποτελεσμάτων.

7.2. Προεπεξεργασία δεδομένων

Σε αυτή την ενότητα θα παρουσιάσουμε την προεπεξεργασία που έγινε στα δεδομένα, πριν χρησιμοποιηθούν στον αλγόριθμο. Αυτή η επεξεργασία έγινε από τον David Gasquez, την υλοποίηση του οποίου επιλέξαμε να χρησιμοποιήσουμε στο πρώτο κομμάτι της λύσης και θα παρουσιάσουμε στην παρούσα ενότητα. Αφού αντιγράψαμε (fork) την υλοποίηση σε δικό μας Kernel, τρέξαμε τον κώδικα και τον αναλύσαμε. Θα παρουσιάσουμε στη συνέχεια τον κώδικα αυτό, σχολιάζοντας κάθε τμήμα του χωριστά. Ο κώδικας όπως παρουσιάζεται εδώ, υπάρχει διαθέσιμος και μπορεί να εκτελεστεί μέσα από την πλατφόρμα της Kaggle, στο Kernel που δημιουργήσαμε για αυτό το σκοπό, στο σύνδεσμο: https://www.kaggle.com/mbx17061/airbnb.

Αρχικά θα πρέπει να εισάγουμε τις βιβλιοθήκες που μας είναι απαραίτητες. Για την επεξεργασία των δεδομένων χρειαζόμαστε τις βιβλιοθήκες Numpy και Pandas της Python:

```
import numpy as np
import pandas as pd
```

Στη συνέχεια θα γίνει η φόρτωση των δεδομένων από το διαγωνισμό της Airbnb. Συγκεκριμένα χρειαζόμαστε και θα εισάγουμε τα αρχεία train και test των χρηστών καθώς και το αρχείο sessions:

```
train_users = pd.read_csv('../input/airbnb-recruiting-new-user-bookings/train_users_2.csv')
test_users = pd.read_csv('../input/airbnb-recruiting-new-user-bookings/test_users.csv')
sessions = pd.read_csv('../input/airbnb-recruiting-new-user-bookings/sessions.csv')
```

Θα ενώσουμε τα δύο σύνολα χρηστών με σκοπό τη δημιουργία ενός συνόλου δεδομένων που θα περιέχει όλους τους χρήστες:

```
users = pd.concat([train_users, test_users], axis=0, ignore_index=True, sort=True)
```

Θα προχωρήσουμε στη διαγραφή της στήλης "date_first_booking". Στα δεδομένα του συνόλου test δεν υπάρχουν τιμές για αυτή τη στήλη, με αυτή τη διαγραφή θα αφαιρέσουμε και τις τιμές των train δεδομένων:

```
users.drop('date_first_booking', axis=1, inplace=True)
```

Ηλικία

Ένα από τα πιο σημαντικά χαρακτηριστικά για προβλέψεις συμπεριφοράς χρηστών είναι η ηλικία. Η δυνατότητα εξαγωγής πληροφοριών σχετικά με την ηλικία των χρηστών είναι απαραίτητη για τις μελλοντικές προβλέψεις μας. Μερικοί χρήστες εισήγαγαν έτος αντί της ηλικίας τους.

Η μετατροπή του έτους σε ηλικία:

```
user_with_year_age_mask = users['age'] > 1000
users.loc[user_with_year_age_mask, 'age'] = 2015 - users.loc[user_with_year_age_mask, 'age']
```

Οριοθέτηση της ηλικίας, αφαιρώντας τιμές μεγαλύτερες από 100 και μικρότερες από 18:

```
users.loc[(users['age'] > 100) | (users['age'] < 18), 'age'] = -1
```

Αντικατάσταση των τιμών NaN στη στήλη age με την τιμή -1:

```
users['age'].fillna(-1, inplace=True)
```

Η μεταβλητή ηλικία έχει μεγάλο πλήθος τιμών. Θα δημιουργήσουμε ομάδες και θα τοποθετήσουμε κάθε χρήστη στην αντίστοιχη ηλικιακή ομάδα:

```
bins = [-1, 20, 25, 30, 40, 50, 60, 75, 100]
users['age_group'] = np.digitize(users['age'], bins, right=True)
```

NaN

Ακολουθεί ο υπολογισμός του αριθμού των άγνωστων στοιχείων για κάθε χρήστη. Αυτό μπορεί να είναι χρήσιμο για τον ταξινομητή όταν αποφασίζει πού θα κάνει κράτηση κάποιος χρήστης.

```
users['nans'] = np.sum([
    (users['age'] == -1),
    (users['gender'] == '-unknown-'),
    (users['language'] == '-unknown-'),
    (users['first_affiliate_tracked'] == 'untracked'),
    (users['first_browser'] == '-unknown-')
], axis=0)
```

Μετατροπή ημερομηνιών στη σωστή μορφή:

```
users['date_account_created'] = pd.to_datetime(users['date_account_created'], errors='ignore')
```

```
users['date_first_active'] = pd.to_datetime(users['timestamp_first_active'], format='%Y%m%d%H%M%S')
```

Μετατροπή σε DatetimeIndex:

```
date_account_created = pd.DatetimeIndex(users['date_account_created'])
date_first_active = pd.DatetimeIndex(users['date_first_active'])
```

Διαίρεση των ημερομηνιών σε ημέρα, εβδομάδα, μήνα και έτος:

```
users['day_account_created'] = date_account_created.day
users['weekday_account_created'] = date_account_created.weekday
users['week_account_created'] = date_account_created.week
users['month_account_created'] = date_account_created.month
users['year_account_created'] = date_account_created.year
users['day_first_active'] = date_first_active.day
users['weekday_first_active'] = date_first_active.weekday
users['week_first_active'] = date_first_active.week
users['month_first_active'] = date_first_active.month
users['year_first_active'] = date_first_active.year
```

Ανάκτηση διαφοράς μεταξύ της ημερομηνίας κατά την οποία δημιουργήθηκε ο λογαριασμός και της ημερομηνίας κατά την οποία ήταν ενεργός για πρώτη φορά ο χρήστης:

```
users['time_lag'] = (date_account_created.values - date_first_active.values).astype(int)
```

Διαγραφή διπλών στηλών:

```
drop_list = [
    'date_account_created',
```

```
    'date_first_active',
    'timestamp_first_active'
]

users.drop(drop_list, axis=1, inplace=True)
```

Πληροφορίες session

Υπάρχουν πολλές πληροφορίες στο αρχείο sessions.csv. Θα επικεντρωθούμε στην καταμέτρηση κάθε ενέργειας χρήστη και κάποια στατιστικά στοιχεία για τα δευτερόλεπτα που έχουν περάσει. Περαιτέρω επεξεργασία θα μπορούσε να είναι ωφέλιμη για τις τελικές προβλέψεις.

Μετονομασία στήλης user_id:

```
sessions.rename(columns = {'user_id': 'id'}, inplace=True)
```

Μέτρηση συχνότητας ενέργειας

Καταμέτρηση επαναλήψεων ενέργειων από κάθε χρήστη.

```
action_count = sessions.groupby(['id', 'action'])['secs_elapsed'].agg(len).unstack()
action_type_count = sessions.groupby(['id', 'action_type'])['secs_elapsed'].agg(len).unst
ack()
action_detail_count = sessions.groupby(['id', 'action_detail'])['secs_elapsed'].agg(len).un
stack()
device_type_sum = sessions.groupby(['id', 'device_type'])['secs_elapsed'].agg(sum).unst
ack()

sessions_data = pd.concat([action_count, action_type_count, action_detail_count, devic
e_type_sum],axis=1, sort=True)
sessions_data.columns = sessions_data.columns.map(lambda x: str(x) + '_count')
```

```
# Προτιμόμενη συσκευή
sessions_data['most_used_device'] = sessions.groupby('id')['device_type'].max()

users = users.join(sessions_data, on='id')
```

Στατιστικά για το Elapsed Seconds

Θα εξαγάγουμε κάποιες πληροφορίες για τα δευτερόλεπτα που έχουν περάσει ανά χρήστη. Δεδομένου ότι δεν είναι γνωστό ακριβώς τι σημαίνει το secs_elapsed, υποθέτουμε ότι είναι παύση μεταξύ των ενεργειών:

```
secs_elapsed = sessions.groupby('id')['secs_elapsed']

secs_elapsed = secs_elapsed.agg(
    {
        'secs_elapsed_sum': np.sum,
        'secs_elapsed_mean': np.mean,
        'secs_elapsed_min': np.min,
        'secs_elapsed_max': np.max,
        'secs_elapsed_median': np.median,
        'secs_elapsed_std': np.std,
        'secs_elapsed_var': np.var,
        'day_pauses': lambda x: (x > 86400).sum(),
        'long_pauses': lambda x: (x > 300000).sum(),
        'short_pauses': lambda x: (x < 3600).sum(),
        'session_length': np.count_nonzero
    }
)

users = users.join(secs_elapsed, on='id')
```

Αποθήκευση των δεδομένων μετά την επεξεργασία που έγινε νωρίτερα :

```
users.set_index('id', inplace=True)
users.loc[train_users['id']].to_csv('processed_train_users.csv')
users.loc[test_users['id']].drop('country_destination', axis=1).to_csv('processed_test_users.csv')
```

7.3. Εκπαίδευση του μοντέλου με XGBoost

Αφού έγινε η προεπεξεργασία των δεδομένων, θα συνεχίσουμε με την παρουσίαση της τελικής λύσης, με τη χρήση του αλγορίθμου XGBoost. Σε αυτή την ενότητα θα παρουσιάσουμε τον κώδικα από την υλοποίηση του διαγωνιζόμενου με το όνομα Sandro, σχολιάζοντας κάθε τμήμα χωριστά όπως κάναμε και στην προηγούμενη ενότητα. Τα παρακάτω τμήματα κώδικα με σύντομη επεξήγηση υπάρχουν επίσης διαθέσιμα και στην πλατφόρμα της Kaggle, στο σύνδεσμο: https://www.kaggle.com/mbx17061/xgboost.

Αρχικά θα γίνει και εδώ η φόρτωση των βιβλιοθηκών. Καθώς ο κώδικας που ακολουθεί βρίσκεται σε διαφορετικό Kernel, κάποιες από τις βιβλιοθήκες πρέπει να τις εισάγουμε εκ νέου. Αυτή τη φορά εκτός από της numpy και pandas, θα εισάγουμε τμήματα της βιβλιοθήκης sklearn καθώς και τη βιβλιοθήκη για τον αλγόριθμο XGBoost.

```
import pandas as pd
import numpy as np
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import KFold
import xgboost
```

Στη συνέχεια θα εισάγουμε τα αρχεία χρηστών, train και test, όπως τα αποθηκεύσαμε μετά την επεξεργασία των δεδομένων στην προηγούμενη ενότητα:

```
train_users = pd.read_csv('../input/processed_train_users.csv', sep=';', decimal=",")
df_test = pd.read_csv('../input/processed_test_users.csv', sep=';', decimal=",")
```

Αποθηκεύουμε τις τιμές από τη στήλη χώρα προορισμού στη μεταβλητή labels και στη συνέχεια την αφαιρούμε από το σύνολο των χρηστών. Αποθηκεύουμε επίσης το ID από το σύνολο χρήστων για τεστ σε ξεχωριστή μεταβλητή. Τέλος καταχωρούμε τον αριθμό των καταχωρίσεων (γραμμών) στο σύνολο εκπαίδευσης χρηστών.

```
labels = df_train['country_destination'].values
df_train = df_train.drop(['country_destination'], axis=1)
id_test = df_test['id']
```

```
piv_train = df_train.shape[0]
```

Ένωση των δύο συνόλων χρηστών σε ένα με το όνομα df_all. Έτσι έχουμε ένα σύνολο (dataframe) με όλα τα δεδομένα.

```
df_all = pd.concat((df_train, df_test), axis=0, ignore_index=True)
```

Αφαιρούμε τη στήλη id και κάνουμε αντικατάσταση των NaN με την τιμή -1.

```
df_all = df_all.drop(['id'], axis=1)
df_all = df_all.fillna(-1)
```

Κωδικοποίηση των μεταβλητών με one-hot-encoding (ohe)

```
ohe_feats = ['gender', 'signup_method', 'signup_flow', 'language', 'affiliate_channel', 'affiliate_provider', 'first_affiliate_tracked', 'signup_app', 'first_device_type', 'first_browser']

for f in ohe_feats:
    df_all_dummy = pd.get_dummies(df_all[f], prefix=f)
    df_all = df_all.drop([f], axis=1)
    df_all = pd.concat((df_all, df_all_dummy), axis=1)
```

Διαχωρισμός του συνόλου στις μεταβλητές X και X_test καθώς και τις μεταβλητής y η οποία θα περιέχει τις ετικέτες.

```
vals = df_all.values
X = vals[:piv_train]
le = LabelEncoder()
y = le.fit_transform(labels)
X_test = vals[piv_train:]
```

Χρήση του ταξινομητή του αλγορίθμου XGBoost. Στη συνέχεια υπολογίζουμε τις προβλέψεις.

```
xgb = XGBClassifier(max_depth=6, learning_rate=0.3, n_estimators=25,
                    objective='multi:softprob', subsample=0.5, colsample_bytree=0.5, seed=0)
xgb.fit(X, y)
y_pred = xgb.predict_proba(X_test)
```

Παίρνουμε τις πέντε τιμές χώρας προορισμού με τη μεγαλύτερη πιθανότητα για κάθε χρήστη.

```
ids = []  #list of ids
cts = []  #list of countries
for i in range(len(id_test)):
    idx = id_test[i]
    ids += [idx] * 5
    cts += le.inverse_transform(np.argsort(y_pred[i])[::-1])[:5].tolist()
```

Δημιουργούμε το αρχείο των αποτελεσμάτων. Το αρχείο αυτό περιέχει τους πέντε πιθανότερους προορισμούς για κάθε χρήστη, βάσει των προβλέψεων του μοντέλου.

```
sub = pd.DataFrame(np.column_stack((ids, cts)), columns=['id', 'country'])
sub.to_csv('sub.csv',index=False)
```

Πίνακας 7.1 Προεπισκόπηση αποτελεσμάτων

1	id	country
2	5uwns89zht	NDF
3	5uwns89zht	US
4	5uwns89zht	FR
5	5uwns89zht	IT
6	5uwns89zht	DE
7	jtl0dijy2j	NDF
8	jtl0dijy2j	US
9	jtl0dijy2j	FR
10	jtl0dijy2j	GB
11	jtl0dijy2j	IT
12	xx0ulgorjt	NDF
13	xx0ulgorjt	US
14	xx0ulgorjt	FR
15	xx0ulgorjt	GB

Στον πίνακα 7.1, φαίνονται οι πρώτες γραμμές των αποτελεσμάτων. Ολόκληρο το αρχείο με τα αποτελέσματα βρίσκεται στο Kaggle στο σύνδεσμο: https://www.kaggle.com/mbx17061/xgboost/output

7.4. Περίληψη

Παρουσιάσαμε σε αυτό το κεφάλαιο, ολοκληρωμένη τη λύση του προβλήματος πρόβλεψης της πρώτης κράτησης ενός νέου χρήστη στην Airbnb. Ο κώδικας της λύσης είναι ένας συνδυασμός λύσεων από αυτούς που χρησιμοποιήθηκαν από τους διαγωνιζόμενους David Gasquez και Sandro αντίστοιχα. Έγιναν κάποιες τροποποιήσεις αφαιρώντας τμήματα και από τις δύο υλοποιήσεις ώστε να αποφευχθούν επαναλήψεις. Τμηματοποιήσαμε τον κώδικα και αναλύσαμε σχολιάζοντας κάθε τμήμα του χωριστά. Μέσω των συνδέσμων για τα Kernel στο Kaggle, υπάρχει η δυνατότητα να εκτελεστεί ο παραπάνω κώδικας.

8. Συμπεράσματα

Η Airbnb, προσπαθώντας να βρει τη βέλτιστη λύση σε ένα πρόβλημα που αφορούσε το σύστημα συστάσεών της, απευθύνθηκε στην κοινότητα του Kaggle. Αφού αναλύσαμε το πρόβλημα, παρουσιάσαμε και επεξεργαστήκαμε τα δεδομένα μας, συνδυάσαμε δύο εκ των υποβληθέντων λύσεων και καταλήξαμε σε ένα μοντέλο το οποίο προβλέπει αυτό που ζητούσε η εταιρεία, μπορούμε να παραθέσουμε τα συμπεράσματά μας.

Από το κεφάλαιο της βιβλιογραφίας έγινε αντιληπτή η σημαντικότητα των συστημάτων συστάσεων, ειδικότερα για της διαδικτυακές εταιρείες τουρισμού. Η εκτίμησή μας είναι ότι τα επόμενα χρόνια θα γίνεται όλο και πιο σημαντικό για τις εταιρείες αυτές να μπορούν να διαχειρίζονται σωστά το μεγάλο όγκο των δεδομένων τους. Επενδύοντας στην επιστήμη των δεδομένων και τη μηχανική μάθηση, θα μπορούν οι εταιρείες να παρέχουν εύκολη και γρήγορη πρόσβαση στις σωστές πληροφορίες, έχοντας με αυτό τον τρόπο ευχαριστημένους χρήστες και κατ' επέκταση πελάτες. Το γεγονός αυτό θα δώσει σε αυτές τις εταιρίες ένα σημαντικό ανταγωνιστικό πλεονέκτημα.

Οι ηγέτες του κλάδου έχουν αντιληφθεί αυτή την πραγματικότητα και η Airbnb σαφώς δεν αποτελεί εξαίρεση. Το ερώτημα που γεννάται σε αυτό το σημείο, είναι γιατί η Airbnb, έχοντας ήδη επενδύσει αρκετά στον τομέα αυτό, δεν ήταν σε θέση να βρει μόνη της τη λύση σε αυτό το πρόβλημα που αντιμετώπιζε. Η απάντηση ίσως κρύβεται στο γεγονός ότι ο διαγωνισμός αυτός, ήταν διαγωνισμός πρόσληψης. Το πιθανότερο λοιπόν, είναι ότι η Airbnb είχε ήδη λύση στο πρόβλημα αυτό, και ο διαγωνισμός ήταν απλά μια δοκιμασία αναζήτησης καλών επιστημόνων, οι οποίοι θα μπορούσαν να αποτελέσουν μελλοντικούς συνεργάτες της εταιρίας. Ένας άλλος λόγος θα μπορούσε να είναι η αξιολόγηση της πλατφόρμας και της κοινότητας, με στόχο τη χρήση τους για μελλοντικά προβλήματα. Φυσικά αν κάποια από τις υποβληθείσες λύσεις ήταν καλύτερη από αυτή που ήδη χρησιμοποιούσε η Airbnb εκείνη τη χρονική στιγμή, θα ήταν ένα επιπλέον κέρδος η βελτίωση του αλγορίθμου της.

Το σενάριο ότι η Airbnb δεν είχε απαραίτητα ως κύριο στόχο την εξεύρεση λύσης, ενισχύεται και από την ποιότητα των δεδομένων που έδωσε η ίδια στο κοινό. Όπως είδαμε στην ανάλυσή μας, υπήρχαν αρκετές ελλείψεις στα δεδομένα καθώς και πολλά σφάλματα. Το γεγονός αυτό δυσκολεύει πολύ την εκπαίδευση του μοντέλου, ειδικά αν

στόχος είναι πραγματικά η βέλτιστη λύση. Έτσι είχαμε, για παράδειγμα, μεγάλο ποσοστό έλλειψης της πληροφορίας για το φύλλο του χρήστη καθώς και λάθος δεδομένα για την ημερομηνία γέννησης. Για το πρώτο μπορούμε να δικαιολογήσουμε την έλλειψη του, καθώς πρόκειται για μια πληροφορία που μάλλον προαιρετικά οι χρήστες μπορούν καταχωρήσουν. Το σφάλμα ωστόσο με την ημερομηνία γέννησης θα μπορούσε να είχε αποφευχθεί, με έλεγχο επικύρωσης των δεδομένων κατά την εισαγωγή της πληροφορίας από το χρήστη. Επίσης κάποια από τα στοιχεία θα ήταν καλό να είχαν μια επιπλέον περιγραφή, καθώς για μερικά από αυτά δεν ήταν ευδιάκριτο από το όνομα τους, ποια πληροφορία περιέχουν.

Θετική ήταν η εμπειρία εργασίας μας μέσα από την πλατφόρμα του Kaggle. Ιδιαίτερα για νέους στο αντικείμενο χρήστες, ο τρόπος που λειτουργεί η πλατφόρμα, η ενεργή και φιλική κοινότητα χρηστών, οι τεχνολογικές δυνατότητες που παρέχει καθώς και η ευχρηστία της, την καθιστούν τον ιδανικό τρόπο για να κάνει κάποιος τα πρώτα του βήματα σε αυτό το αντικείμενο. Η εγγραφή στο σύστημα αρκεί για να έχει κάποιος όλα όσα χρειάζεται ώστε να μπορέσει να πειραματιστεί αντιγράφοντας και τροποποιώντας κώδικα άλλων χρηστών, χωρίς να είναι απαραίτητη καμία επιπλέον εγκατάσταση λογισμικού. Αυτό δίνει φυσικά τη δυνατότητα στους χρήστες να δουλεύουν οποιαδήποτε στιγμή από οποιοδήποτε συσκευή έχει πρόσβαση στο διαδίκτυο. Η πρόβλεψή μας είναι ότι η συγκεκριμένη πλατφόρμα θα αναπτυχθεί ακόμη περισσότερο στο μέλλον, καθώς το αντικείμενο που διαπραγματεύεται, αν και θεωρείται ακόμη αρκετά εξειδικευμένο, στο κοντινό μέλλον θα αφορά όλο και περισσότερες εταιρείες και κλάδους. Η ιδέα των διαγωνισμών άλλωστε είναι εξαιρετική, δίνοντας συνεχώς κίνητρα στην κοινότητα να παραμένει ενεργή και να αναπτύσσεται.

Στο διαγωνισμό της Airbnb η συμμετοχή ήταν μεγάλη και αντίστοιχα πολλά ήταν και τα Kernel που δημοσιοποιήθηκαν με τμήματα κώδικα αναφορικά με το θέμα. Αυτό που συμβαίνει ωστόσο, είναι ότι δεν δημοσιοποιείται ολόκληρη η λύση όπως υποβάλλεται, παρά μόνο ένα τμήμα της. Ο λόγος γι' αυτό είναι ότι οι δημοσιοποιήσεις γίνονται στο διάστημα που είναι ενεργός ο διαγωνισμός και έτσι υπάρχει ο φόβος της αντιγραφής από τους διαγωνιζόμενους. Αυτός ήταν άλλωστε ο λόγος για τον οποίο η λύση που παρουσιάσαμε στην εργασία μας, ήταν συνδυασμός δύο υλοποιήσεων. Μια πρόταση για βελτίωση, θα ήταν η δημοσιοποίηση των καλύτερων λύσεων από το Kaggle, μετά το πέρας του διαγωνισμού. Η υλοποίηση της συγκεκριμένης πρότασης θα βοηθούσε ιδιαίτερα για διδακτικούς σκοπούς αλλά και ως εργαλείο έρευνας.

Η επιστήμη των δεδομένων και η επιχειρηματικότητα έρχονται όλο και πιο κοντά, ειδικά όσο μεγαλώνει ο όγκος των δεδομένων που έχουν οι επιχειρήσεις στη διάθεσή τους. Η τεχνολογική εξέλιξη ενισχύει αυτή τη σύνδεση, καθιστώντας εταιρίες που δεν αξιοποιούν αυτές τις δυνατότητες, μη ανταγωνιστικές. Η βιβλιογραφία δεν έχει επεκταθεί ακόμη όσο θα έπρεπε αναλογικά με τη σημαντικότητα του θέματος. Στο μέλλον θα μπορούσαν να χρησιμοποιηθούν παρόμοια προβλήματα με αυτό που παρουσιάσαμε σε αυτή τη εργασία, από άλλες εταιρίες, αναλύοντας τον τρόπο προσέγγισης του προβλήματος καθώς και τη σημασία της επίλυσής του για την εταιρία.

Κατάλογος αναφορών

Adomavicius, G., Tuzhilin, A., 2005. Towards the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering 17 (6), 734–749.

Borras J., Moreno A., Valls A. (2014), Intelligent tourism recommender systems: a survey, Expert Systems with Applications, 41, 7370-7389

Burke, R. (2002). Hybrid recommender systems: Survey and experiments, User Modeling and User-Adapted Interaction, 12(4), 331–370.

Chen, Z., Lin, F., Liu, H., Liu, Y., Ma, W.-Y., & Wenyin, L. (2002). User Intention Modeling in Web Applications Using Data Mining. World Wide Web: Internet and Web Information Systems, pp. 181-191.

David Gasquez Arcos (2016), Competition de Kaggle Airbnb New User Booking

Fesenmeier D., Woeber K., Werthner H. (2006), Destination Recommendation Systems: Behavioural Foundations and Applications, CAB International

Gao, M., Liu, K., & Wu, Z. (2010). Personalisation in web computing and informatics: Theories, techniques, applications, and future research. Information Systems Frontiers, 12, 607–629.

Garcva-Crespo, A., Chamizo, J., Rivera, I., Mencke, M., Colomo-Palacios, R., & Gomez-Berbvs, J. M. (2009). SPETA: Social pervasive e-tourism advisor. Telematics and Informatics, 26, 306–315.

Hunziker W. Krapf K. (1942), Grundriss der Allgemeinen Fremdenverkehrslehre (Outline of the general teaching of tourism), Seminars für Fremdenverkehr und Verkehrspolitik an der Handels-Hochschule St. Gallen. 1. Zurich: Polygraphischer Verlag AG

Kabassi K., (2010), Personalizing recommendations for tourists, Telematics and Informatics 27, 51-66

Kotsiantis, S. B. (2007), Supervised Machine Learning: A Review of Classification Techniques. *Informatica*, 249-268.

Kouroupetroglou Praxitelis - Nikolaos (2017), Machine Learning Techniques for Short-Term Electric Load Forecasting

Mihajlo Grbovic, Haibin Cheng, (2018) Real-time Personalization using Embeddings for Search Ranking at Airbnb

Nelli Fabio (2015), Python Data Analytics: Data Analysis and Science using pandas, matplotlib

Ricci F., Rokach L., Shapira B. (2011), Introduction to Recommender Systems Handbook, Recommender Systems Handbook, Springer, 1-35

Wang, Y., Chan, S. C.-F., & Ngai, G. (2012), Applicability of Demographic Recommender System to Tourist Attractions: A Case, 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology.

Ηγουμενάκης Ν., Κραβαρίτης Κ. (2004), Τουρισμός: Βασικές Έννοιες, Interbooks

Ιστοσελίδες

https://www.airbnb.com

https://www.kaggle.com

http://www2.unwto.org

https://www.travelocity.com

https://www.expediagroup.com

https://www.tripadvisor.com

https://www.bookingholdings.com

https://en.wikipedia.org

https://medium.com/cracking-the-data-science-interview/the-5-machine-learning-use-cases-that-optimize-your-airbnb-travel-experience-fb027a56e5a5

https://medium.com/airbnb-engineering/how-airbnb-uses-machine-learning-to-detect-host-preferences-18ce07150fa3

https://github.com/davidgasquez/kaggle-airbnb/tree/master/notebooks

https://neilpatel.com/blog/how-airbnb-uses-data-science/