ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΑΡΑΓΩΓΗΣ ΚΑΙ ΔΙΟΙΚΗΣΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΚΑΤΕΥΘΥΝΣΗ:ΟΡΓΑΝΩΣΗ ΚΑΙ ΔΙΟΙΚΗΣΗ

ΠΡΟΒΛΕΨΗ ΤΙΜΩΝ ΚΑΙ ΕΞΟΡΥΞΗ ΓΝΩΜΗΣ ΑΠΟ ΚΡΙΤΙΚΕΣ ΣΕ ΔΙΑΔΙΚΤΥΑΚΕΣ ΠΛΑΤΦΟΡΜΕΣ ΒΡΑΧΥΧΡΟΝΙΑΣ ΜΙΣΘΩΣΗΣ ΑΚΙΝΗΤΩΝ

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ:ΤΣΑΦΑΡΑΚΗΣ ΣΤΕΛΙΟΣ ΕΠΙΚΟΥΡΟΣ ΚΑΘΗΓΗΤΗΣ

ΙΑΚΩΒΟΣ ΣΤΑΥΡΟΥΛΑΚΗΣ

ΧΑΝΙΑ 2020

Περίληψη

Σκοπός της παρούσας μεταπτυχιακής εργασίας είναι η πρόβλεψη των τιμών των χαρακτηριστικών καταλυμάτων που έχουν εισαχθεί στην πλατφόρμα Airbnb στην περιοχή των Αθηνών με την χρήση μεθόδων μηχανικής μάθησης.

Το πρώτο χαρακτηριστικό είναι η τιμή των καταλυμάτων ανά ημέρα όπου παρατηρήθηκε ποιοι παράγοντες επηρεάζουν την τιμή όπως η απόσταση από αξιοθέατα, προηγούμενες κριτικές, ανέσεις δωματίου. Τα κριτήρια για να γίνει ένας οικοδεσπότης Superhost όπως για παράδειγμα η ταχύτητα απάντησης. Τέλος, μελετήθηκε πως συγκεκριμένες λέξεις κλειδιά επηρεάζουν τις κριτικές των καταλυμάτων και συγκεκριμένα αν έχουν θετικό ή αρνητικό αντίκτυπο στην κριτική ενός καταλύματος με την χρήση της μεθόδου Εξόρυξης γνώμης(Sentiment Analysis). Έγινε χρήση των ακόλουθων τεχνικών μηχανικής μάθησης: Απλή γραμμική παλινδρόμηση, Ridge & Lasso Παλινδρόμηση, Τυχαία Δέντρα Αποφάσεων και Τυχαία Δάση Αποφάσεων με την χρήση των προγραμμάτων tableau, SPSS, Microsoft Excel, Dataiku όπου πραγματοποιήθηκε εκτενής σύγκριση των τεχνικών ώστε να προκύψει το βέλτιστο και πιο αντιπροσωπευτικό αποτέλεσμα.

Λέξεις κλειδιά: Μηχανική Μάθηση,Airbnb, Superhost, Sentimental Analysis

Περιεχόμενα

1. Εισαγωγή

Ο άνθρωπος ανέκαθεν ήθελε να έχει την δυνατότητα πρόβλεψης διαφόρων καταστάσεων για μπορεί να έχει συγκριτικό πλεονέκτημα σε σχέση με τους ανταγωνιστές του. Στις μέρες μας πλέον πραγματοποιείται εκτεταμένη χρήση επιστημονικών εργαλείων με στόχο την πρόβλεψη. Το κύριο επιστημονικό εργαλείο πρόβλεψης είναι ο μαθηματικός κλάδος και στην συγκεκριμένη εργασία θα χρησιμοποιηθούν διάφορες τεχνικές μηχανικής μάθησης για να πετύχουμε ένα αξιόπιστο αποτέλεσμα.

Η συγκεκριμένη εργασία έχει συγγραφεί για να βρεθεί η χρησιμότητα διαφόρων τεχνικών μηχανικής μάθησης στον τομέα του τουρισμού και συγκεκριμένα στην πλατφόρμα της εταιρίας Airbnb για την περιοχή της Αθήνας.

Η χρήση των ακόλουθων τεχνικών αποτελεί μια πρωτοποριακή μέθοδο ανάλυσης των δεδομένων στον τουριστικό κλάδο και μπορεί να δώσει άμεσες ενδείξεις για την

ποιότητα των καταλυμάτων αλλά και τρόπους βελτίωσης τους έτσι ώστε να μεγιστοποιηθούν τα κέρδη των ιδιοκτητών.

Αρχικά, πραγματοποιείται μια εισαγωγή στην πλατφόρμα Airbnb που αφορά την ιστορία της εταιρίας, την λειτουργία της αλλά και μια σύντομη ανάλυση του κλάδου του peer to peer τουρισμού.

Στη συνέχεια, έγινε μια βιβλιογραφική προσέγγιση στις μαθηματικές μεθόδους που χρησιμοποιήθηκαν για να παραχθούν τα αποτελέσματα της έρευνας. Ακολούθησε μια περιγραφική ανάλυση των δεδομένων που συλλέχθηκαν και έπειτα μια μοντελοποίηση των δεδομένων με τις μεθόδους μηχανικής μάθησης. Στο τέλος της μελέτης παραθέτονται κάποια ενδεικτικά συμπεράσματα αλλά και μελλοντικές κατευθύνσεις που μπορούν να ακολουθηθούν.

1.1 Η πλατφόρμα του Airbnb στην Ελλάδα και τον Κόσμο

Για όλους όσους αναζητούν μια κατοικία η οποία να θυμίζει το προσωπικό τους σπίτι το Airbnb έχει αναδειχθεί ως η βέλτιστη λύση για καταλύματα ανά τον κόσμο. Η συγκεκριμένη πλατφόρμα από την αρχή της λειτουργίας της έχει γνωρίσει τεράστια επιτυχία αφού δραστηριοποιείται σε περισσότερες από 190 χώρες στον κόσμο. Παρόλο τον ανταγωνισμό έμμεσο ή άμεσο από ανάλογες πλατφόρμες όπως

το Tripadvisor και το Booking έχει καταφέρει να ξεχωρίσει μέσω των πρωτοποριακών χαρακτηριστικών και της απλοποιημένης χρήσης με αποτέλεσμα να έχει αυξήσει δραματικά τον αριθμό των χρηστών του. Η πλατφόρμα Airbnb φιλοξενεί αξιόλογους ταξιδιώτες και οικοδεσπότες οι οποίοι συμβάλλουν στην ανάπτυξη του τουρισμού στην περιοχή τους. Το 97 τοις εκατό της τιμής του καταλύματος παραμένει στον ίδιο τον οικοδεσπότη ενώ η πλατφόρμα δεσμεύει μόλις το 3% για τις υπηρεσίες που προσφέρει (Airbnb, 2015).

Αυτές οι καταχωρίσεις περιλαμβάνουν από μικρά διαμερίσματα έως σπίτια πλήρους μεγέθους και καλύπτουν όλα τα επίπεδα του προϋπολογισμού σε ολόκληρο τον κόσμο, από τη βασική έως την υπερσύγχρονη.

Η Airbnb κατέγραψε περισσότερες από 40 εκατομμύρια αφίξεις επισκεπτών το 2015. Αυτό ήταν περισσότερο από το συνολικό αριθμό των κρατήσεων για τα έξι προηγούμενα χρόνια μετά, όταν η εταιρεία κοινής χρήσης κατοικίας ξεκίνησε για πρώτη φορά στο Σαν Φρανσίσκο το 2008.

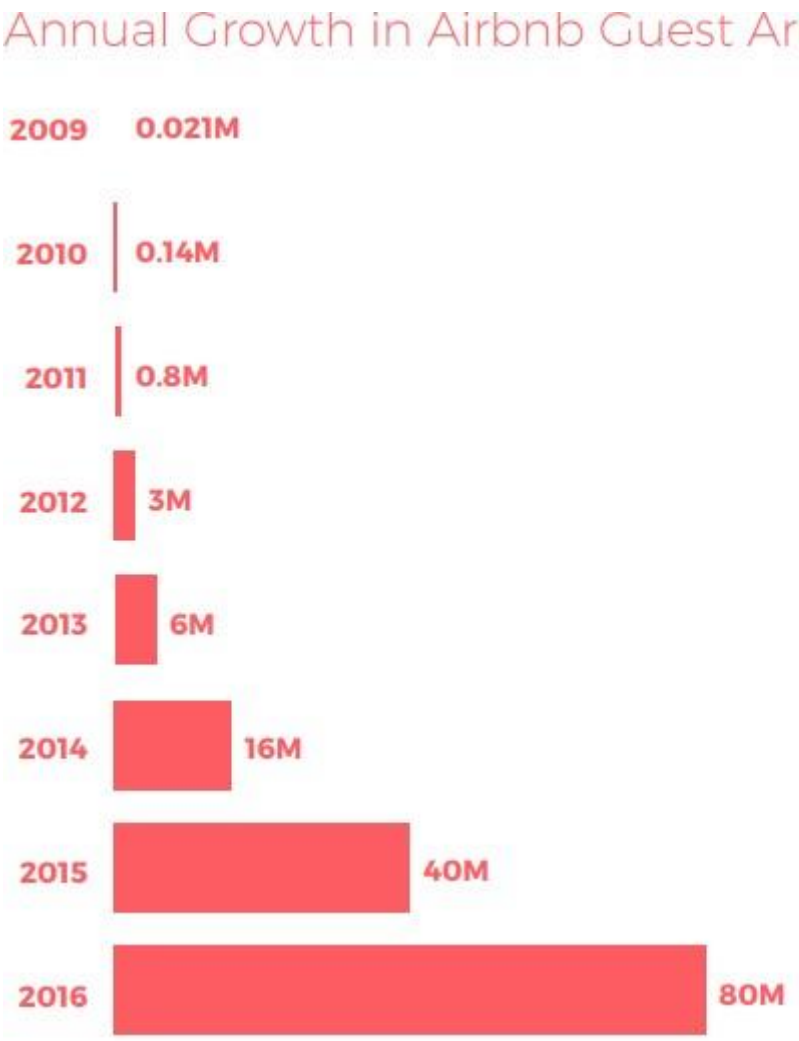
Το 2016, ο αριθμός αυτός διπλασιάστηκε και έπεσε σε 80 εκατομμύρια αφίξεις φιλοξενούμενων παγκοσμίως. Μέχρι σήμερα, η ιστοσελίδα της Airbnb περιλαμβάνει περισσότερα από τρία εκατομμύρια καταλόγους σε 65.000 πόλεις σε 192 χώρες. Αυτό το παγκόσμιο δίκτυο τοπικών οικοδεσποτών είναι το θεμέλιο για την ετικέτα της Airbnb: "Belong Anywhere" (Zervas, et al., 2015) .

Αυτήν την περίοδο υπάρχουν πάνω από ένα εκατομμύριο άνθρωποι που διαμένουν σε ένα κατάλυμα Airbnb σε οποιαδήποτε δεδομένη νύχτα του έτους. Αυτό ισοδυναμεί με μια κατανεμημένη, κινητή πόλη μεσαίου μεγέθους, περίπου το μέγεθος του Σαν Φρανσίσκο ή του Μιλάνου, που κατοικείται από παγκόσμιους πολίτες που θέλουν να εξερευνήσουν τον κόσμο διαμένοντας στο σπίτι κάποιου άλλου.

Η δημοτικότητα του Airbnb αυξάνεται τόσο στους ταξιδιώτες αναψυχής όσο και στους ταξιδιώτες που ταξιδεύουν για επαγγελματικούς λόγους (που αντιπροσωπεύουν περίπου το 10% των αφίξεων και το 15% των ταξιδιωτικών διανυκτερεύσεων). Τα επαγγελματικά ταξίδια τριπλασιάστηκαν πέρυσι.

Φέτος, περίπου το 25% των ταξιδιωτών αναψυχής αναμένεται να πραγματοποιήσουν τουλάχιστον μία διαμονή μέσω της πλατφόρμας Airbnb, από 19% πέρυσι, σύμφωνα με μια έκθεση που πραγματοποιήθηκε στα τέλη του περασμένου έτους από την Morgan Stanley Research. Σύμφωνα με την έκθεση, το 23% των επιχειρηματιών θα χρησιμοποιήσουν το Airbnb φέτος, από 18% πέρυσι.

Όσον αφορά το μέλλον, η ανάπτυξη αναμένεται να συνεχιστεί. Μια μελέτη του 2016 από την εταιρεία παροχής χρηματοοικονομικών υπηρεσιών Cowen και Company με έδρα τη Νέα Υόρκη εκτιμά ότι οι ετήσιες αφίξεις Airbnb θα μπορούσαν ενδεχομένως να αυξηθούν σε μισή δισεκατομμύριο ετήσιες αφίξεις σε πέντε χρόνια και ένα δισεκατομμύριο έως το 2025 (Oates, 2017).



1.1 Ετήσιος αριθμός αφίξεων επισκεπτών σε καταλύματα μέσω της πλατφόρμας Airbnb 2009-2016 (Oates, G., 2017. Demystifying Airbnb for corporate travel managers)

Η Ελλάδα ως ένας από τους κύριους πόλους έλξης της τουριστικής κίνησης στην Μεσόγειο και γενικότερα στην Ευρώπη έχει αρχίσει τα τελευταία χρόνια να υιοθετεί την πλατφόρμα του Airbnb σε όλο και περισσότερα καταλύματα. Ειδικότερα στην Αθήνα όπου γίνεται και η συγκεκριμένη μελέτη παρουσιάζεται αρκετά μεγάλη κινητικότητα από υποψήφιους οικοδεσπότες οι οποίοι επιθυμούν να φιλοξενήσουν στην κατοικία τους άτομα από κάθε μεριά της υφηλίου. Ειδικότερα στο κέντρο της πόλης χωρίς να περιλαμβάνονται τα προάστια όπου επίσης η σχετική δραστηριότητα ανθεί παρόλο που δεν διαθέτουν την αίγλη του κέντρου, ο αριθμός των προς εκμίσθωση ακινήτων ανέρχεται πλέον σε 5.127, όταν μόλις πριν από εννέα μήνες το σχετικό μέγεθος δεν ξεπερνούσε τις 2.500 κατοικίες (Roussanoglou, 2017).

Όπως προκύπτει από την έρευνα , όσο πιο κοντά βρίσκεται το ακίνητο στο ιστορικό κέντρο της πόλης –και δη σε περιοχές που προσφέρουν άμεση πρόσβαση σε μέσα μεταφοράς, και οι οποίες βρίσκονται κοντά σε τουριστικά αξιοθέατα και μουσεία– τόσο υψηλότερο είναι το ενοίκιο. Συνοικίες όπως το Κουκάκι, η Πλάκα, του Μακρυγιάννη, το Θησείο, η Ακρόπολη, το Μοναστηράκι, το Μεταξουργείο και ο Κεραμεικός καταγράφουν τόσο υψηλή ζήτηση τους τελευταίους μήνες, ώστε οι τιμές

των ενοικίων έχουν αρχίσει να ακολουθούν ανοδική πορεία, τη στιγμή που στην υπόλοιπη αγορά τα ενοίκια εξακολουθούν να έχουν πτωτική πορεία. Το ίδιο ισχύει και για ακίνητα τα οποία προσφέρουν υψηλές ποιοτικές προδιαγραφές και θέα στον Λυκαβηττό και στην Ακρόπολη (Roussanoglou, 2017).

Σύμφωνα με την πρόσφατη μελέτη που πραγματοποίησε η Grant Thornton για λογαριασμό του Ξενοδοχειακού Επιμελητηρίου Ελλάδος, υπολογίζεται ότι ο συνολικός αριθμός των ακινήτων που έχουν καταχωρισθεί προς εκμίσθωση μέσω των ηλεκτρονικών πλατφορμών το 2017 ανέρχεται σε 42.155 πανελλαδικά. Επιπλέον 21.716 ακίνητα εκμισθώνονται ως τουριστικά καταλύματα (διαθέτουν δηλαδή το σχετικό σήμα του ΕΟΤ), ανεβάζοντας τον συνολικό αριθμό των κατοικιών και διαμερισμάτων που προορίζονται για εκμετάλλευση σε 63.871. Σε ετήσια βάση, τα ακίνητα αυτά συνιστούν μια αγορά της τάξεως των 1,7 δισ. ευρώ (εκτίμηση για το 2017). Εξ αυτών, περίπου 860 εκατ. ευρώ θα αφορούν τη δαπάνη για τη διαμονή των ξένων επισκεπτών, ενώ τα υπόλοιπα αφορούν τις δαπάνες των ανθρώπων αυτών για τις ημέρες της διαμονής τους, σε σίτιση, ψυχαγωγία κλπ (Roussanoglou, 2017).

Η πόλη της Αθήνας έχει ρυθμό αύξησης των καταλυμάτων 65% περίπου κάθε χρόνο. Συγκεκριμένα το 2011 όπου και ξεκίνησε η δραστηριότητα νοικιαζόντουσαν μέσω Airbnb 32 κατοικίες και σήμερα νοικιάζονται 14.252 κατοικίες. Η μεγάλη ανάπτυξη δε, ήρθε μετά την ταραχώδη χρονιά του 2015. Ενδεικτικά εν καιρώ κρίσης και δημοψηφίσματος νοικιαζόντουσαν 3.318 κατοικίες ενώ ένα χρόνια αργότερα, το 2016 ο αριθμός είχε διπλασιαστεί φτάνοντας τις 6.997 το ίδιο και 2017 όπου έφτασαν τις 11.580.

Τα σημερινά 14.252 καταλύματα των Αθηνών νοικιάζουν 4.516 «ιδιοκτήτες». Ωστόσο η πλειονότητα αυτών που νοικιάζουν είναι απλοί ιδιοκτήτες καταλυμάτων, ποσοστό της τάξης του 80%. Ωστόσο περί τους 1000 ιδιοκτήτες – συγκεκριμένα 922 – έχουν πολλαπλές καταχωρήσεις ακινήτων, δηλαδή είναι είτε διαχειρίστριες εταιρείες είτε επενδυτές ιδιοκτήτες ακινήτων (Insider.gr, 2018).

Το 2019 ο πρόεδρος του Πανελληνίου Συλλόγου Διαχειριστών Ακινήτων - ΠΑΣΥΔΑ Ανδρέας Χίου παρουσίασε στατιστικά στοιχεία που αφορούν τη βραχυχρόνια μίσθωση στην Αθήνα. Συγκεκριμένα, τον Απρίλιο του 2019, στην πρωτεύουσα δραστηριοποιούνταν 8.989 ενεργά καταλύματα στις πλατφόρμες, με τη συντριπτική πλειοψηφία αυτών να είναι ολόκληρα σπίτια ενώ μόλις το 10% μεμονωμένα δωμάτια. Περισσότεροι από 340.000 τουρίστες επισκέφθηκαν έτσι την Αθήνα το 2018, αφήνοντας 89,5 εκατομμύρια ευρώ στην πόλη (Lifo.gr, 2019).

1.2.Περιληπτικά στατιστικά στοιχεία παγκοσμίως για την πλατφόρμα Airbnb (Introduction to Airbnb)

1.2 Ιστορικά στοιχεία της πλατφόρμας Airbnb

Το 2007, οι σχεδιαστές της πλατφόρμας Brian Chesky και Joe Gebbia δεν μπορούσαν να αντέξουν οικονομικά το ενοίκιο στο διαμέρισμα που κατοικούσαν στο Σαν Φρανσίσκο. Για να αυξήσουν τα έσοδα τους, αποφάσισαν να μετατρέψουν τη σοφίτα τους σε χώρο διαμονής δημιουργώντας την δικιά τους ιστοσελίδα. Με την χρήση της συγκεκριμένης ιστοσελίδας πραγματοποίησαν τις 3 πρώτες εκμισθώσεις , όπου ο κάθε ταξιδιώτης πλήρωνε $ 80, και μετά από αυτό το πρώτο Σαββατοκύριακο άρχισαν λαμβάνουν μηνύματα ηλεκτρονικού ταχυδρομείου από μέρη όπως το

Μπουένος Άιρες, το Λονδίνο και η Ιαπωνία τα οποία ζητούσαν πότε θα ήταν διαθέσιμο το κατάλυμα.

Το καλοκαίρι του 2009, το Airbnb δεν είχε καταφέρει να διεισδύσει στην αγορά της Νέας Υόρκης όσο θα επιθυμούσαν οι δημιουργοί του και έτσι οι Gebbia και Chesky χρησιμοποίησαν πειραματικά 24 καταλύματα για να καταλάβουν ποιο ήταν το πρόβλημα. Όπως αποδείχθηκε, οι οικοδεσπότες δεν προσπαθούσαν αρκετά για να αναδείξουν τα καταλύματα τους (Brown, 2016).

Έτσι λοιπόν το ζευγάρι των δημιουργών σκέφτηκε την πιο απλή αλλά αποτελεσματική λύση, νοίκιασε μια φωτογραφική μηχανή αξίας 5.000 δολαρίων και πήγε σε σπίτια λαμβάνοντας επαγγελματικές εικόνες από όσο το δυνατόν περισσότερα καταλύματα της Ν. Υόρκης. Οι νέες φωτογραφίες οδήγησαν σε δύο έως τρεις φορές περισσότερες κρατήσεις στις λίστες της Νέας Υόρκης και μέχρι το τέλος του μήνα τα έσοδα του Airbnb στην πόλη είχαν διπλασιαστεί.

Το 2014 η εταιρία αποφάσισε την επέκταση σε παγκόσμια κλίμακα ώστε να μεγαλώσει την κερδοφορία της. Ένας από τους τρόπους μέσω των οποίων το Airbnb δοκίμασε την παγκόσμια αγορά ήταν μέσω μιας δοκιμής στην Γαλλία. Το Airbnb επέλεξε να προωθηθεί τυχαία στις μισές διεθνείς πόλεις με φυσικές διαφημίσεις και στις άλλες μισές μέσω της χρήσης διαφημίσεων Facebook .Τελικά αποδείχθηκε ότι οι διαφημίσεις μέσω Facebook δεν είχαν τα ανάλογα αποτελέσματα και περιορίστηκαν.

Στα τέλη του 2013, το Airbnb αποφάσισε να ξεκινήσει εκ νέου την πολιτική referral . Μετά από 3 μήνες και 30.000 γραμμές κώδικα, το πρόγραμμα παραπομπών ξεκίνησε ξανά τον Ιανουάριο του 2014.Εξετάστηκαν όλες οι μεταβλητές και αποδείχθηκε ότι όταν η προσφορά από μηνύματα ηλεκτρονικού ταχυδρομείου που έστελνε το Airbnb σε εν δυνάμει άτομα που θα λάμβαναν referrals είχε αλτρουιστικό χαρακτήρα οδηγούσε σε μεγαλύτερο όγκο referrals από τους ίδιους (Brown, 2016).

Το νέο πρόγραμμα παραπομπών της Airbnb οδήγησε σε εκατοντάδες χιλιάδες διανυκτερεύσεις που κρατήθηκαν με παραπομπή των χρηστών το 2014, και οι παραπομπές αυξήθηκαν κατά 25% σε ορισμένες αγορές. Το Airbnb διαπίστωσε ότι οι χρήστες που κάνουν χρήση referrals τείνουν να είναι καλύτεροι από τον μέσο χρήστη αφού τείνουν να παραμένουν δεσμευμένοι με την πλατφόρμα και κλείνουν μελλοντικά ταξίδια.Εκτός αυτού είναι πολύ πιθανό να στείλουν και οι ίδιοι referrals σε άλλους χρήστες.

Τον Ιούλιο του 2014, το Airbnb ανέφερε ότι πάνω από 17 εκατομμύρια επισκέπτες είχαν χρησιμοποιήσει την υπηρεσία, με πάνω από ένα εκατομμύριο επισκέπτες να το χρησιμοποιούν κάθε μήνα (Friedman, 2014). Τον Ιανουάριο του 2015 το Airbnb ανέφερε ότι 30 εκατομμύρια επισκέπτες είχαν χρησιμοποιήσει την υπηρεσία, με σχεδόν 20 εκατομμύρια χρήστες να την χρησιμοποιούν μόνο το

2014 (Chesky, 2015). Αργότερα το 2015 η εταιρεία ανέφερε ότι η ιστοσελίδα είχε 17 εκατομμύρια επισκέπτες μόνο το καλοκαίρι εκείνου του έτους (Airbnb, 2015c). Στις αρχές του 2016 η εταιρεία ανέφερε ότι πάνω από 60 εκατομμύρια επισκέπτες είχαν χρησιμοποιήσει την υπηρεσία (Airbnb, 2016a). Και το καλοκαίρι του 2016 η πλατφόρμα δήλωσε ότι 100 εκατομμύρια επισκέπτες είχαν χρησιμοποιήσει την υπηρεσία (Chafkin & Newcomer, 2016).

Η δημοτικότητα του Airbnb αυξάνεται τόσο στους ταξιδιώτες αναψυχής όσο και στους ταξιδιώτες που ταξιδεύουν για επαγγελματικούς λόγους (οι οποίοι

αντιπροσωπεύουν περίπου το 10% των αφίξεων και το 15% των ταξιδιωτικών διανυκτερεύσεων). Τα επαγγελματικά ταξίδια τριπλασιάστηκαν πέρυσι.

Φέτος, περίπου το 25% των ταξιδιωτών αναψυχής αναμένεται να κλείσουν μια διαμονή στην Airbnb τουλάχιστον μία φορά, από 19 τοις εκατό πέρυσι, σύμφωνα με μια έκθεση στα τέλη του περασμένου έτους από την Morgan Stanley Research. Σύμφωνα με την έκθεση, το 23% των επιχειρηματιών θα χρησιμοποιήσουν την Airbnb φέτος, από 18% πέρυσι (Molla, 2017).

Δημιουργώντας μια φήμη της εξατομίκευσης, της αξιοπιστίας και της εμπιστοσύνης, το Airbnb προσέλκυσε τους χρήστες που πραγματικά πίστευαν στην αξία της επωνυμίας που δημιούργησε η εταιρεία. Για να δημιουργήσει αυτό το οικοσύστημα, το Airbnb χρησιμοποίησε τα έσοδα από αμοιβές συναλλαγών για την εφαρμογή συστημάτων όπως βελτιωμένη επαλήθευση πελατών ,ασφάλιση κλοπής / ζημιών εκατομμυρίων δολαρίων, αυθεντικές κριτικές πελατών και κοινωνικές συνδέσεις. Τα παραπάνω χαρακτηριστικά οδήγησαν την συγκεκριμένη startup στο να γίνει η μεγαλύτερη peer to peer πλατφόρμα ηλεκτρονικής ενοικίασης παγκοσμίως.

1.3 Peer to peer και ανταγωνισμός

Η έννοια της peer to peer οικονομίας προέκυψε περίπου πριν από 15 χρόνια. Από τότε, έχει προωθήσει πολλές εταιρείες με αποτίμηση σε δισεκατομμύρια δολάρια, και ανακατασκεύασε τα θεμέλια πολλών διαφορετικών βιομηχανιών επαναπροσδιορίζοντας τα βασικά επιχειρηματικά τους μοντέλα (Francis, et al., 2016).

Τα κύρια στοιχεία της αγοράς ενοικίασης peer to peer είναι η εμπιστοσύνη μεταξύ των 2 πλευρών αλλά και η άμεση χρήση της τεχνολογίας ώστε να υπάρχει μεγαλύτερη προσβασιμότητα.

Οι συστάσεις και οι κριτικές από τους προηγούμενους χρήστες αποτελούν βασικούς παράγοντες για την διάδοση της peer to peer ενοικίασης και η σύγχρονη μορφή παραπομπών πελατών και εγκρίσεων μέσω της τεχνικής word of mouth . Οι μέσες αξιολογήσεις, ο αριθμός αξιολογήσεων και το feedback των χρηστών διαδραματίζουν σημαντικό ρόλο στις αποφάσεις των πελατών σχετικά με την αγορά και την κατανάλωση υπηρεσιών και υλικών αγαθών . Επιπλέον, ο πολλαπλασιασμός των συνδέσεων στο Διαδίκτυο, των κινητών συσκευών, των εργαλείων συνεργασίας και της άνοδος των κοινωνικών μέσων έχει καταλυτικό ρόλο στην επικοινωνία μεταξύ των χρηστών ώστε να μπορούν να ανταλλάζουν απόψεις για ένα προϊόν ή υπηρεσία.

Η ραγδαία ανάπτυξη των smartphones και tablet έχει βοηθήσει στην διείσδυση της peer to peer οικονομίας στον απλό καταναλωτή . Αυτό σημαίνει ότι οι πελάτες μπορούν να προσφέρουν και να εντοπίζουν αγαθά και υπηρεσίες πιο συχνά - οποτεδήποτε, οπουδήποτε. Ως αποτέλεσμα, οι νεοσύστατες εταιρείες στην εξαρτώνται σε μεγάλο βαθμό από τις εφαρμογές για κινητά.

Σήμερα, πολλές πλατφόρμες κοινής χρήσης χρησιμοποιούν κοινωνικά δίκτυα για να εκτελούν εκστρατείες μάρκετινγκ, να διαχειρίζονται δημόσιες σχέσεις και να διατηρούν επικοινωνίες. Οι συνομιλίες στα μέσα κοινωνικής δικτύωσης μπορούν συχνά να αποκαλύψουν αλλαγές στις αντιλήψεις των καταναλωτών και να συμβάλλουν στη βελτίωση της εμπειρίας των πελατών.

Τα συστήματα ηλεκτρονικής ανταλλαγής (peer-to-peer) περιλαμβάνουν τη μεταφορά των αποτιμημένων πόρων όπως τα αγαθά και οι υπηρεσίες (δηλαδή, κοινωνική ανταλλαγή) μεταξύ ομάδων που πιθανώς δεν έχουν συναντηθεί ποτέ πρόσωπο με πρόσωπο πριν.

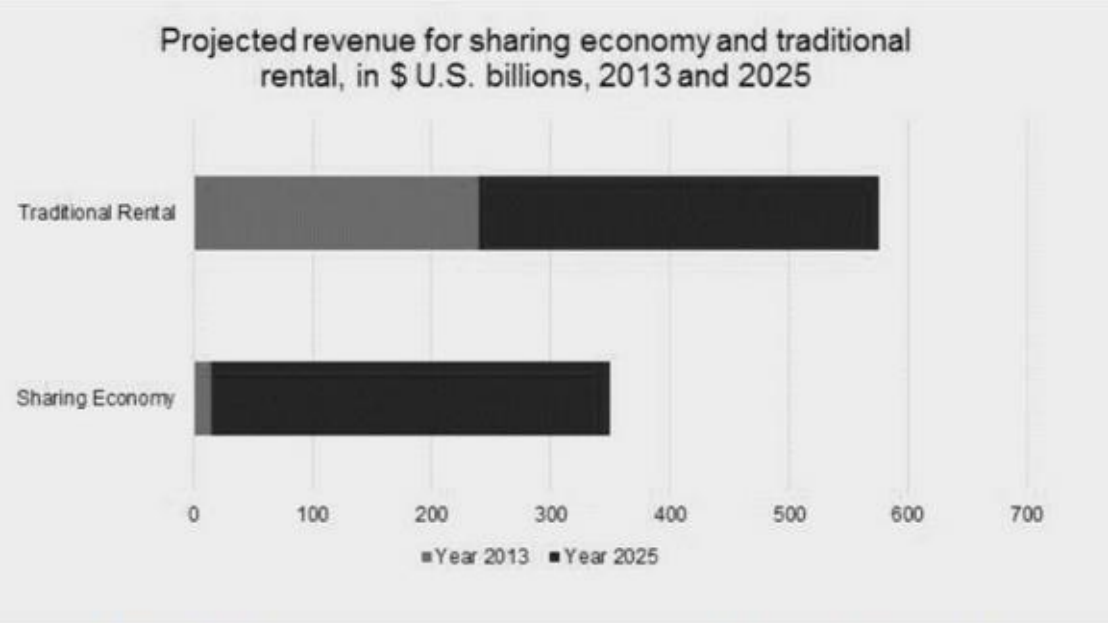
Ο τομέας της peer to peer online ενοικίασης έχει αποκτήσει τεράστια ανάπτυξη τα τελευταία χρόνια. Αυτή η τάση οφείλεται κυρίως σε μεγάλο βαθμό στις αυξήσεις της προσβασιμότητας στο Διαδίκτυο, σε απευθείας σύνδεση με τα κοινωνικά δίκτυα, την κινητή τεχνολογία, τις υπηρεσίες που βασίζονται στην τοποθεσία. Η δυνατότητα ενοικίασης μεγάλου αριθμού προϊόντων και υπηρεσιών οδηγεί στην πιο άμεση επέκταση του δικτύου peer to peer ενοικίασης (Elizaveta, 2016).

Αντί λοιπόν ο καταναλωτής να νοικιάσει μέσω γραφείου ενοικίασης αυτοκινήτων (Hertz, Enterprise) ή εταιρείας κρατήσεων ξενοδοχείων (Hotels.com, Expedia), μπορεί τώρα να χρησιμοποιήσει ένα σύνολο peer to peer αγορών όπου μπορεί να βρει περισσότερη ευελιξία και συχνά μια καλύτερη τιμή. Με την εμφάνιση της peer to peer οικονομίας, υπάρχει αύξηση του ανταγωνισμού στην αγορά ενοικίασης, καθώς υπάρχουν ανταγωνιστές τόσο στην παραδοσιακή αγορά ενοικίασης (Marriott, Westin) όσο και ανταγωνιστές που προσφέρουν παρόμοιες υπηρεσίες στην ηλεκτρονική αγορά peer-to-peer (Airbnb,Wimdu, Roomorama).

Μια μελέτη του 2016 του Ευρωπαϊκού Κοινοβουλίου εκτιμά ότι το θεωρητικό μέγιστο συνολικό οικονομικό όφελος που συνδέεται με την αποτελεσματικότερη χρήση των ικανοτήτων της peer to peer ενοικίασης θα φτάσει την ετήσια κατανάλωση 572 δισ. ευρώ σε ολόκληρη την ΕΕ (Juul, 2017).

Μια έρευνα του Ευρωβαρόμετρου που κυκλοφόρησε επίσης το 2016 έδειξε έντονο ενδιαφέρον για των καταναλωτών για την peer to peer οικονομία: το 52% των ερωτηθέντων γνώριζε τις υπηρεσίες κοινής οικονομίας πλατφόρμες και το 17% είχε χρησιμοποιήσει τέτοιες υπηρεσίες τουλάχιστον μία φορά. Οι ερωτηθέντες ηλικίας 25 ετών και 39 ετών (27%) και εκείνοι που ολοκλήρωσαν την εκπαίδευση ηλικίας 20 ετών και άνω (27%) ήταν πιο πιθανό να χρησιμοποιήσουν αυτές τις πλατφόρμες (Ravi, 2017).

Όπως δείχνει το παρακάτω διάγραμμα , τα επόμενα δέκα χρόνια, η αύξηση των εσόδων από την παραδοσιακή βιομηχανία ενοικίασης θα είναι μικρή σε σύγκριση με την έκρηξη των εσόδων στην peer to peer οικονομία.



Source: "The sharing economy—sizing the revenue opportunity," (Hawksworth et al., 2014)

1.3.Μελλοντικά έσοδα για την παραδοσιακή ενοικίαση και την peer to peer ενοικίαση το 2013 και 2025

Το μέλλον της peer to peer sharing οικονομίας έχει άκρως θετικές προοπτικές. Όμως, παρόλες τις προοπτικές υπάρχουν αρκετά εμπόδια τα οποία θα πρέπει να προσπεραστούν άμεσα.

Η κύρια ανησυχία είναι η κανονιστική αβεβαιότητα. Οι ενοικιαστές δωματίων θα υπόκεινται π.χ. σε φόρους ξενοδοχείων; Στο Άμστερνταμ και την Ελλάδα μέχρι πρόσφατα οι υπάλληλοι χρησιμοποιούν καταχωρήσεις Airbnb για να εντοπίσουν ξενοδοχεία χωρίς άδεια. Ο κίνδυνος είναι ότι παρόλο που ορισμένοι κανόνες πρέπει να επικαιροποιηθούν για την προστασία των καταναλωτών από βλάβες, οι κατεστημένοι φορείς θα προσπαθήσουν να καταστρέψουν τον ανταγωνισμό. Οι άνθρωποι που ενοικιάζουν δωμάτια πρέπει να καταβάλλουν φόρο, φυσικά, αλλά δεν πρέπει να ρυθμίζονται όπως ένα ξενοδοχείο Ritz-Carlton (Economist, 2013).

Η οικονομία κοινής χρήσης αποτελεί ένα ακόμα παράδειγμα της αξίας του διαδικτύου για τους καταναλωτές . Αυτό το αναδυόμενο μοντέλο είναι πλέον μεγάλο και αποδιοργανωτικό ώστε οι ρυθμιστικές αρχές και οι εταιρείες έχουν αρχίσει να το χρησιμοποιούν σε μεγάλο βαθμό. Αυτό αποτελεί ένδειξη των τεράστιων δυνατοτήτων του.

1.4 Λειτουργία της Πλατφόρμας του Airbnb

Η εταιρία Airbnb δεν κατέχει τα καταλύματα αλλά απλά χρεώνει τόσο τους οικοδεσπότες όσο και τους επισκέπτες για τη χρήση της ιστοσελίδας . Οι ταξιδιώτες μπορούν να κάνουν ερωτήσεις ή κρατήσεις απευθείας με τους ιδιοκτήτες ακινήτων. Η εμπειρία του χρήστη είναι προσαρμοσμένη σύμφωνα με τις ανάγκες τόσο των οικοδεσποτών όσο και των φιλοξενουμένων της. Οι οικοδεσπότες ενθαρρύνονται να παρέχουν περισσότερες πληροφορίες στο δικό τους προφίλ Airbnb για να ενισχύσουν την αξιοπιστία τους. Οι κριτικές χρησιμοποιούνται επίσης για την ενίσχυση της φήμης των οικοδεσποτών. Με αυτόν τον τρόπο το Airbnb προσπαθεί να γίνει μια πλατφόρμα

εμπιστοσύνης με την οποία οι άνθρωποι αισθάνονται ασφαλείς ώστε να εκτελούν συναλλαγές.

Η Airbnb παίζει το ρόλο του κυβερνο-πράκτορα, συνδέοντας αγοραστές και πωλητές ομοειδών και λαμβάνει μια ονομαστική αμοιβή πρακτορείου όταν επιτυγχάνεται μια επιτυχημένη συμφωνία μεταξύ ενός οικοδεσπότη και ενός επισκέπτη. Έχει δημιουργήσει μια κοινότητα παρόμοιων ταξιδιωτών, δηλ.
οικοδεσπότες που επιθυμούν να μοιραστούν τις δικές τους ταξιδιωτικές εμπειρίες με άλλους ταξιδιώτες. Για να γίνει αυτό, η Airbnb απαιτεί από τους ταξιδιώτες να δημιουργήσουν τα προφίλ τους στην ιστοσελίδα τους πριν μπορέσουν να προβούν σε κάποιες κρατήσεις. Το προφίλ του ταξιδιώτη είναι προσβάσιμο από τον επιλεγέντα οικοδεσπότη του ώστε να μπορεί να επιλέξει αν θα δεχθεί την κράτηση ή όχι.

Όταν λοιπόν ο οικοδεσπότης αποδεχθεί το κλείσιμο του καταλύματος ο κεντρικός υπολογιστής δέχεται την συναλλαγή και οι φιλοξενούμενοι εκτός του κόστους του καταλύματος έχουν την υποχρέωση να πληρώσουν και το τέλος συναλλαγής που ανέρχεται στο 6%-12% ανάλογα με την περιοχή. Αντίστοιχα για τους οικοδεσπότες το Airbnb χρεώνει ένα μικρότερο ποσό για την συναλλαγή. Μετά την κράτηση της λίστας, ο κεντρικός υπολογιστής λαμβάνει την πληρωμή και το Airbnb εισπράττει τέλος συναλλαγής 3%.

Οι Superhost είναι έμπειροι οικοδεσπότες που αποτελούν το καλό παράδειγμα για άλλους οικοδεσπότες και προσφέρουν εξαιρετικές εμπειρίες στους επισκέπτες τους. Όταν ένας οικοδεσπότης γίνει Superhost, θα εμφανιστεί αυτόματα ένα σήμα στην καταχώρηση και το προφίλ του έτσι ώστε να είναι εμφανής η διαφοροποίηση του σε σχέση με τους υπόλοιπους οικοδεσπότες. Το Airbnb ελέγχει την δραστηριότητα των Superhost τέσσερις φορές το χρόνο έτσι ώστε να διασφαλίζει το υψηλό επίπεδο της υπηρεσίας (Airbnb, 2017).

Για να γίνει ένας απλός οικοδεσπότης superhost θα πρέπει :

1. Να έχει φιλοξενήσει επισκέπτες τουλάχιστον 10 φορές
2. Να διατηρεί ρυθμό απάντησης σε τυχόν ερωτήματα τουλάχιστον 90%
3. Να έχει λάβει κριτικές 5 αστεριών τουλάχιστον στο 80% των κριτικών εφόσον τουλάχιστον 50% των επισκεπτών άφησαν κριτική
4. Τέλος, να έχει ολοκληρώσει τις κρατήσεις χωρίς ακυρώσεις.

Η πλατφόρμα του Airbnb διαθέτει 3 διαφορετικούς βασικούς τύπους καταλυμάτων :

1. Ολόκληρα διαμερίσματα: Ένα ολόκληρο σπίτι, διαμέρισμα . Ο οικοδεσπότης δεν διαμένει στο σπίτι κατά τη διάρκεια της διαμονής του επισκέπτη.
2. Ιδιωτικοί χώροι: Ένας χώρος στο σπίτι του οικοδεσπότη με συγκεκριμένη ιδιωτικότητα. Σε αυτήν την περίπτωση η μίσθωση είναι μικρής διάρκειας.
3. Κοινόχρηστο δωμάτιο: Οι επισκέπτες και οι φιλοξενούμενοι καταλαμβάνουν

τον ίδιο χώρο διαβίωσης με μειωμένη ιδιωτικότητα, που αποτελεί και το αρχικό μοντέλο της εταιρίας.

Τιμολόγηση

Η τιμολόγηση στο Airbnb διεξάγεται εξ ολοκλήρου από τους οικοδεσπότες. Οι επιλογές πληρωμής διαφέρουν, αλλά περιλαμβάνουν σημαντικές πιστωτικές, χρεωστικές και προπληρωμένες κάρτες, PayPal και Πορτοφόλι Google. (WU, et al., 2012)

Τα έξοδα για τους επισκέπτες είναι ως εξής:

Προκαταβολή εγγύησης: Οι επισκέπτες θα πρέπει να καταβάλουν εγγύηση ασφαλείας, η οποία καθορίζεται από τον οικοδεσπότη. Αυτό βοηθά στην κάλυψη συμβάντων που προκύπτουν κατά τη διάρκεια μιας διαμονής, όπως ένα σπασμένο στοιχείο ή ένα κλειδί που δεν έχει επιστρέψει.
Χρέωση υπηρεσιών: Η Airbnb χρεώνει τους επισκέπτες με χρέωση 6% - 12% κάθε φορά που κάνουν κράτηση για μια υπηρεσία.
Τέλος καθαρισμού: Οι οικοδεσπότες έχουν τη δυνατότητα να χρεώσουν αυτό το ποσό για να καθαρίσουν το χώρο τους για έναν επισκέπτη.
Αντίθετα ο οικοδεσπότης έχει τις ακόλουθες κρατήσεις:
Χρέωση υπηρεσιών : Η πλατφόρμα του Airbnb χρεώνει ένα τέλος 3% για υπηρεσίες που παρέχονται στην περιοχή του (WU, et al., 2012). Οι οικοδεσπότες μπορούν να λάβουν την τελική πληρωμή τους με διάφορους τρόπους: ACH / άμεση κατάθεση, τραπεζική μεταφορά ή διεθνή καλώδιο, PayPal, Western Union, Payoneer ή ταχυδρομική επιταγή.

2 Θεωρητικό πλαίσιο

2.1 Μηχανική μάθηση

Ένα από τα κύρια χαρακτηριστικά του ανθρώπου είναι η ικανότητα μάθησης και η βελτίωση της ικανότητας αυτής. Τις τελευταίες δεκαετίες επιστήμονες προσπαθούν να μεταφέρουν την ικανότητα στα υπολογιστικά συστήματα έτσι ώστε να καταστήσουν δυνατή την επίλυση προβλημάτων που είναι πρακτικά αδύνατο να επιλυθούν από τον ίδιο τον άνθρωπο.
Ορισμός
Ο οργανισμός (Society, 2017) αναφέρει ότι η μηχανική μάθηση είναι η τεχνολογία που επιτρέπει στους υπολογιστές να μάθουν απευθείας από παραδείγματα και εμπειρία με την μορφή δεδομένων. Συγκεκριμένα όταν δοθεί ένα πρόβλημα σε ένα σύστημα που χρησιμοποιεί μηχανική μάθηση γίνεται η χρήση μεγάλου όγκου δεδομένων ως παραδείγματα για να πραγματοποιηθεί η εύρεση συγκεκριμένων μοτίβων (Society, 2017).
Ιστορική αναδρομή
1950 - Ο Alan Turing δημιουργεί το "Test Turing" για να διαπιστώσει εάν ένας υπολογιστής έχει πραγματική νοημοσύνη. Για να περάσει η δοκιμή, ένας υπολογιστής πρέπει να είναι σε θέση να ξεγελάσει ένα άτομο να πιστέψει ότι είναι επίσης άνθρωπος.
1952 - Ο Arthur Samuel έγραψε το πρώτο πρόγραμμα εκμάθησης ηλεκτρονικών υπολογιστών. Το πρόγραμμα ήταν το παιχνίδι των πούρων της IBM
1957 - Ο Frank Rosenblatt σχεδίασε το πρώτο νευρωνικό δίκτυο για υπολογιστές (το perceptron), το οποίο προσομοιώνει τις διαδικασίες σκέψης του ανθρώπινου εγκεφάλου.
1967 - Ο αλγόριθμος "πλησιέστερου γείτονα" γράφτηκε, επιτρέποντας στους υπολογιστές να αρχίσουν να χρησιμοποιούν πολύ βασική αναγνώριση προτύπων. Αυτό θα μπορούσε να χρησιμοποιηθεί για τη χαρτογράφηση μιας διαδρομής για τους

μετακινούμενους πωλητές, ξεκινώντας από μια τυχαία πόλη, αλλά εξασφαλίζοντας ότι επισκέπτονται όλες τις πόλεις κατά τη διάρκεια μιας σύντομης περιήγησης.
1979 - Οι σπουδαστές στο Πανεπιστήμιο του Στάνφορντ επινοούν το "Stanford Cart" που μπορεί να πλοηγηθεί στα εμπόδια σε ένα δωμάτιο από μόνο του.
1981 - Ο Gerald Dejong εισάγει την έννοια της Εκμάθησης Βασισμένης Μάθησης (EBL), στην οποία ένας υπολογιστής αναλύει τα δεδομένα εκπαίδευσης και δημιουργεί έναν γενικό κανόνα που μπορεί να ακολουθήσει η απόρριψη ασήμαντων δεδομένων.
1985 - Ο Terry Sejnowski ανακαλύπτει το NetTalk, το οποίο διδάσκει να προφέρει λέξεις όπως και το μωρό.
Η δεκαετία του 1990 - Οι εργασίες για τη μηχανική μάθηση μετατοπίζονται από μια προσέγγιση που βασίζεται στη γνώση σε μια προσέγγιση που βασίζεται σε δεδομένα. Οι επιστήμονες αρχίζουν να δημιουργούν προγράμματα για υπολογιστές για να αναλύουν μεγάλα ποσά δεδομένων και να αντλούν συμπεράσματα - ή να "μαθαίνουν" - από τα αποτελέσματα.
1997 - Ο Deep Blue της IBM κέρδισε τον παγκόσμιο πρωταθλητή στο σκάκι.

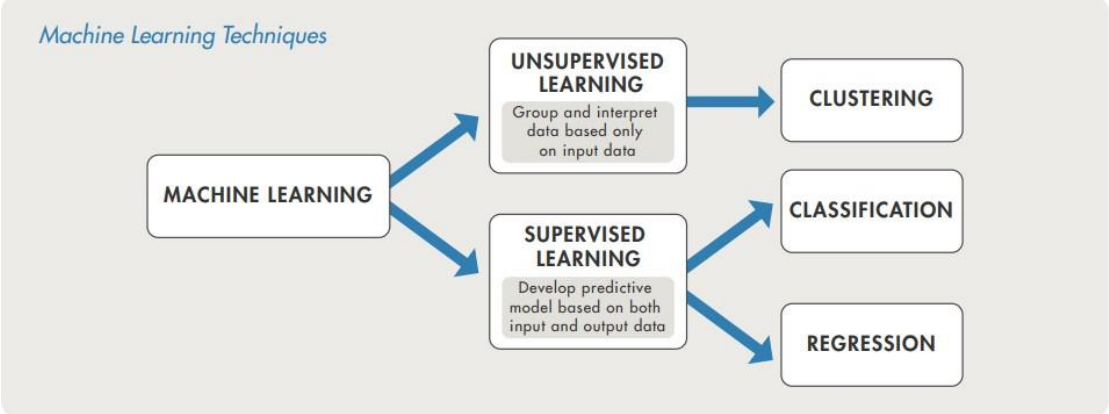
2006 - Ο Geoffrey Hinton νομίζει τον όρο "βαθιά εκμάθηση" για να εξηγήσει νέους αλγόριθμους που επιτρέπουν στους υπολογιστές να "βλέπουν" και να διακρίνουν αντικείμενα και κείμενο σε εικόνες και βίντεο.
2012 - Το X Lab της Google αναπτύσσει έναν αλγόριθμο εκμάθησης μηχανών που είναι σε θέση να περιάγει αυτόνομα βίντεο στο YouTube για να εντοπίσει τα βίντεο που περιέχουν γάτες.
2014 - Το Facebook FB -0.02% αναπτύσσει το DeepFace, έναν αλγόριθμο λογισμικού που είναι σε θέση να αναγνωρίσει ή να επαληθεύσει τα άτομα στις φωτογραφίες στο ίδιο επίπεδο με τον άνθρωπο.
2015 - Η Amazon εγκαινιάζει τη δική της πλατφόρμα εκμάθησης μηχανών.
2016 - Η Microsoft δημιουργεί το Toolkit για την εκμάθηση των υπολογιστών, το οποίο επιτρέπει την αποτελεσματική κατανομή των προβλημάτων μηχανικής μάθησης σε πολλούς υπολογιστές (Marr, 2016).
Πότε χρησιμοποιείται η μηχανική μάθηση και ποιες είναι οι εφαρμογές της
Η μηχανική μάθηση όπως επισημαίνει ο (Shalev-Shwartz & Shai , 2014) χρησιμοποιείται για εργασίες που μπορεί να εκτελεί ο άνθρωπος ή ακόμα και τα ζώα αλλά είναι αρκετά δύσκολο να τις προγραμματίσουμε. Παραδείγματα τέτοιων εργασιών είναι η αναγνώριση φωνής, η οδήγηση κτλ. Σε αυτόν τον τομέα των προβλημάτων ο υπολογιστής έχει την δυνατότητα να επιτύχει ικανοποιητικά αποτελέσματα αρκεί φυσικά να εκτεθεί σε μεγάλο αριθμό παραδειγμάτων ώστε να εξάγει ασφαλή συμπεράσματα.
Αντίθετα, υπάρχουν και προβλήματα τα οποία είναι αδύνατον να εκτελεστούν από τον άνθρωπο όπου σε αυτήν την περίπτωση με την βοήθεια του Data Mining μπορούν να δημιουργηθούν μεγάλες αποθήκες δεδομένων τις οποίες μπορεί να επεξεργαστεί ένα υπολογιστικό σύστημα.
Η μηχανική μάθηση έχει διεισδύσει σε μεγάλο βαθμό στην καθημερινότητα μας και αυτό αποδεικνύεται από τις πολλαπλές εφαρμογές που έχει καταφέρει να αναπτύξει στο εμπόριο, την υγεία την οικονομία κτλ.

Μια από τις σημαντικότερες εφαρμογές στην οικονομία είναι η αναγνώριση προτύπων στα συστήματα εύρεσης ψευδών πιστωτικών καρτών. Χρησιμοποιώντας τα δεδομένα συναλλαγών από μεγάλο αριθμό χρηστών, οι αλγόριθμοι εκπαιδεύονται για να αναγνωρίζουν τα πρότυπα δαπανών. Στη συνέχεια, εάν ένας χρήστης εμφανίζει ένα ασυνήθιστο πρότυπο δαπανών, το σύστημα μπορεί να ειδοποιήσει τον κάτοχο της κάρτας και να μπλοκαριστεί κάθε συναλλαγή.
Η Μηχανική Μάθηση μπορεί να υποστηρίξει προηγμένα συστήματα αναγνώρισης εικόνας. Στις εφαρμογές κοινωνικών μέσων, η αναγνώριση εικόνας μπορεί να χρησιμοποιηθεί για την επισήμανση αντικειμένων ή ατόμων σε φωτογραφίες που έχουν μεταφορτωθεί σε έναν ιστότοπο . Παρόμοια συστήματα αναγνώρισης εικόνων μπορούν επίσης να χρησιμοποιηθούν για την αναγνώριση σαρωμένου χειρόγραφου υλικού, για παράδειγμα για την αναγνώριση των διευθύνσεων στα γράμματα ή στα ψηφία των επιταγών.
Άλλες πιθανές εφαρμογές αποτελούν οι :

- ιατρική διάγνωση: διάγνωση ενός ασθενούς ως πάσχοντος ή μη πάσχοντος από κάποια ασθένεια
- κατάτμηση πελατών: Πρόβλεψη για το ποιοι πελάτες θα ανταποκριθούν σε ένα συγκεκριμένο προϊόν (Schapire, 2008).
- πρόγνωση καιρού: πρόβλεψη , για παράδειγμα, εάν θα βρέξει αύριο ή όχι (Society, 2017).

2.2 Μάθηση υπό επίβλεψη και χωρίς επίβλεψη

Η Μηχανική Μάθηση χωρίζεται σε 2 μεγάλες υποκατηγορίες : Την Μάθηση Υπό Επίβλεψη και την Μάθηση χωρίς Επίβλεψη. Στην Μάθηση Υπό Επίβλεψη , η εστίαση είναι στην ακριβή πρόβλεψη, ενώ στην Μάθηση Χωρίς Επίβλεψη στόχος είναι να βρεθούν ακριβείς συμπαγείς περιγραφές των δεδομένων .Ιδιαίτερα στην Μάθηση Υπό Επίβλεψη, κάποιος ενδιαφέρεται για μεθόδους που έχουν καλές επιδόσεις σε δεδομένα που δεν έχουν επεξεργαστεί πιο πριν. Γι' αυτόν τον λόγο τα δεδομένα χωρίζονται σε 2 σύνολα όπου το ένα χρησιμοποιείται για να εκπαιδεύσει τον αλγόριθμο ενώ το άλλο για δοκιμές (Mathworks, 2016).



2.1.Διαχωρισμός τεχνικών μηχανικής μάθησης(Mathworks, 2016. Introducing Machine Learning)

Ορισμός

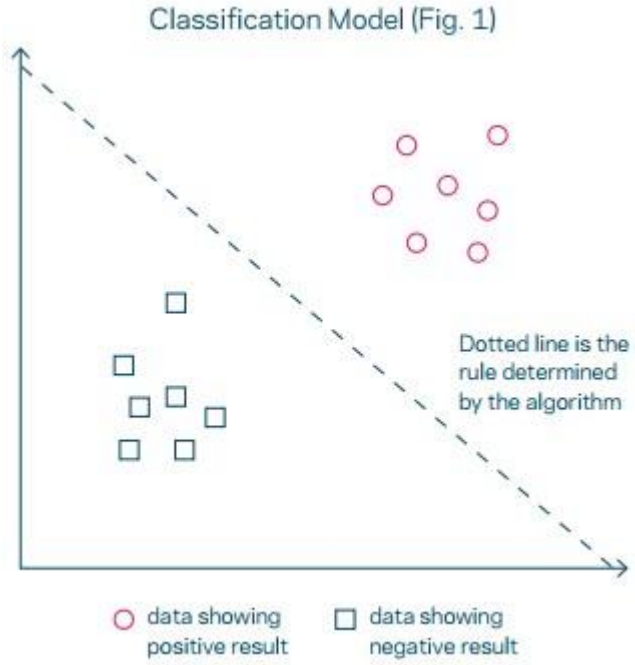
Έχοντας ένα σύνολο δεδομένων D = {(x_n ,y_n), n = 1, . . . , N} στόχος είναι να βρεθεί η σχέση μεταξύ του χ και του y έτσι ώστε όταν δοθεί ένα νέο x* να μπορούμε να υπολογίσουμε το y* με μια δεδομένη ακρίβεια (James, et al., 2017).

Στην συγκεκριμένη εργασία χρησιμοποιήθηκαν τόσο τεχνικές ταξινόμησης όσο και παλινδρόμησης λόγω της ποικιλομορφίας των προβλημάτων που είχαμε να αντιμετωπίσουμε.

2.2.1 Ταξινόμηση

Οι τεχνικές ταξινόμησης δίνουν διακριτές απαντήσεις .Για παράδειγμα, αν ένα μήνυμα ηλεκτρονικού ταχυδρομείου είναι αυθεντικό ή ανεπιθύμητο, ή αν ένας όγκος είναι καρκινικός ή καλοήθης. Τα μοντέλα ταξινόμησης ταξινομούν τα δεδομένα εισόδου σε κατηγορίες. Τυπικές εφαρμογές περιλαμβάνουν την ιατρική απεικόνιση, την αναγνώριση ομιλίας κτλ.

Ο σκοπός του μοντέλου ταξινόμησης είναι να προσδιοριστούν κατηγορίες δεδομένων με διαφορετικά χαρακτηριστικά . Εκπαιδεύουμε το μοντέλο χρησιμοποιώντας μια σειρά από ομάδες δεδομένων. Κάθε φορά που εκπαιδεύουμε τον αλγόριθμο με πρόσθετα δεδομένα, η ακρίβεια της πρόβλεψης βελτιώνεται αισθητά παρέχοντας βέλτιστα αποτελέσματα (Health, 2016).

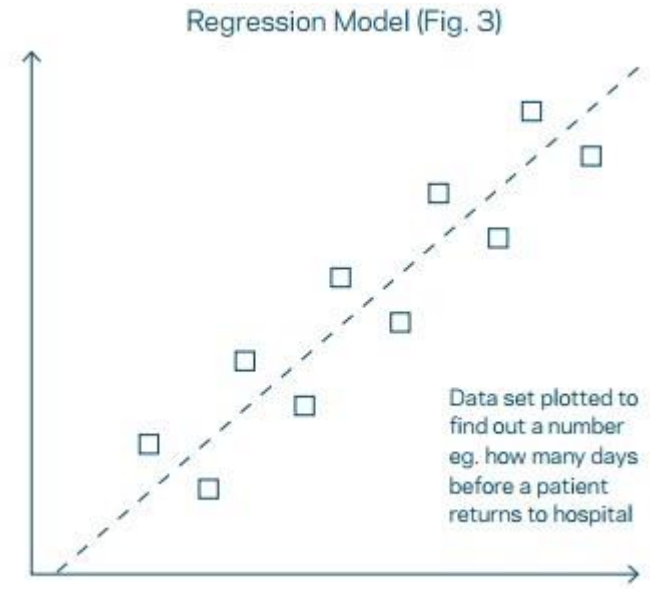


2.2.Γραφική αναπαράσταση του μοντέλου ταξινόμησης(Health, O., 2016. Introduction to Machine Learning in Healthcare)

2.2.2 Παλινδρόμηση

Αν για ένα x, το αποτέλεσμα y είναι μια συνεχής μεταβλητή, αυτό ονομάζεται πρόβλημα παλινδρόμησης.

Οι τεχνικές παλινδρόμησης προβλέπουν συνεχείς απαντήσεις -για παράδειγμα, αλλαγές θερμοκρασίας ή διακυμάνσεις στην ζήτηση ισχύος. Τυπικές εφαρμογές περιλαμβάνουν την πρόβλεψη φορτίου ηλεκτρικού ρεύματος και την αλγοριθμική συναλλαγή (Health, 2016).



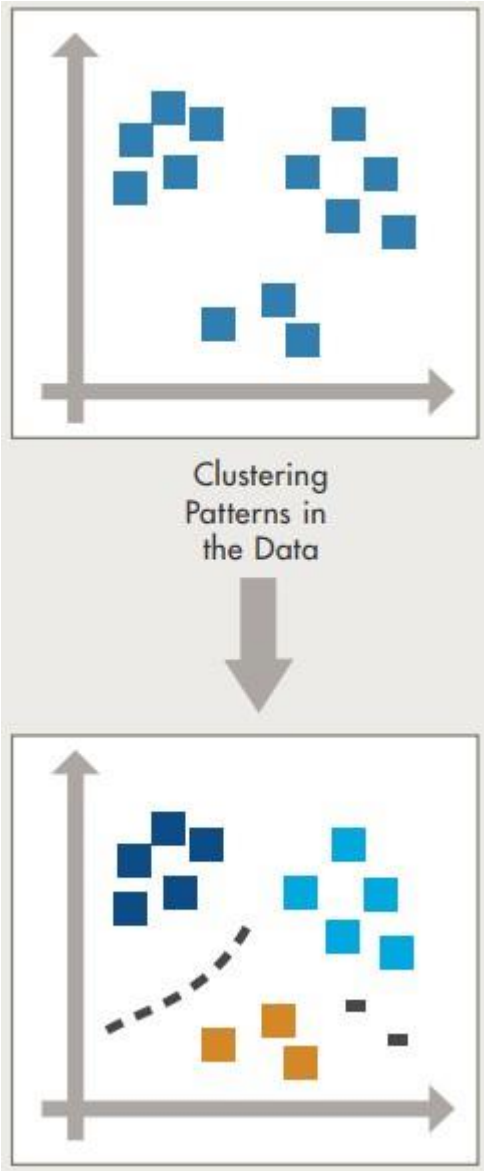
2.3.Γραφική αναπαράσταση του μοντέλου παλινδρόμησης (Health, O., 2016. Introduction to Machine Learning in Healthcare)

2.2.3 Clustering

Η ομαδοποίηση αναφέρεται σε ένα πολύ ευρύ σύνολο τεχνικών για την εύρεση υποομάδων ή συστοιχιών συστάδων σε ένα σύνολο δεδομένων. Όταν συγκεντρώνουμε τις παρατηρήσεις ενός συνόλου δεδομένων, επιδιώκουμε να τις χωρίσουμε σε ξεχωριστές ομάδες έτσι ώστε οι παρατηρήσεις μέσα σε κάθε ομάδα να είναι αρκετά παρόμοιες μεταξύ τους, ενώ οι παρατηρήσεις σε διαφορετικές ομάδες είναι αρκετά διαφορετικές μεταξύ τους (Health, 2016).

Από μια πιθανοτική προοπτική ενδιαφερόμαστε να διαμορφώσουμε τη κατανομή p (x).

Οι εφαρμογές του clustering περιλαμβάνουν ανάλυση γονιδιακής ακολουθίας, έρευνα αγοράς και αναγνώριση αντικειμένων.



2.4.Γραφική απεικόνιση του μοντέλου ομαδοποίησης(Health, O., 2016. Introduction to Machine Learning in Healthcare)

2.3 Εξόρυξη Γνώμης

Η εξόρυξη γνώμης είναι μια τεχνική ανίχνευσης και εξαγωγής υποκειμενικών πληροφοριών σε έγγραφα κειμένου. Σε γενικές γραμμές, η ανάλυση συναισθημάτων προσπαθεί να προσδιορίσει το συναίσθημα ενός συγγραφέα σχετικά με κάποια πτυχή ή τη συνολική συμφραζόμενη πολικότητα ενός εγγράφου. Το συναίσθημα μπορεί να είναι η κρίση, η διάθεση ή η αξιολόγησή του. Ένα βασικό πρόβλημα σε αυτόν τον τομέα είναι η ταξινόμηση των συναισθημάτων ανάλογα με το είδος του προϊόντος που χρειάζεται να αναλυθεί (ταινία, βιβλίο, προϊόν κ.λπ.) (Padmaja & S Sameen , 2013)
Η Συναισθηματική Ανάλυση (Sentiment Analysis), ή Σημασιολογικός Προσανατολισμός (Semantic Orientation), είναι το μέτρο υποκειμενικότητας και γνώμης ενός κειμένου. Συνήθως αναπαρίσταται με τη χρήση ενός παράγοντα αξιολόγησης (π.χ. θετικό ή αρνητικό), συνοδευόμενο από έναν δείκτη ισχύος (ο βαθμός

στον οποίο ένα κείμενο είναι θετικό ή αρνητικό) που αφορά σε ένα ζήτημα, πρόσωπο ή ιδέα (Michailidis, 2017).
Αντικείμενο (Object): Ένα αντικείμενο είναι μια οντότητα που μπορεί να είναι θέμα, προϊόν, γεγονός, άτομο ή οργάνωση. Συνδέεται με το ζεύγος O: (T, A), όπου T είναι μια ιεραρχία των στοιχείων και των υποστοιχείων του αντικειμένου O. A είναι ένα σύνολο χαρακτηριστικών του αντικειμένου O. Κάθε στοιχείο έχει τα δικά του υποστοιχεία και ένα σύνολο χαρακτηριστικών. Ωστόσο, απλά, συχνά χρησιμοποιούμε τον όρο 'χαρακτηριστικό' για να αναπαραστήσουμε τα στοιχεία και τα χαρακτηριστικά τους. Ένα αντικείμενο είναι επίσης ένα χαρακτηριστικό. Μπορούμε να ορίσουμε ένα έγγραφο d, το οποίο μπορεί να είναι μια κριτική ταινίας, ένα blog, ένα μήνυμα στο φόρουμ που αξιολογεί ορισμένα αντικείμενα. Ένα έγγραφο d αποτελείται από ορισμένες προτάσεις, έτσι ώστε d = {s1, s2, s3, s4...}.
Κάτοχος Γνώμης (Opinion Holder): Πρόκειται για ένα άτομο ή έναν οργανισμός που δημοσιεύει τη γνώμη σχετικά με ένα αντικείμενο. Για παράδειγμα, δημιουργός ενός μηνύματος σε φόρουμ, σε blog κ.α.
Σημασιολογικός Προσανατολισμός της Γνώμης: Ο σημασιολογικός προσανατολισμός της γνώμης σχετικά με μια άποψη σημαίνει ότι μπορεί να είναι θετική, αρνητική ή ουδέτερη (Thivaios, 2017).

2.3.1 Εφαρμογές

Ανάλυση συναισθημάτων στις χρηματοπιστωτικές αγορές. Για τους επενδυτές είναι σημαντική η πληροφόρηση των αναλυτών και των άλλων επενδυτών σχετικά με τις μετοχές μιας εταιρείας, για τον εντοπισμό των τάσεων των τιμών.
Ανάλυση συναισθημάτων στα προϊόντα. Μια εταιρεία ενδιαφέρεται για τις αντιλήψεις των πελατών για τα προϊόντα της. Οι πληροφορίες μπορούν να χρησιμοποιηθούν για τη βελτίωση των προϊόντων και τον εντοπισμό νέων στρατηγικών μάρκετινγκ (Smeureaunu & Bucur, 2012).
Τα σχόλια των χρηστών είναι ένα σημαντικό πρόβλημα. Θα μπορούσαμε επίσης να φανταστούμε ότι θα μπορούσαν να διορθωθούν τα σφάλματα στις αξιολογήσεις των χρηστών: υπάρχουν περιπτώσεις όπου οι χρήστες έχουν σαφώς επιλέξει κατά λάθος χαμηλή βαθμολογία όταν η αναθεώρησή τους υποδεικνύει μια θετική αξιολόγηση. Επιπλέον, υπάρχουν κάποιες ενδείξεις ότι οι αξιολογήσεις χρηστών μπορεί να είναι προκατειλημμένες ή αλλιώς να χρειάζονται διόρθωση και αυτοματοποιημένοι ταξινομητές θα μπορούσαν να παρέχουν αυτές τις ενημερώσεις (Pang & Lee, 2008).

2.3.2 Προβλήματα της εξόρυξης γνώμης

Υπάρχουν πολλές προκλήσεις στην εξόρυξη γνώμης. Το πρώτο είναι ότι μια λέξη που θεωρείται θετική σε μια κατάσταση μπορεί να θεωρηθεί αρνητική σε μια άλλη κατάσταση. Εάν ένας πελάτης είπε ότι η διάρκεια ζωής της μπαταρίας του φορητού υπολογιστή ήταν μεγάλη, αυτό θα ήταν μια θετική γνώμη. Εάν ο πελάτης είπε ότι ο χρόνος εκκίνησης του φορητού υπολογιστή ήταν μεγάλος, ωστόσο, αυτό θα ήταν μια αρνητική γνώμη. Αυτές οι διαφορές σημαίνουν ότι ένα σύστημα γνώσης που έχει εκπαιδευτεί για τη συγκέντρωση απόψεων σχετικά με ένα είδος προϊόντος ή χαρακτηριστικού προϊόντος μπορεί να μην λειτουργεί πολύ καλά σε ένα άλλο.

Μια δεύτερη πρόκληση είναι ότι οι άνθρωποι δεν εκφράζουν πάντοτε τις απόψεις τους με τον ίδιο τρόπο. Η πιο παραδοσιακή επεξεργασία κειμένου βασίζεται στο γεγονός ότι οι μικρές διαφορές ανάμεσα σε δύο κομμάτια κειμένου δεν αλλάζουν πολύ το νόημα. Στην άποψη της εξόρυξης, ωστόσο, "η ταινία ήταν μεγάλη" είναι πολύ διαφορετική από "η ταινία δεν ήταν μεγάλη" (TechTarget, 2010).
Πολυγλωσσία: Η ανάπτυξη αλγορίθμων σε λέξεις εκτός της αγγλικής περιλαμβάνει τη δόμηση λεξικών και σωμάτων εκπαίδευσης (training corpora) σε αυτές τις γλώσσες, μία διαδικασία ταυτόχρονα δημιουργική και δαπανηρή (Dempelis, 2014).

2.4 Προηγούμενες Σχετικές Έρευνες

Η πλατφόρμα βραχυχρόνιας ενοικίασης Airbnb έχει απασχολήσει στο παρελθόν αρκετούς ερευνητές οι οποίοι ανέλυσαν τα χαρακτηριστικά των καταλυμάτων και κατάφεραν να ανακαλύψουν πως επηρεάζουν τις τιμές των καταλυμάτων και όχι μόνο.

Το 2015 οι Tang και Sangani (Tang & Sangani, 2015) αποφάσισαν να αναλύσουν τα δεδομένα για τα καταλύματα από την πόλη του San Francisco τα οποία έλαβαν από το project InsideAirbnb. Στόχος τους ήταν να προβλέψουν την κατάλληλη τιμή ενός καταλύματος λαμβάνοντας υπόψιν ιδιαίτερα χαρακτηριστικά στην αγορά της βραχυχρόνιας μίσθωσης όπως και να προβλέψουν την περιοχή στην οποία βρισκόταν ένα κατάλυμα δίνοντας μια εικόνα για τα πολιτιστικά στοιχεία που είναι ορατά μέσω του κειμένου, της εικόνας και των ανέσεων. Με αυτόν τον τρόπο η πλατφόρμα του Airbnb θα μπορούσε να προτείνει στον χρήστη κάποια ανάλογη περιοχή για μελλοντική επίσκεψη. Με την χρήση μεθόδων μηχανικής μάθησης (Support Vector Machine) κατέληξαν ότι μπορούμε να προβλέψουμε επιτυχώς την περιοχή και τις τιμές των καταλυμάτων χρησιμοποιώντας μια σειρά χαρακτηριστικών που εξάγονται από τις καταχωρίσεις. Και οι δύο ταξινομητές που ανέπτυξαν είχαν απόδοση καλύτερη από όσο περίμεναν και δείχνουν ότι η καταγραφή πληροφοριών, χαρακτηριστικών κειμένου και άλλων μπορεί να αξιοποιηθεί αποτελεσματικά για την πρόβλεψη της τοποθεσίας και τιμής.

Το 2017 στόχος των Wang και Nicolau ήταν να προσδιορίσουν τους καθοριστικούς παράγοντες της τιμής των καταλυμάτων σε πλατφόρμες που ανήκουν στην ψηφιακή αγορά. Πάρθηκε ένα δείγμα 180.533 ενοικιαζόμενων διαμερισμάτων σε 33 πόλεις που είναι εισηγμένες στην πλατφόρμα Airbnb και χρησιμοποιήθηκε η μέθοδος της απλής παλινδρόμησης. Είκοσι πέντε επεξηγηματικές μεταβλητές σε πέντε κατηγορίες (χαρακτηριστικά υποδοχής, ιδιότητες τοποθεσίας και ιδιοκτησίας, παροχές και υπηρεσίες, κανόνες ενοικίασης και αξιολογήσεις αξιολόγησης μέσω διαδικτύου) διερευνήθηκαν για να βρεθεί η σχέση μεταξύ της τιμολόγησης και των καθοριστικών παραγόντων. Η μέθοδος ελάχιστων τετραγώνων αποκαλύπτει ότι 24 από τις 25 μεταβλητές που μελετώνται ήταν καλοί προγνωστικοί παράγοντες της τιμής. Έτσι, τα

ευρήματα προσφέρουν πληροφορίες σχετικά με την πολυπλοκότητα της σχέσης τιμής-προσδιοριστικών παραγόντων των ενοικιαζόμενων καταλυμάτων .

Το 2018 οι (Singh, Choudhury, Banerjee, & Manniste, 2018) χρησιμοποίησαν δεδομένα από την ιστοσελίδα Inside Airbnb από το 2008 έως το 2018 για τα καταλύματα στην πόλη της Νέας Υόρκης. Η μελέτη χρησιμοποίησε τις καταχωρίσεις από τη Νέα Υόρκη για να προβλέψει την πιθανότητα να γίνει κάποιος οικοδεσπότης superhost βασιζόμενη στην χρήση κριτικών. Για την επίτευξη αυτού του στόχου δημιουργήθηκαν διάφορα στατιστικά μοντέλα όπως η λογική παλινδρόμηση, το SVM και τα δέντρα αποφάσεων. Το δέντρο απόφασης ήταν το καλύτερο μοντέλο με ποσοστό εσφαλμένης ταξινόμησης 12% και ευαισθησία 52%. Ένα άλλο σύνολο γραμμικών και μη γραμμικών μοντέλων παλινδρόμησης δημιουργήθηκε για να προβλέψει την τιμή εισαγωγής όλων των καταλυμάτων Airbnb στη Νέα Υόρκη και σε αυτή την περίπτωση, το βηματικό μοντέλο παλινδρόμησης υπερέβη τις άλλες εναλλακτικές με τιμή R Squared 55%.

Το 2018 οι (Choudhary, Jain, & Baijal, 2018) συνέλεξαν δεδομένα από την πλατφόρμα του airbnb χρονικά από το 2015 έως και το 2017 και επικεντρώθηκαν στις πόλεις Νέα Υόρκη και Σαν Φρανσίσκο. Ο στόχος τους ήταν διπλός. Αρχικά, ήθελαν να διαπιστώσουν αν μπορούν να αναλύσουν παρελθοντικές καταχωρίσεις για να προτείνουν τη βέλτιστη χρέωση που θα έδινε οικοδεσπότης για τη νέα καταχώριση. Επίσης, είναι αβέβαιο το πότε ένας χώρος έχει την δυνατότητα ενοικίασης ή όταν δεν είναι διαθέσιμος. Επομένως, στόχος τους ήταν οι επισκέπτες να μπορούν να αποφασίσουν εάν θα μετακινηθούν ή θα περιμένουν να είναι διαθέσιμο ένα κατάλυμα ανάλογα με κάποια πιθανότητα. Αρχικά, χρησιμοποίησαν την μέθοδο απλής παλινδρόμησης για να διαπιστώσουν ποιοι παράγοντες ήταν σημαντικοί. Για να βρουν την πιθανότητα ένα κατάλυμα να είναι διαθέσιμο διαχώρισαν τα καταλύματα σε clusters υψηλής και χαμηλής διαθεσιμότητας και χρησιμοποιώντας τον τύπο του καταλύματος αλλά και τον ταχυδρομικό κώδικα δημιούργησαν ένα μοντέλο Bayes.

Το 2019 οι (Cai, Han, & Wu, 2019) είχαν ως στόχο της μελέτης τους να δημιουργήσουν ένα μοντέλο πρόβλεψης τιμών για τα καταλύματα Airbnb στη Μελβούρνη. Χρησιμοποίησαν διάφορες μεθόδους παλινδρόμησης και συνέκριναν τις επιδόσεις τους. Το έργο τους βασίστηκε στο αρχικό πρόβλημα της παλινδρόμησης τιμών, χωρίς να μετατραπεί σε πρόβλημα ταξινόμησης. Εκτός από τις παραδοσιακές μεθόδους μηχανικής μάθησης, ενσωμάτωσαν και δεδομένα κειμένου (από περιγραφές και κριτικές) στο μοντέλο τους. Συγκεκριμένα χρησιμοποίησαν παραδοσιακές μεθόδους μηχανικής μάθησης όπως γραμμική παλινδρόμηση, Support Vector Machine , Παλινδρόμηση τυχαίων δέντρων και τέσσερα νευρωνικά δίκτυα για την παραγωγή των προβλεπόμενων τιμών των καταχωρίσεων. Διαπίστωσαν ότι η μέθοδος Gradient Boosting είχε την καλύτερη απόδοση ενώ δεύτερη σε απόδοση ήρθε η μέθοδος τυχαίων δασών όπου αν υπήρχε μια πιο αυστηρή επιλογή χαρακτηριστικών θα είχε ως αποτέλεσμα την βελτίωση της απόδοσης της συγκεκριμένης μεθόδου.

Το 2019 οι (Kalehbasti, Nikolenko, & Rezaei, 2019) έχοντας στην κατοχή τους ένα σύνολο δεδομένων με 96 χαρακτηριστικά από την πόλη της Νέας Υόρκης για τα καταλύματα της πλατφόρμας του Airbnb συνέταξαν ένα άρθρο το οποίο στόχευε στην ανάπτυξη ενός αξιόπιστου μοντέλου πρόβλεψης των τιμών με τη χρήση της μηχανικής μάθησης και της τεχνικής NLP, τόσο στους ιδιοκτήτες ακινήτων όσο και στους πελάτες

με την αξιολόγηση των τιμών, παρέχοντας ελάχιστες διαθέσιμες πληροφορίες σχετικά με το ακίνητο. Χαρακτηριστικά των ενοικιάσεων, τα χαρακτηριστικά του ιδιοκτήτη και οι αναλύσεις πελατών ήταν οι κύριοι προσδιοριστικοί παράγοντες και μια σειρά από μεθόδους όπως γραμμική παλινδρόμηση έως μοντέλα βασισμένα σε δέντρα, SVR και νευρωνικά δίκτυα) χρησιμοποιήθηκαν για τη δημιουργία του μοντέλου πρόβλεψης. Μεταξύ των μοντέλων που δοκιμάστηκαν, η μέθοδος (SVR) πραγματοποίησε το καλύτερο και παρήγαγε ένα R2 score της τάξης του 69%. Αυτό το επίπεδο ακρίβειας είναι ένα πολλά υποσχόμενο αποτέλεσμα δεδομένης της ετερογένειας του συνόλου των δεδομένων και των εμπλεκόμενων κρυφών παραγόντων, συμπεριλαμβανομένων των προσωπικών χαρακτηριστικών των ιδιοκτητών, τα οποία ήταν αδύνατο να ληφθούν υπόψη.

Το 2019 οι (Keating, Katnic, Hahn, & Yang, 2019) εξέτασαν εμπειρικά καταλύματα φιλοξενίας στην πλατφόρμα Airbnb για να δημιουργήσουν προχωρημένα μοντέλα πρόβλεψης της τιμής ενοικίασης βάσει βασικών παραγόντων και μεταβλητών. Τα δεδομένα που χρησιμοποιήσαν περιλάμβαναν ~ 4000 καταχωρήσεις καταλυμάτων στην περιοχή Seattle με 92 χαρακτηριστικά και ~ 85.000 κριτικές αυτών των καταχωρίσεων. Δημιουργήσαν διάφορα προγνωστικά μοντέλα που χρησιμοποιούν διαφορετικές μεθόδους μηχανικής μάθησης όπως τυχαία δάση και νευρωνικά δίκτυα. Εκτός αυτού , δημιούργησαν ένα μοντέλο sentimental analysis του κειμένου σε μία περιγραφή περιγραφής, βαθμολογώντας θετικά ή αρνητικά κριτικές από το 1 έως το 100. Όμως διαπίστωσαν ότι ο βαθμός αυτός δεν είναι σημαντικός για την πρόβλεψη σχέσεων τιμολόγησης των καταλυμάτων σε συνάρτηση με τις κριτικές τους. Έχοντας τρέξει και αναπτύξει καθένα από τα τρία μοντέλα, συμπεράναν ότι η μέθοδος νευρωνικών δικτύων είχε την μεγαλύτερη ακρίβεια με περιθώριο λάθους τα 32-35 δολάρια.

3 Μεθοδολογικό πλαίσιο

3.1 Μέθοδοι Παλινδρόμησης

3.1.1 Γραμμική Παλινδρόμηση

Η παλινδρόμηση είναι μια στατιστική τεχνική για τον προσδιορισμό της γραμμικής σχέσης μεταξύ δύο ή περισσοτέρων μεταβλητών. Η παλινδρόμηση χρησιμοποιείται κυρίως για την πρόβλεψη και την αιτιώδη συσχέτιση.

Στην απλούστερη (διμεταβλητή) μορφή της, η παλινδρόμηση δείχνει τη σχέση μεταξύ ανεξάρτητης μεταβλητής (X) και εξαρτημένης μεταβλητής (Y), όπως στον παρακάτω τύπο:

Εξίσωση 1

$$y = \hat{\alpha} + \hat{\beta}x$$

Το μέγεθος και η κατεύθυνση αυτής της σχέσης δίδονται από την παράμετρο κλίσης ($\hat{\beta}$) και η κατάσταση της εξαρτώμενης μεταβλητής όταν απουσιάζει η ανεξάρτητη μεταβλητή δίνεται από την παράμετρο ($\hat{\alpha}$). Ένας όρος σφάλματος (u) καταγράφει το ποσό της διακύμανσης που δεν προβλέπεται από τους όρους κλίσης και παρατήρησης. Ο συντελεστής παλινδρόμησης (R2) δείχνει πόσο καλά οι τιμές προσαρμόζονται στα δεδομένα (Seltman, 2018).

Οι συντελεστές $\hat{\alpha}$ και $\hat{\beta}$ δίνονται αντίστοιχα από τους ακόλουθους τύπους:

Εξίσωση 2

$$\hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x},$$

Εξίσωση 3

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

3.1.2 Ridge και Lasso Regression

Οι μέθοδοι παλινδρόμησης Ridge και Lasso αποτελούν ισχυρές τεχνικές οι οποίες έχουν την δυνατότητα να διαχειριστούν με μεγαλύτερη ακρίβεια και χωρίς τα προβλήματα της απλής παλινδρόμησης τα δεδομένα όταν υπάρχει πληθώρα μεταβλητών.

Τέτοια προβλήματα μπορεί να αποτελούν το overfitting που είναι συχνό φαινόμενο στην απλή γραμμική παλινδρόμηση ειδικά όταν στο πρόβλημα χρησιμοποιούνται αρκετές μεταβλητές. Επίσης, προβλήματα μεγάλα σε όγκο τα οποία προκαλούν υπολογιστικές δυσκολίες και απαιτούν βελτιστοποιημένες μεθόδους για να επιλυθούν σε λογικά χρονικά πλαίσια (Tibshirani, 2013).

Παρόλο που οι μέθοδοι Ridge και το Lasso μπορεί να φαίνεται εκ πρώτης όψεως να λειτουργούν με τον ίδιο τρόπο, οι εγγενείς ιδιότητες και οι περιπτώσεις πρακτικής χρήσης διαφέρουν. Και οι 2 λειτουργούν περιορίζοντας την επίδραση των συντελεστών των μεταβλητών ενώ ταυτόχρονα ελαχιστοποιούν το σφάλμα ανάμεσα στις πραγματικές παρατηρήσεις και τις προβλέψεις τους . Αυτές ονομάζονται τεχνικές "κανονικοποίησης". Η βασική διαφορά είναι ο τρόπος με τον οποίο περιορίζουν τους συντελεστές:

Η Ridge Παλινδρόμηση χρησιμοποιεί έναν συντελεστή L2 που προσθέτει πέναλτι ίσο με το άθροισμα των τετραγώνων των συντελεστών.

1)Όταν ο συντελεστής L2=0 τότε το πρόβλημα μετατρέπεται σε ένα πρόβλημα απλής γραμμικής παλινδρόμησης και οι συντελεστές θα παίρνουν τις ίδιες τιμές με αυτούς του απλού προβλήματος.

2)Αν ο συντελεστής L2=∞ τότε οι συντελεστές θα είναι 0

3)Όταν $0 < L2 < \infty$ τότε το μέγεθος του L2 θα καθορίσει το βάρος των συντελεστών. Όσο ο συντελεστής L2 αυξάνεται τόσο αυξάνεται και το bias ενώ όσο μειώνεται τόσο αυξάνεται η διακύμανση.

Η Lasso Παλινδρόμηση χρησιμοποιεί έναν συντελεστή L1 που προσθέτει πέναλτι ίσο με το απόλυτο άθροισμα των τετραγώνων των συντελεστών (Stephanie, 2015). Οι λύσεις Lasso είναι τετραγωνικά προβλήματα προγραμματισμού. Ο στόχος του αλγορίθμου είναι να ελαχιστοποιήσει την ακόλουθη σχέση:

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Σχέση 1

Οι λύσεις Ridge είναι τετραγωνικά προβλήματα προγραμματισμού. Ο στόχος του αλγορίθμου είναι να ελαχιστοποιήσει την ακόλουθη σχέση:

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j$$

Σχέση 2

Η μέθοδος Ridge Παλινδρόμηση έχει ως στόχο την αποφυγή του overfitting χωρίς όμως να αφαιρεί μεταβλητές από το μοντέλο. Αντίθετα η μέθοδος Lasso Παλινδρόμηση έχει την δυνατότητα να μηδενίσει τους συντελεστές των μεταβλητών και επομένως θεωρείται αρκετά χρήσιμη σε προβλήματα με τεράστιο αριθμό μεταβλητών. Ενώ παλιότερα η βηματική παλινδρόμηση ήταν ο κύριος τρόπος επιλογής χαρακτηριστικών με την εξέλιξη της μηχανικής μάθησης οι 2 συγκεκριμένες μέθοδοι απλοποιούν την διαδικασία και δίνουν ακριβέστερα αποτελέσματα (Jain, 2016).

3.1.3 Λογιστική Παλινδρόμηση

Η λογιστική παλινδρόμηση (Logistic regression) αποτελεί στην ουσία ένα μοντέλο ταξινόμησης των τιμών μιας μεταβλητής απόκρισης Υ με βάση τη θεωρία των πιθανοτήτων. Στο μοντέλο αυτό όπου η μεταβλητή Υ συνήθως έχει δυαδικό χαρακτήρα (λαμβάνει δύο τιμές) στοχεύεται η πρόβλεψη της έκβασης αυτής από ένα πλήθος προβλεπτικών μεταβλητών που μπορεί να είναι ονομαστικές, τακτικές ή ποσοτικές (Park& Hyeoun-Ae,2013.).

Διακρίνονται τρεις τύποι λογιστικής παλινδρόμησης ανάλογα με την ιδιαίτερη φύση της εξαρτημένης κατηγορικής μεταβλητής η οποία μπορεί να είναι: 1. Δίτιμη ή δυαδική ή διχοτομική (binary) ή διμερής εξαρτημένη μεταβλητή. Συνίσταται από δύο κατηγορίες, όπως π.χ. είναι οι εκβάσεις επιτυχία/αποτυχία, ΝΑΙ/ΟΧΙ, γεγονός απόν/παρόν.

2. Τακτική (ordinal) μεταβλητή. Η εξαρτημένη μεταβλητή συνίσταται από τρεις ή περισσότερες κατηγορίες μεταξύ των οποίων ισχύει η έννοια της ανισότητας, όπως

π.χ. σε μια ερώτηση της κλίμακας διαφωνώ καθόλου, λίγο, μέτρια, αρκετά, πολύ, στην κατάταξη ενός στρώματος υλικού ως λεπτού, μεσαίου, παχέος. 3. Ονομαστική (Nominal) ή πολυωνυμική (polynomial) ή πολυχοτομική (polychotomus) ή κατηγορική αδιαβάθμητη (non-ordered categorical) ή πολυμερής μεταβλητή απόκρισης. Περιέχει τρεις ή περισσότερες κατηγορίες χωρίς κάποια φυσική διαβάθμιση, όπως π.χ. ο χαρακτηρισμός ενός τρόφιμού ως τραγανού, μαλακού, εύθρυπτου ή του χρώματος αντικειμένων ως ερυθρού, πράσινου, κίτρινου κτλ. (Park, 2013).

Η λογιστική παλινδρόμηση χρησιμεύει στην περιγραφή της σχέσης που αναπτύσσεται μεταξύ μιας ή περισσότερων ανεξάρτητων μεταβλητών (π.χ. ηλικία, φύλο, τοξική συγκέντρωση ουσίας) και μιας δυαδικής μεταβλητής απόκρισης εκφρασμένης ως πιθανότητα δυνάμενη να πάρει μία από δύο τιμές, όπως π.χ. θετική (1) αρνητική (0), παρόν ενδεχόμενο (1) απόν ενδεχόμενο (0), επιζών (1) θανών (0), αρεστός (1) δυσάρεστος (0)

1. Η λογική παλινδρόμηση δεν υποθέτει μια γραμμική σχέση μεταξύ εξαρτημένων και ανεξάρτητων μεταβλητών.

2. Η εξαρτημένη μεταβλητή πρέπει να είναι μια διχοτόμηση (2 κατηγορίες).

3. Οι ανεξάρτητες μεταβλητές δεν χρειάζεται να είναι αλληλένδετες, ούτε κανονικά κατανεμημένες ούτε γραμμικές, ούτε της ίδιας διακύμανσης σε κάθε ομάδα.

Η πιο διαδεδομένη, έκφραση της εξίσωσης της Λογιστικής Παλινδρόμησης είναι:

$$\ln(odds) = \alpha + \beta 1x1 + \beta 2x2 + \beta 3x3 + \cdots . \beta\kappa x\kappa$$

Εξίσωση 4

Το δεξί μέρος της εξίσωσης δημιουργείται από ένα γραμμικό συνδυασμό των ανεξάρτητων μεταβλητών που συμμετέχουν στο μοντέλο της παλινδρόμησης.Το αριστερό μέρος περιέχει τις τιμές της εξαρτημένης μεταβλητής με την μορφή του λογαρίθμου των odd δηλαδή, του λογαρίθμου της σχέσης: odds = prob/(1-prob) .Το odds εναλλακτικά ονομάζεται logt και ο όρος Prob εκφράζει την πιθανότητα να συμβεί το γεγονός που έχει ορισθεί σαν επιτυχία του πειράματος.

Οι συντελεστές των ανεξάρτητων μεταβλητών στην εξίσωση της παλινδρόμησης εκτιμούνται βάση της μεθόδου Μέγιστης Πιθανοφάνειας βάση της μεθόδου αυτής η τιμή των συντελεστών των ανεξάρτητων μεταβλητών είναι αυτή που κάνει τις παρατηρηθήσες τιμές της εξαρτημένης μεταβλητής πιο πιθανές, βάση του σετ των ανεξάρτητων μεταβλητών (Karandreas, 2014).

Τα βήματα κατασκευής του μοντέλου της Λογιστικής Παλινδρόμησης είναι ανάλογα αυτών της γραμμικής παλινδρόμησης.

- Προσδιορίζουμε το μέγεθος του ενδιαφέροντος (εξαρτημένη μεταβλητή) κα ιτο σετ των ανεξάρτητων μεταβλητών που θα συμμετέχουν στην παλινδρόμηση.
- Διερευνούμε τα δεδομένα για τυχόν ύπαρξη ασυνήθιστων κινήσεων όπως, ακραίες τιμές, ελλείπουσες τιμές κ. λ. π.
- Ελέγχουμε την ικανοποίηση των υποθέσεων για την σωστή εφαρμογή της Λογιστικής Παλινδρόμησης.
- Δημιουργούμε την εξίσωσης της παλινδρόμησης.
- Μελετάμε την επίδραση κάθε ανεξάρτητης μεταβλητής στο μοντέλο.
- Εξετάζουμε την ικανοποίηση των υποθέσεων της Τεχνικής και διερευνούμε την πιθανότητα κάποια συγκεκριμένη τιμή να επηρεάζει υπερβολικά τα αποτελέσματα (Park&Hyeoun-Ae,2013.).

Στο σημείο αυτό θα πρέπει να αναφερθεί ότι η Λογιστική Παλινδρόμηση, για την σωστή εφαρμογή της απαιτεί μεγάλο δείγμα, προκειμένου να παράγει αξιόπιστο αποτέλεσμα. Ένας εμπειρικός κανόνας αναφέρει ότι το δείγμα θα πρέπει να είναι 30 φορές μεγαλύτερο από το αριθμό των παραμέτρων που εκτιμά το μοντέλο.

3.2 Τυχαίο Δάσος(Random Forest)

Για να μπορέσουμε να αναφερθούμε στην μέθοδο του τυχαίου δάσους θα πρέπει πρώτα να αναφερθούμε στον αλγόριθμό των δέντρων αποφάσεων.

Τα δέντρα αποφάσεων αποτελούν έναν από τους δημοφιλέστερους τρόπους ταξινόμησης και πρόβλεψης αφού μπορούν να χρησιμοποιηθούν τόσο σε κατηγορικές όσο και συνεχείς μεταβλητές.

Αρχικά διαθέτουμε ένα δείγμα του συνόλου δεδομένων όπου περιλαμβάνει όλα τα χαρακτηριστικά που μας ενδιαφέρουν. Στην συνέχεια ο αλγόριθμος λαμβάνει ως είσοδο κάθε περίπτωση (input vector) και συγκρίνοντας τις τιμές κάθε περίπτωσης τις κατατάσσει σε κλάσεις και δημιουργεί κανόνες οι οποίοι θα χρησιμοποιηθούν για μελλοντικά δεδομένα.

Το σημείο όπου γίνεται η υπόθεση για μία μεταβλητή είναι αυτό στο οποίο η μεταβλητή χωρίζεται (split) ανάμεσα σε δύο τιμές και ταυτόχρονα χωρίζει το training set σε δύο subsets. Σχηματικά, στην απεικόνιση του δένδρου το σημείο αυτό ονομάζεται «node». Από το κάθε node γεννιούνται 2 subsets - «κλαδιά» (branches) .

Οι κάτω κόμβοι είναι κόμβοι "τερματικού".

• Για την παλινδρόμηση, η προβλεπόμενη τιμή σε έναν κόμβο είναι ο μέσος όρος

Των μεταβλητών απόκρισης για όλες τις παρατηρήσεις στον κόμβο.

• Για ταξινόμηση η προβλεπόμενη τάξη είναι η πιο κοινή κλάση στον

κόμβο (ψηφοφορία με πλειοψηφία) (Zaggana, 2012).

Τα δέντρα αποφάσεων μπορούν να χειριστούν τεράστια σύνολα δεδομένων.

Συγκεκριμένα, μπορούν να χειριστούν μικτούς προγνωστικούς παράγοντες - ποσοτικούς και ποιοτικούς .Να αγνοήσουν μεταβλητές οι οποίες δεν επηρεάζουν το μοντέλο .Να διαχειριστούν ελλιπή δεδομένα με βέλτιστο τρόπο .Όμως πολλές φορές η πρόβλεψη και τα σφάλματα δεν είναι ακριβή ειδικότερα όταν δημιουργούνται μεγάλα δέντρα αποφάσεων όπου υπάρχει κίνδυνος overfitting.

Τα τυχαία δάση είναι μια από τις πιο ισχυρές, πλήρως αυτοματοποιημένες τεχνικές μηχανικής μάθησης. Με σχεδόν καμία προετοιμασία δεδομένων ή εμπειρογνωμοσύνη μοντελοποίησης, οι αναλυτές μπορούν εύκολα να αποκτήσουν εκπληκτικά αποτελεσματικά μοντέλα.

Τα τυχαία δάση αναπτύχθηκαν αρχικά από τον οραματιστή UC Berkeley Leo Breiman σε μια μελέτη που δημοσίευσε το 1999, βασιζόμενο σε μια ζωή με σημαντικές συνεισφορές συμπεριλαμβανομένου του δέντρου αποφάσεων CART.

Τα Random Forests είναι ουσιαστικά μία συλλογή από δέντρα αποφάσεων.

Αρχικά, αναπτύσσονται πολλά classification decision trees . Κάθε δέντρο δίνει μία ταξινόμηση – «Το δέντρο ψηφίζει αυτήν την κλάση». Έτσι, κάθε κλάση έχει έναν αριθμό «ψήφων» (votes). Η τελική και οριστική ταξινόμηση γίνεται με το «δάσος» να διαλέγει την κλάση με τις περισσότερες votes

Ο αλγόριθμός είναι ως εξής :

Αναπτύξτε ένα δάσος από πολλά δέντρα. (200-500 αποτελεί έναν ικανοποιητικό αρχικό αριθμό)

Αναπτύξτε κάθε δέντρο σε ένα ανεξάρτητο δείγμα bootstrap * από τα δεδομένα εκπαίδευσης .Όπου ως δείγμα bootstrap ορίζουμε ένα δείγμα από N τυχαίες περιπτώσεις με εναπόθεση. Εφαρμόζοντας την τεχνική αυτή, σε μεθόδους ταξινόμησης, δίνεται η δυνατότητα να δημιουργηθούν από ένα σύνολο δεδομένων περισσότερα από ένα σύνολα δεδομένων εκπαίδευσης. Αυτός είναι και ο λόγος που χρησιμοποιείται στα τυχαία δάση, καθώς θέλουμε να δημιουργήσουμε πολλά διαφορετικά δένδρα και επομένως πολλά σύνολα δεδομένων εκπαίδευσης.(Breiman & Cutler, 2014) .

Σε κάθε κόμβο:

1. Επιλέξτε m μεταβλητές τυχαία από όλες τις πιθανές μεταβλητές M (ανεξάρτητα για κάθε κόμβο).
2. Βρείτε την καλύτερη διάσπαση στις επιλεγμένες μεταβλητές m.

3.Αναπτύξτε τα δέντρα στο μέγιστο βάθος (ταξινόμηση).

4.Υπολογίστε τον μέσο όρο των δέντρων για να λάβετε προβλέψεις για νέα δεδομένα.

5.Συνολική πρόβλεψη εξόδου ως μέση απόκριση(παλινδρόμηση) ή ψηφοφορία με πλειοψηφία (ταξινόμηση) από όλα τα ατομικά εκπαιδευμένα δέντρα.

Ακρίβεια - Τα Τυχαία Δάση είναι ανταγωνιστικά με τις πιο γνωστές μεθόδους μηχανικής μάθησης .

Σταθερότητα: αν αλλάξουμε τα δεδομένα λίγο, τα μεμονωμένα δέντρα μπορεί να αλλάξουν, αλλά το δάσος είναι σχετικά σταθερό επειδή είναι ένας συνδυασμός πολλών δέντρων (Zaggana, 2012).

Η συγκεκριμένη μέθοδος παρουσιάζει και κάποια μειονεκτήματα κυρίως στην παλινδρόμηση.

Η παλινδρόμηση δεν μπορεί να προβλέψει πέρα από το εύρος στα δεδομένα εκπαίδευσης.

Στην παλινδρόμηση, οι ακραίες τιμές συχνά δεν προβλέπονται με ακρίβεια.

3.3 Gradient Tree Boosting

Το Gradient Tree Boosting (Friedman 2001) είναι ένας επαναληπτικός, μη παραμετρικός αλγόριθμος μηχανικής μάθησης που έχει χρησιμοποιηθεί με επιτυχία σε πολλούς τομείς. Παρουσιάζει αξιοσημείωτη ευελιξία στην επίλυση διαφορετικών συναρτήσεων απώλειας.

Ο αλγόριθμος του Gradient Boosting εξελίχθηκε από την εφαρμογή μεθόδων Boosting σε δέντρα παλινδρόμησης. Η γενική ιδέα είναι να υπολογιστεί μια ακολουθία (πολύ) απλών δέντρων, όπου κάθε διαδοχικό δέντρο είναι χτισμένο από τα υπόλοιπα πρόβλεψης του προηγούμενου δέντρου.

Η ενδυνάμωση(boosting) βασίζεται σε αδύναμους μαθητές (υψηλή μεροληψία, χαμηλή διακύμανση). Όσον αφορά τα δέντρα απόφασης, οι αδύναμοι μαθητές είναι ρηχά δέντρα, μερικές φορές ακόμη και μικρά (δέντρα με δύο φύλλα). Η ενδυνάμωση μειώνει το σφάλμα κυρίως με τη μείωση της μεροληψίας (και επίσης σε κάποιο βαθμό τη διακύμανση, με τη συγκέντρωση της παραγωγής από πολλά μοντέλα).

Από την άλλη πλευρά, το τυχαίο δάσος χρησιμοποιεί τα πλήρως αναπτυσσόμενα δέντρα αποφάσεων (χαμηλή προκατάληψη, υψηλή διακύμανση). Αντιμετωπίζει το έργο μείωσης του σφάλματος με τον αντίθετο τρόπο: μειώνοντας τη διακύμανση. Τα δένδρα γίνονται χωρίς συσχετισμό για να μεγιστοποιηθεί η μείωση της διακύμανσης, αλλά ο αλγόριθμος δεν μπορεί να μειώσει τη μεροληψία (η οποία είναι ελαφρώς υψηλότερη από τη μεροληψία ενός μεμονωμένου δένδρου στο δάσος.

Για να αποφευχθεί το overfitting, η ενίσχυση γίνεται συνήθως με αρκετά απλά υπομοντέλα: η κλασσική επιλογή είναι τα μικρά δέντρα απόφασης (Johansson, 2015).

Ο αλγόριθμος είναι ως εξής:

Algorithm 1 Simple boosting algorithm for regression.

let h_0 be a "dummy" constant model
let F_0 be an ensemble just consisting of h_0
for $m = 1, \ldots, M$
  for each pair (x_i, y_i) in the training set
    compute the residual $R(y_i, F_{m-1}(x_i)) = y_i - F_{m-1}(x_i)$
  train a regression sub-model h_m on the residuals
  add h_m to the ensemble: $F_m(x) = F_{m-1}(x) + h_m(x)$
return the ensemble F_M

3.1 Αλγόριθμος Gradient Tree Boosting

Κοινές παράμετροι δέντρου:

Αυτές οι παράμετροι καθορίζουν την τελική συνθήκη για την κατασκευή ενός νέου δέντρου. Συνήθως τις παραμετροποιούμε για να αυξήσουμε την ακρίβεια και να αποτρέψουμε το Overfitting.

- Μέγ. βάθος: πόσο ψηλά μπορεί να αναπτυχθεί ένα δέντρο .Συνήθως απαιτείται βάθος <10 και πολλές φορές καθορίζεται και από τον αριθμό των φύλλων.

- Μέγ. χαρακτηριστικά: πόσα χαρακτηριστικά μπορούν να χρησιμοποιηθούν για την οικοδόμηση ενός δοσμένου δέντρου Το δέντρο δεν χρειάζεται να χρησιμοποιεί όλα τα διαθέσιμα χαρακτηριστικά

- Ελάχ. δείγματα ανά φύλλο: πόσα δείγματα απαιτούνται για την κατασκευή ενός νέου φύλλου ○ Συνήθως θέλουμε <1% των δεδομένων ○ Μερικές φορές ορίζονται από δείγματα ανά διαχωρισμό.

Ρυθμός εκμάθησης: πόσο να προσαρμοστούν τα βάρη των δεδομένων μετά κάθε επανάληψη

- Μικρότερη προσαρμογή είναι καλύτερη αλλά πιο αργή
- Μέγεθος υποσυνόλων: Πόσες μετρήσεις για να εκπαιδευτεί κάθε νέο δέντρο
- Τα δείγματα δεδομένων επιλέγονται τυχαία κάθε επανάληψη
- Αριθμός δένδρων: Πόσα συνολικά δέντρα δημιουργούνται
- Συνήθως περισσότερα είναι καλύτερα, αλλά θα μπορούσαν να οδηγήσουν σε overfitting.

Οφέλη:

- Γρήγορη μέθοδος
- Και η μάθηση και η πρόβλεψη είναι γρήγορες
- Τα χαρακτηριστικά μπορεί να είναι ένα μείγμα κατηγορηματικών και συνεχών δεδομένων
- Καλή απόδοση
- Η εκπαίδευση στα σφάλματα δίνει πολύ καλή ακρίβεια
- μεγάλος αριθμός διαθέσιμου λογισμικού
- Χρησιμοποιούνται πολύ συχνά οι ενισχυμένοι αλγόριθμοι δέντρων
- Υπάρχει διαθέσιμο ένα πολύ καλά υποστηριζόμενο, καλά δοκιμασμένο λογισμικό.
- Προβλήματα:
- Ευαίσθητη μέθοδος στο overfitting και στον θόρυβο
- Θα πρέπει πάντα να πραγματοποιείται crossvalidating (Woodruff, 2017).

3.4 XGBOOST

Το Gradient Boosting είναι σήμερα μια από τις πιο δημοφιλείς τεχνικές για την αποτελεσματική μοντελοποίηση των συνόλων δεδομένων όλων των μεγεθών. Το XGboost είναι μια πολύ γρήγορη και κλιμακούμενη εφαρμογή του Gradient Boosting που έχει πάρει την επιστήμη των δεδομένων με μοντέλα που χρησιμοποιούν το XGBoost κερδίζοντας τακτικά πολλούς διαγωνισμούς ηλεκτρονικών επιστημών δεδομένων και χρησιμοποιείται σε μεγάλη κλίμακα σε διάφορους κλάδους (Chen & Guestrin, 2016).

Είναι ένα εξαιρετικά ευέλικτο και ευέλικτο εργαλείο που μπορεί να λειτουργήσει μέσω των περισσότερων μορφών παλινδρόμησης, ταξινόμησης και ταξινόμησης προβλημάτων. Ως λογισμικό ανοιχτού κώδικα, είναι εύκολα προσβάσιμο και μπορεί να χρησιμοποιηθεί μέσω διαφορετικών πλατφορμών.

Το Extreme Gradient Boosting, αναπτύχθηκε από την Tianqi Chen και τώρα αποτελεί μέρος μιας ευρύτερης συλλογής βιβλιοθηκών ανοιχτού κώδικα που αναπτύχθηκε από την Distributed Machine Learning Community (DMLC). Το XGBoost είναι μια κλιμακούμενη και ακριβής εφαρμογή του gradient tree boosting και έχει αποδείξει ότι προωθεί τα όρια της υπολογιστικής ισχύος για τους ενισχυμένους αλγόριθμους δέντρων καθώς χτίστηκε και αναπτύχθηκε με μοναδικό σκοπό την απόδοση μοντέλου και την υπολογιστική ταχύτητα. Συγκεκριμένα, έχει σχεδιαστεί για να εκμεταλλεύεται κάθε bit της μνήμης και των πόρων υλικού για αλγόριθμους ενίσχυσης δένδρων (Brownlee, 2017).

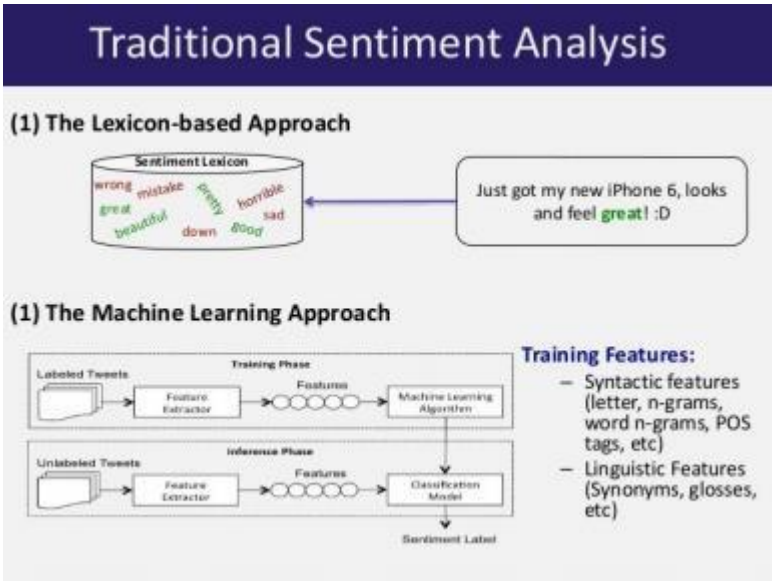
Το XGBoost είναι μία από τις ταχύτερες υλοποιήσεις των gradient boosted δέντρων. Αυτό συμβαίνει αντιμετωπίζοντας μια από τις σημαντικότερες ανεπάρκειες της προηγούμενης μεθόδου: την πιθανή απώλεια για όλους τους πιθανούς διαχωρισμούς δημιουργίας ενός νέου κλάδου (ειδικά εάν εξετάζουμε την περίπτωση όπου υπάρχουν χιλιάδες χαρακτηριστικά και επομένως χιλιάδες πιθανοί διαχωρισμοί). Το XGBoost αντιμετωπίζει αυτήν την αναποτελεσματικότητα εξετάζοντας τη διανομή των χαρακτηριστικών σε όλα τα δεδομένα σε ένα φύλλο και χρησιμοποιεί αυτές τις πληροφορίες για να μειώσει τον χώρο αναζήτησης πιθανών χωρισμάτων των χαρακτηριστικών.

3.5 Μέθοδοι συναισθηματικής ανάλυσης

Η Προσέγγιση Λεξικού (Lexicon Based Approach), υπολογίζει τον συναισθηματικό προσανατολισμό ενός κειμένου, χρησιμοποιώντας το σημασιολογικό προσανατολισμό των ατομικών λέξεων ή φράσεων του κειμένου. Το βοηθητικό λεξικό μπορεί να δημιουργηθεί είτε χειροκίνητα, είτε αυτόματα, χρησιμοποιώντας λέξεις-σπόρους (seedwords).

Η Προσέγγιση Μάθησης (Learning Approach) βασίζεται στη χρήση μεθόδων Μηχανικής Μάθησης (Machine Learning) και συγκεκριμένα με τη χρήση Μάθησης με Επίβλεψη (Supervised Learning). Αυτό σημαίνει ότι προκειμένου να γίνει εφαρμογή αυτής της προσέγγισης, χρειάζεται αρχικά η ύπαρξη ενός συνόλου δεδομένων, ήδη κατηγοριοποιημένων. Η προσέγγιση αυτή δε βασίζεται στην ύπαρξη ή δημιουργία ενός λεξικού, αλλά εκπαιδεύεται στην ανακάλυψη του σημασιολογικού προσανατολισμού, με τη χρήση των αρχικών δεδομένων (Thivaios, 2017).

3.5.1 .Μέθοδοι συναισθηματικής ανάλυσης (Εξόρυξη Γνώμης από Σχόλια Χρηστών στο Twitter σε Πραγματικό Χρόνο)
.



3.5.1 Τεχνική Απεικόνισης

Το Tag Cloud είναι ίσως η πιο συχνή μέθοδος οπτικοποίησης κειμένου. Αποτελείται ουσιαστικά από μια λίστα λέξεων, φράσεων ή περιγραφών, οι οποίες έχουν μέγεθος ανάλογο με τη συχνότητα εμφάνισης κάθε λέξης, δηλαδή το πόσες φορές εμφανίζεται μια συγκεκριμένη λέξη σε ένα κείμενο. Χρησιμοποιούνται συχνά για να τονίσουν τις κύριες πτυχές ενός κειμένου.

3.5.2.Παράδειγμα χρήσης της τεχνικής word cloud

3.6 Περιγραφή δείγματος και διαδικασίας

Στην παρούσα ενότητα θα αναφερθούμε αρχικά στην συλλογή των δεδομένων και επεξεργασία τους. Στην συνέχεια θα μελετήσουμε τις μεταβλητές που θα χρησιμοποιήσουμε ως προς την κατανομή τους, την σχέση που έχουν μεταξύ τους και πως επηρεάζουν τις εξαρτημένες μεταβλητές.

3.6.1 Επισκόπηση δεδομένων

Η συλλογή των δεδομένων πραγματοποιήθηκε από την ιστοσελίδα InsideAirbnb όπου έγινε η συλλογή 3 αρχείων (listings.csv.gz ,calendar.csv.gz, reviews.csv.gz).Σε κάθε ένα από τα 3 dataset είναι σωστά δομημένα με μια μεταβλητή ανά στήλη και σειρά. Μεταβλητές, οι οποίες εμπεριείχαν σύμβολα όπως $ ή ευρώ ή διάφορα άλλα σημεία στίξης διορθώθηκαν ώστε να μπορέσουν να χρησιμοποιηθούν

στην μελέτη. Συγκεκριμένα, η μεταβλητή host_verifications μετατράπηκε σε Number of Verifications όπως αντίστοιχα και η μεταβλητή amenities σε number of amenities που μετατράπηκε από λίστα σε αριθμό. Επίσης, η μεταβλητή host_since η οποία αποτελεί ημερομηνία μετατράπηκε σε μέρες ώστε να μπορέσει να συμπεριληφθεί στην ανάλυση. Τέλος, με την χρήση των μεταβλητών longitude, latitude καταφέραμε να υπολογίσουμε με την βοήθεια του λογισμικού dataiku το geopoint κάθε listing όπως και τις μεταβλητές distancePlaka, DistanceOmonoia ,DistanceAcropolis που μας επιτρέπουν να αποτυπώσουμε τις σχέσεις μεταξύ των μεταβλητών γραφικά σε χάρτη.

Παρόλο που σε γενικές γραμμές τα δεδομένα έχουν έναν ικανοποιητικό βαθμό πληρότητας κάποιες μεταβλητές όπως review_score_rating κτλ έχουν ελλιπείς τιμές. Υπάρχουν πολλές τεχνικές που μπορούν να χρησιμοποιηθούν για να διαχειριστούμε τα δεδομένα που απουσιάζουν από το dataset όπως η χρήση της μέσης τιμής ή η παλινδρόμηση. Μετά από επαναλαμβανόμενες δοκιμές των τεχνικών μέσης τιμής και παλινδρόμησης καταλήξαμε σε αποτελέσματα μη αντιπροσωπευτικά αφού μειώθηκε σημαντικά η συσχέτιση μεταξύ των μεταβλητών. Συνεπώς, επιλέχθηκε η μέθοδος διαγραφής των ελλιπών δεδομένων ανά γραμμές αφού τα δεδομένα μας δεν ακολουθούν κάποιο μοτίβο Missing Completely at Random (MCAR) και παρόλο που η συγκεκριμένη μέθοδος μειώνει τα δεδομένα μας είναι αρκετά για να βγάλουμε συγκεκριμένα συμπεράσματα. Το τελικό μέγεθος των δεδομένων ανέρχεται σε 4000 reviews και 3657 δεδομένα που θα χρησιμοποιηθούν για την πρόβλεψη των εξαρτημένων μεταβλητών τιμή και superhost.

Στην βάση δεδομένων μας εμπεριέχονται 3 πιθανές μεταβλητές για την μοντελοποίηση της τιμής. Η τιμή ανά διανυκτέρευση, η τιμή ανά βδομάδα και η τιμή ανά μήνα. Επιλέξαμε να χρησιμοποιήσουμε την μεταβλητή τιμή ανά διανυκτέρευση αφού τα εβδομαδιαία και μηνιαία δεδομένα έχουν ελλείψεις και οι τιμές διαφοροποιούνται αρκετά λόγω των διαφορετικών εκπτώσεων που έχει την δυνατότητα να πραγματοποιήσει ο κάθε οικοδεσπότης.

3.6.2 Μοντελοποίηση Πρόβλεψης Τιμής Καταλυμάτων

Στο συγκεκριμένο τμήμα της έρευνας θα χρησιμοποιήσουμε τόσο απλές μεθόδους πρόβλεψης όσο και πιο σύγχρονες και πολύπλοκες οι οποίες θα μας δώσουν πιο αντιπροσωπευτικά αποτελέσματα. Σε όλα τα μοντέλα θα χρησιμοποιήσουμε τόσο στον συντελεστή $R2$ όσο και το μέσο όρο των τετραγωνικών σφαλμάτων (MSE). Επίσης, σε όλα τα μοντέλα το 80% του δείγματος χρησιμοποιήθηκε για μάθηση ενώ το υπόλοιπο 20% για testing και συγκεκριμένα το δείγμα χωρίστηκε σε 3 μέρη με την μέθοδο k fold cross validation.

Τα χαρακτηριστικά που αξιοποιήσαμε είναι τα ακόλουθα :

Host_is_Superhost, host_total_listings_count, neighbourhood_cleansed, latitude, longtitude, property_type, room_type, accomodates, bathrooms, bedrooms, beds,

Bed_type, cleaning_fee_amount,

guests_included, extra_people_amount, availability_30, availability_60, availability_90, number_of_reviews, review_scores_rating, reviews_per_month,

DistancePlaka, DistanceOmonoia, DistanceAcropolis, Number of Verifications, Number of Amenities.

3.6.3 Μοντελοποίηση Πρόβλεψης Superhost

Στο συγκεκριμένο στάδιο της μελέτης ο κύριος στόχος είναι η μοντελοποίηση των χαρακτηριστικών που απαιτούνται για να γίνει κάποιος οικοδεσπότης superhost. Παρόλο, που υπάρχουν κάποιες συγκεκριμένες απαιτήσεις τις οποίες οφείλει να ικανοποιεί ο οικοδεσπότης για να γίνει superhost υπάρχουν περαιτέρω χαρακτηριστικά τα οποία διαφοροποιούν τις 2 κατηγορίες ιδιοκτητών.
Για την ανάλυση των χαρακτηριστικών χρησιμοποιήθηκαν 3 μέθοδοι : Η Λογιστική Παλινδρόμηση, το Support Virtual Machine και η μέθοδος Random Forest.
Τα χαρακτηριστικά που χρησιμοποιήθηκαν για την μοντελοποίηση της μεταβλητής Superhost είναι τα ακόλουθα: since_host_since_parsed_days, host_response_rate, number_of_reviews , review_scores_rating , review_scores_accuracy , review_score_cleanliness , review_scores_checkin , review_scores_communication , review_scores_location , review_scores_value , reviews_per_month , number of Verifications **.**

Σε όλα τα μοντέλα θα χρησιμοποιήσουμε τον συντελεστή ROC. Ο συγκεκριμένος συντελεστής μπορεί να οριστεί σαν την πιθανότητα όπου ταξινομητής θα ταξινομήσει μια τυχαία επιλεγμένη θετική παρατήρηση υψηλότερα από μία τυχαία επιλεγμένη αρνητική παρατήρηση.

Επίσης, σε όλα τα μοντέλα το 80% του δείγματος χρησιμοποιήθηκε για μάθηση ενώ το υπόλοιπο 20% για testing και συγκεκριμένα το δείγμα χωρίστηκε σε 3 μέρη με την μέθοδο k fold cross validation.

3.6.4 Εξόρυξη γνώμης για τα καταλύματα της Αθήνας που έχουν εισαχθεί στην πλατφόρμα του Airbnb

Στόχος αυτού του τμήματος της έρευνας είναι η αποκωδικοποίηση της γνώμης των επισκεπτών για τα καταλύματα του δήμου Αθηναίων και συγκεκριμένα να φανούν τόσο τα δυνατά τους σημεία όσο και τυχόν κοινές αδυναμίες οι οποίες θα πρέπει να βελτιωθούν στο μέλλον.

Για το κομμάτι της εξόρυξης γνώμης χρησιμοποιήθηκε το αρχείο δεδομένων reviews.csv το οποίο διαθέτει 124000 κριτικές από καταλύματα που έχουν καταχωρηθεί στην πλατφόρμα του Airbnb στο οποίο αναφέρεται η ταυτότητα του καταλύματος, η ημερομηνία καταχώρησης της κριτικής, το όνομα του επισκέπτη και φυσικά η κριτική του. Κατά την διαδικασία εξόρυξης γνώμης πρέπει να δημιουργηθεί μια καινούργια μεταβλητή η οποία θα λειτουργεί ως δείκτης για το αν μια κριτική είναι θετική ή όχι. Η καινούργια μεταβλητή ονομάστηκε Sentimental και λαμβάνει την τιμή 1 όταν η κριτική είναι θετική και την τιμή 0 όταν είναι αρνητική. Σε γενικές περιπτώσεις 500 κριτικές θα ήταν αρκετές για την ανάλυση μας όμως η πλατφόρμα απορρίπτει σε βάθος χρόνου τα καταλύματα που έχουν μέσο όρο κριτικών κάτω από 3/5 αστέρια. Επομένως , η πλειοψηφία των κριτικών είναι θετικές και για να βγάλουμε

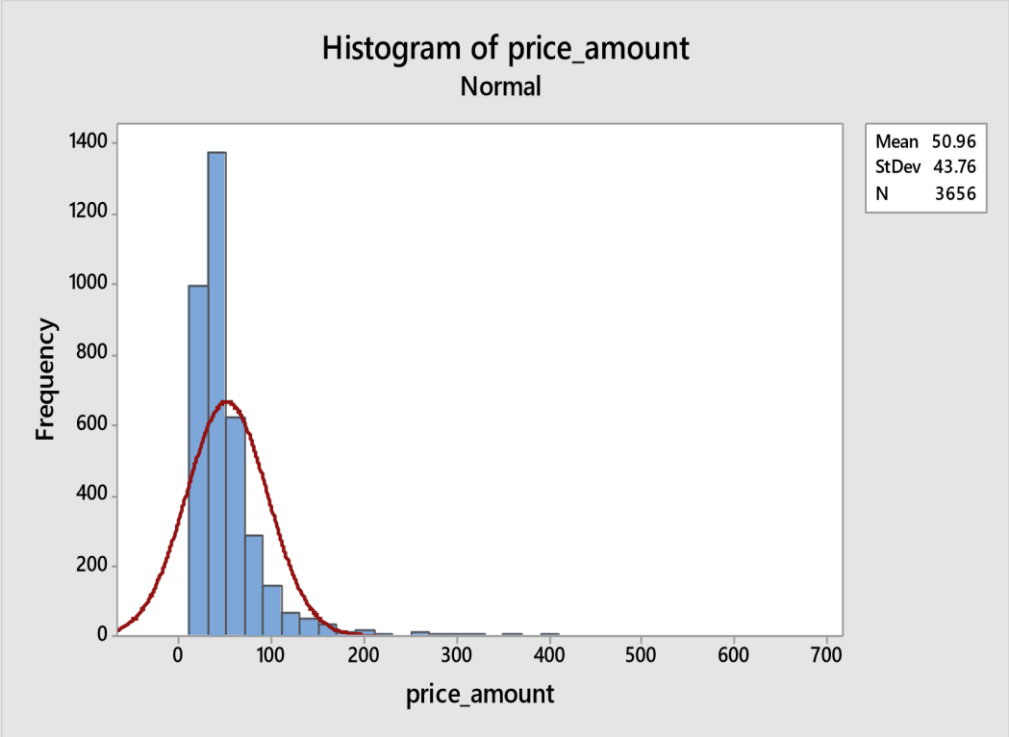
ασφαλή συμπεράσματα καταλήξαμε στην χρήση 4000 κριτικών. Επιπλέον, διορθώθηκαν όποια ορθογραφικά λάθη υπήρχαν και αφαιρέθηκαν κριτικές οι οποίες ήταν σε άλλη γλώσσα εκτός της αγγλικής για να υπάρξουν πιο αντιπροσωπευτικά αποτελέσματα.

Οι μεταβλητές που θα αξιοποιηθούν είναι οι: Comments η οποία περιέχει τις κριτικές και φυσικά η Sentimental η οποία είναι και η εξαρτημένη μεταβλητή που επιθυμούμε να μελετήσουμε. Στην μεταβλητή Comments απορρίπτουμε τις λέξεις που εμφανίζονται στο 0,1% των γραμμών όπως και τις λέξεις που εμφανίζονται με μεγάλο ρυθμό και συγκεκριμένα με συχνότητα άνω του 80% λόγω του ότι τόσο συχνά εμφανιζόμενες λέξεις είναι σύνδεσμοι, άρθρα κτλ. Επίσης, τα Ngrams που θα χρησιμοποιηθούν (δηλαδή πολλές λέξεις που θεωρούνται ως 1 από τον αλγόριθμο) θα είναι από 1 λέξη έως και 3 ώστε να υπάρχει μια σημαντική διασφάλιση ότι δεν θα απορριφθούν μικρές εκφράσεις οι οποίες καθορίζουν μια κριτική. Όπως και στα προηγούμενα τμήματα της έρευνας το 80% των δεδομένων θα χρησιμοποιηθεί ως train set ενώ το υπόλοιπο 20% ως test set. Η μέθοδος ανάλυσης των κριτικών στην πλατφόρμα του Airbnb είναι η λογιστική παλινδρόμηση με την χρήση L2 κανονικοποίησης.

4.Αποτελέσματα

4.1 Περιγραφική ανάλυση

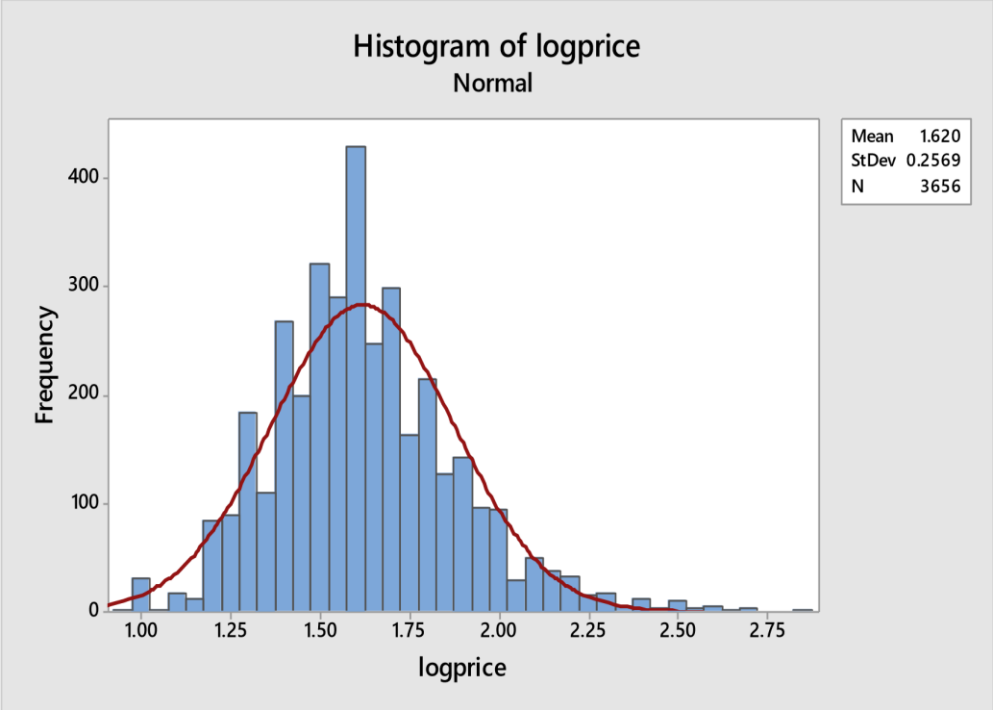
Στο συγκεκριμένο τμήμα της έρευνας παρουσιάζουμε τα αποτελέσματα από την χρήση των τεχνικών μηχανικής μάθησης στην πλατφόρμα Airbnb.



4.1 Ιστόγραμμα της τιμής των καταλυμάτων με fit

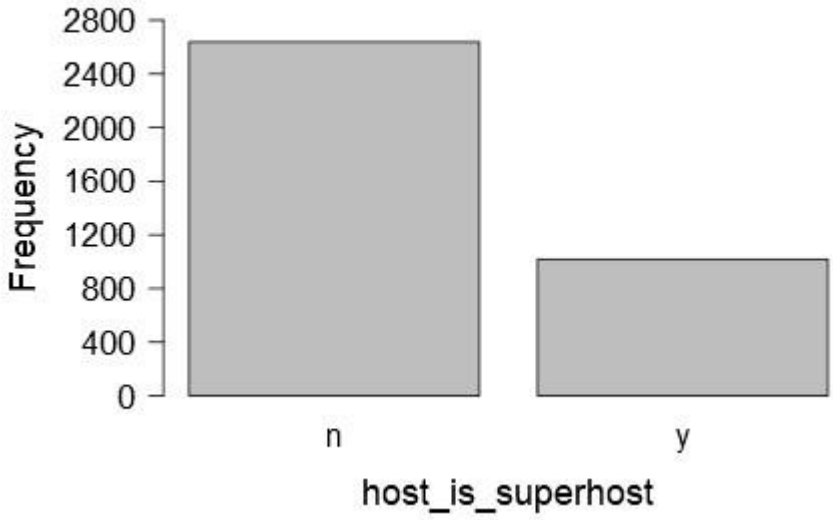
Αρχικά από το ιστόγραμμα της μεταβλητής price_amount προκύπτει ότι η μεταβλητή έχει μεγάλες αποκλίσεις από την κανονική κατανομή με την πλειοψηφία των παρατηρήσεων να βρίσκονται στο εύρος 20-80 και ελάχιστες ακραίες παρατηρήσεις να φθάνουν έως και τα 700 ευρώ. Για να κανονικοποιήσουμε την εξαρτημένη μεταβλητή εφαρμόσαμε μια μετατροπή με βάση τον λογάριθμο έτσι ώστε να επιτύχουμε μια πιο κανονικοποιήμενη κατανομή που θα ανταποκρίνεται πιο βέλτιστα στις τεχνικές πρόβλεψης με επίβλεψη.

Μετά την μετατροπή η εξαρτημένη μεταβλητή ακολουθεί την παρακάτω κατανομή που είναι αρκετά πιο κοντά στην κανονική και μπορεί να δώσει πιο αξιόπιστα αποτελέσματα.



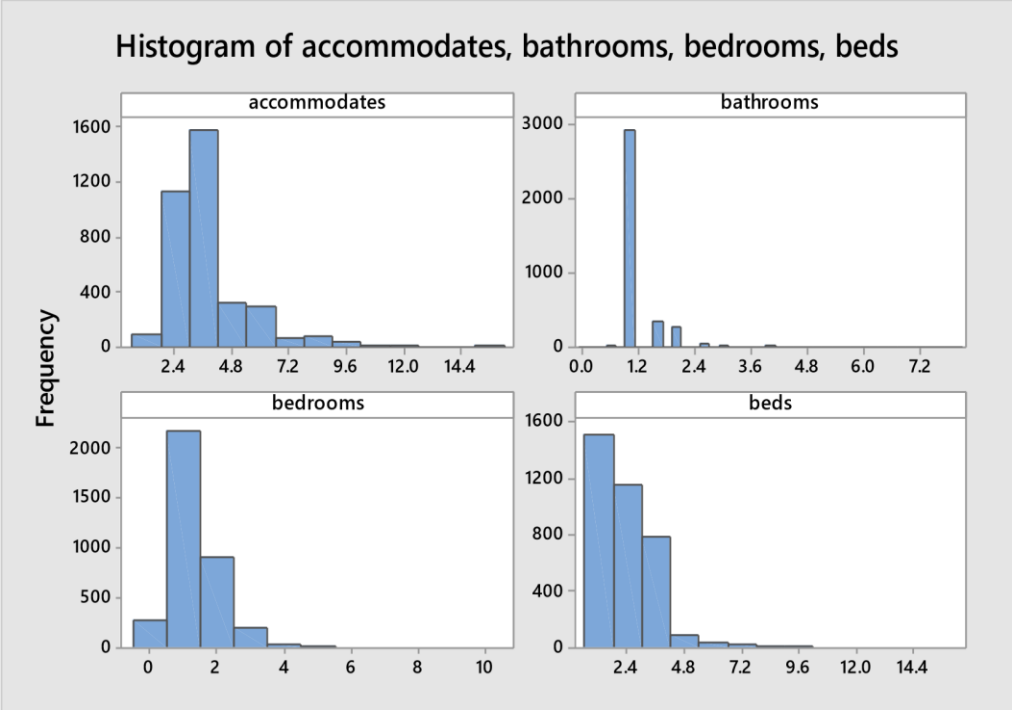
4.2 Ιστόγραμμα της νέας μεταβλητής Logprice με fit

Αντίστοιχα για την μεταβλητή host_is_superhost από τον πίνακα συχνοτήτων παρατηρούμε ότι η πλειοψηφία των οικοδεσποτών δεν είναι superhosts όπου συγκεκριμένα το 72,1% είναι απλοί οικοδεσπότες ενώ το 27,9% είναι superhosts.



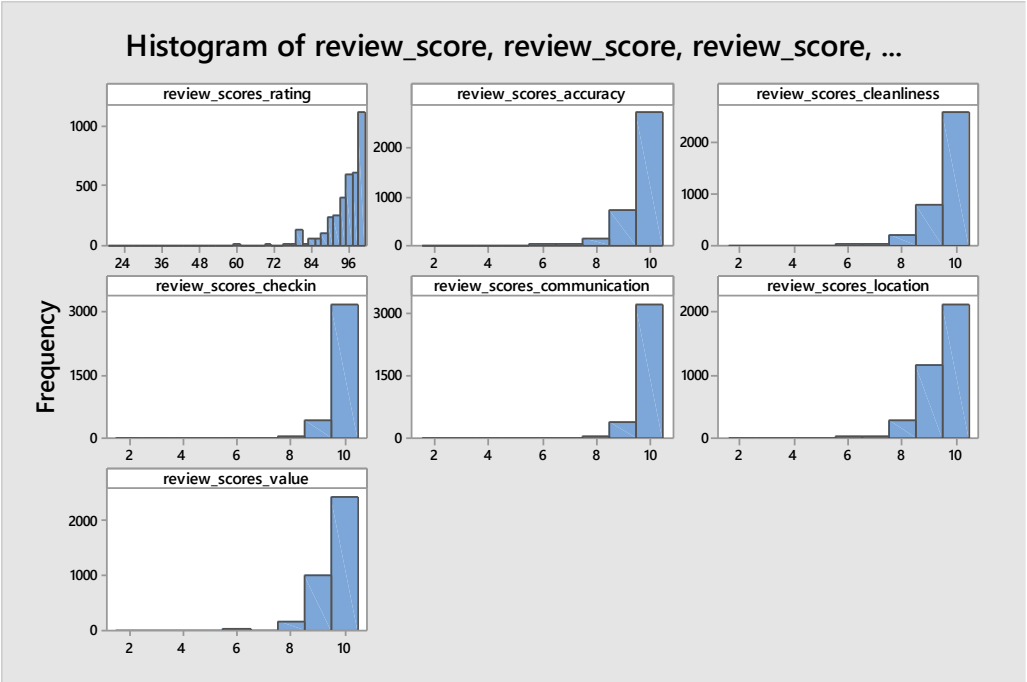
4.3Διάγραμμα της μεταβλητής host_is_superhost

Ανάλογα για τις ανεξάρτητες μεταβλητές έχουμε ότι οι μεταβλητές accommodates,bathrooms,bedrooms,beds ακολουθούν μια λοξότητα προς τα αριστερά με κάποιες ελάχιστες ακραίες τιμές. Συγκεκριμένα, η πλειοψηφία των ακινήτων μπορεί να φιλοξενήσει 2 έως 7 άτομα, το μεγαλύτερο μέρος των ακινήτων(80%) διαθέτει 1 μπάνιο, αντίστοιχα η πλειοψηφία τους διαθέτει 1-2 υπνοδωμάτια και υπάρχουν σπάνιες περιπτώσεις όπου διαθέτουν άνω των 5 κρεβατιών.

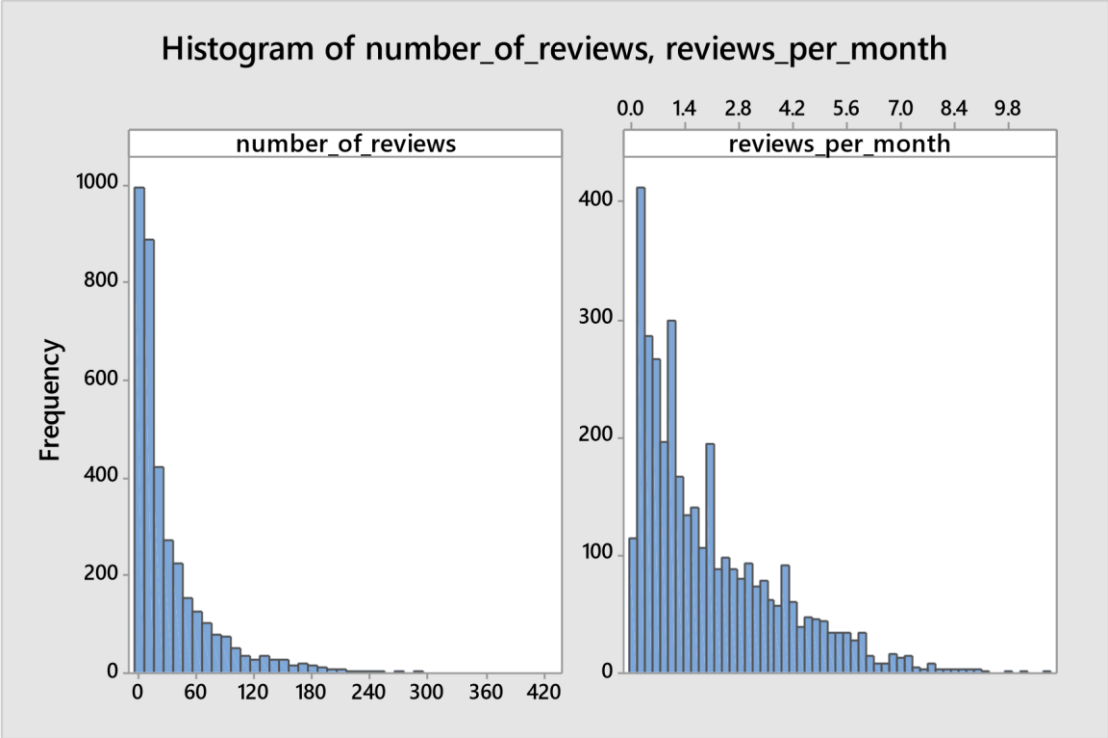


4.4Ιστογράμματα για τις μεταβλητές accommodates, bathrooms ,bedrooms , beds

4.5 Ιστογράμματα για τις μεταβλητές review_scores_rating λαμβάνει τιμές κυρίως άνω του 80. Ανάλογα οι μεταβλητές review_scores_Accuracy, review_scores_cleanliness,review_scores_checkin ,review_scores_communication,_review_scores_location, review_scores_value



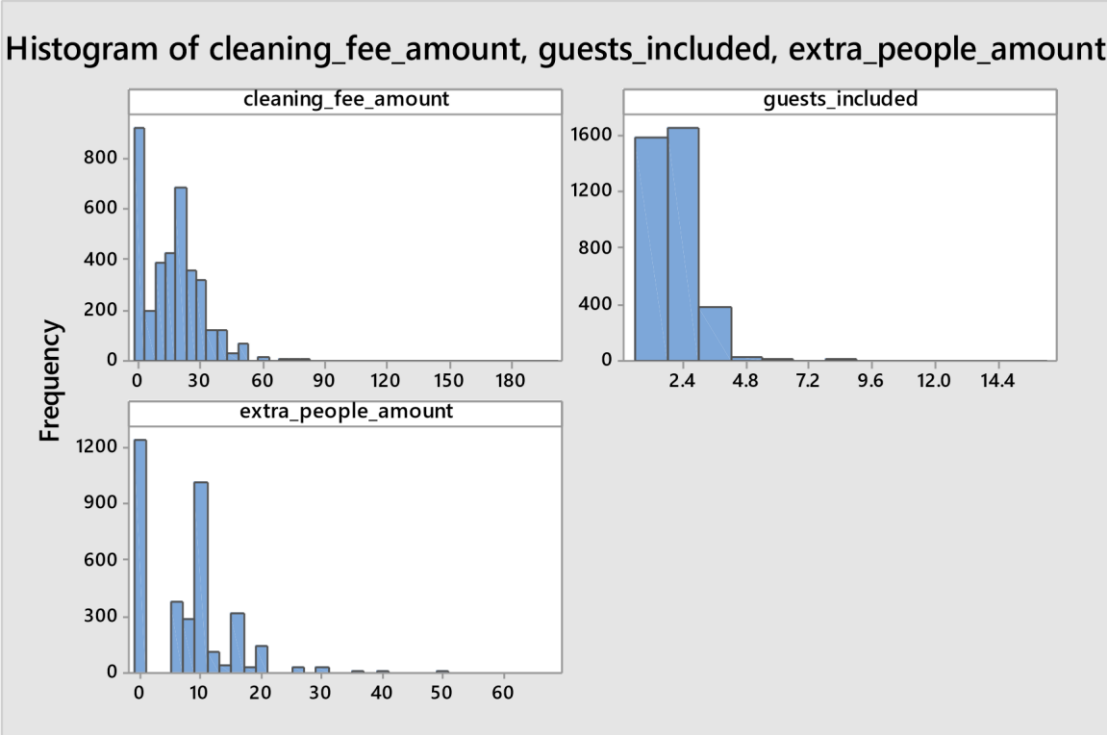
Αντίστοιχα, η κατανομή των μεταβλητών number_of_reviews και reviews_per_month ακολουθεί μια λοξότητα προς τα αριστερά με ελάχιστες ακραίες τιμές να ξεπερνούν τα 200 reviews και 8 reviews/per month αντίστοιχα.



4.6 Ιστογράμματα για τις μεταβλητές number_of_reviews, reviews_per_month

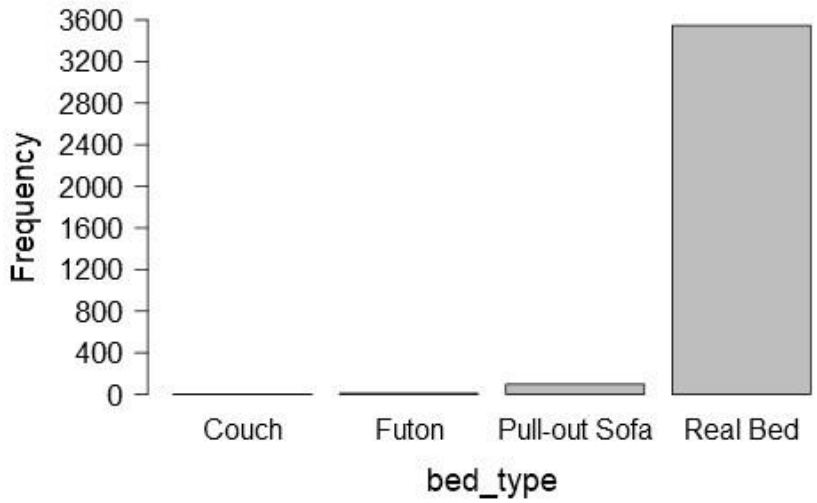
Παρόμοιο μοτίβο ακολουθούν και οι μεταβλητές cleaning_fee_amount, guests_included, extra_people_amount όπου παρουσιάζουν λοξότητα προς τα αριστερά με το μεγαλύτερο μέρος των κατοικιών να μην έχει έξοδα καθαρισμού άνω των 30 ευρώ ,να μην δέχεται πάνω από 5 επισκέπτες και όχι πάνω από 20 επιπλέον άτομα.

Μια εντελώς διαφορετική κατανομή παρουσιάζουν τα δεδομένα των μεταβλητών που αφορούν την βαθμολογία των φιλοξενούμενων στις υπηρεσίες των καταλυμάτων. Συγκεκριμένα, στην κλίμακα 0-100 η μεταβλητή review_scores_rating λαμβάνει τιμές κυρίως άνω του 80. Στο επόμενο ιστόγραμμα οι μεταβλητές **review_scores_Accuracy,review_scores_cleanliness,review_scores_checkin,revie w _scores_communication,_review_scores_location, review_scores_value** στην κλίμακα 0-10 λαμβάνουν τιμές κυρίως άνω του 8 με κάποιες ακραίες τιμές στο εύρος 0-2.



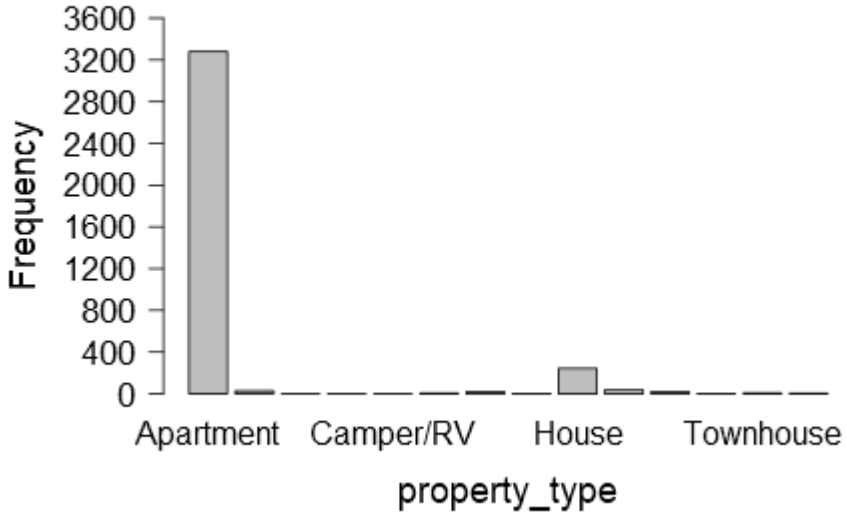
4.7 Ιστογράμματα για τις μεταβλητές cleaning_fee_amount, guests_included, extra_people_amount

Αντίστοιχα, για την μεταβλητή bed_type παρατηρούμε ότι το 97% των παρατηρήσεων είναι κανονικά κρεβάτια, ακολουθεί ο αναδιπλούμενος καναπές με 2,7%,το φουτόν με 0,2% και τέλος ο απλός καναπές με 0,1%.



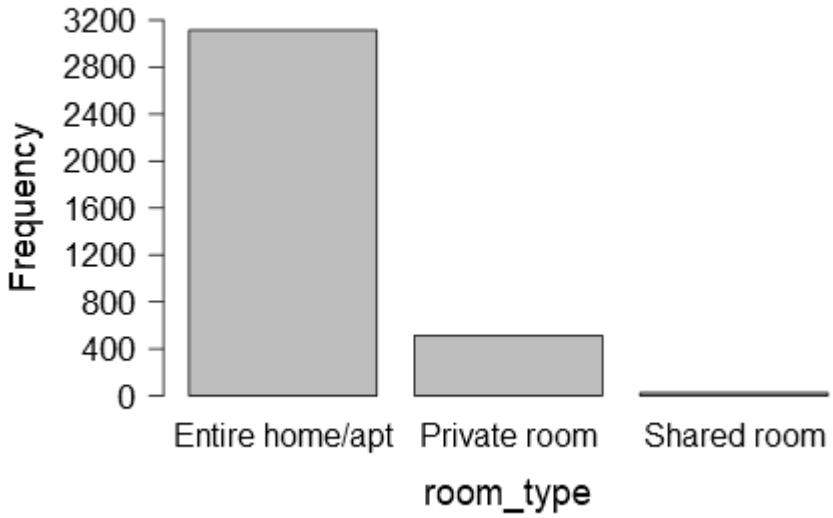
4.8 Διάγραμμα συχνοτήτων τύπου κρεβατιού

Στην μεταβλητή property_type η πλειοψηφία των παρατηρήσεων συγκεντρώνεται στην επιλογή διαμέρισμα με ποσοστό 89,7% και ακολουθεί το απλό σπίτι με ποσοστό 6,7%.Τρίτη επιλογή αποτελεί η σοφίτα με ποσοστό 1% και οι υπόλοιπες επιλογές λαμβάνουν ποσοστά κάτω του 1%.



4.9 Διάγραμμα συχνοτήτων τύπου καταλύματος

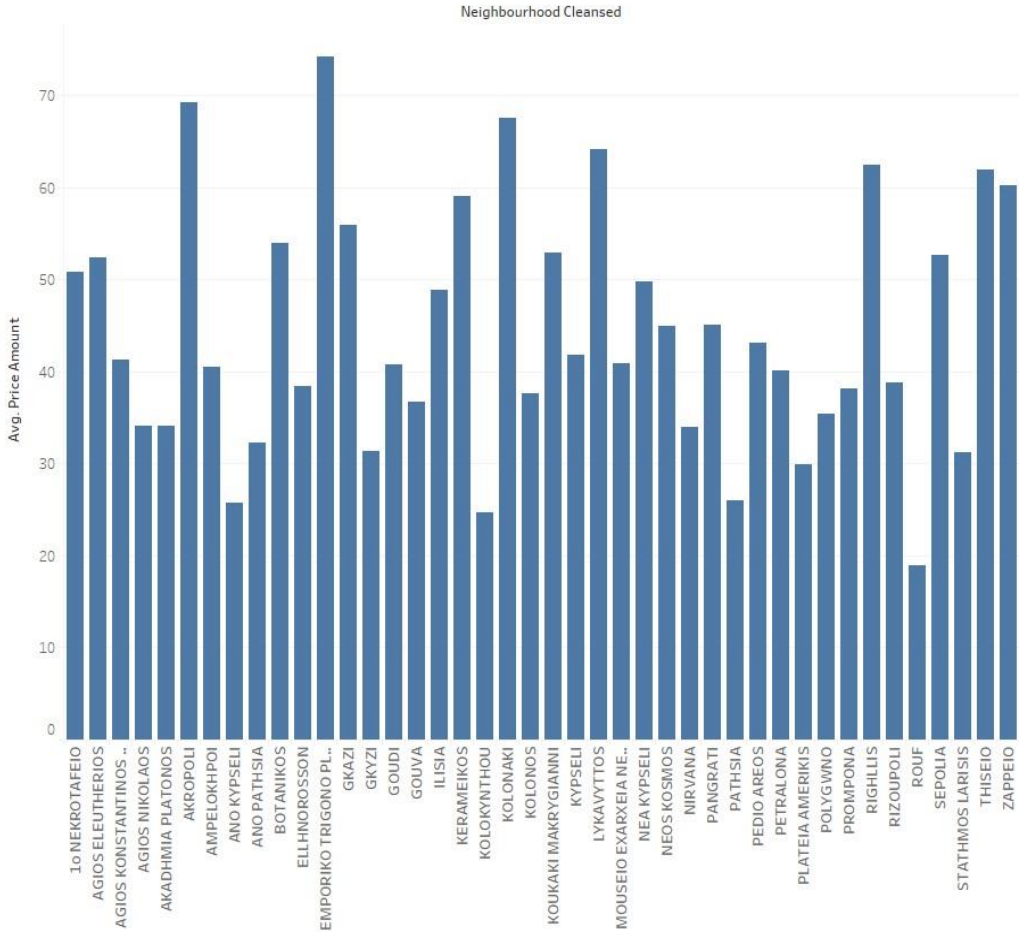
Στο επόμενο διάγραμμα συχνοτήτων η μεταβλητή room_type το 85.2% των επισκεπτών επέλεξε ολόκληρο σπίτι ή διαμέρισμα για τις διακοπές του, το 14.1% ατομικό δωμάτιο ενώ μόλις το 0.7% των επισκεπτών επέλεξε το κοινό δωμάτιο.



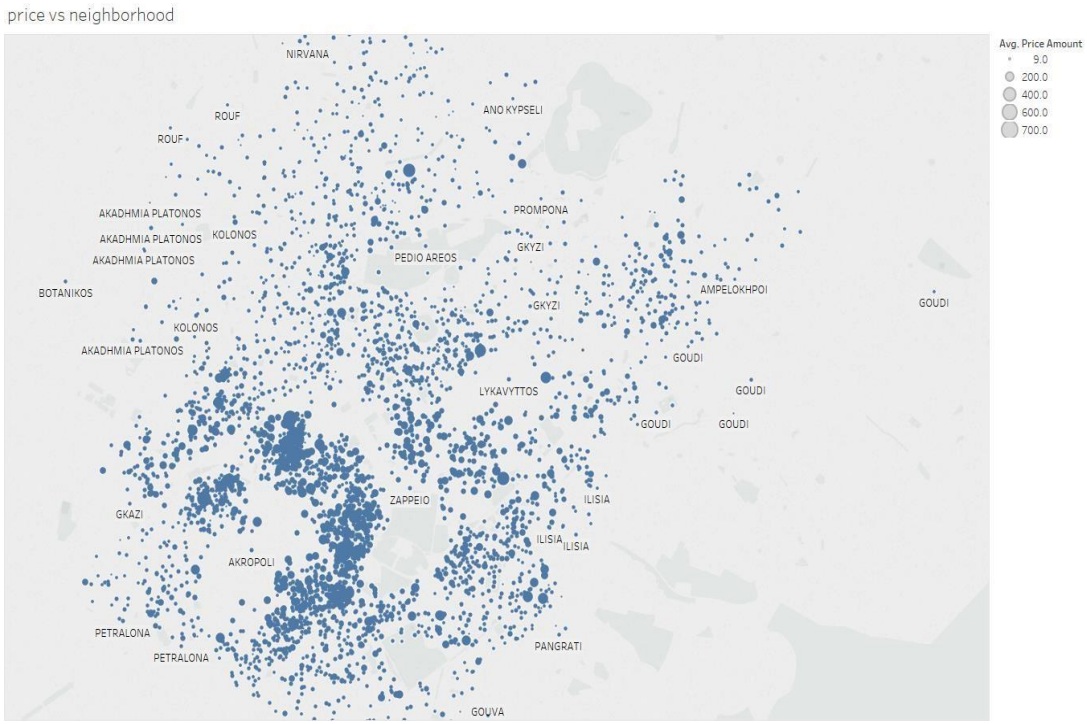
4.10 Διάγραμμα συχνοτήτων για την μεταβλητή τύπος δωματίου

Με την βοήθεια του προγράμματος tableau απεικονίσαμε κάθε ένα από τα ξεχωριστά καταλύματα σε έναν ψηφιακό χάρτη ανάλογα με την τιμή που έχουν ανά διανυκτέρευση. Από τον χάρτη παρατηρούμε ότι τα ακίνητα με τις υψηλότερες τιμές βρίσκονται γύρω από την Ακρόπολη, κοντά στο Ζάππειο και έχουν τιμές άνω των 200 ευρώ εκτός ελαχίστων εξαιρέσεων . Αντίθετα, περιοχές όπως το Γουδί, οι Αμπελόκηποι, το Ρούφ και η Άνω Κυψέλη έχουν τις χαμηλότερες τιμές κάτω από 100 ευρώ. Συγκεκριμένα, οι περιοχές με την μεγαλύτερη μέση τιμή είναι οι Εμπορικό Τρίγωνο Πλάκας με 74,14 ευρώ ανά κατάλυμα, ακολουθούν οι περιοχές Ακρόπολη και Κολονάκι με τιμές 69.19 ευρώ και 67.47 ευρώ αντίστοιχα ενώ αντίθετα οι περιοχές με την χαμηλότερη μέση τιμή καταλυμάτων είναι οι Ρουφ με 19 ευρώ ,Κολοκυνθού με 24,67 και τέλος Άνω Κυψέλη με 25,79 ευρώ.

<Average Price by Neighbourhood>

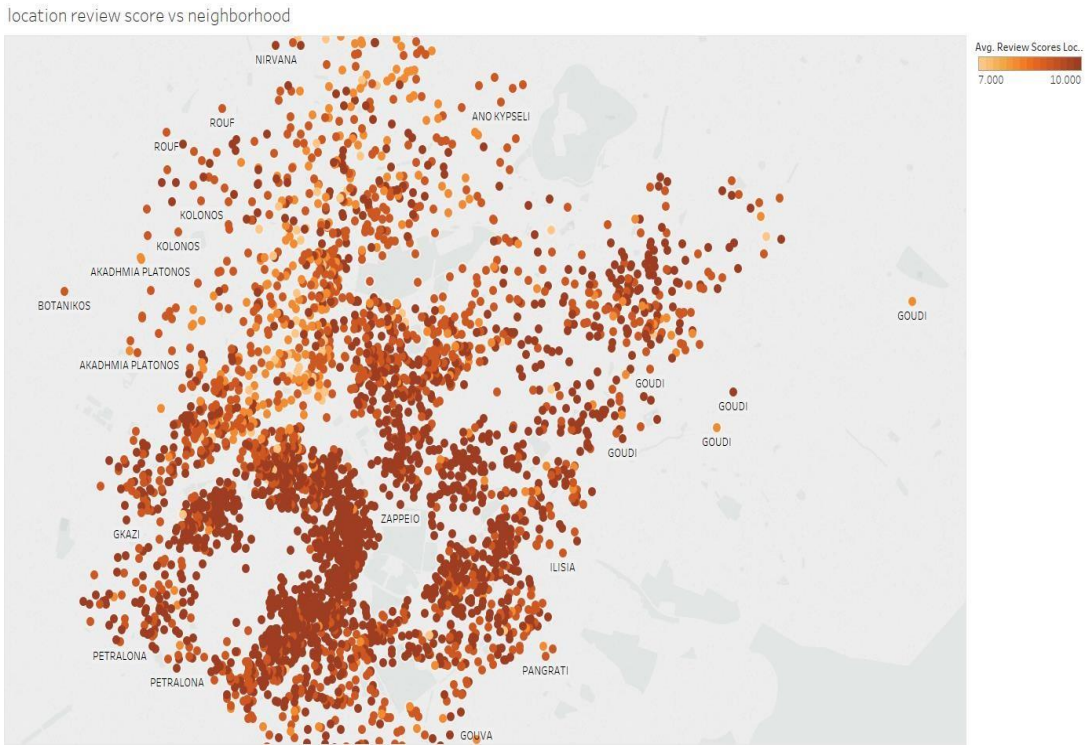


4.11Μέση τιμή αξίας καταλυμάτων για μια διανυκτέρευση ανά περιοχή



4.12 Γεωγραφικός χάρτης της τιμής ανάλογα με την περιοχή

Αντίστοιχα, στην απεικόνιση του χάρτη της περιοχής σε σχέση την βαθμολογία του κάθε καταλύματος ανάλογα με την τοποθεσία του παρατηρούμε ότι οι περιοχές με το πιο σκούρο χρώμα άρα και μεγαλύτερη βαθμολογία είναι η Ακρόπολη, το Κουκάκι,η Πλάκα, Κολωνάκι και Παγκράτι ενώ περιοχές όπως ο άγιος Κωνσταντίνος και η άνω κυψέλη έχουν από τις χαμηλότερες βαθμολογίες. Ήδη από τα γραφήματα, παρατηρούμε ότι υπάρχει συσχέτιση μεταξύ της τιμής και της βαθμολογίας των φιλοξενούμενων για την τοποθεσία.

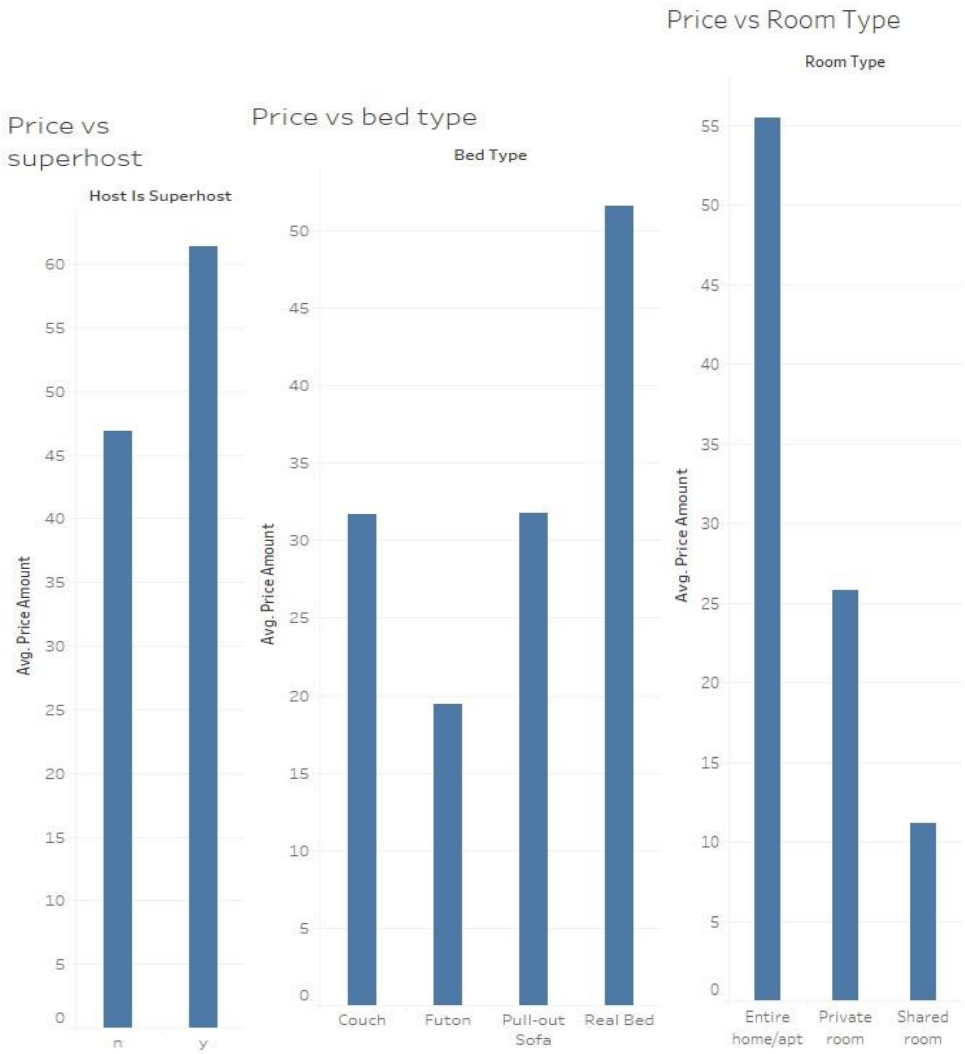


4.13 Γεωγραφικός χάρτης της βαθμολογίας κριτικής σε σχέση με την περιοχή

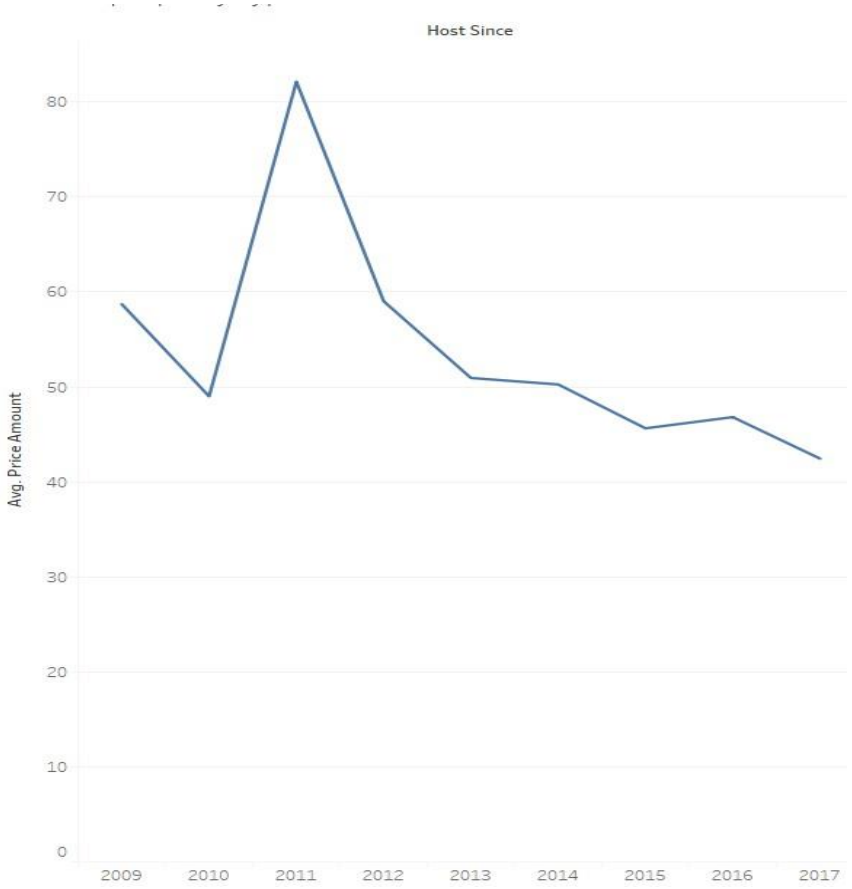
Αντίθετα για την μεταβλητή host_is_superhost δεν μπορούμε από την γραφική αναπαράσταση να καταλήξουμε σε κάποια συμπεράσματα αφού οι οικοδεσπότες είναι ισοκατανεμημένοι σε όλες τις περιοχές.

Συσχέτιση μεταξύ των μεταβλητών

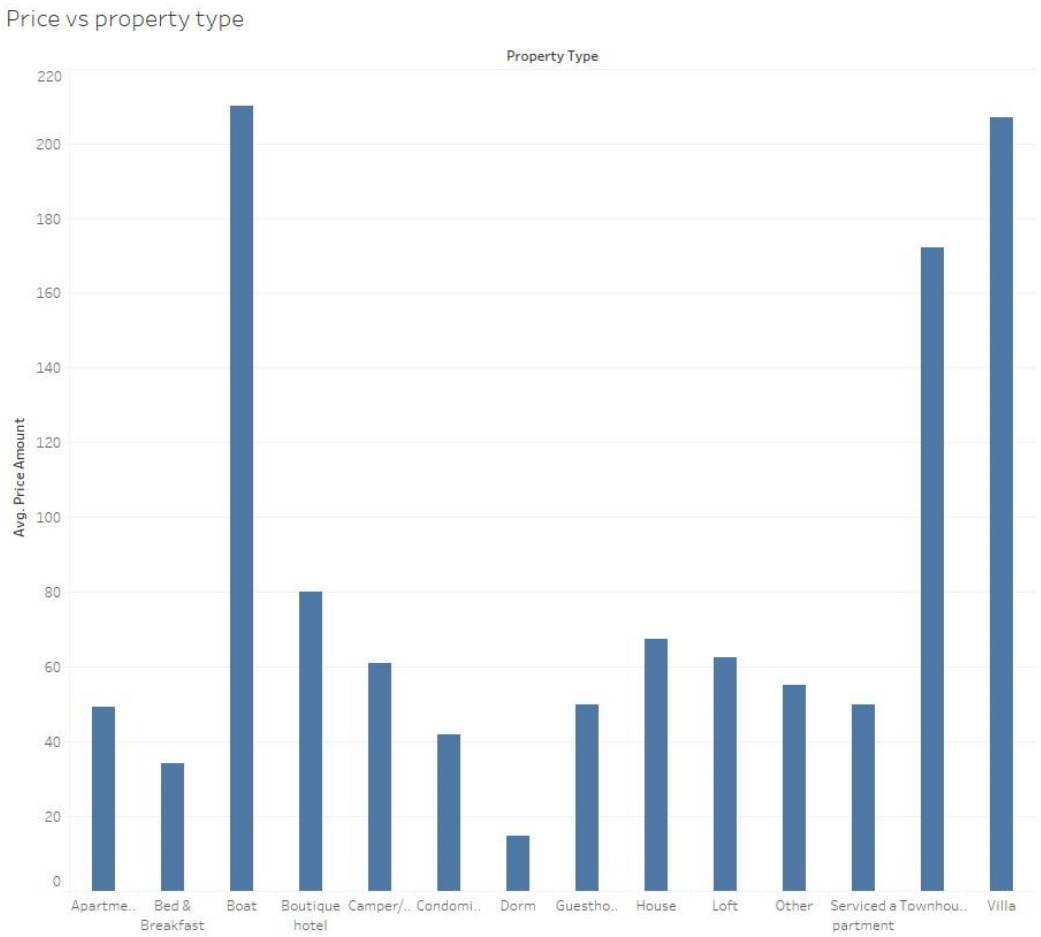
Στην σύγκριση μεταξύ των μεταβλητών τιμή ανά διανυκτέρευση και Host_is_Superhost παρατηρούμε ότι οι οικοδεσπότες που δεν είναι superhosts έχουν χαμηλότερη τιμή στα καταλύματα τους από ότι αυτοί που έχουν την ένδειξη superhost. Συγκεκριμένα, οι απλοί οικοδεσπότες έχουν μέση τιμή στα καταλύματα τους 46.92 ευρώ ενώ οι superhosts έχουν μια αύξηση της τάξης του 30% σε σχέση με τους συμβατικούς οικοδεσπότες. Ανάλογα, την μεγαλύτερη μέση τιμή ανά διανυκτέρευση λαμβάνει το κανονικό κρεβάτι με 51.58 ευρώ ενώ την μικρότερη το απλό στρώμα με 19.44 ευρώ. Επιπλέον, το διάγραμμα δείχνει ότι όσο πιο πρόσφατα έχει εγγραφεί στην πλατφόρμα ένας οικοδεσπότης τόσο χαμηλότερη είναι η τιμή διανυκτέρευσης στην οποία θα προσφέρει το κατάλυμα του. Συγκεκριμένα, οι οικοδεσπότες που είναι εγγεγραμμένοι στην πλατφόρμα από το 2011 διαθέτουν τα καταλύματα τους στην τιμή των 82.14 ευρώ ενώ αυτοί που έγιναν οικοδεσπότες το 2017 ζητούν σχεδόν την μισή τιμή στα 42.45 ευρώ. Τέλος, το διάγραμμα απεικονίζει ποιος τύπος δωματίου έχει την υψηλότερη τιμή στην Αθήνα το 2017. Το ολόκληρο σπίτι έχει μέση τιμή 55.45 ευρώ ενώ το ατομικό δωμάτιο 25.81 ευρώ και το δωμάτιο που μοιράζονται 2 ή περισσότερα άτομα 11.15 ευρώ ανά διανυκτέρευση.



4.14 Διαγράμματα μέσης τιμής καταλυμάτων σε σχέση με τις μεταβλητές Superhost ,τύπο κρεβατιού, τύπο δωματίου



4.15 Διάγραμμα μέσης τιμής καταλύματος και χρόνου οικοδεσπότη

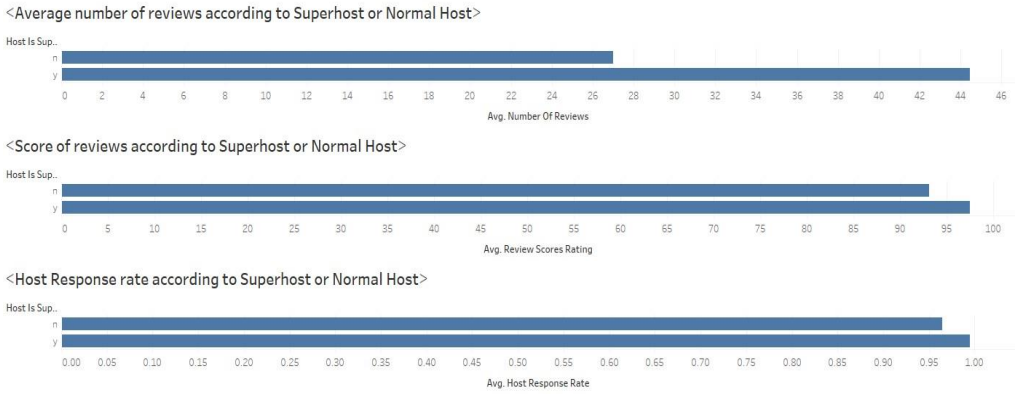


4.16 Διάγραμμα μέσης τιμής καταλύματος και τύπου καταλύματος

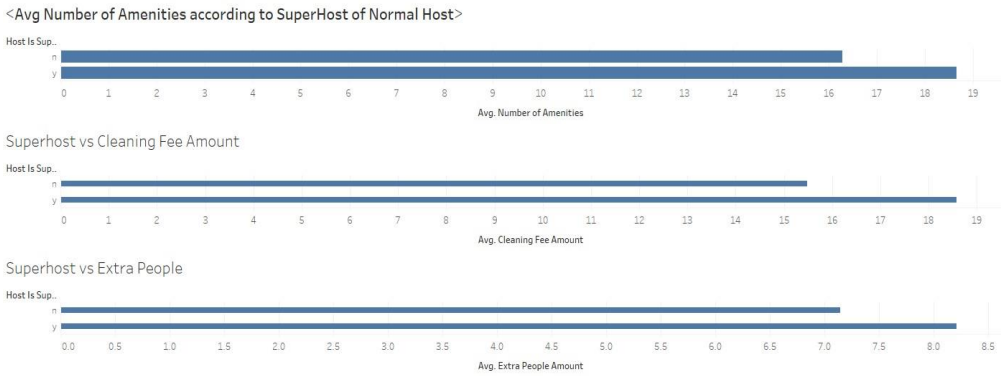
Στο διάγραμμα που απεικονίζει την σχέση μέσης τιμής διανυκτέρευσης με τον τύπο του καταλύματος παρατηρούμε ότι η τιμή της Βάρκας και της Βίλλας ξεχωρίζουν σε σχέση με τον ανταγωνισμό. Συγκεκριμένα, η τιμή της Βάρκας κυμαίνεται στα 210 ευρώ ανά διανυκτέρευση και η τιμή της Βίλλας στα 206.8 ευρώ αντίστοιχα ενώ στο κατώτατο άκρο βρίσκεται ο απλός κοιτώνας με 14.6 ευρώ ανά διανυκτέρευση.

Στο διάγραμμα παρατηρούμε ότι οι οικοδεσπότες που έχουν το σήμα superhost έχουν 44.45 reviews ενώ οι απλοί οικοδεσπότες έχουν 27 reviews. Αντίστοιχα, στο επόμενο διάγραμμα οι οικοδεσπότες με την ένδειξη superhost λαμβάνουν καλύτερη βαθμολογία στις κριτικές των φιλοξενούμενων αφού ο μέσος όρος βαθμολογίας τους είναι 97.44/100 έναντι 93/100 που έχουν οι απλοί οικοδεσπότες. Επιπλέον, οι οικοδεσπότες με την ένδειξη superhost οφείλουν να διατηρούν τον δείκτη απάντησης σε οποιαδήποτε απορία του ενδιαφερόμενου φιλοξενούμενου άνω του 80% για να διατηρήσουν το σήμα και αυτό αποτυπώνεται και στο διάγραμμα όπου οι superhost έχουν 99.5% δείκτη απάντησης ενώ οι απλοί οικοδεσπότες 94.6%. Στο γράφημα παρατηρούμε ότι οι απλοί οικοδεσπότες διαθέτουν μικρότερο αριθμό εξοπλισμού και ανέσεων σε σχέση με τους superhost και συγκεκριμένα ο αριθμός των ανέσεων για τους superhosts ανέρχεται στις 18.6 ενώ για τους απλούς οικοδεσπότες στις 16.2. Όπως θα παρατηρήσουμε και παρακάτω οι επιπλέον ανέσεις και χαρακτηριστικά όπως ο βελτιωμένος χρόνος απόκρισης στα ερωτήματα των φιλοξενουμένων οδηγούν σε μια αύξηση της τιμής των καταλυμάτων για τους superhost επομένως και αυξημένα κέρδη σε σχέση με τους απλούς οικοδεσπότες. Επίσης, αισθητή διαφορά παρουσιάζεται στην μεταβλητή

κόστος καθαρισμού σε σχέση με το αν ο οικοδεσπότης είναι superhost ή όχι. Οι απλοί οικοδεσπότες χρεώνουν κατά μέσο όρο 15.5 ευρώ για τον καθαρισμό του καταλύματος ενώ οι Superhost έχουν την δυνατότητα να αυξήσουν το κόστος στα 18.6 ευρώ. Τέλος, στην μεταβλητή επιπλέον άτομα οι superhosts δέχονται περισσότερα άτομα σε σχέση με τους απλούς οικοδεσπότες και συγκεκριμένα δέχονται 8.2 άτομα επιπλέον στα καταλύματα τους ενώ οι απλοί οικοδεσπότες 7.1 και αυτό το αποτέλεσμα μπορεί να εξηγηθεί εν μέρει από το γεγονός ότι τα καταλύματα που βρίσκονται στην ιδιοκτησία superhost έχουν την δυνατότητα φιλοξενίας περισσότερων ατόμων κατά μέσο όρο σε σχέση με αυτά των συμβατικών Host.



4.17 Διαγράμματα της μεταβλητής Superhost 1



4.18 Διαγράμματα της μεταβλητής Superhost 2

Συσχέτιση

Για να μελετήσουμε την γραμμική σχέση μεταξύ της εξαρτημένης μεταβλητής (price_amount) και των ανεξάρτητων μεταβλητών θα υπολογίσουμε τον συντελεστή συσχέτισης pearson. Από τον παρακάτω πίνακα παρατηρούμε ότι το σύνολο των φυσικών χαρακτηριστικών των καταλυμάτων συσχετίζεται με την τιμή. Συγκεκριμένα, η μεταβλητή Bathrooms παρουσιάζει θετική συσχέτιση 0.603, η μεταβλητή accommodates 0.530, η μεταβλητή bedrooms 0.498 ενώ η μεταβλητή beds 0.381.

Αντίθετα, μεταβλητές όπως οι Availability_60, Availability_90, παρουσιάζουν αρνητική συσχέτιση σε επίπεδο εμπιστοσύνης 95% που μπορεί να εξηγηθεί ως εξής: Όσο μεγαλύτερο είναι το διάστημα διαμονής σε ένα κατάλυμα τόσο μεγαλύτερη είναι η πιθανότητα έκπτωσης της τιμής αφού η πλατφόρμα του Airbnb δίνει την δυνατότητα για εβδομαδιαία και μηνιαία έκπτωση.

Correlation: price_amount, accommodates, bathrooms, bedrooms, beds, availability, ...

	price_amount	accommodates					
0.530	0.000						
bathrooms	0.603	0.512		0.000	0.000		
bedrooms	0.498	0.704	0.561		0.000	0.000	0.000
beds	0.381	0.824	0.444	0.647			
	0.000	0.000	0.000				
0.000							
availability_30	0.022	0.020	0.050	0.046			
	0.183	0.230	0.003				
0.006							
availability_60	-0.039	-0.028	-0.005	-0.001			
	0.018	0.087	0.742				
0.975							
availability_90	-0.059	-0.054	-0.032	-0.027			
	0.000	0.001	0.055				
0.098							
number_of_review	-0.029	0.016	-0.061	-0.069			
	0.076	0.337	0.000				
0.000							
review_scores_ra	0.092	-0.023	0.007	-0.014			
	0.000	0.171	0.655				
0.382							
review_scores_ac	0.060	-0.022	-0.000	-0.019			
	0.000	0.180	0.985				
0.252							
review_scores_ch	0.055	0.004	0.006	-0.012			
	0.001	0.807	0.737				
0.473							
review_scores_co	0.044	-0.012	-0.022	-0.024			
	0.008	0.480	0.186				
0.145							
review_scores_va	0.011	-0.035	-0.028	-0.009			
	0.514	0.032	0.091				

0.576

4.19 Πίνακας συσχετίσεων μεταξύ των μεταβλητών

Row Labels	Average of bathrooms	Average of bedrooms	Average of accommodates	Average of beds
Apartment	1.1	1.3	3.7	2.1
Bed & Breakfast	1.2	1.1	3.1	2.1
Boat	2.0	4.0	8.0	8.0
Boutique hotel	1.5	2.0	6.0	2.0
Camper/RV	1.0	0.0	4.0	2.0
Condominium	1.3	1.9	3.4	2.1
Dorm	1.4	1.0	2.5	2.1
Guesthouse	1.0	0.5	2.0	1.0
House	1.4	1.7	4.1	2.4
Loft	1.1	0.8	3.2	1.7
Other	1.5	1.5	4.2	2.9
Serviced apartment	1.0	1.0	4.0	2.0
Townhouse	1.9	2.1	4.9	3.4
Villa	3.3	3.7	7.5	4.5
Grand Total	**1.2**	**1.4**	**3.7**	**2.1**

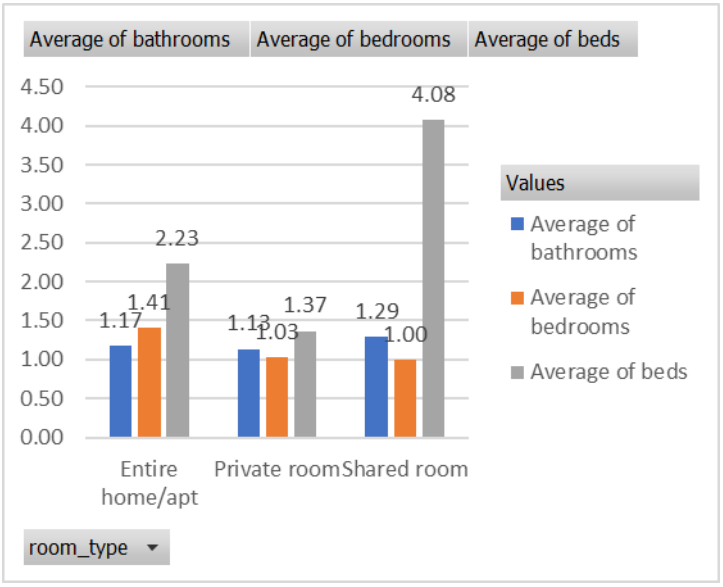
4.2 Πίνακας μπάνιων ,υπνοδωματίων, κρεβατιών, ατόμων που μπορούν να φιλοξενηθούν σε σχέση με τον τύπο του καταλύματος

Στο παραπάνω διάγραμμα παρατηρούμε ότι οι Βίλλες με μεγάλη διαφορά και στην συνέχεια τα ιδιωτικά σκάφη προσφέρουν την δυνατότητα για διαμονή πολλών ατόμων και συγκεκριμένα οι βίλλες διαθέτουν κατά μέσο όρο 3.3 μπάνια , 3.7 υπνοδωμάτια, 4.5 κρεβάτια και μπορούν να φιλοξενήσουν 7.5 άτομα με ακόλουθα τα ιδιωτικά σκάφη όπου διαθέτουν κατά μέσο όρο 2 μπάνια, 4 υπνοδωμάτια,8 κρεβάτια και μπορούν να φιλοξενήσουν 8 άτομα. Αντίθετα, ο τύπος καταλύματος που ενδείκνυται για μικρό αριθμό ατόμων είναι ο ξενώνας ο οποίος διαθέτει κατά μέσο όρο 1 μπάνιο, 0.5 υπνοδωμάτια, 1 κρεβάτι και μπορεί να φιλοξενήσει 2 άτομα.

Row Labels	Average of bathrooms	Average of bedrooms	Average of beds
Entire home/apt	1.17	1.41	2.23
Private room	1.13	1.03	1.37
Shared room	1.29	1.00	4.08
Grand Total	**1.17**	**1.35**	**2.12**

4.3 Πίνακας μπάνιων, υπνοδωματίων και κρεβατιών σε σχέση με τον τύπο δωματίου

Αντιστοίχως , ένα ολόκληρο σπίτι ή διαμέρισμα διαθέτει κατά μέσο όρο 2.23 κρεβάτια, 1.41 υπνοδωμάτια και 1.17 μπάνια και ακολουθεί το ιδιωτικό δωμάτιο που διαθέτει 1.37 κρεβάτια , 1.03 υπνοδωμάτια και 1.13 μπάνια. Τέλος, το κοινό δωμάτιο διαθέτει 4.08 κρεβάτια αφού σε ένα δωμάτιο στεγάζονται πολλά άτομα, έχει 1 υπνοδωμάτιο και ο αριθμός των μπάνιων του είναι 1.29 κατά μέσο όρο.



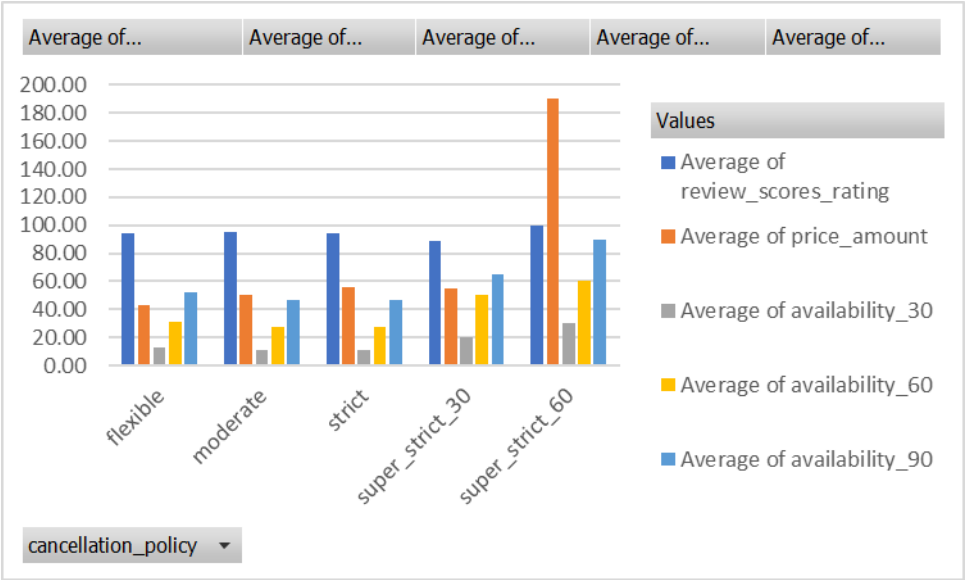
Γράφημα μέσου όρου χαρακτηριστικών

Row Labels	Average of review_scores_rating	Average of price_amount	Average of availability_30	Average of availability_60	Average of availability_90
flexible	94.01	42.84	13.14	31.21	52.04
moderate	94.83	50.58	11.24	27.42	47.12
strict	94.02	56.09	11.22	27.41	47.12
super_strict_30	89.00	55.00	20.00	50.00	65.00
super_strict_60	100.00	190.00	30.00	60.00	90.00
Grand Total	**94.30**	**50.96**	**11.70**	**28.35**	**48.33**

4.5 Πίνακας μέσης τιμής, μέσης τιμής βαθμολογίας, διαθεσιμότητας για 30,60,90 μέρες σε σχέση με την πολιτική ακύρωσης

Στο παραπάνω διάγραμμα αποτυπώνεται η σχέση της ελαστικότητας του οικοδεσπότη σχετικά με το κλείσιμο του καταλύματος σε σχέση με την τιμή του, την βαθμολογία των reviews αλλά και την διαθεσιμότητα του. Παρατηρούμε, ότι όσο πιο αυστηρός είναι ο οικοδεσπότης τόσο αυξάνεται και η τιμή του καταλύματος όπου ο οικοδεσπότης που ακολουθεί χαλαρή πολιτική ακύρωσης έχει μέσης τιμή 42.84 ευρώ ενώ αυτός που ακολουθεί την αυστηρότερη(60 μέρες πριν) 190 ευρώ. Στον τομέα της διαθεσιμότητας ενώ δεν παρατηρούμε μεγάλες διαφορές μεταξύ των flexible, moderate, strict διαπιστώνουμε ότι οι οικοδεσπότες με την πιο αυστηρή πολιτική έχουν μεγαλύτερη διαθεσιμότητα με αποτέλεσμα να μην έχουν τις μέγιστες απολαβές από το κατάλυμα που διαθέτουν στην πλατφόρμα.

Τέλος, σχετικά με το χαρακτηριστικό βαθμολογία των reviews δεν υπάρχει κάποιο συμπέρασμα εκτός ότι τα καταλύματα με την πιο αυστηρή πολιτικής ακύρωσης έχουν την απόλυτη βαθμολογία που αποτελεί ένα λογικό συμπέρασμα αφού αναφερόμαστε σε χώρους με υψηλή τιμή και απαιτητικό οικοδεσπότη που επιλέγει με αυστηρό τρόπο τους καλεσμένους του.



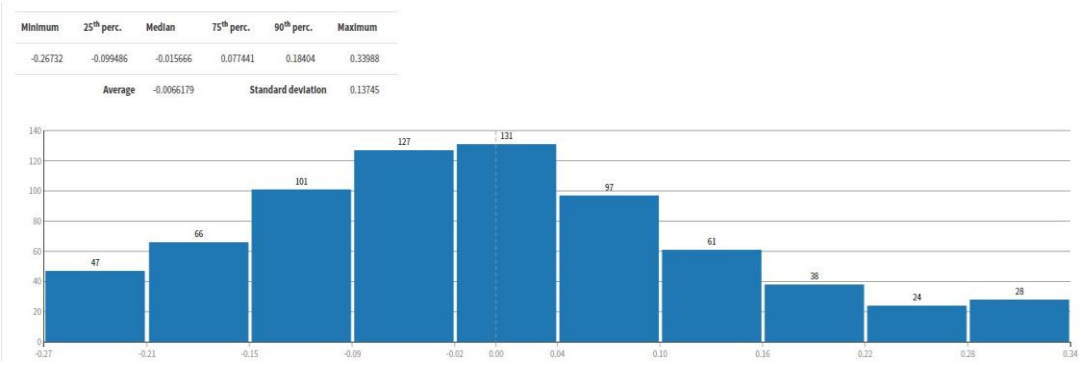
4.2 Αποτελέσματα Μοντελοποίησης Πρόβλεψης Τιμής Καταλυμάτων

Ο πρώτος αλγόριθμος που θα χρησιμοποιήσουμε είναι η απλή γραμμική παλινδρόμηση.

Η συγκεκριμένη μέθοδος όπως παρατηρούμε και στην παρακάτω εικόνα καταφέρνει να εξηγήσει μόνο το 64% της διασποράς των παρατηρήσεων ενώ ο συντελεστής R^2 λαμβάνει την τιμή 0.638 που παρόλο που δεν είναι μια χαμηλή τιμή δεν μπορεί να δώσει τόσο ασφαλή αποτελέσματα. Ο συντελεστής MSE δείχνει την διαφορά μεταξύ των σημείων και της εκτίμησης τους στην γραμμή παλινδρόμησης. Η τιμή του είναι αρκετά χαμηλή (0.021) επομένως μπορούμε να καταλήξουμε ότι το μοντέλο έχει ένα καλό fit στα πραγματικά δεδομένα.Ο συντελεστής RMSLE χρησιμοποιείται όταν δεν θέλουμε να απορρίψουμε την επιρροή μεγάλων διαφορών στις τιμές των μεταβλητών όταν οι μεταβλητές είναι ποσοτικά μεγάλες. Επομένως, από την τιμή 0.056047 συμπεραίνουμε υπάρχει μικρό σφάλμα μεταξύ της εκτίμησης και των πραγματικών τιμών ακόμα και όταν οι τιμές των μεταβλητών λαμβάνουν ποσοτικά υψηλές τιμές. Ο συντελεστής Pearson λαμβάνει την τιμή 0.79975 επομένως και από το συγκεκριμένο μέτρο παρατηρούμε ότι οι πραγματικές τιμές με αυτές που προβλέπει το μοντέλο παλινδρόμησης έχουν ισχυρή θετική συσχέτιση, το οποίο αναμέναμε από το μικρό μέσο τετραγωνικό σφάλμα. Επιπλέον ,παρατηρούμε ότι η κατανομή των σφαλμάτων δεν ακολουθεί ακριβώς την κανονική κατανομή όπου στο δεξί κομμάτι του ιστογράμματος είναι συγκεντρωμένο ένα αρκετά μεγάλο κομμάτι των παρατηρήσεων των σφαλμάτων.

Explained Variance Score Best possible score is 1.0, lower values are worse	0.63951
Mean Absolute Error (MAE) Average of the absolute value of the regression error	0.11356
Mean Average Percentage Error Average of the absolute value of the regression error	7.11%
Mean Squared Error (MSE) Average of the squares of the errors	0.021978
Root Mean Squared Error (RMSE) Root of the above mesure	0.14825
Root Mean Squared Logarithmic Error (RMSLE) Root of the average of the squares of the natural log of the regression error	0.056047
Pearson coefficient Correlation coefficient between actual and predicted values. +1 = perfect correlation, 0 = no correlation, -1 = perfect anti-correlation	0.79975
R2 Score (Coefficient of determination) regression score function	0.63896

4.2.1 Πίνακας με λεπτομερή μέτρα για την μέθοδο της απλής γραμμικής παλινδρόμησης



4.2.2 Κατανομή σφαλμάτων για την μέθοδο της απλής γραμμικής παλινδρόμησης

Ο δεύτερος αλγόριθμος που θα αξιοποιήσουμε είναι μια παραλλαγή της απλής γραμμικής παλινδρόμησης και ονομάζεται Lasso Παλινδρόμηση η οποία βοηθά στην επίλυση των προβλημάτων υπερμοντελοποίησης της απλής γραμμικής παλινδρόμησης με την χρήση του καλύτερου συντελεστή L1 αξιοποιώντας 10 fold cross validation. Αντιστοίχως, στην μέθοδο Ridge Παλινδρόμηση θα χρησιμοποιήσουμε τον καλύτερο συντελεστή L2 αξιοποιώντας και πάλι 10 fold cross validation.

Τόσο η Lasso Παλινδρόμηση όσο και η Ridge Παλινδρόμηση καταφέρνουν να εξηγήσουν το 64% και 60% της διακύμανσης των παρατηρήσεων αντίστοιχα. Ο μέσος όρος των τετραγωνικών σφαλμάτων είναι 0.022 και 0.024 αντίστοιχα ενώ ο συντελεστής R^2 παίρνει τις τιμές 0.638 και 0.605 αντίστοιχα. Η Lasso Παλινδρόμηση έχει ακριβώς τα ίδια αποτελέσματα σε σχέση με την απλή γραμμική παλινδρόμηση ενώ η ridge παλινδρόμηση παρουσιάζει μια μείωση ακρίβειας της τάξης του 0.05 ή

0.5%. Αυτό συμβαίνει διότι η νόρμα L2 που χρησιμοποιείται στην Ridge Παλινδρόμηση μειώνει τα χαρακτηριστικά αφαιρώντας αυτά με χαμηλή σημαντικότητα. Παρόλο, που κάποια χαρακτηριστικά μπορεί να μην είναι τόσο σημαντικά όσο άλλα, προσδίδουν αξία και ακρίβεια στο μοντέλο και δεν είναι συνετό να τα αφαιρέσουμε. Επομένως η μέθοδος ridge Παλινδρόμηση απορρίπτεται.

Explained Variance Score Best possible score is 1.0, lower values are worse	0.63949
Mean Absolute Error (MAE) Average of the absolute value of the regression error	0.11435
Mean Average Percentage Error Average of the absolute value of the regression error	7.17%
Mean Squared Error (MSE) Average of the squares of the errors	0.021990
Root Mean Squared Error (RMSE) Root of the above mesure	0.14829
Root Mean Squared Logarithmic Error (RMSLE) Root of the average of the squares of the natural log of the regression error	0.056074
Pearson coefficient Correlation coefficient between actual and predicted values. +1 = perfect correlation, 0 = no correlation, -1 = perfect anti-correlation	0.80045
R2 Score (Coefficient of determination) regression score function	0.63877

Explained Variance Score Best possible score is 1.0, lower values are worse	0.60675
Mean Absolute Error (MAE) Average of the absolute value of the regression error	0.11872
Mean Average Percentage Error Average of the absolute value of the regression error	7.51%
Mean Squared Error (MSE) Average of the squares of the errors	0.024010
Root Mean Squared Error (RMSE) Root of the above mesure	0.15495
Root Mean Squared Logarithmic Error (RMSLE) Root of the average of the squares of the natural log of the regression error	0.059065
Pearson coefficient Correlation coefficient between actual and predicted values. +1 = perfect correlation, 0 = no correlation, -1 = perfect anti-correlation	0.78059
R2 Score (Coefficient of determination) regression score function	0.60558

4.2.3Πίνακες μέτρων για τις μεθόδους ridge και lasso regression

Στην συνέχεια οι πιο τεχνικές μέθοδοι είχαν καλύτερα αποτελέσματα σε σχέση με την απλή παλινδρόμηση και τις παραλλαγές της. Οι μέθοδοι είναι Random Forest, Gradient Tree Boosting και XG Boost. Στον αλγόριθμο Random Forest έγινε χρήση 100 δέντρων,σε κάθε διαχωρισμό για το δείγμα χρησιμοποιήθηκε το 30% των χαρακτηριστικών, κάθε δέντρο είχε μέγιστο βάθος 27 με στόχο να μεγιστοποιήσουμε την ποιότητα της πρόβλεψης. Στην μέθοδο Gradient Tree Boosting έγινε χρήση 100 σταδίων boosting ώστε να επιτευχθεί καλύτερη απόδοση του μοντέλου, ρυθμός μάθησης 0.1 και μέγιστο βάθος δέντρου 3. Τέλος, στον αλγόριθμο XGBoost ο οποίος αποτελεί μια πιο πολύπλοκη μορφή του αλγορίθμου Gradient Tree Boosting όπου ο

μέγιστος αριθμός δέντρων καθορίστηκε στα 300,το μέγιστο βάθος κάθε δέντρου ήταν 3, ο ρυθμός μάθησης 0.2, επίσης χρησιμοποιήθηκε ο συντελεστής L2 σε μια προσπάθεια να μειωθεί το overfitting και να μην αυξηθεί ο χρόνος ολοκλήρωσης.

Στα αποτελέσματα η μέθοδος Random Forest παρείχε τα χειρότερα αποτελέσματα από τις 3 και συγκεκριμένα το δείκτης επεξήγησης της διασποράς ήταν 0.684,το μέσο τετραγωνικό σφάλμα 0.019 ενώ ο συντελεστής R^2 είχε τιμή 0.683. Η μέθοδος Gradient Tree Boosting είχε καλύτερη ακρίβεια και κατάφερε να επιτύχει καλύτερα αποτελέσματα. Ο δείκτης επεξήγησης της διασποράς ήταν 0.7, το μέσο τετραγωνικό σφάλμα 0.018 και ο συντελεστής R^2 είχε τιμή 0.7.

Explained Variance Score Best possible score is 1.0, lower values are worse	0.68453
Mean Absolute Error (MAE) Average of the absolute value of the regression error	0.10701
Mean Average Percentage Error Average of the absolute value of the regression error	6.72%
Mean Squared Error (MSE) Average of the squares of the errors	0.019268
Root Mean Squared Error (RMSE) Root of the above mesure	0.13881
Root Mean Squared Logarithmic Error (RMSLE) Root of the average of the squares of the natural log of the regression error	0.052464
Pearson coefficient Correlation coefficient between actual and predicted values. +1 = perfect correlation, 0 = no correlation, -1 = perfect anti-correlation	0.82940
R2 Score (Coefficient of determination) regression score function	0.68349

4.2.4Πίνακας μέτρων για την μέθοδο Random Forest

Explained Variance Score Best possible score is 1.0, lower values are worse	0.70085
Mean Absolute Error (MAE) Average of the absolute value of the regression error	0.10559
Mean Average Percentage Error Average of the absolute value of the regression error	6.61%
Mean Squared Error (MSE) Average of the squares of the errors	0.018239
Root Mean Squared Error (RMSE) Root of the above mesure	0.13505
Root Mean Squared Logarithmic Error (RMSLE) Root of the average of the squares of the natural log of the regression error	0.050914
Pearson coefficient Correlation coefficient between actual and predicted values. +1 = perfect correlation, 0 = no correlation, -1 = perfect anti-correlation	0.83825
R2 Score (Coefficient of determination) regression score function	0.70037

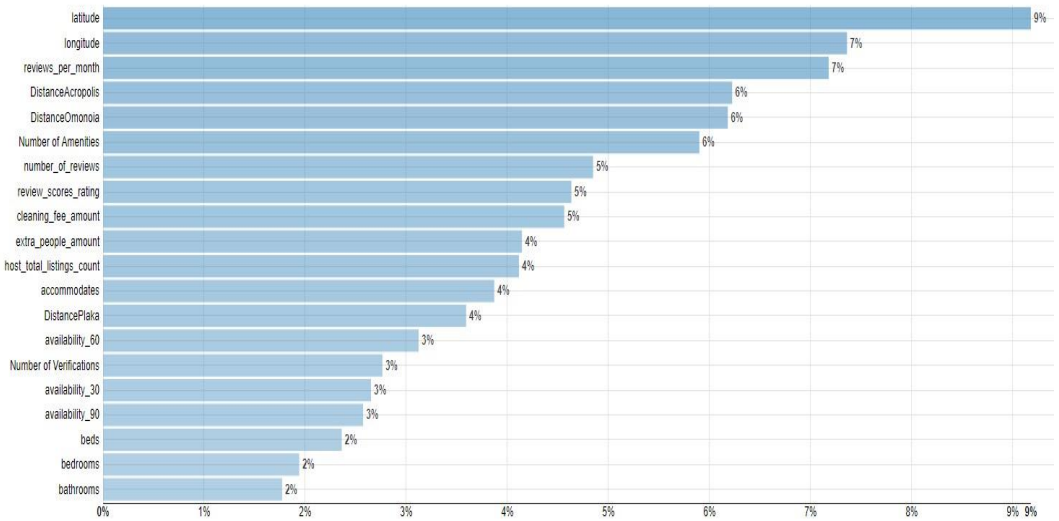
4.2.5Πίνακας μέτρων για την μέθοδο Gradient Tree Boosting

Ο αλγόριθμος XGBoost όπως αναμένονταν παρείχε τα καλύτερα αποτελέσματα από όλες τις μεθόδους που αναλύθηκαν παραπάνω. Επομένως θα χρησιμοποιηθεί και ως το βασικό μοντέλο πρόβλεψης της τιμής καταλυμάτων στο Airbnb. Ο δείκτης επεξήγησης της διασποράς ήταν 0.709 , το μέσο τετραγωνικό σφάλμα ήταν 0.017 ενώ ο συντελεστής R^2 ήταν 0.709 .

Explained Variance Score Best possible score is 1.0, lower values are worse	0.70952
Mean Absolute Error (MAE) Average of the absolute value of the regression error	0.10354
Mean Average Percentage Error Average of the absolute value of the regression error	6.48%
Mean Squared Error (MSE) Average of the squares of the errors	0.017721
Root Mean Squared Error (RMSE) Root of the above mesure	0.13312
Root Mean Squared Logarithmic Error (RMSLE) Root of the average of the squares of the natural log of the regression error	0.050245
Pearson coefficient Correlation coefficient between actual and predicted values. +1 = perfect correlation, 0 = no correlation, -1 = perfect anti-correlation	0.84291
R2 Score (Coefficient of determination) regression score function	0.70890

4.2.6Πίνακας μέτρων για τον αλγόριθμο XGboost

Οι μεταβλητές που επηρεάζουν την τιμή ενός καταλύματος στο Airbnb όπως παρουσιάζονται στον παρακάτω πίνακα είναι οι latitude με ποσοστό 9%, longitude και reviews_per_month με ποσοστό 7%,DistanceAcropolis,DistanceOmonoia και Number of Amenities με ποσοστό 6%,number_of_reviews,review_scores_rating και cleaning_fee_amount με ποσοστό 5%,extra_people_amount
, Host_total_listings_count, accommodates, DistancePlaka με ποσοστό 4%, availability_60, Number of Verifications, availability_30, availability_90 με ποσοστό 3% και τέλος beds, bedrooms, bathrooms με ποσοστό 2%.
Τα παραπάνω νούμερα υποδεικνύουν ότι η απόσταση από συγκεκριμένους προορισμούς όπως η Ακρόπολη και η Ομόνοια αλλά και η γενικότερη θέση του καταλύματος είναι καθοριστικοί παράγοντες για την διαμόρφωση της τιμής του. Το ίδιο ισχύει και για τις κριτικές που έχει το κάθε κατάλυμα αφού όσο περισσότερες κριτικές αλλά και καλύτερη βαθμολογία τόσο πιο πιθανό είναι να είναι περιζήτητό.



4.2.7Διάγραμμα σημαντικότητας μεταβλητών για την πρόβλεψη τιμής καταλύματος

4.3 Αποτελέσματα μοντελοποίησης μεταβλητής Superhost

Η πρώτη μέθοδος που χρησιμοποιήθηκε ήταν η Λογιστική Παλινδρόμηση για το πρόβλημα ταξινόμησης. Επειδή ο απλός αλγόριθμος της λογιστικής παλινδρόμησης είναι ευαίσθητος σε λάθη και υπάρχει μια μικρή πιθανότητα overfitting ένα όριο κανονικοποίησης στα βάρη των χαρακτηριστικών. Από την ανάλυση της Λογιστικής Παλινδρόμησης προκύπτει ένα σχετικά ικανοποιητικό μοντέλο με συντελεστή ROC-AUC 0.791, επίσης το επίπεδο Accuracy κυμαίνεται στο 0.743 που δείχνει το ποσοστό των σωστών προβλέψεων είτε αυτές ήταν θετικές είτε αρνητικές και θεωρείται ικανοποιητικό ενώ αντίθετα το επίπεδο Precision λαμβάνει την τιμή 0.552 και αντικατοπτρίζει το ποσοστό των θετικών προβλέψεων που ήταν όντως θετικές.

Threshold-dependent (current threshold = `0.3500`)	
Accuracy Proportion of correct predictions (positive and negative) in the test set	0.7431
Precision Proportion of positive predictions that were indeed positive (in the test set)	0.5529
Recall Proportion of actual positive values found by the classifier	0.6651
F1 Score Harmonic mean between Precision and Recall	0.6039
Hamming loss Fraction of labels that are incorrectly predicted (the lower the better)	0.2569
Matthews Correlation Coefficient Correlation coefficient between actual and predicted values. +1 = perfect, 0 = no correlation, -1 = perfect anti-correlation	0.4200
Threshold-independent	
Log loss Error metric that takes into account the predicted probabilities (the lower the better)	0.4908
ROC - AUC Score Area under the ROC; from 0.5 (random model) to 1 (perfect model)	0.7912

4.3.1Πίνακας μέτρων για την μέθοδο λογιστική παλινδρόμηση

Η δεύτερη μέθοδος που θα χρησιμοποιηθεί για την μοντελοποίηση της μεταβλητής Superhost στους οικοδεσπότες είναι το Support Vector Machine. Είναι αρκετά αποτελεσματική μέθοδος σε μελέτες με πολλά χαρακτηριστικά όπως η συγκεκριμένη. Από την ανάλυση του Support Vector Machine τα αποτελέσματα είναι βελτιωμένα σε σχέση με την λογιστική παλινδρόμηση και συγκεκριμένα η τιμή του συντελεστή ROC AUC είναι 0.841,το επίπεδο Accuracy λαμβάνει την τιμή 0.745 και είναι ίδιο με αυτό της λογιστικής παλινδρόμησης ενώ ο συντελεστής Precision λαμβάνει την τιμή 0.546 η οποία είναι ελάχιστα χαμηλότερη από αυτή της λογιστικής παλινδρόμησης. Παρόλο, που το συγκεκριμένο μοντέλο επιτυγχάνει μεγαλύτερο ROC-AUC συντελεστή, συνεχίζει να παρουσιάζει τα προβλήματα που έχει και η προηγούμενη μέθοδος όπως χαμηλό συντελεστή Precision.

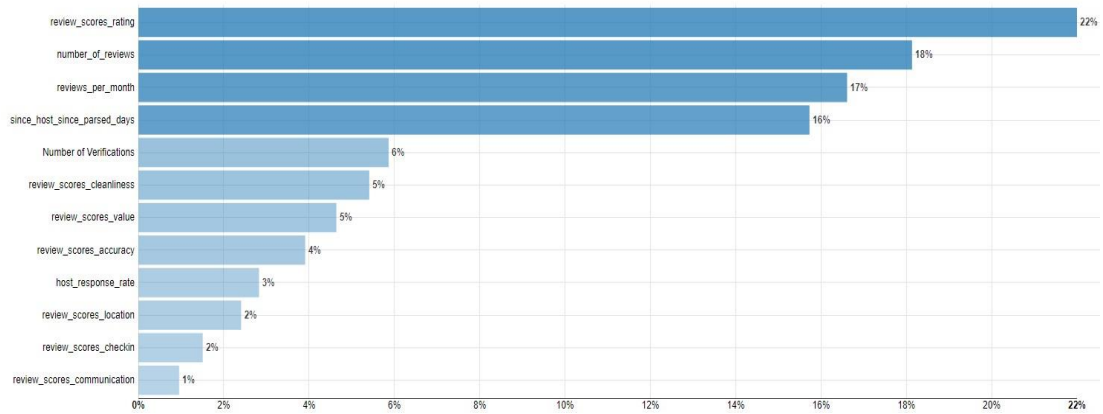
Threshold-dependent (current threshold = `0.2000`)	
Accuracy Proportion of correct predictions (positive and negative) in the test set	0.7458
Precision Proportion of positive predictions that were indeed positive (in the test set)	0.5460
Recall Proportion of actual positive values found by the classifier	0.8113
F1 Score Harmonic mean between Precision and Recall	0.6528
Hamming loss Fraction of labels that are incorrectly predicted (the lower the better)	0.2542
Matthews Correlation Coefficient Correlation coefficient between actual and predicted values. +1 = perfect, 0 = no correlation, -1 = perfect anti-correlation	0.4868
Threshold-independent	
Log loss Error metric that takes into account the predicted probabilities (the lower the better)	0.4405
ROC - AUC Score Area under the ROC; from 0.5 (random model) to 1 (perfect model)	0.8414

4.3.2Πίνακας μέτρων για την μέθοδο SVM

Η τελευταία και πιο αποτελεσματική μέθοδος που χρησιμοποιήθηκε ήταν ο αλγόριθμος Random Forest. Η ταξινόμησή μέσω Τυχαίου Δάσους πραγματοποιείται μέσω πολλών δέντρων αποφάσεων όπου κάθε δέντρο απόφασης συμμετέχει στην διαδικασία και στο τέλος της διαδικασίας επιλέγεται η τάξη που έχει «ψηφιστεί» από τα περισσότερα δέντρα. Ο αριθμός των δέντρων στον συγκεκριμένο αλγόριθμο ήταν 100 και το μέγιστο βάθος κάθε δέντρου ήταν 15. Ο συγκεκριμένος αλγόριθμος έδωσε και τα καλύτερα αποτελέσματα με τον συντελεστή ROC-AUC να λαμβάνει τιμή 0.877,τον συντελεστή Accuracy 0.83 και τον συντελεστή Precision 0.695.

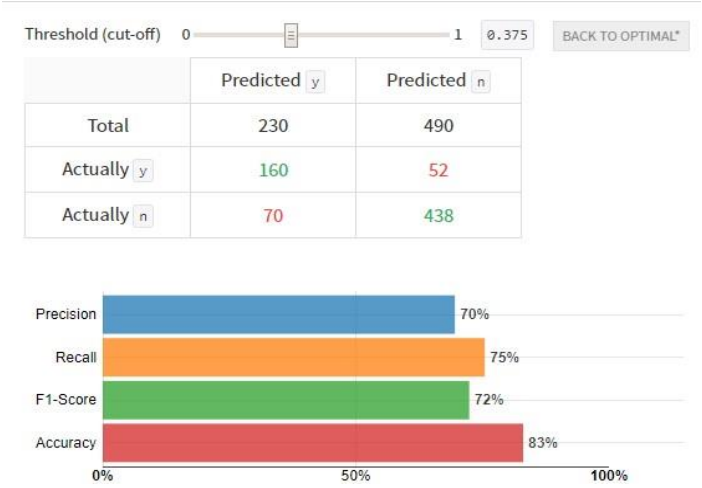
Τα χαρακτηριστικά που επηρεάζουν την επιλογή του σήματος Superhost είναι η βαθμολογία των κριτικών που έχει ένας οικοδεσπότης με ποσοστό 22%,ο αριθμός των κριτικών με ποσοστό 18%, ακολουθούν οι κριτικές ανά μήνα με ποσοστό 17%, το χρονικό διάστημα για το οποίο ο οικοδεσπότης διαθέτει το κατάλυμα του στο Airbnb με ποσοστό 16% . Στην συνέχεια τα λιγότερο σημαντικά χαρακτηριστικά είναι ο αριθμός των πιστοποιήσεων με ποσοστό 6%, η βαθμολογία της καθαριότητας όπως και το αν το κατάλυμα ανταποκρίνεται στην τιμή του με ποσοστό 5%,η ακρίβεια με 4%, η ανταπόκριση του ιδιοκτήτη σε τυχόν απορίες με ποσοστό 3% ,η ακρίβεια της τοποθεσίας με ποσοστό 2% και τέλος η βαθμολογία του Checkin αλλά και πόσο εύκολη είναι η επικοινωνία με τον οικοδεσπότη με ποσοστά 2% και 1% αντίστοιχα.

Threshold-dependent (current threshold = `0.3750`)	
Accuracy Proportion of correct predictions (positive and negative) in the test set	0.8306
Precision Proportion of positive predictions that were indeed positive (in the test set)	0.6957
Recall Proportion of actual positive values found by the classifier	0.7547
F1 Score Harmonic mean between Precision and Recall	0.7240
Hamming loss Fraction of labels that are incorrectly predicted (the lower the better)	0.1694
Matthews Correlation Coefficient Correlation coefficient between actual and predicted values. +1 = perfect, 0 = no correlation, -1 = perfect anti-correlation	0.6031
Threshold-independent	
Log loss Error metric that takes into account the predicted probabilities (the lower the better)	0.4289
ROC - AUC Score Area under the ROC; from 0.5 (random model) to 1 (perfect model)	0.8771



4.3.3Διάγραμμα σημαντικότητας μεταβλητών για την πρόβλεψη της ένδειξης Superhost

Κατά την ταξινόμηση δημιουργείται η πιθανότητα ένα αντικείμενο να ανήκει σε μια συγκεκριμένη κλάση όπως στην περίπτωση μας όπου για την μεταβλητή host_is_superhost η οποία ανήκει στην κλάση y. Το όριο cut-off είναι ο αριθμός μετά τον οποίο η πρόβλεψη θεωρείται θετική. Εάν του δώσουμε υψηλή τιμή θα έχει ως αποτέλεσμα να προβλέπει ότι ο οικοδεσπότης είναι όντως superhost ενώ αν του δώσουμε χαμηλή τιμή θα κάνει την συγκεκριμένη πρόβλεψη πιο συχνά από όσο πρέπει. Στην συγκεκριμένη περίπτωση προσαρμόστηκε το όριο ώστε να μεγιστοποιείται το F1 Score το οποίο αποτελεί τον αρμονικό μέσο μεταξύ του precision και accuracy και αποτελεί ιδανική λύση όταν το σύνολο των δεδομένων δεν είναι ισορροπημένο . Στον παρακάτω πίνακα στο δοκιμαστικό dataset με 720 παρατηρήσεις προέβλεψε ότι οι 230 θα ήταν superhost ενώ στην πραγματικότητα ήταν 160 και 490 ότι ήταν απλοί οικοδεσπότες ενώ ήταν 438. Αυτή η απόκλιση στην πρόβλεψη και ειδικότερα στην πρόβλεψη του superhost οφείλεται στο γεγονός ότι στα δεδομένα μας το ποσοστό των superhost επί του συνόλου ήταν αρκετά μικρό.



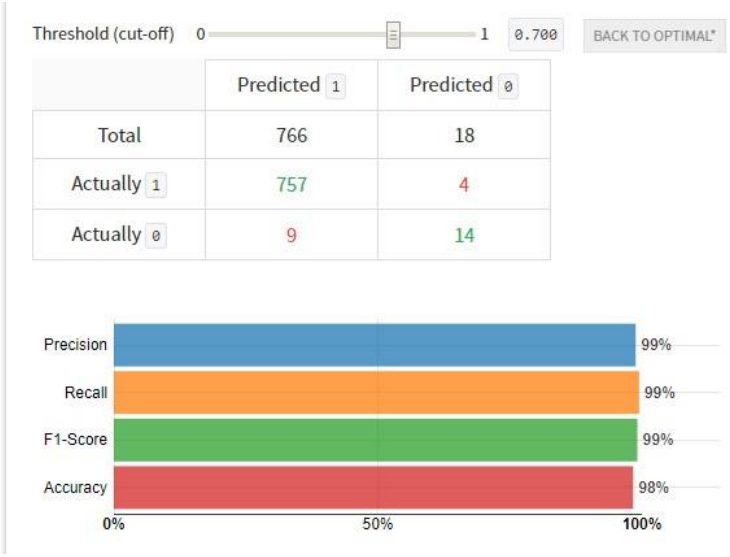
4.3.4Confusion Matrix για την μέθοδο Random Forest

4.4 Αποτελέσματα εξόρυξης γνώμης για τα καταλύματα της Αθήνας που έχουν εισαχθεί στην πλατφόρμα του Airbnb

Ο συντελεστής ROC-AUC λαμβάνει την τιμή 0,924 η οποία βρίσκεται πολύ κοντά στην μονάδα και δείχνει ότι το ποσοστό των σωστών θετικών προβλέψεων σε σχέση με τις λάθος είναι αρκετά μεγαλύτερο. Αντίστοιχα τα μεγέθη Accuracy, Precision και F1 Score λαμβάνουν σχεδόν άριστες τιμές για την έρευνα μας και δείχνουν ότι το μοντέλο μας έχει την ικανότητα να ξεχωρίζει τις λέξεις που αναφέρονται σε αρνητικές κριτικές και αυτές που αναφέρονται σε θετικές με σημαντική επιτυχία. Η ακρίβεια του μοντέλου φαίνεται και από την εικόνα όπου παρουσιάζεται το confusion matrix. Στον συγκεκριμένο πίνακα παρατηρούμε ότι στο test set από τις 766 θετικές κριτικές το μοντέλο προέβλεψε τις 757 ως θετικές ενώ μόνο 9 ως αρνητικές. Από τις 18 αρνητικές κριτικές προέβλεψε σωστά ως αρνητικές τις 14 ενώ 4 λανθασμένα ως θετικές. Το μεγαλύτερο σφάλμα που παρατηρείται στην πρόβλεψη των αρνητικών κριτικών οφείλεται όπως προαναφέρθηκε στην έλλειψη μεγάλου πλήθους αρνητικών κριτικών στα δεδομένα λόγω της πολιτικής της πλατφόρμας του Airbnb.

Accuracy Proportion of correct predictions (positive and negative) in the test set	0.9834
Precision Proportion of positive predictions that were indeed positive (in the test set)	0.9883
Recall Proportion of actual positive values found by the classifier	0.9947
F1 Score Harmonic mean between Precision and Recall	0.9915
Hamming loss Fraction of labels that are incorrectly predicted (the lower the better)	0.0166
Matthews Correlation Coefficient Correlation coefficient between actual and predicted values. +1 = perfect, 0 = no correlation, -1 = perfect anti-correlation	0.6799
Threshold-independent	
Log loss Error metric that takes into account the predicted probabilities (the lower the better)	0.0782
ROC - AUC Score Area under the ROC; from 0.5 (random model) to 1 (perfect model)	0.9242

4.4.1Πίνακας μέτρων για την μέθοδο Λογιστική παλινδρόμηση



4.4.2Confusion Matrix για την μέθοδο Λογιστική Παλινδρόμηση

Από την χρήση του μοντέλου της λογιστικής παλινδρόμησης παρατηρούμε ότι το μοντέλο μας θεωρεί ως αρνητικές μεταβλητές ενδεικτικά τις λέξεις “bad”, “not” , “poor” , “without” , “terrible”, “problems” με συντελεστές λογιστικής παλινδρόμησης -1,57 , -1,46 , -1,1 , -1,08 , -1,06 , -1,03 ενώ ως θετικές μεταβλητές τις λέξεις ”great” , “nice” , “very”,“perfect”, “amazing” με συντελεστές 1,51 , 1,25 , 0,93 , 0,85 , 0,68 αντίστοιχα.

Variable	Coefficient
comments contains bad	-1.5774
comments contains great	1.5119
comments contains not	-1.4684
comments contains nice	1.2505
comments contains poor	-1.1013
comments contains without	-1.0813
comments contains terrible	-1.0626
comments contains problems	-1.0378
comments contains wifi	-0.9686
comments contains dirty	-0.9483
comments contains very	0.9336
comments contains perfect	0.8502
comments contains la	0.8366
comments contains didn	-0.8129
comments contains noise	-0.7593
comments contains tres	0.7123
comments contains little	-0.6856
comments contains amazing	0.6818
comments contains metro	0.6797

4.4.3Σημαντικότητα μεταβλητών κατά την αξιολόγηση κριτικών

5. Συμπεράσματα

Σε αυτήν την εργασία πραγματοποιήθηκαν 3 τμήματα πρόβλεψης. Αρχικά, μελετήθηκαν τα χαρακτηριστικά που επηρεάζουν την τιμή των καταλυμάτων στην πλατφόρμα του Airbnb στην περιοχή της Αθήνας. Στη συνέχεια, μελετήθηκαν τα χαρακτηριστικά που μετατρέπουν έναν απλό οικοδεσπότη σε superhost. Τέλος, πραγματοποιήθηκε εξόρυξη γνώσης από τις κριτικές των καταλυμάτων στην περιοχή της Αθήνας.

Για την μοντελοποίηση των τιμών μελετήθηκαν αρκετά διαφορετικά μοντέλα όπως η απλή παλινδρόμηση, τα τυχαία δάση αλλά οι μέθοδοι Gradient Boosting και XGBoost μας έδωσαν τα καλύτερα και πιο ρεαλιστικά αποτελέσματα.

Για να διαπιστώσουμε ποια χαρακτηριστικά σχετίζονται με το αν ένας οικοδεσπότης έχει την ικανότητα να γίνει superhost, χρησιμοποιήσαμε ανάλυση δεδομένων και βρήκαμε ότι ο αριθμός των κριτικών ,η τοποθεσία, η επικοινωνία κτλ. ήταν κάποια από τα κύρια χαρακτηριστικά που διαχώριζαν έναν απλό οικοδεσπότη από έναν superhost. Στην συνέχεια πάλι με την χρήση μαθηματικών μεθόδων πρόβλεψης όπως τα τυχαία δάση και XGBoost καταφέραμε να επιβεβαιώσουμε τα ευρήματα της προηγούμενης ανάλυσης.

Στο τρίτο και τελευταίο τμήμα της μελέτης διαπιστώσαμε ποιες λέξεις στις κριτικές που έκαναν οι πελάτες είχαν θετικό ή αρνητικό αντίκτυπο με αποτέλεσμα να δημιουργηθεί ένα μοντέλο που θα μπορούσε να προβλέψει ανάλογες συμπεριφορές στο μέλλον. Λέξεις όπως “great”, “nice” θεωρήθηκαν από το μοντέλο μας σαν θετικοί παράμετροι ενώ αντίθετα λέξεις όπως “bad”,”dirty” θεωρήθηκαν ως αρνητικοί παράμετροι.

Συμπερασματικά λοιπόν μπορούμε να καταλήξουμε ότι οι τεχνικές μηχανικής μάθησης που χρησιμοποιήσαμε δώσανε μια ικανοποιητική απόδοση. Τα αποτελέσματα της έρευνας είναι άκρως ενθαρρυντικά και αποδεικνύουν ότι το μοντέλο αποδίδει και έτσι μπορεί να χρησιμοποιηθεί ουσιαστικά πάνω στη πρόβλεψη της τιμής των καταλυμάτων αλλά και στους παράγοντες που καθορίζουν αν κάποιος οικοδεσπότης θα είναι superhost ή όχι . Επομένως, μπορούμε να καταλήξουμε ότι η χρήση των συγκεκριμένων μοντέλων μπορεί να επεκταθεί και στον εμπορικό κλάδο και να χρησιμοποιηθεί άμεσα είτε από ιδιώτες είτε από επιχειρήσεις που δραστηριοποιούνται στον τομέα του τουρισμού.

6. Μελλοντικές κατευθύνσεις - Επίλογος

Στο μελλοντικό έργο, ο τελικός στόχος μας είναι να δημιουργηθεί ένα σύστημα πληροφόρησης για τους χρήστες Airbnb, στο οποίο ο χρήστης θα εισάγει το ποσό που θα είναι διατεθειμένος να ξοδέψει και το σύστημα θα προσφέρει μια καλή επιλογή. Για να επιτευχθεί αυτός ο στόχος, μπορεί να χρησιμοποιηθεί ένας συνδυασμός της εξόρυξης γνώσης μέσω των κριτικών μαζί με την πρόβλεψη των τιμών/superhost, για να εξασφαλιστούν καλύτερα αποτελέσματα.

Τέλος, παρατηρήθηκε ότι οι τιμές των καταλυμάτων διαφέρουν αρκετά ανάλογα με την εκάστοτε γειτονιά και θα ήταν ενδιαφέρον να κατανοηθούν σε βάθος οι συνδιακυμάνσεις μεταξύ των τιμών και των περιοχών αφού θα μπορούσε να υπάρξει άμεση εφαρμογή και σε άλλους κλάδους όπως το real estate ή σε κάποιο πολεοδομικό σχεδιασμό της πόλης.

Βιβλιογραφία

Άρθρα σε περιοδικά

Brown, M., 2016. The Making of Airbnb. *Boston Hospitality Review*, 4(1), pp. 33-38.

Cai, T., Han, K., & Wu, H. (2019). Melbourne Airbnb Price Prediction. 6, pp.1-12.

Chen, T. & Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System. *KKD*, pp.8-11.

Choudhary, P., Jain, A., & Baijal, R. (2018). Unravelling AirbnbPredicting Price for New Listing. 10, pp.1-18.

Juul, M., 2017. Tourism and the sharing economy. *Members' Research Service*, January,4, p. 10.

Kalehbasti, P. R., Nikolenko, L., & Rezaei, H. (2019). Airbnb Price Prediction Using Machine Learning and Sentiment Analysis. 7, pp.1-8.

Keating, J., Katnic, E., Hahn, C., & Yang, R. (2019). PREDICTIVE MODELING ON AIRBNB LISTING PRICES. 7, pp.1-7.

Padmaja, S. & S Sameen, F., 2013. Opinion Mining and Sentiment Analysis –An Assessment of Peoples' Belief: A Survey. *International Journal of Ad hoc, Sensor & Ubiquitous Computing,* 4(1), p. 33.

Pang, B. & Lee, L., 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval,* 2(1), pp. 1-135.

Park, H.-A., 2013. An Introduction to Logistic Regression: From Basic Concepts to Interpretation with Particular Attention to Nursing Domain. *J Korean Accad Nurs,* 43(2), pp. 154-164.

Singh, M., Choudhury, S., Banerjee, R., & Manniste, A. (2018). Airbnb New York City: Demystifying the Superhost Program. 7, p.9.

SMEUREANU, I. & BUCUR, C., 2012. Applying Supervised Opinion Mining Techniques on Online User Reviews. *Informatica Economică,* 16(2), pp. 81-91.

Tang, E., & Sangani, K. (2015). Neighborhood and Price Prediction for San Francisco Airbnb Listings. 5, pp.1-6.

Wang, D., & Nicolau, J. (2017). Price Determinants of Sharing Economy Based Accommodation Rental: A Study of Listings from 33 Cities on Airbnb.com. Journal of Hospitality Management, 38, pp.13.

Zervas, G., Proserpio, D. & Byers, J. W., 2015. A First Look at Online Reputation on Airbnb,1, pp.3-7.

Βιβλία

Elizaveta, R., 2016. Peer-to-peer as a travel accommodation option and the customer value. Saimaa University of Applied Sciences.

James, G., Witten, D., Hastie, T. & Tibshirani, R., 2017. *An Introduction to Statistical Learning.* 8th επιμ. New York: Springer.

Schapire, R., 2008. *COS 511: Theoretical Machine Learning,* Princeton: Princeton University.

Shalev-Shwartz, S. & Shai, B.-D., 2014. *Understanding Machine Learning: From Theory to Algorithms.* 1st επιμ. Cambridge: Cambridge University Press.

Νικόλαος, Δ., 2014. Γλωσσολογικές πηγές για τεχνικές εξόρυξης γνώμης (opinion mining) προσαρμοσμένες στις ιδιαιτερότητες της Νέας Ελληνικής, Πατρα: ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ.

ΕΛΕΝΗ, Ζ., 2012. «ΑΝΑΓΝΩΡΙΣΗ ΚΑΙ ΚΑΤΑΤΑΞΗ ΟΝΟΜΑΤΩΝ ΟΝΤΟΤΗΤΩΝ ΣΕ ΕΛΛΗΝΙΚΑ ΚΕΙΜΕΝΑ ΜΕ ΧΡΗΣΗ ΤΥΧΑΙΩΝ ΔΑΣΩΝ, Πάτρα: Πανεπιστήμιο Πατρών.

Γιάννης, Θ., 2017. Μελέτη και αξιολόγηση τεχνικών εξόρυξης πολιτικής γνώμης σε tweets. Πάτρα: ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ.

Αλέξανδρος, Κ., 2014. *Ανάλυση λογιστικής Παλινδρόμησης σε δεδομένα από έρευνα αγοράς,* s.l.: ΤΕΙ Ανατολικής Μακεδονίας-Θράκης.

Δημήτριος, Μ., 2017. *Εξόρυξη Γνώμης από Σχόλια Χρηστών στο Twitter σε Πραγματικό Χρόνο.* Θεσσαλονίκη: ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ.

Αναφορές στο Διαδίκτυο

Airbnb, 2017. *Airbnb.*
Ανάκτηση: https://www.airbnb.gr/help/article/828/what-is-a-superhost

Airbnb, T., χ.χ. *The Airbnb Community Compact.*
Ανάκτηση: https://www.airbnbaction.com/wp-content/uploads/2015/11/AirbnbCommunity-Compact.pdf
Brownlee, J., 2017. *A Gentle Introduction to XGBoost for Applied Machine Learning.*
Ανάκτηση: https://machinelearningmastery.com/gentle-introduction-xgboost-appliedmachine-learning/

Breiman, L. & Cutler, A., 2014. Random forests for beginners, s.l.: Salford Systems.
Ανάκτηση : https://gitee.com/it-ebooks/it-ebooks-2018-11to12/raw/master/RANDOM%20FORESTS%20FOR%20BEGINNERS.pdf

Economist, L., 2013. *The rise of the sharing economy.*
Ανάκτηση: https://www.economist.com/leaders/2013/03/09/the-rise-of-the-sharingeconomy

Francis, A., Yamijala, R., Thangudu, J. K. & Adhikary, P., 2016. *The Sharing Economy: Implications for Property &Casualty Insurers.*

Ανάκτηση: https://www.cognizant.com/whitepapers/The-Sharing-Economy-Implicationsfor-Property-and-Casualty-Insurers-codex1820.pdf

Health, O., 2016. *Introduction to Machine Learning in Healthcare*.

Ανάκτηση από : http://web.Orion health.com/rs/981-HEV-035/images/Introduction_to_Machine_Learning.pdf

Insider.gr. (2018, July 8). Πώς διαμορφώνεται ο χάρτης της Airbnb στην Ελλάδα. Ανάκτηση από Insider.gr: https://www.insider.gr/epiheiriseis/toyrismos/88791/pos-diamorfonetai-o-hartis-tis-airbnb-stin-ellada

JAIN, A., 2016. A Complete Tutorial on Ridge and Lasso Regression in Python. Ανάκτηση: https://www.analyticsvidhya.com/blog/2016/01/complete-tutorial-ridge-lassoregression-python/

Johansson, R, An intuitive explanation of gradient boosting. Ανάκτηση από : http://www.cse.chalmers.se/ richajo/dit865/files/gb_explainer.pdf

Lifo.gr. (2019, May 9). Πόσα Airbnb μισθώνονται στην Αττική:Χιλιάδες τουρίστες και εκατομμύρια έσοδα. Ανάκτηση από Lifo.gr: https://www.lifo.gr/now/greece/236736/posa-airbnb-misthonontai-stin-attiki-xiliades-toyristes-kai-ekatommyria-esoda

Mathworks, 2016. *Introducing Machine Learning.* Ανάκτηση από : https://www.mathworks.com/content/dam/mathworks/tag-team/Objects/i/88174_92991v00_machine_learning_section1_ebook.pdf

Marr, B., 2016. Forbes. Ανάκτηση: https://www.forbes.com/sites/bernardmarr/2016/02/19/a-short-history-ofmachine-learning-every-manager-should-read/#79a9560715e7

Molla, R., 2017. Recode. Ανάκτηση: https://www.recode.net/2017/7/19/15949782/airbnb-100-million-stays-2017threat-business-hotel-industry

Oates, G., 2017. *Demystifying airbnb Corporate travel managers*. Ανάκτηση από : https://kwaliteit.toerismevlaanderen.be/sites/default/files/atoms/files/Demystifying%20Airbnb%20for%20corporate%20travel%20managers.pdf

Ravi, S., 2017. The Current and Future State of the Sharing Economy, New Delhi: Brookings Institution India Center. Ανάκτηση : https://www.brookings.edu/wp-content/uploads/2016/12/sharingeconomy_032017final.pdf

Society, T. R., 2017. *Machine learning: the power and promise of computers that learn by example Ανάκτηση από:* https://royalsociety.org/~/media/policy/projects/machine-learning/publications/machine-learning-report.pdf

Roussanoglou, N., 2017. ekathimerini.com Απογειώθηκαν οι μισθώσεις μέσω Airbnb στο κέντρο της Αθήνας. Ανάκτηση:

http://www.kathimerini.gr/913973/article/oikonomia/epixeirhseis/apogeiw8hkan-oimis8wseis-mesw-airbnb-sto-kentro-ths-a8hnas

Roussanoglou, N., 2017. ekathimerini.com Athens properties soar on Airbnb. Ανάκτηση:
http://www.ekathimerini.com/219266/article/ekathimerini/business/athensproperties-soar-on-airbnb

Roussanoglou, N., 2017. kathimerini.gr Πλάκα, Εξάρχεια και κουκάκι πρώτα στις μισθώσεις τύπου Airbnb. Ανάκτηση:
http://www.kathimerini.gr/932213/article/oikonomia/ellhnikhoikonomia/plaka-e3arxeia-kai-koykaki-prwta-stis-mis8wseis-typoy-airbnb

Seltman, H. J., 2018. *Experimental Design and Analysis.*
Ανάκτηση: http://www.stat.cmu.edu/~hseltman/309/Book/Book.pdf

Stephanie, 2015. Statisticshowto Lasso-Regression. Ανάκτηση:
http://www.statisticshowto.com/lasso-regression/

TechTarget, 2010. opinion mining (sentiment mining).

Ανάκτηση: https://searchbusinessanalytics.techtarget.com/definition/opinion-miningsentiment-mining

Tibshirani, R., 2013. *Modern Regression 2: The Lasso. Ανάκτηση από : https://www.stat.cmu.edu/~ryantibs/datamining/lectures/17-modr2.pdf*.

Woodruff, K., 2017. *Introduction to boosted decision trees, Ανάκτηση από: https://indico.fnal.gov/event/15356/contributions/31377/attachments/19671/24560/DecisionTrees.pdf*.

WU, S., LEE, F. & REYNARD, J., 2012. *Airbnb Ανάκτηση από:*
https://www.coursehero.com/file/12776376/Airbnb/

Λάρισσας, Τ,. Εισαγωγή στη Χρήση του SPSS for Windows. Ανάκτηση:
http://www.lib.teiher.gr/webnotes/seyp/spss/Kef12.pdf