ΔΙΕΘΝΕΣ ΠΑΝΕΠΙΣΤΗΜΙΟ ΤΗΣ ΕΛΛΑΔΟΣ

ΠΡΟΒΛΕΨΕΙΣ ΕΣΟΔΩΝ ΠΕΛΑΤΩΝ ΤΟΥ GOOGLE ANALYTICS

ΑΡΙΣΤΟΤΕΛΗΣ ΠΟΖΙΔΗΣ

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ: ΑΛΚΙΒΙΑΔΗΣ ΤΣΙΜΠΙΡΗΣ

ΔΙΕΘΝΕΣ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΤΗΣ ΕΛΛΑΔΟΣ

Διεθνές Πανεπιστήμιο Της Ελλάδος

Τμήμα Μηχανικών Πληροφορικής και Τηλεπικοινωνιών

Κατεύθυνση Μηχανικών Λογισμικού

ΠΡΟΒΛΕΨΕΙΣ ΕΣΟΔΩΝ ΠΕΛΑΤΩΝ ΤΟΥ GOOGLE ANALYTICS

ΑΡΙΣΤΟΤΕΛΗΣ ΠΟΖΙΔΗΣ

ΑΕΜ : 3997

ΤΣΙΜΠΙΡΗΣ ΑΛΚΙΒΙΑΔΗΣ

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ

3 Δεκεμβρίου 2019

Περίληψη

Στόχος αυτής της πτυχιακής εργασίας ήταν η ανάλυση ενός συνόλου δεδομένων πελατών του Google Merchandise Store και η πρόβλεψη εσόδων ανά πελάτη αποδεικνύοντας έτσι τον κανόνα 80/20.

Η συμβολή της ανάλυσης δεδομένων στην απόκτηση νέας γνώσης και στη διαδικασία λήψης αποφάσεων είναι ιδιαίτερα σημαντική. Γενικά, αν θεωρηθεί ότι μελετάται ένα αντικείμενο ενός ευρύτερου συστήματος, τότε η ανάλυση των δεδομένων που έχουν συλλεχθεί σε συνδυασμό με την αρχική γνώση του αντικειμένου οδηγούν σε νέες γνώσεις.

Ο κανόνας 80/20 η αλλιώς αρχή Pareto θεωρήθηκε αρχικά μια μαθηματική φόρμουλα για να προβλέπει το ποσοστό των ατόμων που είχαν το υψηλότερο εισόδημα σε μια κοινωνία. Αργότερα, όταν άρχισε να εξετάζεται καλύτερα, βρέθηκε ότι μπορούσε να έχει εφαρμογές σχεδόν παντού. Ο κανόνας 80/20 αποδεικνύεται αληθινός για πολλές επιχειρήσεις, μόνο ένα μικρό ποσοστό πελατών παράγει το μεγαλύτερο μέρος των εσόδων δείχνουν τα αποτελέσματα από τις αναλύσεις δεδομένων. Τα αποτελέσματα αυτά οδηγούν στην εξαγωγή χρήσιμων συμπερασμάτων που θα βοηθήσουν τους υπεύθυνους στη λήψη των κατάλληλων μέτρων και αποφάσεων για τη βελτίωση των επενδύσεων σε στρατηγικές προώθησης.

Abstract

The purpose of this thesis was to analyze a set of Google Merchandise Store customer data and predict customer revenue thus demonstrating the 80/20 rule.

The contribution of data analysis to the acquisition of new knowledge and decision-making is particularly important. Generally, if an object is considered a wider system, then the analysis of the data collected in conjunction with the object's original knowledge leads to new knowledge.

The 80/20 rule or the Pareto principle was initially considered a mathematical formula to predict the percentage of people who had the highest income in a society. Afterwards, when it was analyzed better, it was found that it could have applications almost everywhere. The 80/20 rule is accurate for many businesses, with only a small percentage of customers generating the bulk of the revenue resulting from data analysis. These results lead to useful conclusions that will help managers to take appropriate measures and decisions to improve investment in promotional strategies.

Ευχαριστίες

Η παρούσα πτυχιακή εργασία με θέμα "Προβλέψεις εσόδων πελατών του Google Analytics" πραγματοποιήθηκε στα πλαίσια του τμήματος Μηχανικών Πληροφορικής του Ανώτατου Τεχνολογικού Εκπαιδευτικού Ιδρύματος Κεντρικής Μακεδονίας το έτος 2019.

Η παρούσα πτυχιακή εργασία είναι το αποτέλεσμα μιας σειράς αλληλεπιδράσεων με διάφορα άτομα, καθένα από τα οποία έπαιξε σημαντικό ρόλο στην εξέλιξή της.

Θα ήθελα στο σημείο αυτό να εκφράσω τις θερμές μου ευχαριστίες σε όλους όσους με στήριξαν στην προσπάθεια αυτή:

Και πρώτα απ' όλα, στον επιβλέποντα καθηγητή της πτυχιακής μου εργασίας, κύριο Αλκιβιάδη Τσιμπίρη για την αμέριστη υποστήριξη, τις ουσιώδεις συμβουλές καθώς και την ενθάρρυνση που μου παρείχε σε όλη την διάρκεια της εργασίας αυτής.

Επίσης θα ήθελα να ευχαριστήσω θερμά τους συναδέλφους μου στην πρακτική μου άσκηση στην εταιρεία ASML NETHERLANDS B.V., για την καθοδήγηση τους αλλά και για τον χρόνο που διέθεσαν για να με βοηθήσουν στην ολοκλήρωση αυτής της πτυχιακής εργασίας.

Τέλος, θα επιθυμούσα να ευχαριστήσω όλους εκείνους που στάθηκαν στο πλευρό μου και με συμβούλευαν σε όλοι την διάρκεια των τεσσάρων αυτών ετών: την οικογένειά μου και ορισμένους πάρα πολύ κοντινούς μου ανθρώπους, χωρίς τους οποίους τίποτα από όσα έχω καταφέρει μέχρι σήμερα δεν θα ήταν πραγματικότητα.

Αριστοτέλης Ποζίδης

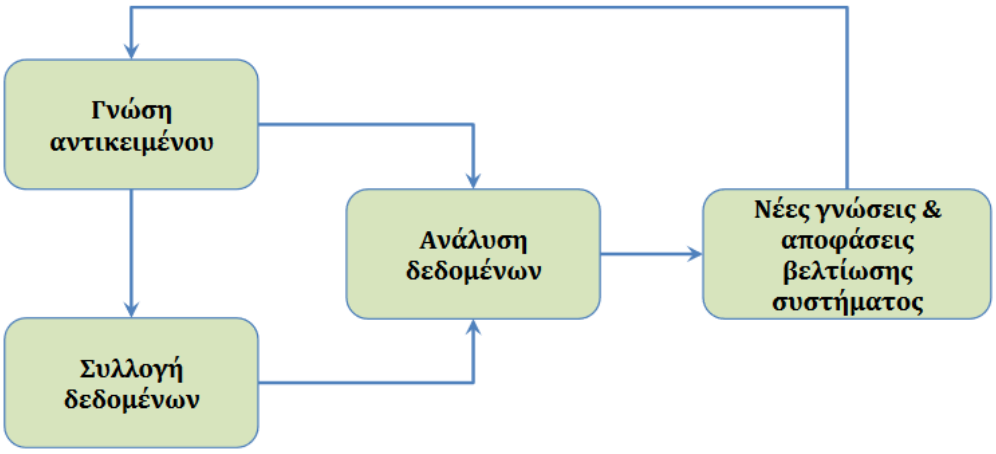
Σέρρες, 2019

Περιεχόμενα

Κεφάλαιο 1

Εισαγωγή

Για κάθε επιχείρηση το πεδίο της ανάλυσης δεδομένων είναι απαραίτητο, διότι οδηγεί σε έξυπνες επιχειρηματικές λύσεις που κατά περίπτωση μπορεί να αποδειχθούν αποδοτικότερες, να αποφέρουν υψηλότερα κέρδη και να εξασφαλίσουν ένα μόνιμα ικανοποιημένο πελατολόγιο. Η ανάλυση δεδομένων είναι η διαδικασία αξιολόγησης δεδομένων χρησιμοποιώντας αναλυτικά και στατιστικά εργαλεία για την ανακάλυψη χρήσιμης πληροφορίας που θα συντελέσει στη λήψη επιχειρηματικών αποφάσεων. Υπάρχουν διάφοροι μέθοδοι ανάλυσης δεδομένων, όπως η εξόρυξη δεδομένων, η ανάλυση κειμένου, η επιχειρησιακή ευφυΐα όπως και η απεικόνιση δεδομένων. [1]



Σχήμα 1.1: Data Analysis

Για την απόκτηση νέας γνώσης πρέπει να ακολουθηθεί η ροή του σχήματος 1.1. Ξεκινώντας λοιπόν από την συλλογή δεδομένων όπου χωρίς αυτά δεν θα ήταν δυνατή η

ανάλυση και η βελτιστοποίηση του εκάστοτε συστήματος, ακολουθεί η ανάλυση των συλλεγμένων δεδομένων όπου έχει ως αποτέλεσμα τις νέες γνώσεις και αποφάσεις και τέλος η απόκτηση γνώσεις για το αντικείμενο της ανάλυσης. Όμως δεν τερματίζεται σε αυτό το σημείο η διαδικασία διότι η νέα γνώση γίνεται ο λόγος για νέα συλλογή δεδομένων ή εκτενέστερη ανάλυση δεδομένων.

1.1 Κίνητρα έρευνας

Η Επιστήμη των Δεδομένων είναι ένα διεπιστημονικό πεδίο του οποίου αντικείμενο είναι η εξαγωγή της γνώσης από αδόμητα ή δομημένα δεδομένα. [2] [3] Αποτελεί επέκταση διάφορων επιστημονικών πεδίων όπως της στατιστικής, της ανάλυσης προγνωστικών (predictive analytics), της μηχανικής μάθησης (machine learning) και της εξόρυξης γνώσης (data mining). Με τις τεχνολογικές προόδους τις δύο τελευταίες δεκαετίες, σε συνδυασμό, εν μέρη με την έκρηξη του διαδικτύου, έχει προκύψει μια νέα μορφή ανάλυσης δεδομένων. Στην παρούσα πτυχιακή εργασία θα παρουσιαστεί μια εφαρμογή η οποία δέχεται ως είσοδο ένα μεγάλο αρχείο δομημένων δεδομένων (*.csv) και αυτοματοποιεί κάποιες διεργασίες της επιστήμης αυτής, όπως την ανάλυση προγνωστικών, την στατιστική ανάλυση και την εξόρυξη δεδομένων. Ύστερα από αυτές τις διεργασίες η εφαρμογή έχει ως έξοδο μία σειρά διαφόρων γραφημάτων, τα οποία βοηθούν στην αναπαράσταση των δεδομένων αυτών και στην σύνοψη σημαντικών συμπερασμάτων.

1.2 Ερευνητικές ερωτήσεις

Η παρούσα πτυχιακή εργασία έχει ως κύριο στόχο την απάντηση ορισμένων ερωτημάτων διαχείρισης μεγάλου όγκου δεδομένων και την σωστή και σαφή αναπαράσταση τους με γραφήματα στατιστικής ανάλυσης. Η κύρια ερώτηση που θα απαντηθεί μέσω της παρούσας πτυχιακής είναι κατά πόσο ισχύει ο κανόνας του 80/20. Δηλαδή αν το μεγαλύτερο μέρος των εσόδων του καταστήματος προέρχεται από το 20 τις εκατό των επισκεπτών του καταστήματος. Επίσης θα γίνει ξεκάθαρα κατανοητό εάν πρέπει να αναπαριστούμε τα δεδομένα για καλύτερη και ευκολότερη κατανόησή τους με μεθόδους διαγραμματικής αναπαράστασης στατιστικών μεγεθών.

Κεφάλαιο 2

Θεωρητική Επισκόπηση

2.1 Η επιστήμη των δεδομένων (Data Science)

Η επιστήμη των δεδομένων χρησιμοποιεί εκτεταμένες τεχνικές και θεωρίες από διάφορους τομείς όπως τα μαθηματικά, την έρευνα, την επιστήμη της πληροφορίας και την επιστήμη των υπολογιστών. Στην πρακτική προσέγγιση περιλαμβάνει την ανάλυση σημάτων, τα προγνωστικά μοντέλα, τη μηχανική μάθηση, τη στατιστική, την εξόρυξη δεδομένων, τις βάσεις δεδομένων, τον προγραμματισμό αλλά, και τέλος, την τεχνητή νοημοσύνη. Οι μέθοδοι διαχείρισης των μεγάλων δεδομένων (big data) έχουν πιθανώς το μεγαλύτερο ενδιαφέρον της συγκεκριμένης επιστήμης, παρόλο που οι μέθοδοι που χρησιμοποιούνται στην επιστήμη δεδομένων δεν αφορούν αποκλειστικά μεγάλους όγκους δεδομένων. [4] Ειδικότερα, η επιστήμη δεδομένων προέκυψε από το συνδυασμό σημαντικών εξελίξεων σε δυο υπό-περιοχές της πληροφορικής τα τελευταία 15 χρόνια. Πρώτον τη σημαντική πρόοδο που σημειώθηκε σε αλγορίθμους και τεχνικές μηχανικής μάθησης και γενικότερα τεχνικές τεχνητής νοημοσύνης βασισμένες σε στατιστικές αρχές και δεύτερον, στην περιοχή της διαχείρισης δεδομένων, που οδήγησαν μέσω νέων αλγορίθμων, αρχιτεκτονικών και συστημάτων σε τάξεις μεγέθους βελτίωση της ταχύτητας επεξεργασίας τεράστιων, ετερογενών, συνεχών μεταβαλλόμενων όγκων δεδομένων. Οι σύγχρονες επεξεργαστικές δυνατότητες συνδυασμένες με τον όγκο των δεδομένων δημιούργησαν ένα ενάρετο κύκλο ανάπτυξης υπολογιστικών τεχνικών που στηρίζονται στην επαναληπτική βελτίωση,τη πρόβλεψη και τέλος, τη λήψη αποφάσεων. [5]

2.2 Μεγάλα δεδομένα (Big Data)

Τα μεγάλα δεδομένα η αλλιώς Big Data είναι ένα πεδίο που αντιμετωπίζει τους τρόπους ανάλυσης, συστηματικής απόσπασης πληροφοριών ή άλλους τρόπους αντιμετώπισης πακέτων δεδομένων που είναι πολύ μεγάλα ή περίπλοκα για να αντιμετωπιστούν από το παραδοσιακό λογισμικό εφαρμογών επεξεργασίας δεδομένων. Δεδομένα με πολλές περιπτώσεις (σειρές) προσφέρουν μεγαλύτερη στατιστική ισχύ, ενώ δεδομένα με μεγαλύτερη πολυπλοκότητα (περισσότερα χαρακτηριστικά ή στήλες) μπορεί να οδηγήσουν σε υψηλότερο ποσοστό ψευδών ανακαλύψεων. [6] Οι μεγάλες προκλήσεις περιλαμβάνουν τη συλλογή, την αποθήκευση, την ανάλυση και την αναζήτηση δεδομένων καθώς και τη κοινή χρήση, τη μεταφορά, την απεικόνιση, την ιδιωτικότητα των πληροφοριών και την πηγή δεδομένων. Τα μεγάλα δεδομένα συνδέθηκαν αρχικά με τρεις βασικές έννοιες: την έννοια του όγκου, της ποικιλίας και της ταχύτητας. [7] Όταν διαχειριζόμαστε τα μεγάλα δεδομένα, ενδέχεται να μην συλλέγουμε δείγματα αλλά απλά να παρατηρούμε και να παρακολουθούμε την εξέλιξη. Επομένως, τα μεγάλα δεδομένα περιλαμβάνουν μεγέθη που υπερβαίνουν την ικανότητα του παραδοσιακού λογισμικού να επεξεργάζεται εντός αποδεκτού χρόνου [8] και αξίας. [9] Ο τρέχων όρος μεγάλα δεδομένα τείνει να αναφέρεται στη χρήση προγνωστικών αναλύσεων, αναλύσεων συμπεριφοράς χρηστών ή ορισμένων άλλων προηγμένων μεθόδων ανάλυσης δεδομένων που εξάγουν αξία από δεδομένα και σπανίως σε ένα συγκεκριμένο μέγεθος συνόλου δεδομένων. Δεν υπάρχει αμφιβολία ότι οι διαθέσιμες ποσότητες δεδομένων είναι πράγματι μεγάλες, αλλά αυτό δεν είναι το πιο σχετικό χαρακτηριστικό αυτού του νέου οικοσυστήματος δεδομένων. [10] Η ανάλυση των συνόλων δεδομένων μπορεί να εντοπίσει νέες συσχετίσεις για τις τάσεις των επιχειρήσεων, καταπολέμηση της εγκληματικότητας κ.ο.κ. [11]

2.3 Τεχνικές Εξόρυξης Δεδομένων (Data Mining)

Εξόρυξη δεδομένων (ή ανακάλυψη γνώσης από μεγάλο όγκο δεδομένων) [12] είναι η εξεύρεση μιας (ενδιαφέρουσας, αυτονόητης, μη προφανούς και πιθανόν χρήσιμης) πληροφορίας ή προτύπων από μεγάλο όγκο δεδομένων με χρήση αλγορίθμων ομαδοποίησης ή κατηγοριοποίησης και των αρχών της στατιστικής, της τεχνητής νοημοσύνης, της μηχανικής μάθησης και των συστημάτων βάσεων δεδομένων. Στόχος της εξόρυξης

δεδομένων είναι η πληροφορία που θα εξαχθεί και τα πρότυπα που θα προκύψουν να έχουν δομή κατανοητή προς τον άνθρωπο έτσι ώστε να τον βοηθήσουν να πάρει τις κατάλληλες αποφάσεις. Υπάρχουν αρκετές τεχνικές εξόρυξης γνώσης που αναλύονται παρακάτω.

2.3.1 Νευρωνικά Δίκτυα (Neural Networks)

Τα τεχνητά νευρικά δίκτυα (ANN) ή τα συστήματα σύνδεσης είναι υπολογιστικά συστήματα που εμπνέονται, αλλά δεν ταυτίζονται, με τα βιολογικά νευρωνικά δίκτυα που αποτελούν τον εγκέφαλο των ζωντανών όντων. Αυτά τα συστήματα 'μαθαίνουν' να εκτελούν εργασίες εξετάζοντας παραδείγματα, γενικά χωρίς να προγραμματίζονται με συγκεκριμένους κανόνες. Για παράδειγμα, στην αναγνώριση εικόνας μπορεί να μάθουν να εντοπίζουν εικόνες που περιέχουν γάτες αναλύοντας παραδείγματα εικόνων που έχουν επισημανθεί με το χέρι ως 'γάτα' ή 'μη γάτα' και χρησιμοποιώντας τα αποτελέσματα για τον εντοπισμό τους σε άλλες εικόνες. Το κάνουν αυτό χωρίς προηγούμενη γνώση, για παράδειγμα, ότι έχουν γούνα, ουρές, μουστάκια και πρόσωπα που μοιάζουν με γάτες. Και άυτού, παράγουν αυτόματα τα αναγνωριστικά χαρακτηριστικά από τα παραδείγματα που επεξεργάζονται.

Ένα ANN βασίζεται σε μια συλλογή από συνδεδεμένες μονάδες ή κόμβους που ονομάζονται τεχνητοί νευρώνες, οι οποίοι μοντελοποιούν τους νευρώνες σε έναν βιολογικό εγκέφαλο. Κάθε σύνδεση, όπως οι συνάψεις σε έναν βιολογικό εγκέφαλο, μπορεί να μεταδώσει ένα σήμα σε άλλους νευρώνες. Ένας τεχνητός νευρώνας που λαμβάνει ένα σήμα στη συνέχεια επεξεργάζεται αυτό και μπορεί να σηματοδοτήσει νευρώνες που συνδέονται με αυτόν.

Στις υλοποιήσεις του ANN, το 'σήμα' σε μια σύνδεση είναι ένας πραγματικός αριθμός και η έξοδος κάθε νευρώνα υπολογίζεται από κάποια μη γραμμική συνάρτηση του αθροίσματος των εισόδων του. Οι συνδέσεις ονομάζονται άκρα. Οι νευρώνες και τα άκρα έχουν συνήθως ένα βάρος που προσαρμόζεται ως έσοδα της μάθησης. Το βάρος αυξάνει ή μειώνει τη δύναμη του σήματος σε μια σύνδεση. Οι νευρώνες μπορεί να έχουν ένα φίλτρο τέτοιο ώστε ένα σήμα να αποστέλλεται μόνο αν το συνολικό σήμα διασχίζει αυτό το όριο. Τυπικά, οι νευρώνες συσσωματώνονται σε στρώματα. Τα διαφορετικά στρώματα μπορούν να εκτελούν διαφορετικούς μετασχηματισμούς στις εισόδους τους. Τα σήματα

μετακινούνται από το πρώτο στρώμα (το στρώμα εισόδου) στο τελευταίο στρώμα (το στρώμα εξόδου), πιθανώς μετά από πολλαπλές διαδρομές και μετατροπές.

Ο αρχικός στόχος της προσέγγισης ANN ήταν να λυθούν τα προβλήματα με τον ίδιο τρόπο που θα λυνόντουσαν από έναν ανθρώπινο εγκέφαλο. Ωστόσο, με την πάροδο του χρόνου, η προσοχή μεταφέρθηκε στην εκτέλεση συγκεκριμένων καθηκόντων, οδηγώντας σε αποκλίσεις από τη βιολογία. Τα ANN έχουν χρησιμοποιηθεί σε διάφορα καθήκοντα, όπως οραματισμό στον υπολογιστή, αναγνώριση ομιλίας, μηχανική μετάφραση, φιλτράρισμα κοινωνικών δικτύων, βιντεοπαιχνίδια, ιατρική διάγνωση και ακόμη και σε δραστηριότητες που παραδοσιακά θεωρούνται αποκλειστικά για τον άνθρωπο, όπως ζωγραφική [13]

2.3.2 Δέντρα Αποφάσεων (Decision Trees)

Ένα δέντρο απόφασης είναι ένα εργαλείο υποστήριξης αποφάσεων που χρησιμοποιεί ένα γράφημα τύπου δέντρου ή ένα μοντέλο αποφάσεων και τις πιθανές συνέπειες τους, συμπεριλαμβανομένων των εξελίξεων τυχαίων γεγονότων, του κόστους πόρων και της χρησιμότητας. Είναι ένας τρόπος προβολής ενός αλγορίθμου που περιέχει μόνο δηλώσεις υπό όρους ελέγχου.

Ένα δέντρο απόφασης είναι μια δομή που μοιάζει με διάγραμμα ροής flowchart στην οποία κάθε εσωτερικός κόμβος αντιπροσωπεύει μια 'δοκιμή' σε ένα χαρακτηριστικό (π.χ. μια ρίψη νομισμάτων που έχει ως αποτέλεσμα κεφαλή ή γράμματα), κάθε κλάδος αντιπροσωπεύει το αποτέλεσμα της δοκιμής, και κάθε κόμβος φύλλων αντιπροσωπεύει ένα χαρακτηριστικό (η απόφαση λαμβάνεται αφού υπολογιστούν όλα τα χαρακτηριστικά). Οι διαδρομές από ρίζα (root) σε φύλλο αντιπροσωπεύουν κανόνες ταξινόμησης.

Οι αλγόριθμοι μάθησης που βασίζονται σε δέντρα αποφάσεων θεωρούνται από τις καλύτερες μεθόδους μάθησης με επίβλεψη που χρησιμοποιούνται ως επί το πλείστον. Οι μέθοδοι που βασίζονται σε δέντρα ενισχύουν τα προγνωστικά μοντέλα με υψηλή ακρίβεια, σταθερότητα και ευκολία ερμηνείας. Σε αντίθεση με τα γραμμικά μοντέλα, χαρτογραφούν τις μη γραμμικές σχέσεις αρκετά καλά. Είναι προσαρμόσιμα στην επίλυση κάθε είδους προβλήματος (ταξινόμηση ή παλινδρόμηση). Οι αλγόριθμοι των δέντρων αποφάσεων αναφέρονται ως CART (δέντρα ταξινόμησης και παλινδρόμησης).

2.3.3 Συναρτήσεις Απόστασης (Distance functions)

Στα μαθηματικά, μια συνάρτηση μέτρησης ή απόστασης είναι μια μέθοδος που ορίζει μια απόσταση μεταξύ κάθε ζεύγος στοιχείων ενός συνόλου. Ένα σετ με μια μέτρηση ονομάζεται μετρικός χώρος. [14] Μια μέτρηση επάγει μια τοπολογία σε ένα σετ, αλλά δεν μπορούν να δημιουργηθούν όλες οι τοπολογίες με μια μέτρηση. Ένας χώρος του οποίου η τοπολογία μπορεί να περιγραφεί από μια μέτρηση ονομάζεται μετρήσιμη.

Μια σημαντική πηγή μετρήσεων στη διαφορική γεωμετρία είναι οι μετρικές τάσεις, οι διηλεκτρικές μορφές που μπορούν να οριστούν από τους εφαπτομενικούς φορείς μιας διαφοροποιήσιμης πολλαπλής μιας κλίμακας. Ένας μετρικός τανυστής επιτρέπει τις αποστάσεις κατά μήκος των καμπυλών που πρέπει να προσδιοριστούν μέσω της ολοκλήρωσης, και έτσι καθορίζει μια μέτρηση. Ωστόσο, κάθε μετρική δεν προέρχεται από έναν μετρικό τανυστή με αυτόν τον τρόπο.

Μια μέτρηση σε ένα σετ X είναι μια συνάρτηση (που ονομάζεται συνάρτηση απόστασης ή απλά απόσταση)

$$d : X \times X \to [0, \infty)$$

όπου $[0, \infty)$ είναι το σύνολο των μη αρνητικών πραγματικών αριθμών και για όλα τα $x, y, z \in X$, πληρούνται οι ακόλουθες προϋποθέσεις:

1. $d(x, y) \geq 0$	μη αρνητικότητα ή αξίωμα διαχωρισμού [15]
2. $d(x, y) = 0 \Leftrightarrow x = y$	ταυτότητα των αδιάκριτων [16]
3. $d(x, y) = d(y, x)$	συμμετρία [17] [18]
4. $d(x, z) \leq d(x, y) + d(y, z)$	δευτερογενή ή τριγωνική ανισότητα [19] [20]

Οι συνθήκες 1 και 2 ορίζουν μαζί μια θετικά καθορισμένη λειτουργία. Η πρώτη προϋπόθεση υπονοείται από τους άλλους.

Μια μέτρηση ονομάζεται υπερμετρική αν ικανοποιεί την ακόλουθη ισχυρότερη εκδοχή της τριγωνικής ανισότητας όπου τα σημεία δεν μπορούν ποτέ να πέσουν «μεταξύ» άλλων σημείων:

$$d(x, z) \leq \max(d(x, y), d(y, z))$$

για όλα τα $x, y, z \in X$

Μια μέτρηση d στο X ονομάζεται εγγενής αν οποιαδήποτε δύο σημεία x και y στο X μπορούν να ενωθούν με μία καμπύλη με μήκος αυθαίρετα κοντά στο d(x, y).

Για σύνολα στα οποία έχει οριστεί μια προσθήκη + : $X \times X \longrightarrow X$, d ονομάζεται μεταβλητή μετάφρασης εάν

$$d(x, y) = d(x + a, y + a)$$

για όλα τα $x, y, z \in X$

2.3.4 Συσχετίσεις (Correlations)

Η συσχέτιση είναι ένα στατιστικό μέτρο που υποδεικνύει το βαθμό στον οποίο δύο ή περισσότερες μεταβλητές κυμαίνονται μεταξύ τους. Ο θετικός συσχετισμός δείχνει το βαθμό στον οποίο οι μεταβλητές αυτές αυξάνονται ή μειώνονται παράλληλα, ενώ μια αρνητική συσχέτιση δείχνει το βαθμό στον οποίο μια μεταβλητή αυξάνεται καθώς η άλλη μειώνεται. [21]

Το πιο γνωστό μέτρο εξάρτησης μεταξύ δύο ποσοτήτων είναι ο συντελεστής συσχέτισης του Pearson, κοινώς ονομάζεται απλά ο 'συντελεστής συσχέτισης'. Λαμβάνεται διαιρώντας τη συν-διακύμανση των δύο μεταβλητών με το γινόμενο των τυπικών αποκλίσεων τους. Ο Karl Pearson ανέπτυξε το συντελεστή από μια παρόμοια αλλά ελαφρώς διαφορετική ιδέα από τον Francis Galton. [22] Ο συντελεστής συσχέτισης είναι ένα στατιστικό μέτρο του βαθμού στον οποίο οι μεταβολές στην τιμή μιας μεταβλητής προβλέπουν αλλαγή στην τιμή άλλης. Όταν η διακύμανση μιας μεταβλητής προβλέπει αξιόπιστα μια παρόμοια διακύμανση σε μια άλλη μεταβλητή, το συχνότερο συμπέρασμα είναι ότι η αλλαγή σε μία τιμή προκαλεί την αλλαγή στην άλλη. Ωστόσο, ο συσχετισμός δεν συνεπάγεται με αιτιώδη συνάφεια διότι μπορεί να υπάρχει, για παράδειγμα, ένας άγνωστος παράγοντας που επηρεάζει και τις δύο μεταβλητές παρομοίως.

Ο συντελεστής συσχέτισης του πληθυσμού $\rho_{X,Y}$ μεταξύ δύο τυχαίων μεταβλητών X και Y με τις αναμενόμενες τιμές μ_X και μ_Y και τυπικές αποκλίσεις σ_X και σ_Y ορίζεται ως

$$\rho_{X,Y} = corr(X, Y) = \frac{conv(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

όπου E είναι ο χειριστής της αναμενόμενης τιμής, cov σημαίνει συν-διακύμανση και corr είναι μια ευρέως χρησιμοποιούμενη εναλλακτική ένδειξη για τον συντελεστή συ-

σχέτισης. Ο συσχετισμός Pearson ορίζεται μόνο εάν και οι δύο τυπικές αποκλίσεις είναι πεπερασμένες και θετικές. Μια εναλλακτική φόρμουλα είναι

$$\rho_{X,Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E(X)^2}\,\sqrt{E(Y^2) - E(Y)^2}}$$

2.3.5 Κατάταξη (Classification)

Η κατάταξη είναι μια διαδικασία που σχετίζεται με την κατηγοριοποίηση στην οποία οι ιδέες και τα αντικείμενα αναγνωρίζονται, διαφοροποιούνται και κατανοούνται. Θεωρείται περίπτωση εποπτευόμενης μάθησης, δηλαδή η εκμάθηση όπου είναι διαθέσιμο ένα εκπαιδευτικό σύνολο σωστά προσδιορισμένων παρατηρήσεων. [23]

Συχνά, οι μεμονωμένες παρατηρήσεις αναλύονται σε ένα σύνολο ποσοτικοποιημένων ιδιοτήτων, γνωστών ως επεξηγηματικές μεταβλητές ή χαρακτηριστικά. Αυτές οι ιδιότητες μπορούν να είναι κατηγορηματικές (π.χ. για τύπο αίματος), κανονικές (π.χ. 'μεγάλες', 'μέτριες' ή 'μικρές'), ή πραγματικής αξίας (π.χ. μετρήσεις). Άλλοι ταξινομητές δουλεύουν συγκρίνοντας παρατηρήσεις με προηγούμενες παρατηρήσεις μέσω μιας συνάρτησης ομοιότητας ή απόστασης. Ένας αλγόριθμος που εφαρμόζει την κατάταξη, ειδικά σε μια συγκεκριμένη εφαρμογή, είναι γνωστός ως ταξινομητής. Ο όρος 'ταξινομητής' αναφέρεται επίσης μερικές φορές στη μαθηματική συνάρτηση, που εφαρμόζεται από έναν αλγόριθμο ταξινόμησης, ο οποίος χαρτογραφεί δεδομένα εισόδου σε μια κατηγορία.

2.3.6 Συσταδοποίηση (Clustering)

Η συσταδοποίηση είναι η ομαδοποίησης ενός συνόλου αντικειμένων με τέτοιο τρόπο ώστε αντικείμενα στην ίδια ομάδα (Cluster) να είναι περισσότερο όμοια (με κάποια έννοια) μεταξύ τους παρά με εκείνα σε άλλες ομάδες (Clusters). Πρόκειται για ένα βασικό καθήκον διερευνητικής εξόρυξης δεδομένων και μιας κοινής τεχνικής για την ανάλυση στατιστικών δεδομένων που χρησιμοποιείται σε πολλούς τομείς, συμπεριλαμβανομένης της μηχανικής μάθησης, της αναγνώρισης προτύπων, της ανάλυσης εικόνων, της ανάκτησης πληροφοριών, της βιοπληροφορικής, της συμπίεσης δεδομένων και των γραφικών υπολογιστών.

Η συσταδοποίηση δεν είναι ένας συγκεκριμένος αλγόριθμος, αλλά το γενικό καθήκον που πρέπει να επιλυθεί. Μπορεί να επιτευχθεί με διάφορους αλγόριθμους που διαφέρουν σημαντικά στην κατανόησή τους για το τι αποτελεί σύμπλεγμα και πώς να τα ορίσει αποτελεσματικά. Οι δημοφιλείς έννοιες των συμπλεγμάτων περιλαμβάνουν ομάδες με μικρές αποστάσεις μεταξύ μελών του συμπλέγματος, πυκνές περιοχές του χώρου δεδομένων, διαστήματα ή συγκεκριμένες στατιστικές κατανομές. Επομένως, η συσταδοποίηση μπορεί να διατυπωθεί ως πρόβλημα πολλαπλών στόχων βελτιστοποίησης.

2.3.7 Παλινδρόμηση (Regression)

Η ανάλυση παλινδρόμησης είναι ένα σύνολο στατιστικών διεργασιών για την εκτίμηση των σχέσεων μεταξύ μιας εξαρτώμενης μεταβλητής (συχνά αποκαλούμενης «μεταβλητής έκβασης») και μιας ή περισσοτέρων ανεξάρτητων μεταβλητών (συχνά αποκαλούμενων «προγνωστικών», «συνεκτικών» ή «χαρακτηριστικών»). Η πιο συνηθισμένη μορφή ανάλυσης παλινδρόμησης είναι η γραμμική παλινδρόμηση (Linear Regression), στην οποία ένας ερευνητής βρίσκει τη γραμμή (ή μια πιο περίπλοκη γραμμική συνάρτηση) που ταιριάζει περισσότερο με τα δεδομένα σύμφωνα με ένα συγκεκριμένο μαθηματικό κριτήριο. Για παράδειγμα, η μέθοδος των συνήθων ελαχίστων τετραγώνων (Mean Squared Error) υπολογίζει τη μοναδική γραμμή (ή υπερπλησία) που ελαχιστοποιεί το άθροισμα των τετραγωνικών αποστάσεων μεταξύ των αληθινών δεδομένων και εκείνης της γραμμής. Για συγκεκριμένους μαθηματικούς λόγους, αυτό επιτρέπει στον ερευνητή να εκτιμήσει την υπό όρους προσδοκία (ή τη μέση τιμή του πληθυσμού) της εξαρτημένης μεταβλητής όταν οι ανεξάρτητες μεταβλητές λαμβάνουν ένα δεδομένο σύνολο τιμών. Οι λιγότερο συνήθεις μορφές παλινδρόμησης χρησιμοποιούν ελαφρώς διαφορετικές διαδικασίες για την εκτίμηση εναλλακτικών παραμέτρων θέσης (π.χ., ποσοτική παλινδρόμηση ή ανάλυση απαραίτητων συνθηκών) ή εκτίμηση της προσδοκίας υπό όρους σε μια ευρύτερη συλλογή μη γραμμικών μοντέλων (π.χ. μη παραμετρική παλινδρόμηση).

Κεφάλαιο 3

Διαχείρισή Δεδομένων

Η εφαρμογή δέχεται ως είσοδο δύο σειρές δεδομένων, το σετ δεδομένων εκπαίδευσης και το σετ δεδομένων εξέτασης. Και οι δύο σειρές είναι δομημένες που αυτό σημαίνει ότι όλα τα δεδομένα έχουν τιμές στις ίδιες στήλες (columns) με την ίδια ακριβώς σειρά, τις οποίες στήλες θα τις αναλύσουμε παρακάτω. Κάθε εγγραφή δηλαδή σειρά (row) αντιπροσωπεύει μια καινούργια επίσκεψη στο Gstore αλλά αυτό δεν σημαίνει απαραίτητα ότι είναι κάποιος καινούργιος χρήστης, μπορεί κάλλιστα ο ίδιος χρήστης να επισκέφτηκε το κατάστημα μερικά λεπτά νωρίτερα.

3.1 Σύνολα δεδομένων

Ένα σύνολο δεδομένων (Dataset) είναι μια συλλογή δεδομένων. Στην περίπτωση των πινακοποιημένων δεδομένων, ένα σύνολο δεδομένων αντιστοιχεί σε έναν ή περισσότερους πίνακες βάσεων δεδομένων, όπου κάθε στήλη ενός πίνακα αντιπροσωπεύει μια συγκεκριμένη μεταβλητή και κάθε σειρά αντιστοιχεί σε μια δεδομένη εγγραφή του εν λόγω συνόλου δεδομένων. Το σύνολο δεδομένων παραθέτει τιμές για κάθε μια από τις μεταβλητές, όπως το ύψος και το βάρος ενός αντικειμένου, για κάθε μέλος του συνόλου δεδομένων. Κάθε τιμή είναι γνωστή ως δεδομένο. Τα σύνολα δεδομένων μπορούν επίσης να αποτελούνται από μια συλλογή εγγράφων ή αρχείων.

Αρκετά χαρακτηριστικά καθορίζουν τη δομή και τις ιδιότητες μιας ομάδας δεδομένων. Αυτές περιλαμβάνουν τον αριθμό και τους τύπους των χαρακτηριστικών ή μεταβλητών

και διάφορα στατιστικά μέτρα που εφαρμόζονται σε αυτά, όπως η τυπική απόκλιση και η κύρτωση [24].

Οι τιμές μπορεί να είναι αριθμοί, όπως πραγματικοί αριθμοί ή ακέραιοι, για παράδειγμα, που αντιπροσωπεύουν το ύψος ενός ατόμου σε εκατοστά, αλλά μπορεί επίσης να είναι ονομαστικά δεδομένα (δηλ. Δεν αποτελούνται από αριθμητικές τιμές), για παράδειγμα, που αντιπροσωπεύουν την εθνότητα ενός ατόμου. Γενικότερα, οι τιμές μπορεί να είναι οποιουδήποτε είδους που περιγράφεται ως επίπεδο μέτρησης. Για κάθε μεταβλητή, οι τιμές είναι κανονικά όλες του ίδιου είδους. Εντούτοις, μπορεί επίσης να υπάρχουν ελλείπουσες τιμές, οι οποίες πρέπει να αναφέρονται με κάποιο διαφορετικό τρόπο (NAN).

Στα στατιστικά δεδομένα, τα σύνολα δεδομένων προέρχονται συνήθως από τις πραγματικές παρατηρήσεις που λαμβάνονται με δειγματοληψία ενός στατιστικού πληθυσμού και κάθε σειρά αντιστοιχεί στις παρατηρήσεις για ένα στοιχείο αυτού του πληθυσμού. Τα σύνολα δεδομένων μπορούν επιπλέον να δημιουργηθούν με αλγόριθμους για τον έλεγχο ορισμένων ειδών λογισμικού. Ορισμένα σύγχρονα λογισμικά στατιστικής ανάλυσης, όπως το SPSS, εξακολουθούν να παρουσιάζουν τα δεδομένα τους με την κλασική σειρά δεδομένων. Εάν τα δεδομένα λείπουν ή είναι ύποπτα, μπορεί να χρησιμοποιηθεί μια μέθοδος καταλογισμού για την ολοκλήρωση ενός συνόλου δεδομένων

Στη μηχανική μάθηση, ένα κοινό καθήκον είναι η μελέτη και η κατασκευή αλγορίθμων που μπορούν να μάθουν και να κάνουν προβλέψεις στα δεδομένα. Αυτοί οι αλγόριθμοι λειτουργούν με την πραγματοποίηση προβλέψεων ή αποφάσεων που βασίζονται σε δεδομένα, δημιουργώντας ένα μαθηματικό μοντέλο από δεδομένα εισόδου. Τα δεδομένα που χρησιμοποιούνται για την κατασκευή του τελικού μοντέλου προέρχονται συνήθως από πολλαπλά σύνολα δεδομένων. Συγκεκριμένα, τρία σύνολα δεδομένων χρησιμοποιούνται συνήθως σε διάφορα στάδια της δημιουργίας του μοντέλου. [25] [26]

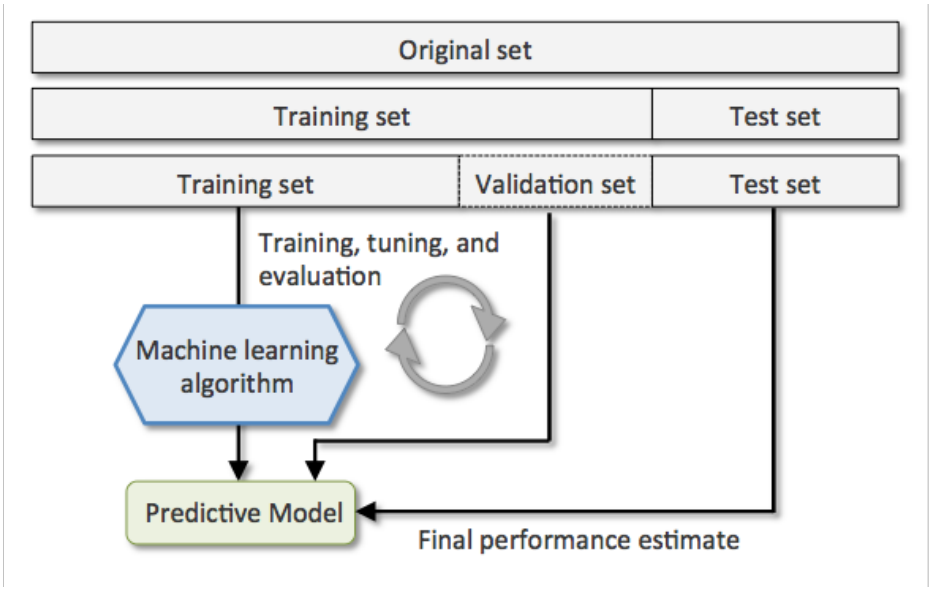
Το μοντέλο αρχικά ταιριάζει σε ένα σύνολο δεδομένων κατάρτισης, [27] που είναι ένα σύνολο παραδειγμάτων που χρησιμοποιούνται για την προσαρμογή των παραμέτρων (π.χ. βάρη των συνδέσεων μεταξύ νευρώνων σε τεχνητά νευρωνικά δίκτυα) του μοντέλου. [28] Το μοντέλο (π.χ. ένας αφελής ταξινομητής Bayes) εκπαιδεύεται στο σύνολο δεδομένων κατάρτισης χρησιμοποιώντας μια εποπτευόμενη μέθοδο εκμάθησης (π.χ. κατάβαση κλίσης ή κατάβαση στοχαστικής κλίσης). Στην πράξη, το σύνολο δεδομένων

κατάρτισης αποτελείται συχνά από ζεύγη ενός διανύσματος εισόδου (ή βαθμωτού) και από τον αντίστοιχο φορέα εξόδου (ή βαθμωτού), ο οποίος συνήθως συμβολίζεται ως στόχος (ή ετικέτα). Το τρέχον μοντέλο τρέχει με το σύνολο δεδομένων κατάρτισης και παράγει ένα αποτέλεσμα, το οποίο στη συνέχεια συγκρίνεται με το στόχο, για κάθε διάνυσμα εισόδου στο σύνολο δεδομένων κατάρτισης. Με βάση το αποτέλεσμα της σύγκρισης και τον ειδικό αλγόριθμο μάθησης που χρησιμοποιείται, οι παράμετροι του μοντέλου ρυθμίζονται. Η τοποθέτηση μοντέλου μπορεί να περιλαμβάνει τόσο μεταβλητή επιλογή όσο και εκτίμηση παραμέτρων.

Διαδοχικά, το προσαρμοσμένο μοντέλο χρησιμοποιείται για την πρόβλεψη των απαντήσεων για τις παρατηρήσεις σε μια δεύτερη δέσμη δεδομένων που ονομάζεται σύνολο δεδομένων επικύρωσης. Το σύνολο δεδομένων επικύρωσης παρέχει μια αμερόληπτη αξιολόγηση ενός μοντέλου που ταιριάζει στο σύνολο δεδομένων εκπαίδευσης, ενώ ρυθμίζει τα υπερπαραμετρικά μοντέλα (π.χ. ο αριθμός των κρυφών μονάδων σε ένα νευρικό δίκτυο). Τα σύνολα δεδομένων επικύρωσης μπορούν να χρησιμοποιηθούν για την τακτοποίηση με την έγκαιρη διακοπή: σταματήστε την κατάρτιση όταν αυξάνεται το σφάλμα στο σύνολο δεδομένων επικύρωσης, καθώς αυτό αποτελεί ένδειξη υπερφόρτωσης στο σύνολο δεδομένων κατάρτισης. Αυτή η απλή διαδικασία περιπλέκεται στην πράξη από το γεγονός ότι το σφάλμα του συνόλου δεδομένων επικύρωσης μπορεί να κυμαίνεται κατά τη διάρκεια της εκπαίδευσης, παράγοντας πολλαπλά τοπικά ελάχιστα. Αυτή η επιπλοκή έχει οδηγήσει στη δημιουργία πολλών ad-hoc κανόνων για να αποφασιστεί πότε ξεκίνησε πραγματικά η υπερφόρτωση. [29]

Τέλος, το σύνολο δεδομένων δοκιμής είναι ένα σύνολο δεδομένων που χρησιμοποιείται για να παρέχει μια αμερόληπτη αξιολόγηση ενός τελικού μοντέλου που ταιριάζει στο σύνολο δεδομένων κατάρτισης. Εάν τα δεδομένα στο σύνολο δεδομένων δοκιμών δεν έχουν χρησιμοποιηθεί ποτέ στην εκπαίδευση (για παράδειγμα σε διασταυρωμένη επικύρωση), το σύνολο δεδομένων δοκιμής ονομάζεται επίσης σύνολο δεδομένων holdout.

Όπως αναφέραμε και παραπάνω τα δεδομένα που διαχειρίζεται η εφαρμογή της παρούσας πτυχιακής είναι δομημένα και θα τα εξετάσουμε παρακάτω με λεπτομέρεια.



Σχήμα 3.1: Model Creation

3.1.1 Σετ δεδομένων εκπαίδευσης (train set)

Το σετ δεδομένων εκπαίδευσης είναι η κύρια πηγή πληροφορίας της εφαρμογής. Το οποίο σημαίνει ότι πέρα από ότι είναι βασικό τμήμα της εφαρμογής αλλά χωρίς αυτό δεν θα ήταν εφικτή καμία από τις αναπαράστασής στατιστικών γραφημάτων ούτε η πρόβλεψη των εσόδων ανά πελάτη του Gstore. Μέσα σε αυτή την συλλογή πληροφορίας κρύβονται όλα όσα χρειαζόμαστε για να παρουσιάσουμε την ¨κίνηση¨ των χρηστών και τις προτιμήσεις τους, έτσι με αυτόν τον τρόπο μπορούμε να προβλέψουμε την επόμενη του προτίμηση.

Τα δεδομένα αυτά λοιπόν είναι σε μορφή ".csv" τα οποία είναι αρχικά των λέξεων "Comma Seperated Values" που σημαίνουν ¨τιμές χωρισμένες με κόμμα¨. Έτσι κάθε κόμμα χωρίζει την μια στήλη από την άλλη και σε περίπτωση που δεν υπάρχει τιμή ανάμεσα στα κόμματα δεν υπάρχει τίποτα αλλά είναι πολύ σημαντικό να μην παραλειφθεί καμία από αυτές διότι αλλιώς δεν θα είναι σωστά τα δεδομένα και οι τελευταίες στήλες δεν θα έχουν τιμές το οποίο θα οδηγήσει σε σφάλματα.

Οι στήλες του σετ δεδομένων εκπαίδευσης είναι οι εξής:

- fullVisitorId → Ένα μοναδικό αναγνωριστικό για κάθε χρήστη του Google Merchandise Store.
- channelGrouping → Το κανάλι μέσω του οποίου ήρθε ο χρήστης στο Κατάστημα.

- date → Την ημερομηνία κατά την οποία ο χρήστης επισκέφθηκε το κατάστημα.
- device → Οι προδιαγραφές για τη συσκευή που χρησιμοποιούνται για την πρόσβαση στο κατάστημα.
- geoNetwork → Αυτή η ενότητα περιέχει πληροφορίες σχετικά με τη γεωγραφία του χρήστη.
- socialEngagementType → Είδος δέσμευσης, είτε 'κοινωνικά δεσμευμένος' είτε 'μη κοινωνικά δεσμευμένος'.
- totals → Αυτή η ενότητα περιέχει συνολικές τιμές κατά τη διάρκεια της περιόδου σύνδεσης.
- trafficSource → Αυτή η ενότητα περιέχει πληροφορίες σχετικά με την πηγή επισκεψιμότητας από την οποία προέρχεται η περίοδος σύνδεσης.
- visitId → Ένα αναγνωριστικό για αυτή τη σύνοδο. Αυτό είναι μέρος της αξίας που συνήθως αποθηκεύεται ως cookie _utmb. Αυτό είναι μοναδικό μόνο για τον χρήστη. Για ένα εντελώς μοναδικό αναγνωριστικό, θα πρέπει να χρησιμοποιήσετε ένα συνδυασμό fullVisitorId και visitId.
- visitNumber → Ο αριθμός της σύνδεσης για αυτόν τον χρήστη. Εάν πρόκειται για την πρώτη συνεδρία, τότε αυτό έχει οριστεί σε 1.
- visitStartTime → Η χρονική σήμανση (εκφράζεται ως χρόνος POSIX).
- hits → Αυτή η σειρά και τα ένθετα πεδία έχουν συμπληρωθεί για όλους τους τύπους επισκέψεων. Παρέχει μια καταγραφή όλων των επισκέψεων σελίδας.
- customDimensions → Αυτή η ενότητα περιέχει όλες τις προσαρμοσμένες ιδιότητες σε επίπεδο χρήστη ή σε επίπεδο σύνδεσης που έχουν οριστεί για μια περίοδο λειτουργίας. Αυτό είναι ένα επαναλαμβανόμενο πεδίο και έχει μια καταχώρηση για κάθε διάσταση που έχει οριστεί.
- totals → Αυτό το σύνολο των στηλών περιλαμβάνει ως επί το πλείστον συγκεντρωτικά δεδομένα υψηλού επιπέδου.

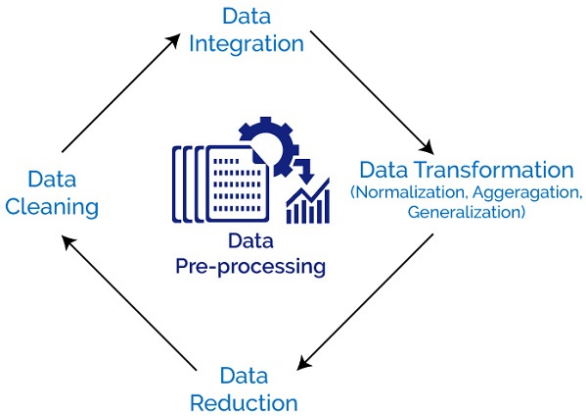
3.1.2 Σετ δεδομένων εξέτασης (test set)

Το σετ δεδομένων εξέτασης είναι ο τρόπος ελέγχου αν το μοντέλο που εκπαιδεύτηκε κάνει σωστές προβλέψεις. Με την βοήθεια αυτού μπορούμε να βγάζουμε μια τιμή για την ακρίβεια με την οποία κάνει προβλέψεις το εκπαιδευμένο μοντέλο. Οι στήλες του σετ δεδομένων εξέτασης είναι ίδιες με αυτές του σετ εκπαίδευσης.

Κεφάλαιο 4

Επεξεργασία Δεδομένων

4.1 Προ-επεξεργασία δεδομένων



Σχήμα 4.1: Data Pre-Processing

4.1.1 Χρησιμότητα Προ-επεξεργασίας

Η προ-επεξεργασία των δεδομένων αποτελεί ένα αναγκαίο στάδιο στη διαδικασία εξόρυξης γνώσης. Στο κεφάλαιο αυτό παρουσιάζονται τα βασικά προβλήματα και οι αντίστοιχες τεχνικές αντιμετώπισης τους που σχετίζονται με την προ-επεξεργασία των δεδομένων. Γίνεται αναφορά στο καθαρισμό των δεδομένων, στην ολοκλήρωση τους και στο μετασχηματισμό τους. Επίσης παρουσιάζονται μέθοδοι μείωσης διαστάσεων και επιλογής σημαντικών χαρακτηριστικών. Για να έχουμε ποιοτικά αποτελέσματα από την εξόρυξη γνώσης χρειαζόμαστε ποιοτικά δεδομένα. Πιο συγκεκριμένα, τα δεδομένα δεν

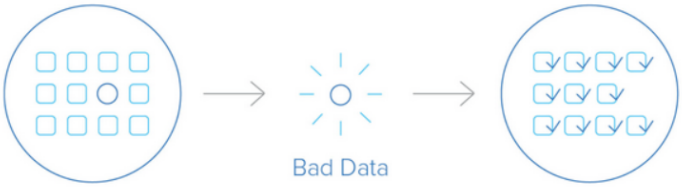
είναι ολοκληρωμένα, δηλαδή λείπουν τιμές, περιέχουν σφάλματα ή είναι αντιφατικά. Ύστερα από την προ-επεξεργασία τα δεδομένα πρέπει να είναι συνεπή, ενοποιημένα και ποιοτικά.

4.1.2 Καθαρισμός δεδομένων (Data Cleansing)

Ο καθαρισμός δεδομένων είναι η διαδικασία ανίχνευσης και διόρθωσης (ή αφαίρεσης) διεφθαρμένων ή ανακριβών αρχείων από ένα σετ, πίνακα ή βάση δεδομένων και αναφέρεται στην αναγνώριση ελλιπών, εσφαλμένων, ανακριβών ή άσχετων τμημάτων των δεδομένων και στη συνέχεια αντικατάσταση, ή διαγραφή των βρόμικων ή χονδροειδών δεδομένων. [30] Ο καθαριμός δεδομένων μπορεί να πραγματοποιηθεί διαδραστικά με εργαλεία κατακερματισμού δεδομένων ή ως επεξεργασία παρτίδας μέσω δέσμης ενεργειών.

Κύριες εργασίες καθαρισμού δεδομένων :

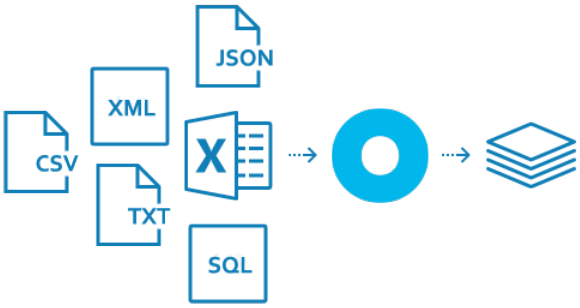
- Άμεση κτήση δεδομένων (data acquisition) και μετα-δεδομένων (metadata).
- Συμπλήρωση των ελλιπών τιμών (missing data).
- Μετατροπή των ονομαστικών τιμών (nomimal values) σε αριθμητικές τιμές (numerical values).
- Αναγνώριση των υπερβολικά υψηλών τιμών (outliers) και εξομάλυνση δεδομένων με θόρυβο.
- Διόρθωση ασυνεπειών στα δεδομένα.



Σχήμα 4.2: Data Cleansing

4.1.3 Ενοποίηση δεδομένων (Data Integration)

Η ενοποίηση δεδομένων περιλαμβάνει το συνδυασμό δεδομένων που διαμένουν σε διαφορετικές πηγές και παρέχει στους χρήστες μια ενιαία εικόνα αυτών. [31] Αυτή η διαδικασία καθίσταται σημαντική σε μια ποικιλία καταστάσεων που περιλαμβάνουν τόσο εμπορικές (όπως όταν δύο παρόμοιες εταιρείες πρέπει να συγχωνεύσουν τις βάσεις δεδομένων τους) όσο και επιστημονικές (συνδυάζοντας τα αποτελέσματα της έρευνας από διαφορετικά αποθετήρια βιοπληροφορικής, για παράδειγμα). Η ενσωμάτωση δεδομένων εμφανίζεται με αυξανόμενη συχνότητα καθώς ο όγκος (δηλαδή μεγάλα δεδομένα [32] και η ανάγκη ανταλλαγής υφιστάμενων δεδομένων εκρήγνυται [33]. Έχει γίνει το επίκεντρο εκτεταμένης θεωρητικής εργασίας και πολλά ανοιχτά προβλήματα παραμένουν ανεπίλυτα. Η ενσωμάτωση δεδομένων ενθαρρύνει τη συνεργασία μεταξύ εσωτερικών και εξωτερικών χρηστών.



Σχήμα 4.3: Data Integration

4.1.4 Μετασχηματισμός δεδομένων (Data transformation)

Ο μετασχηματισμός δεδομένων είναι η διαδικασία μετατροπής δεδομένων από μία μορφή σε άλλη, συνήθως από τη μορφή ενός συστήματος πηγής στην απαιτούμενη μορφή ενός συστήματος προορισμού. Ο μετασχηματισμός δεδομένων αποτελεί συστατικό στοιχείο των περισσότερων εργασιών ενσωμάτωσης δεδομένων και διαχείρισης δεδομένων, όπως η ανταλλαγή δεδομένων και η αποθήκευση δεδομένων.

Ο στόχος της διαδικασίας μετασχηματισμού δεδομένων είναι η εξαγωγή δεδομένων από μια πηγή, η μετατροπή τους σε μια χρησιμοποιήσιμη μορφή και η παράδοσή

τους σε έναν προορισμό. Αυτή η όλη διαδικασία είναι γνωστή ως ETL (Extract, Load, Transform). Κατά τη διάρκεια της φάσης εξαγωγής, τα δεδομένα αναγνωρίζονται και συλλέγονται από πολλές διαφορετικές τοποθεσίες ή πηγές σε ένα μόνο αποθετήριο. [34]

Τα δεδομένα που εξάγονται από την τοποθεσία προέλευσης είναι συχνά ωμά και δεν μπορούν να χρησιμοποιηθούν στην αρχική τους μορφή. Για να ξεπεραστεί αυτό το εμπόδιο, τα δεδομένα πρέπει να μετατραπούν. Αυτό είναι το βήμα της διαδικασίας ETL που προσθέτει την μεγαλύτερη αξία στα δεδομένα επιτρέποντάς τους να εξορύσσονται για επιχειρηματική ευφυΐα. Κατά τη διάρκεια του μετασχηματισμού, έχουν ληφθεί ορισμένα βήματα για τη μετατροπή τους στην επιθυμητή μορφή. Σε ορισμένες περιπτώσεις, τα δεδομένα πρέπει πρώτα να καθαριστούν πριν μετατραπούν. Ο καθαρισμός δεδομένων προετοιμάζει τα δεδομένα για μετασχηματισμό, επιλύοντας ασυνέπειες ή ελλείπουσες τιμές. Μόλις καθαριστούν τα δεδομένα, συμβαίνουν τα ακόλουθα βήματα στη διαδικασία μετασχηματισμού:

- Ανακάλυψη δεδομένων Data discovery → Το πρώτο βήμα στη διαδικασία μετασχηματισμού δεδομένων συνίσταται στην αναγνώριση και κατανόηση των δεδομένων στη μορφή πηγής τους. Αυτό επιτυγχάνεται συνήθως με τη βοήθεια ενός εργαλείου δημιουργίας προφίλ δεδομένων. Αυτό το βήμα σας βοηθά να αποφασίσετε τι πρέπει να συμβεί στα δεδομένα, προκειμένου να το φτάσετε στην επιθυμητή μορφή.

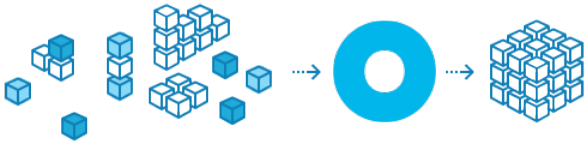
- Χαρτογράφηση δεδομένων Data mapping → Κατά τη διάρκεια αυτής της φάσης, προγραμματίζεται η πραγματική διαδικασία μετασχηματισμού.

- Δημιουργία κώδικα Generating code → Για να ολοκληρωθεί η διαδικασία μετασχηματισμού, πρέπει να δημιουργηθεί ένας κώδικας για την εκτέλεση της εργασίας μετασχηματισμού. Συχνά αυτοί οι κώδικες δημιουργούνται με τη βοήθεια ενός εργαλείου ή μιας πλατφόρμας μετασχηματισμού δεδομένων.

- Εκτέλεση του κώδικα Executing the code → Η διαδικασία μετασχηματισμού δεδομένων που έχει προγραμματιστεί και κωδικοποιηθεί τεθεί τώρα σε κίνηση και τα δεδομένα μετατρέπονται στην επιθυμητή έξοδο.

- Ανασκόπηση Review → Τα μετασχηματισμένα δεδομένα ελέγχονται για να βε-

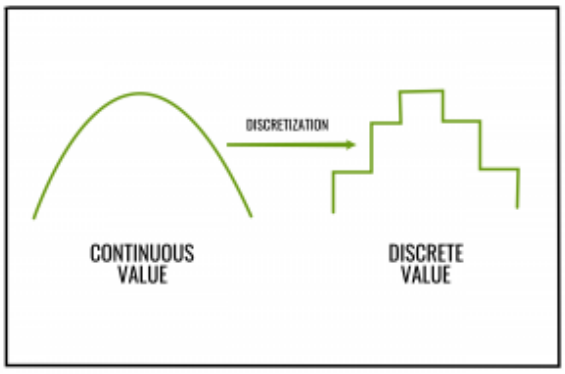
βαιωθείτε ότι έχουν μορφοποιηθεί σωστά.



Σχήμα 4.4: Data Transformation

4.1.5 Διακριτοποίηση δεδομένων (Data discretization)

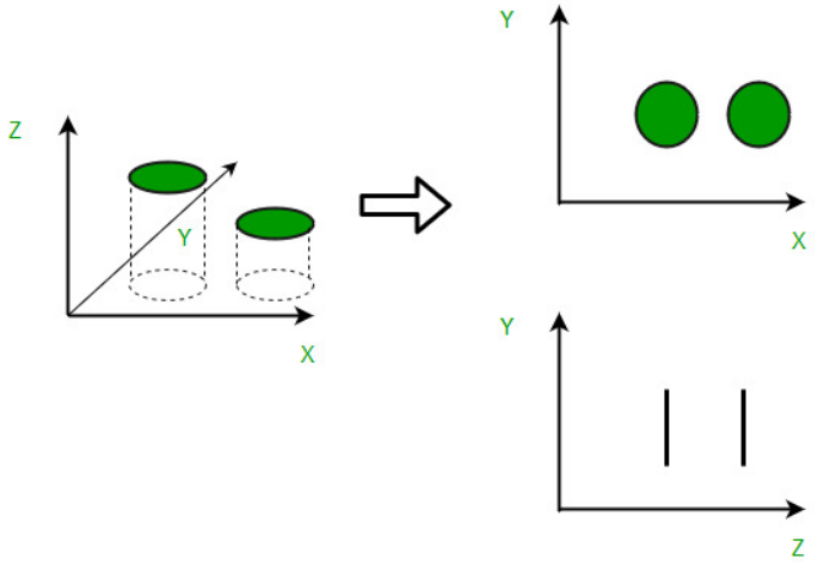
Στην στατιστική και την εκμάθηση μηχανών, η διακριτοποίηση αναφέρεται στη διαδικασία μετατροπής ή διαίρεσης συνεχών χαρακτηριστικών, χαρακτηριστικών ή μεταβλητών σε διακριτοποιημένα ή ονομαστικά χαρακτηριστικά, μεταβλητές ή διαστήματα. Αυτό μπορεί να είναι χρήσιμο όταν δημιουργούμε λειτουργίες μαζικής πιθανότητας - τυπικά, σε εκτίμηση πυκνότητας. Γενικά και η συσσώρευση πρόκειται για μια μορφή διακριτοποίησης, όπως στην κατασκευή ενός ιστόγραμμα. Όποτε τα συνεχή δεδομένα είναι διακριτοποιημένα, υπάρχει πάντα κάποιο ποσό σφάλματος διακριτοποίησης. Ο στόχος είναι να μειωθεί το ποσό σε επίπεδο που θεωρείται αμελητέο για τους σκοπούς μοντελοποίησης.



Σχήμα 4.5: Data Discretization

4.1.6 Μείωση διαστάσεων και δεδομένων (Dimensionality and Data reduction)

Στη στατιστική, στη μηχανή μάθηση και στη θεωρία της πληροφορίας, η μείωση των διαστάσεων είναι η διαδικασία μείωσης του αριθμού τυχαίων μεταβλητών που εξετάζονται με την απόκτηση ενός συνόλου βασικών μεταβλητών. Οι προσεγγίσεις μπορούν να χωριστούν σε επιλογή χαρακτηριστικών και εξαγωγή χαρακτηριστικών.



Σχήμα 4.6: Dimensionality Reduction

Επιλογή χαρακτηριστικών (Feature selection)

Οι προσεγγίσεις επιλογής λειτουργιών προσπαθούν να βρουν ένα υποσύνολο των μεταβλητών εισόδου (που ονομάζονται επίσης χαρακτηριστικά ή χαρακτηριστικά). Οι τρεις στρατηγικές είναι: η στρατηγική του φίλτρου (π.χ. κέρδος πληροφοριών), η στρατηγική περιτυλίγματος (π.χ. αναζήτηση που καθοδηγείται από την ακρίβεια) και η ενσωματωμένη στρατηγική (τα επιλεγμένα χαρακτηριστικά προσθέτουν ή αφαιρούνται ενώ παράγεται το μοντέλο βάσει σφαλμάτων πρόβλεψης).

Η ανάλυση δεδομένων, όπως η παλινδρόμηση ή η ταξινόμηση, μπορεί να γίνει στο μειωμένο χώρο με μεγαλύτερη ακρίβεια απ ́ότι στον αρχικό χώρο [3].

Εξαγωγή χαρακτηριστικών (Feature extraction)

Στη μηχανική μάθηση, στην αναγνώριση προτύπων και στην επεξεργασία εικόνων, η εξαγωγή χαρακτηριστικών ξεκινά από ένα αρχικό σύνολο μετρημένων δεδομένων και παράγει παραγόμενες αξίες (χαρακτηριστικά) που προορίζονται να είναι ενημερωτικές και μη περιττές, διευκολύνοντας τα επακόλουθα βήματα μάθησης και γενίκευσης και σε μερικές περιπτώσεις σε καλύτερες ανθρώπινες ερμηνείες. Η εξαγωγή χαρακτηριστικών σχετίζεται με τη μείωση των διαστάσεων.

Ας υποθέσουμε ότι τα δεδομένα που πρέπει να μειωθούν συνίστανται από πλειάδες ή φορείς δεδομένων που περιγράφονται από n χαρακτηριστικά. Η βασική ανάλυση των στοιχείων, Principal Components Analysis ή PCA (που ονομάζεται επίσης η μέθοδος Karhunen-Loeve ή K-L), αναζητά kn διαστάσεων φορείς που μπορούν να χρησιμοποιηθούν καλύτερα για την ερμηνεία των δεδομένων. Επομένως, τα αρχικά δεδομένα προβάλλονται σε πολύ μικρότερη κλίμακα, με αποτέλεσμα τη μείωση των διαστάσεων. Σε αντίθεση με την ποικιλία υποσυνόλων χαρακτηριστικών, η οποία μειώνει το μέγεθος των χαρακτηριστικών, διατηρώντας ένα υποσύνολο του αρχικού συνόλου των χαρακτηριστικών, ο PCA συνδυάζει την ουσία των χαρακτηριστικών δημιουργώντας ένα επιπλέον μικρότερο σύνολο μεταβλητών. Τα αρχικά δεδομένα μπορούν στη συνέχεια να προβάλλονται σε αυτό το μικρότερο σετ. Η PCA αποκαλύπτει συχνά σχέσεις που δεν είχαν προηγουμένως υποψιαστεί και έτσι επιτρέπει ερμηνείες που συνήθως δεν θα προέκυπταν. [35]

Η βασική διαδικασία είναι η εξής:

- Τα δεδομένα εισόδου ομαλοποιούνται έτσι ώστε κάθε χαρακτηριστικό να εμπίπτει στην ίδια κλίμακα. Αυτό το βήμα συμβάλλει στη διασφάλιση ότι τα χαρακτηριστικά με μεγάλους τομείς δεν θα κυριαρχούν σε ιδιότητες με μικρότερους τομείς.
- Ο PCA υπολογίζει k ορθονομικούς (και ορθογώνιους και κανονικοποιημένους) φορείς που παρέχουν μια βάση για τα κανονικοποιημένα δεδομένα εισόδου. Αυτοί είναι φορείς μονάδας που κάθε σημείο κατευθύνεται κατευθείαν προς τους άλλους. Αυτοί οι φορείς αναφέρονται ως τα κύρια συστατικά. Τα δεδομένα εισόδου είναι ένας γραμμικός συνδυασμός των κύριων στοιχείων.
- Τα βασικά συστατικά ταξινομούνται κατά σειρά μείωσης της σημασίας ή της α-

ντοχής. Χρησιμεύουν ως ένα νέο σύνολο αξόνων για τα δεδομένα, παρέχοντας σημαντικές πληροφορίες σχετικά με τη διακύμανση. Δηλαδή, οι διατεταγμένοι άξονες είναι τέτοιοι ώστε ο πρώτος άξονας να δείχνει τη μεγαλύτερη διακύμανση μεταξύ των δεδομένων, ο δεύτερος άξονας να δείχνει την επόμενη υψηλότερη διακύμανση και ούτω καθεξής. Για παράδειγμα, η Εικόνα 2.17 παρουσιάζει τα δύο πρώτα κύρια στοιχεία, Ψ1 και Ψ2, για το δεδομένο σύνολο δεδομένων που αντιστοιχούσαν αρχικά στους άξονες Χ1 και Χ2. Αυτές οι πληροφορίες βοηθούν στην αναγνώριση ομάδων ή μοτίβων μέσα στα δεδομένα.

- Επειδή τα εξαρτήματα ταξινομούνται σύμφωνα με τη φθίνουσα σειρά σπουδαιότητας, το μέγεθος των δεδομένων μπορεί να μειωθεί εξαλείφοντας τα ασθενέστερα συστατικά, δηλαδή εκείνα με μικρή διακύμανση. Χρησιμοποιώντας τα ισχυρότερα βασικά στοιχεία, θα πρέπει να είναι δυνατή η ανανέωση μιας καλής προσέγγισης των αρχικών δεδομένων.

Κεφάλαιο 5

Υλοποίηση εφαρμογής

Η εφαρμογή που αναπτύχθηκε στα πλαίσια αυτής της πτυχιακής εργασίας είχε ως κύριο σκοπό την εισαγωγή μίας σειράς δεδομένων του Gstore και την αναπαράσταση αυτών με κάποιων ειδών στατιστικών γραφημάτων όπως επίσης και η πρόβλεψη των εσόδων του κάθε χρήστη ξεχωριστά. Πέρα από αυτά όμως εκτελέστηκαν και μια σειρά από άλλες τεχνικές διαχείρισής δεδομένων που θα αναπτυχθούν με λεπτομέρεια παρακάτω όπως και οι τεχνολογίες που χρησιμοποιήθηκαν για την ανάπτυξη αυτής της εφαρμογής.

5.1 Γλώσσα προγραμματισμού Python

Σχήμα 5.1: Python

Η γλώσσα προγραμματισμού που επιλέχθηκε για την υλοποίηση αυτής της εφαρμογής είναι η Python. Η Python είναι μια ερμηνευμένη γλώσσα (interpreted) υψηλού επιπέδου (high-level) γενικής χρήσης (general-purpose). Δημιουργήθηκε από τον Guido van Rossum και κυκλοφόρησε για πρώτη φορά το 1991, η φιλοσοφία σχεδίασης της

Python τονίζει την αναγνωσιμότητα του κώδικα με τη αξιοσημείωτη χρήση σημαντικών κενών (significant whitespace). Οι γλωσσικές κατασκευές και η αντικειμενοστραφής προσέγγιση στοχεύουν να βοηθήσουν τους προγραμματιστές να γράψουν έναν σαφή, λογικό κώδικα για μικρά και μεγάλα έργα.

5.1.1 Χαρακτηριστικά και φιλοσοφία

Η Python είναι μια γλώσσα προγραμματισμού πολλαπλών υποδειγμάτων. Ο προγραμματισμός με αντικειμενοστραφούς προγραμματισμό και ο δομημένος προγραμματισμός υποστηρίζονται πλήρως και πολλά από τα χαρακτηριστικά του υποστηρίζουν τον λειτουργικό προγραμματισμό και τον προγραμματισμό προσανατολισμένων στις πτυχές (συμπεριλαμβανομένου του μεταπρογράμματος και των μεταοβθεστς (μαγικές μεθόδους». Πολλά άλλα παραδείγματα υποστηρίζονται μέσω επεκτάσεων, συμπεριλαμβανομένου του σχεδιασμού με σύμβαση και λογικού προγραμματισμού. Η Python χρησιμοποιεί δυναμική πληκτρολόγηση και συνδυασμό συνδυασμού αναφορών και συλλέκτη σκουπιδιών που ανιχνεύει κύκλους για διαχείριση μνήμης. Διαθέτει επίσης δυναμική ανάλυση ονομάτων (καθυστερημένη σύνδεση), η οποία δεσμεύει τα ονόματα μεθόδων και μεταβλητών κατά την εκτέλεση του προγράμματος. Ο σχεδιασμός της Python προσφέρει κάποια υποστήριξη για λειτουργικό προγραμματισμό στην παράδοση Lisp. Έχει φίλτρο, χάρτη και μειώνει τις λειτουργίες. κατανόηση λίστας, λεξικά, σύνολα και εκφράσεις γεννήτριας. Η τυπική βιβλιοθήκη έχει δύο ενότητες (itertools και functools) που εφαρμόζουν λειτουργικά εργαλεία δανεισμένα από Haskell και Standard ML. Η βασική φιλοσοφία της γλώσσας συνοψίζεται στο έγγραφο Zen of Python (PEP 20).

5.1.2 Βιβλιοθήκες της Python

Όπως σε κάθε γλώσσα προγραμματισμό, έτσι και στην Python ένας ορισμός της βιβλιοθήκης μπορεί να είναι ο ακόλουθος: μια συλλογή προ-συγκεντρωμένων και μη πτητικών ρουτινών που χρησιμοποιούνται από τα προγράμματα. Αυτές οι ρουτίνες, μερικές φορές αποκαλούμενες ενότητες, μπορούν να περιλαμβάνουν δεδομένα διαμόρφωσης, τεκμηρίωση, πρότυπα μηνυμάτων, υπορουτίνες, κλάσεις, τιμές ή προδιαγραφές

Σχήμα 5.2: Python Libraries

τύπου. Το μεγαλύτερο πλεονέκτημα μιας βιβλιοθήκης και ο λόγος για τη χρήση της είναι η δυνατότητα επαναχρησιμοποίησης της συμπεριφοράς. Όταν χρησιμοποιείτε μια βιβλιοθήκη, το πρόγραμμα αποκτά τη συμπεριφορά που εφαρμόζεται μέσα σε αυτή τη βιβλιοθήκη χωρίς να χρειάζεται ο προγραμματιστής να ξαναγράψει την ίδια τη συμπεριφορά. Τέτοιου είδους βιβλιοθήκες υπάρχουν σε αποθετήρια λογισμικού όπως το PyPI. Το PyPI είναι ένα αποθετήριο λογισμικού για τη γλώσσα προγραμματισμού Python. Το PyPI βοηθά στην εύρεση και να εγκαταστήσετε λογισμικό που αναπτύχθηκε και μοιράστηκε από την κοινότητα της Python. Οι συντάκτες πακέτων χρησιμοποιούν την PyPI για τη διανομή του λογισμικού τους και για την εγκατάσταση αυτών χρησιμοποιείτε η εντολή pip μέσω της κονσόλας (CMD ή Terminal). Παρακάτω θα αναλύσουμε τις πιο σημαντικές βιβλιοθήκες που χρησιμοποιήθηκαν κατά την ανάπτυξη της παρούσας εφαρμογής.

Pandas

Είναι μια βιβλιοθήκη λογισμικού γραμμένη για τη γλώσσα προγραμματισμού Python για χειρισμό και ανάλυση δεδομένων. Συγκεκριμένα, προσφέρει δομές δεδομένων και λειτουργίες για τον χειρισμό αριθμητικών πινάκων και χρονοσειρών. Το όνομα προέρχεται από τον όρο 'δεδομένα πίνακα' "panel data", έναν οικονομετρικό όρο για σύνολα δεδομένων που περιλαμβάνουν παρατηρήσεις για πολλαπλές χρονικές περιόδους για τα ίδια άτομα. Η βιβλιοθήκη είναι εξαιρετικά βελτιστοποιημένη για απόδοση, με κρίσιμες διαδρομές κώδικα γραμμένες σε Cython ή C. Οι κύριες λειτουργίες της βιβλιοθήκης είναι η εξής:

Σχήμα 5.3: Pandas

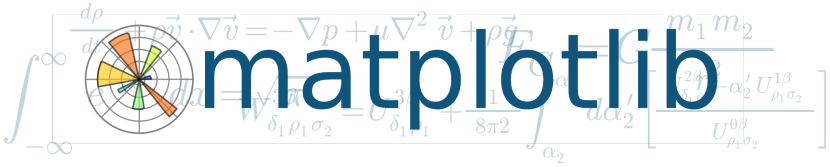
- Αντικείμενο DataFrame για χειρισμό δεδομένων με ενσωματωμένη ευρετηρίαση.
- Εργαλεία για την ανάγνωση και τη γραφή δεδομένων μεταξύ δομών δεδομένων εντός μνήμης και διαφορετικών μορφών αρχείων.
- Ευθυγράμμιση δεδομένων και ολοκληρωμένη διαχείριση δεδομένων που λείπουν.
- Επαναδιαμόρφωση και περιστροφή των συνόλων δεδομένων.
- Ετικετοποίηση με βάση το τεμαχισμό, φανταχτερή ευρετηρίαση και υποσύνολο μεγάλων συνόλων δεδομένων.
- Εισαγωγή και διαγραφή στήλης δομής δεδομένων.
- Ομαδοποίηση με τον κινητήρα, επιτρέποντας τις λειτουργίες split-apply-combine σε σύνολα δεδομένων.
- Το σύνολο δεδομένων συγχωνεύεται και ενώνεται.
- Ιεραρχικός άξονας ευρετηρίασης για την εργασία με δεδομένα μεγάλης διαστάσεως σε μια δομή δεδομένων χαμηλότερων διαστάσεων.
- Λειτουργικότητα χρονολογικών σειρών: Γενική χρονική περίοδος [4] και μετατροπή συχνότητας, μετακίνηση στατιστικών παραθύρων, γραμμικές παλινδρομήσεις κινούμενων παραθύρων, μετατόπιση ημερομηνίας και υστέρηση.
- Παρέχει φιλτράρισμα δεδομένων.

Numpy

Σχήμα 5.4: Numpy

Είναι μια βιβλιοθήκη για τη γλώσσα προγραμματισμού Python, προσθέτοντας υποστήριξη για μεγάλες, πολυδιάστατες συστοιχίες, μαζί με μια μεγάλη συλλογή μαθηματικών λειτουργιών υψηλού επιπέδου για να λειτουργούν σε αυτές τις συστοιχίες. Ο πρόγονος του NumPy, Numeric, δημιουργήθηκε αρχικά από τον Jim Hugunin με τη συμβολή πολλών άλλων προγραμματιστών. Το 2005, ο Travis Oliphant δημιούργησε το NumPy με την ενσωμάτωση χαρακτηριστικών του ανταγωνιζόμενου Numarray σε Numeric, με εκτεταμένες τροποποιήσεις. Το NumPy είναι λογισμικό ανοιχτού κώδικα και έχει πολλούς συνεργάτες.

Matplotlib



Σχήμα 5.5: Matplotlib

Το Matplotlib είναι μια βιβλιοθήκη σχεδίασης για τη γλώσσα προγραμματισμού Python και την αριθμητική επέκταση NumPy. Με την βοήθεια της βιβλιοθήκης αυτής έχουμε την δυνατότητα να δημιουργήσουμε plots, histograms, power spectra, bar charts, errorcharts, scatterplots κλπ.

Σχήμα 5.6: Plotly

Plotly

Η γραφική βιβλιοθήκη Python (plotly.py) είναι μια διαδραστική βιβλιοθήκη σχεδίασης ανοιχτού κώδικα που υποστηρίζει πάνω από 40 μοναδικούς τύπους χαρτών που καλύπτουν ένα ευρύ φάσμα στατιστικών, οικονομικών, γεωγραφικών, επιστημονικών και τρισδιάστατων περιπτώσεων χρήσης.

Seaborn

Σχήμα 5.7: Seaborn

Το Seaborn είναι μια βιβλιοθήκη για τη δημιουργία στατιστικών γραφικών στην Python. Είναι χτισμένο πάνω από το matplotlib και είναι στενά συνδεδεμένο με τις δομές δεδομένων του pandas. Παρακάτω είναι μερικές από τις λειτουργίες που προσφέρει:

- Ένα API προσανατολισμένο στο σύνολο δεδομένων για την εξέταση σχέσεων μεταξύ πολλαπλών μεταβλητών
- Ειδική υποστήριξη για τη χρήση κατηγοριολογικών μεταβλητών για την εμφάνιση παρατηρήσεων ή συγκεντρωτικών στατιστικώνΕπιλογές για την απεικόνιση μο-

νοδιάστατων ή διμερών διανομών και για τη σύγκρισή τους μεταξύ υποσυνόλων δεδομένων

- Αυτόματη εκτίμηση και σχεδίαση μοντέλων γραμμικής παλινδρόμησης για διαφορετικές μεταβλητές που εξαρτώνται από το είδος
- Βολική θέα στη συνολική δομή σύνθετων συνόλων δεδομένων
- Υψηλού επιπέδου αφαιρέσεις για τη διαμόρφωση πλέγματος πολλαπλών οικόπεδων που σας επιτρέπουν να δημιουργήσετε εύκολα σύνθετες απεικονίσεις
- Συνοπτικός έλεγχος στο στυλ του ματπλοτλιβ με διάφορα ενσωματωμένα θέματα
- Εργαλεία για την επιλογή παλέτας χρωμάτων που αποκαλύπτουν πιστά τα πρότυπα στα δεδομένα σας

Το Seaborn στοχεύει να κάνει την απεικόνιση κεντρικό μέρος της διερεύνησης και κατανόησης των δεδομένων. Οι λειτουργίες σχεδίασης με βάση το σύνολο δεδομένων λειτουργούν σε πλαίσια δεδομένων και συστοιχίες που περιέχουν ολόκληρα σύνολα δεδομένων και εκτελούν εσωτερικά την απαραίτητη σημασιολογική χαρτογράφηση και στατιστική συσσωμάτωση για την παραγωγή ενημερωτικών γραφικών.

LightGBM

Microsoft
LightGBM

Σχήμα 5.8: LightGBM

Το LightGBM είναι ένα πλαίσιο ενίσχυσης κλίσης που χρησιμοποιεί αλγόριθμους μάθησης βασισμένους σε δέντρα αποφάσεων. Είναι σχεδιασμένο για να είναι κατανεμημένο και αποδοτικό με τα ακόλουθα πλεονεκτήματα:

- Ταχύτερη εκπαίδευση και υψηλότερη απόδοση.

- Λιγότερη χρήση της μνήμης.
- Καλύτερη ακρίβεια.
- Υποστήριξη παράλληλης μάθησης και μάθησης GPU.
- Δυνατότητα χειρισμού δεδομένων μεγάλης κλίμακας.

Μερικές από τις μετρήσεις που υποστηρίζει το LightGBM είναι οι εξής:

- Ποσοστό σφάλματος ταξινόμησης (Classification error rate)
- Κατανομή (Poisson)
- (Kullback-Leibler)
- L1 loss, L2 loss
- Log loss

SkLearn (Scikit-Learn)

Σχήμα 5.9: Scikit-Learn

Η Scikit-learn είναι μια βιβλιοθήκη στην Python που παρέχει πολλούς αλγόριθμους μάθησης χωρίς επίβλεψη και επίβλεψη. Είναι χτισμένο πάνω στις τεχνολογίες που προαναφέραμε ήδη, όπως το NumPy, τα Pandas και το Matplotlib. Οι λειτουργίες που παρέχει η Scikit-learn είναι οι εξής:

- Οπισθοδρόμηση (Regression), συμπεριλαμβανομένης της γραμμικής και λογικής παλινδρόμησης.
- Κατάταξη (Classification), συμπεριλαμβανομένων των K-Nearest Neighbours.
- Ομαδοποίηση (Clustering), συμπεριλαμβανομένων των K-Means και K-Means++.
- Επιλογή μοντέλου (Model selection).

- Προεπεξεργασία (Preprocessing), συμπεριλαμβανομένης της Κανονικοποίησης Min-Max.

5.2 Γραφικό περιβάλλον

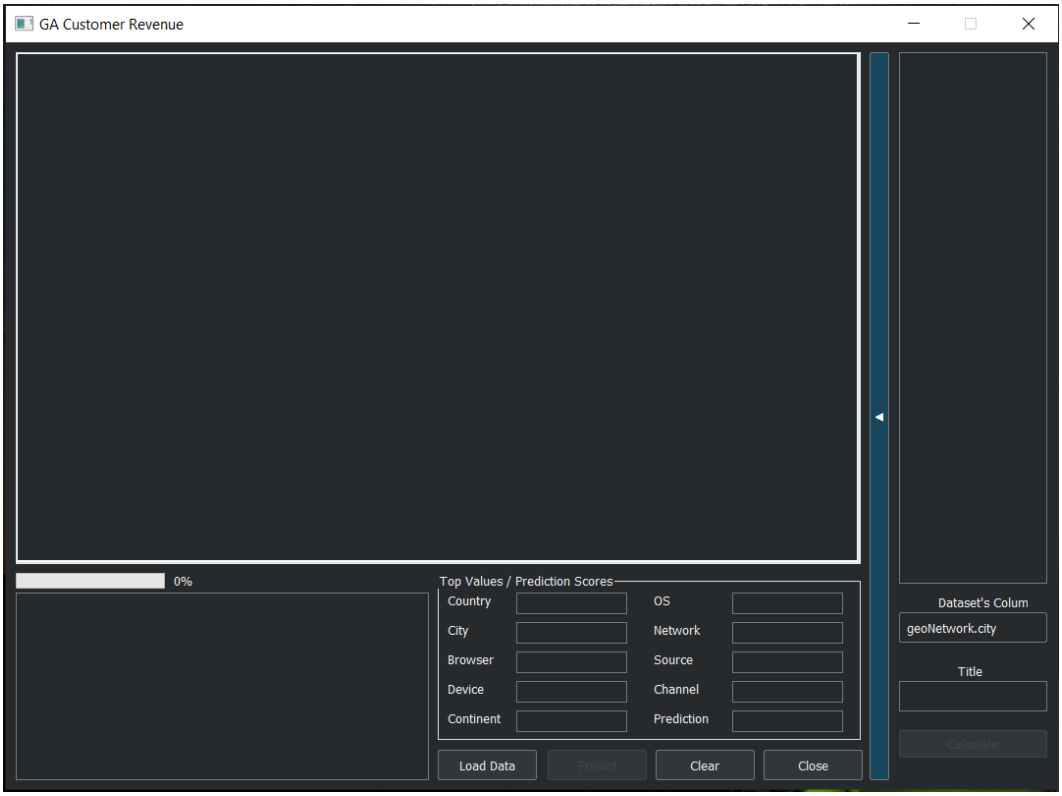
Σχήμα 5.10: PyQt

Για την δημιουργία του γραφικού περιβάλλοντος (GUI) χρησιμοποιήθηκε το PyQt5. Το PyQt είναι ένα εργαλείο GUI widgets. Πρόκειται για μια διασύνδεση Python για το Qt, μία από τις πιο ισχυρές και δημοφιλείς βιβλιοθήκες GUI πολλαπλών πλατφορμών. Το PyQt αναπτύχθηκε από την RiverBank Computing Ltd. Το PyQt API είναι ένα σύνολο μονάδων που περιέχουν μεγάλο αριθμό κατηγοριών και λειτουργιών. Ενώ η μονάδα Χtδρε περιέχει λειτουργίες εκτός GUI για εργασία με αρχεία και καταλόγους, η μονάδα QtGUI περιέχει όλα τα γραφικά στοιχεία ελέγχου. Επιπλέον, υπάρχουν λειτουργικές μονάδες για εργασία με XML (QtXml), SVG (QtSvg) και SQL (QtSql). Το PyQt είναι συμβατό με όλα τα δημοφιλή λειτουργικά συστήματα, συμπεριλαμβανομένων των Windows, Linux και Mac OS.

5.3 Επισκόπηση εφαρμογής

Η εφαρμογή αποτελείτε από τρεις βασικές ενότητες όπως διακρίνουμε και στο σχήμα 5.11. Η πρώτη και πιο βασική ενότητα είναι το κεντρικό πλαίσιο στο οποίο εμφανίζονται τα γραφήματα το καθένα σε διαφορετική καρτέλα. Η δεύτερη ενότητα είναι αυτή

κάτω από το πλαίσιο γραφημάτων η οποία χωρίζεται σε δύο υπό-ενότητες, την πρόοδο της εφαρμογής και τις ανώτερες τιμές όπως και τις τιμές πρόβλεψης. Στην τρίτη και τελευταία ενότητα υπάρχει ένα πλαίσιο αποτελεσμάτων που έχουν να κάνουν ειδικά με το σετ δεδομένων, όπως παραδείγματος χάρη τις τιμές που μετρήθηκαν ανά γράφημα. Επίσης στο κάτω μέρος τις τελευταίας ενότητας υπάρχουν δύο πλαίσια κειμένου και ένα κουμπί με τα οποία δίνεται η δυνατότητα να επιλεχθεί μία στήλη από το σετ δεδομένων και να δημιουργηθεί ένα γράφημα με την προεπιλεγμένη στήλη του σετ ¨συνολικές συναλλαγές¨ (totals.transactionRevenue) και με τον τίτλο που θα δοθεί στο δεύτερο πλαίσιο κειμένου. Παρακάτω θα εξετάσουμε και θα αναλύσουμε τον τρόπο με τον οποίο δουλεύει η παρούσα πτυχιακή.



Σχήμα 5.11: GA Customer Revenue Prediction

5.3.1 Φόρτωση δεδομένων

Όπως είναι φανερό στο σχήμα 5.12 πατώντας το κουμπί ¨φόρτωση δεδομένων¨ (Load Data) η εφαρμογή εντοπίζει το αρχείο train.csv το οποίο είναι το σετ εκπαίδευσης της εφαρμογής, ελέγχει το μέγεθός του και ξεκινώντας το διάβασμα του φορτώνει την μπάρα προόδου (progress bar). Σε αυτό το σημείο να τονίσουμε ότι κατά την διάρκεια

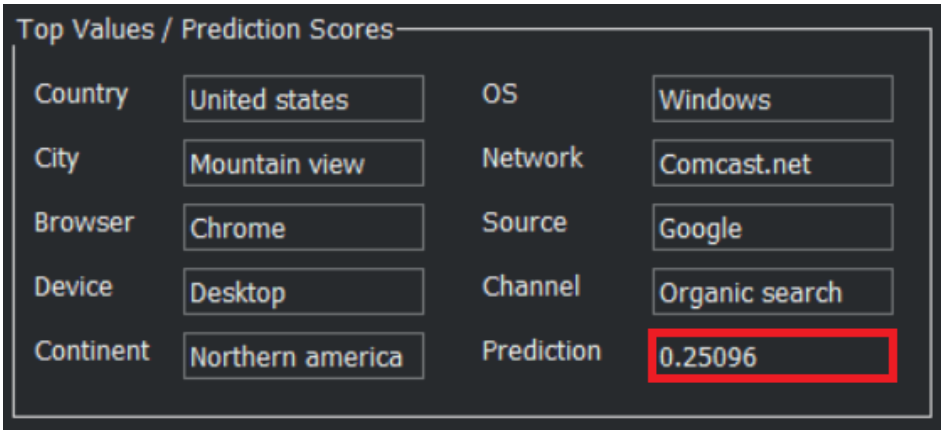
φόρτωσης του σετ δεδομένων η εφαρμογή στο παρασκήνιο διαμορφώνει τις τιμές (Normalazing), γεμίζει τις μηδενικές τιμές (NAN) ανάλογα το είδος τιμών που εκπροσωπεί η στήλη π.χ. (ακεραίους (int) τις συμπληρώνει με το 0, τύποι δεδομένων αληθείας (booleans) τις συμπληρώνει με το ¨ψευδής¨ (false), χαρακτήρες (characters) τις συμπληρώνει με το κενό (¨ ¨), κινητής υποδιαστολής (floating-point numbers) τις συμπληρώνει με το 0.0 και αλφαριθμητικές συμβολοσειρές (strings) τις συμπληρώνει επίσης με κενό (¨ ¨)), ξεφορτώνεται τις στατικές στήλες (Constant Columns) που αυτό σημαίνει αν μια στήλη έχει σε όλες τις γραμμές της την ίδια ακριβώς τιμή τότε είναι στατική και τέλος την στήλη ημερομηνίας (Date) την διαμορφώνει από την μορφή ημερομηνίας των linux σε κανονική μορφή 'DD - MM – YYYY' και ύστερα σε διαφορετικές στήλες [(Day) μέρα, (Month) μήνας και (Year) χρόνος].

Αμέσως μετά την διαδικασία αυτή η εφαρμογή αποθηκεύει πλέον το καθαρισμένο-διαμορφωμένο σετ δεδομένων σε ένα ξεχωριστό αρχείο *.csv με όνομα train_cleaned.csv με αυτόν τον τρόπο στην επόμενη χρήση της θα φορτώσει πλέον το καθαρισμένο σετ χωρίς να χρειαστεί ξανά η παραπάνω διαδικασία. Το επόμενο βήμα είναι η αναπαράσταση γραφημάτων και η ανάδειξη των συχνότερων τιμών που εμφανίζονται στο σετ δεδομένων.

Σχήμα 5.12: GA Customer Revenue Prediction

5.3.2 Πρόβλεψη (Prediction)

Ύστερα από την φόρτωση δεδομένων και την εμφάνιση των διαγραμμάτων ενεργοποιείτε το κουμπί της πρόβλεψης (Predict). Με αυτό η εφαρμογή εντοπίζει και φορτώνει το αρχείο test.csv το οποίο είναι το σετ δεδομένων εξέτασης, εκτελώντας ακριβώς τα ίδια βήματα με αυτά του σετ εκπαίδευσης (καθαρισμός, διαμόρφωση και τέλος αποθήκευση με όνομα test_cleaned.csv. Ο κύριος σκοπός του Predict όμως είναι να εκπαιδεύσει ένα μοντέλο το οποίο να μπορεί να προβλέψει από το σετ εξέτασης ποιοι από τους χρήστες θα επιφέρουν κέρδος. Επίσης στο τέλος της εκπαίδευσης θα μας επιστρέψει και μία τιμή εκτίμησης για το πόσο ορθά εκπαιδεύτηκε το μοντέλο με την μέθοδο του μέσου τετραγωνικού σφάλματος (Root Mean Squared Error [RMSE]). Το RMSE είναι το μέσο τετραγωνικό σφάλμα ενός εκτιμητή και μετρά τον μέσο όρο των τετραγώνων των σφαλμάτων, δηλαδή τη μέση τετραγωνική διαφορά μεταξύ των εκτιμώμενων τιμών και των πραγματικών τιμών.



Σχήμα 5.13: Root Mean Squared Error Value (R-Squared)

Αυτή η τιμή ονομάζεται R-Squared και κυμαίνεται από 0 έως 1 και αναφέρονται συνήθως ως ποσοστά από 0% έως 100%. Ο τύπος για το R-Squared είναι ο εξής:

$$R^2 = 1 - \frac{\text{Explained Variation}}{\text{Total Variation}}$$

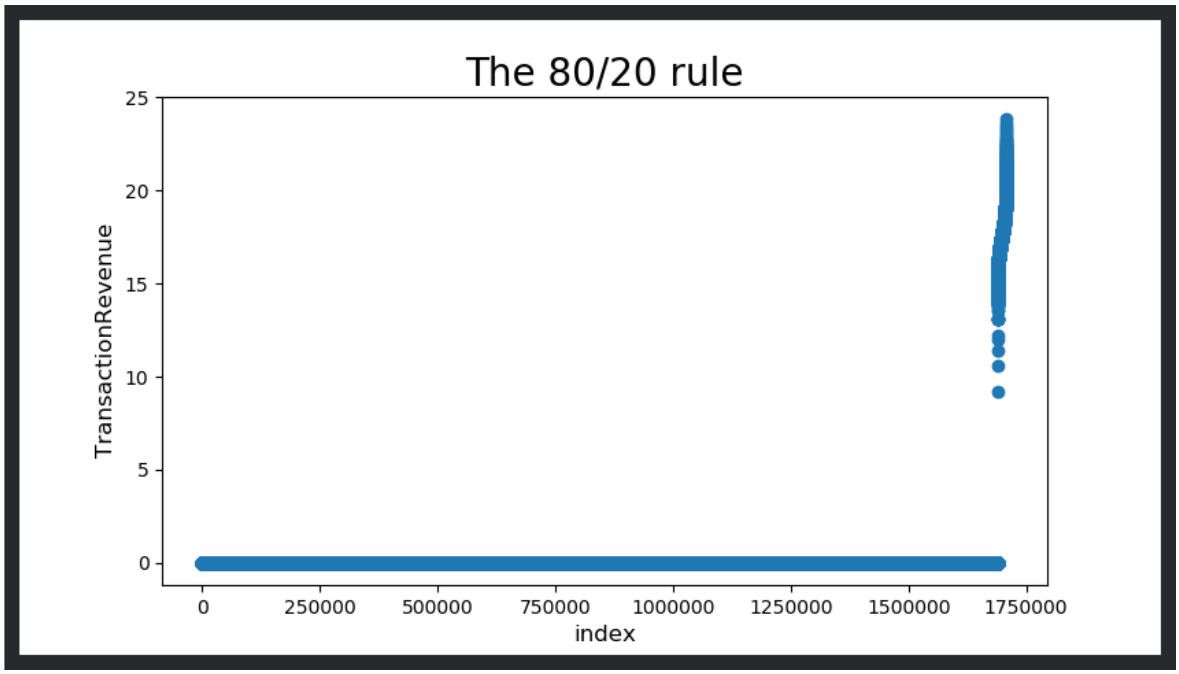
Άρα στην περίπτωσή μας 0.865 ή 86.5%

Κεφάλαιο 6

Αποτελέσματα

6.0.1 Perato, Κανόνας 80/20

Perato



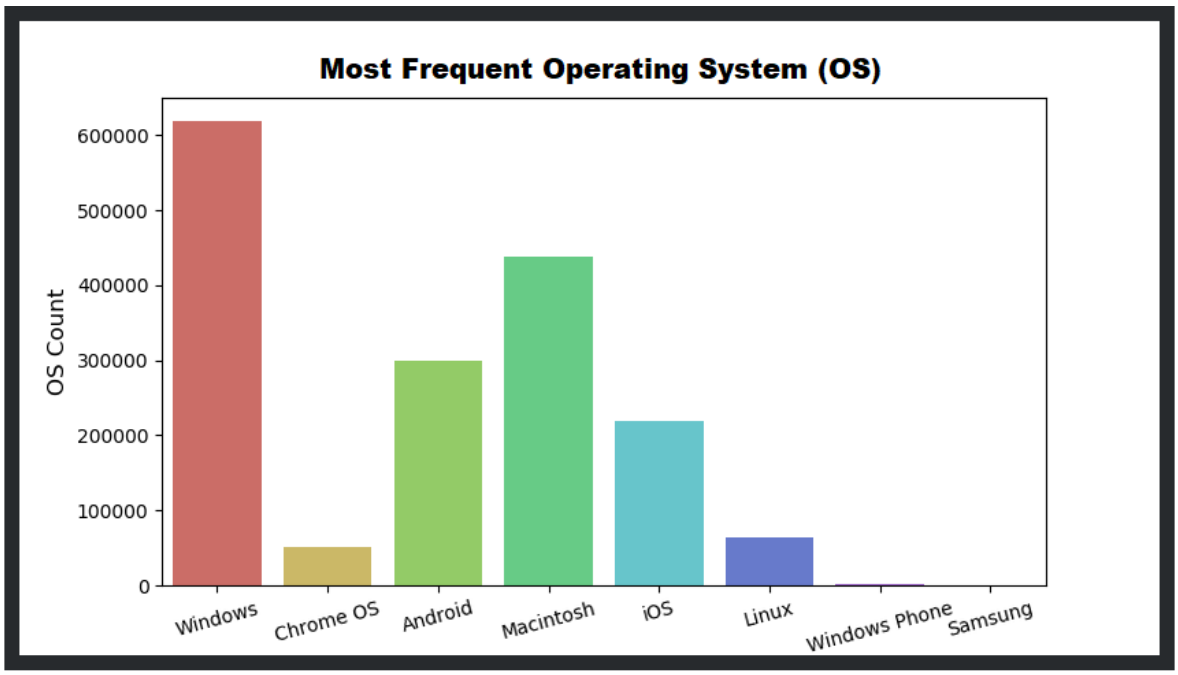
Σχήμα 6.1: Perato (Rule 80/20)

Στο σχήμα 6.1 είναι φανερό ότι ισχύει ο κανόνας Perato (80/20). Στο σετ δεδομένων υπήρχαν περίπου 1.750.000 εγγραφές από τις οποίες ένα πολύ μικρό ποσοστό επέφερε κέρδος στην επιχείρηση (περίπου 23%). Ως αποτέλεσμα ο κανόνας 80/20 ισχύει δηλώνοντας ότι, για πολλά γεγονότα, περίπου το 80% των επιπτώσεων προέρχεται α-

πό το 20% των αιτιών και στην δικιά μας περίπτωση ισχύει ότι το το 77% των κερδών προέρχεται από το 23% των πελατών.

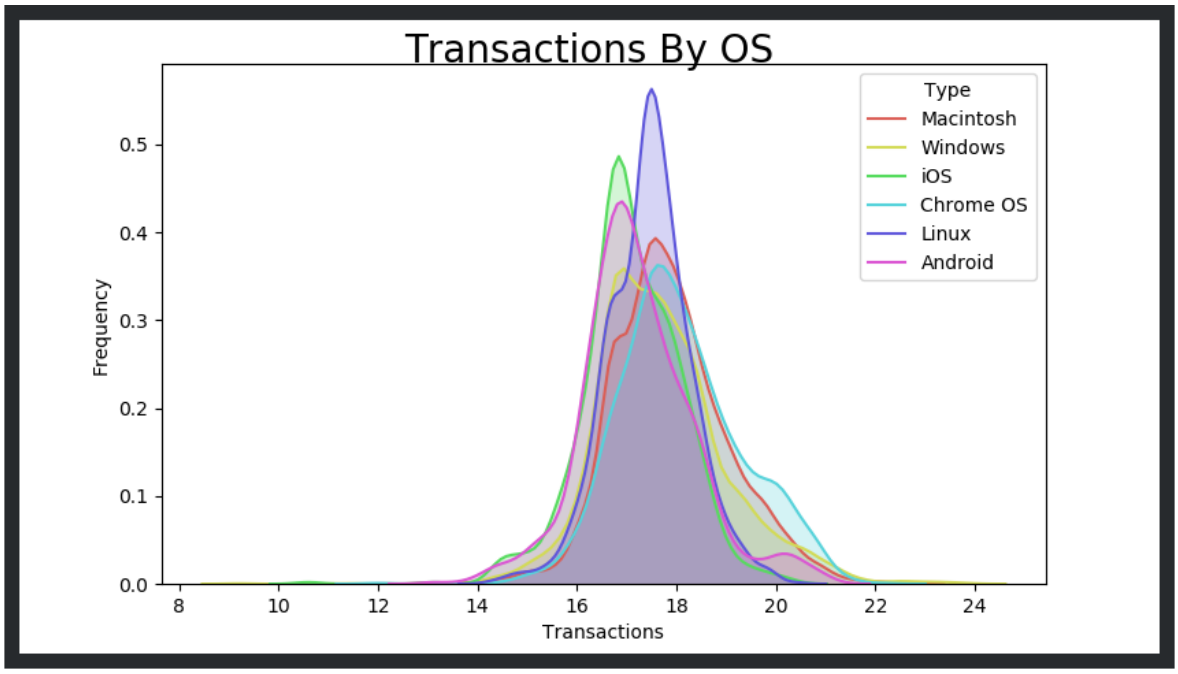
6.0.2 Γραφήματα με βάση τις προτιμήσεις των χρηστών

Operating System



Σχήμα 6.2: OS Count

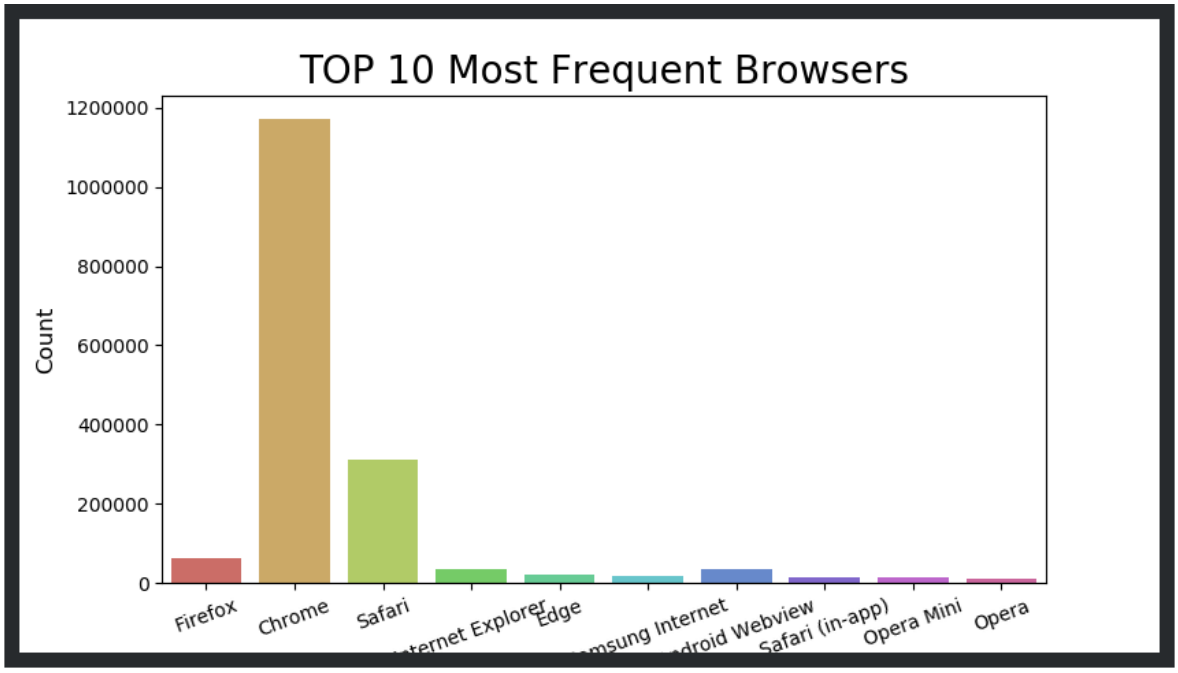
Στο σχήμα 6.2 παρατηρούμε ότι το λειτουργικό σύστημα (OS) που χρησιμοποιούν η πλειοψηφία των χρηστών που επισκέπτονται το Google Merchandise Store είναι το λειτουργικό σύστημα Windows με αριθμό περίπου 610000. Αμέσως μετά είναι το Macintosh με περίπου στους 450000 χρήστες και στην συνέχεια το λειτουργικό σύστημα των κινητών τηλεφώνων Android με 300000 χρήστες. Παρατηρούμε επίσης ότι το iOS ακολουθεί με λίγο παραπάνω από 200000 χρήστες και στο τέλος με μικρή διαφορά το Linux με το Chrome OS που έχουν αριθμό χρηστών κάτω από 100000.



Σχήμα 6.3: Revenue Per OS

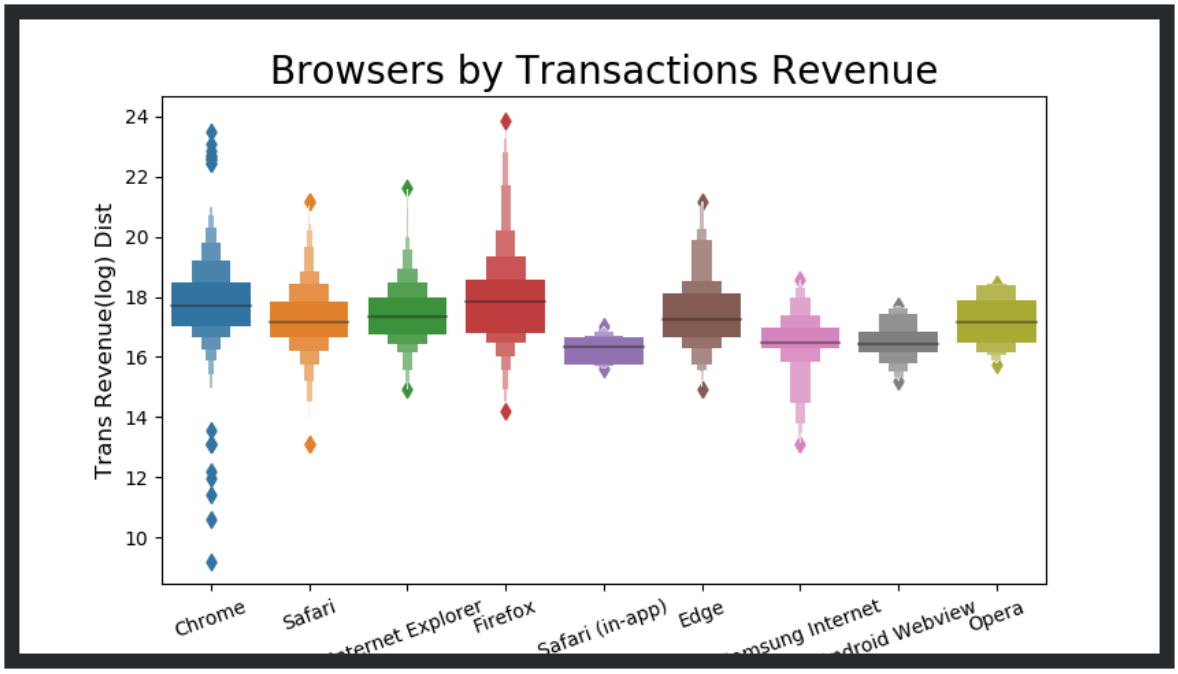
Αφού εξετάσαμε το σχήμα 6.2 στην συνέχεια θα αναλύσουμε πόσες συναλλαγές πραγματοποιήθηκαν ανά λειτουργικό σύστημα που αναπαρίσταται γραφικά στο σχήμα 6.3. Παρατηρούμε ότι ενώ το λειτουργικό σύστημα Linux και iOS έχουν από τους μικρότερους αριθμούς χρηστών η συχνότητα συναλλαγών που πραγματοποιούνται από αυτά είναι η μεγαλύτερη από όλα τα υπόλοιπα. Η συχνότητα και των δύο είναι κοντά στο 50% με το πρώτο να είναι κατά ελάχιστα περισσότερο. Στην συνέχεια έχουμε Android με Macintosh στην συχνότητα γύρω στο 40% και τέλος Windows με Chrome OS στην συχνότητα γύρω στο 35%.

Browser Usage



Σχήμα 6.4: Browser Count

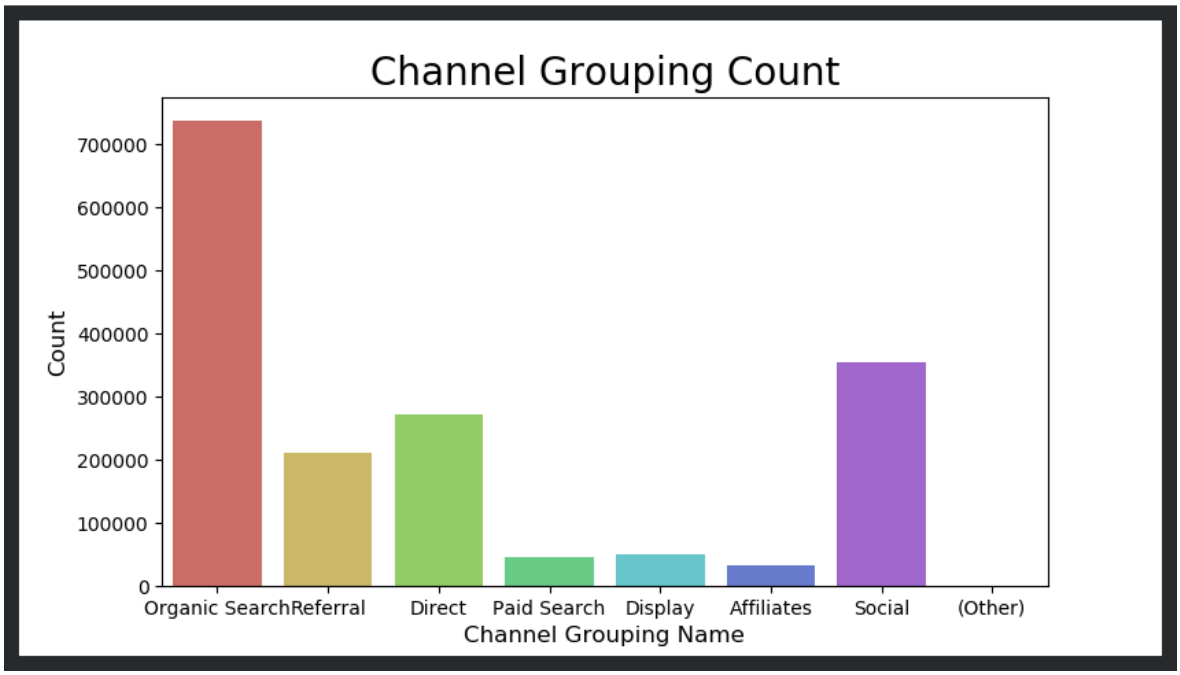
Στο σχήμα 6.4 παρατηρούμε ποιά λογισμικά πλοήγησης χρησιμοποιούν οι χρήστες του Gstore. Είναι φανερό ότι η πλειοψηφία χρησιμοποιεί το Chrome με αριθμό γύρω στους 1200000. Δεύτερος πιο συχνός φυλλομετρητής είναι το Safari αλλά με πολύ μικρότερο αριθμό από τον πρώτο στους 300000 χρήστες. Τέλος, το Firefox και Android Webview με αριθμό χρηστών μικρότερο από τις 50.000.



Σχήμα 6.5: Revenue Per Browser

Ύστερα από την εξέταση του σχήματος 6.4 που είχαν ως βάση την χρήση των φυλλομετρητών, στο σχήμα 6.5 αναλύονται οι συναλλαγές που πραγματοποιήθηκαν από αυτούς. Παρατηρούμε ότι ενώ το Firefox άνηκε στις χαμηλότερες ομάδες χρήσης, έχουμε τις περισσότερες συναλλαγές από αυτό. Στην συνέχεια το Chrome που είναι λογικό διότι άνηκε στην πλειοψηφία χρήσης και τέλος σχεδόν ισόβαθμα το Safari, Edge και Internet Explorer.

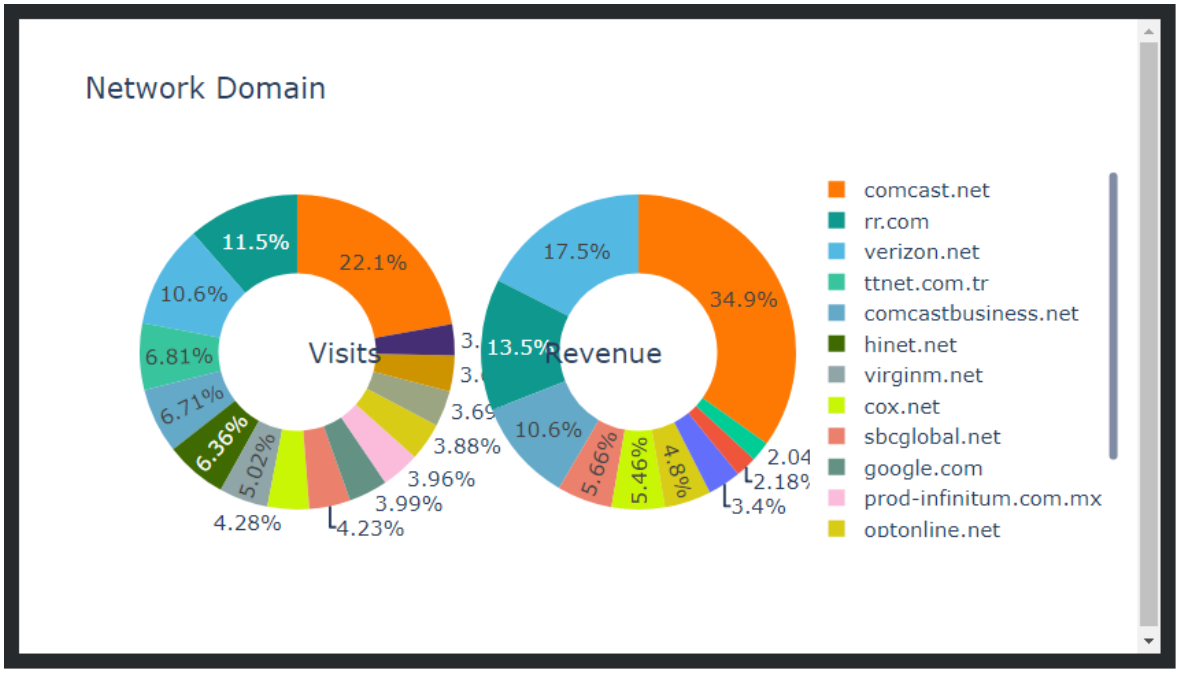
Channel Usage



Σχήμα 6.6: Channel Count

Το σχήμα 6.6 αναπαριστά τις πηγές από τις οποίες βρήκαν οι χρήστες το κατάστημα Gstore. Παρατηρούμε ότι οι περισσότεροι χρήστες που επισκέφτηκαν το κατάστημα, με αριθμό πάνω από 700000, το βρήκαν από οργανική αναζήτησή (Organic Search), δηλαδή η αναζήτηση μίας συγκεκριμένης λέξης, η φράσης κλειδί σε οποιαδήποτε μηχανή αναζήτησης. Αμέσως μετά είναι από κοινωνικά μέσα με αριθμό επισκέψεων πάνω από 350000. Επίσης παρατηρούμε ότι βρίσκονται πολύ κοντά η επίσκεψη από παραπομπή (Referral) με 200000, που σημαίνει ότι ο χρήστης επισκέφθηκε το κατάστημα από άλλους ιστότοπους, αποκλείοντας προφανώς τις μηχανές αναζήτησης, και η απευθείας (Direct) με 250000, που σημαίνει ότι οι χρήστες πληκτρολόγησαν τη διεύθυνση (URL) και υποδεικνύει επίσης άτομα που είχαν το κατάστημα στους σελιδοδείκτες τους. Τέλος παρατηρούμε ότι οι επισκέψεις από μισθωτά μέσα όπως διαφημίσεις (Display), καμπάνιες που πληρώνονται ανά επίσκεψη (Paid Search) και κυκλοφορία λόγω μάρκετινγκ θυγατρικών (Affiliates) έχουνε πολύ μικρή απόδοση με τιμές που κυμαίνονται κάτω από τις 50000.

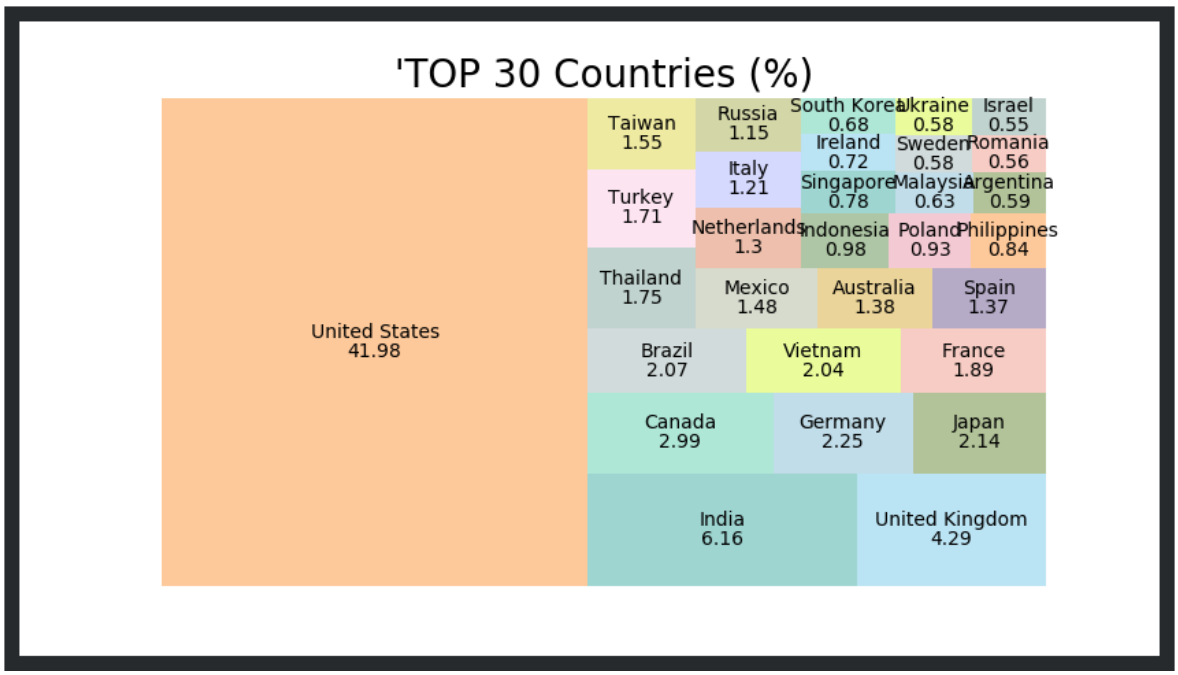
Domain Usage



Σχήμα 6.7: Total Revenue Per Domain

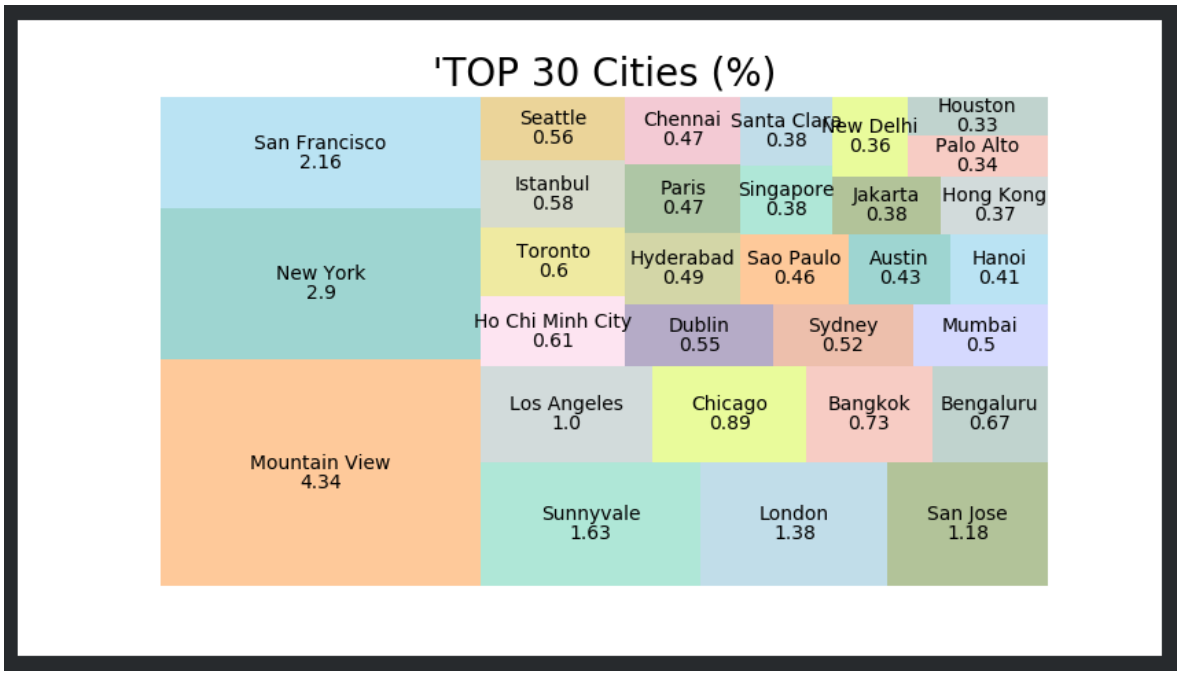
Στο σχήμα 6.7 εξετάζεται από ποίες διευθύνσεις διαδικτύου έγιναν οι περισσότερες επισκέψεις και πόσο τις εκατό (%) προσοδοφόρες ήταν αυτές. Στην πρώτη θέση βρίσκεται η διεύθυνση comcast.net με ποσοστό επίσκεψης 22.1% και ποσοστό κέρδους από αυτές τις επισκέψεις 34.9%. Στην δεύτερη θέση η διεύθυνση rr.com με ποσοστό επίσκεψης 11.5% και ποσοστό κέρδους 13.5% και τέλος στην τρίτη θέση με ποσοστό επίσκεψης 10.6% το verizon.net αλλα με ποσοστό κέρδους μεγαλύτερο από αυτό του rr.com 17.5%.

6.0.3 Γραφήματα με βάση τα γεωγραφικά σημεία



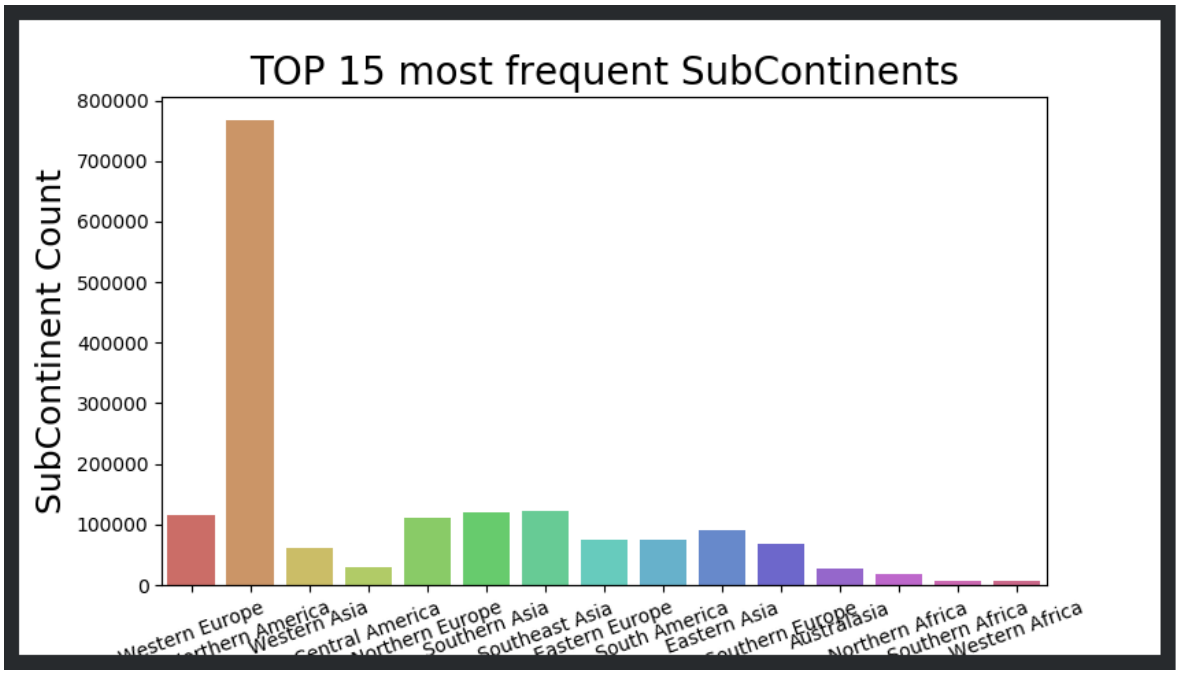
Σχήμα 6.8: Most active Countries

Το σχήμα 6.8 αναπαριστά τις χώρες που είναι πιο ενεργές στο Gstore. Οι Ηνωμένες Πολιτείες της Αμερικής είναι στην πρώτη θέση με 41.98%, στην δεύτερη θέση η Ινδία με 6.16% και τέλος στην τρίτη θέση η Αγγλία με 4.29%. Ακολουθεί ο Καναδάς με 2.99% και στην σύνεχεια με πολύ μικρή διαφορά η Γερμανία, η Ιαπωνία, η Βραζιλία και το Βιετνάμ με διακύμανση από 2.04% - 2.25%.



Σχήμα 6.9: Most active Cities

Το σχήμα 6.9 αναπαριστά τις πόλεις που είναι πιο ενεργές στο Gstore. Το Mountain View είναι στην πρώτη θέση με 4.34%, στην δεύτερη θέση η Νέα Υόρκη με 2.9% και τέλος στην τρίτη θέση το San Francisco με 2.16%. Ακολουθούν με πολύ μικρή διαφορά το Sunnyvale, το Λονδίνο και το San Jose με διακύμανση από 1.18% - 1.63%.



Σχήμα 6.10: Most active SubContinents

Στο σχήμα 6.10 αναλύονται οι πιο ενεργοί ήπειροι. Είναι ξεκάθαρο ότι έχουμε την πλειοψηφία των χρηστών με αριθμό πάνω από 750000, ανήκουν στην Νότια Αμερική. Στην συνέχεια ακολουθούν η Δυτική Ευρώπη, η Βόρεια Ευρώπη, η Νότια Ασία, η Νοτιοανατολική Ασία και η Ανατολική Ασία με αριθμό χρηστών να κυμαίνονται γύρω από τις 100000. Τέλος παρατηρούμε ότι οι Αφρικάνικές ήπειροι έχουν πάρα πολύ μικρό αριθμό χρηστών όπως και η Αυστραλία.

Κεφάλαιο 7

Συμπεράσματα

Το βασικό συμπέρασμα που λαμβάνουμε ύστερα από την περάτωση της παρούσας πτυχιακής εργασίας είναι λοιπόν ότι πίσω από μεγάλο όγκο δεδομένων κρύβεται ακόμη πιο μεγάλος όγκος γνώσης. Από την στιγμή που οποιαδήποτε εταιρεία επιθυμεί να αποδώσει αποτελεσματικά στον τεράστιο ανταγωνισμό πρέπει να ξεκινήσει να χρησιμοποιεί τα δεδομένα που συλλέγει. Χρησιμοποιώντας τα δεδομένα εννοούμε τη σωστή διαχείρισή τους και αναπαράσταση τους έτσι ώστε να μπορούν να γίνουν εύκολα κατανοητά χωρίς περαιτέρω προσπάθεια. Με αυτόν το τρόπο θα μπορέσουν να διαχειριστούν καλύτερα τους πελάτες τους, να εξελίσσονται και να παίρνουν πιο σωστά τις επιχειρηματικές τους αποφάσεις. Με την εφαρμογή αυτής της πτυχιακής επιτυγχάνουμε αυτό το σκοπό, της απλοποίησης του μεγάλου φόρτου δεδομένων και τις αναπαραστάσεις αυτού με γραφήματα που γίνονται κατανοητά από όλους.

Κεφάλαιο 8

Μελλοντικές Επεκτάσεις

Η εφαρμογή της πτυχιακής αυτής θα μπορούσε να επεκταθεί και να γίνει πιο ευέλικτη, εννοώντας ότι θα μπορεί να διαχειριστεί κάθε είδους δεδομένα και όχι κάποιο συγκεκριμένο σετ δεδομένων. Έτσι θα μπορεί οποιοσδήποτε να χρησιμοποιεί την εφαρμογή με δικά του δεδομένα, ορίζοντας ποίες στήλες να συσχετίζονται για την δημιουργία γραφημάτων, το είδος των γραφημάτων και τι αποτελέσματα να έχει στις κύριες οθόνες. Επίσης θα ήταν δυνατόν να δημιουργηθεί και ιστοσελίδα η οποία θα έδειχνε τα γραφήματα και θα δινόταν μια επεξήγηση αυτών. Με αυτό το τρόπο θα μπορούσαν να έχουν την εφαρμογή αυτή σε κάποιον απομακρυσμένο κεντρικό διακομιστή και οι χρήστες να μετέφεραν τα δεδομένα τους σε αυτόν μέσω της ιστοσελίδας και να έβλεπαν ως έξοδο τα γραφήματα και τα αποτελέσματα των δεδομένων τους.

Παράρτημα Α΄

Κώδικας εφαρμογής

Α΄.0.1 GaPredictionMain Module

Αυτή η ενότητα εμπεριέχει συναρτήσεις που έχουν να κάνουν με το γραφικό περιβάλλον, για παράδειγμα τις λειτουργίες των κουμπιών, την εμφάνισή των γραφημάτων σε καρτέλες όπως και τα μηνύματα που εμφανίζονται στο χρήστη.

Α΄.0.2 CleaningDF Module

Αυτή η ενότητα είναι υπεύθυνη για τον καθαρισμό του συνόλου δεδομένων (dataset), την διόρθωση των ημερομηνιών στην σωστή μορφή και την αντικατάσταση μεγάλων συμβολοσειρών (strings). Βασικές συναρτήσεις είναι "`filling_na_values`" που συμπληρώνει τις τιμές που λείπουν από το σύνολο δεδομένων με τις προεπιλεγμένες τιμές και η "`discover_constant_columns`" που ανακαλύπτει τις στατικές στήλες και τις διαγράφει με την μέθοδο "`drop_constant_columns`".

Α΄.0.3 DataFrameLoader Module

Αυτή η ενότητα είναι υπεύθυνη για την φόρτωση των δεδομένων σε δομή δεδομένων της γλώσσας python (pandas dataframe). Μετά την φόρτωση των δεδομένων καλούνται οι συναρτήσεις της ενότητας CleaningDf για να καθαριστούν, να συμπληρωθούν και να διορθωθούν.

Α΄.0.4 Plots Module

```
figure = plt.figure(figsize=(10, 10))
plt.scatter(range(df.shape[0]), np.sort(df["totals.transactionRevenue"].values))
plt.xlabel('index', fontsize=12)
plt.ylabel('TransactionRevenue', fontsize=12)
plt.title('The 80/20 rule', fontsize=20)
plt.savefig('plots/80_20.png', format='png', bbox_inches='tight')
```

Σχήμα Α΄.1: Pareto Rule

```
figure = plt.figure(figsize=(10, 10))
sns.countplot(df[df['device.browser']
             .isin(df['device.browser']
                   .value_counts()[:10].index.values)]['device.browser'],
             palette="hls")  # It's a module to count the category's
plt.title("TOP 10 Most Frequent Browsers", fontsize=20)  # Adding Title and setting the size
plt.xlabel("Browser Names", fontsize=12)  # Adding x label and setting the size
plt.ylabel("Count", fontsize=12)  # Adding y label and setting the size
plt.xticks(rotation=20)  # Adjust the x ticks, rotating the labels
plt.savefig('plots/browser.png', format='png', bbox_inches='tight')
```

Σχήμα Α΄.2: Most used browsers

```
figure = plt.figure(figsize=(10, 10))
city_tree = round((city_tree[:30] / len(df['geoNetwork.city']) * 100), 2)

g = squarify.plot(sizes=city_tree.values, label=city_tree.index,
                  value=city_tree.values,
                  alpha=.4, color=color)
g.set_title("'TOP 30 Cities (%)", fontsize=20)
g.set_axis_off()
plt.savefig('plots/cities.png', format='png', bbox_inches='tight')
```

Σχήμα Α΄.3: Most frequent cities

```
figure = plt.figure(figsize=(10, 10))
# let explore the browser used by users
sns.countplot(df["channelGrouping"], palette="hls")  # It's a module to count the category's
plt.title("Channel Grouping Count", fontsize=20)  # setting the title size
plt.xlabel("Channel Grouping Name", fontsize=12)  # setting the x label size
plt.ylabel("Count", fontsize=12)  # setting the y label size
plt.savefig('plots/channel.png', format='png', bbox_inches='tight')
```

Σχήμα Α΄.4: Most used channel groupings

```
fig, ax = plt.subplots(figsize=(12, 18))
lgb.plot_importance(model, max_num_features=50, height=0.8, ax=ax)
ax.grid(False)
plt.title("LightGBM - Feature Importance", fontsize=15)
fig.savefig('plots/importance.png', format='png', bbox_inches='tight')
```

Σχήμα Α΄.5: Future importance

```
figure = plt.figure(figsize=(10, 7))
sns.countplot(df[df['device.operatingSystem']
                 .isin(df['device.operatingSystem']
                       .value_counts()[:8].index.values)]['device.operatingSystem'],
                palette="hls")  # It's a module to count the category's
plt.title("Operational System used Count", fontsize=20)  # setting the title size
plt.xlabel("Operational System Name", fontsize=6)  # setting the x label size
plt.ylabel("OS Count", fontsize=12)  # setting the y label size
plt.xticks(rotation=15)  # Adjust the x ticks, rotating the labels
plt.savefig('plots/os.png', format='png', bbox_inches='tight')
```

Σχήμα Α΄.6: Most used operating system

```
figure = plt.figure(figsize=(10, 10))
figure.suptitle("Transactions By OS", fontsize=20)
figure.subplots_adjust(top=0.93, wspace=0.3)
ax = figure.add_subplot(1, 1, 1)
ax.set_xlabel("Transactions")
ax.set_ylabel("Frequency")
g = sns.FacetGrid(df[(df['device.operatingSystem']
                      .isin(df['device.operatingSystem']
                            .value_counts()[:6].index.values)) & df['totals.transactionRevenue'] > 0],
                  hue='device.operatingSystem', palette="hls")
g.map(sns.kdeplot, 'totals.transactionRevenue', shade=True, ax=ax)
ax.legend(title='Type')
figure.savefig('plots/transaction_os.png', format='png', bbox_inches='tight')
```

Σχήμα Α΄.7: Transactions by operating system

```
figure = plt.figure(figsize=(10, 10))
country_tree = round((df["geoNetwork.country"].value_counts()[:30]
                      / len(df['geoNetwork.country']) * 100), 2)

g = squarify.plot(sizes=country_tree.values, label=country_tree.index,
                  value=country_tree.values,
                  alpha=.4, color=color)
g.set_title("'TOP 30 Countries (%)", fontsize=20)
g.set_axis_off()
plt.savefig('plots/countries.png', format='png', bbox_inches='tight')
```

Σχήμα Α΄.8: Most frequent countries

```
figure = plt.figure(figsize=(10, 10))
sns.countplot(df[df['geoNetwork.subContinent']
             .isin(df['geoNetwork.subContinent']
                   .value_counts()[:15].index.values)]['geoNetwork.subContinent'],
             palette="hls")  # It's a module to count the category's
plt.title("TOP 15 most frequent SubContinents", fontsize=20)  # setting the title size
plt.xlabel("subContinent Names", fontsize=18)  # setting the x label size
plt.ylabel("SubContinent Count", fontsize=18)  # setting the y label size
plt.xticks(rotation=20)  # Adjust the x ticks, rotating the labels
plt.savefig('plots/sub_continent.png', format='png', bbox_inches='tight')
```

Σχήμα Α΄.9: Most frequent subcontinents

Τα παραπάνω γραφήματα έχουν κατανοητό κώδικα με σχολιασμούς που τονίζουν την λειτουργία κάθε συνάρτησης. Χρησιμοποιήθηκαν οι στήλες του σετ δεδομένων για την αναπαράσταση αυτών των γραφημάτων τα οποία εμφανίζονται σε ένα δισδιάστατο σύστημα αξόνων. Στο τέλος κάθε μεθόδου το γράφημα αποθηκεύεται σε μορφή ".PNG" κάτω από τον φάκελο /plots/.

Α΄.0.5 Prediction Module

```
def predict_revenue_at_session_level(ui, train_df, test_df):
    print("Variables not in test but in train : ", set(train_df.columns).difference(set(test_df.columns)))
    train_df = train_df.drop(["trafficSource.campaignCode"], axis=1)

    # Impute 0 for missing target values
    train_df["totals.transactionRevenue"].fillna(0, inplace=True)
    train_y = train_df["totals.transactionRevenue"].values
    train_id = train_df["fullVisitorId"].values
    test_id = test_df["fullVisitorId"].values

    # label encode the categorical variables and convert the numerical variables to float
    cat_cols = ["channelGrouping", "device.browser", "device.deviceCategory", "device.operatingSystem",
                "geoNetwork.city", "geoNetwork.continent", "geoNetwork.country", "geoNetwork.metro",
                "geoNetwork.networkDomain", "geoNetwork.region", "geoNetwork.subContinent",
                "trafficSource.adwordsClickInfo.adNetworkType", 'trafficSource.adwordsClickInfo.gclId',
                "trafficSource.medium", "trafficSource.source", 'trafficSource.adwordsClickInfo.isVideoAd',
                'trafficSource.isTrueDirect', 'trafficSource.campaign', 'trafficSource.adwordsClickInfo.page',
                'trafficSource.referralPath', 'trafficSource.adwordsClickInfo.slot', 'trafficSource.keyword']

    for col in cat_cols:
        print(col)
        lbl = preprocessing.LabelEncoder()
        lbl.fit(list(train_df[col].values.astype('str')) + list(test_df[col].values.astype('str')))
        train_df[col] = lbl.transform(list(train_df[col].values.astype('str')))
        test_df[col] = lbl.transform(list(test_df[col].values.astype('str')))
        ui.update_progressbar(ui.get_progressbar_status() + 1)

    num_cols = ["totals.hits", "totals.pageviews", "visitNumber", "visitStartTime", 'totals.bounces', 'totals.newVisits']
    for col in num_cols:
        train_df[col] = train_df[col].astype(float)
        test_df[col] = test_df[col].astype(float)

    train_df['date'] = pd.to_datetime(train_df['date'], format='%Y/%m/%d')
    # Split the train dataset into development and valid based on time
    dev_df = train_df[train_df['date'] <= datetime.date(2017, 5, 31)]
    val_df = train_df[train_df['date'] > datetime.date(2017, 5, 31)]
    dev_y = np.log1p(dev_df["totals.transactionRevenue"].values)
    val_y = np.log1p(val_df["totals.transactionRevenue"].values)

    dev_X = dev_df[cat_cols + num_cols]
    val_X = val_df[cat_cols + num_cols]
    test_X = test_df[cat_cols + num_cols]
```

Σχήμα Α΄.10: Predict Revenue At Session Level

```python
# custom function to run light gbm model
def run_lgb(train_X, train_y, val_X, val_y, test_X):
    params = {
        "objective": "regression",
        "metric": "rmse",
        "num_leaves": 30,
        "min_child_samples": 100,
        "learning_rate": 0.1,
        "bagging_fraction": 0.7,
        "feature_fraction": 0.5,
        "bagging_frequency": 5,
        "bagging_seed": 2018,
        "verbosity": -1
    }

    lgtrain = lgb.Dataset(train_X, label=train_y)
    lgval = lgb.Dataset(val_X, label=val_y)
    model = lgb.train(params, lgtrain, 1000, valid_sets=[lgval], early_stopping_rounds=1000, verbose_eval=1000)

    pred_test_y = model.predict(test_X, num_iteration=model.best_iteration)
    pred_val_y = model.predict(val_X, num_iteration=model.best_iteration)
    return pred_test_y, model, pred_val_y

# Training the model #
pred_test, model, pred_val = run_lgb(dev_X, dev_y, val_X, val_y, test_X)

pred_val[pred_val < 0] = 0
val_pred_df = pd.DataFrame({"fullVisitorId": val_df["fullVisitorId"].values})
val_pred_df["transactionRevenue"] = val_df["totals.transactionRevenue"].values
val_pred_df["PredictedRevenue"] = np.expm1(pred_val)
val_pred_df = val_pred_df.groupby("fullVisitorId")["transactionRevenue", "PredictedRevenue"].sum().reset_index()
value = np.sqrt(metrics.mean_squared_error(np.log1p(val_pred_df["transactionRevenue"].values),
                                           np.log1p(val_pred_df["PredictedRevenue"].values)))
print(value)

sub_df = pd.DataFrame({"fullVisitorId": test_id})
pred_test[pred_test < 0] = 0
sub_df["PredictedLogRevenue"] = np.expm1(pred_test)
sub_df = sub_df.groupby("fullVisitorId")["PredictedLogRevenue"].sum().reset_index()
sub_df.columns = ["fullVisitorId", "PredictedLogRevenue"]
sub_df["PredictedLogRevenue"] = np.log1p(sub_df["PredictedLogRevenue"])
sub_df.to_csv("DataSets/prediction.csv", index=False)

showPlots.display_feature_importance(ui, model)

return round(value, 5)
```

Σχήμα Α΄.11: Predict Revenue At Session Level

Παράρτημα Β΄

Ακρωνύμια και συντομογραφίες

Gstore Google Merchandise Store

Csv Comma Seperated Values

ANN Artificial Neural Networks

CART Classification And Regression Tree

Conv Covariance

Corr Correlation Coefficient

SPSS Statistical Package for the Social Sciences

POSIX Portable Operating System Interface

ETL Extract, Load, Transform

PCA Principal Components Analysis

PEP20 Zen of Python

ML Machine Learning

CMD Command Prompt

PyPi Python Package Index

Pip Pip Installs Packages

Pandas Python Data Analysis Library

SkLearn Scikit-Learn

NAN Not A Number

R-Squared R^2

C Programming Language C

Cython Python With C-inspired Syntax And Performance

API Application Program Interface

GPU Graphics Processing Unit

XML Extensible Markup Language

SVG Scalable Vector Graphics

SQL Structured Query Language

OS Operating System

DD Day

MM Month

YY Year

Int Integer

RMSE Root Mean Squared Error

iOS iPhone Operating System

URL Uniform Resource Locator

K-L Karhunen-Loeve

Haskell Programming Language

Df DataFrame

Bibliography

[1] Jay Sridhar, Makeuseof, What is data analysis, February 12, 2018

[2] Dhar, V. «"Data science and prediction"». ACM, December 2013, Vol. 56 No.12, Pages 64-73

[3] «The key word in "Data Science" is not Data, it is Science. Simply Statistics». November 8, 2016.

[4] Definition of "Data Science", Wikipedia, June 16, 2017

[5] Dykes, Brent. «Data Storytelling: The Essential Data Science Skill Everyone Needs». Forbes. November 8, 2016.

[6] Breur, Tom (July 2016). "Statistical Power Analysis and the contemporary "crisis" in social sciences". Journal of Marketing Analytics. 4 (2–3): 61–65. doi:10.1057/s41270-016-0001-3. ISSN 2050-3318

[7] Laney, Doug (2001). "3D data management: Controlling data volume, velocity and variety". META Group Research Note. 6 (70).

[8] Li, Rita; Li, Herru (29 January 2018). "Have Housing Prices Gone with the Smelly Wind? Big Data Analysis on Landfill in Hong Kong". Sustainability. MDPI AG. 10 (2): 341. doi:10.3390/su10020341. ISSN 2071-1050.

[9] Marr, Bernard (6 March 2014). "Big Data: The 5 Vs Everyone Must Know".

[10] Crawford, Kate (21 September 2011). "Six Provocations for Big Data". Social Science Research Network: A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society. doi:10.2139/ssrn.1926431.

[11] "Data, data everywhere". The Economist. 25 February 2010. Retrieved 9 December 2012.

[12] Eleni Gkolemi. Cryptography and Data Mining. June 2016

[13] Bethge, Matthias; Ecker, Alexander S.; Gatys, Leon A. (26 August 2015). "A Neural Algorithm of Artistic Style".

[14] Čech, Eduard (1969). Point Sets. New York: Academic Press. p. 42.

[15] Conway, John (1990), A course in functional analysis, Springer Verlag, ISBN 0-387-97245-5

[16] Forrest, Peter (Fall 2008). "The Identity of Indiscernibles". In Edward N. Zalta (ed.). The Stanford Encyclopedia of Philosophy. Retrieved 2012-04-12.

[17] F. N. David, M. G. Kendall & D. E. Barton (1966) Symmetric Function and Allied Tables, Cambridge University Press.

[18] Joseph P. S. Kung, Gian-Carlo Rota, & Catherine H. Yan (2009) Combinatorics: The Rota Way, §5.1 Symmetric functions, pp 222–5, Cambridge University Press, ISBN 978-0-521-73794-4 .

[19] Fekete, M. (1923). "Über die Verteilung der Wurzeln bei gewissen algebraischen Gleichungen mit ganzzahligen Koeffizienten". Mathematische Zeitschrift. 17 (1): 228–249. doi:10.1007/BF01504345.

[20] de Bruijn, N.G.; Erdös, P. (1952). "Some linear and some quadratic recursion formulas. II". Nederl. Akad. Wetensch. Proc. Ser. A. 55: 152–163. doi:10.1016/S1385-7258(52)50021-0. (The same as Indagationes Math. 14.) See also Steele 1997, Theorem 1.9.2.

[21] Margaret Rouse, Ivy Wigmore, WhatIs, https://whatis.techtarget.com/definition/correlation

[22] Rodgers, J. L.; Nicewander, W. A. (1988). "Thirteen ways to look at the correlation coefficient". The American Statistician. 42 (1): 59–66. doi:10.1080/00031305.1988.10475524. JSTOR 2685263.

[23] Alpaydin, Ethem (2010). Introduction to Machine Learning. MIT Press. p. 9. ISBN 978-0-262-01243-0.

[24] Jan M. Żytkow, Jan Rauch (1999). Principles of data mining and knowledge discovery. ISBN 978-3-540-66490-1.

[25] Ron Kohavi; Foster Provost (1998). "Glossary of terms". Machine Learning. 30: 271–274.

[26] Bishop, Christopher M. (2006). Pattern Recognition and Machine Learning. New York: Springer. p. vii. ISBN 0-387-31073-8. "Pattern recognition has its origins in engineering, whereas machine learning grew out of computer science. However, these activities can be viewed as two facets of the same field, and together they have undergone substantial development over the past ten years."

[27] James, Gareth (2013). An Introduction to Statistical Learning: with Applications in R. Springer. p. 176. ISBN 978-1461471370.

[28] Ripley, Brian (1996). Pattern Recognition and Neural Networks. Cambridge University Press. p. 354. ISBN 978-0521717700.

[29] Prechelt, Lutz; Geneviève B. Orr (2012-01-01). "Early Stopping — But When?". In Grégoire Montavon; Klaus-Robert Müller (eds.). Neural Networks: Tricks of the Trade. Lecture Notes in Computer Science. Springer Berlin Heidelberg. pp. 53–67. doi:10.1007/978-3-642-35289-8_5. ISBN 978-3-642-35289-8.

[30] Wu, S. (2013), "A review on coarse warranty data and analysis", Reliability Engineering and System, 114: 1–11, doi:10.1016/j.ress.2012.12.021

[31] Maurizio Lenzerini (2002). "Data Integration: A Theoretical Perspective" (PDF). PODS 2002. pp. 233–246.

[32] Talend https://www.talend.com/resources/what-is-data-integration/

[33] Frederick Lane (2006). "IDC: World Created 161 Billion Gigs of Data in 2006".

[34] Shana Pearlman, Talend, https://www.talend.com/resources/data-transformation-defined/

[35] Andrei Tuță, https://towardsdatascience.com/data-mining-101-dimensionality-and-data-reduction-2a8fa427b092

3 ΔΕΚΕΜΒΡΙΟΥ 2019