

Τμήμα Οργάνωσης και Διοίκησης Επιχειρήσεων  
Σχολή Επιστημών Διοίκησης Επιχειρήσεων

Προπτυχιακή Εργασία

στο μάθημα

«Ειδικά Θέματα Πληροφοριακών Συστημάτων ΙI»

Κοβούσογλου Γεώργιος, Α.Μ.: bba17003

Κουμεντάκης Κωνσταντίνος, Α.Μ.: bba17010

Θεσσαλονίκη 2020

Περίληψη

Εισαγωγή

Η Instacart είναι μια αμερικανική εταιρεία που λειτουργεί ως υπηρεσία αυθημερόν παράδοσης ειδών παντοπωλείου. Οι πελάτες επιλέγουν τα προϊόντα που επιθυμούν να αγοράσουν μέσω μιας διαδικτυακής εφαρμογής από διάφορους εμπόρους λιανικής και παραδίδονται από έναν προσωπικό αγοραστή. Η υπηρεσία της Instacart παρέχεται κυρίως μέσω μιας εφαρμογής smartphone, διαθέσιμη σε πλατφόρμες iOS και Android, εκτός από τον ιστότοπο της.

Το 2017 η Instacart διοργάνωσε έναν διαγωνισμό Kaggle και παρείχε στην κοινότητα ένα δείγμα συνόλου δεδομένων με πάνω από 3 εκατομμύρια παραγγελίες παντοπωλείων από περισσότερους από 200.000 χρήστες Instacart. Οι παραγγελίες περιλάμβαναν 32 εκατομμύρια είδη καλαθιού και 50.000 μοναδικά προϊόντα. Ο στόχος του διαγωνισμού ήταν οι συμμετέχοντες να προβλέψουν ποια από τα προϊόντα που αγοράστηκαν προηγουμένως θα παραγγελθούν στην επόμενη παραγγελία ενός χρήστη.

Το Instacart παρουσιάζει τα δεδομένα για κάθε παραγγελία ως προηγούμενα (prior), εκπαίδευσης (train) ή δοκιμαστικά (test). Οι προηγούμενες παραγγελίες περιγράφουν την προηγούμενη συμπεριφορά ενός χρήστη, ενώ οι εντολές εκπαίδευσης και δοκιμών αφορούν τη μελλοντική συμπεριφορά που πρέπει να προβλέψουμε.

Για τις παραγγελίες εκπαίδευσης, η Instacart αποκαλύπτει τα αποτελέσματα (δηλαδή, τα παραγγελθέντα προϊόντα), ενώ για τις παραγγελίες δοκιμής δεν έχουμε αυτήν την πληροφορία. Επιπλέον, η μελλοντική παραγγελία κάθε χρήστη μπορεί να είναι είτε εκπαιδευτική είτε δοκιμαστική.

Αυτό είναι ένα πρόβλημα ταξινόμησης επειδή πρέπει να προβλέψουμε εάν κάθε ζεύγος χρήστη και προϊόντος θα ξαναπαραγγελθεί ή όχι. Αυτό υποδηλώνεται από την δυαδική τιμή της επαναπαραγγελούμενης μεταβλητής.

Ως αποτέλεσμα, πρέπει να βρούμε και να υπολογίσουμε διάφορες μεταβλητές πρόβλεψης (X) που θα περιγράφουν τα χαρακτηριστικά ενός προϊόντος και τη συμπεριφορά ενός χρήστη σχετικά με ένα ή περισσότερα προϊόντα. Θα το κάνουμε αναλύοντας τις προηγούμενες παραγγελίες. Στη συνέχεια, θα χρησιμοποιήσουμε τους χρήστες εκπαίδευσης για να δημιουργήσουμε ένα μοντέλο πρόβλεψης και τους δοκιμαστικούς χρήστες για να κάνουμε την πραγματική μας πρόβλεψη. Ως αποτέλεσμα, δημιουργούμε έναν πίνακα ως εξής και εκπαιδεύουμε έναν αλγόριθμο με βάση τις μεταβλητές πρόβλεψης (X) και τη μεταβλητή απόκρισης (Y).

Μεθοδολογία

Αρχικά εισάγαμε πακέτα python και τα δεδομένα τα οποία τα αναδιαμορφώσαμε. Στην συνέχεια δημιουργήσαμε τις παρακάτω περιγραφόμενες μεταβλητές πρόβλεψης:

* user\_days\_since\_last\_order: Πόσες ημέρες πέρασαν από την τελευταία παραγγελία του κάθε χρήστη.
* product\_average\_position: Μέση θέση του προϊόντος στην παραγγελία.
* user\_product\_last\_5\_orders: Πόσες φορές στις τελευταίες παραγγελίες πάρθηκε ένα προϊόν από έναν χρήστη.
* user\_product\_last\_5\_orders\_ratio: Το ποσοστό της προηγούμενης μεταβλητής.
* user\_product\_last\_5\_orders\_max\_days: Στις τελευταίες 5 παραγγελίες το μεγαλύτερο διάστημα που πέρασε μεταξύ των παραγγελιών του κάθε χρήστη.
* user\_product\_last\_5\_orders\_median\_days: Στις τελευταίες 5 παραγγελίες το μέσο διάστημα που πέρασε μεταξύ των παραγγελιών του κάθε χρήστη.
* user\_product\_average\_position: Μέση θέση του προϊόντος στις παραγγελίες του κάθε χρήστη.
* user\_days\_since\_last\_order: Για κάθε χρήστη οι μέρες που πέρασαν από την τελευταία του παραγγελία.
* user\_product\_last\_5\_orders\_median\_day\_of\_week: η οποία δείχνει την μέση ημέρα της εβδομάδας των πέντε τελευταίων παραγγελιών ενός χρήστη.
* user\_product\_last\_5\_orders\_median\_hours: η οποία δείχνει την μεσαία ώρα της παραγγελίας των πέντε τελευταίων παραγγελιών ενός χρήστη.
* user\_days\_since\_last\_order: Οι μέρες που πέρασαν από την τελευταία παραγγελία ενός χρήστη. Για τη δημιουργία της χρησιμοποιήθηκε ο πίνακας orders και συγκεκριμένα μόνο οι εγγραφές με eval\_set “train” και “test”.
* user\_product\_days\_since\_prior\_order: Οι μέρες που πέρασαν από την τελευταία παραγγελία ενός προϊόντος από έναν χρήστη. Για τη δημιουργία της χρησιμοποιήθηκε αθροιστική συνάρτηση στην days\_since\_prior\_order ανά χρήστη και η user\_days\_since\_last\_order.
* user\_product\_orders\_since\_prior\_order: Οι παραγγελίες που πέρασαν από την τελευταία παραγγελία ενός προϊόντος από έναν χρήστη. Για τη δημιουργία της χρησιμοποιήθηκε η order\_number\_back.
* aisle\_order : το πλήθος των παραγγελιών που έγιαναν σε κάθε διάδρομο.
* department\_order: το πλήθος των παραγγελιών που έγιναν σε κάθε κατάστημα.
* one\_shot\_ratio: με ενδιάμεσες μεταβλητές unique\_customers και customers\_one\_shot .
* mean\_one\_shot\_ratio\_of\_aisle: το μέσο one\_shot\_ratio του κάθε διαδρόμου.
* mean\_one\_shot\_ratio\_of\_department: αντίστοιχα για τα καταστήματα.
* user\_product\_orders\_since\_last\_order\_div\_mean\_orders\_between\_purchases: Ο αριθμός των παραγγελιών που πέρασαν από την τελευταία φορά που ένας χρήστης παράγγειλε ένα προϊόν προς τον μέσο αριθμό παραγγελιών που περνούν μεταξύ συνεχόμενων παραγγελιών του προϊόντος από τον χρήστη. Για την δημιουργία της χρησιμοποιήθηκαν η user\_product\_orders\_since\_last\_order και η user\_product\_frequency\_after\_1st, που ουσιαστικά είναι ο μέσος που αναφέρθηκε παραπάνω αντεστραμμένος.
* user\_product\_mean\_days\_between\_orders: Ο μέσος αριθμός ημερών που περνούν μεταξύ συνεχόμενων παραγγελιών ενός προϊόντος από έναν χρήστη.
* user\_product\_days\_since\_last\_order\_div\_mean\_days\_between\_purchases: Ο αριθμός των ημερών που πέρασαν από την τελευταία φορά που ένας χρήστης παράγγειλε ένα προϊόν προς τον μέσο αριθμό ημερών που περνούν μεταξύ συνεχόμενων παραγγελιών του προϊόντος από τον χρήστη. Για την δημιουργία της χρησιμοποιήθηκαν η user\_product\_days\_since\_last\_order και η user\_product\_mean\_days\_between\_orders
* παραγγελιών ενός χρήστη που περιείχαν ένα προϊόν.

Έπειτα δημιουργήσαμε τα train, τα test DataFrames και προχωρήσαμε στην δημιουργία του προγνωστικού μας μοντέλου. Το μοντέλο Machine Learning που δημιουργήσαμε βασίζεται στον αλγόριθμο XGBoost. Στο Virtual Machine της Google με την χρήση GPU υπολογίσαμε τις καλύτερες τιμές των παραμέτρων με την βοήθεια του GridSearch έπειτα από πολλές δοκιμές. ΒΑΛΕ ΤΟΝ ΠΙΝΑΚΑ ΜΕ ΤΙΣ BEST PARAMETER

Συμπεράσματα

Καταληκτικά, ο νικητής του διαγωνισμού είχε F1 score 0.4091 και εμείς καταφέραμε να δημιουργήσουμε ένα μοντέλο με F1 score ……, μνήμης ….. που απαιτεί …. λεπτά για να εκτελεστεί.

ΒΑΛΕ ΤΗΝ ΕΙΚΟΝΑ ΜΕ ΤΟ ΣΚΟΡ

Η σημαντικότητα των μεταβλητών όπως φαίνεται στην παρακάτω εικόνα μας δείχνει ότι οι μεταβλητές …………….. είχαν την μεγαλύτερη συνεισφορά στην δημιουργία του μοντέλου.

ΒΑΛΕ ΤΗΝ ΕΙΚΟΝΑ ΤΟΥ FEATURE IMPORTANCE