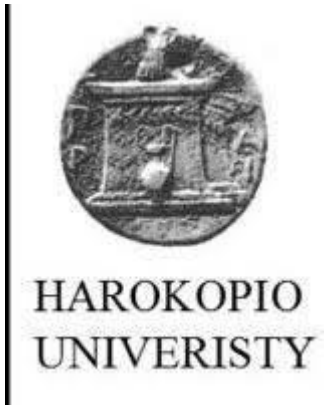
Τεχνητή Νοημοσύνη

Εργασία 2η

Καθηγητής: Χρήστος Δίου

Παναγιώτης Φωτεινόπουλος – 2021154

Εαρινό Εξάμηνο 2023 – 2024



Ερωτήσεις αναφοράς

1. Σε αυτό το ερώτημα υλοποιώ ένα eda.py το οποίο θα κάνει την διερεύνηση των δεδομένων με φόρτωση των δεδομένων από το αρχείο project2_dataset.scv και το ουσιαστικά είναι ένα explanatory file. Τα αποτελέσματα μου είναι ότι συνολικά οι εγγραφές είναι 12,330, το ποσοστό όπου οι χρήστες αγόρασαν είναι 15.47% και η ευστοχία του μοντέλου για την πρόβλεψη ότι ο χρήστης δεν θα αγοράσει είναι 84.53%.

2. Το δεύτερο ερώτημα θα δημιουργήσει την συνάρτηση prepare_data για να προετοιμάσει το σύνολο των δεδομένων μας. Ουσιαστικά, αυτό που γίνεται είναι ότι προετοιμάζουμε τα δεδομένα πριν την υλοποίηση του μοντέλου με το να αφαιρέσουμε κάποια χαρακτηριστικά για απλούστευση, να κάνουμε τις Boolean τιμές αριθμητικές, να εφαρμόσουμε One-hot encoding στις Region, TrafficType, VisitorType όπου χρησιμοποιώ την get_dummies όπως προτείνατε όπου χρησιμοποιούνται dummy variables που είναι δυαδικές μεταβλητές που χρησιμοποιούνται για την αναπαράσταση κατηγορικών δεδομένων. Ακόμη, χωρίστηκε η μεταβλητή στόχος Revenue από τις υπόλοιπες και χρησιμοποιώντας την train_test_split του sklearn χωρίζουμε τα datasets σε σύνολο εκπαίδευσης και δοκιμής.

3. Το τρίτο ερώτημα επιπλέον θα ορίσει 70% και 30% σύνολο εκπαίδευση και δοκιμής αλλά και θέτω το random_state σε 42.Επιπλέον, χρησιμοποιώ MinMaxScaler της sklearn για να κανονικοποιήσω γραμμικά τα δεδομένα.

4. Στο τέταρτο ερώτημα θα υλοποιήσω το μοντέλο όπου θα χρησιμοποιηθεί η LogisticRegression όπου έτσι θα δούμε με γραμμική παλινδρόμηση αν μία μεταβλητή στόχος είναι κατηγορική. Χρησιμοποιώντας τον λογάριθμο όπως αναφέρθηκε στο μάθημα των αποδόσεων ως εξαρτημένη μεταβλητή. Γενικότερα, η LogisticRegression προβλέπει την πιθανότητα εμφάνισης ενός δυαδικού γεγονότος χρησιμοποιώντας την συνάρτηση logit. Επιπλέον, όπως εξηγήσαμε στο μάθημα το μοντέλο της λογιστικής παλινδρόμησης / γραμμική ταξινόμηση είναι ότι με την λογιστική συνάρτηση/ σιγμοειδής συνάρτηση σ(x-x0) = 1 / (1 + e^-z) κοιτάζω πόσο απέχει το x από το x0 και όπως είπαμε είναι μία δυαδική ταξινόμηση (Linear Classification) όπου ορίζω ένα 0 ή 1 και ο στόχος είναι να ψάξω ένα αποδοτικό threshold.

5. Ο πίνακας σύγχυσης σε αυτό το ερώτημα λειτουργεί σαν εκτιμητής για την απόδοση του μοντέλου ταξινόμησης. Όπως αναφέραμε και στο μάθημα το στοιχείο (i,j) υποδηλώνει πόσα δείγματα τα οποία ανήκουν στην κλάση i και ο ταξινομητής εκτιμά ότι ανήκουν στην κλάση j. Συνεπώς, στην διαγώνιο θα δούμε τα σωστά ταξινομημένα δείγματα. Ειδικότερα, παρουσιάζεται ο αριθμός των πραγματικών κλάσεων που ταξινομήθηκαν για πρόβλεψη, διαιρώντας τα σε 2 κατηγορίες όπου είναι οι True Positives (TP) και True Negatives (TN), και επίσης τις False Positives (FP) και False Negatives (FN). Έτσι, οι TP δείχνουν τα δείγματα που προβλέφθηκαν σωστά ως θετικά, ενώ οι TN προβλέφθηκαν σωστά σαν αρνητικά αντιστοίχως. Οι FP είναι λανθασμένες προβλέψεις των θετικών δηλαδή τα λάθος θετικά και οι FN αντίστοιχα είναι λανθασμένες προβλέψεις των αρνητικών. Στον δικό μου πίνακα φαίνεται ότι για το σύνολο δοκιμής υπάρχουν 3060 πραγματικές αρνητικές προβλέψεις (TN), 64 ψευδείς θετικές προβλέψεις (FP), 371 ψευδείς αρνητικές προβλέψεις (FN) και 204 πραγματικές θετικές προβλέψεις (TP). Ενώ για το σύνολο εκπαίδευσης βλέπουμε 7137 πραγματικές αρνητικές προβλέψεις (TN), 161 ψευδείς θετικές προβλέψεις (FP), 822 ψευδείς αρνητικές προβλέψεις (FN), 511 πραγματικές θετικές προβλέψεις (TP).

   Για την βελτίωση του μοντέλου θα μπορούσαμε να κάνουμε ακόμη καλύτερη ανάλυση των χαρακτηριστικών εξετάζοντας ποια είναι τα σημαντικότερα και κάποια ίσως να τα αφαιρούσαμε ή να προσθέταμε άλλα χαρακτηριστικά αν ήταν δυνατόν. Ακόμη, θα μπορούσαμε να κάνουμε χρήση άλλων μοντέλων μηχανικής μάθησης, όπως μηχανές διανυσμάτων (SVM) για να δούμε αν

μπορούν να προσφέρουν καλύτερη ακρίβεια στις προβλέψεις μας https://www.geeksforgeeks.org/support-vector-machine-algorithm/).