Θεωρία Αποφάσεων
Εργαστηριακή Άσκηση

Αθανάσιος Ανδρικόπουλος
Χειμερινό Εξάμηνο 2020

Κάθε ομάδα θα πρέπει να επιλέξει μία από τις παρακάτω ασκήσεις προς υλοποίηση. Κάθε μια από τις ασκήσεις θα μπορεί να ανατεθεί το πολύ σε τέσσερις ομάδες. Τα παραδοτέα κάθε ομάδας θα είναι ο κώδικας υλοποίησης της άσκησης σε γλώσσα προγραμματισμού Python (later than v.3.0), στον οποίο θα απαιτείται εκτενής σχολιασμός σε κάθε μια εντολή που χρησιμοποιείται. Επιπλέον είναι απαραίτητη η συγγραφή τεχνικής αναφοράς στην οποία θα περιγράφεται το πρόβλημα, οι μέθοδοι που χρησιμοποιήθηκαν, τα αποτελέσματα που προέκυψαν και οποιαδήποτε άλλη πληροφορία που η ομάδα θεωρεί απαραίτητη. Ο παραδοτέος κώδικας θα πρέπει να υπάρχει στο version control tool GitHub. Η ημερομηνία παράδοσης ορίζεται η **20η Φεβρουαρίου 2021** και ώρα **23.59**, χωρίς δυνατότητα παράτασης.

Εργαστηριακή Άσκηση 1

Έχουμε διαθέσιμα τα δεδομένα μιας αλυσίδας καταστημάτων και είναι σημαντικό να προβλέψουμε τις πωλήσεις στα διάφορα τμήματα των υποκαταστημάτων. Δεδομένου ότι μπορεί να υπάρχουν πολλοί παράγοντες που μπορούν να επηρεάσουν τις πωλήσεις για κάθε τμήμα, καθίσταται επιτακτική ανάγκη να προσδιορίσουμε τους βασικούς παράγοντες που συμβάλλουν στην προώθηση των πωλήσεων και τις χρησιμοποιούν για την ανάπτυξη ενός μοντέλου που μπορεί να βοηθήσει στην πρόβλεψη των πωλήσεων με κάποια ακρίβεια.

Έχουμε εβδομαδιαία δεδομένα πωλήσεων για 45 καταστήματα και 98 τμήματα για μια περίοδο 3 ετών. Επιπλέον, έχουμε συγκεκριμένες πληροφορίες για το κατάστημα και τη γεωγραφία, όπως το μέγεθος του καταστήματος, το ποσοστό ανεργίας, τη θερμοκρασία, κλπ. Χρησιμοποιώντας αυτούς τους παράγοντες, χρειαζόμασταν να αναπτύξουμε ένα μοντέλο παλινδρόμησης που μπορεί να προβλέψει τις πωλήσεις και είναι επίσης υπολογιστικά αποτελεσματικό και επεκτάσιμο. Θα πρέπει να ληφθεί υπόψιν η εποχικότητα αλλά άλλες σημαντικές μεταβλητές όπως οι διακοπές και ο τύπος τμήματος για να έχουμε καλύτερη ακρίβεια.

Υπόμνημα Dataset 1

Store: Μοναδικός κωδικός καταστήματος
Dept: Κωδικός Τμήματος
Date: Ημερομηνία
Weekly Sales: Αριθμός εβδομαδιαίων πωλήσεων
IsHoliday: True/False, ανάλογα αν πρόκειται για ημέρα αργίας ή όχι
Temperature: Θερμοκρασία
Unemployment: Ποσοστό ανεργίας

Εργαστηριακή Άσκηση 2

Έχουμε συλλέξει τα δεδομένα μιας αλυσίδας καταστημάτων για το έτος 2013, για 1559 προϊόντα και για 10 διαφορετικά υποκαταστήματα που διαθέτει. Επίσης έχουμε διαθέσιμα και συγκεκριμένα χαρακτηριστικά για καθένα από τα προϊόντα και τα καταστήματα. Ο σκοπός είναι να φτιάξουμε ένα μοντέλο πρόβλεψης των πωλήσεων κάθε προϊόντος ανά κατάστημα. Το μοντέλο μας θα μπορούσε να χρησιμοποιηθεί από την ομάδα της εν λόγω εταιρίας για να προβλέψει τις πωλήσεις που θα έχει και να οργανώσει το διαθέσιμο απόθεμα που θα πρέπει να έχει. Θα ήταν χρήσιμο να λάβουμε υπόψιν τα χαρακτηριστικά του προϊόντος, της θέση τους μέσα στο κατάστημα, τον πληθυσμό που επισκέπτεται το κατάστημα κλπ.

Υπόμνημα Dataset 2

Item_Identifier: Μοναδικός Κωδικός προϊόντος
Item_weight: Βάρος προϊόντος
Item_Fat_Content: Ποσοστό λίπους του προϊόντος
Item_Visibility: Μεγαλύτερη ορατότητα σημαίνει ότι οι πελάτες είναι πιο πιθανό να δουν πραγματικά το προϊόν
Item_Type: Κατηγορία προϊόντος
Outlet_Identifier: Μοναδικός κωδικός καταστήματος
Outlet_Establisment_Year: Χρονιά ίδρυσης του καταστήματος
Outlet_Size: Μέγεθος καταστήματος, άμεσα συνδεδεμένο με τον αριθμό των πωλήσεων

Outlet_Location_Type: Tier 1~3, Τα καταστήματα Tier1 απευθύνονται σε μεγαλύτερο πληθυσμό και γενικά το εισόδημα των πελατών είναι υψηλότερο
Outlet_Type: Supermarket Type1/2/3 ή Grocery Store.
Item_Outlet_Sales: Η εξαρτημένη μεταβλητή μας το αποτέλεσμα που θέλουμε να εξηγήσουμε

Εργαστηριακή Άσκηση 3

Στόχος της άσκησης είναι να δημιουργηθεί ένα music recommendation system. Στο dataset που έχουμε στη διάθεση μας έχουμε ανά χρήστη, το τραγούδι που άκουσε. Καλείστε να φτιάξετε ένα σύστημα το οποίο θα προτείνει τραγούδια κατάλληλα για τον κάθε χρήστη , σύμφωνα με τα προηγούμενα τραγούδια που έχει ακούσει και περιέχονται στο dataset. Επιπλέον θα ήταν χρήσιμη η ανάλυση δεδομένων από την οποία θα προκύψει κατηγοριοποίηση και οπτικοποίηση αυτών. Για παράδειγμα, θα μπορούσε να γίνει οπτικοποίηση των δημοφιλέστερων καλλιτεχνών κ.λπ.

Υπόμνημα Dataset 3

user_id: το id του χρήστη
artistname: Το όνομα του καλλιτέχνη
trackname: Το όνομα του τραγουδιού
playlistname: Το όνομα της playlist (κατηγορία)

Εργαστηριακή Άσκηση 4

Ένας διευθυντής σε μια τράπεζα παρατηρεί ότι όλο και περισσότεροι πελάτες εγκαταλείπουν τις υπηρεσίες της πιστωτικής τους κάρτας. Θα ήταν χρήσιμο αν κάποιος μπορούσε να προβλέψει ποιοι πελάτες πρόκειται να εγκαταλείψουν , ώστε να μπορούν να πηγαίνουν προληπτικά στον πελάτη για να τους παρέχουν καλύτερες υπηρεσίες και να αναστρέψουν την απόφαση των πελατών προς την αντίθετη κατεύθυνση

Σύμφωνα με το dataset που σας δίνεται, να εκπαιδεύσετε το σύστημα σας ώστε να προβλέπει ποιοι πελάτες είναι πιθανό να εγκαταλείψουν σύντομα. Το σύνολο δεδομένων αποτελείται από 10.000 πελάτες που αναφέρουν την ηλικία τους, τον μισθό τους, την οικογενειακή τους κατάσταση, το όριο πιστωτικών καρτών, την κατηγορία πιστωτικών καρτών κ.λπ.

Υπόμνημα Dataset 4

CLIENTNUM: το id του χρήστη
Attrition_Flag: (Existing Customer/Attrited Customer) Αν ο πελάτης έχει εγκαταλήψει την υπηρεσία
Customer Age: Η ηλικία του χρήστη

Gender: Φύλο
Education_Level: Μορφωτικό επίπεδο
Marital_Status: Οικογενειακή κατάσταση χρήστη
Dependent_Count: Αριθμός εξαρτώμενων παιδιών
Income_Category: Εισόδημα
Card_category: Κατηγορία κάρτας
Months_on_book: Μήνες που ο πελάτης είναι εγγραμμένος
Months_Inactive: Μήνες που ο πελάτης είναι ανενεργός

Εργαστηριακή Άσκηση 5

Σκοπός της εργαστηριακής άσκησης είναι να δημιουργηθεί ένα σύστημα που να αξιολογεί την ποιότητα του κρασιού, ανάλογα με κάποια συστατικά και χαρακτηριστικά του. Έχετε στη διαθεσή σας ένα training dataset (Wine_Training) που περιέχει ορισμένα συστατικά και χαρακτηριστικά του κάθε κρασιού καθώς και την τελική βαθμολογία της ποιότητας του κρασιου (quality). Καλείστε να φτιάξετε και να εκπαιδεύσετε ένα τέτοιο σύστημα, το οποίο τελικά θα δέχεται ως είσοδο τις τιμές για αυτά τα χαρακτηριστικά του κρασιού και θα επιστρέφει βαθμολογημένη την ποιότητά του. Στη διάθεση σας έχετε και το testing dataset (Wine_Testing).

Υπόμνημα Dataset 5

fixed acid, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol: συστατικά/χαρακτηριστικά του κρασιού
quality: βαθμολογία της ποιότητας του κρασιού

Εργαστηριακή Άσκηση 6

Σε αυτή την άσκηση έχετε στη διάθεση σας ένα dataset που περιλαμβάνει τις αξιολογήσεις 6040 χρηστών για 3900 ταινίες. Ο σκοπός αυτής της άσκησης είναι η δημιουργία ενός συστήματος που να προτείνει ταινίες στους χρήστες σύμφωνα με τις ταινίες που έχουν παρακολουθήσει και τις αξιολογήσεις που έχουν κάνει. Επιπλέον θα ήταν χρήσιμη η ανάλυση δεδομένων από την οποία θα προκύψει κατηγοριοποίηση και οπτικοποίηση αυτών. Για παράδειγμα, θα μπορούσε να γίνει οπτικοποίηση των ταινιών που παρακολουθούν οι χρήστες ανάλογα με την ηλικία, το φύλο τους κλπ.

Υπόμνημα Dataset 6

movie_id: Το id της ταινίας
title: Ο τίτλος της ταινίας
genres: Το είδος της ταινίας

user_id: το id του χρήστη
rating: Η βαθμολογία που ο χρήστης έχει βάλει στην ταινία (1-5)
timestamp: Ημερομηνία και ώρα
gender: Το φύλο του χρήστη
age: Το id της ηλικία του χρήστη
occupation: Το id της δουλειάς του χρήστη
age_desc: Το εύρος της ηλικίας του χρήστη
occupation_desc: Η περιγραφή της δουλειάς τους χρήστη

Εργαστηριακή Άσκηση 7

Σκοπός της εργαστηριακής άσκησης είναι να δημιουργηθεί ένα σύστημα που θα επεξεργάζεται social media posts και θα αξιολογεί το περιεχόμενο τους, ως θετικό ή αρνητικό. Έχετε στη διαθεσή σας δύο datasets, (SocialMedia_Positive, SocialMedia_Negative) που περιέχουν posts που χαρακτηρίζονται ως θετικα και αρνητικά αντίστοιχα. Καλείστε να φτιάξετε και να εκπαιδεύσετε ένα τέτοιο σύστημα, το οποίο τελικά θα δέχεται ως είσοδο ένα post και θα αναγνωρίζει τη γενικότερη κατηγορία στην οποία ανήκει. Μπορείτε να χρησιμοποιήσετε ένα μέρος των δεδομένων datasets για training του συστήματος, και ένα άλλο μέρος τους για testing.

Υπόμνημα Dataset 7

ID: μοναδικό ID κάθε post
Text: post
Sentiment: Χαρακτηρισμός του περιεχομένου του post (positive/negative)

Εργαστηριακή Άσκηση 8

Σκοπός της εργαστηριακής άσκησης είναι να δημιουργηθεί ένα job recommendation system. Έχετε στη διαθεσή σας ένα training dataset (JobsDataset_Training) που περιέχει περιγραφές θέσεων εργασίας (job description), τον τίτλο της κάθε θέσης (job title), και την γενικότερη κατηγορία στην οποία ανήκει κάθε θέση (query). Καλείστε να φτιάξετε και να εκπαιδεύσετε ένα τέτοιο σύστημα, το οποίο τελικά θα δέχεται ως είσοδο ένα job description και θα αναγνωρίζει τη γενικότερη κατηγορία στην οποία ανήκει. Στη διάθεση σας έχετε και το testing dataset (JobsDataset_Testing).

Υπόμνημα Dataset 8

Query: Κατηγορία που ανήκει η θέση εργασίας
Job Title: Τίτλος θέσης εργασίας

Description: Περιγραφή της θέσης εργασίας

Εργαστηριακή Άσκηση 9

Σκοπός της εργαστηριακής άσκησης είναι να δημιουργηθεί ένα σύστημα που θα χαρακτηρίζει την εντύπωση που προκάλεσε μια ταινία στους θεατές της ως θετική ή αρνητική. Έχετε στη διαθεσή σας ένα dataset (Movies Dataset) που περιέχει κριτικές θεατών για ταινίες (Summary), καθώς και την αξιολόγηση για το αν η κριτική είναι θετική ή αρνητική (Sentiment). Καλείστε να φτιάξετε και να εκπαιδεύσετε ένα τέτοιο σύστημα, το οποίο τελικά θα δέχεται ως είσοδο κριτικές θεατών και θα αναγνωρίζει την κατηγορία στην οποία ανήκουν. Μπορείτε να χρησιμοποιήσετε ένα μέρος του δεδομένου dataset για training του συστήματος, και ένα άλλο μέρος του για testing.

Υπόμνημα Dataset 9

Summary: Κριτική θεατών για ταινίες
Sentiment: Αξιολόγηση της κριτικής (1: θετική, 0:αρνητική)