Εργασία για το μάθημα Τεχνικές Εξόρυξης Δεδομένων
Εαρινό εξάμηνο , Ακ. Έτος 2021-22
Πρόβλεψη διάρκειας ταξιδιού (ομαδική εργασία)

Η Νέα Υόρκη είναι γεμάτη με μονόδρομους, μικρούς παράδρομους και έναν σχεδόν ανυπολόγιστο αριθμό πεζών σε κάθε δεδομένη χρονική στιγμή. Για να μην αναφέρουμε τον αριθμό των αυτοκινήτων/μοτοσυκλετών/ποδηλάτων που φράζουν τους δρόμους. Συνδυάστε το με μια τρελή βιασύνη για να φτάσετε από το σημείο Α στο σημείο Β, είναι πολύ πιθανό ότι δεν θα είστε στην ώρα σας.

Η λύση για να φτάσετε από τον προορισμό Α στον Β όταν ζείτε σε μια πόλη όπως η Νέα Υόρκη (χωρίς να χάσετε το μυαλό σας) είναι εύκολη: πάρτε ταξί/Uber/Lyft/κ.λπ. Δεν χρειάζεται να αγχώνεστε για την κίνηση ή τους πεζούς και θα έχετε λίγο χρόνο για να κάνετε κάτι άλλο, όπως να προλάβετε τα email. Αν και αυτό ακούγεται αρκετά απλό, δεν σημαίνει ότι θα φτάσετε στον προορισμό σας εγκαίρως. Επομένως, πρέπει να έχετε τον οδηγό σας να κάνει το συντομότερο δυνατό ταξίδι. Όταν λέμε το συντομότερο, μιλάμε για χρόνο. Εάν μια διαδρομή Α είναι Χ χιλιόμετρα *μακρύτερη*, αλλά σας πηγαίνει εκεί Υ λεπτά *γρηγορότερα* από τη διαδρομή Β, προτιμήστε να ακολουθήσετε αυτήν.

Για να γνωρίζουμε ποια διαδρομή είναι η καλύτερη για να ακολουθήσουμε, πρέπει να είμαστε σε θέση να προβλέψουμε πόσο θα διαρκέσει το ταξίδι όταν ακολουθούμε μια συγκεκριμένη διαδρομή. Επομένως, **ο στόχος αυτού της εργασίας είναι να προβλέψει τη διάρκεια κάθε ταξιδιού στο σύνολο δεδομένων δοκιμής, με δεδομένες τις συντεταγμένες έναρξης και λήξης.**

Σας δίνονται 3 αρχεία, (train.csv, test.csv, target.csv). Θα χρησιμοποιήσετε τη βιβλιοθήκη Pandas και συγκεκριμένα την συνάρτηση `read_csv` για να διαβάσετε τα δεδομένα.

Οι στήλες στα αρχεία περιγράφονται παρακάτω:

id - ένα μοναδικό αναγνωριστικό για κάθε ταξίδι
vendor_id - έναν κωδικό που υποδεικνύει τον πάροχο που σχετίζεται με την εγγραφή ενός ταξιδιού
pickup_datetime - ημερομηνία και ώρα δέσμευσης του μετρητή
dropoff_datetime - ημερομηνία και ώρα απενεργοποίησης του μετρητή
passenger_count - ο αριθμός των επιβατών στο όχημα (τιμή που εισήγαγε ο οδηγός)
pickup_longitude - το γεωγραφικό μήκος όπου ξεκίνησε ο μετρητής
pickup_latitude - το γεωγραφικό πλάτος όπου ξεκίνησε ο μετρητής
dropoff_longitude - the longitude where the meter was disengaged
dropoff_latitude - το γεωγραφικό μήκος όπου απενεργοποιήθηκε ο μετρητής
store_and_fwd_flag - Αυτή η τιμή υποδεικνύει εάν το αρχείο ταξιδιού διατηρήθηκε στη μνήμη του οχήματος πριν αποσταλεί στον προμηθευτή επειδή το όχημα δεν είχε σύνδεση με τον διακομιστή (Y=store and forward; N=not a store and forward trip)
trip_duration - διάρκεια του ταξιδιού σε δευτερόλεπτα
Προφανώς το dropoff_datetime και το trip_duration είναι διαθέσιμα μόνο για το σύνολο του train set.

Αναφορές

Scikit-Learn
Pandas
Numpy
Seaborn

Ερωτήματα

Το πρώτο σετ ερωτημάτων αφορά την **ανάλυση των δεδομένων** για την κατανόηση τους και την εξαγωγή συμπερασμάτων σχετικά με αυτά. Το δεύτερο σετ ερωτημάτων αφορά στη **συσταδοποίηση** (clustering) και **πρόβλεψη** του χρόνου ταξιδίου.

Ανάλυση Δεδομένων

1) Χρησιμοποιήστε την **.describe()** στο pandas DataFrame του train set για να πάρετε μία συνοπτική εικόνα των στατιστικών του dataset. Θα χρειαστείτε να αφαιρέσετε κάποια **outliers** (ακραία data samples) για την καλύτερη

διαχείριση των δεδομένων. Χρησιμοποιώντας τη στήλη “trip_duration” ως βάση, αφαιρέστε τα δεδομένα τα οποία βρίσκονται περισσότερο από 2 τυπικές αποκλίσεις μακρυά από τη μέση τιμή για αυτό το feature. (Μπορεί να σας βοηθήσει και η κλάση StandardScaler από το sklearn εδώ)

2) Επίσης, **τα όρια της πόλης της NY** με γεωγραφικά μήκη και πλάτη είναι:
   city_long_border = (-74.03, -73.75)
   city_lat_border = (40.63, 40.85)
   Αφαιρέστε όλα τα σημεία που πέφτουν εκτός αυτών των ορίων για να επικεντρωθείτε στην πόλη της NY.
3) Χρησιμοποιήστε την συνάρτηση **pd.to_datetime()** της βιβλιοθήκης pandas για να αλλάξετε τη μορφοποίηση των μεταβλητών ημερομηνίας (συγκεκριμένα τις τιμές για τις στήλες `pickup_datetime` και `dropoff_datetime`) τόσο για τα train όσο και για τα test δεδομένα. Χρησιμοποιήστε το μετασχηματισμένο πεδίο datetime και δημιουργήστε τις παρακάτω στήλες (και για το pickup και για το dropoff)
   - ***day_period*** - το όνομα της περιόδου της ημέρας i.e. morning, evening, συγκεκριμένα δημιουργήστε 4 περιόδους ‘Morning’ (από 6:00 am μέχρι 11:59 pm), ‘Afternoon’ (από 12pm μέχρι 3:59 pm), ‘Evening’ (από 4:00 pm μέχρι 9:59 pm), και ‘Late Night’ (από 10:00 pm to 5:59 am)
   - ***day_name*** - ημέρα της εβδομάδας i.e. Monday, Tuesday, etc…
   - ***month*** - μήνας
   - ***hour*** - η ώρα σε 24h format
   - ***year*** - η χρονιά
   - ***season*** - η εποχή

   Με τις νέες στήλες που δημιουργήσατε φτιάξτε τα παρακάτω ιστογράμματα (μόνο για τα train)
   - Ιστογραμμα για τις πιο πολυσύχναστες pickup days
   - Ιστογραμμα για τις πιο πολυσύχναστες dropoff days
   - Ιστόγραμμα για Trips per day_period (για pickup και dropoff)
   - Ιστόγραμμα για Trips per month (για pickup και dropoff)

   Στη συνέχεια σχεδιάστε με όποιο τρόπο θέλετε τα παρακάτω
   - Μέση διάρκεια ταξιδιού ανά pickup hour
   - Μέση διάρκεια ταξιδιού ανά day_period

- Μέση διάρκεια ταξιδιού ανά Day of Week

4) Σχεδιάστε ένα **απλό ιστόγραμμα για τη στήλη trip duration**, χωρίζοντας τα δεδομένα σε 100 bins. Στη συνέχεια εφαρμόστε τα παρακάτω
   a) Εφαρμόστε **log transformation** στη στήλη trip duration και βάλτε τα αποτελέσματα σε μία νέα στήλη train['**log_trip_duration**']
   b) Σχεδιάστε το ιστόγραμμα για την νέα στήλη.

5) Σχεδιάστε την **μέση διάρκεια ταξιδιού** για τους διαθέσιμους παρόχους.

6) Μελετήστε τον αντίκτυπο που μπορεί να έχει **ο αριθμός των επιβατών ανά ταξίδι στον χρόνο τους ταξιδιού**. Για να διαπιστώσετε εάν πράγματι υπάρχει σημαντική επιρροή στον χρόνο ταξιδιού, μπορείτε να ομαδοποιήσετε τους μέσους χρόνους ταξιδιού με βάση το “passenger_count”.

7) Στη συνέχεια θα προστεθεί μία νέα στήλη **στα train και στα test δεδομένα**, η οποία θα υπολογίζει την συντομότερη απόσταση μεταξύ δύο σημείων στην επιφάνεια της γης, χρησιμοποιώντας την απόσταση Manhattan. (Μπορείτε να αξιοποιήσετε την υλοποίηση του sklearn)
   Στη συνέχεια σχεδιάστε τα παρακάτω διαγράμματα
   - Μέση απόσταση διαδρομής ανά ημέρα της εβδομάδας
   - Απόσταση διαδρομής και χρόνος διαδρομής

8) Χρησιμοποιήστε την απόσταση και το χρόνο για να υπολογίσετε τη **μέση ταχύτητα**. Συγκεκριμένα δημιουργήστε γραφήματα για να περιγράψετε πώς η ώρα της ημέρας, η ημέρα της εβδομάδας και ο μήνας του έτους επηρεάζουν τη μέση ταχύτητα.

9) Χρησιμοποιώντας τις στήλες lat, long δημιουργήστε δύο **scatter plots** για τα σημεία pickup και dropoff.

Συσταδοποίηση και πρόβλεψη διάρκειας ταξιδιού

1) Η συσταδοποίηση θα γίνει με χρήση του αλγορίθμου clustering ***K-Means***. Το clustering θα πρέπει να υλοποιηθεί μόνο για τα pickup σημεία με βάση τις συντεταγμένες Latitude και Longitude από το train set. Για την επιλογή του βέλτιστου αριθμού των clusters μπορείτε είτε να δοκιμάσετε μερικές τιμές

εμπειρικά είτε να δοκιμάσετε με την μέθοδο ***elbow*** (δείτε και τις διαφάνειες του scikit-learn-intro σχετικά). Αφού εκπαιδεύσετε τον K-means με το βέλτιστο αριθμό κέντρων, απεικονίστε σε ένα **scatter plot** με τα Latitude και Longtitude ως άξονες με:

a) Τα σημεία του train-set (με χρώμα το αντίστοιχο cluster που ανήκουν)
b) Τα κεντροειδή των clusters

2) Η πρόβλεψη της διάρκειας ταξιδιού, trip_duration, θα γίνει με χρήση του ***RandomForestRegressor*** από την βιβλιοθήκη sklearn. Η στήλη στην οποία θα γίνει η πρόβλεψη θα είναι η **log_trip_duration** που δημιουργήσατε σε προηγούμενο ερώτημα. Προσοχή: για το ερώτημα αυτό τα train και τα test δεδομένα πρέπει να έχουν τις ίδιες στήλες από features. Ενδεικτικά μπορείτε να αξιοποιήσετε τις:
   a) vendor_id
   b) passenger_count
   c) pickup_longitude
   d) pickup_latitude
   e) dropoff_longitude
   f) dropoff_latitude
   g) pickup_day_name
   h) pickup_month
   i) pickup_season
   j) pickup_day_period
   k) pickup_hour
   l) manhattan_distance

   Επίσης χρησιμοποιήστε την ***feature_importances_*** https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html και βρείτε ποιά είναι τα πιο σημαντικά χαρακτηριστικά (μπορείτε να τα παρουσιάσετε και σε διάγραμμα)

3) Με χρήση της ***GridSearchCV*** δοκιμάστε να βελτιώσετε το μοντέλο σας βρίσκοντας το καλύτερο σύνολο παραμέτρων. Πειραματιστείτε με τις παραμέτρους **max_depth**, **max_features**, **n_estimators**.
4) Εκπαιδεύστε πάλι τον ***RandomForestRegressor*** με τις νέες παραμέτρους.
5) Μετρήστε την απόδοση προβλέποντας τα test δεδομένα χρησιμοποιώντας το αρχείο target.csv τις παρακάτω μετρικές **MAE**, **MSE (!** Θυμηθείτε να κάνετε τον log μετασχηματισμό και στη στήλη *trip_duration* του target_csv!)

Bonus:

1) Χρησιμοποιώντας τη βιβλιοθήκη folium , φτιάξτε έναν χάρτη της Νέας Υορκης όπου θα φαίνονται είτε τα pickup είτε τα drop-off σημεία. Σε περίπτωση που ο χάρτης αργεί πολύ να εμφανίσει τα σημεία, χρησιμοποιήστε ένα μικρότερο υποσύνολο του train set.
2) Αν θέλετε μπορείτε να πειραματιστείτε περαιτέρω με τα δεδομένα. Για παράδειγμα:
    a) Να αναλύσετε περαιτέρω τα δεδομένα **εξάγοντας νέα συμπεράσματα**, όπως να βρείτε τον πιο συχνό vendor, να βρείτε ώρες που υπάρχει traffic congestion κ.α.. Παρουσιάστε σχόλια ή/και γραφικές σχετικά με αυτά.
    b) Να δημιουργήσετε **άλλα features** που πιστεύετε ότι μπορεί να βοηθήσουν το μοντέλο πρόβλεψης, όπως ένα feature για το αν η διαδρομή έγινε το ΣΚ ή όχι, καθώς και άλλα αντίστοιχα. Παρουσιάστε τα ανανεωμένα αποτελέσματα με σχετικά σχόλια.

Παραδοτέο:

Η εργασία πρέπει να εκπονηθεί με τις ίδιες ομάδες που είχατε στην πρώτη εργασία.

Ο φάκελος scr της εργασίας είναι ο φάκελος στον οποίο θα γράψετε τον κώδικά σας και είναι και αυτός που θα παραδώσετε (δηλαδή δεν θα παραδώσετε εκ νέου τα τα δεδομένα εκπαίδευσης/δοκιμής). Θα ανεβάσετε στο eclass ένα φάκελο της μορφής scr_sdixxxx. (όπου sdi το ΑΜ ενός εκ των ατόμων της ομάδας).

Ο κώδικάς σας πρέπει να περιέχει **ΥΠΟΧΡΕΩΤΙΚΑ ένα Ipython notebook** με το οποίο θα μπορεί κάποιος να τρέξει την εργασία σας βήμα-βήμα. Μπορείτε να έχετε και *.py αρχεία με τις συναρτήσεις σας αλλά η εργασία πρέπει να τρέχει από ένα notebook. Στο notebook μπορείτε σε όποια σημεία κρίνετε απαραίτητο να εισάγετε visualizations με τον τρόπο που θα εξηγήσουμε στα φροντιστήρια. Το notebook αποτελεί και την ολοκληρωμένη αναφορά για την εργασία σας (δεν θα παραδώσετε τίποτα σε doc, pdf) , σχεδιάστε το με προσοχή, να θυμάστε να γράψετε μία περιγραφή σε κάθε βήμα για το τι κάνει ο κώδικάς σας σε κάθε κελί. Το notebook πρέπει να παραδοθεί “**τρεγμένο**” με τα αποτελέσματα εμφανή.

Επίσης για όλα τα ερωτήματα (στο μέτρο του εφικτού-χρήσιμου), **τα ποιοτικά σχόλια είναι απαραίτητα** σχετικά με το τι παρατηρείτε, ποιό συμπέρασμα προκύπτει από τις γραφικές κλπ. Ειδικά και για το δευτερο σετ ερωτημάτων (clustering και regression) δεν μας ενδιαφέρει τόσο η απόδοση των μοντέλων, όσο η σωστή εκτέλεση των βημάτων και η ποιοτική περιγραφή για αυτά.

Για πιο γρήγορες δοκιμές, μπορείτε να ξεκινήσετε την εργασία σας με λίγα από τα δεδομένα και στο τέλος να την ξανατρέξετε με όσο το δυνατόν πιο πολλά από τα δεδομένα.