TAFE NSW

BACHELOR OF DATA ENGINEERING

ITDAT301A BIG DATA AND ADVANCED DATABASE CONCEPTS

PROJECT 1

NYC Taxi Trip Problem

**PREPARED BY:**

Ashitosh Dhungana

SUMMARY

The NYC Taxi Trip Duration Prediction project aims to predict the duration of taxi trips in New York City using features like pickup and drop-off locations, pickup time, and passenger count. This model aims to enhance dispatching, route planning, and fare estimations by precisely forecasting trip durations.

The dataset appears to have included trip information such as timestamps, journey duration, and pickup and drop-off locations that will be used by the machine learning model later on to predict the required outcomes.

Table of Contents

Introduction……………………………4

Classification and Sanitization……….4

Feature Engineering………………….5

Explanatory Data Analysis……………6

Analysis and Findings………………..9

Machine Learning……………………12

Conclusion……………………………17

References……………………………17

Introduction:

The New York taxi trip duration operation is a project that is basically undertaken to develop a machine learning model capable of accurately predicting taxi trip durations in New York City.

This project's goal is to forecast how long a taxi ride in New York City will take depending on a number of variables, including the time of day, day of the week, passenger count, and the locations of pickup and drop-off. Both passengers and taxi drivers can gain from accurate trip duration forecasts, which can enhance dispatching, time management, and fare estimates.

Accurate trip time predictions are now crucial for improving user experience, expediting fare calculations, and optimizing fleet management due to the ride-hailing and urban mobility services' exponential expansion. The goal of this research is to develop a model that can produce these predictions with a high level of accuracy by utilising historical data and cutting-edge machine learning techniques.

Variables Classification and Sanitization:

Classification:

Id(object): Unique identifier for each trip

vendor\_id(int 64): Identifier for the provider of the taxi service

pickup\_datetime(object): Describes the timestamp when the trip began

dropoff\_datetime(object):Describes the time when the trip ended

trip\_duration(int64): Target variable, also the total duration of the trip in seconds

passenger\_count(int64): Total passengers in the taxi during the trip

pickup\_longitude (float64): Longitude coordinate of the pickup location.

pickup\_latitude (float64): Latitude coordinate of the pickup location.

dropoff\_longitude (float64): Longitude coordinate of the dropoff location.

dropoff\_latitude (float64): Latitude coordinate of the dropoff location.

store\_and\_fwd\_flag(object): Notifies if the trip data was stored before sending (Y) or sent immediately (N)

Sanitization:

To curate the dataset and to initiate preprocessing, the following method was followed:

The dataset was initially loaded and split into appropriate columns. Missing values in the dataset were checked and the column that required was converted to datetime format.

Fig 1: Removing null values

Feature Extraction and Engineering:  
In order to enhance the dataset, the process of feature engineering was initiated by identifying new variables that incorporate contextual, temporal, and spatial information pertinent to trip durations. Using the Haversine formula, which calculates the straight-line distance between pickup and drop-off locations based on geographic coordinates, firstly trip distance was determined.

From this, the computation of average speed of each trip, giving insight into potential traffic or route conditions was also performed. Furthermore, creation of flags for rush hour and weekend trips, indicating times when travel patterns and traffic flow may differ was also carried forward Additionally, the introduction of a passenger-distance interaction variable, which considers how the number of passengers and trip distance together might impact trip duration was also formed.

A screenshot of a computer program

Description automatically generated

Fig 2: Feature Engineering

Outlier Detection and Explanatory Data Analysis:

The distribution and correlations between the main features were shown using histograms, pair plots, and a correlation heatmap. Prior to using the univariate and bivariate analysis methods, this aids in a better comprehension of the data.

Also, outliers were detected using the interquartile range (IQR) method, and their presence was visualized using box plots. Identifying outliers is crucial as they can significantly impact the performance of clustering algorithms.

A screenshot of a computer

Description automatically generated

A graph with a blue rectangle

Description automatically generated

Fig 3: Outlier Detection

The distribution of trip\_duration variable shows that there lies presence of extreme outliers as it is clearly visible, as some data points extend far beyond the typical range of trip durations.

Similarly, the distribution and the range of average journey lengths become much more apparent after the outliers are eliminated. It is simpler to determine the median and interquartile range (IQR) of the data since the plot displays a clearly defined box with whiskers that extend to acceptable trip durations. Since the extreme values that would have distorted the results are no longer evident, this enhanced visualization facilitates more accurate analysis and model training.

If additional feature is derived from trip\_duration, such as average\_speed\_kmh, outliers in trip\_duration could skew the derived feature. For example, a very high trip\_duration value could lead to an unusually low average\_speed\_kmh, which may not be meaningful and could impact model accuracy.

Similarly, the use of correlation matrix was pursued to depict how various features in the dataset are correlated to each other. Also, log transformation is used to handle skewness in variable’s distribution as there were more outliers present in the beginning. This helped to make the data more distributed and stabilized the variance.

The features that are highly correlated are log\_trip\_duration and trip\_distance because it shows that when one increases the other also tends to increase.

A screenshot of a computer

Description automatically generated

Fig 4: Correlation matrix

Analysis and Findings:

While the median journey duration stays constant throughout the day, the analysis shows that there is more fluctuation and a greater number of outliers in the late-night and early-morning hours. This can be the result of varied trip features or less consistent travel schedules at these hours. By reducing skewness, the log scale facilitates trend identification and variability comparison across various pickup periods.

A graph of a diagram

Description automatically generated with medium confidenceFig 5: Box plot visualizing the trip duration across various pickup hours

Also when compared to the frequency of trips by the hour of the day for both pickup and drop-off times, it can be seen that the data provide a clear view of taxi demand patterns by hour, showing peaks in the late afternoon/evening, with a typical decline during late-night and early-morning hours. This data is valuable for understanding high-demand periods and planning accordingly, especially for resource distribution or pricing models in the transportation industry.

The likelihood for this to happen is because the morning (around 7–9 AM) and evening (around 5–7 PM) peaks align with typical commute times for work and school. People often use taxis as a reliable form of transport during these hours, especially in urban areas.

Therefore, it can be clearly noted that these trends deliver the fact that peak hours are primarily indicative of normal commuter behavior, with supplementary influences from urban living patterns. Since the data accurately depicts these typical human behaviors, the trends seen are typical of most taxi datasets.

A graph of different colored bars

Description automatically generated with medium confidenceFig 6: Bar chart describing frequency of trips

Similarly, when there’s a bivariate analysis done on the following dataset, some important results have been drawn as well.

When two features, trip duration and trip distance were analyzed together, there is a general upward trend, suggesting that longer trip distances tend to have longer durations. This correlation aligns with common sense, as longer trips typically require more time to complete. Shorter durations and shorter distances seem to be associated with a dense cluster of data points. Given that most taxi rides in cities are often somewhat brief, this is to be expected.

A blue dotted graph with white text

Description automatically generated with medium confidenceFig 7 :Comparison between two features

Furthermore, it has also been suggested that The average passenger count for Vendor ID 2 is higher than that for Vendor ID 1. This suggests that Vendor 2 generally transports more passengers per trip on average compared to Vendor 1.

This disparity could also result from differing operational practices, policies, or types of vehicles in use by each vendor. Vendor 2 might use larger vehicles or target areas with more demand for multiple-passenger trips.

A graph of a number of blue squares

Description automatically generated

Fig 8: Difference of passenger counts between two vendor

Machine Learning and Integration:

For this project, the model chosen was XGBoost Regressor, a powerful gradient boosting algorithm because of its anility to use both gradient boosting and regularization techniques to reduce overfitting and improve prediction accuracy.

In this dataset, where variables like time, location, distance, and other parameters may interact in non-linear ways to effect journey duration, XGBoost's ability to capture complicated interactions between features is very helpful.

While Linear Regression was thought for the prediction, they assume a linear relationship between features and the target variable. In this case, trip duration is influenced by multiple non-linear interactions, which linear regression might struggle to capture effectively.

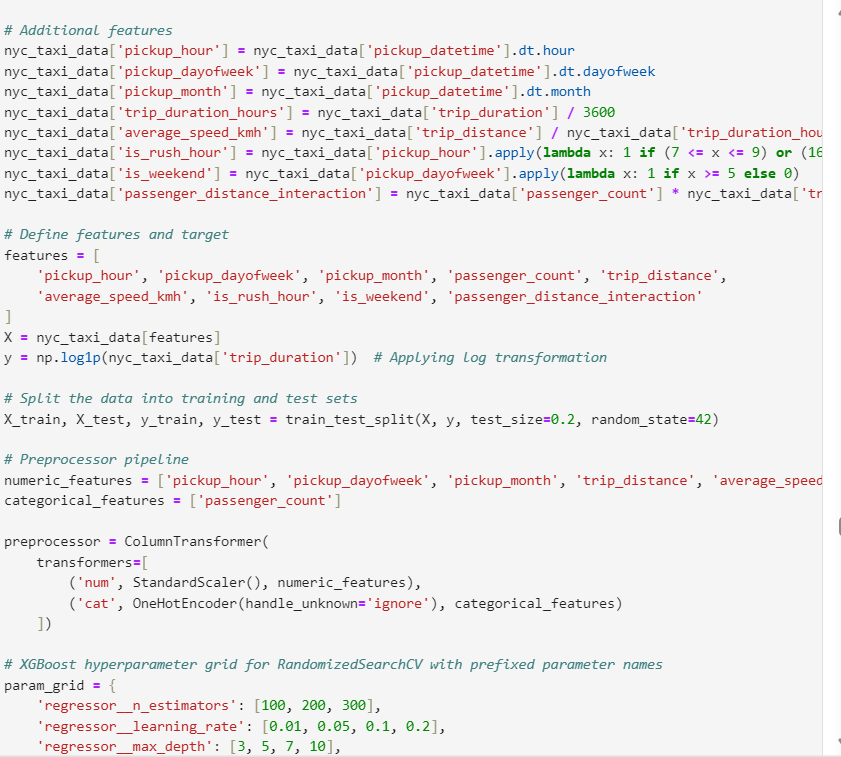


Fig 9 : Model Training

A screen shot of a graph

Description automatically generated

Fig 10: Metrics resulted from the model training

From the Mean Absolute Error, it can be seen that on average, the model’s prediction are off by about 48 secs which can relatively be considered as a good performance.

An R² score of 0.966 means that 96.6% of the variance in the trip duration can be explained by the model. This is a strong score, indicating that the model captures most of the patterns in the data well.

The graph plots actual trip durations on the x-axis and predicted trip durations on the y-axis, with the red dashed line representing perfect predictions. Most of the points are close to the red line, indicating that the model's predictions are generally accurate with few points deviating from the line.

When comparison between the distribution of actual and predicted trip duration is made post model training, it is evident that the blue line represents the actual trip duration distribution, and the orange line represents the predicted distribution from the model. The close overlap of these two lines indicates that the model's predictions closely follow the true distribution of trip durations.

To normalize the distribution and lessen the impact of outliers or extreme values, log transformation was used. When the original data is significantly skewed—for example, there are a lot of short trips and few lengthy journeys, this method is frequently employed.

The tight match between the real and predicted distributions in this graphic indicates that the model has learned the trip time distribution successfully overall. This points to a good model fit, especially following the log transformation, which enhanced predicted accuracy over the range and lessened the influence of extreme values.

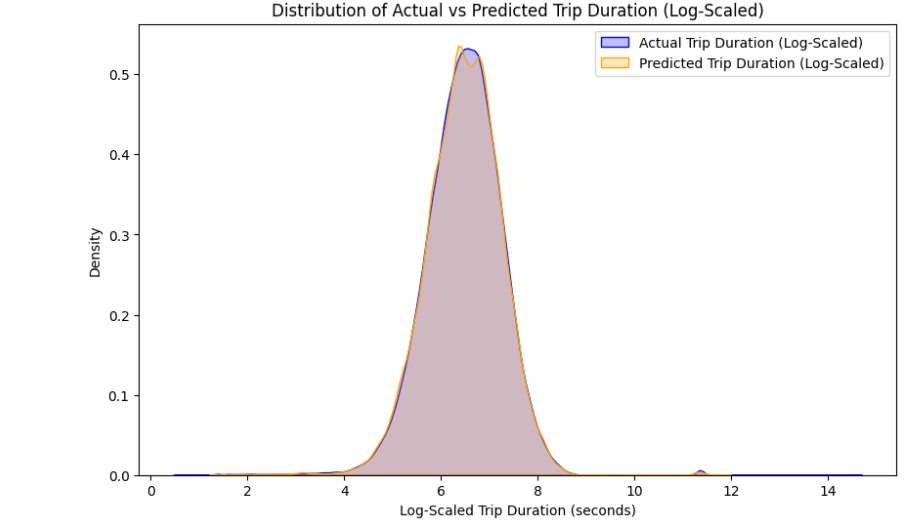


Fig 11: Model’s Prediction

Finally, when weekend and weekday trip predicted by the model shows that most of the data points are clustered close to the diagonal line, particularly for shorter trips. This suggests that the model performs well for weekend trips of typical durations.

The right plot shows a more concentrated clustering near the origin, as weekday trips generally seem to have shorter actual durations than weekend trips.

The model appears to perform better on weekday trips, likely due to more consistent travel patterns, such as commuting times.

This also proves that the model effectively captures the majority of trip durations, as evidenced by most points aligning with the diagonal line. This indicates that the model can reliably predict standard trip durations based on the provided features (such as trip distance etc.).However, there are some outliers with significantly high actual durations, where the model underestimates the trip duration (shown as points deviating above the line).

A screenshot of a graph

Description automatically generated

Fig 12: Model’s Prediction for trip duration

Conclusion:

To conclude, using a range of parameters, such as trip distance, passenger count, and other temporal variables, the goal of this study was to create a prediction model for calculating the length of taxi trips.

The model exhibits good accuracy on average trip lengths after being fine-tuned through substantial feature engineering and hyperparameter optimization, especially on weekdays when travel patterns are more predictable.

It has been classified by the model that:

The model accurately captures most trip durations, particularly shorter and moderate-duration trips.

Weekday trips are more predictable than weekend trips due to their routine nature, while weekend trips showed greater variability.

Trip duration outliers pose a problem for the model and point to regions that need work, particularly on weekends.

References:

Ramraj, S., Uzir, N., Sunil, R. and Banerjee, S., 2016. Experimenting XGBoost algorithm for prediction and classification of different datasets. *International Journal of Control Theory and Applications*, *9*(40), pp.651-662.

Donovan, B. and Work, D.B., 2014. New york city taxi data (2010-2013). *Dataset, http://dx. doi. org/10.13012/J8PN93H8*.

Huberty, C.J. and Morris, J.D., 1992. Multivariate analysis versus multiple univariate analyses.