Trip Duration Prediction Project Report

1 Introduction

The NYC Taxi Duration Prediction competition on Kaggle challenges participants to build a model that predicts the total ride duration of taxi trips in New York City. The primary dataset is provided by the NYC Taxi and Limousine Commission and includes information such as pickup time, geo-coordinates, number of passengers, and other variables.

2 Shape of the data

2.1 Target Variable (Trip Duration):

The target variable distribution see Figure 3 looks like Gaussian distribution, there is a long right tail (right-skewed) which means that there are a few very long trips that may be outliers compared to the majority of trips.

The peak of the distribution is around 5: This means that most trips are between 150 seconds and 1000 seconds (about 2.5 and 16.7 minutes) long

The max trip duration took around 58771 minutes is approximately 40 days and 1111 minutes so definitely an outlier.

Note: we perform

$$\log(1+x)$$

transformation to Visualise better and **to help us with modeling large values.**

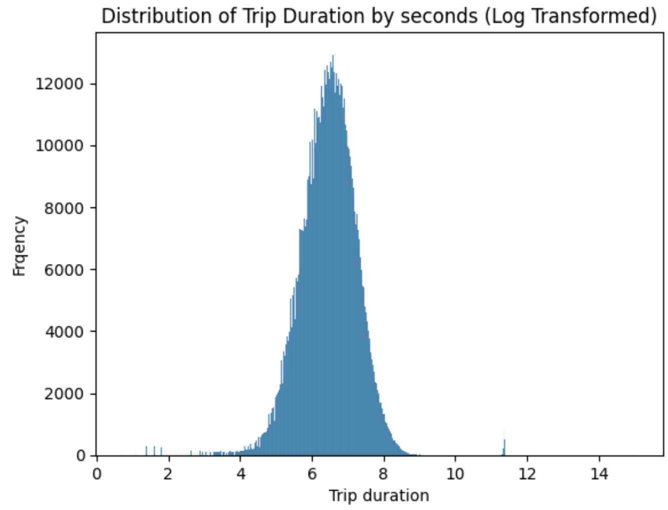


Figure 1: Trip duration distribution

2.2 Discrete Numerical Feature

Vendor ID and passenger count are our Discrete Numerical Features.

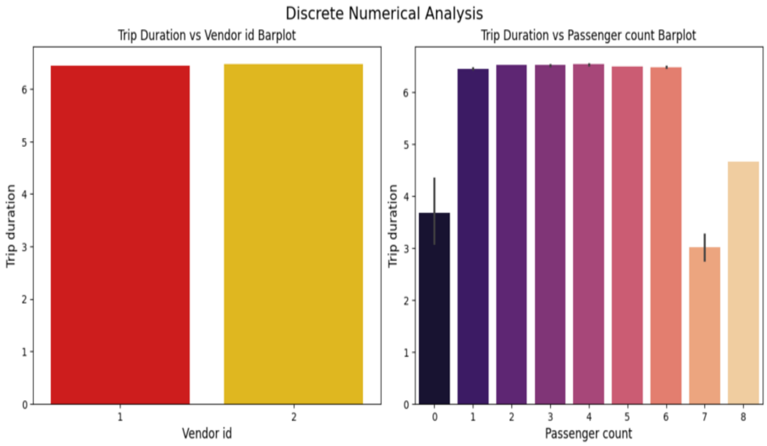


Figure 2: Discrete Numerical Analysis

Trip duration and Vendor ID: It's difficult to discern any clear patterns or trends between trip duration and vendor ID from the bar chart. The bars appear to be spread out relatively evenly across the x-axis, suggesting no significant difference in trip duration among vendors.

When the number of passenger groups from [1 to 6] take constant trip duration and the number of passenger groups from [7 to 8] take less trip duration, Possible explanations:

Vehicle Capacity: The vehicles used by both vendors may have a maximum capacity of 6 passengers. When there are 6 or fewer passengers, the vehicles are operating at their maximum capacity, and the trip duration remains constant because the vehicles are fully utilized.

Vehicle Type: It's also possible that the two vehicle vendors have different types of vehicles in their fleets. One vendor might have larger vehicles capable of accommodating more passengers, while the other vendor might have smaller vehicles. The larger vehicles can comfortably accommodate 7 to 8 passengers, resulting in shorter trip durations.

Trip purpose: Trips with 7-8 passengers might be more likely to be for specific purposes (e.g., airport shuttle, group sightseeing) that involve shorter distances or routes optimised for efficiency.

If vendor of the taxi not effect with trip duration so idea of vehicle type and vehicle capacity not correct so we need to use a boxplot to detect If we just dealing with some random noise or passengers groups from [7 to 8] just travel less than another groups.

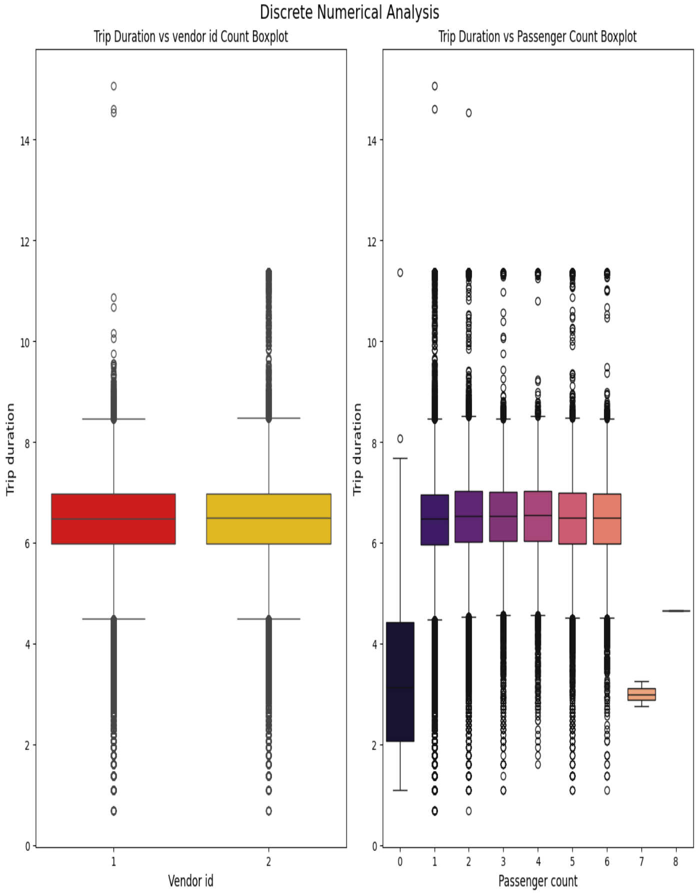


Figure 3: Discrete Numerical Analysis (Boxplot)

Trip duration distribution: The distribution of trip durations appears to be positively skewed for both groups, meaning there are more shorter trips than longer trips. This is especially evident for the group with lower passenger counts (left boxplot).

Median trip duration: The median trip duration, represented by the horizontal line within the box, is longer for the group with higher passenger counts (right boxplot). This suggests that trips with more passengers tend to take longer overall.

Outliers: There are a few outliers for both groups, represented by the circles beyond the whiskers. These are individual trips that were much longer than the majority of trips in their respective groups and this **Supports conclusion passengers groups from [7 to 8] just travel less than another groups because Trip purpose.**

2.3 Geographical Feature

Using pickup latitude, pickup longitude, dropoff latitude and dropoff longitude we can calculate **haversine distance**

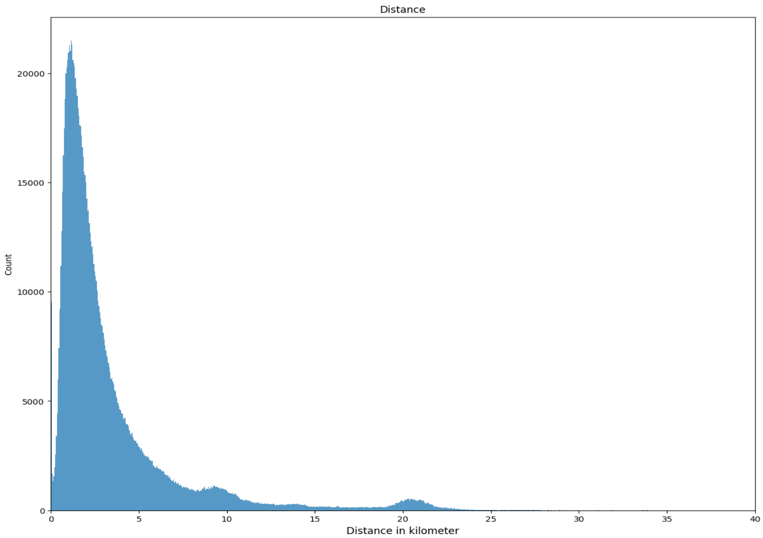


Figure 4: Haversine distance

This distribution's right skewed So we can use a transformation

$$y' = \log(x)$$

or

$$y' = \sqrt{y}$$

which can improve the performance of our Linear model

Looks like most of the trip goes from less than 1 kilometre to 25 kilometres. Also, we have the trip duration and distance so we can calculate the speed of the trip

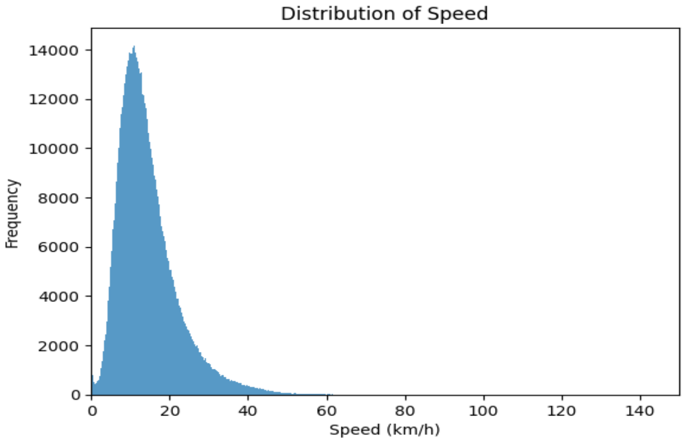


Figure 5: Distribution speed of the trips

2.4 Temporal/Time-date Analysis

Using data/time we can get new information like Months-day-Morning or afternoon or night-season for each trip and now we can try to find how it affects trip duration.

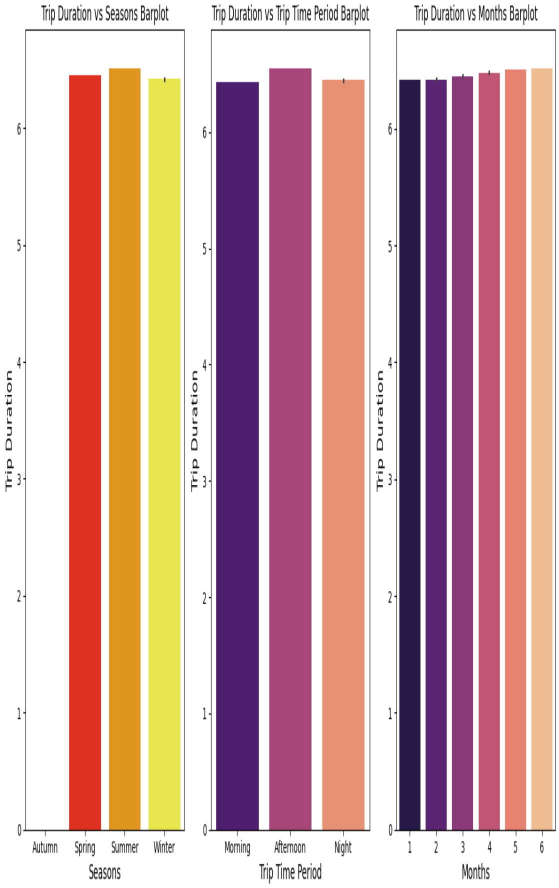


Figure 6: Time-date Analysis

The longer trip durations during summer might be attributed to vacations and holidays 6, which lead to increased traffic on the roads. Longer trip durations in the afternoon can be explained by higher levels of crowding during that time of day. April, May, and July experience longer trip durations compared to other months.

Trip durations are generally longer on weekends than on weekdays (Saturday and Sunday) 7.This is likely due to the fact that people are more likely to take longer trips for purposes on weekends.

Trip durations are generally shorter during the morning and evening rush hours. This is likely due to the fact that people are more likely to be making shorter trips for commuting purposes during these times.

Trip durations are generally longer in the middle of the day.This is likely due to the fact that people are more likely to be making longer trips for shopping, errands, or other activities during these times.

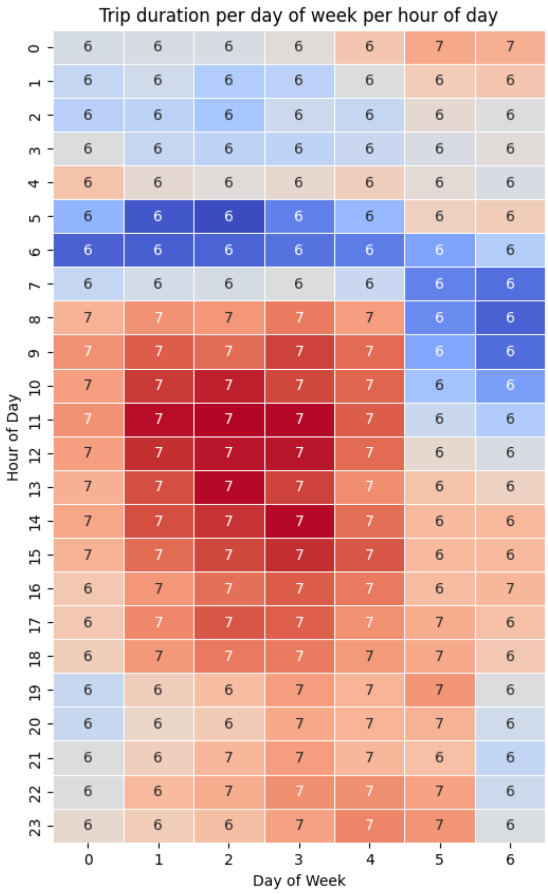


Figure 7: Time-date Analysis

3 Correlation Analysis

Their positive relation trip duration (in Seconds log-Transformed) with pickup longitude, dropoff longitude and passenger count. Figure 8

There negative relation trip duration (in Seconds log-Transformed) with pickup latitude and dropoff latitude. Figure 8

Their strong positive relation trip duration (in Seconds log-Transformed) with distance. Figure 9

Their negative relation trip duration (in Seconds log-Transformed) with speed km/h. Figure 9

Conclusion we have a pretty good relationship between the distance features and trip duration so we can use it in modeling. Figure 9

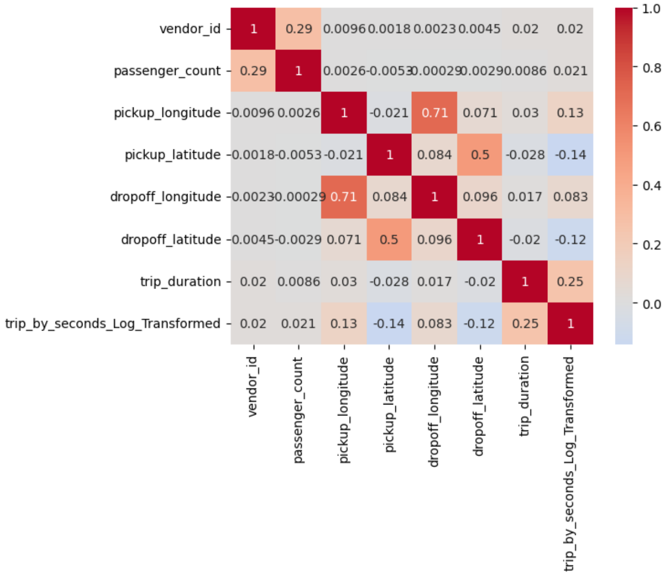


Figure 8: Correlation Analysis

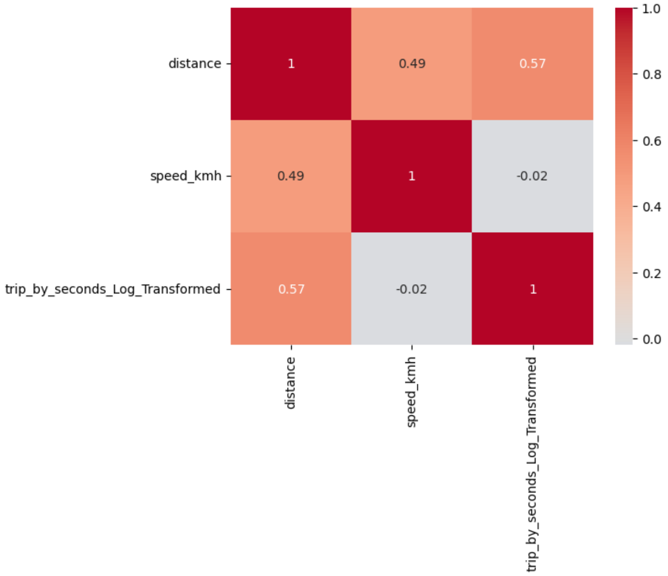


Figure 9: Correlation Analysis

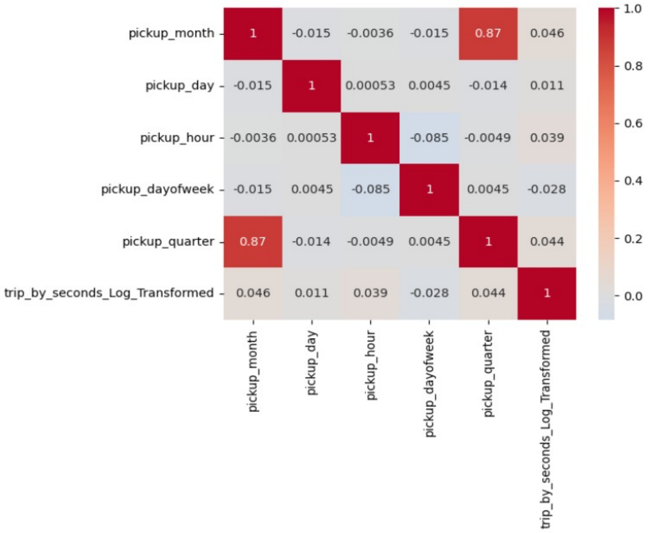


Figure 10: Correlation Analysis

4 Modeling

The data pipeline splits the feature into categorical and numerical features.

We perform One hot encoding for the categorical feature and For the numerical feature We scale the data by standard scalar then do Polynomial Features (degree=6) finally We use log transformation for the data. **As previously discussed We perform log transformation because the distance data is right skewed.**

4.1 Results

Metric	Train	Validation
RMSE	0.4376	0.4427
R^2	0.6968	0.6938

Table 1: Performance metrics for the model

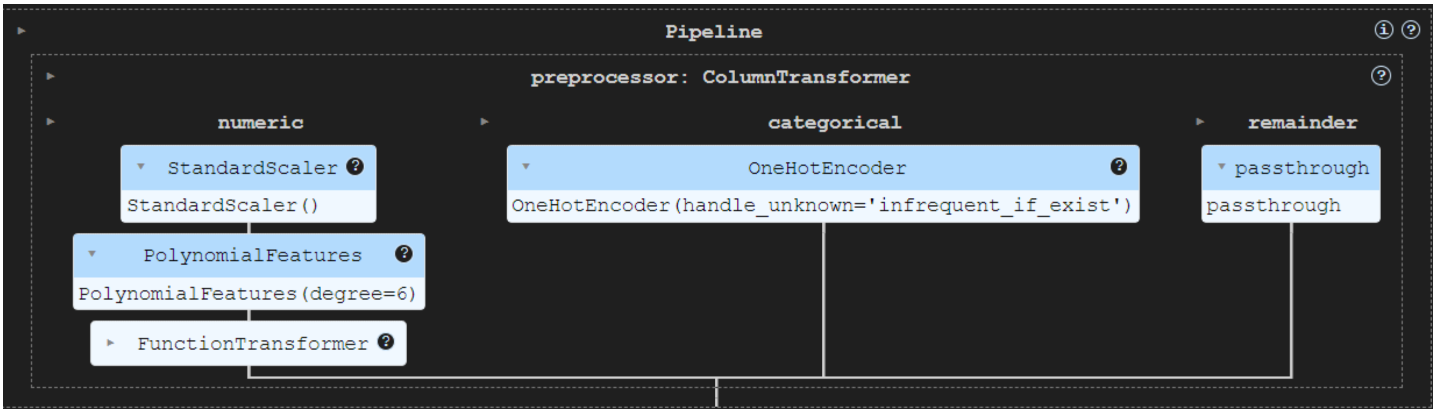
4.2 Lessons and future work

Having a type of version control for the data or model scores is very beneficial for error analysis and verifying assumptions. In this project, we observed the following insights:

- Feature selection consistently improves model performance.
- Outlier removal for intra-trip duration does not improve model performance.

Building a separate pipeline to estimate the speed feature and using it as an input for predicting trip duration did not significantly improve the model's performance.

Linear models often exhibit high bias. Given the characteristics of this dataset, it is important to explore more complex algorithms. For heterogeneous tabular data, techniques such as XGBoost and ensemble methods can be particularly effective.



Pipeline

preprocessor: ColumnTransformer

numeric

StandardScaler

StandardScaler()

PolynomialFeatures

PolynomialFeatures(degree=6)

FunctionTransformer

categorical

OneHotEncoder

OneHotEncoder(handle_unknown='infrequent_if_exist')

remainder

passthrough

passthrough

Figure 11: Data pipeline