Nyc Taxi Trip Time Prediction

By

Manmohan Madhu

**Abstract**

This project aims to train a model which can predict the time taken for a taxi trip in New York City.

**Problem Statement**

The task is to build a model that predicts the total ride duration of taxi trips in New York City. The primary dataset is one released by the NYC Taxi and Limousine Commission, which includes pickup time, geo-coordinates, number of passengers, and several other variables.

**Introduction**

The dataset provided contained details of a particular trip, such as pickup and dropoff datetime, coordinates, trip duration etc. Using this supervised machine learning models were trained and predictions done were evaluated.

**Challenges Faced**

* The datetime values were stored as strings. These were converted to datetime data type.
* Pickup and dropoff coordinates were given. These were converted using the Haversine formula to get distance.
* There were outliers in trip duration and distance.
* One hot encoding had to be performed for categorical variables.

**The Approach Used to Solve the Problem**

The pickup and dropoff datetime were stored as string datatype. After converting this to datetime format, day of the week and month were extracted and analysis done on the rest of the features. Id, passenger count and store and forward features were dropped after learning that they had no effect from the analysis. An additional distance feature was added after calculating the same from the pickup and dropoff coordinates. Outliers were treated and the models were trained.

**Libraries used for analysis**

1. Pandas : To load the data into a dataframe object and analyze.
2. Matplotlib : To help visualize the data.
3. Seaborn : For added functionality to matplotlib.
4. Numpy : To use the numpy functions in analysis.

From sklearn

1. Test train split
2. Grid search CV
3. Linear Regression
4. XGB Regressor
5. Ridge Regressor
6. Minmax scaler
7. Mean squared error

**Dataset**

The dataset is based on the 2016 NYC Yellow Cab trip record data made available in Big Query on Google Cloud Platform. The data was originally published by the NYC Taxi and Limousine Commission (TLC). The data was sampled and cleaned for the purposes of this project.

Data Description:

* id - a unique identifier for each trip
* vendor\_id - a code indicating the provider associated with the trip record
* pickup\_datetime - date and time when the meter was engaged
* dropoff\_datetime - date and time when the meter was disengaged
* passenger\_count - the number of passengers in the vehicle (driver entered value)
* pickup\_longitude - the longitude where the meter was engaged
* pickup\_latitude - the latitude where the meter was engaged
* dropoff\_longitude - the longitude where the meter was disengaged
* dropoff\_latitude - the latitude where the meter was disengaged
* store\_and\_fwd\_flag - This flag indicates whether the trip record was held in vehicle memory before sending to the vendor because the vehicle did not have a connection to the server - Y=store and forward; N=not a store and forward trip
* trip\_duration - duration of the trip in seconds

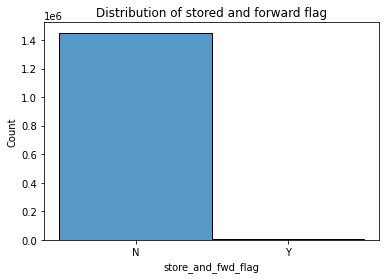
**Dataset preparation before analysis**

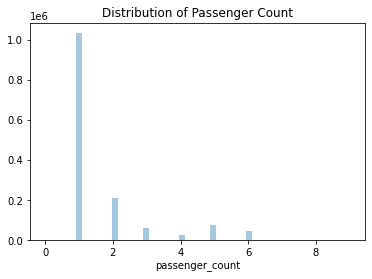
* The dataset did not contain null values.
* The pickup and dropoff datetimes were strings. This was converted to datetime datatype.
* Day of the week was extracted from datetime.
* Time was divided into 4 parts of the day and stored accordingly.

**Plots**

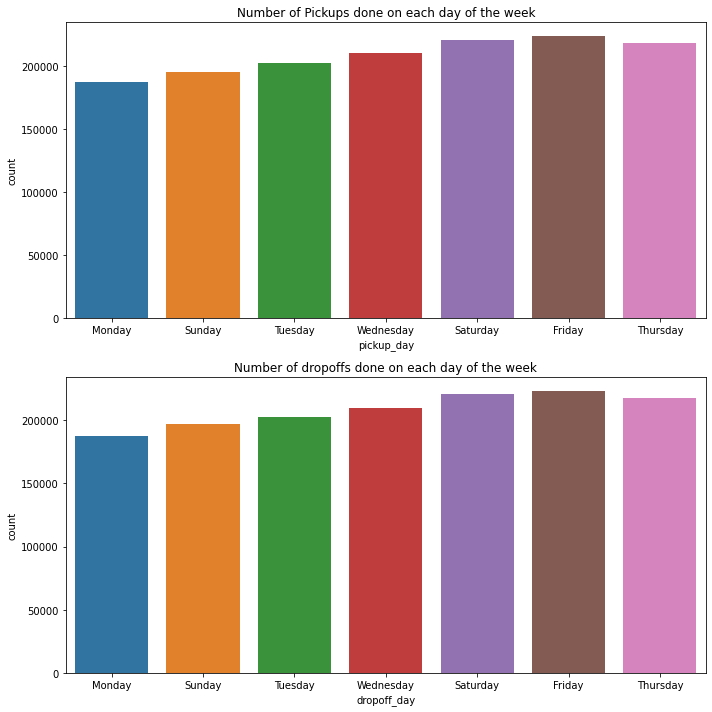
Univariate analysis of variables

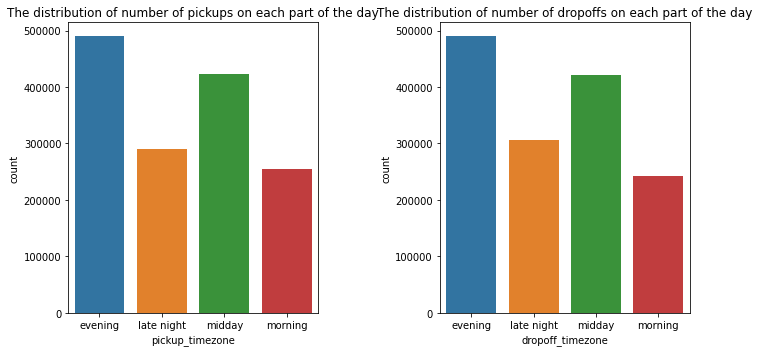




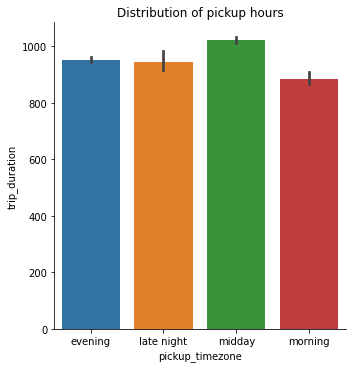


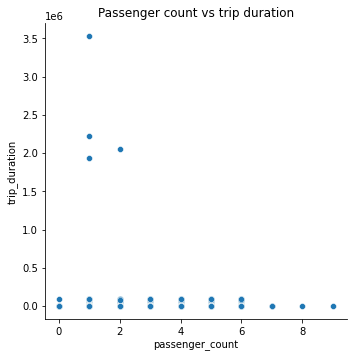


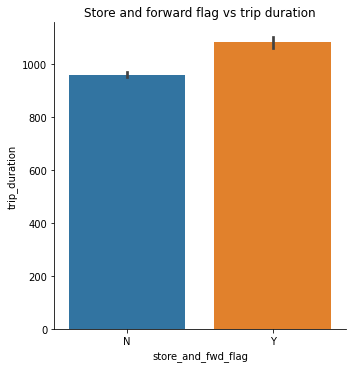


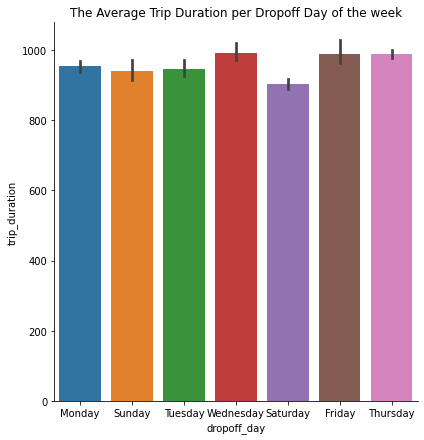


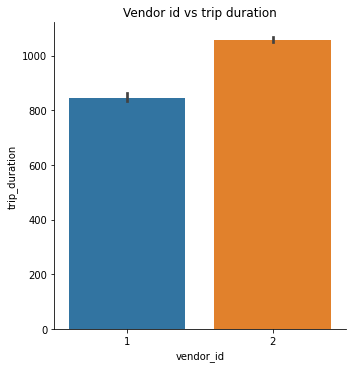
Bivariate analysis of variables











**Analysis**

* Passenger count and store forward flag does not show any trend with respect to trip duration.
* Trip duration has extreme outliers.
* Most trips are taken on Fridays and the least on Mondays.
* Higher number of trips are taken in the evenings and midday.
* Vendor 2 provides more trips than vendor 1.

**Feature Engineering**

* A new feature called distance was added by using the Haversine formula with inputs being the pickup and dropoff coordinates.
* Month, time and the day of the week were extracted from datetimes.
* One hot encoding of categorical variables were performed.
* Outliers in distance and trip duration were handled by deleting rows with zeros and deleting rows which crossed a threshold.

**Training of the model**

* The entire dataset is split into two sets, train and test, by a 80:20 ratio.
* Three modeling libraries were used namely, LinearRegression, XGB Regressor and Ridge regression.
* Hyperparameter tuning: Grid search CV is used to optimize the neg\_mean\_squared\_error for the Ridge regression.

**Evaluation of the models**

* Linear regression model predicted with an RMSE of 275.83 on the test data.
* XGB Regressor predicted with an RMSE of 262.45 and Ridge regression with a value of 275.83.

**Conclusion**

The lowest RMSE obtained using the above mentioned three models is 260 seconds which is negligible compared to our range of nearly 2000 seconds. Therefore, we can confidently use this model for predictions which can be used for improving operations.