**NYC Taxi Time Trip Prediction**

**Ritik Vaidande**

**Data Science Practitioner**

**AlmaBetter, Bangalore**

**Abstract:**

A typical taxi company faces a common problem of efficiently assigning the cabs to passengers so that the service is smooth and hassle free. One of main issue is determining the duration of the current trip so it can predict when the cab will be free for the next trip.

Machine learning has been of significant help as it has helped businesses in abundant ways. We will use Machine Learning to efficiently build a model with a real-world dataset of Yellow Taxi

Service of NYC which will predict the estimated time duration of a tax trip for a given Pick up location, Drop location, Date, and Time.

1. **Problem Statement**

Our task is to build a model that predicts the total ride duration of taxi trips in New York City. Our 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.

1. **NYC Taxi Trip Time Prediction Dataset Insight**

* The dataset contains 1458644 observations with 11 features. It contains features to find out
* location pickup and drop off points, timestamp of trip events, passenger traveled with, necessary
* metrics to make predictions and draw conclusions. Let us look through our features,
* • Id: a unique identifier for each trip
* • vendor\_id: code representing the provider associated with trip
* • 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\_latitude: the latitude where meter was engaged
* • pickup\_longitude: the latitude where meter was engaged
* • dropoff\_latitude: the latitude where meter was disengaged
* • dropoff\_longitude: the latitude where meter was disengaged
* • store\_and\_fwd\_flag: 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 trip in seconds (target variable)

1. **Steps involved**

* **Data Overview**

• id is nominal as well as vendor\_id, but we will check if vendor\_id is of any use as a feature as it contains only values i.e., 1, 2

• pickup/dropoff\_latitude and pickup\_dropoff\_longitude have represented a coordinate

• store\_and\_fwd\_flag is categorical column with Y & N as values

• There is pickup/dropoff\_datetime column which we’ve to convert in datetime format

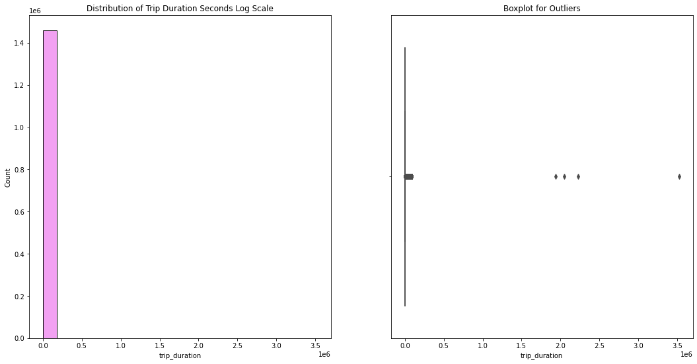
* **Cleaning the Dataset**

There are no missing or duplicated values present in our dataset.

* **Exploratory Data Analysis**

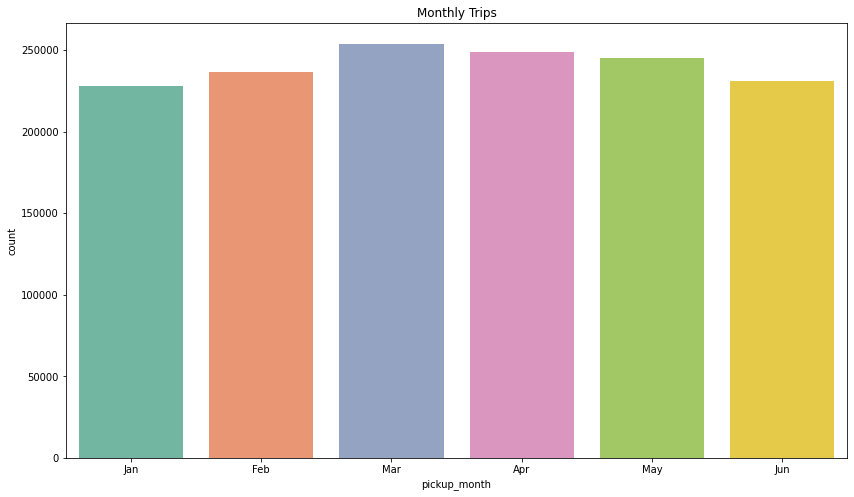
**1. Target Feature analysis**

We begin our EDA by first checking the distribution of our dependent variable i.e. trip duration. We observed that the data is highly positively skewed. We also plotted the box plot and observed that there are many outliers present in the variable.



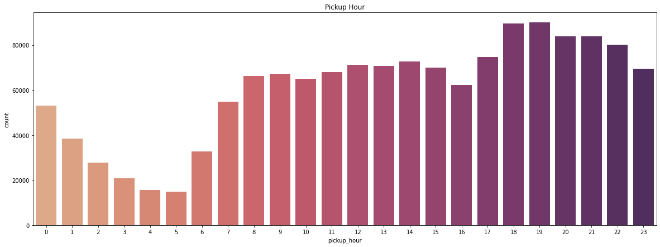
Distribution of Trip Duration and Box Plot for Outliers

**2. Monthly Trips analysis:** Month March crosses the 25k mark with the most number of trips in first-half of the year as it’s the busiest month in the first half of the year while.



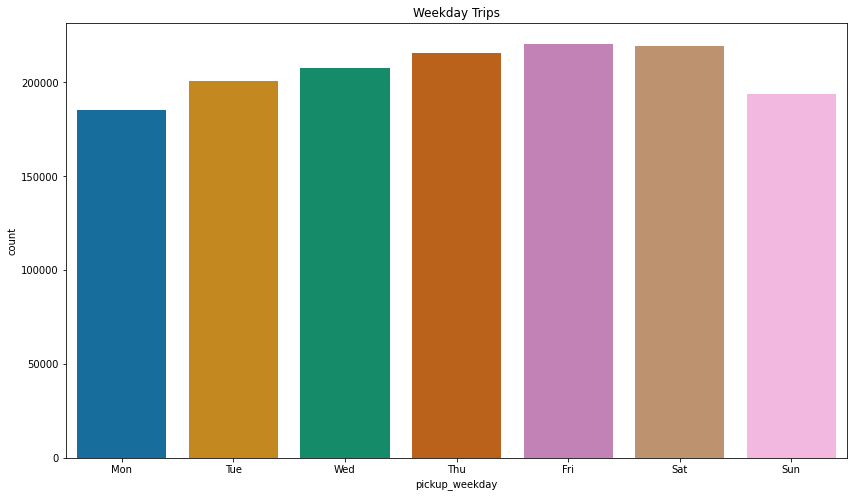
Monthly Trips

**3. Pickup Hour analysis:** The Rush Hours in NYC are around 6pm to 7pm usually most of the offices at till 5pm. While around 3am to 5am is no rush on the streets this shows People prefer taking an taxi more after daylight.



Pickup Hour

1. **Weekday trips analysis:** People in NYC go mostly out on weekends as on Friday and Saturday the trips percentage is high this shows that New Yorker’s prefer going out on weekends. While on Monday the trips are less compared to other days.

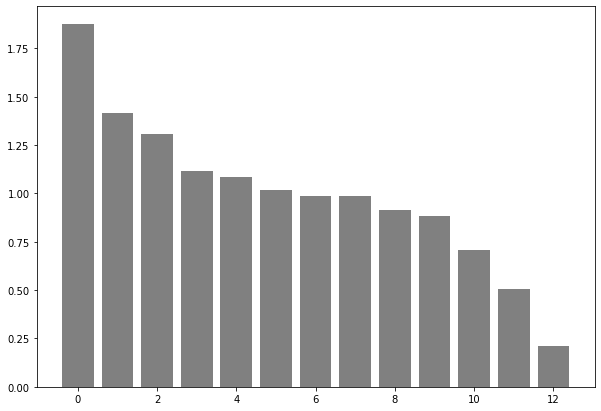


Weekday trips

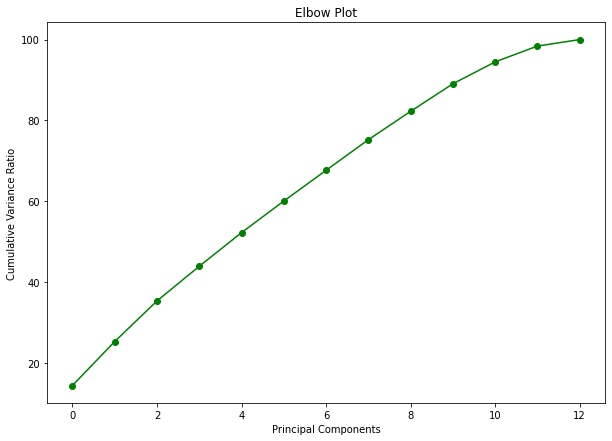
* **Modelling**

**Feature Preparation**

PCA is an Dimensionality Reduction Technique. It is also a Feature extraction Technique. By PCA we create new features from old (Original) Features but the new features will always be independent of each other. So, its not just Dimensionality Reduction Process, we are even eliminating Correlation between the Variables.



At 12th component our PCA model seems to go Flat without explaining much of a Variance.

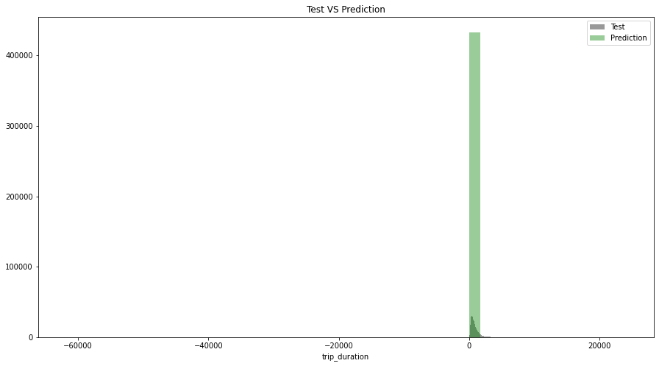
****

By looking at the Elbow plot, 12 is likely to be the required number of components.

**Training**

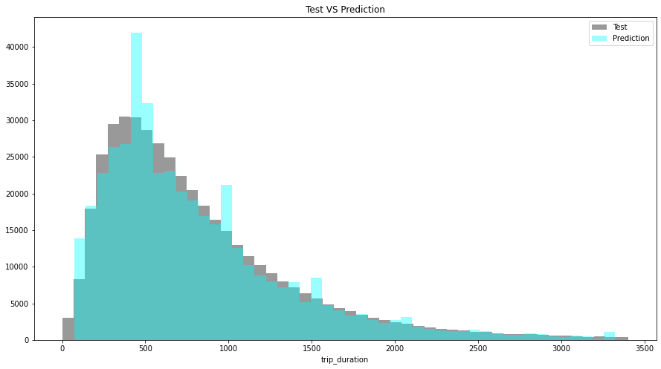
Total 5 Machine Learning Regression Models are implemented.

Linear Regression: Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.



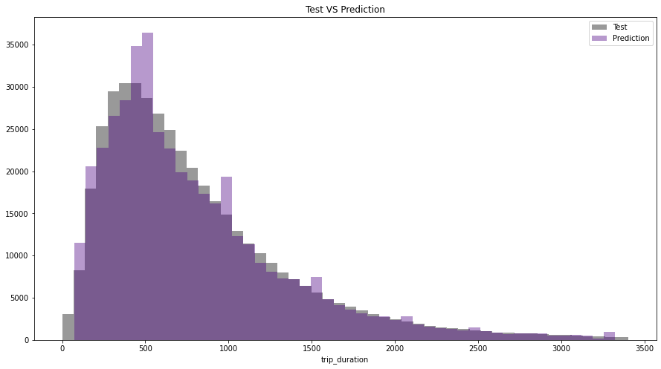
By looking at the test vs prediction graph of linear regression we can see that linear regression is not working well thus linear regression is not suited for this data.

Decision Trees: Decision Tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.



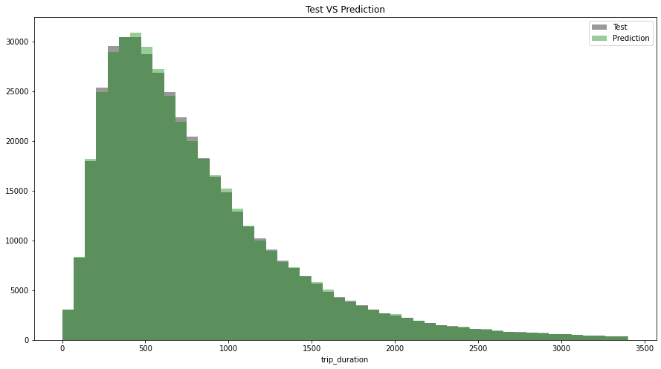
Decision Tree has performed well has compared to Linear Regression.

Random Forest: Random Forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.



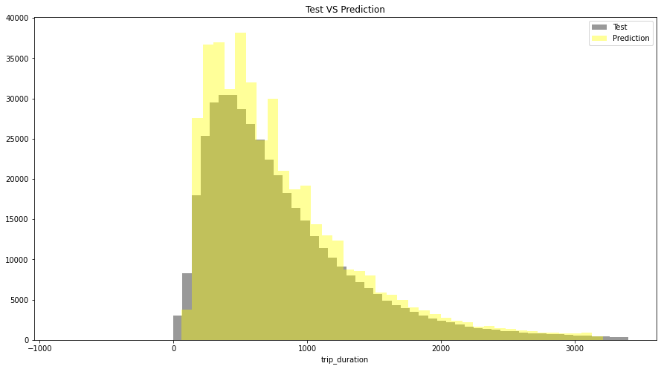
Random Forest has performed slightly better than Decision Trees.

Extra Trees Regressor: Extra Trees is an ensemble machine learning algorithm that combines the predictions from many decision trees. It is related to the widely used random forest algorithm. It can often achieve as-good or better performance than the random forest algorithm, although it uses a simpler algorithm to construct the decision trees used as members of the ensemble.



Extra Tress Regressor appears to be the optimal model.

XGBoost: XGBoost is a powerful approach for building supervised regression models. The validity of this statement can be inferred by knowing about its (XGBoost) objective function and base learners. The objective function contains loss function and a regularization term. It tells about the difference between actual values and predicted values, i.e how far the model results are from the real values.



XGBoost seems slightly less effective than Tree Based Models.

**6. Conclusion**

* Mostly 1 or 2 passengers avail the cab. The instance of large group of people travelling together is rare.
* Most trips were taken on Friday and Monday being the least.
* Fridays and Saturdays are those days in a week when peoples prefer to roam in the city.
* The highest average time taken to complete a trip are for trips started in between 2 pm to 5 pm and the least are the ones taken between 5 am to 7 am.
* Linear Regression doesn’t work well on this data.
* The optimal model is Extra Trees Regressor.

**References**

1. Towards Data Science
2. StackOverflow
3. Medium
4. GitHub