**PROJECT REPORT**

**TAXI FARE PREDICTION**

**by**

Somashree Sahoo



## Introduction

#### Taxi is an indispensable component of public transportation system in dense metropolitan areas such as New York City (NYC). All types of taxis are licensed by the New York City Taxi and Limousine Commission (TLC), which oversees for-hire vehicles, taxis. The iconic taxicabs come in two colors the apple green taxis and the yellow taxis. According to data in the Census Transportation Planning Package (CTPP), commuting trips by taxi account for 2% of all commuting trips by public transportation within NYC. According to a report published by the Taxi and Limousine Commission (TLC) in 2006, NYC taxi cabs carry 11% of passengers who travel in modes of public transportation (Schaller Consulting, 2006). The most recent TLC report of 2016 states that yellow and green taxis together carry about 474,000 trips per day (NYCTLC, 2016). This statistic can be roughly validated based on data from other sources: the population of NYC is about 8.5 million, the daily person trip rate is about 2.5 trips/person for NYC residents, and about 3% of all-purpose trips use taxi or other similar shared-ride service, which gives 0.64 million taxi trips per day (PSB, 2017, U.S. Census Bureau, 2018). The significance of taxi for NYC is also revealed by the number of active taxi drivers (over 50,000) and the millions of taxi passengers they serve per year (NYCTLC, 2016).



***Abstract:***

In the last few years, the number of for-hire vehicles operating in NY has grown from 63,000 to more than 100,000. However, while the number of trips in app-based vehicles has increased from 6 million to 17 million a year, taxi trips have fallen from 11 million to 8.5 million. Hence, the NY Yellow Cab organization decided to become more data centric. Then we have apps like Uber, OLA, Lyft, Gett, etc. how do these apps work? After all, that set price is not a random guess.

#### *Problem Statement*:

Given pickup and drop-off locations, the pickup timestamp, and the passenger count, the objective is to predict the fare of the taxi ride using Random Forest.

#### *Scope:*

● Prepare and analyze data

● Perform feature engineering wherever applicable

● Check the distribution of key numerical variables

● Training a Random Forest model with data and check its performance

● Perform hyperparameter tuning

#### *Dataset Information:*

## unique\_id = A unique identifier or key for each record in the dataset

## date\_time\_of\_pickup = The time when the ride started

## longitude\_of\_pickup = Longitude of the taxi ride pickup point

## latitude\_of\_pickup = Latitude of the taxi ride pickup point

## longitude\_\_of\_dropoff =Longitude of the taxi ride drop-off point

## latitude\_of\_dropoff =Latitude of the taxi ride drop-off point

## no\_of\_passenger =count of the passengers during the ride

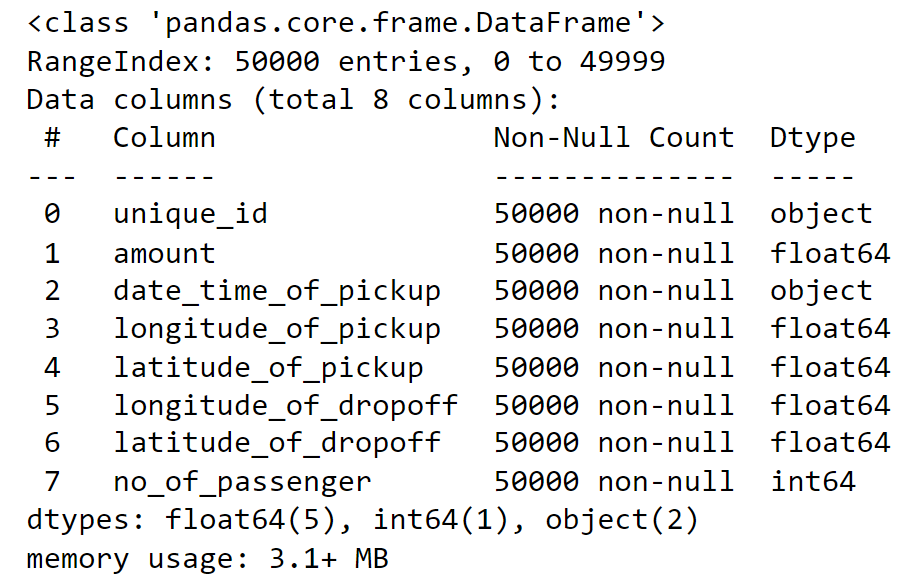
## amount = (target variable) dollar amount of the cost of the taxi ride

Objective:

Machine learning is applied to predict the fare amount. We have used **Random Forest Regressor** to analyze the data.

**EDA and Business Implication**

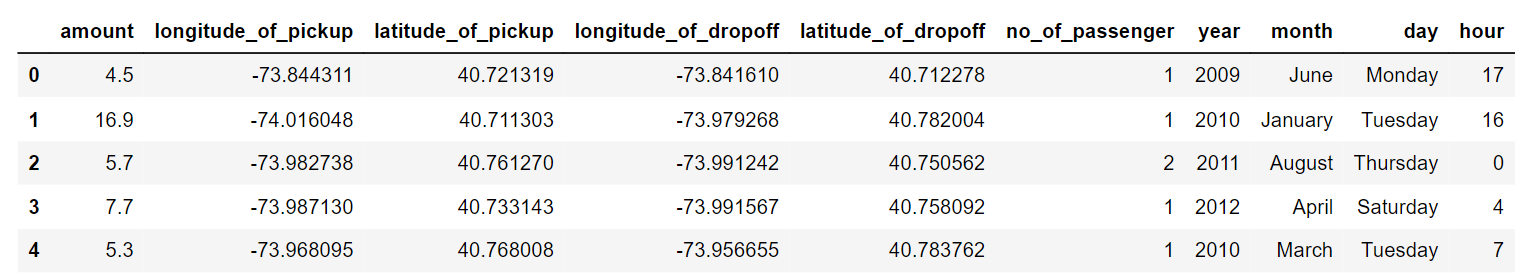
For data description and summary:



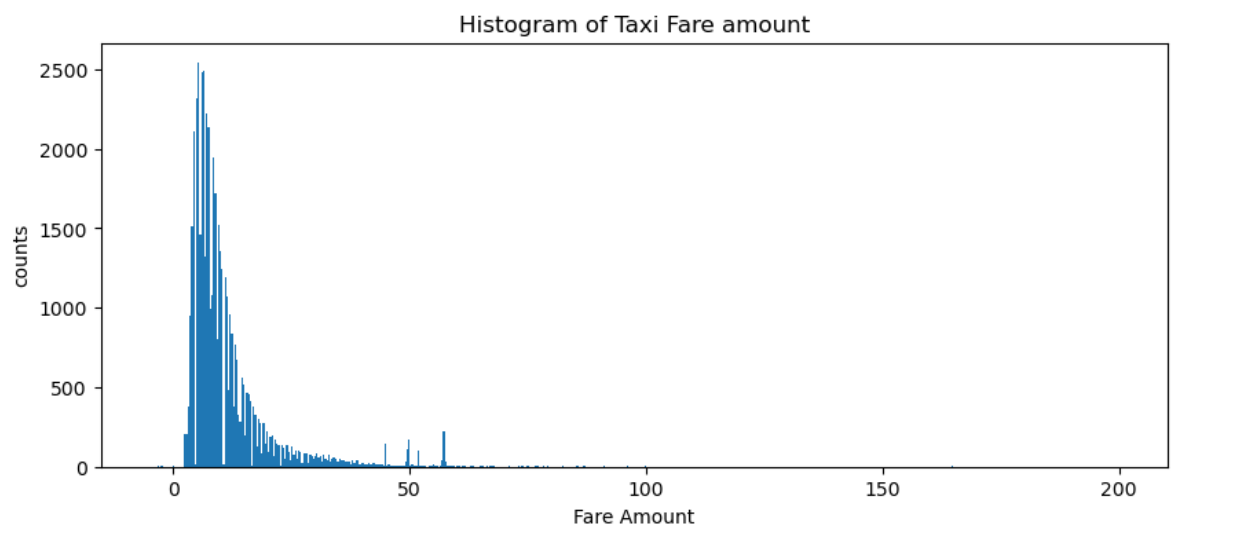
The data types are appropriately listed. It contains information of Number of passengers, Timestamp, Latitude and Longitude for Pickups and Drop-offs and the ***Fare Amount*** which is our ***Target variable***. Number of passengers is ordinal data. Unique id and Date time of pickup are object data types and the rest are numerical float data types. We could also observe that there are no Null values. There are no unwanted punctuation marks, spaces, prefixes or suffixes. Also, the dataset contains 50000 Rows and 8 columns. The existing unique id might not be necessary so we are dropping the column.

As the date time of pickup column contains object datatype but data contains in *date and time format* so will change the datatypes to date-time format and also separate it and create year, month, day and hour columns.

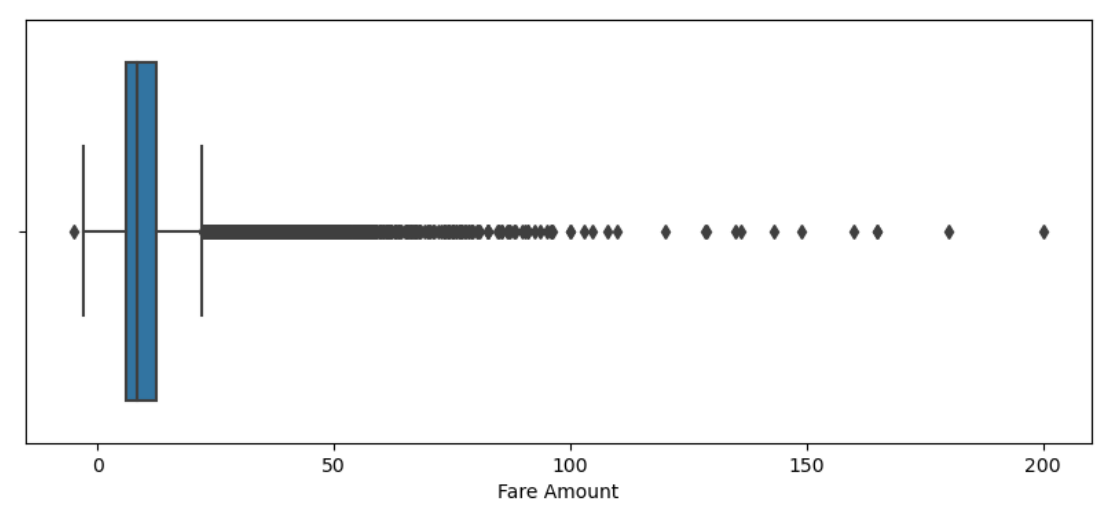
***Printing top 5 Rows of the Dataset****:*



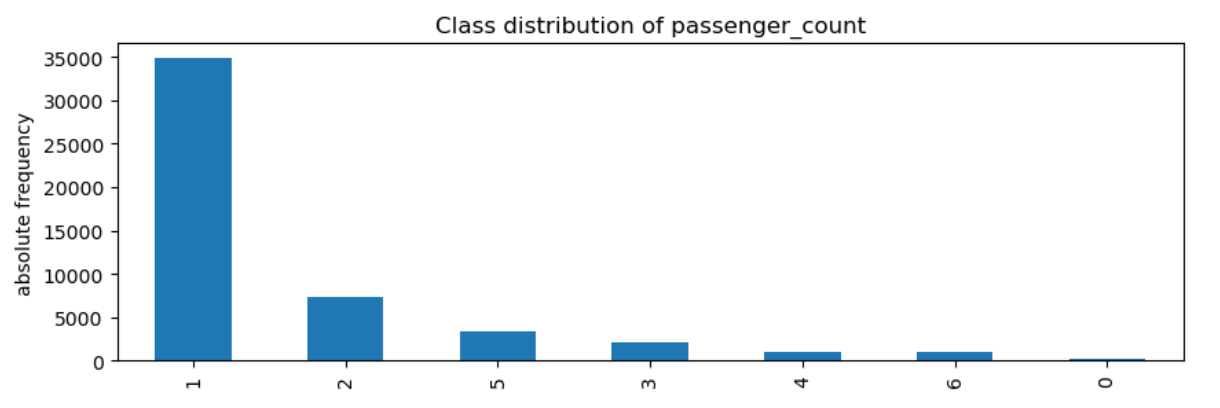
**Visualization:**



From the histogram we observe that the maximum taxi fare is between 0 to 50 and the highest paid amount for taxi trip is 200 which is the outliers and shown below

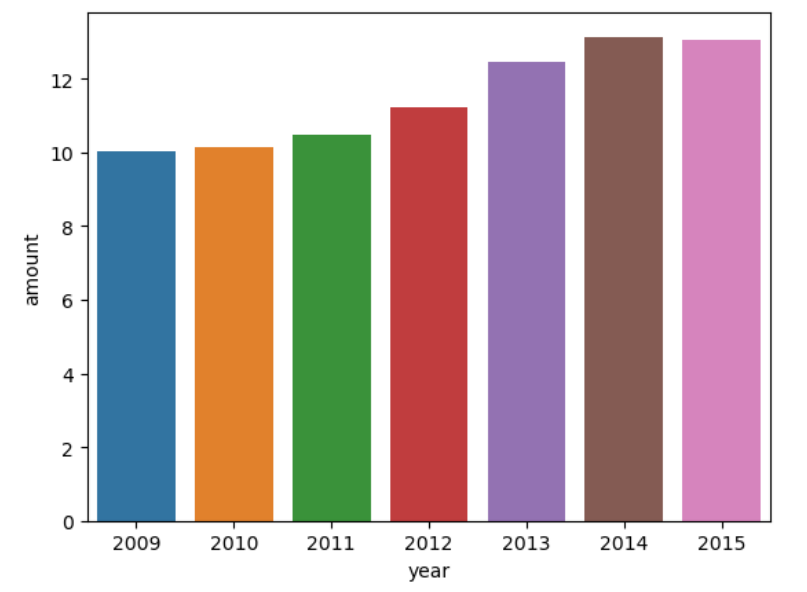


* As fare amount is our target variable so we will not treat the outliers.



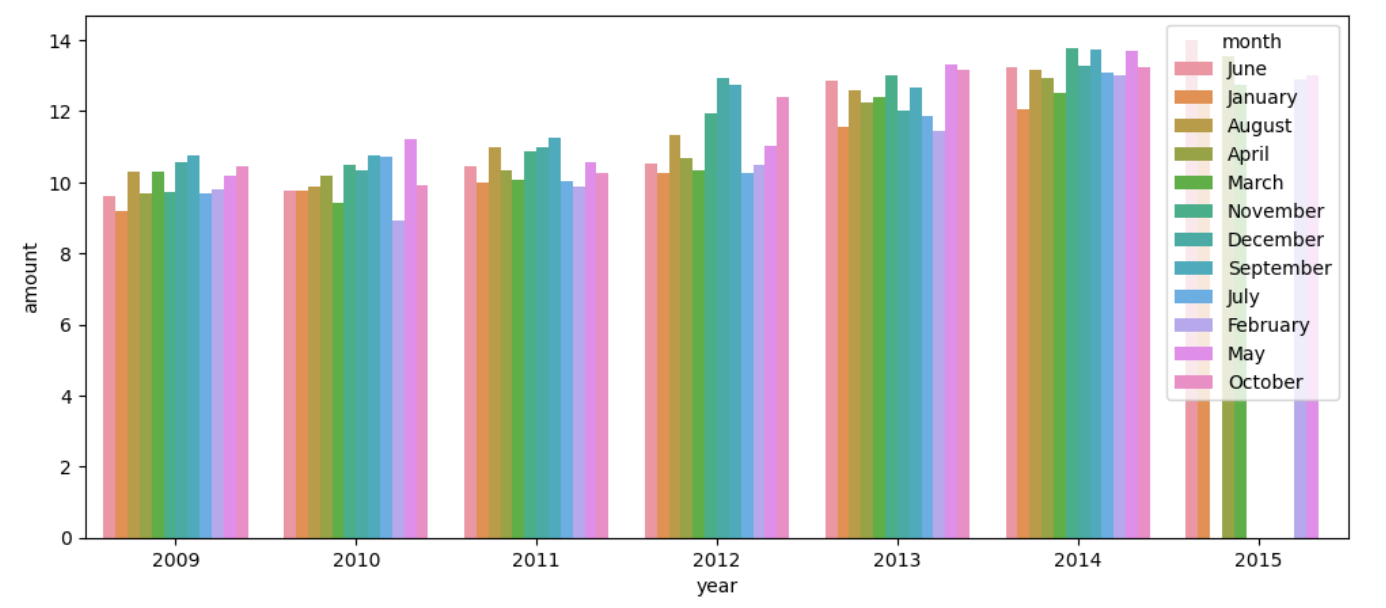
Here we can see that based on the analysis of taxi booking data in New York City (NYC), it has been observed that the majority of people prefer to travel alone, and the number of people traveling in groups is relatively low. The majority of people prefer to travel alone when booking taxis in NYC due to the **convenience and flexibility it offers**. Many people who travel alone do so because they want to have control over their travel itinerary, including the route, the pace of travel, and the stops along the way.

The **cost factor** also plays a significant role in the low number of people traveling in groups as the cost per person increases when traveling in a group. Furthermore, the availability of other transportation options such as public transport, ride-hailing services, and car-sharing services also contributes to the low number of people traveling in groups. Taxi companies in NYC need to be aware of these preferences and cater to the needs of both solo travelers and groups to remain competitive and attract customers.



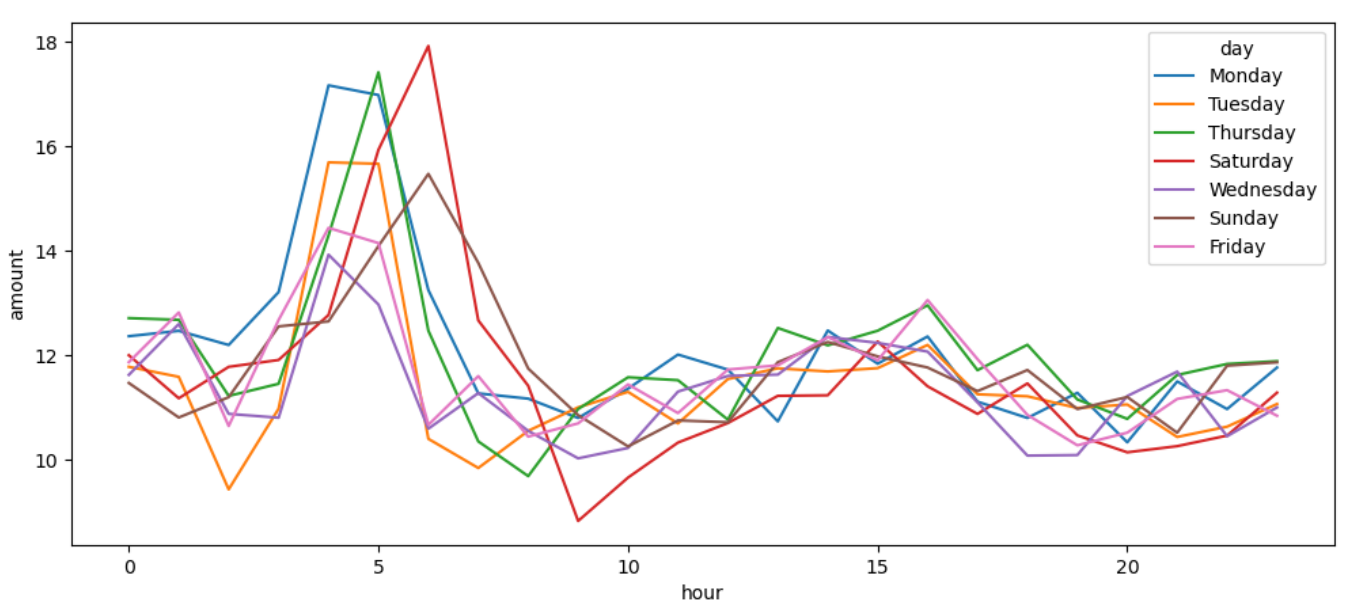
Based on the analysis of taxi fare data in New York City (NYC) from 2009 to 2015, it has been observed that the taxi fare amount gradually increased from 2009 to 2014, but in 2015, the fare amount slightly decreased. In this report, we will explore the reasons for this trend and provide insights into the factors that influence the fare amount in NYC.

The taxi fare amount in NYC gradually increased from 2009 to 2014 due to various factors such as rising fuel costs and inflation. However, in 2015, the fare amount slightly decreased due to the drop in fuel prices and the rise of ride-hailing services such as Uber and Lyft. The TLC regularly reviews and adjusts the taxi fare structure to ensure that it is in line with the cost of living and reflects the changing economic conditions. Taxi companies in NYC need to be aware of these factors and adjust their fare rates accordingly to remain competitive and attract customers.



Based on the analysis of taxi booking data in New York City (NYC) over the past several years, it has been observed that the months of **September, November, and December** are the peak months and **may, June, October** are also other high-demand months for taxi bookings. In this report, we will explore the reasons for this trend specifically in the context of NYC and provide insights into the patterns of taxi bookings during these months.

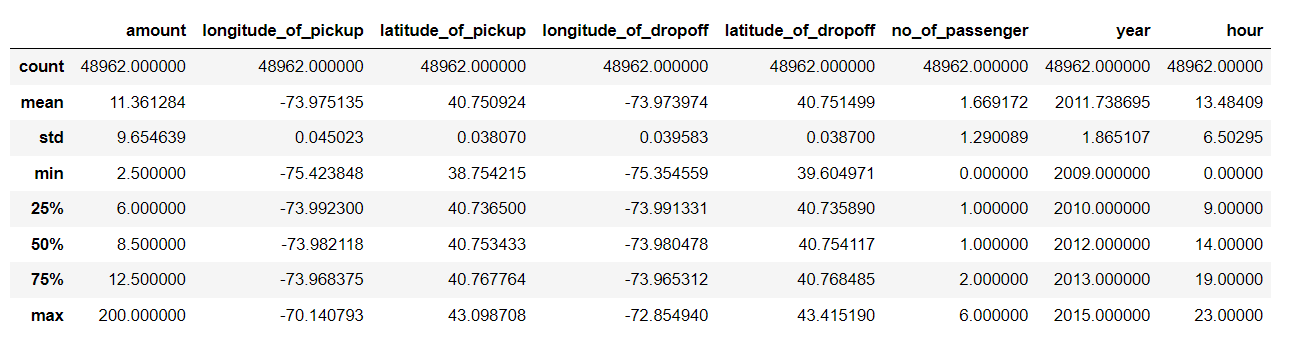
These months are peak months for taxi bookings in NYC due to various factors such as favorable **weather conditions, major festivals and events, and the start of the academic year**. The demand for taxi services is highest during weekends, public holidays, and evening hours. This information can be useful for taxi companies in NYC to optimize their services and resources during these peak months to ensure customer satisfaction and maximize revenue.



Based on the analysis of taxi booking data in New York City (NYC), it has been observed that the majority of people book taxis to travel on Mondays, Thursdays, and Saturdays, and during the time slots of 4:00 AM to 8:00 AM and 1:00 PM to 3:00 PM.

The majority of people book taxis to travel on Mondays, Thursdays, and Saturdays, and during the time slots of 4:00 AM to 8:00 AM and 1:00 PM to 3:00 PM in NYC. These trends are primarily driven by work and travel patterns, leisure activities, and coincide with rush hour and lunch hour periods. Taxi services in NYC need to be aware of these preferences and adjust their operations and availability accordingly to meet the needs of their customers.

**Data Summary:**

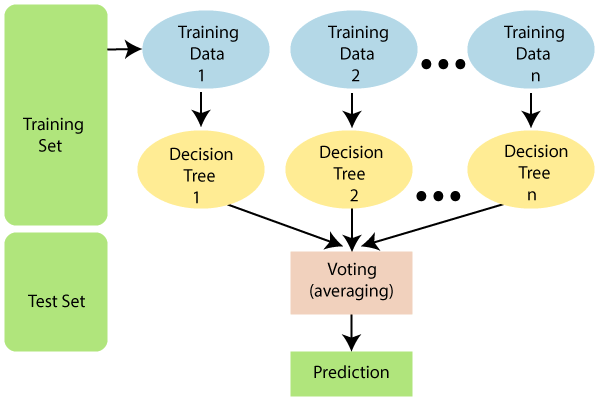
****

The minimum fare amount of 2.50 is the base fare set by the TLC for taxis in NYC. This means that even if a passenger takes a very short ride, they will be charged at least 2.50 as the base fare. It is set to ensure that drivers are able to make a minimum profit, even for very short rides. The maximum fare amount of 200.0 may be for specific types of rides or during certain circumstances, and may be subject to a fare cap.

The longitude of pickup and drop-off locations of taxi rides in NYC falls within the range of -78 to -70, while the latitude falls within the range of 37 to 45. This information can be used by taxi services to identify areas of high demand and optimize their operations to better serve their customers.

Based on the analysis of taxi trips in New York City (NYC), it has been observed that the minimum count of passengers per trip is 1, while the maximum count is 6. The count of passengers per taxi trip in NYC falls within the range of 1 to 6. This information can be used by taxi services to optimize their operations and better serve their customers.

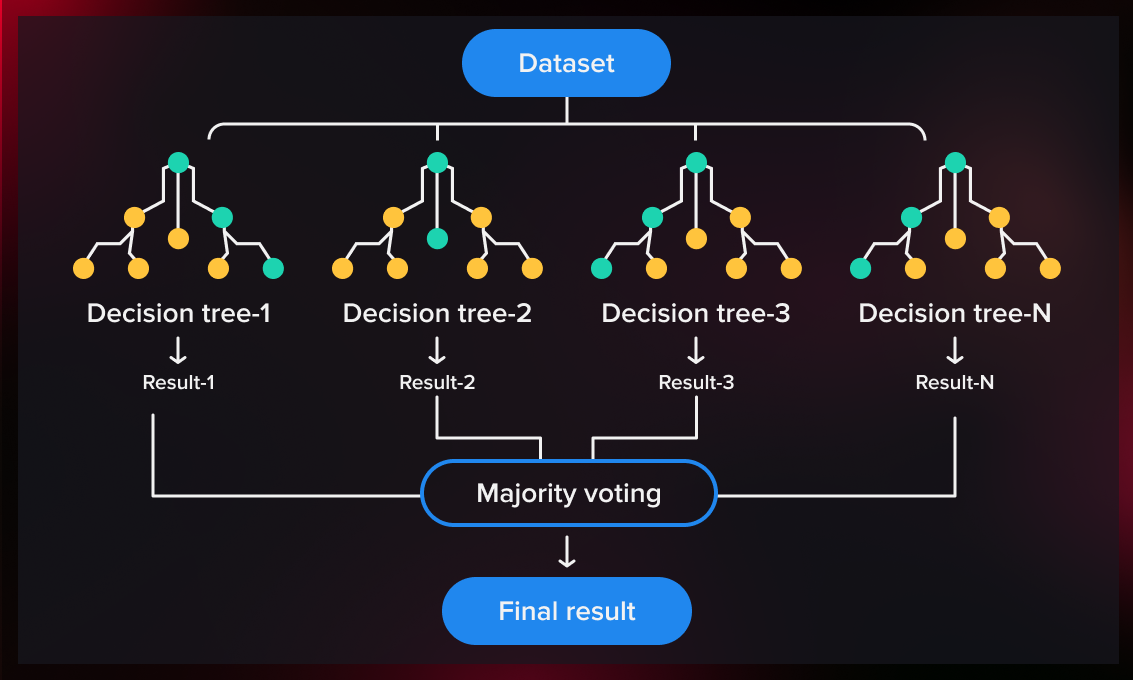
**MODEL BUILDING:**

****

The objective of the Random Forest Regressor model is to predict a continuous numerical value, such as an amount, longitude of pickup and drop off, or latitude of pickup and drop off, based on a set of input features. The model aims to learn a mapping between the input features and the output variable by constructing a collection of decision trees, where each tree provides a prediction for a given input instance.

The goal of the model is to produce accurate and reliable predictions on new and unseen data, while also generalizing well to different input patterns and distributions. To achieve this objective, the model employs an ensemble learning approach that combines the predictions of multiple decision trees, each trained on a different subset of the input features and training data.

**Random Forest Regressor:**

****

The Random Forest Regressor model is a powerful and flexible algorithm that can handle a wide range of input data types and distributions. It is commonly used in applications such as stock price prediction, sales forecasting, and weather modeling, where accurate numerical predictions are critical for decision-making.

The model's performance is evaluated using standard regression metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). These metrics quantify the difference between the predicted output and the actual output, and a lower value indicates better performance.

Overall, the objective of the Random Forest Regressor model is to learn a mapping between the input features and the output variable that can accurately and reliably predict the output value for new and unseen data.

Here is a step-by-step process for implementing Random Forest Regressor:

We Import the necessary libraries such as NumPy, pandas, and scikit-learn packages and load the dataset and preprocess the dataset, including handling missing data and converting categorical data into numerical data.

Further steps include:

1. Split the dataset into training and testing sets.
2. Create an instance of the Random Forest Regressor class from the scikit-learn library.
3. Train the model using the fit method with the training data.
4. Predict the target variable using the predict method with the testing data.
5. Evaluate the model using metrics such as mean squared error, mean absolute error, and R-squared.
6. Tune hyperparameters of the model such as the number of estimators, maximum depth, and minimum samples split to optimize model performance.

|  |  |  |
| --- | --- | --- |
| Train/Test | Train | Test |
| Accuracy | 0.9684 | 0.8189 |
| MSE | 3.0169 | 15.9852 |
| RMSE | 1.7369 | 3.9981 |

An **accuracy** of 0.9684 on the **training data** indicates that the model correctly predicted **96.84%** of the target variable values in the training dataset. An accuracy of 0.8189 on the testing data indicates that the model correctly predicted **81.89%** of the target variable values in the **testing dataset**.

An **RMSE of 1.7369** on the training data indicates that on average, the model's predictions on the **training d**ataset were off by 1.7369 units. An **RMSE of 3.9981** on the **testing data** indicates that on average, the model's predictions on the testing dataset were off by 3.9981 units.

Overall, these metrics suggest that the ***model is performing well*** on the training data but may ***be overfitting since its performance drops on the testing data***. Therefore, it may be necessary to adjust the hyperparameters of the model or apply ***regularization techniques*** to improve its generalization performance.

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |

**GridSearchCV:Top of Form**

Here I used GridSearchCV to tune the hyperparameters of Random Forest Regressor model to improve its generalization performance.

The hyperparameters that I selected through GridSearchCV are-

'Max\_depth': 26

‘Min\_samples\_leaf': 16

'Min\_samples\_split': 50

'n\_estimators': 250.

These hyperparameters specify the maximum depth of the decision trees in the random forest, the minimum number of samples required to be at a leaf node, the minimum number of samples required to split an internal node, and the number of trees in the forest**.**

|  |  |  |
| --- | --- | --- |
| Train/Test | Train | Test |
| Accuracy | 0.7959 | 0.7926 |
| MSE | 19.50 | 18.30 |
| RMSE | 4.416 | 4.278 |
|  |  |  |

**Model Performance:**

The model's performance improved after applying regularization with GridSearchCV, achieving an accuracy score of 79% on both the training and testing datasets. This suggests that the model is no longer overfitting to the training data and is generalizing better to new, unseen data**.**

The RMSE values of 4.416 for the training dataset and 4.278 for the testing dataset suggest that the model has an error of around 4 dollars in its predictions. This means that on average, the predicted fare amount could be off by around 4 dollars from the actual fare amount. While this error may seem small, it could still be significant in some cases, especially for short trips where the fare amount is low. Overall, the model seems to be performing well but could potentially be further optimized to reduce the error in its predictions.

**Conclusion:**

The Random Forest model has shown good performance in predicting taxi fares based on pickup and drop-off locations, pickup timestamp, and passenger count. The accuracy score of 79% and low RMSE values suggest that the model can be relied upon to make fare predictions.

Therefore, the NYC Taxi Committee could consider implementing this model in their fare prediction system to improve accuracy and efficiency. Additionally, they could further fine-tune the model by incorporating more relevant features such as traffic congestion, weather conditions, and special events. Furthermore, the model can also be used to optimize taxi routes, reduce travel time and fuel consumption and improve the overall efficiency of the taxi service.



**Recommendations:**

Moreover, it may be worth considering **collaborations with app-based ride-sharing** services to expand their reach and offer a more comprehensive range of services to customers.

*They could consider* ***developing a mobile app*** *that allows passengers to easily request and pay for rides, track their driver's location in real-time, and provide feedback on their experience.*

Another recommendation could be to improve the quality of service by investing in ***driver training programs*** and implementing quality control measures to ensure that drivers provide safe and efficient rides to their customers. They could also consider offering incentives or rewards to drivers who consistently provide high-quality service.

Finally, they could also consider ***offering competitive pricing and flexible payment options to attract customers.*** By offering different pricing tiers and payment options, such as the ability to pay with cash or credit card, they could appeal to a wider range of customers and potentially increase their ridership.

This can not only help in increasing revenue but also foster healthy competition and innovation within the industry.

**Thank you**

Bottom of Form