**NYC Taxi Trip Time Prediction**

**Vinayak Marathe, Riya Patel,**

**Muskan Kasere**

**Abstract - New York City taxi rides form the core of the traffic in the city of New York. The many rides taken every day by New Yorkers in the busy city can give us a great idea of traffic times, road blockages, and so on. In New York City many of people commute to different regions of city via taxi. A lot of streets and roads in New York City are quite busy due to traffic jams. Predicting the duration of a taxi trip is very important since a user would always like to know precisely how much time it would require of him to travel from one place to another. Here, we are prediction trip duration of taxi so that the user will know how much time it will take to commute from one place to other and also there are app-based taxi such as uber, ola etc. Decisions has to be taken by the user for opting which one to choose based on trip duration.**

**Keywords: Prediction algorithms, Linear Regression, Ridge and Lasso Regression, Random Forest Regressor, mse, mae, rmse, r2, adjusted r2**

1. **INTRODUCTION**

The NYC taxi trip time prediction project involves building a machine learning model to predict the time it takes fora taxi to travel from one location to another in New York City. The model contains independent and dependent variables or features which includes pickup time, drop-off time, geo- coordinates, and number of passengers, trip duration, and several other variables which we have derived in our project using the given information. In our project, to build the model, a large dataset of historical taxi trip records is used to train the model. Various machine learning algorithms, such as Linear Regression, Regularization Techniques and Random Forest, can be applied to the data to build a predictive model.

Once the model is trained, it can be used to predict the trip time for new, unseen taxi trips in NYC. This information can be useful for taxi drivers and passengers, as well as for transportation planning and optimization. The results of the project can also be used to understand the factors that influence taxi trip times in NYC, such as traffic patterns, weather conditions, and time of day. This information can be used to make improvements to the city's transportation infrastructure and to develop more efficient transportation systems.

In order to explain the nature of our data, the data is collected and initialized in the first stage then we have checked our data with the shape, size and type. Then at the second stage, we are focusing on our variable and understanding the behaviour of our variables. On the next stage, we have performed data wrangling to identify some pattern of our data with different Exploratory Data Analysis (EDA) techniques, accompanied with the visualizations of the New York City’s map, storytelling and experimenting with the different charts. Again, at the fourth stage, based on our chart experiments, Hypothesis Testing is performed. We defined three hypothetical statements from the dataset. And in that three questions, we obtain final conclusion about the statements through the code and statistical testing. And then at the next stage, we have done some feature manipulation and found most important feature for our dataset i.e. trip distance as well as we have also done some feature engineering to encode our categorical column and to detect the outliers. In the last and final stage of the project, we have implement our model using different type of regression algorithms which further evaluated using different evaluation matrices in order to check the accuracy of our model and then get the report back with the maximum accurate model. In the end, we have come up with the achieved business insights and conclusion of our project.

1. **Problem Statement**

The problem statement for the NYC taxi trip time prediction project is to accurately predict the time it takes for a taxi to travel from one location to another in New York City. The objective is to develop a machine learning model that takes into account various factors. The challenge lies in capturing the complex relationships between the various factors that influence taxi trip times and accurately predicting the trip duration for any given trip.

The solution to this problem will have practical applications for taxi drivers and passengers, as well as for transportation planning and optimization. Accurate taxi trip time predictions can help drivers plan their routes more effectively and reduce the time and cost of travel for passengers. It can also be used to improve the city's transportation infrastructure and to develop more efficient transportation systems.

1. **METHODOLOGY**

The proposed methodology’s implementation begins with downloading the dataset. Then data wrangling and feature manipulation is executed as a step of pre-processing of data. After this, the data is analysed and different model is executed. Then we have done the hypothesis testing with 3 different hypothetical statement. At last, all the business insights carried out in this project.

**A. Datasets**

All the information of NYC Taxi Trip are here which we use to predict the taxi time duration. The dataset consists of every expertise that is helpful and attributes that are no longer useful. So, in pre-processing the beneficial statistics is chosen and statistics cleansing is performed to get the best possible attributes. Let’s understands our variables:

* Id – A unique identifier for each trip
* Vendor-id – A code indicating the provider with the trip rec
* 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
* 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 – 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\_duration – duration of the trip in seconds.

**B. Pre-Processing**

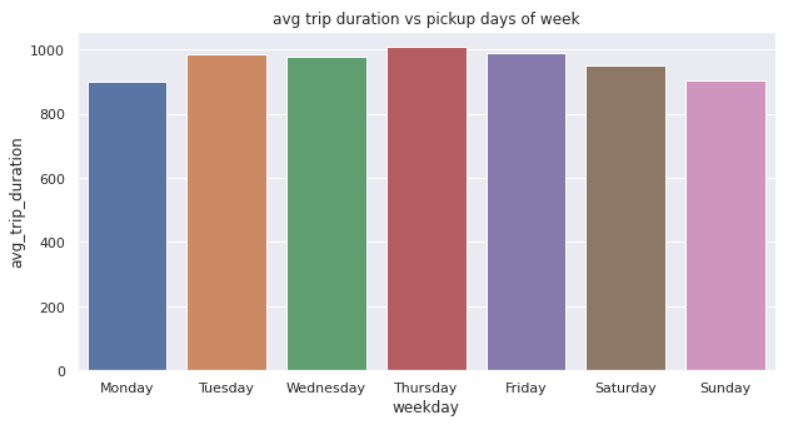
As at this pre-processing level, this is the necessary step; significant data is derived from the dataset of NYC Taxi Trip. This section is compulsory because the raw data is now not reliable and unfinished, so pre-processing is carried out for greater steps to render geared up raw data. In this technique at some stage in pre-processing, 11 attributes are used to apprehend the area of New York City.

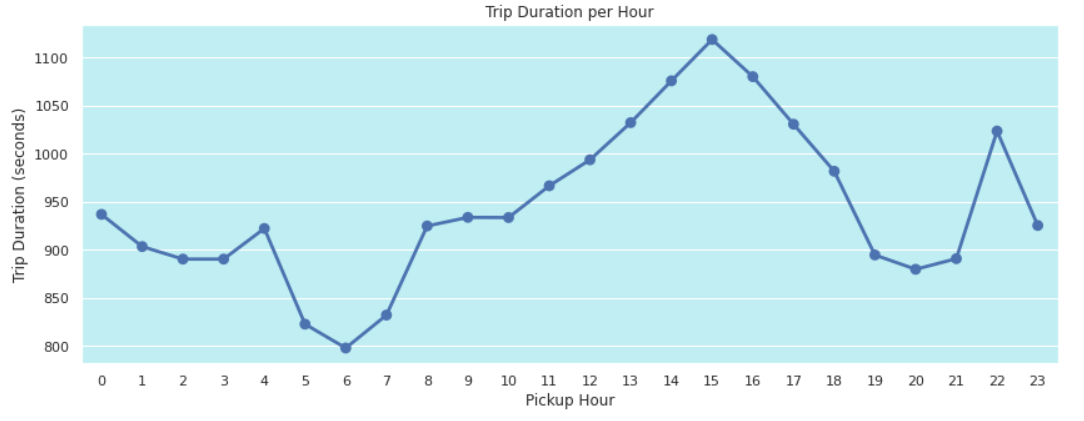
These 11 attributes encompass id, vendor id, pickup datetime, passenger count, drop-off longitude, trip duration and many more. The attribute’s values are normalized and descriptive statistical analysis is done.

**C. Data Wrangling & Visualization**

The quality of data performs a fundamental role, and the most cautiously depicted trouble to be. For this research, our data is already clean with zero null values. And then we performed some usual statistics and pre-processing and found some useful insights. Next, we have visualized our facts graphically (as seen in the fig.1 below) as the dataset is in tabular shape and it is hard to appear at and understand the data in this or any other way. Data visualization helps in grasping the style of the data. Data visualization in this method is a graphical illustration of the data. In this analysis, the utilization of bar charts, box plot, dist. plots, scatter plots, point plots, pie charts, histograms and sub plots, the cleaned records obtained thru pre-processing is visualized. It makes it easy to maintain the attribute's tough relationship with the beneficial resource of graphical representation.

As mentioned above, this visualization performs an essential attribute in data exploration. A variety of parameters of the dataset plotted based on the available attributes as viewed in figures below.





**Fig.1 EDA based on attributes**

**D. Hypothesis Testing**

We have defined three hypothetical statements from the dataset. In the next three questions, perform hypothesis testing to obtain final conclusion about the statements through our code and statistical testing.

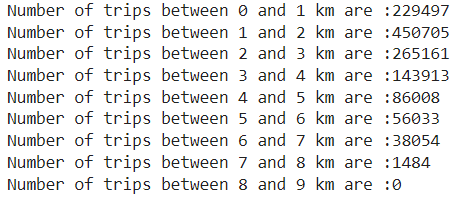
1. There is a significant difference in the trip duration between trips with a single passenger and trips with multiple passengers. Here,
   * **Null hypothesis:** There is no significant difference in the trip duration between trips with a single passenger and trips with multiple passengers.
   * **Alternative hypothesis:** There is a significant difference in the trip duration between trips with a single passenger and trips with multiple passengers.
   * **Result:** The p-value of 9.89e-28 signifies that the probability of observing such extreme differences in trip duration between single passenger trips and multi-passenger trips by chance is very low. Therefore, we can reject the null hypothesis and conclude that there is a significant difference in trip duration between these two groups.
2. Trips with a higher distance have a longer duration. Here,
   * **Null Hypothesis:** The distance of the trip does not significantly affect the duration of the trip.
   * **Alternative Hypothesis:** The distance of the trip significantly affects the duration of the trip.
   * **Result:** A p-value of distance variable 0.0 indicates that the probability of obtaining the observed sample result assuming that null hypothesis is true is extremely small (essentially zero). Therefore, we reject Null Hypothesis and conclude that there is a significant relationship between the two variables.
3. Trips during weekdays have a longer duration compared to weekends.

* **Null Hypothesis:** The day of the week does not significantly affect the duration of the trip.
* **Alternative Hypothesis:** The day of the week significantly affects the duration of the trip, with longer trip durations on weekdays compared to weekends.
* **Result:** The t-statistic of 3.9008 indicates that the difference between the means of the two groups (weekday and weekend trip durations) is 3.9008 times greater than the standard error of the difference between the means.

The p-value of 9.586970042130944e-05 (0.00009587) is less than the commonly used significance level of 0.05, which suggests strong evidence against the null hypothesis. Therefore, we reject Null hypothesis and conclude that there is a statistically significant difference between the means of weekday and weekend trip durations.

**E. Feature Engineering & Data Pre-processing**

In the feature engineering, we have done some manipulation with our dataset where we have entitled new and most important column named **trip distance** which shows us the distance between our ride using latitudes and longitudes given. From which we found number of rides in a particular interval of distance which is shown in the fig below:



**Fig. - No. of Rides**

Then**,** we handle outliers with IQR method for skewed data in this the data points that are less than the 25th percentile are replaced with its 25th percentile value and the data points at are greater than the 75th percentile are replaced by 75th percentile value. And then after dropped some data for better performance of the model, where we found some data i.e. Short Trips: Trip duration is less than a minute and Long Trips: Trip duration is greater than 36 minute and 40 seconds. We kept only the fields which are in Normal Trip: Trip duration is between 1 - 36:40 min. By doing this we are removing less than 5% of our data. In that way data will be cleaner and have less outliers. In the end, finally we found out 9 independent features which are important. All the features which we left with having some importance and none of them are 0. So, It validates the features make sense and heading towards the right direction. Important Features which we kept are trip\_distance, dropoff\_latitude, pickup\_latitude, month, pickup\_longitude, dropoff\_longitude, trip\_distance, pickup\_hour, weekday\_num.

**F. ML Model Implementation**

Considering various Machine Learning models that provide reliable and improved accuracy for prediction based use-cases, Linear Regression, Ridge and Lasso Regression and Random Forest are taken into consideration.

* **Linear Regression -** Linear regression is a machine learning algorithm used to predict outcomes in datasets with continuous data. It can be used to predict the duration of taxi trips in New York City by using data from previous rides, such as pickup and dropoff location, time of day, and other factors. Linear regression models will identify relationships between input features and output (trip duration) in order to accurately predict future trip duration.
* **Ridge Regression -** Ridge regression is a type of linear regression used in machine learning. It is used to reduce the effect of collinearity in a regression model by adding a penalty term to the cost function. In New York City taxi trip duration prediction, ridge regression can be used to identify important factors that can impact the duration of a taxi trip such as traffic patterns, distance travelled, time of day, and weather conditions. The algorithm can be used to train the model to predict the duration of a trip given these factors. By incorporating the penalty term, ridge regression can help the model avoid overfitting and therefore more accurately predict the duration of a taxi trip.
* **Lasso Regression -** Lasso regression is a type of linear regression that uses the L1 regularization technique. This technique works by introducing penalization terms in the objective function to shrink some coefficients towards zero, which can effectively reduce the model complexity and improve the generalization ability of the model.

In the context of New York City taxi trip duration prediction, Lasso regression could be used to identify important factors that affect the trip duration, such as traffic conditions, time of day, and pick-up/drop-off locations. By incorporating these factors into the model, Lasso regression can help to improve the accuracy of the predictions and better capture the nuances of the data.

* **Random Forest -** Random Forest regression is a machine learning algorithm that can be used to predict taxi trip duration in New York City. The Random Forest algorithm works by building multiple decision trees and aggregating their results, thus providing more accurate predictions. In this case, the model would use features such as pickup and dropoff location, time of day, and pickup hour to predict taxi trip duration. Additionally, the model could also use taxi-level features such as driver experience, vehicle type, and route taken to enhance its predictive power if it is given. With the right data and optimization techniques, Random Forest regression can provide a powerful and reliable tool for predicting taxi trip durations in New York City.

In this implementation, we find that the importance of each feature is then normalized such that the sum of all feature importance is equal to 1.0. Therefore, a higher feature importance value indicates that the feature is more important for the model's prediction. In our case most important feature is trip distance almost 52 % important. Feature importance’s can be used as an explain ability tool to gain insights into which features are driving the predictions made by the model. By analysing the feature importance scores, we can identify the most important features and understand their relative contribution to the model's overall performance. This information can be used to guide feature selection and engineering efforts to improve the model's accuracy and generalization.

1. **RESULT**

From the four Model, we have chosen Random Forest Regressor Because

* It reduces RMSE from 291.67 to 213.228
* Reduces MAE from 217.92 to 151.78
* Improves R2 from 0.48 to R2 score: 0.7208182963366805
* Improve Adjusted r2 from 0.48 to 0.7208163191991594

We got the 72% of the accuracy using the Random Forest Algorithm.

1. **FUTURE WORK**

Future directions for improving taxi trip time prediction:

1. **Incorporating more data:** Taxi trip time prediction models can be improved by incorporating additional data sources, such as weather conditions, traffic patterns, and events happening in the city. This additional data can help the model better understand the factors that impact trip times.
2. **Using advanced machine learning techniques:** Advanced machine learning techniques, such as deep learning and ensemble learning, can be applied to taxi trip time prediction. These techniques can help capture complex patterns in the data that may be difficult to detect using traditional modelling approaches.
3. **Developing real-time prediction models:** Real-time prediction models can help provide more accurate estimates of trip times as conditions change. These models can be used to update taxi driver and passenger apps in real-time, providing more accurate and up-to-date information on trip times.
4. **Improving the accuracy of pick-up and drop-off locations:** Accurately predicting pick-up and drop-off locations can help improve the accuracy of trip time predictions. This can be accomplished through improved geocoding techniques or the use of more accurate location data.
5. **Incorporating user feedback:** User feedback can help improve the accuracy of trip time predictions over time. This feedback can be used to refine the model and identify areas where improvements can be made.

Overall, improving taxi trip time prediction is an important area of research that can have significant benefits for transportation services in NYC. By incorporating additional data sources, using advanced machine learning techniques, developing real-time prediction models, improving location accuracy, and incorporating user feedback, it may be possible to improve the accuracy of trip time predictions and provide better transportation services for all.

1. **CONCLUSION**

In conclusion, predicting taxi trip time accurately is an important task for optimizing transportation services in NYC. There have been many efforts to improve the accuracy of trip time predictions, including the use of advanced machine learning techniques, incorporating additional data sources, developing real-time prediction models, improving location accuracy, and incorporating user feedback.

Improving the accuracy of taxi trip time predictions has the potential to provide significant benefits:

* For transportation services in NYC
* Including reducing wait times for passengers
* Optimizing driver routes and
* Improving overall transportation efficiency