Project Title

Uber Fare Prediction

TEAM MEMBERS:

Sailesh Pilla (11593815)

Yaswanth Sai Satish Sreerama (11601948)

Kishore Kumar Paila (11600316)

Project Proposal Link: GitHub

Idea Description:

Our Project is about Uber, the largest taxi company in the world, serving thousands of consumers every day. I noticed some open-source data from Kaggle. Due to the service's global reach, it is necessary to precisely compute fare pricing considering several factors, including distance, weather, time, and demand.

Goals and Objectives:

Goals of this project is to predict fare accordingly in different places with respect to different variables by training the model first and then test it on test dataset.

Objectives:

1. Thoroughly understand the dataset and determine whether the data is correct or not, and if not, perform a clean-up.
2. Create a regression model over the training data that has been separated.
3. Evaluate the model and compare R2 and Root mean squared error scores to find the projected model error so that you may understand how accurate your predictions were.

Motivation:

The reason I wanted to know what type of algorithm or model they were using to see what all elements play a role in forecasting price and see which dimension plays a significant function in accurately predicting the fare is that we utilize uber a lot in our daily lives to get to our destinations.

Significance:

This is accomplished by predicting fare quickly whenever a user enters his pickup and destination; it calculates it automatically considering all the factors using a machine learning algorithm. It attracts users when he can see the amount that he needs to pay before booking; it will give user to allocate or plan their budget accordingly and they can choose the type of vehicle accordingly. Here, utilizing ML models increases the likelihood that the prediction will be accurate. If it is incorrect, we can compute the error and improve our model's training by using more accurate data or by omitting irrelevant data. A better prediction can be made with the right training data.

Literature Survey:

We are attempting to comprehend how Uber operates, how the uber fare is computed, and what additional criteria are taken into account to estimate the fare using ML models. Given that there are numerous variables that affect how fare is predicted, we will attempt to use a multi-linear regression.

1. Data Exploration

2. EDA (Explanatory Data Analysis)

3. Data Pre-Processing

4. Feature Selection

5. Splitting of Data

6. Applying Models and Performing Prediction

Features:

- key - a unique identifier for each trip
- fare_amount - the cost of each trip in USD
- pickup_datetime - date and time when the meter was engaged
- passenger_count - the number of passengers in the vehicle
- 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

Expected Outcome:

Using the data's longitude, latitude, and passenger count, predict the fare amount. To improve our ability to predict the fare amount variable from the given dataset, we can extract new features from the dataset.

Increment 1

Related Work (Background):

Attained data set contains over two lakh travels and spans the years 2009 to 2015.

Dataset: Uber.csv

Detail Design of features:

Given Features:

- key - a unique identifier for each trip
- fare_amount - the cost of each trip in usd
- pickup_datetime - date and time when the meter was engaged
- passenger_count - the number of passengers in the vehicle
- 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

Created Features:

- Distance (including pickup latitude and longitude. Latitude and longitude of drop off)

Analysis:

- After inspecting the dataset, I discovered that it is in csv format with comma separated values, which I must read in. As I'm doing it in Google Colab, I'm uploading my dataset using Google Colab library and using browse function to upload dataset and save it in a dataframe. I wanted to run a Multiple Linear Regression Model and estimate the fare of an Uber dataset, where each feature represents one unique representation.
- The key represents the transaction number for that particular trip, which will be distinct. If we want to know how a particular trip went, we can check it with the key to find out how long it took, when the pickup time was, how many passengers traveled, what kind of car was used more frequently, and whether long or short distances were traveled.
- Before removing null data or filling it with mode values, all necessary analysis must be completed, starting with checking the data's head and null values.
- Once the necessary preprocessing processes have been completed, we can build new features, such as distance calculations utilizing pickup and dropoff latitude and longitude with all four features, and then drop the other four features to reduce the number of columns.
- In order to see the surge in fare depending on time by pickup time column, we can create detailed analysis by creating features like type of car using passenger count and another feature like time of travel and creating bin's in the feature with morning, noon, evening, and night. Then we can drop this column and use the newly created feature as it will give more insights for the prediction.
- Once all necessary fields have been made available, dummification or one-hot encoding will be used to turn continuous data into categorical variables. As soon as the data is prepared, we divide the X and Y variables into independent and dependent ones in a ratio of 70:30 for the train and test runs.
- Once the data has been divided, the model will need to be fitted using a multiple linear regression model once the data has been trained. Now, in order to determine residual errors, we predict the data for the train fare amount for which we already have a number. We then compare it to the projected value and note any differences.
- Then use the same model on a test dataset to examine how it will predict, how accurate it will be with observed and predicted values, and how the train and test data are fitted. At this point, you should calculate the MSE, RMSE, and R Square metrics to determine how well your model predicts. Additionally, you

may utilize any fresh data to predict, estimate its accuracy, and visually assess the performance of your model by drawing a best fit line.

Implementation: Work Completed:

- **Description:** After obtaining a dataset, we completed the necessary preprocessing procedures, followed by the required data analysis, visualization, and train and test split data.

Responsibility:

- **Sailesh** conducted the necessary research to identify the best dataset from which to make the prediction. He then uploaded the dataset and performed a few preprocessing operations, such as computing null values, reading the dataset, and cleaning the dataset.
- **Kishore** completed the feature engineering portion by adding new features, deleting unnecessary columns, performing the required categorical encoding, dummifying data using a single hot encoding operation or label encoder, and completing the required visualizations to gain insights into the removal of unnecessary features.
- **Yaswanth** completed the necessary visualizations, the train-test split, the multiple linear regression model, and determined the model's accuracy to apply the model to any additional dataset.

References:

[1]Fares Dataset

[2]Fare Amount Prediction

[3]Comparing Fares