Mini Project

Regression Analysis on Fare Amount Prediction

Charan
Data science trainee at Almabetter

Flow of the Presentation

- Problem Statement
- Exploratory Data Analysis
- Data Preparation
- Feature Selection
- Model Implementation
- Evaluation of model
- conclusion

Problem Statement

Currently The fare of a taxi ride is function of the mileage and the duration of the ride (sum of drop charge, distance charge and time charge). The drop charge is constant and the distance can easily be estimated but evaluating the duration is not a trivial task. It is the result of complex traffic processes that are nonlinear.

The problem here is we need to analyze the data and create an efficient model that will estimate the fare prices accurately

Data Pipeline

Data processing-: At first phase checked for null values and changed the datetime containing column in dataset.

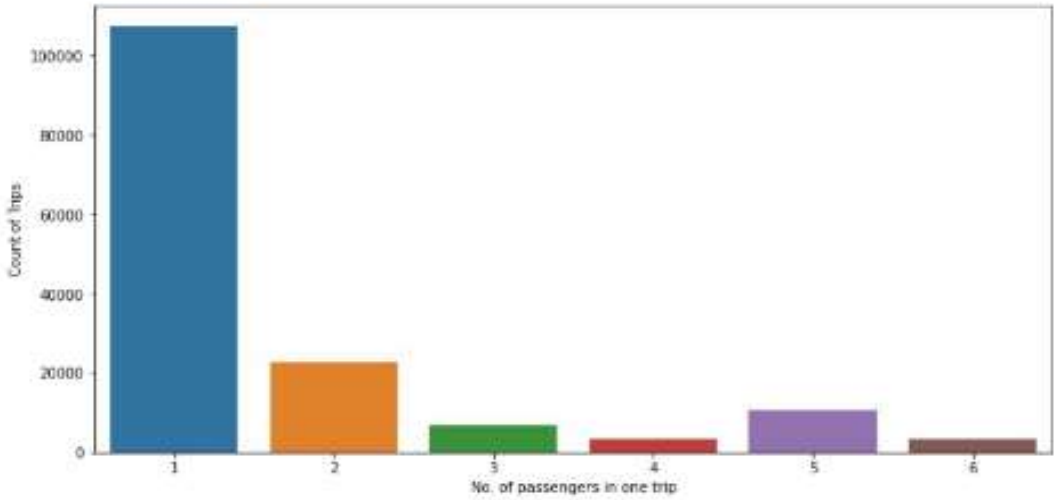
- ❖ Separate Features : Separate dependent and Independent variables. Divide them in different columns.
- ❖ EDA: Exploratory Data analysis was done on the features selected in the phase. (Google map visualization)
- ❖ Data processing-2: Preparing the new dataframe with selected columns And dropped dummy variables.
- ❖ EDA: Exploratory Data analysis was done on the features selected in the phase. (Univariate and scatter plot)
- ❖ Create a model: Finally in this part created models, trained and tested it on the available dataset.

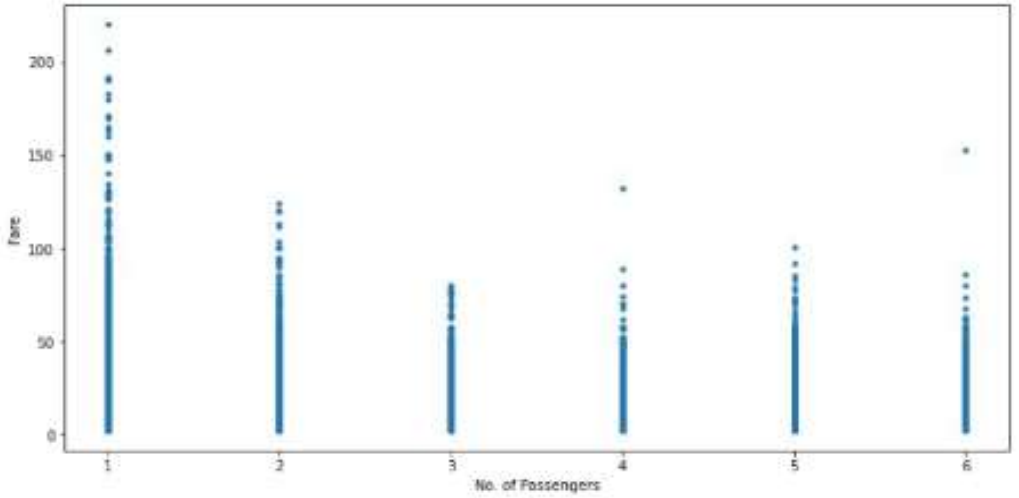
Data Description

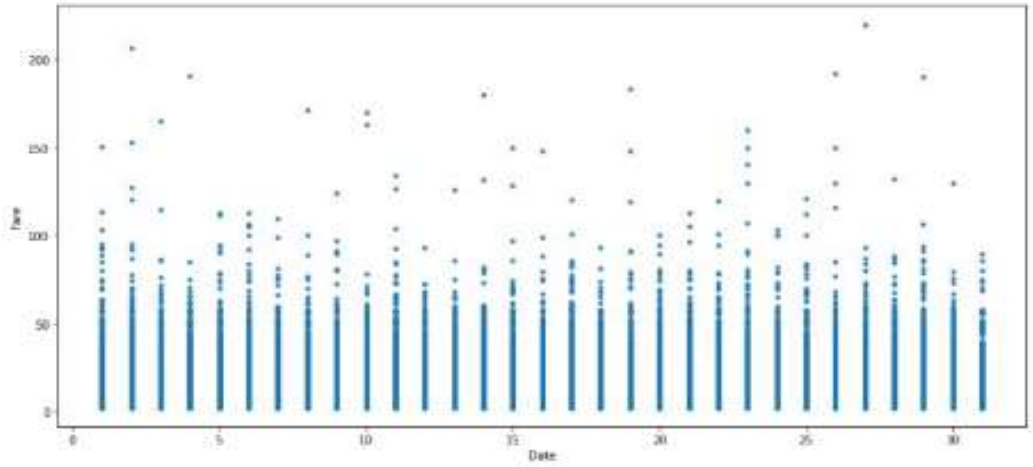
Attribute Information:

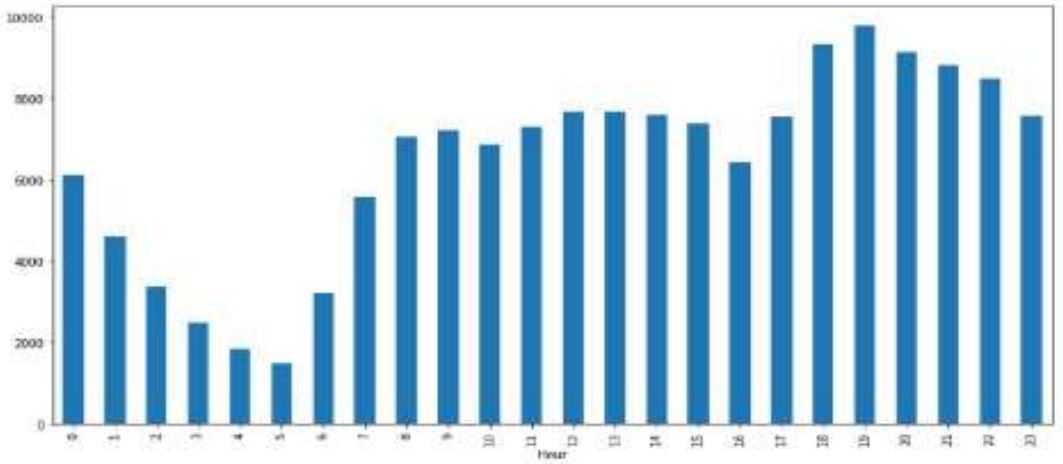
- key - a unique identifier for each trip
- fare_amount - the cost of each trip in usd
- pickup_datetime - date and time when the meter was engaged
- passenger_count - the number of passengers in the vehicle (driver entered value)
- pickup_longitude - the longitude where the meter was engaged
- pickup_latitude - 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

Exploratory Data Analysis

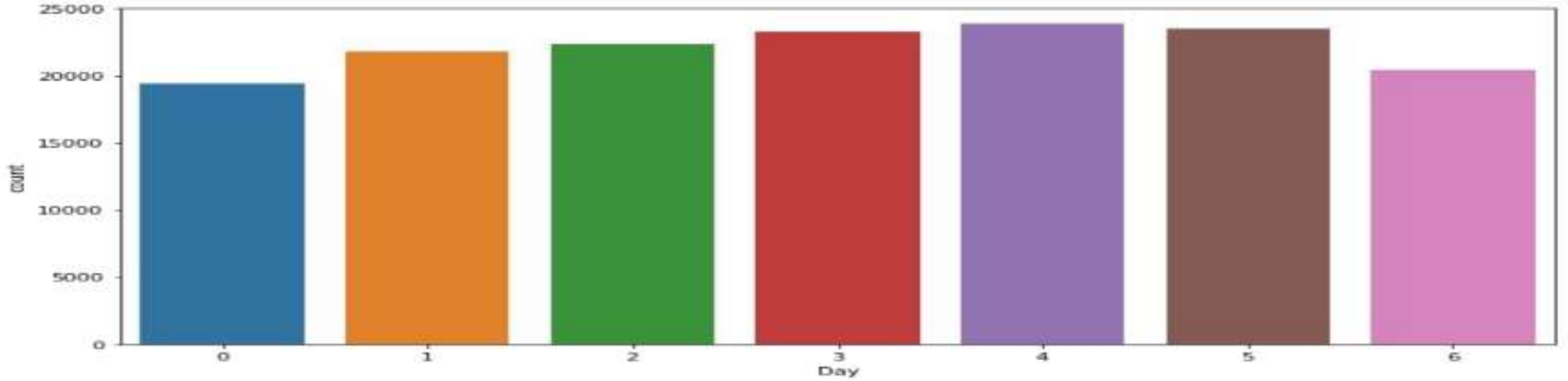


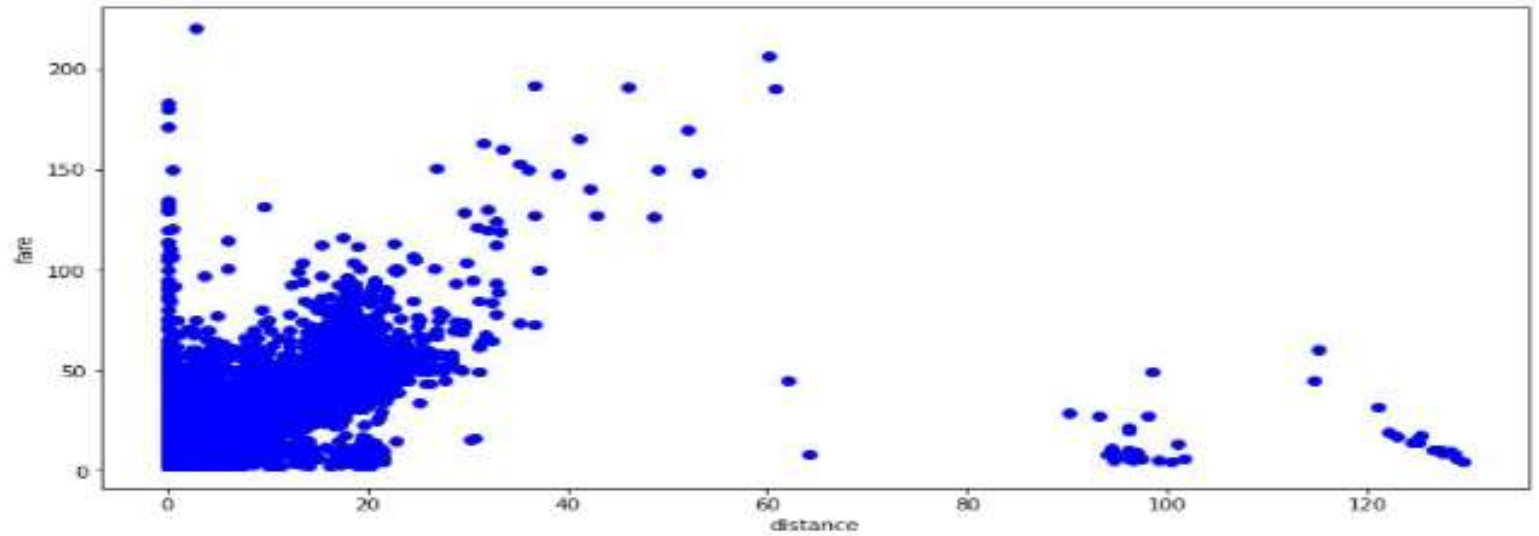




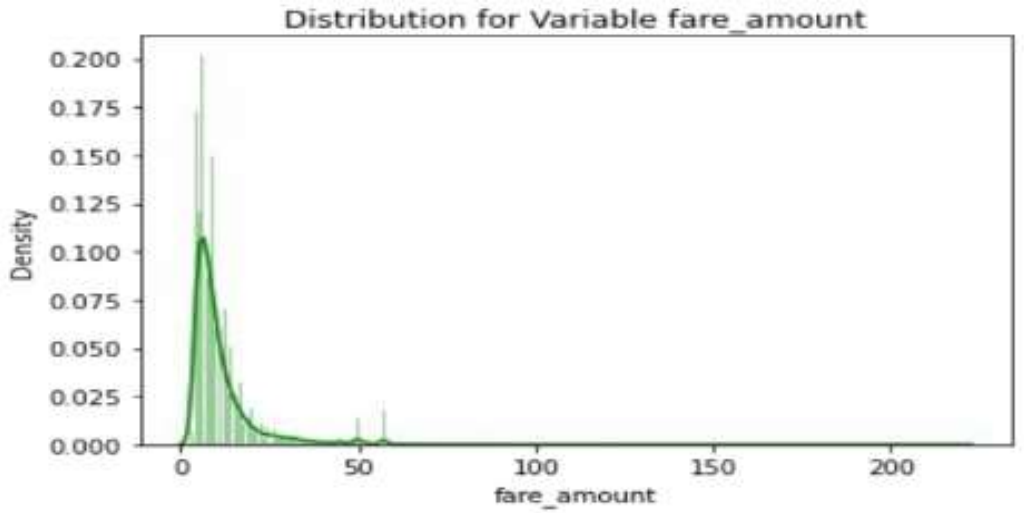


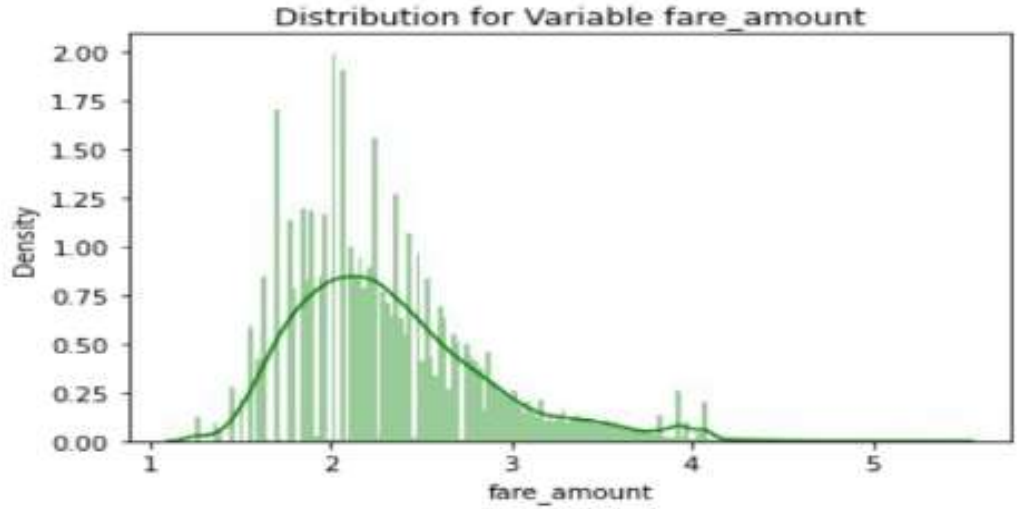
Exploratory Data Analysis

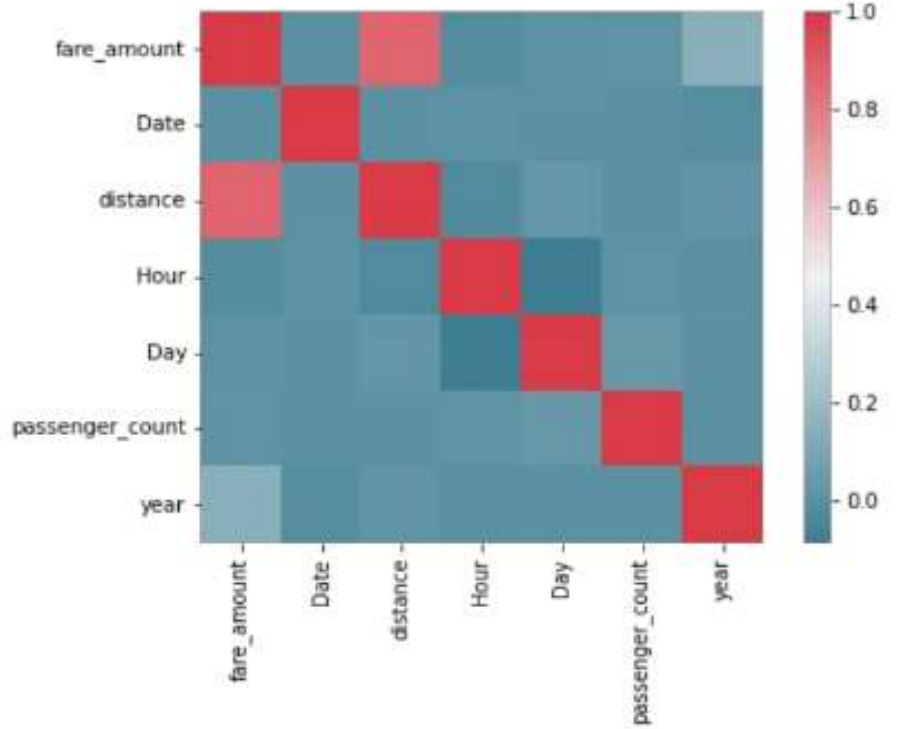




Exploratory Data Analysis







Data Preparation

- ➢ Preparing the new Dataframe with selected columns
- ➢ Variance inflation factor was used to check the correlation among the variables.
- ➢ Few features are dropped.
- ➢ Divided the dataset into train and test set in the ratio of 90:10
- ➢ StandardScaler was used to scale the data

```
#size of train and test datasets
print(f'Size of X_train is: {X_train.shape}')
print(f'Size of X_test is: {X_test.shape}')
print(f'Size of y_train is: {y_train.shape}')
print(f'Size of y_test is: {y_test.shape}')
```

```
Size of X_train is: (139075, 7)
Size of X_test is: (15453, 7)
Size of y_train is: (139075,)
Size of y_test is: (15453,)
```

TRAINING THE MODEL

FOR MODELING WE TRIED VARIOUS REGRESSION ALGORITHMS LIKE:

*LINEAR REGRESSION

*XGBOOST REGRESSOR

EVALUATING THE MODEL

AFTER COMPLETING THE MODEL, IF THE DIFFERENCE BETWEEN ACTUAL VALUE AND THE PREDICTED VALUE IS VERY SMALL THEN IT IS CONSIDERED TO BE A GOOD MODEL AND IT CAN BE USED TO MAKE FUTURE PREDICTIONS

FEW METRIC TOOLS

* R2(R – SQUARED)
* MAE (MEAN ABSOLUTE ERROR)

Conclusion

➢ We calculate the numbers of cabs with respect to Hours, weekly, monthly, day wise and yearly.
➢ Value counts of these cabs between 6pm to 10pm is pretty high.
➢ It shows very less counts on weekend and high counts during working days.
➢ In the months of March, April and May we have a high number of cabs count.
➢ Use of these cabs in year 2012 and 2013 was much higher than other years.
➢ We assumed that 7 in the day of week as the Sunday and stores are mostly closed during Sundays.
➢ single travelling passengers are most frequent travellers.At the sametime we can also conclude that highest Fare are coming from single & double travelling passengers.
➢ Linear Regression and Xgboost are used to train the model.
➢ Both models performed well but Linear Regression performed better. It gave R-squared score of 81% and 0.17 of MAE score.


THANK
YOU