Taxi Fare Prediction

This project “TAXI FARE PREDICTION USING REGRESSION ALGORITHMS” is a research work to find the distance between two geolocation coordinates and predict the fare rate to reach the destination from one location to another using machine learning techniques. Prediction of fare rate for taxi ride for specified distance is a desirable task. This project would be more helpful for all the consumers, who are actually looking for a cab ride. To build a machine learning model, Kaggle taxi dataset of New York City is used.

The Dataset is analysed and pre-processed before training the machine learning models like Multiple Linear regression, Random Forest regression and XGBoost regression models. The train test split is 75% and 25%. Comparison is made based on the result of the individual model’s accuracy rate. Finally, comparatively a model with high accuracy is considered for predicting the taxi fare rate between two locations. Furthermore, in future it can be developed as a live application in both web application and mobile application.

DATASET DESCRIPTION

Dataset consists of two files namely, train.csv which consists of input features and target fare_amount values for the training set it consists of about 55 million rows. test.csv which consists of input features for the test set about 10 thousand rows. Supervised Learning was a goal to predict fare_amount with the help of attributes like pickup_datetime,drop_latitude, drop_longitude, drop_latitude, drop_longitude, passenger count.

To predict the duration and fare amount three supervised techniques such as Extreme Gradient Boosting, Linear Regression and RandomForest have been implemented in this project.

Random Forest and Extreme Gradient Boosting are ensemble models that combine the decisions from multiple models to improve the overall performance of prediction. The taxi dataset contains various details about the taxi ride with given as input to the approaches and prediction is done. Dataset contains totally 8 labelled attributes which include the dependent variable fare amount.

Attribute	Description
Key	Year, time of pickup and id.
Pickup_datetime	Year and time of pickup.
Pickup_longitude	Geographic coordinate that specifies the east–west position of a point on the Earth's surface.
Pickup_latitude	Geographic coordinate that specifies the north–south position of a point on the Earth's surface.
Dropoff _longitude	Geographic coordinate that specifies the east–west position of a point on the Earth's surface.
Dropoff _latitude	Geographic coordinate that specifies the north–south position of a point on the Earth's surface.
Passenger_count	Number of passengers.
Fare_amount	Fare amount of a taxi ride.

Tools and Technology Used: -

1. Python
2. Jupyter Notebook 5.6

Data Preprocessing

1. Handling Missing Values(Null Values)
2. Removing Outliers (Box plot)

Algorithms experimented

1.XGBoost regression.

2. Linear Regression.

3.Random Forest Model.

Result and conclusion

Regression algorithms like XGBoost, Random Forest, and linear regression were used. The XGBoost is accurate for fare prediction. Predictive analysis helps in identifying estimated fare and reduces the excess fares. In terms of accuracy, XGBoost algorithm works better than Random Forest algorithm and Linear regression.

Model	Accuracy Score(in %)	RMSE (in %)
XG-Boost	86.88	3.25
Random Forest	70.57	5.27
Multilinear Regression	61.16	6.06

To further improve the prediction accuracy, more variabilities need to be considered and modelled. Although the rides in hour and average speed in hour work as proxies for traffic, more modelling on the effect of location is needed. These quantities could be calculated for different areas to further model local effects of traffic. Also, modelling traffic and the effect of location in between pickup and dropoff points should be considered as well as difference in drivers' speed.