Taxi Fare Prediction

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

Dataset Information:

Column	Description
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 dropoff point
latitude_of_dropoff	Latitude of the taxi ride dropoff point
no_of_passenger	count of the passengers during the ride
amount	(target variable)dollar amount of the cost of the taxi ride

Scope:

- Prepare and analyse data
- Perform feature engineering wherever applicable
- Check the distribution of key numerical variables
- Training a Random Forest model with data and check it's performance
- Perform hyperparameter tuning

Learning Outcome:

The students will get a better understanding of how the variables are linked to each other and how the EDA approach will help them gain more insights and knowledge about the data that we have and train Random Forest model with the data. Also, use GridSerachCV to get best hyperparameters to build optimized Random Forest model for prediction.