Report on Taxi Project

Introduction

This report provides an analysis of a taxi dataset, containing information on taxi trips taken in New York City. The dataset contains various features related to each taxi trip, such as trip distance, fare amount, payment type, tolls amount, pickup and dropoff locations, and more.

The objective of this project is to develop a model that can accurately predict the fare amount for a given taxi trip based on the available features. To achieve this goal, we will perform exploratory data analysis (EDA) to identify the most significant features affecting the fare amount and then build a machine learning model using these features.

Data

The dataset used in this project consists of 1.5 million taxi trips taken in New York City from January 2013 to December 2014. The data is provided by the New York City Taxi and Limousine Commission (TLC) and can be downloaded from their website.

The dataset contains 21 columns, including trip distance, rate code, store and forward flag, payment type, fare amount, extra, MTA tax, tip amount, tolls amount, improvement surcharge, total amount, pickup location ID, dropoff location ID, year, month, day, day of week, hour of day, trip duration, and calculated total amount.

For our analysis, we removed all columns except for the following features: payment type, trip distance (in kilometers), rate code, tolls amount, and tip amount. The target variable for our prediction model is fare amount. We also converted the trip distance from miles to kilometers and the trip duration from hours to minutes.

EDA

In the EDA phase, we performed various analyses to understand the relationships between the features and the fare amount. We started by examining the distribution of the fare amount and found that it follows a right-skewed distribution, with most fares falling between $0 and $50. We also looked at the distribution of other features, such as trip distance and tip amount, and found that they also follow a similar pattern.

Next, we explored the relationship between the fare amount and each of the selected features. We found that the fare amount increases with the trip distance, but the relationship is not linear. We also found that the fare amount is higher for trips with a rate code of 2 or 3 compared to trips with rate code 1. Additionally, we observed that the fare amount is higher for trips with tolls amount and tip amount.

We also examined the relationship between the fare amount and the time of day, day of the week, and month. We found that the fare amount is generally higher during rush hours and on weekends compared to other times.

Model Building

After completing the EDA, we built a machine learning model to predict the fare amount for a given taxi trip. We used the Random Forest Regression algorithm for this task and achieved an accuracy score of 91%.

The features used in the model are payment type, trip distance, rate code, tolls amount, and tip amount. These features were selected based on their significant impact on the fare amount, as determined from the EDA phase.

Conclusion

In conclusion, we performed an analysis of a taxi dataset containing information on taxi trips taken in New York City. Through exploratory data analysis, we identified the most significant features affecting the fare amount and built a machine learning model that accurately predicts the fare amount based on these features. Our findings suggest that trip distance, rate code, tolls amount, and tip amount are the most important factors affecting the fare amount for a given taxi trip. The model developed in this project can be used to predict the fare amount for future taxi trips, allowing for better planning and budgeting by taxi passengers.