DATA SCIENCE PROJECT

GROUP NAME: ISTANBULLS

Anil Egin - 3149068 / Eren Karsavuranoğlu - 3164647 / Berkay Demirkazık - 3174016

Doruk Efe Kanber - 3163051 / Hamza Abdella Kedir - 3178572

Predicting Tip Amounts in NYC Yellow Taxi Rides

1. INTRODUCTION

In this project, we analyze a subset of data from the NYC Taxi & Limousine Commission, specifically focusing on a subset of data on yellow taxi rides in May 2015. The objective is to predict the tip amount given to drivers for rides paid with a credit card. We will work on a machine learning model to predict the tipping amount by considering input features such as trip distance, fare amount, and passenger numbers. This project is crucial to understanding the factors, including tipping, that can help improve driver satisfaction and service quality.

2. CLEANING, TIDYING AND PREPROCESSING DATA

We began by thoroughly examining the dataset columns. Using a correlation matrix and heatmap, we assessed how variables relate to each other and their relevance to the 'tip amount' target variable.

We checked for missing values to ensure data completeness and found none requiring imputation. Entries with zero values in critical columns like 'fare_amount' and 'length_time' were removed, as they didn't represent valid trip records. Additionally, to enhance the reliability of our analysis, we carefully examined the dataset for extremely unrealistic values across every variable. Subsequently, we omitted these values, attributing them to outliers or potential systematic errors.

To enhance analysis clarity, categorical variables like 'pickup_BoroCode' and 'dropoff_BoroCode' were converted into integers. We then created meaningful categorical features from them.

We introduced the "ratio_tip" to detect any unusual tipping behavior compared to the fare amount. Similarly, the "dist_time" feature helped spot rides that were unusually fast or slow by comparing trip distance to ride duration. Additionally, we formed the "pair" variable by combining pickup and drop-off borough codes, revealing any peculiar spatial patterns in NYC taxi rides. These engineered features were vital in identifying and fixing outliers and inconsistencies within the dataset.

We also calculated summary statistics for numerical features to grasp data trends. To understand relationships between variables, we gained further insights through visualizations like histograms, box plots, and scatter plots.

3. FEATURE ENGINEERING

In our analysis, we introduced a new feature called 'trip,' which we derived from the combination of 'pickup_BoroCode' and 'dropoff_BoroCode'. Our hypothesis was that tips should be higher for taxi rides that were picked up and dropped off from areas with good socio-economic status. To create this feature, we first converted the categorical variables into integers. Then, we applied a specific mathematical formula:

$$trip = pickup\_BoroCode * 5 - (5 - dropoff\_BoroCode)$$

This calculation generates a unique identifier for each route combination. By quantifying the routes in this manner, we aim to provide a robust indicator that enriches our model's understanding of how the interplay between different boroughs impacts tipping behavior.

4. MODELS

Our approach began with thoroughly understanding the dataset columns to determine significant variables influencing tips. The key feature that had the most effect on the tips was fare_amount. Other important futures were trip_distance and length_time.

We have done hyper-parameter tuning for each model with much smaller data, specifically, 30k data entries out of 300k+, using RandomizedSearchCV for faster computation. Once we found the best parameters for each model, we deployed XGBoost, Random Forest, and Linear

Regression using the entire dataset with standardization. We have ended up with 0.49, 0.49, and 0.575 MAE scores, respectively. We then combined its predictions with other models' predictions to improve accuracy. We calculated ensemble weights based on the inverse MAE of each model, ensuring that models with lower error influenced the final prediction more heavily. Unfortunately, it did not yield a better result, giving 0.49 MAE on test data. Each of those models performed inefficiently, resulting in much higher test errors, which proved we had overfitted in each of these models and made weak predictions on unseen data.

To avoid overfitting, we built a neural architecture involving an input layer with 200 neurons and a ReLU activation function, followed by an output layer designed for regression tasks. Although this architecture resulted in higher MAE in its training, it achieved better generalization and lower MAE scores on unseen data.

5. CONCLUSION

Our analysis of NYC taxi tip data provided valuable insights into passenger tipping behavior. We found that factors such as fare amount, trip distance, and length of the ride significantly influence tips, while outliers offered further details about unusual tipping patterns. Retaining outliers helped us understand anomalies that represent unique tipping behaviors and proved to be helpful in the generalization of models. Although early models exhibited some overfitting, the neural network demonstrated better generalization and predictive reliability on unseen data.

To advance this project further, incorporating additional contextual data such as weather conditions, traffic congestion, and significant city events could provide deeper insights into variations in tipping behavior, as these factors can profoundly influence taxi usage patterns and passenger behavior. Additionally, exploring more sophisticated model architectures, such as deep learning models that effectively capture sequential patterns and time-series relationships, would enhance predictive accuracy. These improvements are expected to provide a more granular understanding of tipping behaviors and improve the robustness and reliability of predictive models in real-world scenarios.