Phase 2 Task 3: Machine Learning Insights

Team 2: Tsyr Rau Chen, Arailym Duisengali, Sheikh Noohery, Lo Ying Wu, Kuan Rong Yang

Factors Influencing Predictions

The models predict tip amounts based on trip duration, distance, and encoded city information. However, correlation analysis shows only weak linear relationships between tips and both duration and distance; tipping behavior is also relatively consistent across cities. This suggests that other unmeasured factors are driving tips that are not captured in this data. Another key challenge is the highly right-skewed tip distribution, which includes many small tips and a few large tips. We intentionally retained these large tips because they represent genuine spending behavior, excluding them could result in the loss of important information about high-tip patterns and customer groups, which is crucial for building realistic models and gaining business insights.

Business Implications

Because the existing features don't strongly point toward tipping, richer data collection is justified. Some high-value additions could be time of day, means of payment, customer demographics, and quality of service. This knowledge of tipping can aid pricing strategies, customer service personalization, and targeted promotions, especially for trips or locations with more variable or higher tipping.

ML Process Insights

Among the models compared, degree-2 polynomial regression has a slight edge over linear and random forests, indicating there is some non-linearity. The underperformance of the random forests could be a consequence of a rather small feature space or sample size. Because tips are so skewed, this likely biases the predictions to underpredict large tips. There is no indication of severe overfitting or underfitting, with error metrics quite stable across models given the features.

Data Bias and Model Limitations

The strong skew in tip amount creates a certain imbalance in that most samples are small tips, which biases the model toward predicting lower values and makes hitting higher tips accurately much more difficult. Without strong predictive features, performance is also capped, highlighting the need for better data collection. Neither very poor nor exceptional scores in this regard signal fundamental algorithmic flaws. Instead, they signal that feature engineering and quality of data are the main areas for improvement.