Predicting Customer Tip

Introduction to Problem & Data

Problem Statement:

Tipping behavior in restaurants plays a significant role in shaping the financial outcomes of service staff and the customer experience. However, tipping is influenced by various factors, including the total cost of the meal, the dining party's characteristics, and the situational context, making it challenging to predict accurately. For this project, I aim to develop a predictive model that estimates the tip amount left by customers based on several features such as total bill, gender, smoker status, and dining context.

This model can provide restaurant managers and staff with actionable insights to better understand the tipping habits of their customers, enabling them to tailor their service strategies. Accurate tip predictions could also help restaurants optimize staff scheduling and evaluate the impact of different dining scenarios on customer generosity. Beyond operational improvements, this analysis contributes to a broader understanding of consumer behavior in hospitality settings, offering valuable implications for the service industry.

Dataset Description:

The dataset used for this project is sourced from Kaggle in csv format, providing a collection of data that captures detailed tipping behavior in restaurant scenarios. It contains 7 variables and 244 rows, providing insights into features such as the total bill, tip amount, gender, smoker status, day and time of dining, and the size of the dining party. The data is straightforward and clean, but additional preprocessing and feature engineering may be required to ensure optimal model performance.

Key challenges in building a predictive model for tip amounts include capturing non-linear relationships between features and the target variable. For instance, the total bill and tip amount may exhibit a non-linear relationship influenced by interactions with factors like party size or dining time. Multicollinearity, such as the potential correlation between total bill and party size, can further complicate regression models by causing instability in coefficients and reducing interpretability. Nevertheless, this dataset offers a strong foundation for predicting tip amounts and identifying key influencing factors, with proper exploratory data analysis (EDA) and advanced modeling techniques serving to overcome these challenges.

Exploratory Data Analysis

The dataset does not contain null values, so it is clean and ready for analysis. It consists of 244 dining records, with total bills ranging from 3.07 to 50.81, party sizes ranging from 1 to 6, and tip amounts ranging from 1.0 to 10.0. The mean tip amount is 2.99, and the median is 2.9. From the histogram of tip amounts, the distribution appears to be right-skewed, as most tips fall between 2 to 4 dollars, with only a few higher tip values exceeding 6 dollars. This suggests that smaller bills with relatively small tips dominate the dataset, which aligns with a casual dining setting where spending is relatively low. However, we need to be aware that the small sample size may introduce some bias, as it might not fully capture diverse dining behaviors or tipping patterns.

In addition to tip amount, tip percentage is also an important feature because it standardizes the relationship between the tip and the total bill, providing a clearer measure of customer tipping behavior. Therefore, I added a new column, tip percentage, to the dataframe for further analysis. The tip percentage ranges from a minimum of 3.56% to a maximum of 71.03%, with a mean of 16.08% and a median of 15.48%. This indicates that most customers leave a tip close to 15%-16% of the total bill. The histogram of tip percentage shows a right-skewed distribution. The majority of tips fall between 10% and 20%, while a few extreme values, such as 70%, are outliers. Overall, the distribution suggests that tipping habits are relatively consistent among customers, with a few exceptions.

We first examine the categorical features, including sex, smoker, day, and time, to explore their relationship with tips.

We examine the categorical features, including sex, smoker, day, and time, to explore their relationship with tips. The sex category indicates the gender of the individual paying the bill. From the distribution, there are 157 male and 87 female customers in the dataset, indicating a significant imbalance in gender representation. This may be influenced by cultural practices where males are more likely to be the primary payer in dining situations. Males tend to leave higher tip amounts but lower tip percentages compared to females. This likely occurs because males dine in larger parties or order more expensive meals, leading to higher bills, while adhering to a fixed tipping behavior. However, this does not mean males are more generous, as females leave a slightly higher percentage relative to the total bill.

The smoker category indicates whether the party includes smokers (Yes) or non-smokers. Non-smoking parties (151) outnumber smoking parties (93). The average tip amount is similar for both groups. However, smokers leave a slightly higher tip percentage, though the difference is minimal and not conclusive.

Day category indicates the day of the week when the dining occurred. The dataset shows that most records are from Saturday (87) and Sunday (76), followed by Thursday (62) and Friday (19). This suggests that dining activity peaks on weekends, while weekdays see fewer customers. The lower count on Friday could be because

diners tend to avoid early dinners before transitioning into weekend activities. The bar plot indicates that tip amounts are slightly higher on Sundays, while Saturdays, Thursdays, and Fridays show similar lower values. However, the tip percentage is relatively consistent across all days, with Sunday and Friday showing slightly higher percentages. This suggests that tipping behavior may not vary significantly by day.

The time feature specifies whether the dining experience occurred during lunch or dinner. The dataset shows that most records are for dinner (176), with far fewer for lunch (68), suggesting that dining out is more common in the evening. The bar plot reveals that the average tip amount is higher during dinner compared to lunch, likely due to higher total bills for evening meals. However, the tip percentage is nearly identical for lunch and dinner, indicating that diners maintain consistent tipping behavior regardless of the time of day.

Then we use correlation matrix and scatter plots to examine the numerical features.

The scatterplot of total bill vs tip shows a positive correlation, indicating that higher bills generally result in higher tip amounts. However, the second scatterplot of total bill vs tip percentage reveals an inverse trend, where tip percentages slightly decrease as the total bill increases. This suggests that customers may tip a fixed proportion for smaller bills but reduce the percentage for larger bills, possibly due to psychological or financial considerations.

The relationship between party size and both tip amount and tip percentage reveals interesting patterns. For tip amount, larger parties generally leave higher tips, but variability exists within group sizes. For tip percentage, values cluster around 10-20%, with smaller groups showing more variability, likely due to individual habits, while larger groups display more consistency, suggesting standardized tipping practices.

In summary, the exploratory data analysis reveals key relationships between features like total bill, party size, time, and tipping behavior. Larger bills and party sizes tend to correlate with higher tip amounts, while tip percentages remain relatively stable but show variability for smaller groups and outliers. For the following Modeling section, our primary focus will be on predicting tip amount, as it directly influences restaurant revenue and provides actionable insights into actual income. Accurately forecasting tip amounts can help restaurants optimize management strategies and improve revenue predictions. Additionally, we will briefly explore tip percentage, as it offers valuable insights into customer behavior and tipping habits. This metric can also serve as an indicator of service quality and dining experiences. By examining both tip amount and percentage, we aim to gain a well-rounded understanding of tipping dynamics and their implications for restaurant operations.

Modeling & Interpretations-Tip Amount

Baseline Model

To predict tip amount, I employed several models to identify the one that best captures the variation in the data and fluctuations in tipping behavior. For each model, I applied an 80-20 train-test split, training on 80% of the data and testing on the remaining 20%. I evaluated the performance of each model using metrics such as mean squared error (MSE), comparing the results against a baseline MSE. To establish the baseline, I calculated the mean tip amount from the dataset. The resulting baseline MSE is 1.9066.

Multiple regression model

I chose to build a multiple regression model because I wanted to use independent variables to predict the tip amount, as I believed these predictors collectively influence the tipping behavior. Multiple linear regression allowed me to analyze the relationships between the tip amount and each predictor while also accounting for their combined effect. Overall, my multiple regression model performed better than baseline, with a training MSE of 0.94 and a testing MSE of 1.32, showing good generalization. Feature importance analysis indicates total_bill as the dominant predictor, while size and day have minor impacts, and sex, smoker, and time contribute negligibly. This confirms that tip amounts are primarily driven by the total bill.

K-Nearest Neighbors Regression Model

I chose to try the k-nearest neighbors regression model because KNN predicts based on the similarity of data points in the feature space. Since tipping behavior could be influenced by localized patterns, such as total bill size, party size, and dining characteristics, KNN is well-suited for capturing these relationships. By grouping similar observations, the model can effectively identify patterns in tips that may not be linear or global across the dataset.

My KNN model performed similarly to the multiple regression model. The training MSE for KNN was 0.969, and the testing MSE was 1.374, which are close to the results from the multiple regression model. This indicates that the relationships between the predictors and the target variable (tip amount) are predominantly linear, limiting KNN's advantage in capturing non-linear patterns. The most important feature in this model was again total_bill, which significantly influences the tip amount.

Decision Tree Regression Model

I chose a decision tree regression model to capture non-linear relationships in the data and to provide a clear, interpretable structure for understanding feature influence on tip amounts. Decision trees can identify patterns that other linear models might miss while remaining easy to interpret.

The model outperformed all the models above. But the model's performance on the training data (MSE = 0.37) was significantly better than on the testing data (MSE = 1.11), indicating slight overfitting. This may be due to the tree's tendency to fit specific patterns in the training set. Feature importance analysis highlights total_bill as the most significant predictor, followed by size, which aligns with logical expectations.

Random Forest Model

For my final model, I extended the decision tree approach to a random forest regression model. Random forests combine multiple decision trees, which helps improve accuracy and reduce overfitting by leveraging ensemble methods.

Overall, the random forest model outperformed previous models, with a training MSE of 0.689 and a testing MSE of 1.191, making it the most successful in predicting tip amounts. The model effectively captured non-linear patterns in the data while balancing training and testing performance.

Once again, total_bill emerged as the most influential feature, followed by size and smoker. Other features, such as time and day, showed minimal importance, indicating their limited role in predicting tip amounts.

Modeling & Interpretations-Tip Percentage

I selected the baseline and random forest models to predict tip percentage as it reflects customer behavior and service quality. The baseline model, using the mean tip percentage, had a high MSE of 37.15, showing it failed to capture any variability.

The random forest model improved on this, with an MSE of 16.54 for training but 64.30 for testing, indicating significant overfitting. Total bill was the most influential predictor, while other features like smoker and day had minor impacts.

The large test error suggests high variability in tip percentages and potential dataset limitations, such as outliers and insufficient features. This highlights the need for more robust data or additional predictors (e.g., service ratings) to better explain tipping behavior and improve performance.

Next Steps & Discussion

Summary of Findings

In my analysis of tip amount and tip percentage, the models showed varying predictive capabilities. Random Forest Regression performed the best for tip amount, while tip percentage predictions highlighted challenges due to variability in customer behavior.

Key Findings:

1) Model Performance:
   Random Forest excelled in predicting tip amount, capturing complex patterns and outperforming other models. For tip percentage, while it improved upon the baseline, its performance indicated limitations in generalizing to unseen data.
2) Feature Importance:
   Total bill was the dominant predictor across all models, followed by party size and specific days. Features like smoker and time had minimal influence, suggesting tipping behavior is driven more by spending habits than categorical traits.
3) Challenges and Improvements:
   The variability in tip percentage and imbalanced data (e.g., fewer records for larger parties or certain days) limited model accuracy. Further improvement could involve adding service quality metrics or external factors like time of year.

In conclusion, focusing on tip amount provides actionable insights for revenue optimization, while exploring tip percentage helps uncover behavioral trends, supporting a more comprehensive understanding of tipping dynamics.

Next Steps/Improvements

To enhance the predictive capabilities of the model and gain deeper insights into tipping behavior, I would propose enlarging the dataset size and incorporating additional features as follows:

1) Service Quality Metrics:
   Including data on customer ratings of service quality, such as friendliness, attentiveness, or speed of service, could help capture how the level of service directly influences tipping behavior.
2) Restaurant Type and Ambiance:
   Adding information on the type of restaurant (e.g., casual, fine dining, or fast

food) and its ambiance (e.g., decor, noise level) could provide context on how the dining environment impacts tipping tendencies.

3) Food Quality:
   Features such as food taste ratings and customer satisfaction with meals could reveal how perceived value for money influences tipping decisions.
4) Customer Demographics:
   Integrating demographic information such as customer age group, income level, or dining frequency could help explore patterns in tipping habits among different customer segments.

By integrating these additional factors, the model would better account for the complexities of tipping behavior, resulting in more accurate predictions and valuable insights for restaurant managers. This approach could lead to actionable recommendations for improving service, adjusting pricing strategies, and creating better customer experiences.