DEPARTMENT OF INFORMATION ENGINEERING AND MATHEMATICAL SCIENCES

ENGINEERING MANAGEMENT

Predicting Tips from Taco Delivery Data: A Model Comparison Study

Supervisor:
Prof. Giovanna Maria Dimitri

Candidate:
Gülce Çelik
Mat. 157341

Academic Year 2024-2025

Contents

1.Abstract

This study aims to analyze tipping behavior in taco deliveries and to **predict the tip amount based on various order and restaurant attributes**. Using a dataset of taco sales from 2024 to 2025, we conducted exploratory data analysis to understand distributions, identify outliers, and examine relationships between variables. After preprocessing and feature engineering, three regression models were developed: Linear Regression, Decision Tree, and Neural Network. These models were trained using 5-fold cross-validation, and their performance was evaluated using RMSE and R^2 metrics. Additionally, two different sets of input features were tested to assess the impact of variable selection. The goal is to determine which features have the most predictive power and to identify the most effective model for estimating tip amounts.

2.Introduction and Background

Tipping behavior is a critical aspect of customer interaction in the food delivery industry. Understanding the factors that influence how much customers tip can help restaurants optimize service quality, pricing, and delivery operations. In this project, we explore a dataset of taco sales collected between 2024 and 2025, which includes detailed information such as taco type and size, order price, number of toppings, delivery duration, location, restaurant name, and whether the order was placed on a weekend.

Our primary aim is to model the tip amount (Tip....) as a continuous variable, using these available features. Before modeling, comprehensive exploratory data analysis was conducted to understand data distributions, identify outliers, and examine potential relationships between features. Visualization techniques and correlation analyses revealed patterns in how tips vary across taco types, restaurant branches, and delivery characteristics.

Instead of relying on a single hypothesis or engineered classification like "high tip," this study approaches the problem from a pure regression perspective. The goal is to identify which variables best explain the variation in tip amounts and to evaluate which machine learning model can predict tips most effectively. The analysis tests three models—Linear Regression, Decision Tree, and Neural Network—using 5-fold cross-validation and two different sets of input features.

Similar studies have been conducted in the context of ride-sharing and online food delivery platforms such as Uber, DoorDash, and Grubhub. These analyses have identified several variables that influence tipping behavior, including delivery speed, service price, time of day, and customer satisfaction ratings. For instance, research on Uber has shown that longer wait times or peak-hour pricing can affect both the likelihood and size of tips. Likewise, studies on online food delivery platforms suggest that food quality, platform interface, and delivery

accuracy all play roles in tip decisions. While most prior research focuses on general delivery behavior, this project narrows the focus to a specific domain—taco deliveries—allowing for more targeted and actionable insights.

3.Dataset

The dataset used in this project consists of 1,000 taco delivery records collected during the years 2024 and 2025. Each row represents a unique order, and the dataset includes detailed information about the restaurant, delivery process, order contents, and tip amount. The target variable is **Tip ($)**, a continuous numerical value that indicates how much the customer tipped for a given order. **Main variables include:**

- **Order ID**: Unique identifier for each transaction.
- **Restaurant Name**: The restaurant that fulfilled the order (e.g., *El Taco Loco*, *Casa del Taco*).
- **Location**: The city where the order was delivered (e.g., *Austin*, *Dallas*).
- **Order Time / Delivery Time**: Timestamp of when the order was placed and delivered.
- **Delivery Duration (min)**: Total delivery time in minutes.
- **Taco Type**: Type of taco ordered (*Beef Taco*, *Chicken Taco*, etc.).
- **Taco Size**: Size category of the taco (e.g., *Regular*, *Large*).
- **Toppings Count**: Number of extra toppings selected.
- **Distance (km)**: Distance between restaurant and customer in kilometers.
- **Price ($)**: Total price of the order before the tip.
- **Weekend Order**: Boolean indicator showing whether the order occurred on a weekend.

Data Characteristics:

- Number of rows: 1,000 (each representing a unique order)
- Number of columns: 13
- Variable types:
  - 3 numerical (Tip ($), Price ($), Distance (km))
  - 3 integer (Toppings Count, Delivery Duration (min), Order ID)
  - 6 categorical (Restaurant Name, Location, Taco Size, Taco Type, Order Time, Delivery Time)
  - 1 boolean (Weekend Order)

Data Quality:

- There are no missing values in the dataset.
- Outliers were detected using both Z-score and IQR methods, especially in price and tip columns.
- Outliers were capped to reduce skewness and prevent them from negatively impacting model performance.

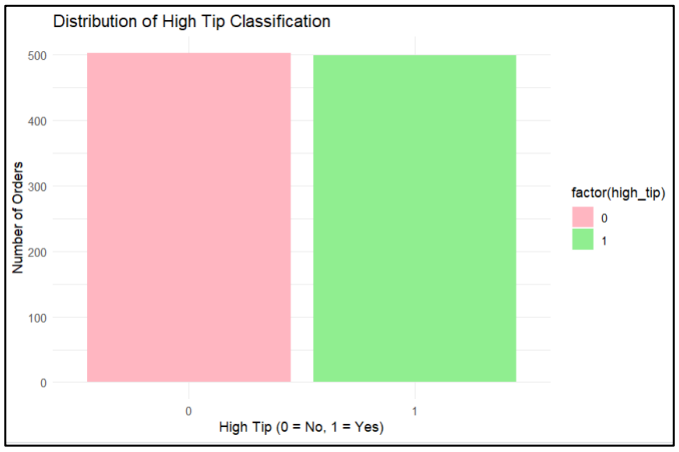

Figure 1: Balanced Distribution of High vs. Low Tips

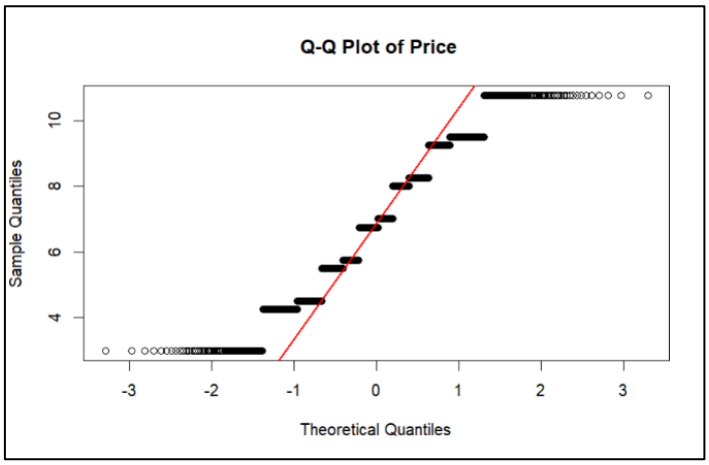

Figure 2: Q-Q Plot Showing Non-Normal Price Distribution

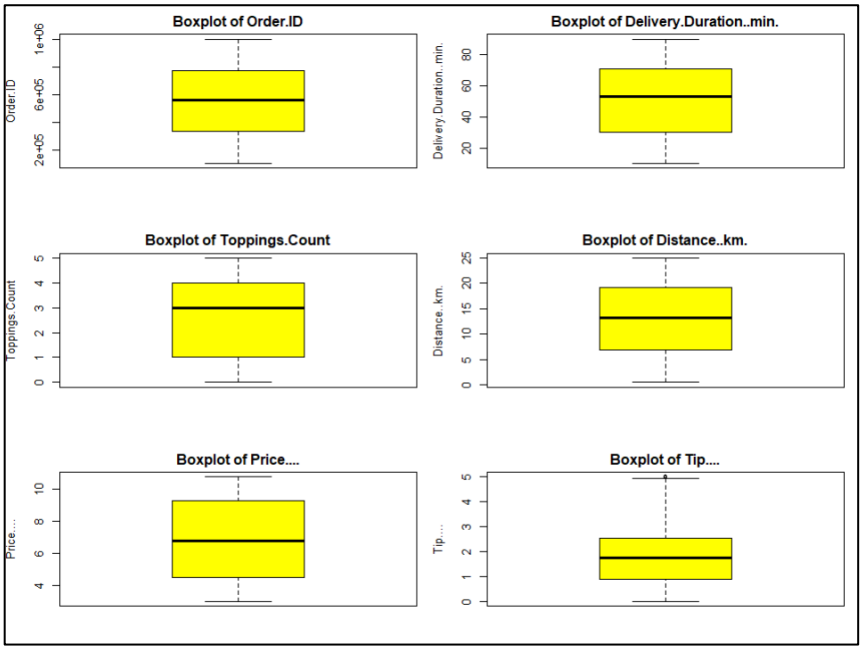

Figure 3: Boxplots of Numeric Features Before Outlier Capping

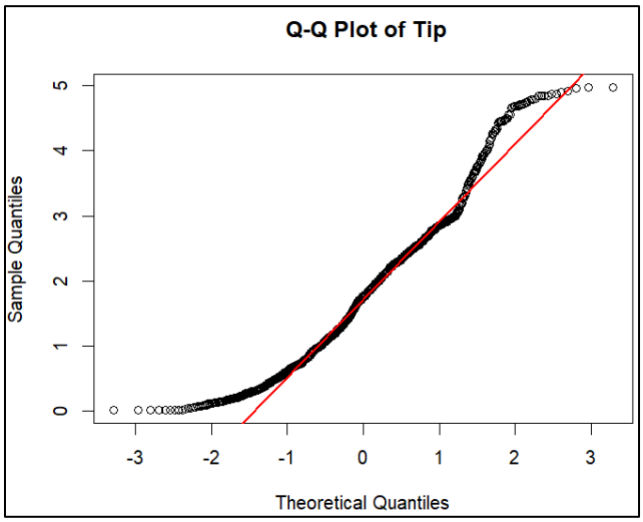

Figure 4: Q-Q Plot of Tip Indicating Right Skew

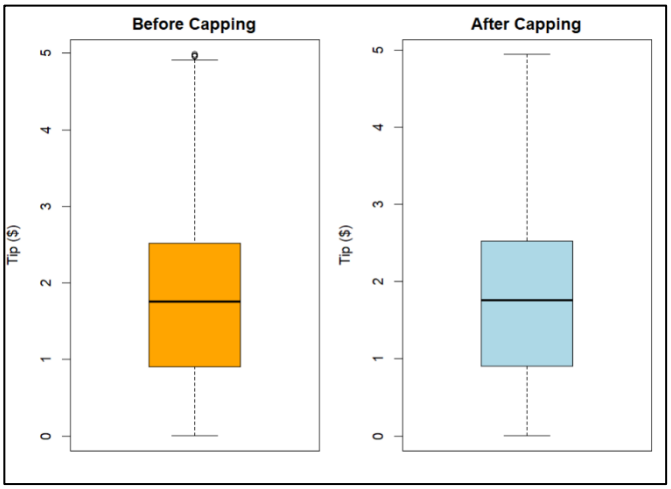

Figure 5: Tip Distribution Before and After Outlier Capping

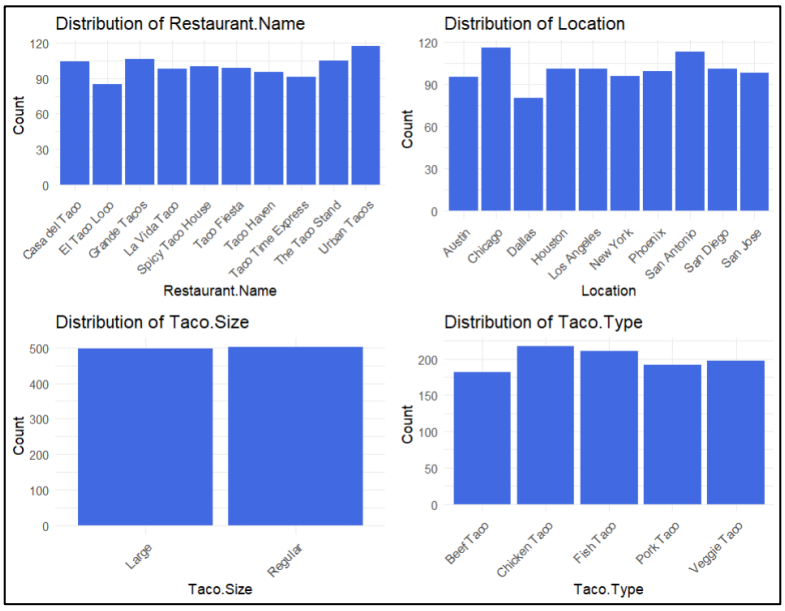

Figure 6: Distribution of Categorical Variables in the Dataset

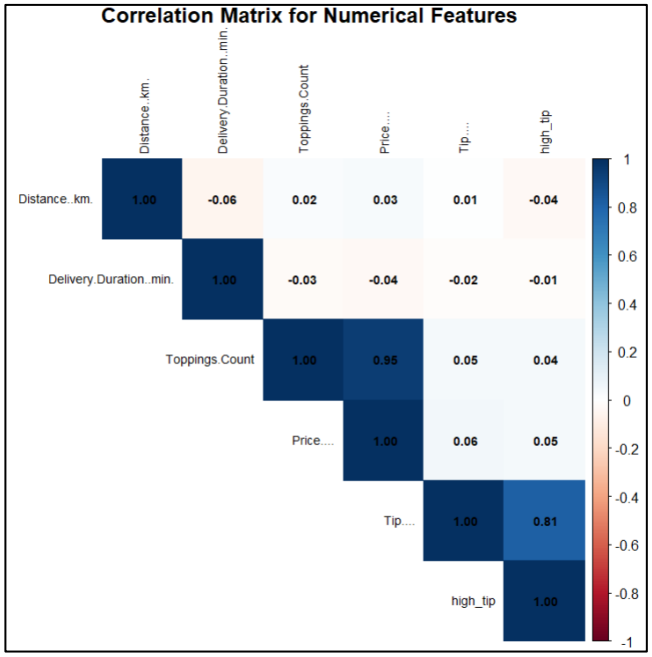

Figure 7: Correlation Matrix of Numerical Variables

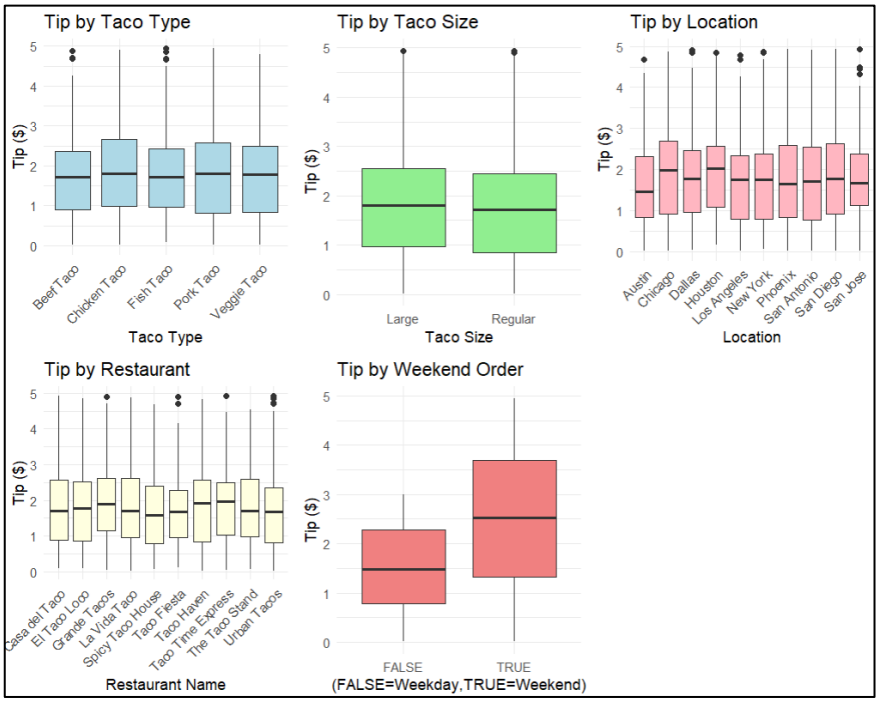

Figure 8: Impact of Categorical Variables on Tip Amounts

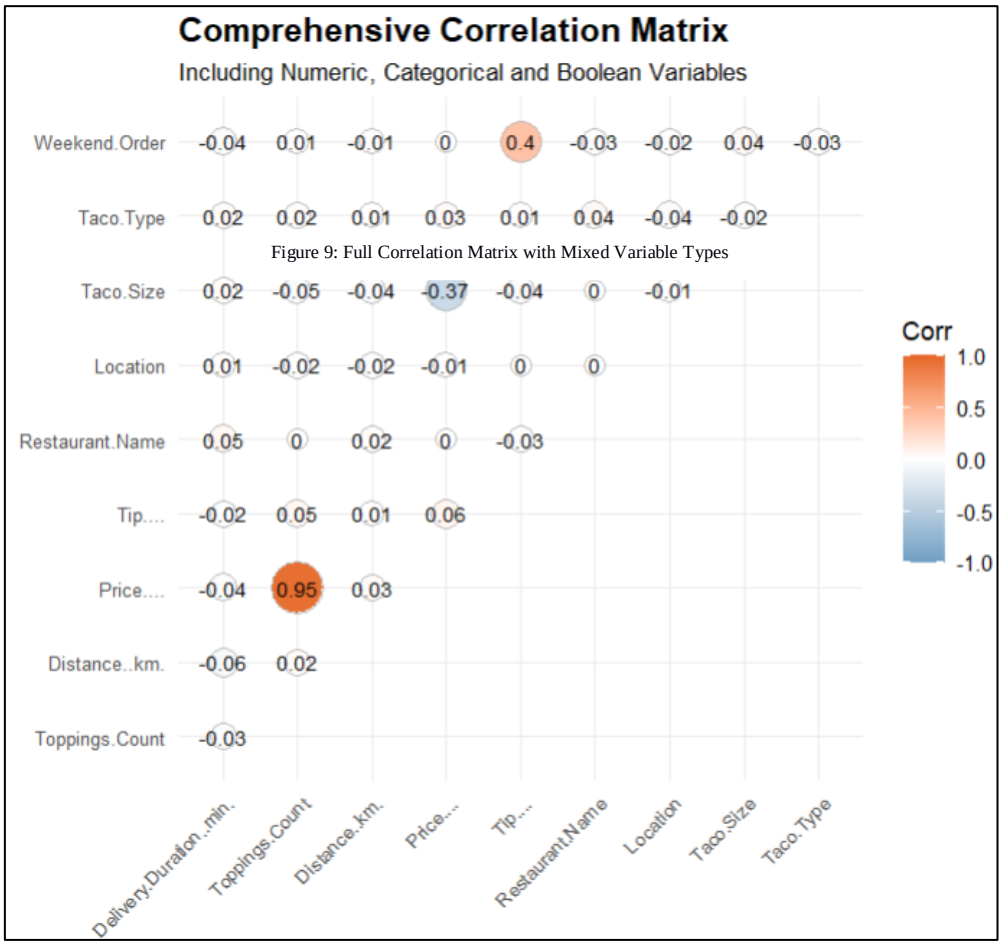

Figure 9: Full Correlation Matrix with Mixed Variable Types

Figure 9: Full Correlation Matrix with Mixed Variable Types

4.Methods

This project followed a structured data science workflow that includes data preprocessing, exploratory analysis, and predictive modeling. The main goal was to predict tip amounts based on order and delivery-related features.

1. Preprocessing

- Outlier Detection and Capping:
  Outliers were detected using both **Z-score** and **IQR** methods across all numerical features. Instead of removing them, we capped extreme values at reasonable bounds to preserve data while reducing skewness.
- Missing Values:
  The dataset had no missing values, so no imputation was necessary.
- Feature Engineering:
  Time-related variables (order/delivery time) were excluded from modeling. Boolean variables (e.g., Weekend Order) were converted to numeric, and categorical variables were transformed using **one-hot encoding**. Gini Index was used to measure the diversity of categorical variables, helping to identify which ones have enough variation to be informative for modeling
- Normalization:
  For the Neural Network model, numeric features were standardized using **Z-score normalization** to ensure proper training.

2. Modeling Approach

Three regression models were applied and compared:

3. Linear Regression
   A simple and interpretable model used as a baseline. It helps understand linear relationships between tip amount and features.
4. Decision Tree Regression
   A non-linear, rule-based model that captures feature interactions without requiring normalization. Useful for handling complex relationships and feature hierarchies.
5. Neural Network
   A multi-layer perceptron with one hidden layer. Selected to explore the potential of capturing non-linear relationships between inputs and tips, especially when feature interactions are subtle.

Each model was evaluated using **5-fold cross-validation**, and performance was measured with **RMSE** and **R^2** metrics. Additionally, two different feature sets were used to compare the impact of variable selection.

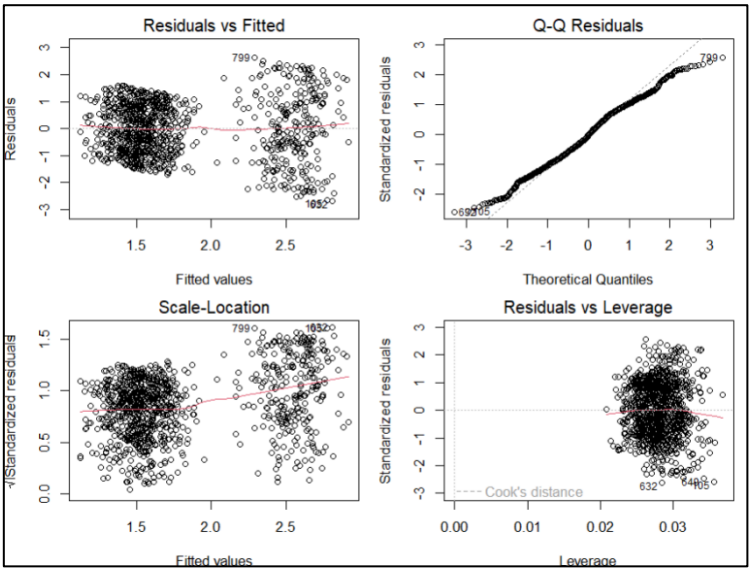

Figure 10: Diagnostic Plots for Linear Regression Model

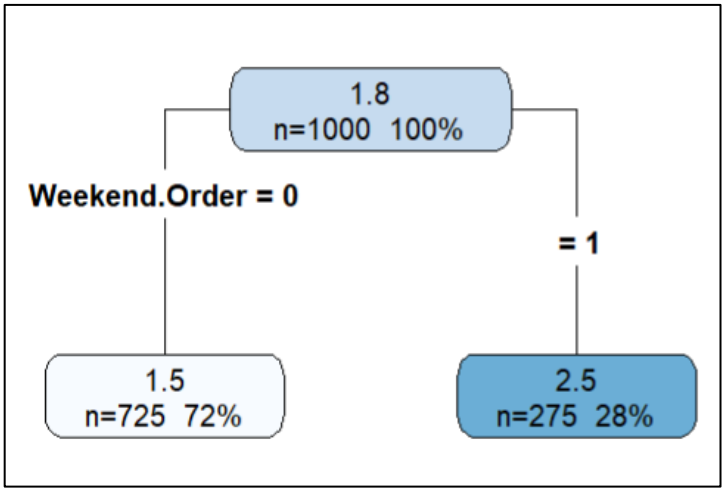

Figure 11: Decision Tree Split Showing Tip Differences by Weekend Order

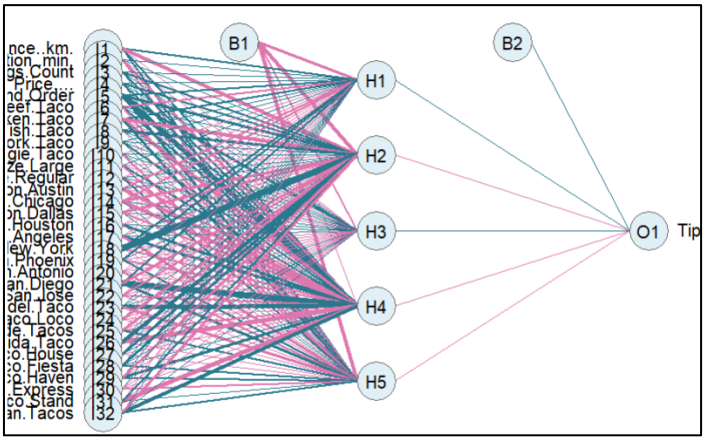

Figure 12: Neural Network Architecture for Tip Prediction

5.Results and Conclusions

Three regression models were trained and evaluated using 5-fold cross-validation to predict tip amounts: **Linear Regression**, **Decision Tree**, and **Neural Network**. Each model was tested with two different feature sets, and model performance was assessed using **Root Mean Squared Error (RMSE)** and **R-squared (R^2)** metrics.

- **Model Performance (Feature Set 1)**

Model	Avg. RMSE	Avg. R^2
Linear Regression	~1.05	~0.14
Decision Tree	~1.07	~0.11
Neural Network	~1.25	~ –0.23

- **Model Performance (Feature Set 2**

Model	Avg. RMSE	Avg. R^2
Linear Regression	~1.04	~0.16
Decision Tree	~1.05	~0.13
Neural Network	~1.22	~ –0.19

Key Findings and Interpretation

- **Linear Regression** consistently outperformed other models in both feature sets. Despite modest R^2 values, it showed the most stable and interpretable behavior.
- **Decision Tree Regression** had slightly worse performance and exhibited overfitting tendencies in smaller feature sets.
- **Neural Network** performed the worst in both accuracy and generalization, likely due to insufficient complexity tuning or noisy feature influence.

Why Were the Results So Limited?

A **comprehensive correlation matrix** (see Figure 9) revealed that **`Tip ($)` has very weak correlations with all other variables**. For example, the correlation between tip and price is only ~0.24, and most other variables show near-zero correlation.

This suggests that **tip behavior may be influenced by latent psychological or external factors** that are not present in the dataset — such as:

- Customer satisfaction or mood
- Quality of interaction with the delivery driver
- Weather or time pressure
- Interface or promotion effects in the ordering app

These limitations explain why models struggled to capture meaningful patterns in tip prediction, regardless of algorithm.

Additional Experiment: Predicting Price Instead of Tip

To test whether modeling failure was due to the data or the methods, we ran a secondary experiment: predicting **`Price ($)`** using only basic features (e.g., `Toppings.Count`). The results were dramatically better:

- Models achieved **R^2 values above 0.90**, confirming that the pipeline and training approach are valid.
- This showed that the **problem is not the modeling**, but rather the **lack of predictive information for the tip target**.

We also excluded variables like `Restaurant Name`, which would trivially reveal price due to fixed menu items. Even then, prediction remained highly accurate — especially when using features logically linked to pricing structure.

Conclusion

Although tip prediction results were modest, the study provided important insights:

- The best-performing model was **Linear Regression**, due to its simplicity and robustness.
- **Low R^2 values were expected**, given that the available features lack strong relationships with tipping behavior.
- **External or behavioral data** would be necessary to meaningfully improve model accuracy.

Future Work

To enhance tip prediction models in future studies, the following steps are recommended:

- Add **time-based features** (hour of order, delivery delays).
- Include **behavioral or satisfaction-related features** (ratings, feedback).
- Incorporate **external data sources**, such as **weather**, **traffic conditions**, or **repeat customer flags**.
- Test **ensemble methods** or **boosted trees** with deeper feature engineering.

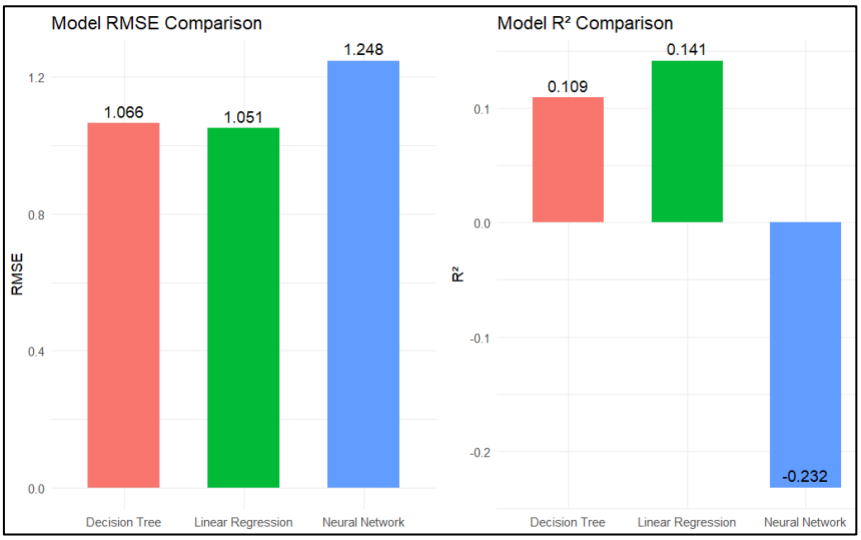

Figure 13: Performance Comparison of Regression Models (RMSE and R^2)

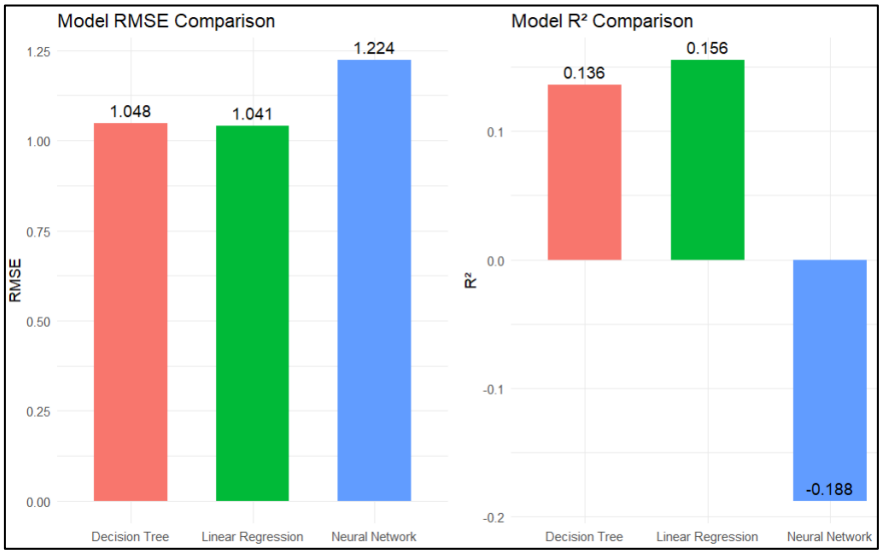

Figure 14: Model Performance with Alternative Feature Selection (RMSE and R^2)

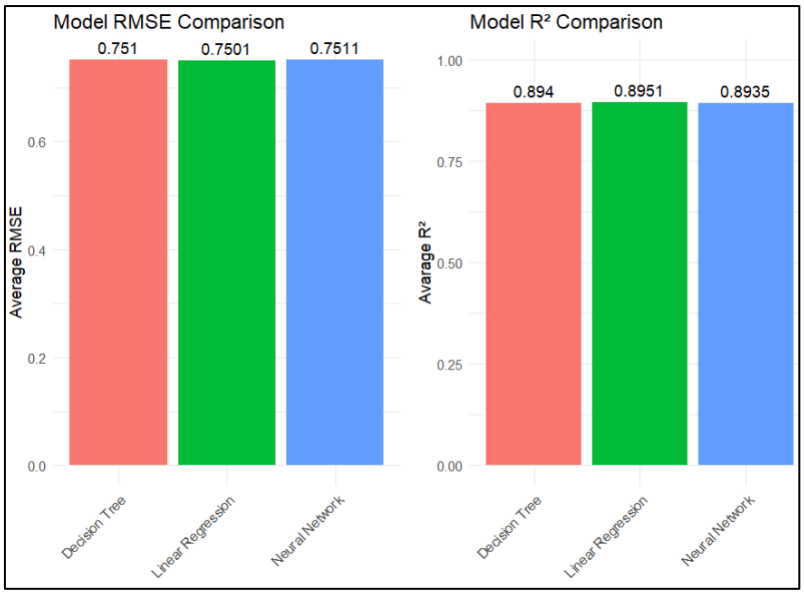

Figure 15: Model Performance Comparison for Price Prediction (RMSE and R^2)

6.Reference

1. **Chandar, B., Gneezy, U., List, J. A., & Muir, I. (2019).**
The Drivers of Social Preferences: Evidence from a Nationwide Tipping Field Experiment.
National Bureau of Economic Research (NBER).
https://www.nber.org/papers/w26380

2.**Wired Magazine (2019).**
Your Secret Uber Tipping Behavior, Exposed.
Wired.com — Summary of NBER's tipping study.
https://www.wired.com/story/your-secret-uber-tipping-behavior-exposed

3. **Liu, Y., et al. (2024).**
The Effect of Dynamic AI Pricing on Tipping Behavior: Empirical Evidence from Uber.
Social Science Research Network (SSRN).
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5080000

4.**Ayessa (Kaggle).**
Waiter Tips – Complete EDA and Regression Model.
Kaggle Notebook.
https://www.kaggle.com/code/ayessa/waiter-tips-complete-eda-and-regression-model

5.Kuhn, M. (2023). *caret Package Documentation*.
https://topepo.github.io/caret/

6. **Atharva Soundankar (2024).**
Taco Sales Dataset (2024–2025) – Simulated dataset for educational purposes.
https://www.kaggle.com/datasets/atharvasoundankar/taco-sales-dataset-20242025

7.DataCamp (n.d.). Linear Regression in R Tutorial.
https://www.datacamp.com/tutorial/linear-regression-R

8. Nugent, C. (n.d.). Introduction to Machine Learning in R – Tutorial Notebook.
https://www.kaggle.com/code/camnugent/introduction-to-machine-learning-in-r-tutorial