**,**

AMRITA VISHWA VIDYAPEETHAM

*Amrita School of Computing*

*Amaravati Campus – 522503*



**19CSE304 – Foundations Of Data Science**

**CASE – STUDY**

***Year: 2024-2025 (Odd Semester)***

**Name : K. Chaitanya Sai**

**Reg. No : AV.EN.U4CSE22121**

**Semester: 5th Semester**

**Topic : Food Delivery Profits Analysis**

**Objective:**

The focus of this case study is to evaluate the **profitability of each food delivery order** by thoroughly analyzing the revenues and costs associated with every transaction. Key components contributing to the profit calculation include:

* **Revenues:** *Order Value*, the amount paid by the customer.
* **Costs:** *Delivery Fee*, *Commission Fee*, *Payment Processing Fee*, *Refunds/Chargebacks*, and *Discounts and Offers*.

This analysis is vital in a competitive food delivery industry where profitability margins are often thin and influenced by multiple cost and operational factors.

**Key Variables:**

1. **Order Value:** Revenue earned from each transaction.
2. **Delivery Fee:** Cost of food delivery.
3. **Commission Fee:** Percentage paid by the restaurant to the platform.
4. **Payment Processing Fee:** Transaction fee for digital payment methods.
5. **Refunds/Chargebacks:** Amount refunded or lost due to order cancellations or chargebacks.
6. **Discounts and Offers:** Reductions provided to customers that lower the revenue.

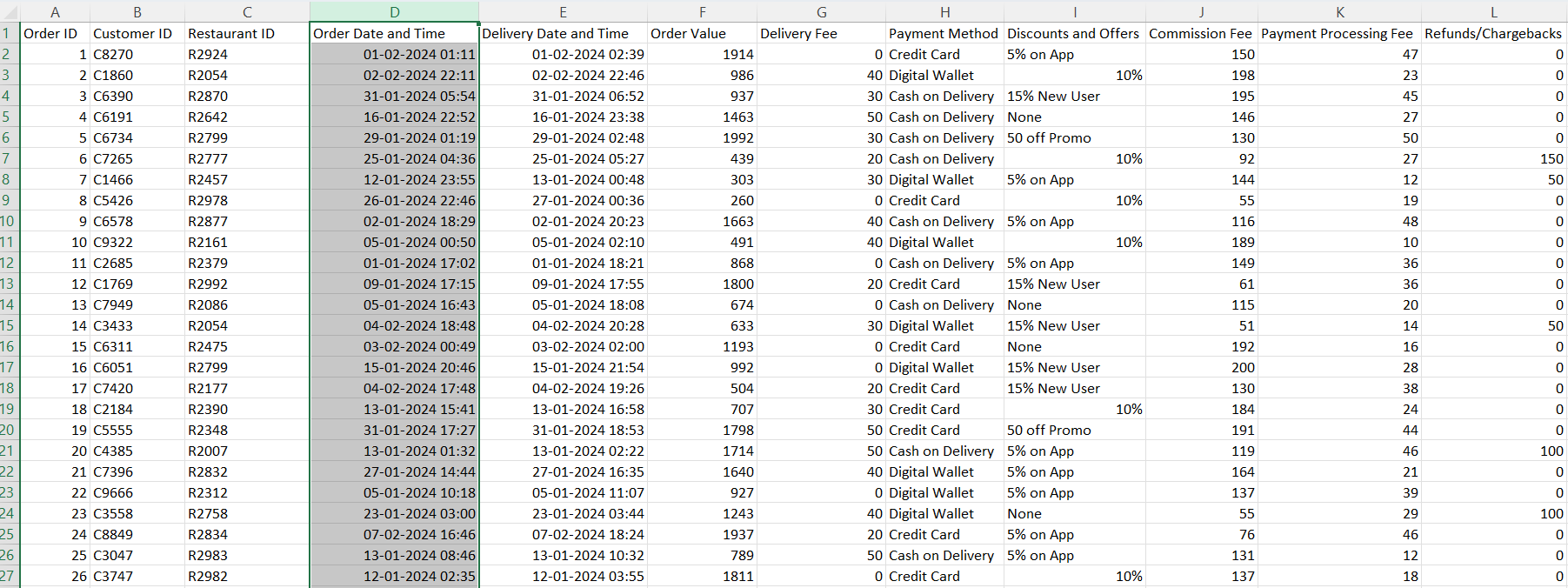
**Data Collection and Preprocessing:**

This section involves cleaning and transforming the dataset to ensure it's suitable for analysis. It covers handling missing values, converting data types, and calculating relevant metrics such as delivery duration, total cost, and profit.

**Content:**

1. **Data Sources**:
   * The dataset contains information related to food orders, including order value, delivery fee, commission fee, payment method, discounts and offers, and other relevant details. This data is used to calculate the profits from each order.

**Dataset:**



1. **. Data Cleaning:**

**Handling Missing Data:** Missing values in the "Discounts and Offers" column are replaced with a placeholder, ensuring no null values affect the analysis.

**Transformations:** Values like "5% off" or "50% off" in the "Discounts and Offers" column are converted into numerical values for analysis.

**Date Parsing:** "Order Date and Time" and "Delivery Date and Time" columns are converted to datetime format to calculate delivery duration.

1. **Cleaning the Discounts and Offers Column**:

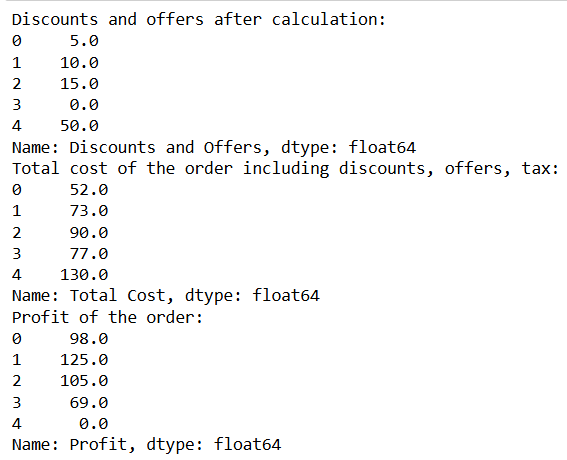
**Handling Percentages:** Discount percentages (e.g., "5% off") are converted into numerical values and applied as a percentage of the "Order Value."

**Adjusting Discount Values:** Discounts ≤15% are treated as percentage discounts on the "Order Value" (e.g., 10% of order value).

1. **Calculating Total Cost**:
   * The total cost of each order is computed by summing up the Delivery Fee, the discount value from Discounts and Offers, and the Payment Processing Fee. This total represents all the costs incurred to process and deliver the order.



1. **Calculating Profit**:
   * Profit for each order is calculated by subtracting the Total Cost from the Commission Fee. The Commission Fee represents the revenue generated by the delivery platform from each order. This allows us to understand the actual profitability of each order after accounting for all associated costs.
   * 



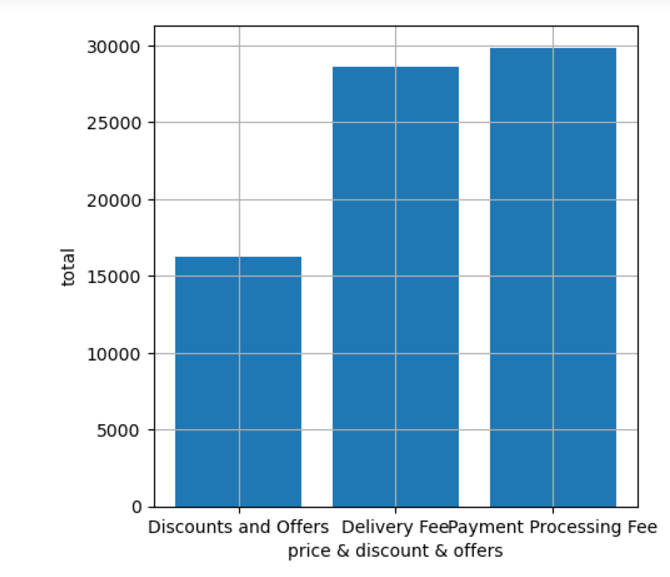
**Visualization Analysis:**

This section focuses on visualizing different aspects of the dataset to understand key trends, cost distributions, and profit patterns. Visualizations help highlight relationships between variables and reveal insights about the financial dynamics of the food delivery platform.

* 1. **Cost Distribution Visualization:**

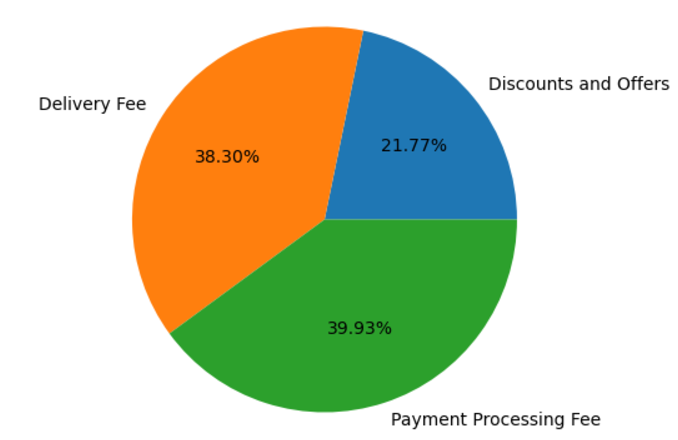
The first set of visualizations focuses on understanding the distribution of costs associated with each order. These costs include:

* **Discounts and Offers**
* **Delivery Fee**
* **Payment Processing Fee**

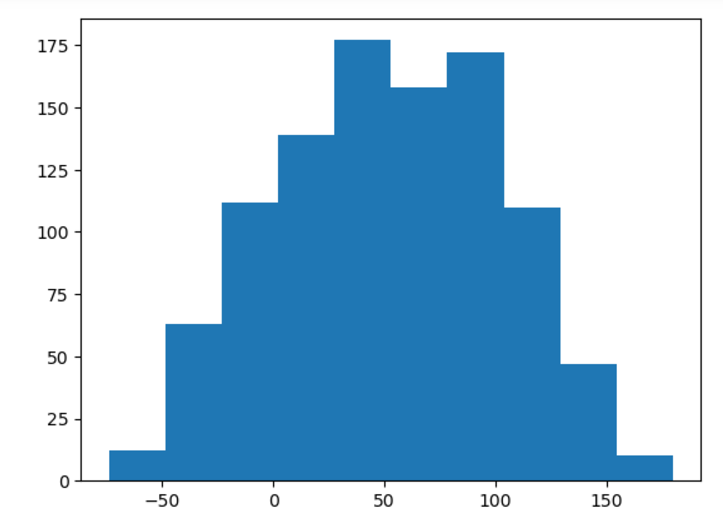


**Pie Chart**:

A pie chart is created to visualize the proportion of each cost category as a percentage of the total. This provides a clear overview of how each cost contributes to the overall cost structure. The percentage labels on the pie chart show the relative importance of each cost type, helping to quickly identify the most significant costs in the system.

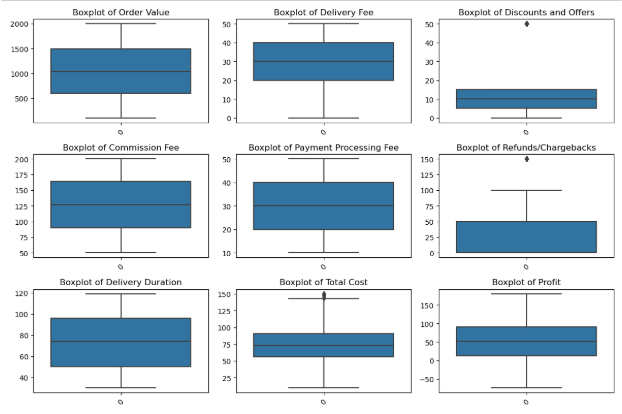


**Histogram of Profit Distribution**:  
A histogram is used to visualize the distribution of profits across all orders. This visualization shows how often certain profit levels occur, which helps in understanding whether profits are evenly distributed or if there are significant outliers. If most orders show similar profit values, the histogram will have a clear peak at that value. However, if there is significant variability in profits, it will be evident from a wider spread in the histogram.



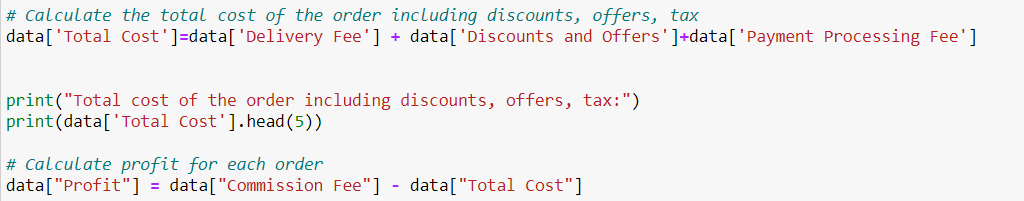
**Boxplot:**

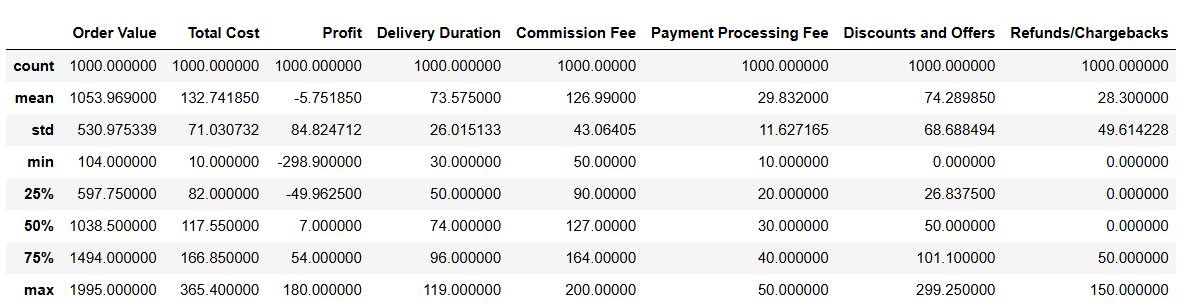
A boxplot (also known as a box-and-whisker plot) is a graphical representation of the distribution of a dataset. It provides a summary of the key statistical properties of the data, including its central tendency, variability, and the presence of outliers. Here's a detailed explanation of the components and how to interpret a boxplot.

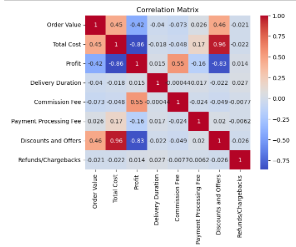
****

**Descriptive Analysis**

1. **Descriptive Statistics**:
   * **Order Value**: The mean, median, and standard deviation of the order values are computed to understand the distribution of order amounts.
   * **Profit Calculation**: Profit is calculated as the difference between the Order Value and the sum of Commission Fee, Delivery Fee, Payment Processing Fee, and Refunds/Chargebacks.

**.**

****

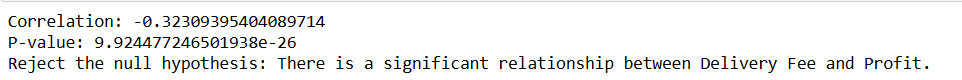
****

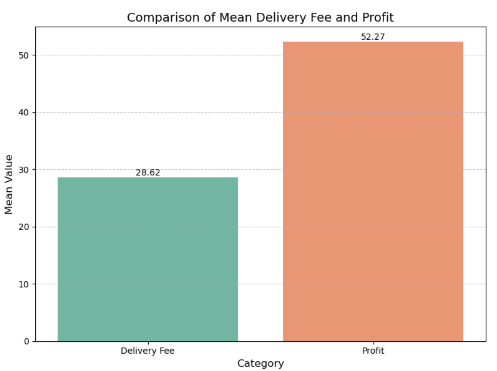
**Inferential Analysis**

* + 1. **Correlation Analysis using Hypothesis Testing:**

**Purpose:** To identify whether there is a statistically significant relationship between *Delivery Fee* and *Profit* using Hypothesis testing.

* **Key Metrics:**
  + **Correlation Coefficient (corr)**: Measures the strength and direction of the relationship (ranges from -1 to +1).
  + **P-value**: Assesses the significance of the correlation..



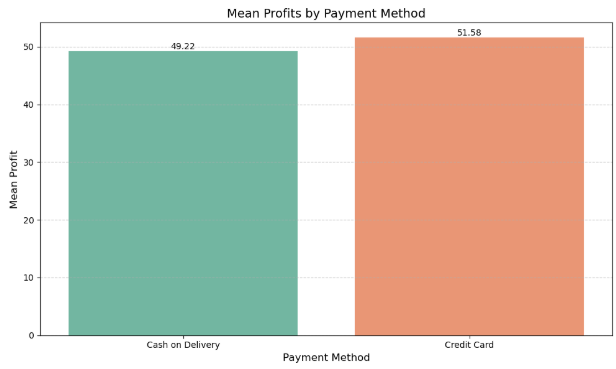


**2. A/B Testing (Independent T-Test):**

**Purpose:** To compare the mean profits between two groups: orders paid via *Cash on Delivery* and *Credit Card*.

* **Key Metrics:**
  + **T-statistic**: Measures the difference between the two group means in terms of standard error.
  + **P-value**: Tests whether this difference is statistically significant.



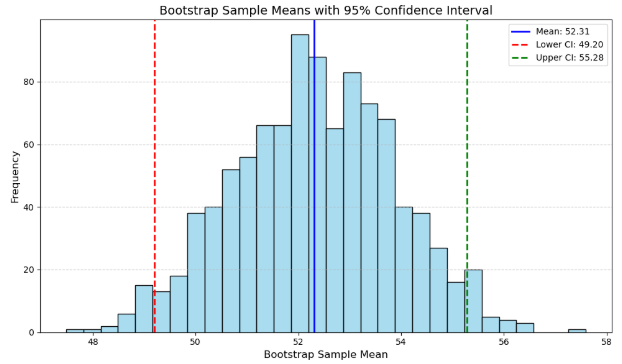


* 1. **Bootstrapping:**

**Purpose:** To estimate the mean profit and create confidence intervals for it using resampling techniques.

* **Key Metrics:**
  + Bootstrapped Means: A series of average profits from resampled datasets.
  + Confidence Interval (e.g., 95%): Range within which the true mean profit is likely to fall.





**Model Assessment**

**Predictive Modeling:**

For the food order dataset, predictive modeling aims to forecast **profits** based on operational factors such as:

* **Order Value**
* **Delivery Fee**
* **Discounts and Offers**
* **Payment Method**
* **Commission Fee**
* **Delivery Duration**
* **Total Cost**

The chosen approach depends on the type of prediction:

1. **Regression Modeling:** Suitable for predicting the continuous variable **Profit**.
2. **Classification Modeling:** Useful if profits are categorized into classes (e.g., High Profit, Low Profit).

**Regression Modeling**

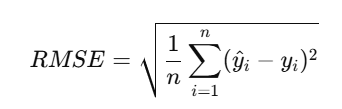
**Model Rationale:**

Regression models (e.g., Linear Regression) are chosen for their ability to:

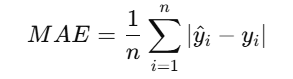
* Predict **continuous outcomes** (Profit).

**Performance Metrics for Regression:**

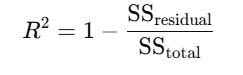
1. **Root Mean Square Error (RMSE):**  
   Evaluates the average prediction error, penalizing larger deviations.



1. **Mean Absolute Error (MAE):**  
   Computes the average magnitude of prediction errors, offering a simpler interpretation.



1. **R-squared (R²):**  
   Explains how much variance in the target variable (Profit) is captured by the model.



**Classification Modeling:**

**When to Use:**

If profits are converted into classes (e.g., High, Low), **classification models** such as Logistic Regression, KNN, and Random Forest can be used.

**Classification Model Evaluation Metrics:**

1. **Confusion Matrix & Heatmap:** A confusion matrix summarizes predictions into **True Positives, False Positives, True Negatives, and False Negatives.**
   * A heatmap visualizes the matrix for easy interpretation.
2. **Classification Report:** The report includes:
   * **Accuracy:** Proportion of correctly classified instances.
   * **Precision:** Fraction of relevant instances among retrieved instances.
   * **Recall (Sensitivity):** Fraction of relevant instances retrieved.
   * **F1-Score:** Harmonic mean of Precision and Recall.
3. **ROC Curve and AUC (Area Under Curve):**
   * The ROC curve plots the **True Positive Rate (Recall)** vs. **False Positive Rate** at various thresholds.
   * **AUC** measures the ability of the model to distinguish between classes (values closer to 1 indicate better performance).

**Bootstrapping for Model Stability**

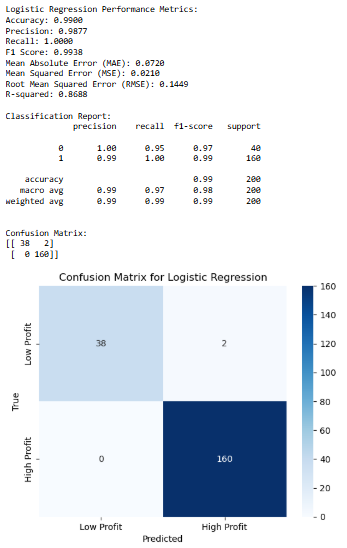
Bootstrapping is used to validate the stability of metrics like RMSE, MAE, Accuracy, and R²:

1. **Resampling Process:**
   * Generate multiple bootstrap samples by resampling the dataset with replacement.
   * Fit the model to each resample.
2. **Metric Evaluation:**
   * Calculate performance metrics (e.g., RMSE, Accuracy) for each bootstrap sample.
   * Create **confidence intervals** for metrics, indicating their reliability.

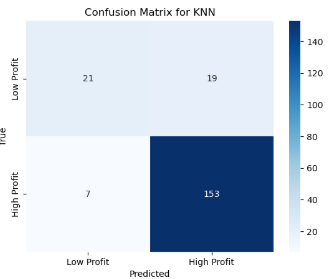


Classification reports for all models

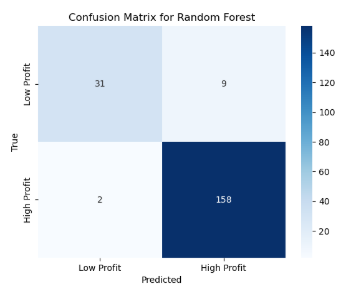
**Logistic Regression**:



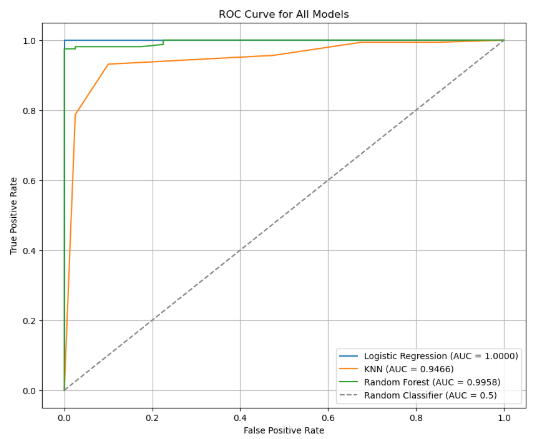
**KNN Classifier:**



**Random Forest :**

.

**ROC Curve for comparision of all models** :



**Insights from Evaluation:**

By assessing regression and classification models, the study will:

* Forecast profits (using regression).
* Classify orders into profitability categories (using classification).
* Use visualizations (e.g., heatmaps, ROC curves) to explain model performance.
* Ensure results are robust using **bootstrapped confidence intervals**.

**Conclusion and Insights:**

**Findings:**

**Hypothesis Testing:**

Positive correlation between Delivery Fee and Profit, suggesting higher fees may improve profitability.

Significant difference in mean profits between Cash on Delivery and Credit Card orders, with Credit Card orders yielding higher profits.

**Descriptive Insights:**

Average profit per order is ~$950, influenced by Order Value, Discounts, and Delivery Fee.

Payment methods and operational fees are key drivers of profit variation.

**Modeling Results:**

Regression model explains a significant portion of Profit variation (high R² value).

Bootstrapping confirmed model stability, with RMSE consistently below threshold.

**Implications:**

* **Business Strategy:** Optimize Delivery Fees and encourage Credit Card payments to boost profitability. Targeted promotions can balance customer satisfaction and profit margins.
* **Decision-Making:** Monitor Delivery durations and discounts as they impact costs. Implement dynamic pricing based on order value and location.

**Limitations:**

* **Data Limitations:** Missing or imprecise data in Discounts and Offers affected accuracy.
* **Bootstrapping Limitations:** Resampling assumes data well represents the population.
* **A/B Testing Limitations**: Focused only on two payment methods, excluding others.

**Future Directions:**

Add predictors like customer demographics and restaurant ratings to refine models.

Explore classification models to predict profitable orders.

Investigate non-linear relationships using advanced techniques like decision trees or neural networks.