PROJECT REPORT

Brewing Success - Classifying Coffee Shop Revenue Patterns

1. Introduction

In today's competitive market, coffee shops must understand the key drivers of revenue to optimise operations and improve profitability. This project focuses on developing a classification model that predicts daily revenue categories based on various business metrics. By identifying key factors that influence revenue fluctuations, this model will help coffee shop owners and managers make informed decisions about staffing, inventory management, and marketing strategies.

2. Problem Statement

The objective of this project is to classify coffee shop revenue patterns into four categories:

- **Low:** Revenue below $1000
- **Medium:** Revenue between $1000 and $2000
- **High:** Revenue between $2000 and $3500
- **Very High:** Revenue over $3500

By identifying and analyzing the key factors that drive revenue fluctuations, this model will provide valuable insights for coffee shop owners and managers. It will enable them to make more informed and data-driven decisions regarding staffing levels, inventory optimization, and targeted marketing strategies. Moreover, understanding these dynamics will help improve operational efficiency, enhance customer experience, and ultimately increase profitability. By leveraging the model's predictions, businesses can proactively address potential issues and capitalise on opportunities to maximize revenue growth.

3. Dataset Overview

The dataset, sourced from Kaggle, contains various socio-economic and business indicators that influence daily revenue.

Source:
https://www.kaggle.com/datasets/himelsarder/coffee-shop-daily-revenue-prediction-dataset

Features:

1. Number of Customers Per Day: The total number of customers visiting the coffee shop on any given day.
2. Average Order Value ($): The average dollar amount spent by a customer during their visit.
3. Operating Hours Per Day: The total number of hours the coffee shop is open for business per day.
4. Number of Employees: The number of employees working on a given day. This can influence service speed, customer satisfaction, and ultimately, sales.

5. Marketing Spend Per Day ($): The amount of money spent on marketing campaigns or promotions on any given day.
6. Location Foot Traffic (people/hour): The number of people passing by the coffee shop per hour. It is a variable indicative of the shop's location and its potential to attract customers.
7. Daily Revenue ($): The total revenue generated by the coffee shop each day.It is calculated as a combination of customer visits, average spending, and other operational factors like marketing spend and staff availability.

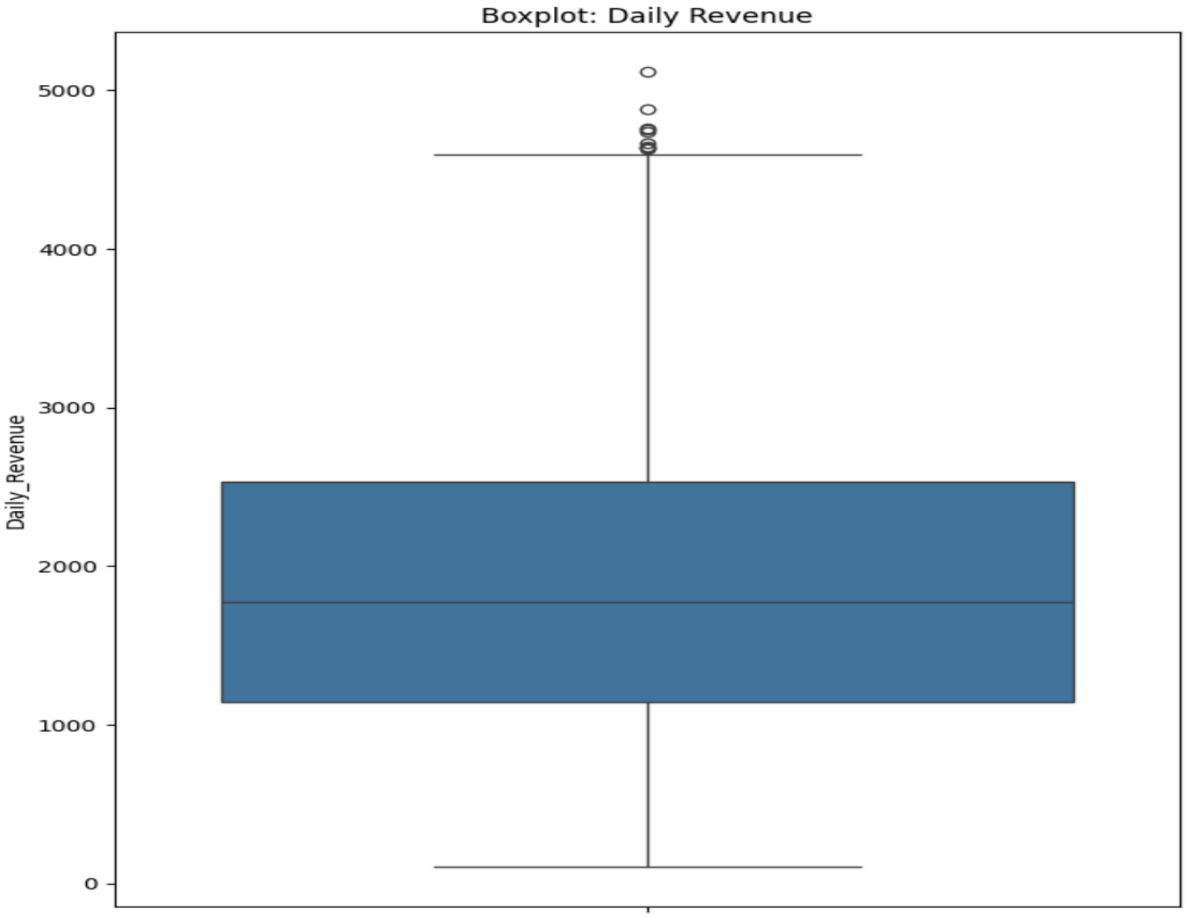
4. Data Wrangling

We are working with a compact dataset consisting of just 7 features and 2,000 entries. All the columns belong to a numeric datatype with no null values and no duplicate rows, which is ideal for the modelling.

Statistical analysis reveals an outlier value in the revenue that is highly unlikely to be correct. Therefore that row has been dropped, leaving our dataset at 1999 rows and 7 columns.

BOXPLOTS:

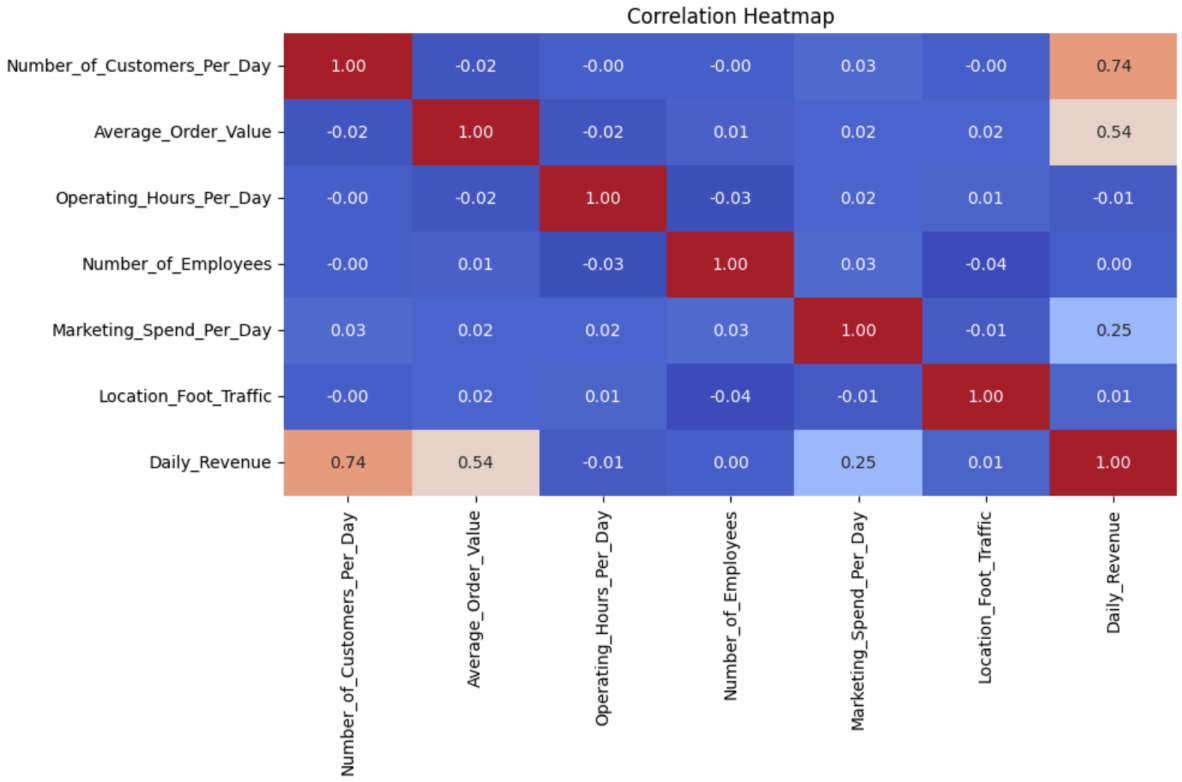
1. **Revenue**



The boxplot for our target variable suggests that most of the values for daily revenue are between $1000 and $3000 (IQR), but there are values that are outliers in our data. Those can be considered as the very high revenue class, to understand their patterns. Although outliers, they are an important part of this analysis and thus not considered to be dropped.

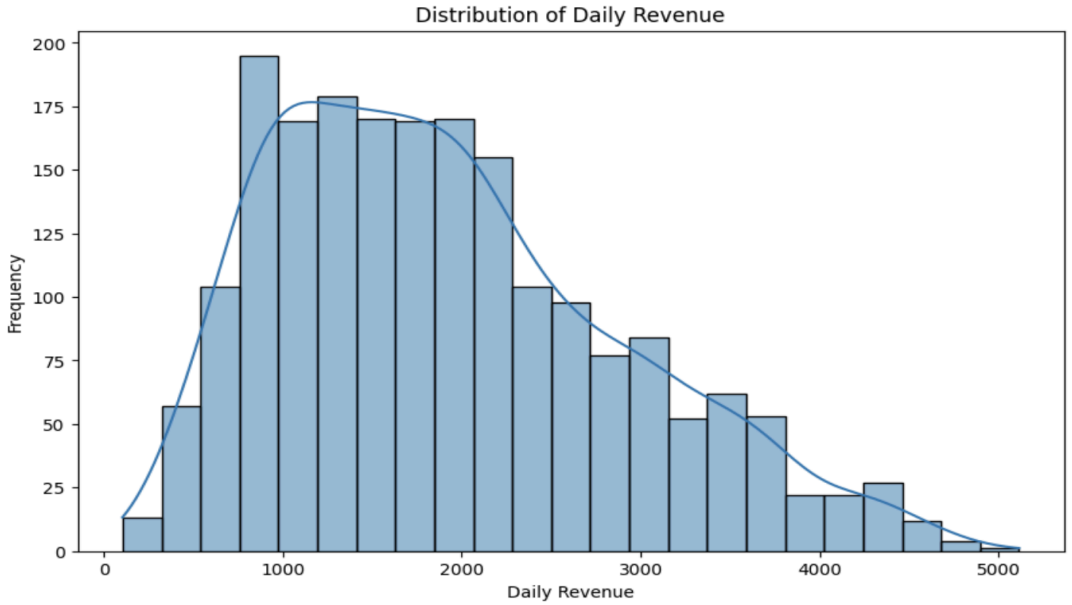
5. Exploratory Data Analysis (EDA)

1. Correlation Heatmap



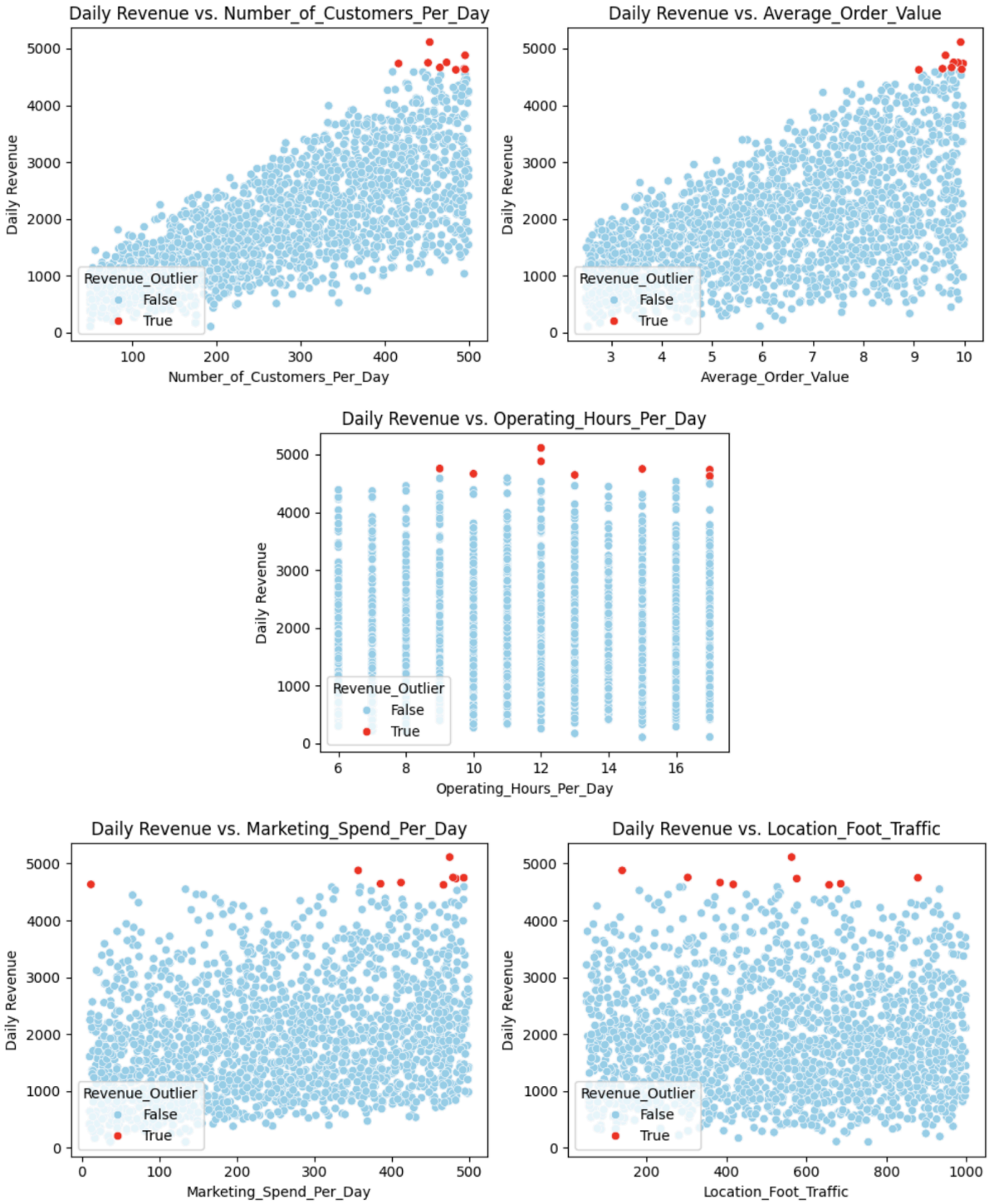
It can be clearly seen that the correlation between various features is very low, except for a few.
The daily revenue is highly correlated with the number of customers per day, which is a sensible correlation. It is also affected by the average order value and the money spent on marketing to some extent.

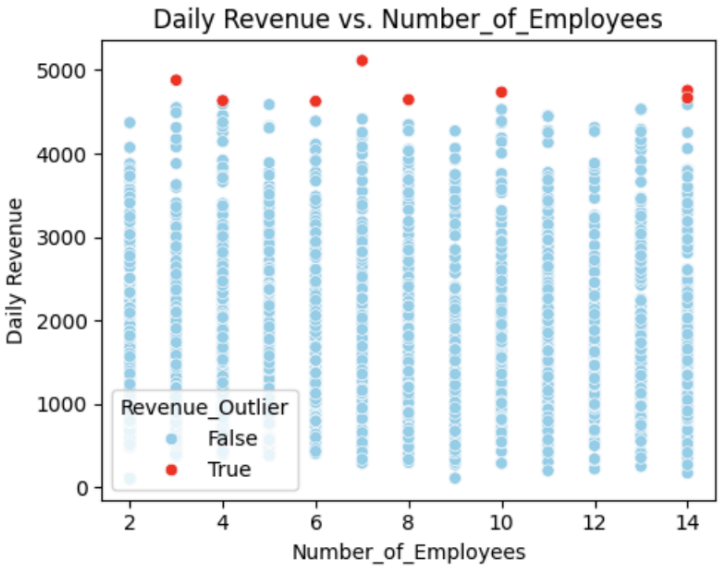
2. Revenue Distribution



The values of daily revenue seem to be mostly right-skewed based on the normal distribution plot. Most of the daily revenue is between $1000 and $2500. There are some values that represent extremely low and high daily revenue days and understanding the reason behind it could be how we can improve on the average daily turnover.

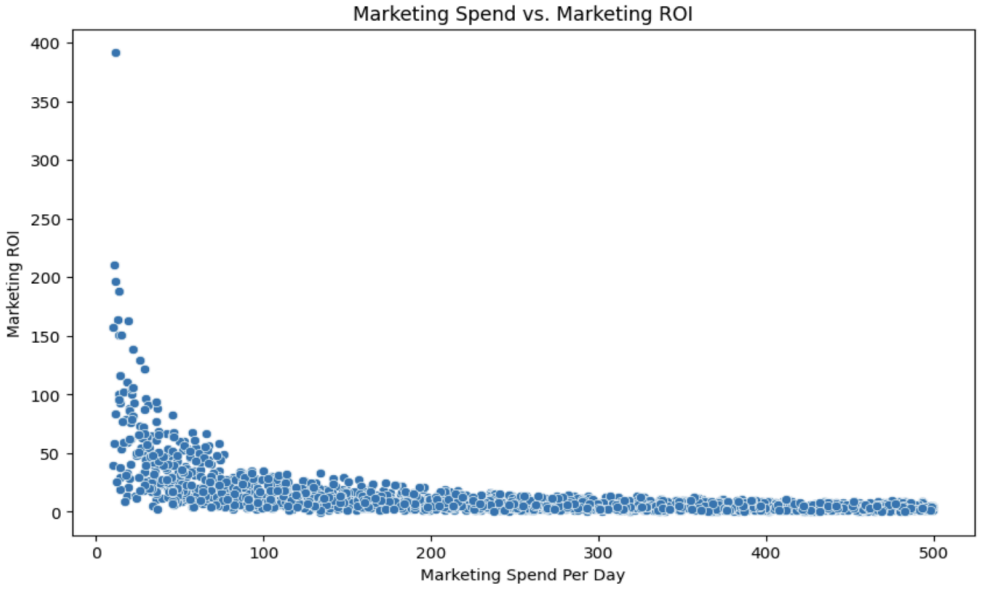
3. Daily Revenue and Understanding Revenue Outliers





a. **Operating Hours:** Outliers are spread across different operating hours. Operating longer doesn't directly create extreme revenue days. Other factors (e.g., more customers or higher Average order value) are likely more influential.
b. **Marketing Spend:** High-revenue outliers coincide with higher marketing spend (~400-500 dollars). However, there are also many non-outlier days with similar marketing spend. This suggests that marketing alone doesn't guarantee record-breaking revenue. Other factors (e.g., foot traffic and order value) likely play a bigger role. We can definitely make a mark of controlling the marketing spend.
c. **Location Foot Traffic:** Outliers appear at varying foot traffic levels (not just the highest). This means that foot traffic alone doesn't define extreme revenue days. A combination of high Average Order Value and high customer count is more critical.
d. **Number of Employees:** Revenue outliers don't consistently align with high employee counts either. This suggests that adding more employees doesn't necessarily boost revenue to outlier levels. Limiting the number of employees to save money can be considered as well.

4. Return on Investment (ROI) vs. Marketing spend



It can be clearly seen that As marketing spending increases, ROI drops sharply. Beyond 100–150 Dollars per day, most ROI values are below 10, meaning that spending more doesn't significantly increase revenue. Therefore, suggestions can be made to limit our marketing spend to a maximum of 150 Dollars.

6. Feature Engineering

This step involves **transforming raw data into a structured, meaningful format** by normalizing and scaling features, encoding categorical variables, and creating new informative variables from the already existing ones to improve model performance and interpretability.

1. Created Interaction Features: Some features were created For analysing the revenue classes, we might require some other variables that will more appropriately describe our revenue classes:
   a. Revenue per customer, to understand how much value a single customer adds to the revenue.
   b. Marketing Efficiency, to get a sense of how much we can spend on the marketing area, so as to be more efficient and have a reasonable return on investment.
   c. Employee Productivity could be useful for understanding factors such as the number of orders processed per employee, speed of service, or overall sales generated per staff member.
   d. Foot Traffic Conversion does also help understand revenue turnovers as well. A higher value indicates that a larger proportion of passersby are becoming paying customers, suggesting good location appeal, marketing effectiveness, or store attractiveness. A lower value may indicate that despite high foot traffic, fewer people are converting into customers, possibly due to pricing, competition, or lack of engagement strategies.

2. Normalized Numerical Features: Scaling features ensures that all variables have a comparable range, preventing models from being biased toward larger numerical values. This is especially important for distance-based algorithms like KNN and SVM, as well as gradient-based models for faster and more stable convergence. Scaling has been performed on the features in this dataset. The target variable has not been included for it needs to be encoded into classes and separated from the features for the purpose of modelling.

3. Categorized and Encoded Target Variable: The target variable, Daily_Revenue into 4 classes Low, Medium, High and Very High. The classification ins done based on the revenue values and not the count based division. Very High class has been included to understand the behaviour based on higher range of revenue values that are over $3500. After that the classes are encoded into numeric values for the purpose of modelling.

4. Train-Test Split: The dataset needs to be split into two datasets at random sets to first train the model using the training dataset and then to evaluate the model performance using the test dataset. For this project, the dataset has been split with 80% of the data used for training and 20% for evaluating the performance of the trained model.

7. Model Selection & Training

For classifying **coffee shop revenue patterns**, the following models are used:

- **XGBoost:** A powerful gradient-boosting algorithm that builds multiple decision trees sequentially, optimizing performance while reducing overfitting.
- **Random Forest:** An ensemble of decision trees that improves accuracy and robustness by averaging multiple predictions.
- **K-Nearest Neighbors (KNN):** A distance-based algorithm that classifies data points based on the majority class of their nearest neighbours.
- **Support Vector Machine (SVM):** A model that finds the optimal hyperplane to separate revenue categories, effective for high-dimensional data.

Each model has been evaluated to determine the best fit for classifying revenue levels.

8. Model Evaluation & Comparison

Performance metrics used to compare and evaluate the model performances are:

- **Accuracy:** Measures the overall correctness of the model. Useful when the dataset is balanced but may be misleading for imbalanced classes.

  Accuracy = Correct Predictions ÷ Total Predictions

- **Precision:** Indicates how many predicted positive cases are actually correct. High precision means fewer false positives, which is crucial when misclassification is costly.

  Precision = (True Positives) ÷ (True Positives + False Positives)

- **Recall (Sensitivity):** Measures how well the model captures actual positive cases. High recall means fewer false negatives, which is important when missing a positive case is critical.

  Recall = (True Positives) ÷ (True Positives + False Negatives)

- **F1-Score:** The harmonic mean of precision and recall, balancing both metrics. Useful when there is an imbalance between precision and recall.

  F1-Score = 2 × (Precision × Recall) ÷ (Precision + Recall)

Each metric provides unique insights, ensuring a comprehensive evaluation of the model's performance.

COMPARISON TABLE :

	Model	Accuracy	Precision	Recall	F1-Score
0	XGBoost	0.950	0.950355	0.950	0.950086
1	Random Forest	0.910	0.910767	0.910	0.910246
2	K-Nearest Neighbors	0.830	0.830696	0.830	0.829600
3	SVM	0.915	0.915794	0.915	0.914969

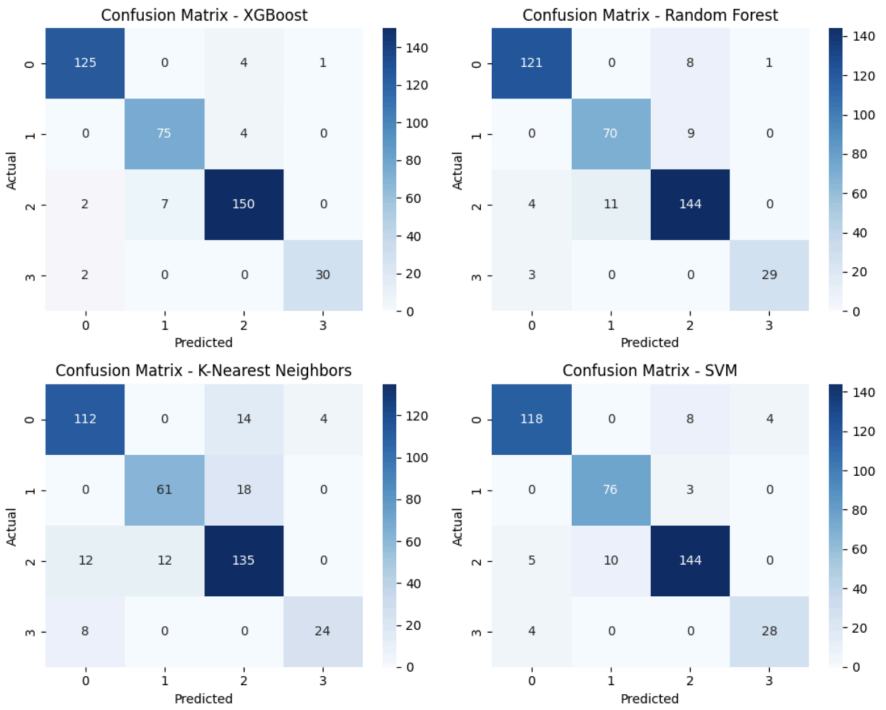
KEY INSIGHTS:

XGBoost outperforms all models, achieving the highest **accuracy (95%)**, **precision**, **recall**, and **F1-score**, making it the best choice for classifying revenue patterns.

SVM performs slightly better than Random Forest (91.5% vs. 91.0%), suggesting it handles decision boundaries well in this dataset.

KNN has the lowest performance (83%), indicating it struggles with distinguishing revenue categories effectively, possibly due to feature scaling or data distribution.

CONFUSION MATRIX



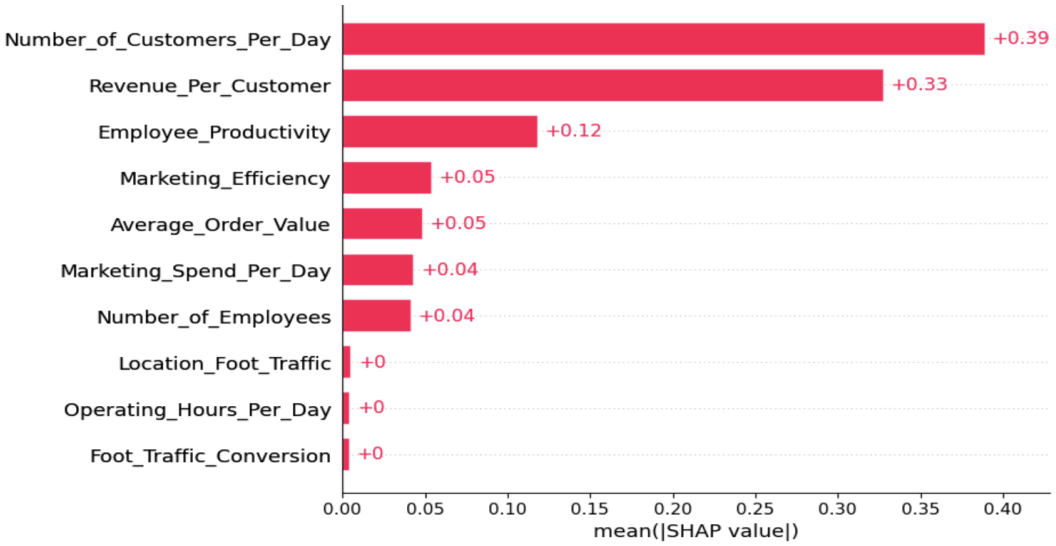
The **confusion matrix is essential** for diagnosing classification errors, improving model performance, and selecting the best algorithm for a given problem.

In this project, **XGBoost** performs best, making minimal misclassifications. It correctly predicts most High (0), Low (1), and Medium (2) revenue cases.

Random Forest is slightly less accurate, with more misclassification between Medium (2) and High (0) revenue classes.
K-Nearest Neighbors (KNN) struggles the most, misclassifying High (0) as Medium (2) and vice versa, indicating that it has difficulty distinguishing between these revenue levels.
SVM performs better than KNN and is close to Random Forest, but still has misclassification between Medium (2) and High (0) revenue categories.

CONCLUSION: XGBoost is the most effective model, with the best balance of precision, recall, and accuracy. Random Forest and SVM are strong alternatives, but they require further tuning to reduce misclassification between revenue classes. KNN is not well-suited for this dataset, as it struggles with class separation. The confusion matrix helps identify specific misclassifications, guiding potential improvements like feature engineering or hyperparameter tuning.

9. Feature Importance & SHAP Analysis



This is a SHAP (SHapley Additive exPlanations) summary bar chart that shows the average importance of each feature in predicting the target variable (likely revenue class in your case). The SHAP values quantify the contribution of each feature to the model's predictions.

Key Observations from the SHAP Graph:

1. **Number of Customers Per Day (+0.39 SHAP value)**
   This is the most influential factor in determining revenue patterns. A higher number of daily customers strongly impacts revenue classification.
2. **Revenue Per Customer (+0.33 SHAP value)**
   The second most important factor. Higher revenue per customer significantly boosts overall revenue.
3. **Employee Productivity (+0.12 SHAP value)**
   Employee efficiency contributes moderately to revenue. More productive employees likely help serve more customers or improve service quality.
4. **Marketing Efficiency, Average Order Value, Marketing Spend Per Day, and Number of Employees (SHAP values ~0.04 - 0.05)**

These features have a minor but noticeable impact. Good marketing efficiency and a higher average order value can contribute positively to revenue.

5. **Location Foot Traffic, Operating Hours Per Day, and Foot Traffic Conversion (+0 SHAP value)**
   These features have little to no impact in the current model, suggesting they might not be strong indicators of revenue patterns or their effects are already captured by other factors.

Conclusion:

The model relies most on daily customer count and revenue per customer to classify revenue levels. Employee productivity plays a secondary role, while marketing efforts and order value provide marginal contributions. Surprisingly, foot traffic and operating hours do not seem to influence revenue classification significantly, indicating potential data limitations or redundancy with other features.

10. Conclusion & Recommendations

Findings:

1. XGBoost is the best-performing model with 95% accuracy.
2. Employee Productivity, Revenue Per Customer, and Customer Count drive revenue.
3. Marketing Efficiency matters more than raw marketing spend.
4. Foot Traffic alone does not significantly impact revenue unless converted into sales.

Business Recommendations:

1. **Optimize workforce efficiency** to maximize daily revenue.
2. **Focus on customer spending** strategies rather than increasing foot traffic alone.
3. **Enhance marketing efficiency** instead of increasing the budget.
4. **Improve order value per customer** by promoting premium offerings.