RESTAURANT REVENUE PREDICTION

Insights from Data analysis

Group 10

Chathurangi Nethmini – 16012

Kusara Udayana – 16073

Amindu Yohan - 16385

Abstract

The restaurant industry operates in a highly competitive environment where revenue generation is influenced by a wide range of factors. Understanding these factors is essential for making informed business decisions and achieving financial success. This project is divided into two main parts. In the first part, we conduct a comprehensive analysis to uncover the key variables that significantly affect restaurant revenue. Based on the insights gained, we provide practical recommendations to help restaurant owners improve their earnings. In the second part, we focus on building a predictive model that estimates restaurant revenue using the identified influential factors. This model aims to assist both existing restaurant owners and aspiring entrepreneurs in making more accurate revenue forecasts, thereby supporting strategic planning and investment decisions.

Table of Contents

List of Figures

List of Tables

Introduction

In today's fast-paced world, dining out is more than just a meal, it's an experience. Behind every successful restaurant lies not only delicious food but also smart decisions driven by data. Yet, many restaurant owners find themselves struggling to understand what truly drives their revenue. Is it the location, the size of the seating area, or the chef's experience? With rising competition and customer expectations, making blind decisions can be risky. This project steps into the world of restaurant data to uncover the hidden patterns behind high-performing businesses. By combining insightful analysis with predictive modeling, we aim to turn raw data into a powerful guide for current and future restaurant owners seeking to boost their revenue with confidence and clarity.

Description of the Question

- **What are the factors that affect the revenue of a restaurant?**
  This includes identifying the most influential features such as location, seating capacity, marketing efforts, customer engagement, and service quality that contribute to revenue performance.
- **What kind of improvements can be made to increase revenue for existing restaurants?**
  We aim to suggest actionable recommendations based on data analysis, such as optimizing seating layouts, targeting high-impact marketing channels, or enhancing service experiences.
- **What are some new startup ideas for restaurants that could lead to high revenue generation?**
  By studying the data and trends, we explore innovative restaurant concepts and business models that are more likely to succeed and generate higher revenue in today's market.

Description of the Dataset

This project was conducted using a publicly available dataset found on Kaggle, titled "Restaurant Revenue Prediction Dataset". The dataset contains 8,368 unique records and includes 17 variables. The description of each variable used in this project is presented in the table below.

Variable Name	Variable Type	Description
Name	Quantitative	Index number of the restaurant.
Location	Qualitative (Ordinal)	weather restaurant located in rural, suburban or downtown.
Cuisine	Qualitative (Nominal)	The type of cuisine offered (Japanese, Mexican, Italian, Indian, French, American)
Rating	Quantitative	The average rating of the restaurant
Seating Capacity	Quantitative	The number of seats available in the restaurant
Average Meal Price	Quantitative	The average price of a meal at the restaurant

Marketing Budget	Quantitative	The marketing budget allocated for the restaurant
Social Media Followers	Quantitative	The number of social media followers.
Chef Experience Years	Quantitative	The number of years of experience of the head chef
Number of Reviews	Quantitative	The total number of reviews the restaurant has received.
Avg Review Length	Quantitative	The average length of reviews.
Ambience Score	Quantitative	A score representing the ambience of the restaurant.
Service Quality Score	Quantitative	A score representing the quality of service.
Parking Availability	Qualitative (Nominal)	Indicates if parking is available (Yes/No)
Weekend Reservations	Quantitative	The number of reservations made on weekends
Weekday Reservations	Quantitative	The number of reservations made on weekdays
Revenue	Quantitative	The total revenue generated by the restaurant.

Table 1- Description of the Dataset

Data Pre-processing

The dataset was checked for duplicates and missing values. No duplicate records or missing values were found. Using the Isolation Forest algorithm, 59 outliers were identified**.** However, no data points were removed.

Feature Engineering

- **Meal Price Category**: A new categorical variable was derived using the Average Meal Price, making it easier to group restaurants by pricing levels and provide targeted suggestions for revenue improvement.
- **Review Quality Score**: This variable was created by combining the Rating and Average Review Length, reflecting both the quality and depth of customer feedback.
    - Review Quality Score = Rating × Average Review Length
- Dataset spitted randomly 70% as training for the descriptive analysis and model training, 30% as testing for model evaluation.

Main Results of Descriptive Analysis

Cluster Analysis

To identify similar restaurant groups and simplify revenue analysis, we performed cluster analysis on the dataset.

- To identify groups of similar restaurants, we applied K-Prototypes clustering, suitable for mixed data types. Using the elbow method and minimum gamma value, we found the optimal number of clusters to be two. (the-k-prototype-as-clustering-algorithm-for-mixed-data-type-categorical-and-numerical, n.d.)

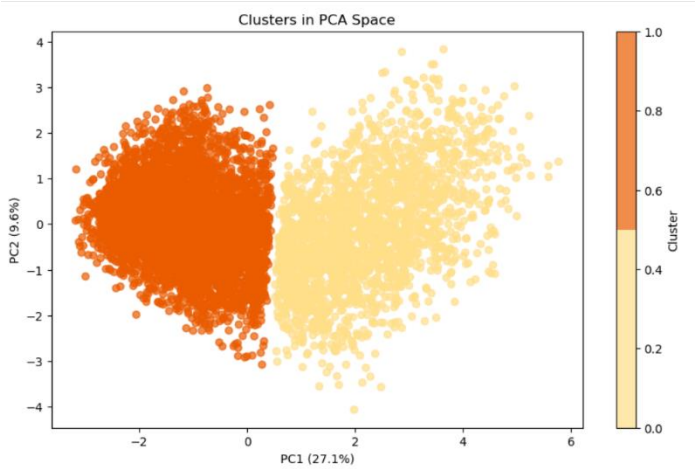

Figure 1- PCA plot to display clusters

- We then performed Principal Component Analysis and plotted the PCA scores, coloring the points by cluster. The visualization confirmed the presence of two distinct clusters. (Holbert, 2023)
- To assess the meaningfulness of the clusters, we conducted Chi-square tests of association between the clusters and each categorical variable. At a 5% significance level, a significant association was found between Cluster and Location. Cramér's V indicated a strong association (value = 0.8739) between cluster and location.
- Since Rural and Suburban restaurants mainly fell into one cluster and Downtown into the other, we named the clusters Nonurban and Urban.

Descriptive Analysis for Urban Restaurants

- The revenue values show a wide range and non-normal distribution. To handle this, we applied a log transformation. Although it did not fully normalize the distribution (as shown in the first plot), it helped manage the range, so log revenue was used for further analysis.

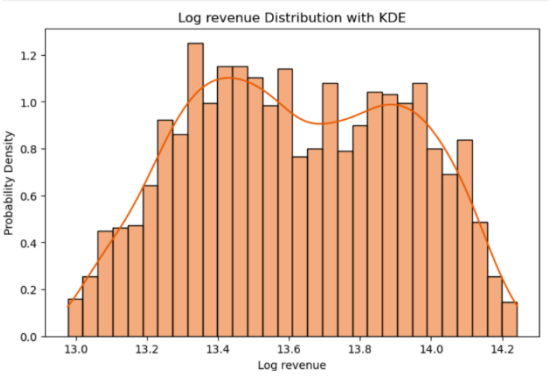

Figure 2- Distribution of log Revenue for Urban restaurants

- Japanese and French cuisines have the highest median log revenues, while Mexican and Indian have the lowest. A Kruskal-Wallis test confirmed that these median differences are statistically significant at the 5% level.
- There is a slight positive correlation between the number of seats and log revenue, indicating that larger restaurants tend to earn more.
- Marketing budget and social media followers are strongly positively correlated with each other, suggesting that restaurants investing more in marketing also tend to have a higher social media presence.

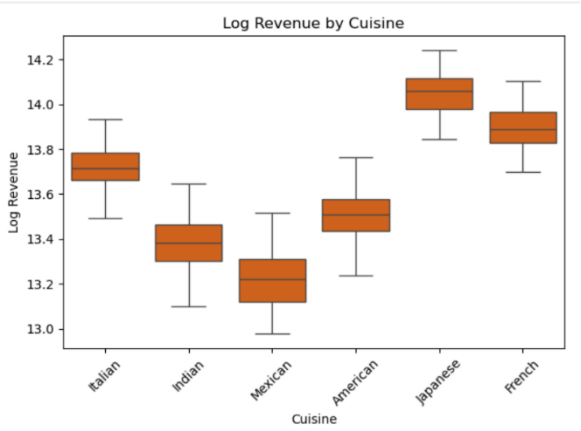

Figure 3- Box plots of log revenue by Cuisine Urban restaurants

Descriptive Analysis for Non-urban Restaurants

- The original revenue data in non-urban restaurants had a wide, non-normal distribution. A log transformation reduced the range and improved manageability, so **log revenue** was used for further analysis.
- Japanese and French cuisines have the highest median log revenues, while Mexican and Indian have the lowest. A Kruskal-Wallis test confirmed that these median differences are statistically significant at the 5% level.
- The scatter plot shows a positive relationship between seating capacity and log revenue, supported by a Pearson correlation of 0.60. This indicates that larger seating areas are moderately associated with higher revenue.

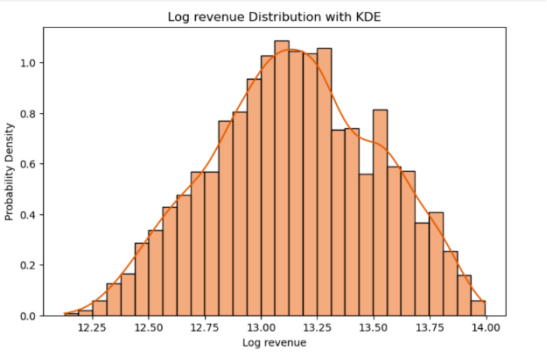

Figure 4-Distribution of log Revenue for Non-urban restaurants

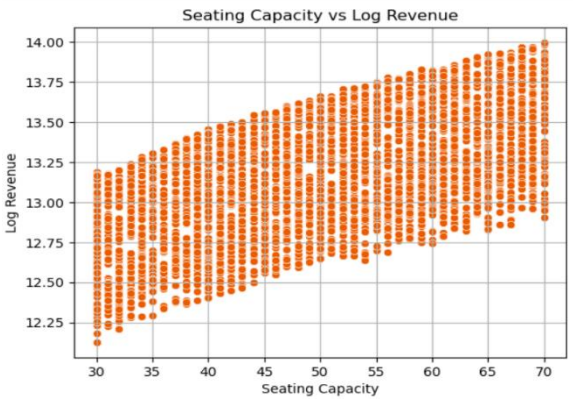

Figure 5- Scatterplot for log revenue vs Seating Capacity for non-urban

Main Results of Advanced Analysis

To better capture the variation in revenue, we fitted models separately for the two identified clusters: urban and non-urban. We began the modeling process with multiple linear regression. However, due to the non-normality observed in the log-transformed revenue and multicollinearity issues identified during the descriptive analysis, we explored regularized linear models such as Lasso, Ridge, and Elastic Net. Since the performance of all three models was nearly identical, we decided to include only the Elastic Net model in this report for simplicity and consistency. To further improve model performance and account for potential non-linear relationships between predictors and revenue, we also implemented tree-based methods, including Random Forest and XGBoost. The table below summarizes the performance of each model across both clusters.

MODEL	Urban Restaurants				Non-urban Restaurants			
	Train		Test		Train		Test	
	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2
MLR	0.0037	0.9591	0.0038	0.9568	0.0042	0.9681	0.0045	0.9647
Elastic-net	0.0037	0.9591	0.0038	0.9575	0.0042	0.9682	0.0045	0.9647
Random-forest	0.0017	0.9807	0.0039	0.9559	0.0019	0.9852	0.0041	0.9677
XGBoost	0.0031	0.9657	0.0037	0.9577	0.0027	0.9797	0.0040	0.9687

Table 2- Summary of Model performance

Best Model

Across both urban and non-urban clusters, all models demonstrated similar performance. We selected the Elastic Net model as the best choice for both clusters, as it effectively handles multicollinearity and performs feature selection. While Multiple Linear Regression offers higher interpretability, it was not chosen due to issues such as non-normality in the target variable and its inability to handle multicollinearity. Additionally, Elastic Net was preferred over ensemble methods due to its greater computational efficiency and better interpretability. The following are the feature importance graphs for the two clusters.

For Urban (Best model -Elastic net)

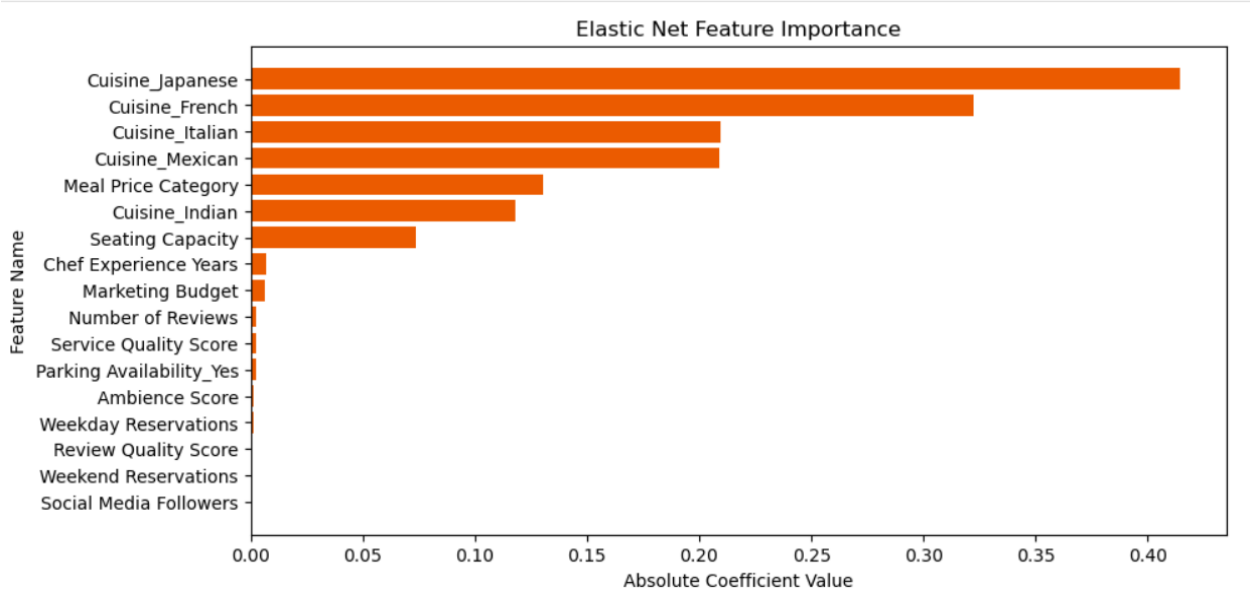

Figure 6-Feature importance graph for Urban Restaurants

The Elastic Net feature importance graph for the urban cluster shows that cuisine type - especially Japanese and French has the greatest impact on predictions, followed by Italian, Mexican, meal price, Indian cuisine, and seating capacity. In contrast, customer feedback and promotional factors have minimal influence, highlighting the importance of cuisine and pricing in urban restaurant performance.

For non-urban (Best model -Elastic net)

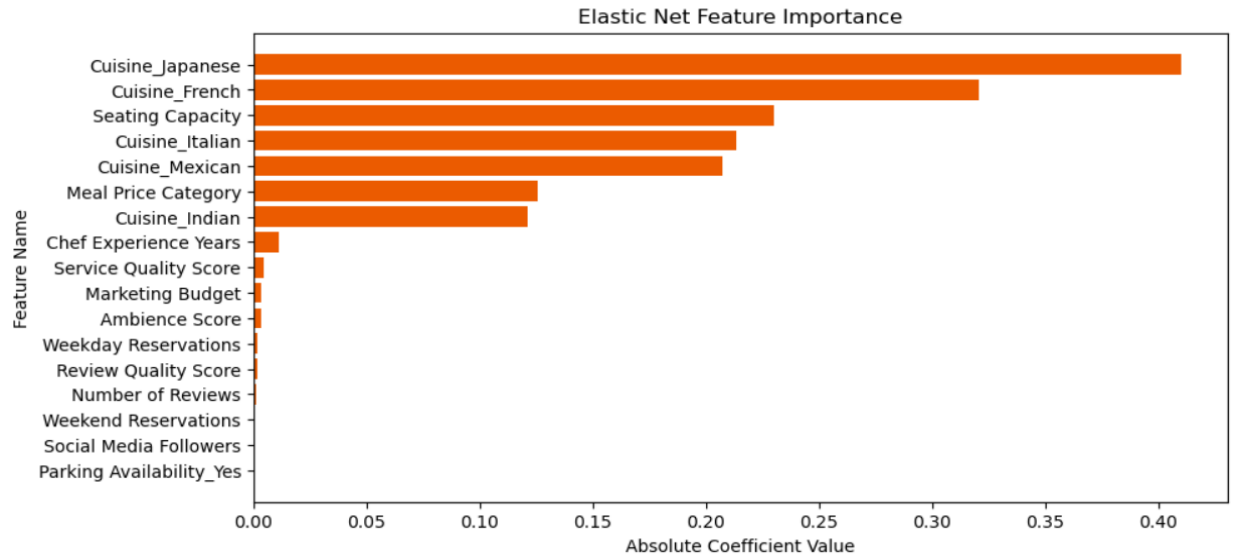

Figure 7- Feature importance graph for Non-Urban Restaurants

The Elastic Net feature importance graph for the non-urban cluster shows that cuisine type - especially Japanese and French along with seating capacity and meal price, are key drivers of restaurant performance. In contrast, service, location and marketing factors have a minimal impact.

Issues encountered and proposed solutions

- **Clustering with Mixed Data**: K-Means requires numerical variables. Since our dataset includes categorical data, we used K-Prototypes, which handles both types.
- **PCA with Ordinal Variables**: PCA isn't ideal for ordinal data. We ordinal encoded those variables and treated them as numeric without standardization.

Discussion and Conclusion

Feature Reduction and Model Performance

Some variables showed little to no impact on revenue prediction, so they were removed to simplify the model. After fitting the Elastic Net model with the reduced feature set, performance remained nearly the same. Therefore, we decided to use the simpler, reduced model for prediction due to its efficiency and interpretability.

Urban Cluster

We refitted the model using the following features: Cuisine Type, Seating Capacity, Meal Price Category, Chef Experience, and Marketing Budget. The partial dependence plots for the best model in the urban cluster reveal that Japanese, French, and Italian cuisines have a strong positive influence on revenue. Meal price category and seating capacity also contribute positively. In contrast, Mexican and Indian cuisines show a slight negative relationship, while chef experience and marketing budget have minimal effects.

Non-Urban Cluster

We refitted the model using the following features: Cuisine Type, Seating Capacity, Meal Price Category, Chef Experience, Location and Service quality score. The partial dependence plots for the non-urban cluster show that Japanese, French, and Italian cuisines positively influence revenue, along with seating capacity and meal price category. In contrast, Mexican and Indian cuisines have

a slight negative impact. The location variable shows minimal effect, likely due to the dataset already being clustered. Also, chef experience has minimal effects. These insights highlight key revenue drivers for non-urban restaurants.

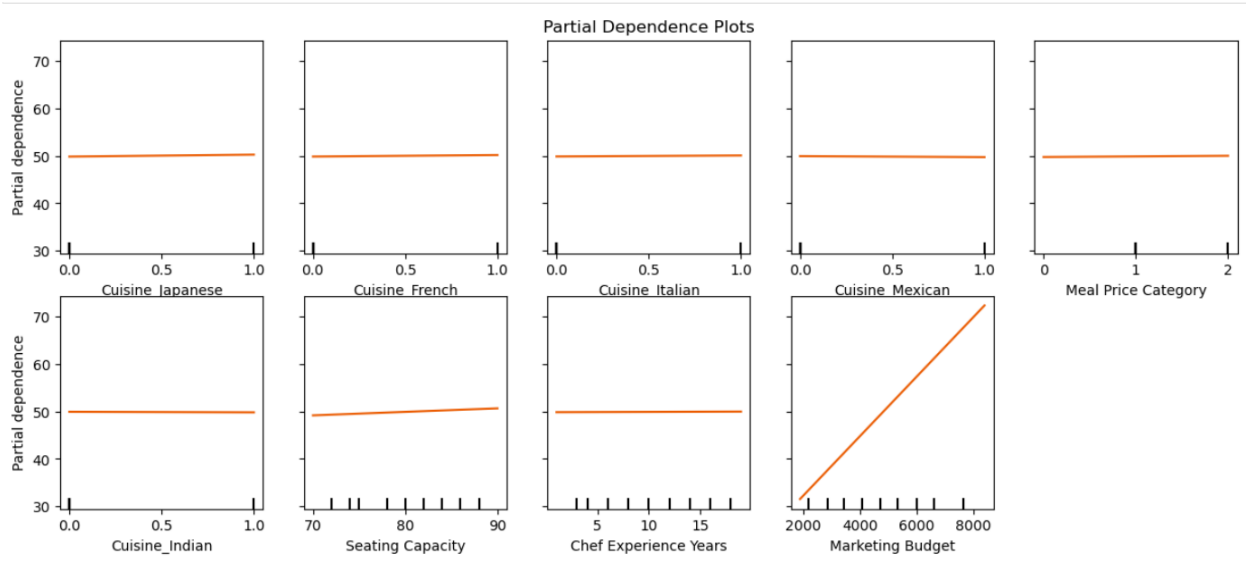

Figure 8- Partial Dependency Plots obtained from elastic net for urban cluster

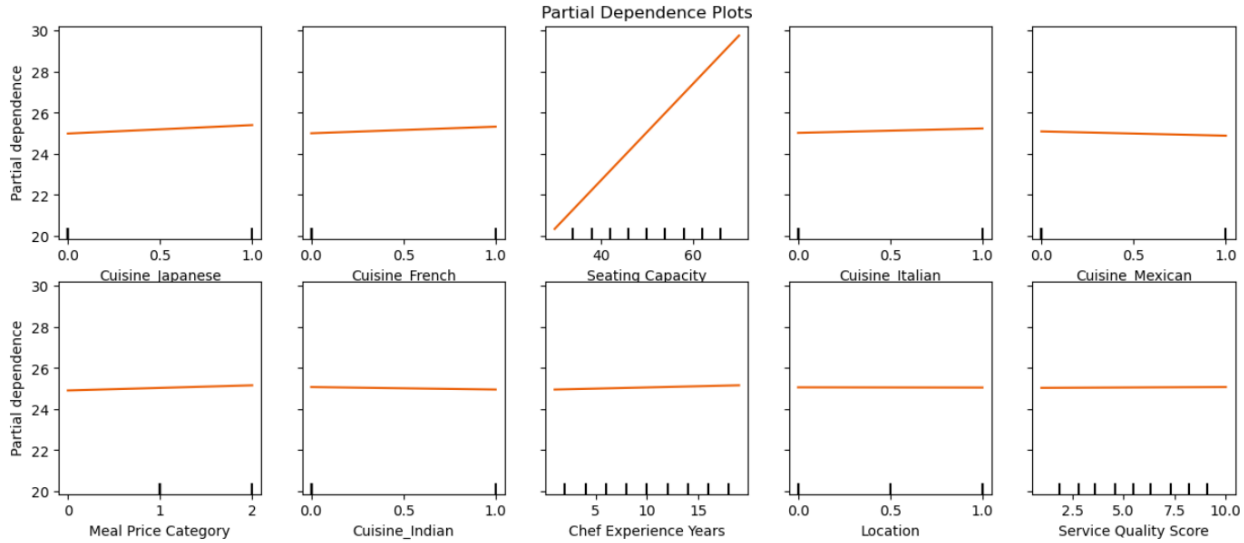

Figure 9- Partial Dependency Plots obtained from elastic net for non-urban cluster

Appendix

https://drive.google.com/drive/folders/1TLuYvCMzHZqNObcM_GEirNCaYQi8eavs?usp=sharing

References

Holbert, C. (2023, 08). *Clustering on Principal Component Analysis*. Retrieved from www.cfholbert.com: https://www.cfholbert.com/blog/cluster-pca/

the-k-prototype-as-clustering-algorithm-for-mixed-data-type-categorical-and-numerical. (n.d.). Retrieved from towardsdatascience.com: https://towardsdatascience.com/the-k-prototype-as-clustering-algorithm-for-mixed-data-type-categorical-and-numerical-fe7c50538ebb/