

A

Project Report

On

**“****Price Optimization and Customer Satisfaction Prediction”**

**DA 204o Data Science in Practice**(August 2024 Term)  
  
  
**Authors and Email Addresses:**   
Barani Ranjan S: baranis@iisc.ac.in  
Ishika Saxena ishikasaxena@iisc.ac.in  
Prasanna Kumar B V: prasannabv@iisc.ac.in  
Siva S: sivas12@iisc.ac.in

MTech (Online) - DSBA  
Indian Institute of Science Bangalore

**Project Supervisor**Pandarasamy Arjunan

**Index**

1. **Introduction**  
   1.1. Project Objective  
   1.2. Significance of the Study
2. **Dataset Overview**  
   2.1. Dataset Description  
   2.2. Key Features  
   2.2.1. Input Variables  
   2.2.2. Target Variables  
   2.3. Preprocessing Steps
3. **Problem Statement**  
   3.1. Restaurant Industry Challenges  
   3.2. How Data Science Can Solve the Problem  
   3.2.1. Regression Model for Price Optimization  
   3.2.2. Classification Model for Predicting Customer Satisfaction  
   3.2.3. Exploratory Data Analysis (EDA)
4. **Methodology**  
   4.1. Classification Model (XGBClassifier)  
   4.1.1. Objective and Approach  
   4.1.2. Model Setup and Data Splitting  
   4.1.3. Hyperparameter Tuning  
   4.1.4. Model Evaluation and Results  
   4.2. Regression Model for Pricing Optimization  
   4.2.1. Objective and Approach  
   4.2.2. Model Evaluation and Results
5. **Results and Insights**  
   5.1. Classification Model Results  
   5.2. Regression Model Results  
   5.3. Key Insights  
   5.3.1. Customer Satisfaction Insights  
   5.3.2. Pricing Insights  
   5.4. Recommendations for Restaurants
6. **Challenges**
7. **Introduction**

**1.1 Project Objective**

The restaurant industry is increasingly competitive, with customer satisfaction playing a pivotal role in determining success. This project utilizes data science techniques to address two main challenges that restaurants face:

* **Predicting Customer Satisfaction:** Using restaurant attributes such as pricing, delivery time, and cuisine type, the project aims to predict whether a customer will rate the restaurant highly (rating ≥ 4) or not (rating < 4).
* **Optimizing Pricing Strategies:** The project aims to recommend optimal price ranges that can maximize customer satisfaction while maintaining profitability, taking into account factors like cuisine, delivery time, and location.

**1.2 Significance of the Study**

* **Enhancing Customer Experience:** By predicting customer satisfaction, restaurants can better understand the factors that lead to high ratings, enabling them to refine their offerings and improve overall customer experience.
* **Optimizing Revenue Strategies:** The regression model will help optimize pricing strategies, ensuring that the restaurant charges the right price to maximize revenue without sacrificing customer satisfaction.
* **Informed Decision-Making:** These data-driven models empower restaurant managers to make informed, strategic decisions related to pricing, customer experience, and operational improvements, thereby improving both customer ratings and financial performance.

1. **Dataset Overview**

**2.1 Dataset Description**

The dataset used in this project contains over 26,000 records of restaurant data. It includes various attributes related to restaurant operations, customer ratings, pricing, delivery times, and cuisine types. This dataset provides a comprehensive view of how different factors influence customer satisfaction and pricing strategies.

**2.2 Key Features**

* **Input Variables:**
  + **average\_price:** The average price of a typical meal at the restaurant, providing insight into the restaurant's pricing strategy.
  + **average\_delivery\_time:** The time taken for a typical order to be delivered, which may impact customer satisfaction.
  + **Cuisine Type Indicators:** Categorical variables representing the cuisine offered by the restaurant. These include:
    - South Indian
    - North Indian
    - Fast Food
    - Street Food
    - Biryani
    - Bakery
* **Target Variables:**
  + **For Classification:**
    - A binary target variable indicating customer satisfaction. A value of **1** represents a rating of 4 or higher, while **0** represents a rating below 4. The original rating column was transformed into this binary outcome for classification.
  + **For Regression:**
    - The continuous variable representing the **optimal price**, which the regression model predicts based on operational factors.

**2.3 Preprocessing Steps**

1. **Handling Missing Values:** There are no missing/duplicate values in the dataset, which seems to be well-maintained.
2. **Feature Normalization:** Numerical features, such as average price and average delivery time, were normalized using **StandardScaler** to standardize the scale and improve model performance.
3. **Target Variable Transformation:** A threshold was applied to the original rating column (ratings ≥ 4) to create a binary target variable for classification (High Rating = 1, Low Rating = 0).
4. **TBD:** Additional preprocessing steps, such as encoding categorical variables or handling outliers, will be performed based on further exploration of the data.

**3. Exploratory Data Analysis**

**3.1 Exploratory Data Analysis (EDA)**

Through **EDA**, data patterns and correlations can be uncovered. For example, EDA can identify how factors such as pricing, cuisine type, and delivery time influence customer ratings across different cities. These insights help restaurants tailor strategies to specific market segments and operational conditions.

**4. Data Modelling**

**4.1 Classification Model (XGBClassifier)**

**4.1.1 Objective and Approach**

The objective of the classification model is to predict whether a restaurant will receive a high rating (≥4) or a low rating (<4) based on operational features such as price, delivery time, and cuisine type. To address this task, the **XGBoost Classifier** was selected due to its robustness, scalability, and performance on imbalanced datasets.

**4.1.2 Model Setup and Data Splitting**

* **Features:** The model is trained using the following features:
  + **average\_price:** The average price of a meal at the restaurant.
  + **average\_delivery\_time:** Time taken for a typical order to be delivered.
  + **Cuisine Types:** Categorical variables representing cuisine types (e.g., South Indian, Fast Food, Biryani).
* **Target Variable:** High Rating (binary outcome: 1 for ratings ≥ 4, 0 for ratings < 4).
* **Data Split:** The data is divided into 80% training and 20% testing, ensuring the model is trained on a substantial portion of the data and evaluated on a separate, unseen dataset.
  + **Training Set:**
    - Class 1 (High Rating): 12,203 records
    - Class 0 (Low Rating): 9,921 records
  + **Test Set:**
    - Class 1 (High Rating): 3,103 records
    - Class 0 (Low Rating): 2,429 records

**4.1.3 Hyperparameter Tuning**

* **GridSearchCV** was used to find the best hyperparameters for the XGBoost model, with a focus on optimizing **ROC AUC** scoring to improve classification performance.
* **Best Hyperparameters:**
  + **colsample\_bytree:** 1.0 (Uses all features for each tree).
  + **learning\_rate:** 0.01 (Slow convergence to prevent overfitting).
  + **max\_depth:** 5 (Limits tree depth to avoid complexity).
  + **n\_estimators:** 200 (Increases ensemble size to improve performance).
  + **scale\_pos\_weight:** 1 (Adjusts the weight of the positive class for imbalanced datasets).
  + **subsample:** 0.8 (Uses 80% of the data for each tree to reduce overfitting).

**4.2 Regression Model (Pricing Optimization)**

**4.2.1 Objective and Approach**

The goal of the regression model is to recommend optimal price ranges that maximize both customer satisfaction and profitability, taking into account various operational factors. Multiple regression models were evaluated to determine the best model for price optimization.

**4.2.2 Model Evaluation and Results  
  
Mean Absolute Error (MAE):** Measures the average magnitude of errors in the predictions, giving an idea of how far the predicted prices are from actual values.

**Mean Squared Error (MSE):** Calculates the average squared difference between predicted and actual values. It emphasizes larger errors more than MAE.

**Root Mean Squared Error (RMSE):** The square root of MSE, which brings the error metric back to the original scale of the data.

**R-squared (R²):** Indicates the proportion of variance in the target variable explained by the model. A higher R² suggests a better fit.

**Models Evaluated & their Result :**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **MAE** | **MSE** | **RMSE** | **R²** |
| Random Forest | 65.62 | 7335.76 | 85.65 | 0.03 |
| Gradient Boosting | 64.53 | 7199.27 | 84.85 | 0.04 |
| XGBoost | 64.58 | 7199.75 | 84.85 | 0.04 |
| Support Vector Regressor (SVR | 61.08 | 7745.82 | 88.01 | -0.03 |

**Best Performing Model:**

Gradient Boosting emerged as the best-performing model, with the lowest MAE (64.53) and the highest R² (0.04). While R² indicates that the model explains only a small portion of the variance, it still provides valuable insights into the optimal price points.

**XGBoost** showed very similar performance to Gradient Boosting, also achieving an R² of 0.04 and a slightly higher MAE (64.58).

The Support Vector Regressor (SVR) performed poorly with a negative R², indicating that it did not fit the data well for this use case.

**5. Reflection**

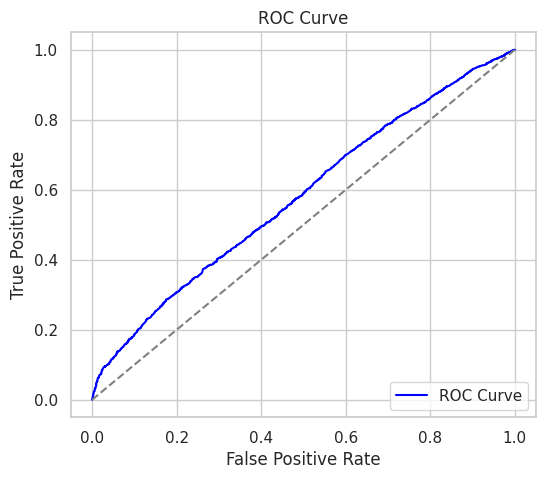
**5.1 Classification Model Results**

* **Classification Report:**
  + **Accuracy:** 57.1%
  + **Precision (Class 1):** 58.2%
  + **Recall (Class 1):** 83.5%
  + **F1-Score (Class 1):** 68.7%
  + **Precision (Class 0):** 52.6%
  + **Recall (Class 0):** 23.5%

A graph with green lines

Description automatically generated

* + **F1-Score (Class 0):** 32.7%
  + **ROC AUC:** 0.5796 (Indicating moderate model performance, as AUC value is closer to 0.5, which is random guessing).



* **Confusion Matrix:**

A blue squares with white text

Description automatically generated

* + **True Positives (TP):** 2,587
  + **False Positives (FP):** 1,857
  + **True Negatives (TN):** 572
  + **False Negatives (FN):** 516

The model performed better at identifying high satisfaction (Class 1), as indicated by the higher recall for Class 1 (83.5%) but lower performance in predicting low satisfaction (Class 0) due to its lower recall (23.5%).

**5.2 Regression Model Results**

* **Best Model:** Gradient Boosting
  + **R²:** 0.04 (indicating the model explains only a small portion of variance).
  + **MAE:** 64.53
* **Other Models:** XGBoost performed similarly to Gradient Boosting, while the Support Vector Regressor (SVR) performed poorly with a negative R², indicating it was not suitable for this task.

The regression model provides useful insights for optimal price recommendations, but its predictive power is limited by the model’s current setup and features.

**5.3 Key Insights**

**5.3.1 Customer Satisfaction Insights**

* **Delivery Time:** Restaurants with faster delivery times tend to receive higher ratings. Faster service is highly valued by customers.
* **Pricing Strategy:** High pricing can lead to lower customer satisfaction. Pricing should be optimized within a reasonable range based on operational factors.

**5.3.2 Pricing Insights**

* The regression model suggests that pricing adjustments based on delivery time, cuisine type, and customer preferences can help maximize satisfaction and profitability.
* Pricing too high for specific operational conditions (e.g., longer delivery times) can result in poor ratings.

**5.4 Recommendations for Restaurants**

* **Use Classification Model:** Predict whether a customer will rate the restaurant highly and adjust operational strategies (e.g., pricing, delivery time) accordingly. This helps improve customer satisfaction and ratings.
* **Use Regression Model:** Optimize pricing dynamically based on factors like delivery time and cuisine type to align with customer satisfaction, maximizing both profitability and positive customer experiences.

**6. Challenges**

* **Imbalanced Data:** The dataset had a slight class imbalance (more high ratings), which led to lower recall for the low rating class (Class 0). Adjusting the **scale\_pos\_weight** in XGBoost helped address this, but further improvements are needed.
* **Model Performance:** Low **R²** values for the regression models suggest limited predictive power. This can be improved by incorporating more features or using more complex models.
* **Dynamic Contexts:** The dataset does not account for dynamic factors like **seasonal changes**, **holidays**, or **real-time customer reviews**, which could impact customer satisfaction and pricing.