Table of Contents

List of Figures

List of Tables

List of Abbreviations

EDA	Exploratory Data Analysis
GridSearchCV	Grid Search with Cross-Validation
MTU	Munster Technological University
ROC-AUC	Receiver Operating Characteristic – Area Under the Curve
SMOTE	Synthetic Minority Over-sampling Technique
XGBoost	Extreme Gradient Boosting

Abstract

Customer satisfaction is considered one of the most important metrics in the restaurant business. It supports customer retention, encourages brand loyalty, and significantly influences how a brand is perceived, especially in online environments. Based on over ten years of experience as a professional chef and now as a data science student, this project explores how customer satisfaction can be predicted using a data-driven approach.

In the past, satisfaction was mostly measured through face-to-face interactions or personal opinions. Today, however, there is much more data available that can support more informed decision-making. For this project, a large synthetic dataset was used, representing a variety of customers and service-related features such as food quality, service rating, wait time, visit frequency, and whether the customer used online or in-person services

Several data analysis techniques were applied using Python, including visual exploration, statistical testing, and machine learning. Two classification models Decision Tree and Random Forest were used to predict whether a customer is likely to be highly satisfied. To address the class imbalance in the dataset, SMOTE was used to balance the training data, and Stratified 5-Fold Cross-Validation was applied to improve evaluation reliability.

The results showed that customers who participate in loyalty programs, use online services, or rate food and service highly are more likely to report high satisfaction. This study highlights how data science can help restaurants improve the customer experience and support better strategic decisions, moving beyond traditional or intuition-based approaches.

1. Introduction

With more than ten years of experience in the restaurant industry as a professional chef, customer satisfaction has always been a key focus of my work. In a field where competition is high and customer loyalty can be difficult to maintain, the ability to keep guests satisfied is essential for business success. Traditionally, satisfaction in restaurants has been driven by two main elements: the quality of the product and the quality of the service. These were the essential tools used to measure and influence how customers felt about their dining experience.

However, over time, customer expectations have changed, and so has the way restaurants operate. In today's fast paced and digital world, businesses now have access to more data than ever before. With the rise of technology, online platforms, reservation systems, and feedback tools, customer behaviour can be tracked and analysed in ways that were not possible in the past. This gives restaurants a new opportunity to better understand what their customers need and how to adapt their services accordingly.

Customer satisfaction not only affects immediate feedback, such as online reviews or word-of-mouth recommendations, but it also plays a long term role in brand reputation, repeat visits, and overall customer loyalty. Positive reviews and high ratings on platforms like Google, Yelp, or TripAdvisor can boost a restaurant's visibility and attract new customers. On the other hand, a few negative experiences can lead to loss of business, especially in a market where customers have many options and are quick to share their opinions online.

Despite the availability of data, many restaurants still rely on traditional methods to evaluate customer satisfaction. Surveys, comment cards, and face to face feedback are useful, but they are often limited, subjective, and difficult to process in real time. In most cases, valuable insights are lost simply because the information is not collected or analysed properly. From my own experience, I have seen that many operational decisions are still made based on intuition, seasonal trends, or assumptions rather than data.

This study was developed with the goal of applying data science and machine learning techniques to improve how customer satisfaction is predicted and understood in the restaurant industry. A synthetic dataset was used to simulate different customer profiles and dining experiences, including factors like food quality, service quality, wait time, ambiance, group size, spending habits, and whether the customer used digital services like delivery or online reservations.

The purpose of this project is to identify the key variables that have the most influence on whether a customer reports a high level of satisfaction. By understanding these patterns, restaurant managers and owners can apply more targeted strategies to improve the overall dining experience, reduce operational waste, and improve customer retention.

Various data science techniques and machine learning models were applied using tools such as Python and associated libraries. programming code were adapted from examples provided in Géron (2022), Kaggle, Scikit-learn documentation, and Stack Overflow discussions. Each step in the process from data cleaning and preprocessing to model evaluation was designed to reflect a real-world approach to solving business problems through data. The final objective is not only to build an accurate prediction model, but also to provide clear and practical insights that can be applied in the restaurant industry today.

2. Literature Review

In recent years, as customer satisfaction has become a central goal across many industries, the use of machine learning to predict consumer behaviour has gained attention. These technologies have been increasingly adopted to support data analysis and help identify patterns in customer satisfaction, allowing businesses to design more effective strategies for improving service and retaining clients. In this section, key literature is reviewed to explore how customer satisfaction is defined and measured in the restaurant industry, and how machine learning and data driven methods have been applied in this context.

2.1 Importance of satisfaction in Restaurant industry

In service based industries, maintaining a high level of customer satisfaction is considered essential for business success. These types of business models rely heavily on the quality of service provided and the overall experience offered to customers. This principle applies not only to restaurants, but also to service focused businesses such as salons, consultancies, and hospitality venues. However, in a competitive economy, satisfying a wide range of customers, each with their own preferences and expectations can be more complex than it appears.

Several studies have been carried out to explore the key factors that influence customer satisfaction in restaurants. For example, Malik et al. (2013) examined how service quality and product quality influence satisfaction. It was found that customer interaction with staff, along with the quality of food served, plays a significant role in shaping the overall dining experience. Their findings support the idea that satisfaction is multi dimensional and goes beyond just food.

Other studies have emphasised the role of restaurant atmosphere. Pecotić et al. (2014), for instance, examined how interior design elements such as plate presentation, background music, and furniture layout can influence customer perception. Their research showed that depending on the restaurant's target audience and concept, design choices can have a strong impact on satisfaction levels. This suggests that strategies to improve customer satisfaction should be tailored to specific business contexts and customer groups.

More recently, Hafeez et al. (2020) identified a wider set of factors that affect satisfaction in dining environments.

Their research listed the following key areas:

- Food Quality – A core factor determining how customers perceive value.
- Service Quality – Including staff responsiveness, friendliness, and professionalism.
- Ambience and Environment – Cleanliness, lighting, and music all contribute to comfort.
- Price Fairness – Customers evaluate whether the price is appropriate for the experience received.
- Location Accessibility – Affects convenience and likelihood of return visits.
- Waiting Time – Long wait times can negatively affect customer mood and perception.

These and other studies reinforce that satisfaction in the restaurant industry is not determined by one factor alone, but by a mix of product quality, service, environment, pricing, and overall expectations.

2.2 Data in the industry and the models

While data collection has become easier thanks to digital tools such as online ordering platforms and customer feedback systems, the use of that data for analysis and decision-making in the restaurant industry remains limited compared to other service sectors. For example, hotels often apply data analysis to forecast demand or personalise guest experiences. In contrast, many restaurants continue relying on seasonal patterns or intuition when making operational decisions, rather than using data-based strategies.

Even though data driven methods have shown clear advantages, such as reducing food waste or managing staffing more efficiently many restaurants do not yet use them fully. For instance, during the COVID-19 pandemic, the need to adapt quickly led to some progress. In a study by Henríquez-Ramírez et al. (2021), online food ordering patterns were analysed to understand how customer preferences changed during lockdown. However, this study was conducted a year after the lockdown began, showing that the restaurant industry often responds more slowly to change than other sectors.

One challenge is that restaurants deal with a wide range of customer needs. Analysing this kind of data takes time and resources, which many small or mid-size restaurants do not have. In practice, it is common to see decisions still being made by “instinct” or tradition, rather

than supported by data. In contrast, sectors like hotels even when they use only a small amount of data tend to see the benefits quickly.

Several academic studies have explored how machine learning and data science can be used to understand satisfaction and improve services. For example, Monroy Ceseña (2021) analysed 49 restaurants in Todos Santos, Mexico, to study how service quality influenced satisfaction across different genders. The results showed a strong link between high service ratings and satisfied customers, but found no significant difference between how men and women rated the experience. The type of restaurant also didn't appear to affect satisfaction levels in this study.

More advanced techniques have also been tested. Lee et al. (2021) used machine learning models to examine what makes some online restaurant reviews more helpful than others. They worked with a large dataset from Yelp, that had, nearly 1.5 million reviews and tested several algorithms including multivariate linear regression, support vector machines, random forest, and extreme gradient boosting (XGBoost).in such study, was found that the best results were achieved with XGBoost. Interestingly, they found that who wrote the review (for example, their credibility or profile history) had more influence than what the review actually said. This suggests that customer trust plays a major role in how opinions are formed, not just the content itself.

Another study by Aisyah Larasati et al. (2012) compared two predictive models neural networks and logistic regression to forecast satisfaction. Their model could predict satisfaction with around 80% accuracy on known data and 70% accuracy on new data. The top tree predictors were:

- Overall satisfaction with service,
- Speed of staff response,
- General service excellence.

These examples show that machine learning and statistical tools can help restaurants better understand what makes customers satisfied and what changes can lead to improvements. Still, these tools are not widely used in practice, and the industry has room to grow when it comes to adopting data based strategies.

3. Methodology

Previous research has examined restaurant customer satisfaction by focusing on various factors such as interior design (Pecotić et al., 2014), service quality (Malik, 2013), and customer demographics like gender (Monroy, 2013).

In this section, the process of how to build a model for predicting customer satisfaction is described. This methodology includes a series of steps: gaining a thorough understanding of the dataset, performing data cleaning and preprocessing, conducting exploratory data analysis (EDA), implementing machine learning models, and evaluating their performance.

The EDA phase involved both visual and statistical exploration of the data. Proportion-based plots were used to compare satisfaction levels across categorical, ordinal, and continuous features, while Chi-square tests were conducted to statistically validate the strength of associations between categorical variables and the satisfaction outcome.

The goal of this phase is to determine which model performs best in predicting customer satisfaction based on the available variables.

Some visualisation and model evaluation functions were adapted from examples provided in Géron (2022), as well as from publicly available resources such as Kaggle, Scikit-learn documentation and Stack Overflow discussions.

3.1 Dataset Description

The dataset used for this project is the "Predict Restaurant Customer Satisfaction" dataset created by Rabie El Kharoua (2024) and published on Kaggle. This dataset is synthetic, meaning that it was generated to simulate realistic customer behaviour and restaurant experiences. Even though it is not real world data, it provides a useful base for applying machine learning models and exploring how certain variables may influence customer satisfaction.

The dataset contains information about different aspects of restaurant visits, including customer demographics, visit details, spending behaviour, and satisfaction ratings. Each row represents a single customer's visit and includes multiple features that could help predict whether that customer was highly satisfied or not.

The features are grouped into three main categories:

- **Demographic Information**

- CustomerID: A unique ID for each customer (this column was later dropped before modelling).
- Age: The age of the customer.
- Gender: Gender of the customer (Male/Female).
- Income: Annual income of the customer in USD.

- **Visit-Specific Variables**
  - Visit Frequency: How often the customer visits the restaurant (Daily, Weekly, Monthly, Rarely).
  - Average Spend: Average amount spent per visit in USD.
  - Preferred Cuisine: Type of cuisine preferred (Italian, Chinese, Indian, Mexican, American).
  - Time of Visit: Typical time the customer visits (Breakfast, Lunch, Dinner).
  - Group Size: Number of people in the group.
  - Dining Occasion: The reason for dining (Casual, Business, Celebration).
  - Meal Type: Type of meal (Dine-in or Takeaway).
  - Online Reservation: Whether the customer booked online (1 = Yes, 0 = No).
  - Delivery Order: Whether the customer placed a delivery order (1 = Yes, 0 = No).
  - Loyalty Program Member: Whether the customer is part of the restaurant's loyalty program (1 = Yes, 0 = No).
  - Wait Time: Time the customer waited before receiving their meal (in minutes).
- **Satisfaction Ratings**
  - Service Rating: Rating given by the customer for service (scale 1 to 5).
  - Food Rating: Rating for the food (scale 1 to 5).
  - Ambiance Rating: Rating for the environment or ambiance (scale 1 to 5).
- **Target Variable**
  - High Satisfaction: This is the main target used in this project. It is a binary variable showing if the customer was highly satisfied (1) or not (0). The variable was likely generated based on other rating and behavioural features in the dataset.

This dataset gives a complete picture of a restaurant experience by including both customer characteristics and service-related feedback. What makes it useful is the variety of features available, which allows for a broader analysis of what influences satisfaction. Although the data is artificial, here it is assumed that the values are realistic and represent the kind of patterns commonly seen in the industry.

By using this dataset, it becomes possible to build and test machine learning models and see how accurately they can predict customer satisfaction based on specific variables. It also provides an opportunity to explore the relationships between different factors and how they may affect the overall experience of restaurant guests.

3.2 Data Cleaning and Preprocessing

Since the dataset used in this study was synthetic, no major cleaning was required. A quick review (see appendix 1) of the data revealed that there were no missing values or inconsistencies, which allowed the preparation process to begin immediately. However, a few preprocessing steps were still necessary before the data could be used for machine learning.

Some variables in the dataset, such as Gender, Meal Type, and Visit Frequency, were categorical and needed to be converted into numerical format. Simple label encoding was applied, assigning a numeric code to each category so that the machine learning models could interpret them correctly. For instance, the gender categories "male" and "female" were replaced with numeric values, as were "breakfast," "lunch," and "dinner" in the Time of Visit column.

Other variables, including Income, Average Spend, Group Size, and Wait Time, were already numerical but required standardisation. Standard scaling was performed to bring all values into a similar range. This ensured that features with larger numerical ranges, such as income, did not dominate those with smaller ranges, such as group size. By scaling the data, the models were able to treat all features more fairly.

The original dataset contained 19 columns, but the Customer ID column was removed because it only served as an identifier and did not contribute to the predictive analysis. After encoding and cleaning, the final dataset consisted of 18 variables, which included both the original numerical values and the encoded categorical features.

Overall, this preprocessing step was a crucial part of the data preparation process. Although the dataset did not require deep cleaning, transforming the data into a machine learning ready format and ensuring proper scaling were essential for building effective models.

3.3 Exploratory Data Analysis (EDA)

To gain a comprehensive understanding of customer behaviour and how different variables relate to satisfaction levels, Exploratory Data Analysis (EDA) was conducted. This stage was essential to uncover trends, detect patterns, and guide feature selection prior to model development (Géron, 2022). Visualisations such as histograms, bar plots, and correlation heatmaps were created using Python libraries including Matplotlib and Seaborn. These visual tools supported the analysis by making it easier to detect patterns across the dataset.

Figure 3.1 presents proportional bar plots comparing the levels of HighSatisfaction (0 = not satisfied, 1 = highly satisfied) across key categorical features, including Gender, Visit Frequency, Preferred Cuisine, Time of Visit, Dining Occasion, Meal Type, Online Reservation, Delivery Order, and Loyalty Program Membership.

Weekly visitors reported the highest satisfaction rate at 20%, followed by daily visitors at 15%, suggesting that frequent customers may be more engaged or familiar with the dining experience. Celebration occasions showed a satisfaction rate of 19%, notably higher than business and casual visits at 9% and 11% respectively, which may reflect heightened expectations or improved service quality during special events. Dine-in customers had a satisfaction rate of 18%, while takeaway customers reported only 8%, indicating a potential gap in experience or perceived service depending on dining type.

Similarly, online reservations and loyalty program membership were linked to higher satisfaction levels, with reservation users reporting 24% and loyalty members 19%, highlighting the possible influence of digital convenience and customer retention strategies. In contrast, variables such as Gender, Preferred Cuisine, and Time of Visit exhibited minimal variation in satisfaction rates, suggesting they may play a less significant role in determining overall satisfaction. These patterns offer valuable insights for feature selection in the modelling stage and point to key areas that could influence how customers perceive their

dining experience.

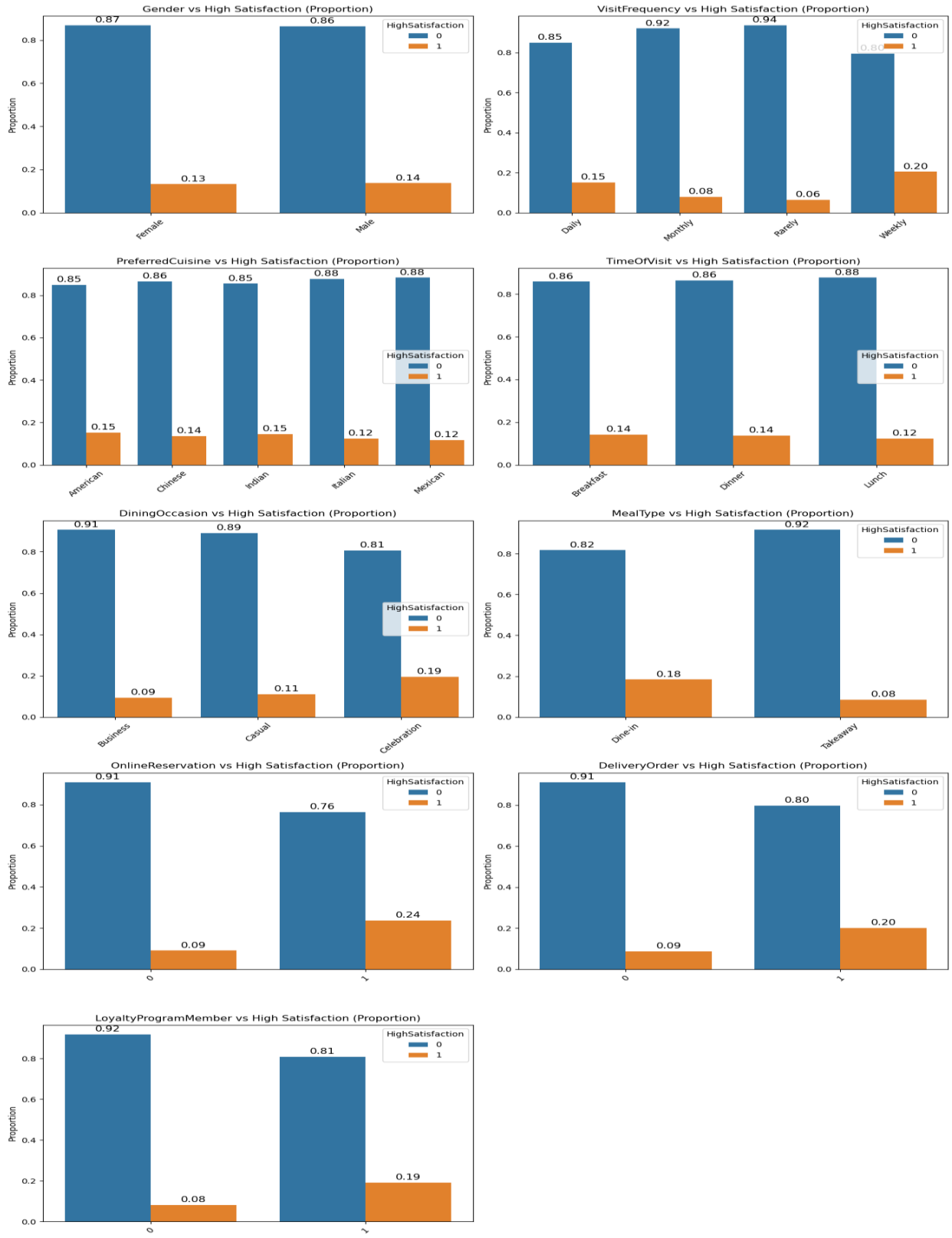

Figure 3.1 Proportion of High Satisfaction Across Categorical Features

Figure 3.2 shows the proportional distributions of continuous numerical variables, including Age, Income, Average Spend, and Wait Time. These plots offer a clearer understanding of customer diversity and behaviour tendencies within the dataset.

Age is distributed fairly evenly between 18 and 70 years, indicating a broad and varied customer base. Income appears relatively uniform, with a slight concentration toward higher earnings but without significant skew, suggesting a balanced distribution of purchasing power. The Average Spend per visit follows a bell-shaped curve, centred around $100 to $125, which points to a consistent spending pattern among customers. Wait Time shows a mild concentration at the lower end (under 20 minutes), yet the distribution remains broad, reflecting variability in service efficiency.

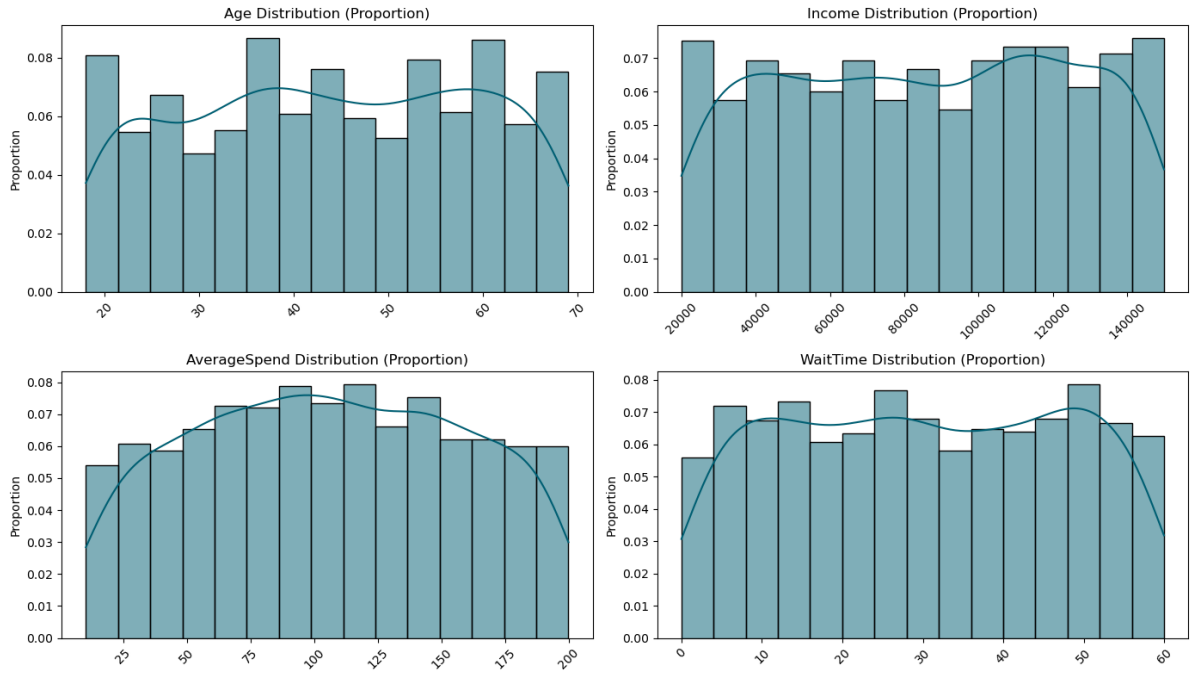

Figure 3.2 Proportional Distribution of Continuous Numerical Features

Figure 3.3 presents normalized histograms for ordinal features including Group Size, Service Rating, Food Rating, and Ambiance Rating. These variables are treated as discrete numerical values with a natural order ranging from low to high.

Group Size ranges from 1 to 9 and displays a relatively balanced distribution, with a slight concentration between 4 and 6 diners, suggesting that small to medium-sized groups are the most typical. The distributions for Food Rating, Service Rating, and Ambiance Rating appear evenly spread across their 1 to 5 scales, indicating that customer evaluations in these

categories are widely varied, with no strong gathering toward either high or low scores.

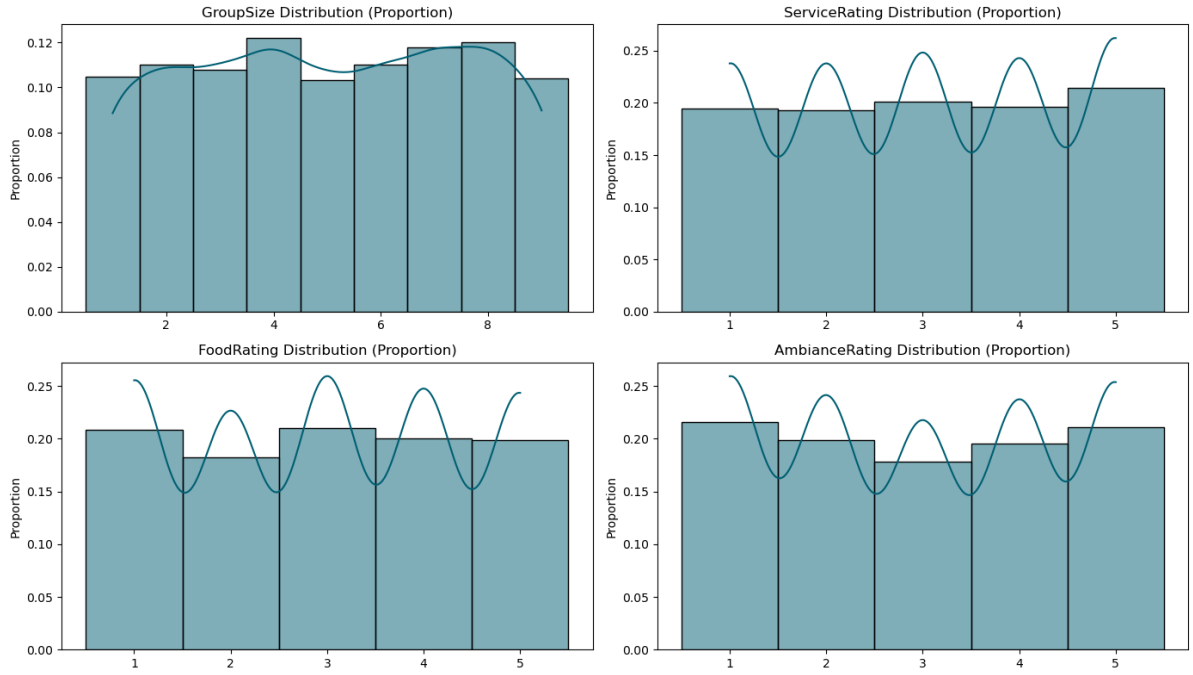

Figure 3.3 Proportional Distribution of Discrete Numerical Features

To complement the visual analysis and statistically validate the observed relationships between categorical features and customer satisfaction, Chi-square tests of independence were conducted. This non-parametric test evaluates whether the distribution of satisfaction levels (HighSatisfaction = 0 or 1) significantly differs across the categories of each feature. The test compares the observed frequencies of satisfied versus not satisfied customers with the frequencies that would be expected if there were no association between the variables.

Each categorical variable was cross tabulated with the target variable, and a Chi-square statistic was calculated to determine whether the observed differences in satisfaction proportions across groups were statistically significant. A low p-value (typically < 0.05) indicates a significant association between the feature and customer satisfaction.

The results are presented in Table 3.1. Behavioral and engagement-related features, such as Visit Frequency, Dining Occasion, Meal Type, Online Reservation, Delivery Order, and Loyalty Program Membership, all showed statistically significant or marginal associations with HighSatisfaction. For example, Visit Frequency had a Chi^2 value of 21.39 with a p-value less than 0.001, suggesting that satisfaction levels vary meaningfully based on how often customers visit. Likewise, customers who use delivery services or participate in the loyalty

program tend to show higher satisfaction levels, as confirmed by significant p-values in both cases.

In contrast, variables such as Gender, Time of Visit, and Preferred Cuisine did not show statistically significant associations. Their high p-values indicate that the distribution of satisfied versus not satisfied customers remains relatively constant across different categories within these features, suggesting a weaker or negligible relationship with the outcome variable.

Feature	Chi² Value	df	p-value
Visit Frequency	21.39	3	< 0.001
Dining Occasion	13.54	2	< 0.001
Meal Type	6.33	2	<0.042
Online Reservation	14.87	1	< 0.001
Delivery Order	10.76	1	<0.001
Loyalty Program	12.91	1	< 0.001
Gender	0.12	1	0.732
Time of Visit	0.03	2	0.986
Preferred Cuisine	1.74	4	0.783

Table 3.1 Chi-Square Test Results

3.4 ML Models

After preparing and exploring the dataset, two machine learning models were selected to predict customer satisfaction: Decision Tree and Random Forest. These models were chosen because they are easy to interpret, work well with both numerical and categorical data, and can handle non-linear relationships between variables. Another reason for selecting tree-based models is their ability to rank features by importance, which can help identify which factors most influence satisfaction. (Geron, 2022)

The main goal of these models was to predict the binary target variable High Satisfaction, where 1 indicates a highly satisfied customer and 0 means otherwise. The prediction task is treated as a classification problem.

The dataset was prepared by encoding categorical variables using one-hot encoding, scaling numerical features, and then splitting the data into training and testing sets using an 80/20 split with stratified sampling to preserve class balance. Given the class imbalance observed during exploratory analysis (only 13% of records reflected high satisfaction.

SMOTE (Synthetic Minority Over-sampling Technique) was introduced to oversample the minority class and provide a more balanced dataset for training.

To assess model performance, some configurations were tested:

- Using the full feature set vs. a reduced set based on statistical significance (Chi-square test results).
- Training with and without SMOTE.
- Applying GridSearchCV for hyperparameter tuning.

Models were evaluated using common classification metrics: Accuracy, Precision, Recall, F1-score and ROC-AUC (see appendix 4). Additionally, stratified 5-Fold Cross-Validation was used to validate the robustness of model performance.

This section sets the foundation for the results discussed in Section 4, where each model configuration is evaluated and compared in terms of its ability to identify both satisfied and not satisfied customers.

4. Results

After training and evaluating the models, this section presents the results obtained from the Decision Tree and Random Forest classifiers under multiple configurations. Both models were assessed using a hold-out test set and evaluated using standard classification metrics, including accuracy, precision, recall, F1-score, and ROC-AUC. Additionally, stratified 5-Fold Cross-Validation was used to provide a more reliable estimate of model performance.

To establish a baseline, a naive classifier that always predicts the majority class (not satisfied) was tested. This model achieved an accuracy of 86.7%, but both recall and F1-score were 0.00 for the satisfied class (Class 1), highlighting its complete inability to detect the minority class. This result reinforces the importance of using more sophisticated models capable of learning from feature patterns.

To assess the effect of both SMOTE balancing and hyperparameter tuning, models were evaluated under four scenarios: without SMOTE and without GridSearch, with SMOTE only, with both SMOTE and GridSearch, and using reduced features based on significance testing. This comprehensive comparison highlights how different strategies influence model performance.

4.1 Decision Tree

The Decision Tree is a supervised learning algorithm that recursively splits the dataset into subsets based on feature values, forming a tree structure of decisions that leads to a final classification. It is known for its simplicity, interpretability, and capability to handle both numerical and categorical features. Due to its structure, the Decision Tree is sensitive to imbalanced data and can benefit from strategies such as SMOTE and hyperparameter tuning.

In the baseline configuration, the model was trained without any resampling or tuning. It achieved an overall accuracy of 74%, which might seem acceptable, but its precision, recall, and F1-score for the minority class were quite low. Specifically, recall stood at 0.28, indicating that fewer than one-third of highly satisfied customers were correctly identified. The F1-score of 0.22 suggests a poor balance between precision and recall. The ROC-AUC score of 0.545 confirms the model's limited ability to distinguish between the two classes.

When SMOTE was applied to balance the dataset, performance improved. The model now achieved an accuracy of 78%, with better results for the minority class: recall increased to

0.40, and the F1-score rose to 0.32. This demonstrates that SMOTE helps the Decision Tree better detect patterns related to highly satisfied customers by mitigating the class imbalance.

With the addition of GridSearchCV for hyperparameter tuning, the model's performance improved further. Using the best parameters found, the Decision Tree reached an accuracy of 80%. More importantly, recall for the minority class increased to 0.50, and F1-score to 0.40, showing a more balanced classification. The ROC-AUC score of 0.67 supports this improvement in predictive capability. This configuration provided the best overall performance for the Decision Tree model. (see appendix 2)

4.2 Random Forest

The Random Forest classifier is an ensemble learning technique that constructs multiple decision trees using different subsets of the training data and features. It aggregates the predictions from these trees to form a final output. This process reduces overfitting and improves generalization, making it a popular choice for classification tasks involving both linear and non-linear relationships.

In its basic configuration without SMOTE or tuning, the Random Forest achieved a high accuracy of 87.0%. However, this accuracy was driven almost entirely by the model's performance on the majority class. It failed to meaningfully detect satisfied customers, achieving a recall of only 0.03 and an F1-score of 0.05 for Class 1. Despite the high overall accuracy, this model offers little value in identifying highly satisfied customers.

When SMOTE was used to address class imbalance, the Random Forest improved significantly. The model achieved 89% accuracy, but more importantly, it recorded a recall of 0.40 and an F1-score of 0.50 for the minority class. This balanced performance made it a viable model for customer satisfaction prediction.

Further improvement was observed with the inclusion of GridSearchCV. By tuning parameters such as the number of estimators and maximum depth, the model achieved an accuracy of 90.0%, precision of 0.75, recall of 0.38, and F1-score of 0.50 for the minority class. Although recall slightly decreased compared to the untuned model, the gain in precision and overall balance of metrics made this configuration the best-performing model overall. The ROC-AUC of 0.754 supports its improved discrimination capability. (see appendix 3)

4.3 Summary and Model Comparison

The naive baseline model achieved high accuracy by predicting the majority class but failed completely in identifying satisfied customers (recall = 0.00). This reinforces the importance of assessing multiple evaluation metrics when working with imbalanced data.

While the Decision Tree model with SMOTE and hyperparameter tuning achieved the highest recall for the minority class (50%), it did so with lower precision (33%) and overall accuracy (80%). This means it correctly identified more highly satisfied customers but also produced more false positives. In contrast, the Random Forest model reached the highest overall accuracy (90%) and precision (75%) for the minority class, although it recalled slightly fewer satisfied customers (38%). Given the stronger balance between precision, recall, and F1-score, the Random Forest with SMOTE and tuning was considered the best overall model. It offered higher confidence in its predictions while maintaining reasonable sensitivity, making it more practical for real-world applications where both correct identification and prediction reliability are important.

Model	***SMOTE***	***GridSearch***	***Accuracy***	***Precision***	***Recall***	***F1 Score***	***ROC-AUC***
Decision Tree	-	-	0.74	0.19	0.28	0.22	0.54
Random Forest	-	-	0.87	1.00	0.03	0.05	0.74
Decision Tree	present	-	0.78	0.27	0.40	0.32	0.61
Random Forest	present	-	0.89	0.67	0.40	0.50	0.76
Decision Tree	present	present	0.80	0.33	0.50	0.40	0.67
Random Forest	present	present	0.90	0.75	0.38	0.50	0.75

Table 4.3 Summary and Model Comparison

While not the main focus of this study, these comparisons demonstrate how model performance can change depending on data preprocessing and tuning. In future work, more advanced techniques such as XGBoost could be explored as suggested in the work of Lee et al. (2021).

5. Discussion & Conclusion

This This study explored how machine learning models can be used to predict customer satisfaction in the restaurant industry using a synthetic dataset. The goal was to uncover key factors behind high satisfaction and to assess how different models perform when trying to identify highly satisfied customers.

The EDA stage showed a strong imbalance in the target variable, with only 13% of customers being classified as highly satisfied. Important relationships were observed between satisfaction and factors such as visit frequency, reservation use, food and service ratings, and membership in a loyalty program.

The Decision Tree model offered simplicity and clear visual explanations. However, its performance for detecting satisfied customers was limited at first. With SMOTE and parameter tuning, it improved, especially in recall and F1-score.

The Random Forest model showed stronger baseline performance. It was more stable and reliable, especially when SMOTE and tuning were applied. Among all the tested configurations, Random Forest with SMOTE and GridSearch was the top performer. It achieved the highest scores across most evaluation metrics, making it the best option for satisfaction prediction.

The naive model, which only predicted the majority class, helped confirm the importance of going beyond accuracy. Although its accuracy was high, it completely failed to detect any satisfied customers.

Even though the dataset used in this study is synthetic, it successfully mirrors real-world challenges like data imbalance and complex customer behaviour. The analysis showed that features linked to customer behaviour and experience (such as service quality, frequency of visits, and loyalty engagement) are more useful than demographics when predicting satisfaction.

In summary, this project showed how data science can support better decisions in the restaurant industry. Machine learning can help identify what matters most to customers and how businesses can improve their services. Future work could include testing more advanced models like XGBoost or applying these methods to real-world customer data to further validate and refine the findings.

References

- Shujah Alam Malik & Tahir Mumtaz Awan. (2013). MEASURING SERVICE QUALITY PERCEPTIONS OF THE CUSTOMERS OF RESTAURANTS IN PAKISTAN. In International Journal for Quality Research (Vol. 7, Issue 2, pp. 187–200). http://www.ijqr.net/journal/v7-n2/2.pdf
- Pecotic, M., Bazdan, V., Samardzija, J. (2014), Interior design in restaurants as a factor influencing customer satisfaction. RIThink. Retrieved from: https://rithink.hr/brochure/pdf/vol4_2014/10-14.pdf
- Hafeez, I., Rizwan, M., Rehman, W. U., & Noreen, U. (2020). Factors affecting customers' satisfaction in restaurants industry in Pakistan. International Journal of Scientific and Research Publications, 10(1), 369–375. Retrieved from https://www.researchgate.net/publication/338884092.
- Henríquez-Ramírez, J., Asipuela-Girón, J. A. and Sánchez-González, I. P. (2021) Online Consumer Behavior and Factors Influencing Purchase Decisions in Restaurants. 593 Digital Publisher CEIT, 6(6), pp. 391–404. doi: 10.33386/593dp.2021.6.783.
- Monroy Ceseña, M. A. (2021) Service quality in restaurants of Todos Santos (Mexico) by gender concept. Journal Universidad & Empresa, 23(40), 1-30 https://doi.org/10.12804/revistas.urosario.edu.co/empresa/a.8229
- Lee, M., Kwon, W. and Back, K.-J. (2021), "Artificial intelligence for hospitality big data analytics: developing a prediction model of restaurant review helpfulness for customer decision-making", International Journal of Contemporary Hospitality Management, Vol. 33 No. 6, pp. 2117-2136. https://doi.org/10.1108/IJCHM-06-2020-0587
- Aisyah Larasati, Camille DeYong, Lisa Slevitch, The Application of Neural Network and Logistics Regression Models on Predicting Customer Satisfaction in a Student-Operated Restaurant, Procedia - Social and Behavioral Sciences, Volume 65, 2012, Pages 94-99, ISSN 1877-0428, https://doi.org/10.1016/j.sbspro.2012.11.097. (https://www.sciencedirect.com/science/article/pii/S1877042812050823)
- Rabie El Kharoua. (2024). Predict Restaurant Customer Satisfaction Dataset [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/8743147
- Géron, A., 2022. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. 3rd ed. Sebastopol, CA: O'Reilly Media.
- Scikit-Learn (2019). User guide: contents — scikit-learn 0.22.1 documentation. [online] Scikit-learn.org. Available at: https://scikit-learn.org/stable/user_guide.html.

Appendix

Appendix 1: Statistical tables

Statistic	*Age*	*Income (USD)*	*Average Spend (USD)*	*Group Size*	*Wait Time (min)*
Count	*1,500*	*1,500*	*1,500*	*1,500*	*1,500*
Mean	*43.8*	*85,922*	*105.7*	*5.0*	*30.2*
Std Dev	*15.0*	*38,183*	*52.4*	*2.6*	*17.2*
Min	*18*	*20,012*	*10.3*	*1*	*0.0*
25th Pctl	*31.8*	*52,444*	*62.3*	*3*	*15.2*
Median	*44*	*85,811*	*104.6*	*5*	*30.0*
75th Pctl	*57*	*119,159*	*148.6*	*7*	*45.3*
Max	*69*	*149,875*	*200.0*	*9*	*60.0*

Table 1.1. Summary Statistics for Numerical Variables

This table provides key descriptive statistics for continuous variables in the dataset, including Age, Income, Average Spend, Group Size, and Wait Time. These metrics help illustrate the distribution, central tendency, and spread of values among customers.

Statistic	*Online Reservation*	*Delivery Order*	*Loyalty Program Member*	*Service Rating*	*Food Rating*	*Ambiance Rating*	*High Satisfaction*
Count	*1,500*	*1,500*	*1,500*	*1,500*	*1,500*	*1,500*	*1,500*
Mean	*0.30*	*0.41*	*0.48*	*3.04*	*3.00*	*2.99*	*0.13*
Std Dev	*0.46*	*0.49*	*0.50*	*1.42*	*1.42*	*1.45*	*0.34*
Min	*0*	*0*	*0*	*1*	*1*	*1*	*0*
25th Pctl	*0*	*0*	*0*	*2*	*2*	*2*	*0*
Median	*0*	*0*	*0*	*3*	*3*	*3*	*0*
75th Pctl	*1*	*1*	*1*	*4*	*4*	*4*	*0*
Max	*1*	*1*	*1*	*5*	*5*	*5*	*1*

Table 1.2. Summary Statistics for Binary and Categorical Variables

This table shows descriptive statistics for binary and categorical features, such as Online Reservation, Delivery Order, Loyalty Program Membership, and satisfaction ratings. The values indicate how these features are distributed and highlight general trends among customers.

Appendix 2: Decision Three Model

Table 4.1 his table presents the performance metrics of the Decision Tree model used to classify customer satisfaction. The model evaluated two classes: Class 0 (Not Satisfied) and Class 1 (Highly Satisfied). Metrics such as precision, recall, F1-score, and support (the number of samples) are reported for each class. In addition, overall performance metrics including accuracy, macro average, weighted average, and ROC-AUC score are included to provide a comprehensive evaluation of the model.

Figure 4.1 displays the confusion matrices for the Decision Tree model evaluated under three configurations: without SMOTE, with SMOTE, and with both SMOTE and GridSearchCV. Each matrix compares actual versus predicted classifications for customer satisfaction. Class 0 represents "Not Satisfied" customers, and Class 1 represents "Highly Satisfied" customers. The diagonal values show correct predictions, while off-diagonal values indicate misclassifications. Improvements can be observed in the third matrix, where tuning and resampling led to more balanced performance, particularly for identifying satisfied customer

Decision Tree Model					
Configuration	**Accuracy**	**Precision (Class 1)**	**Recall (Class 1)**	**F1-Score (Class 1)**	**ROC-AUC**
Without SMOTE	0.78	0.27	0.40	0.32	0.617
With SMOTE	0.78	0.27	0.40	0.32	0.617
With SMOTE + GridSearchCV	0.80	0.33	0.50	0.40	0.672

Table 4.1 Decision Tree model Performance Summary

Note: Class 1 (C1) represents the "Highly Satisfied" group. ROC-AUC score reflects the model's ability to differentiate between classes.

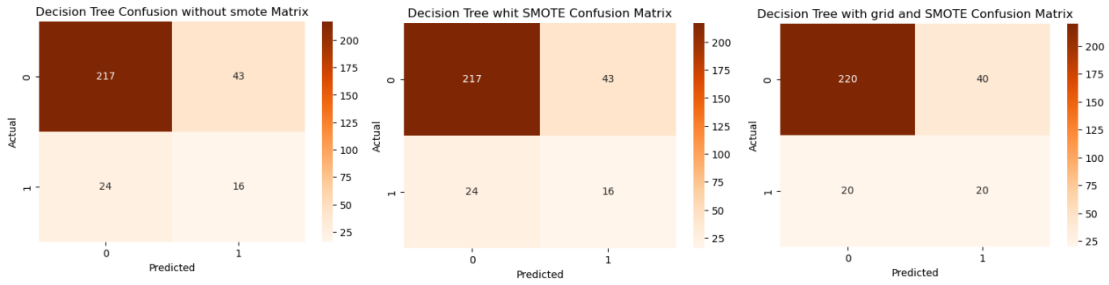

Figure 4.1. Decision Tree Confusing Matrix

Appendix 3: Random Forest Model

Table 4.2. classification performance of the Random Forest model in predicting customer satisfaction. As with the Decision Tree, the analysis considers two classes: Class 0 (Not Satisfied) and Class 1 (Highly Satisfied). Reported metrics include precision, recall, and F1-score for each class, along with overall model performance indicators such as accuracy and ROC-AUC score. These values help assess the model's ability to correctly identify both majority and minority class customers under different configurations—without SMOTE, with SMOTE, and with GridSearchCV optimization plus SMOTE.

Figure 4.2. *The confusion matrices illustrate the Random Forest model's performance across three configurations: without SMOTE, with SMOTE, and with both GridSearchCV tuning and SMOTE. In each case, the model performed very well in identifying dissatisfied customers (Class 0). The best performance for the minority class (Class 1 – highly satisfied) occurred after applying both SMOTE and tuning. However, some difficulty remained in achieving high recall for the satisfied class, emphasizing the persistent challenge of class imbalance despite improvements from model optimization.*

Random Forest Model							
Configuration	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**ROC-AUC**	**Class 1 Recall**	**Class 1 F1 score**
Without SMOTE	0.89	0.67	0.40	0.50	0.769	0.40	0.50
With SMOTE	0.89	0.67	0.40	0.50	0.769	0.40	0.50
With GridSearch + SMOTE	0.90	0.75	0.38	0.50	0.754	0.38	0.50

Table 4.2 Random Forest Models Summary

Note: All metrics refer to the performance of the model on the test set. Class 1 refers to "Highly Satisfied" customers. GridSearch tuning further increased precision, while recall remained stable.

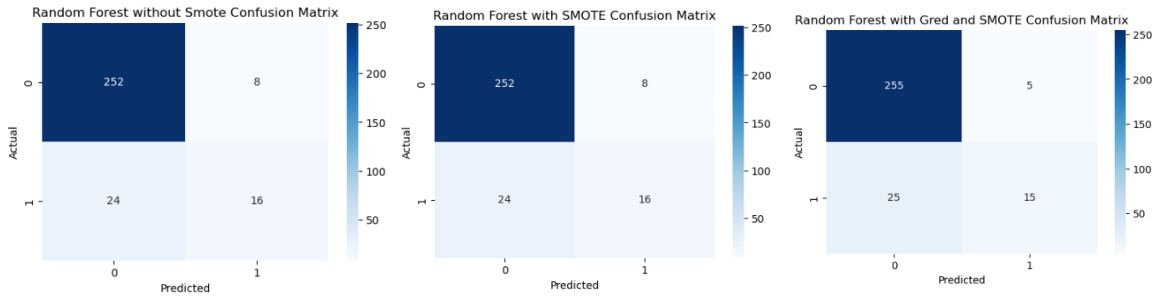

figure 4.2. Random forest Confusion Matrix

Appendix 4: Evaluation Metrics Description

*- **Accuracy**: The proportion of total correct predictions over all predictions made. It measures overall performance but can be misleading in imbalanced datasets.*

$$Accuracy = (TP + TN) / (TP + TN + FP + FN)$$

***- Precision**: The proportion of positive predictions that were actually correct. It reflects how reliable the positive predictions are.*

$$Precision = TP / (TP + FP)$$

***- Recall (Sensitivit)**: The proportion of actual positives that were correctly identified. It is crucial when missing positive cases is costly.*

$$Recall = TP / (TP + FN)$$

***- F1-Score:** Precision and Recall. It provides a balance between the two, especially useful in imbalanced datasets.*

$$F1 = 2 * (Precision * Recall) / (Precision + Recall)$$

***- ROC-AUC** (Receiver Operating Characteristic – Area Under the Curve): Measures the model's ability to distinguish between the positive and negative classes. A higher score indicates better classification performance.*

Where:

***TP** = True Positives, **TN** = True Negatives, **FP** = False Positives, **FN** = False Negatives*