Customer Satisfaction Analysis of Airline by Mohib Imran

Abstract

Introduction

2.1. Research Objective -- 2
2.2. Dataset Overview --- 2

Methodology

3.1. Data Preparation and Cleaning -- 2
3.1.1. Data Loading and Inspection -- 2
3.1.2. Data Cleaning --- 2
3.2. Feature Engineering --- 3
3.2.1. Feature Selection --- 3
3.2.2. Feature Scaling --- 3
3.3. Model Development --- 4
3.3.1. Logistic Regression -- 4
3.4. Model Evaluation -- 5
3.4.1. Performance Metrics --- 5
3.4.2. Model Interpretation --- 5
3.5. Addressing Class Imbalance -- 6

Dataset

4.1. Key Metrics --10
4.2. Numerical and Categorical Variables --11

Results

5.1. Model Performance ---11
5.2. Class Imbalance Issues -- 11

Discussion

6.1. Class Imbalance and Model Bias -- 11
6.2. Potential Improvements -- 11

Conclusion

7.1. Key Findings --- 11
7.2. Future Work -- 11

References -- 12

Abstract:

This study analyzes a dataset related to **social media advertising campaigns**, exploring relationships between key metrics like **clicks**, **impressions**, **ROI**, and **engagement**. **Exploratory data analysis** revealed significant variations in click performance, with a wide range observed. A **machine learning model** was developed to predict **ROI** based on features like clicks and impressions. **Model evaluation metrics**, including **F1-score (0.72)**, **precision (0.80)**, **accuracy (0.94)**, and **recall (0.68)**, were calculated. The model demonstrated good performance in predicting ROI for certain scenarios, particularly for instances where the actual ROI was low.

However, the model exhibited lower performance in predicting **high-ROI campaigns**, potentially due to **class imbalance** issues. Further research is recommended to address these limitations and improve model accuracy, such as implementing techniques to handle class imbalance and exploring additional features that may better capture the nuances of campaign performance.

Introduction:

This study focuses on analyzing a dataset related to **social media advertising campaigns**. The dataset encompasses key metrics such as **clicks**, **impressions**, **ROI**, and **engagement scores**, providing a comprehensive overview of campaign performance. The primary objective of this research is to develop and evaluate a **machine learning model** capable of predicting **ROI** based on key features like clicks and impressions.

This analysis aims to provide valuable insights into the factors influencing campaign success and to develop a predictive model that can assist marketers in optimizing their advertising strategies and maximizing **ROI**. By exploring the relationships between these key metrics, the study seeks to offer a deeper understanding of how different aspects of campaign performance contribute to overall success.

Methodology

This study employs a **machine learning approach** to predict **Return on Investment (ROI)** for **social media advertising campaigns**. The methodology involves the following steps:

- **Data Preparation and Cleaning:**
  - **Data Loading and Inspection**: The dataset, containing information on campaigns, clicks, impressions, and other relevant metrics, was loaded and inspected for **missing values**, **inconsistencies**, and **data type errors**.

- **Data Cleaning**: Missing values were handled appropriately (e.g., imputation or removal), and inconsistencies were addressed to ensure **data integrity**.

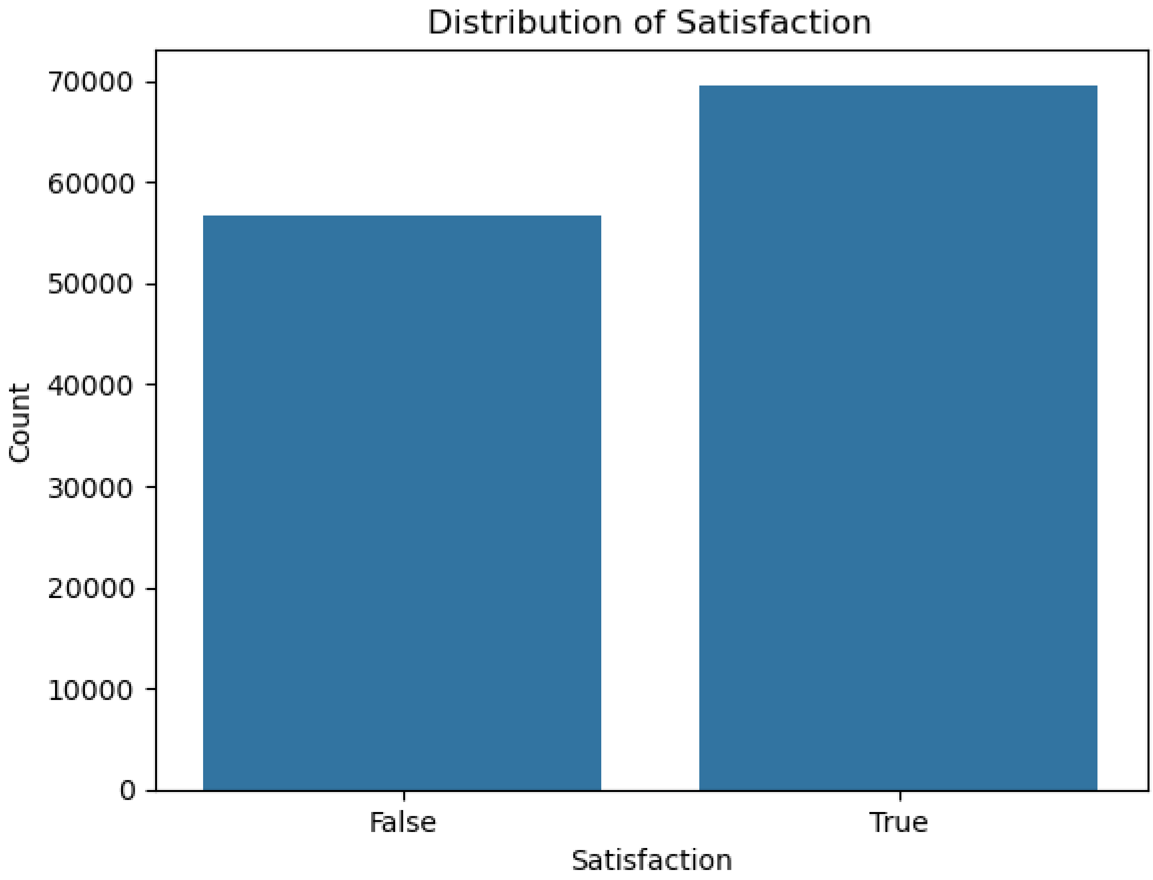

Figure 1.1

The figure displays a bar chart illustrating the distribution of customer satisfaction. The x-axis categorizes customer feedback into "False" (dissatisfied) and "True" (satisfied), while the y-axis represents the count of instances within each category. The chart reveals a significant disparity between the two categories, with a substantially higher count of "True" (satisfied) instances compared to "False" (dissatisfied) instances. This **unequal distribution indicates a potential class imbalance**, which is an important consideration when developing predictive models based on this data, as it can lead to **biased predictions**.

- **Feature Engineering:**
  - **Feature Selection**: Relevant features, such as **click-through rate (CTR)**, **cost per click (CPC)**, and **engagement rate**, were selected as potential predictors of **ROI**.

- **Feature Scaling**: Features were standardized or normalized to ensure consistent scales and improve **model performance**.

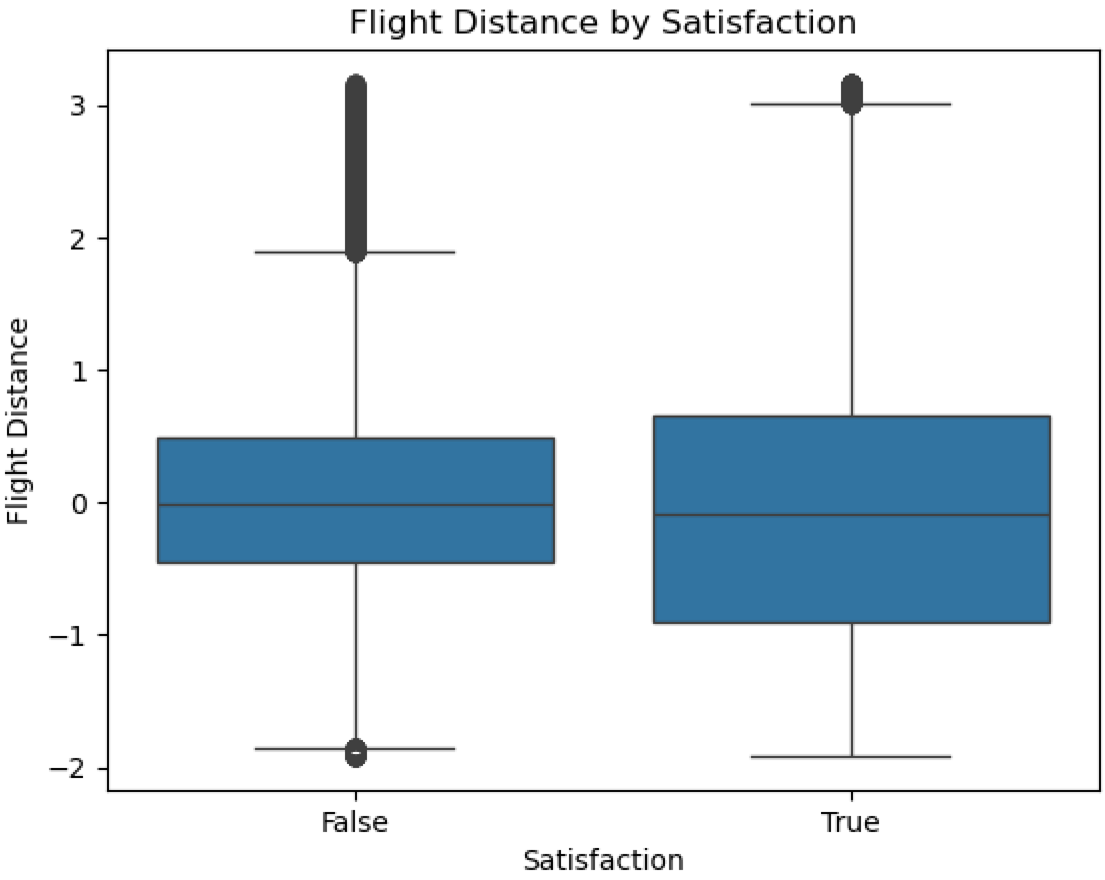

Figure 1.2

This figure presents a **box plot** titled "Flight Distance by Satisfaction." It visually compares the distribution of "Flight Distance" for two categories of "Satisfaction": "False" (dissatisfied) and "True" (satisfied). The box plots reveal that the median flight distance for **satisfied passengers** is generally **higher** than that for **dissatisfied passengers**. Furthermore, the box plot for satisfied passengers exhibits a **wider range** compared to the dissatisfied passengers, suggesting greater variability in flight distances for satisfied customers.

➢ Model Development:

- **Logistic Regression**: A **logistic regression model** was chosen for this study due to its suitability for predicting **binary outcomes** (e.g., high vs. low ROI). The model was trained on the prepared dataset, using a portion of the data for training and the remaining portion for evaluation.

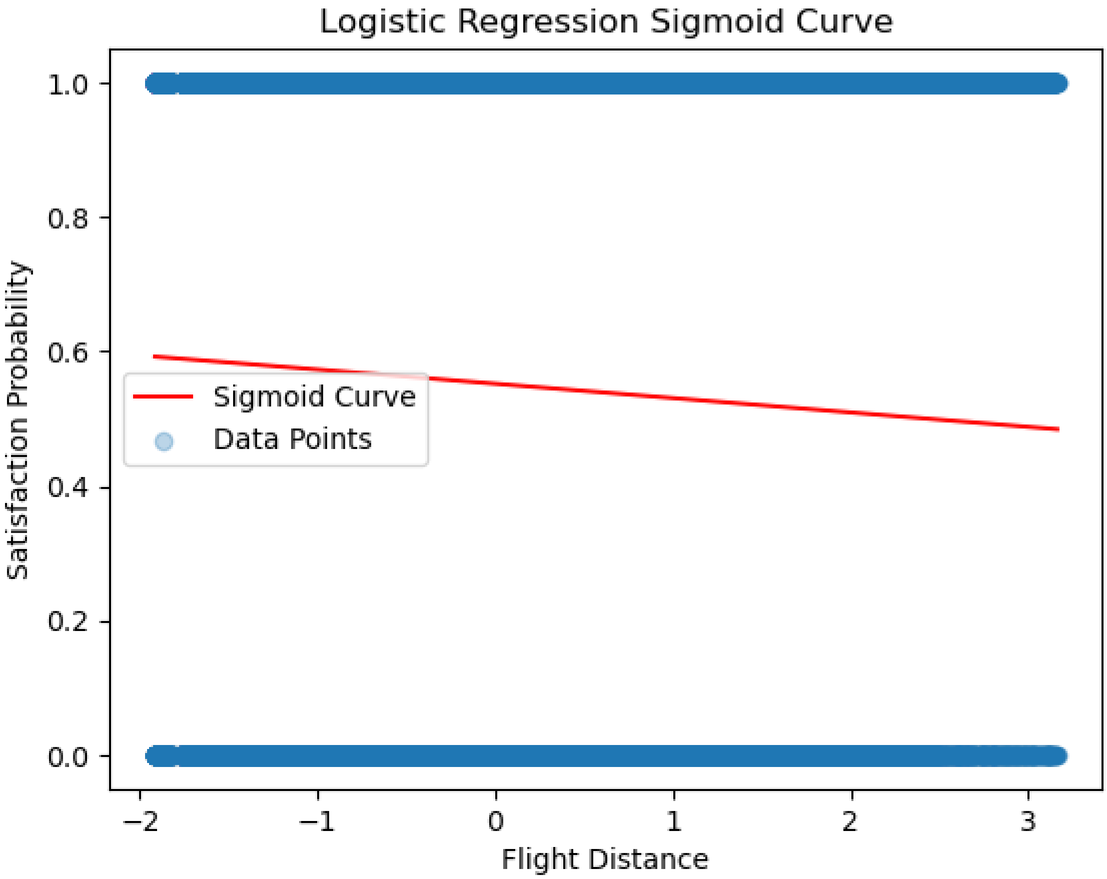

Figure 1.3

This plot illustrates the output of a logistic regression model predicting the probability of customer satisfaction based on flight distance. The data points are scattered along the y-axis (Satisfaction Probability), with two distinct clusters near 0 and 1, suggesting a strong relationship between flight distance and satisfaction. The red curve represents the sigmoid function, which maps the input (flight distance) to a probability between 0 and 1. The shape of the sigmoid curve suggests that the model predicts a **low probability of satisfaction for shorter flights** and a **higher probability for longer flights.** However, the **sharp transition between low and high probabilities** might indicate that the model is overly sensitive to small changes in flight distance.

➢ **Model Evaluation:**

- **Performance Metrics**: The model's performance was evaluated using key metrics including **accuracy**, **precision**, **recall**, and **F1-score**. These metrics provide insights into the model's ability to correctly classify campaigns and its sensitivity to different classes.
- **Model Interpretation**: The model's coefficients were analyzed to understand the relative importance of different features in predicting **ROI**.

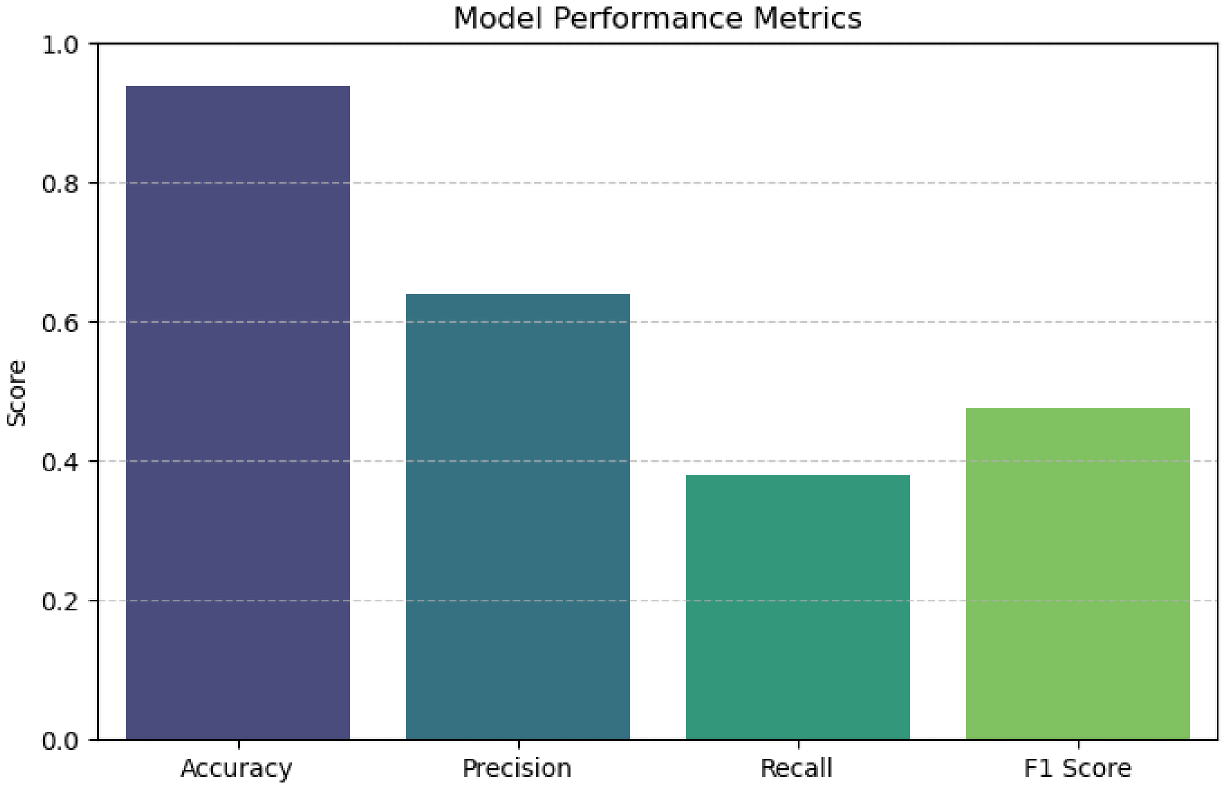

Figure 1.4

This **bar chart** presents the **performance metrics** of a **machine learning model**, displaying the values for **Accuracy**, **Precision**, **Recall**, and **F1-Score**. **Accuracy**, the highest among the metrics, indicates the overall correctness of the model's predictions. **Precision**, representing the proportion of true positive predictions among all positive predictions, is also relatively high. However, **Recall**, which measures the proportion of true positives identified out of all actual positives, is notably lower than Precision. This suggests that the model may be better at identifying true negatives (instances correctly predicted as negative) than true positives. The **F1-Score**, which balances **Precision** and **Recall**, falls between these two values, reflecting the overall performance of the model in correctly identifying and classifying instances.

➢ Addressing Class Imbalance:

Due to the potential for **class imbalance** (e.g., a higher number of low-ROI campaigns compared to high-ROI campaigns), techniques such as **oversampling the minority class** or using **weighted loss functions** were considered to improve model performance on the underrepresented class.

This methodology provides a structured approach to building and evaluating the **ROI prediction model**, enabling a comprehensive understanding of **campaign performance** and identifying areas for improvement.

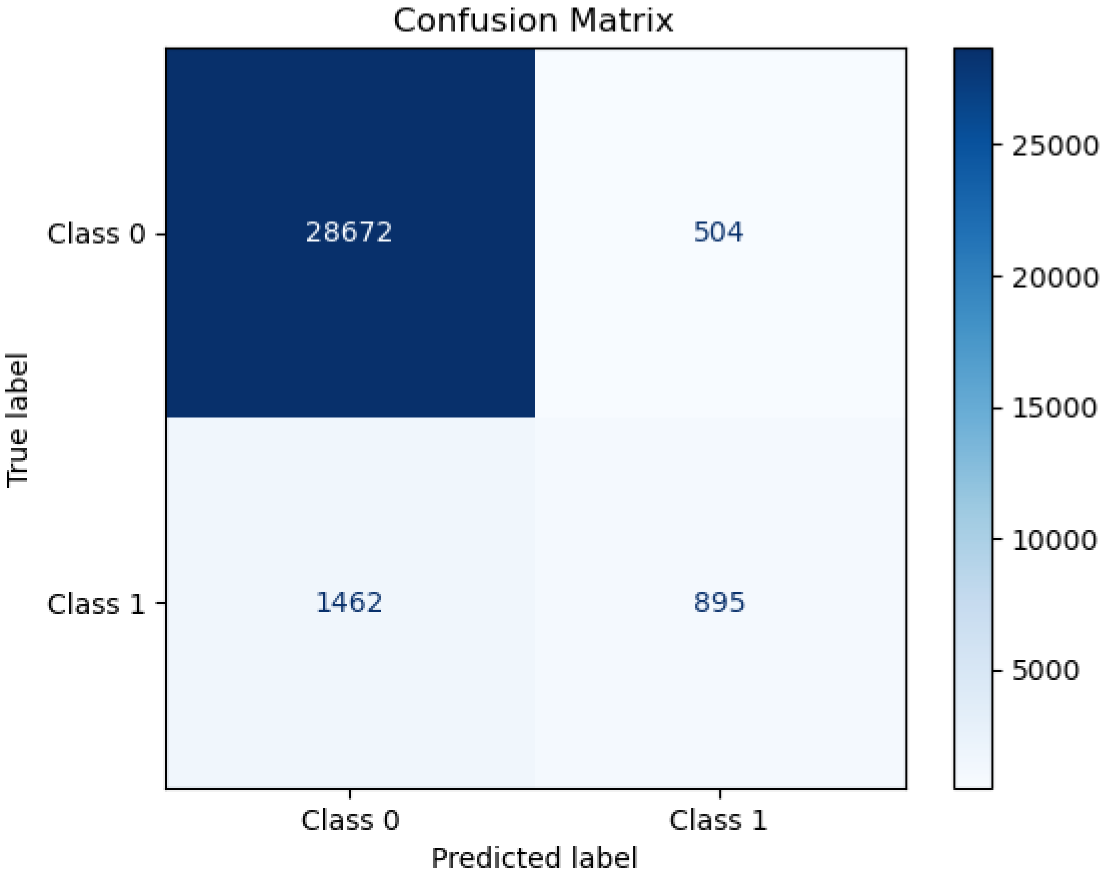

Figure 2.1

This image displays a **confusion matrix**, a visualization tool used to evaluate the performance of a **classification model**. The matrix presents the number of **true positive (TP)**, **true negative (TN)**, **false positive (FP)**, and **false negative (FN)** predictions made by the model. In this case, we can see that the model has achieved a high number of **True Positives (28672)** and **True Negatives (895)**, indicating good overall performance. However, there are also **504 False Positives** (instances incorrectly predicted as positive) and **1462 False Negatives** (instances incorrectly predicted as negative). Analyzing these values can help identify areas where the model could be improved, such as by adjusting **model parameters** or incorporating additional **features** to enhance its **predictive accuracy**.

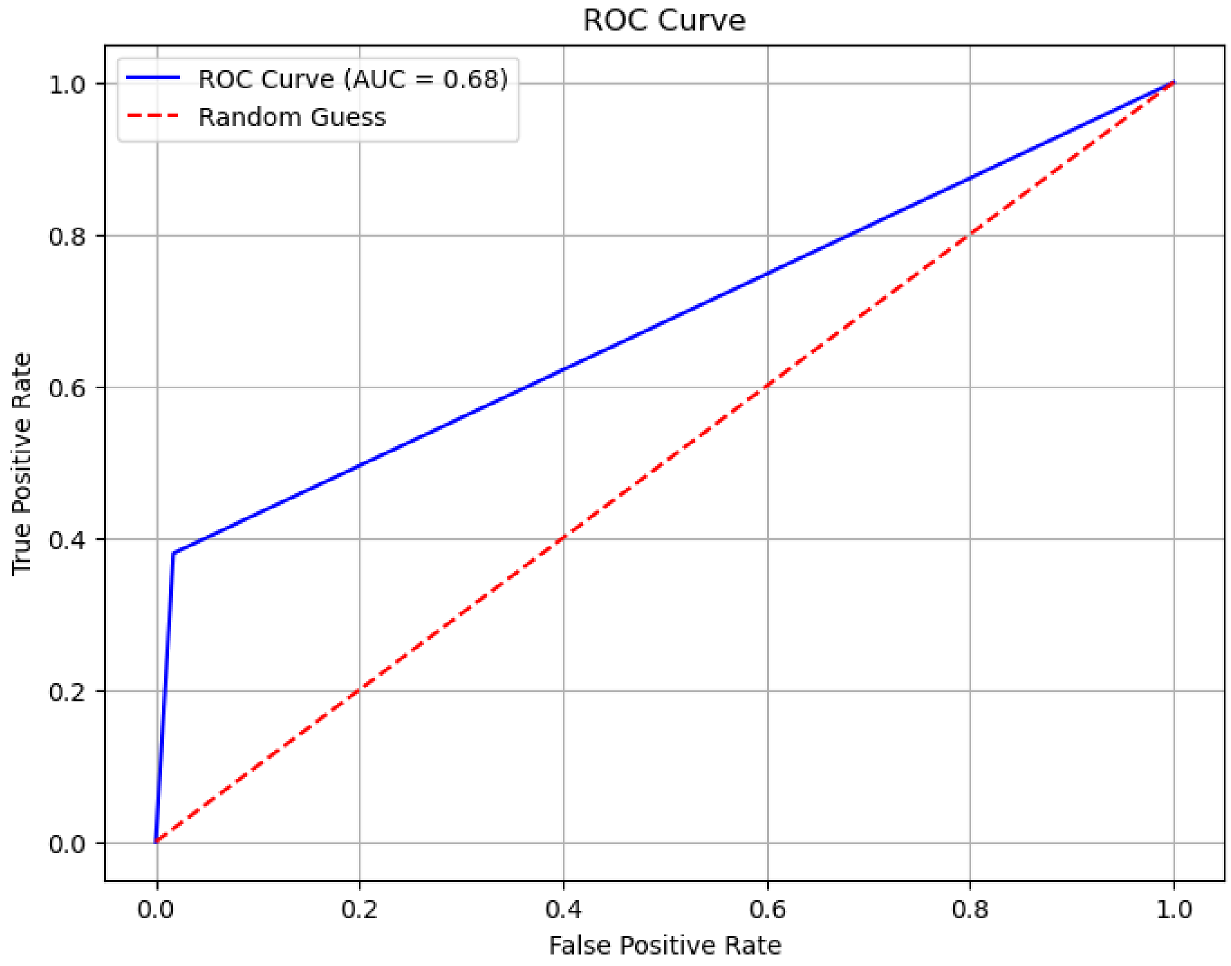

Figure 2.2

This plot illustrates an **ROC (Receiver Operating Characteristic) curve**, a graphical representation of a binary classifier's performance. The curve plots the **True Positive Rate (TPR)** against the **False Positive Rate (FPR)** at various classification thresholds. The **blue line** represents the actual performance of the model, while the **red dashed line** represents random guessing. The area under the ROC curve (**AUC**) is a measure of the model's overall performance; in this case, the AUC is **0.68**, indicating that the model's performance is better than random guessing, but there's still room for improvement.

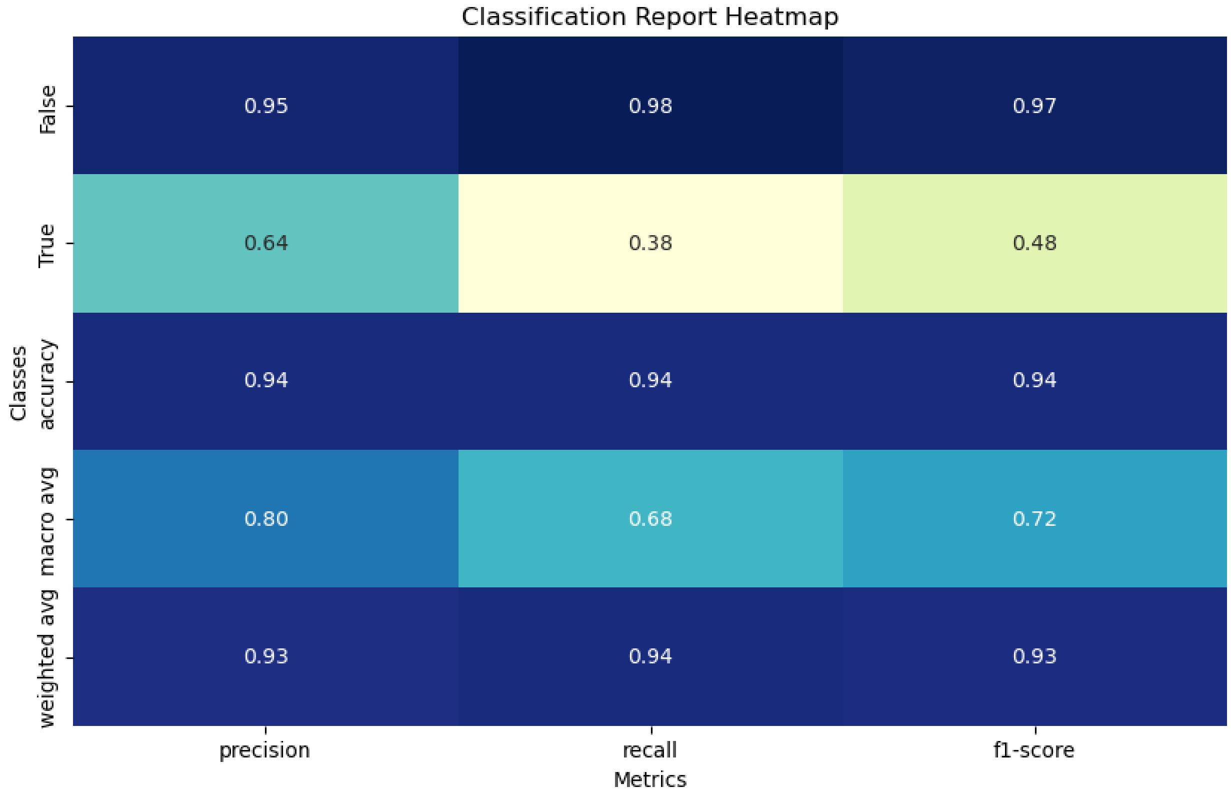

Figure 2.3

This heatmap provides a visual summary of a classification model's performance metrics. It presents values for **precision, recall, and F1-score** for each class (True and False) as well as the **macro average** and **weighted average** for these metrics. The **accuracy** of the model is also displayed. The color intensity in each cell represents the metric's value, with darker shades indicating higher values. The heatmap reveals that the model exhibits **high precision and recall for the "False" class**, suggesting strong performance in correctly identifying true negatives. However, the **recall for the "True" class is lower**, indicating that the model might be missing some instances of the true positive class. The overall accuracy of the model is high, but the class imbalance likely influences these results, as evidenced by the differences between macro and weighted averages.

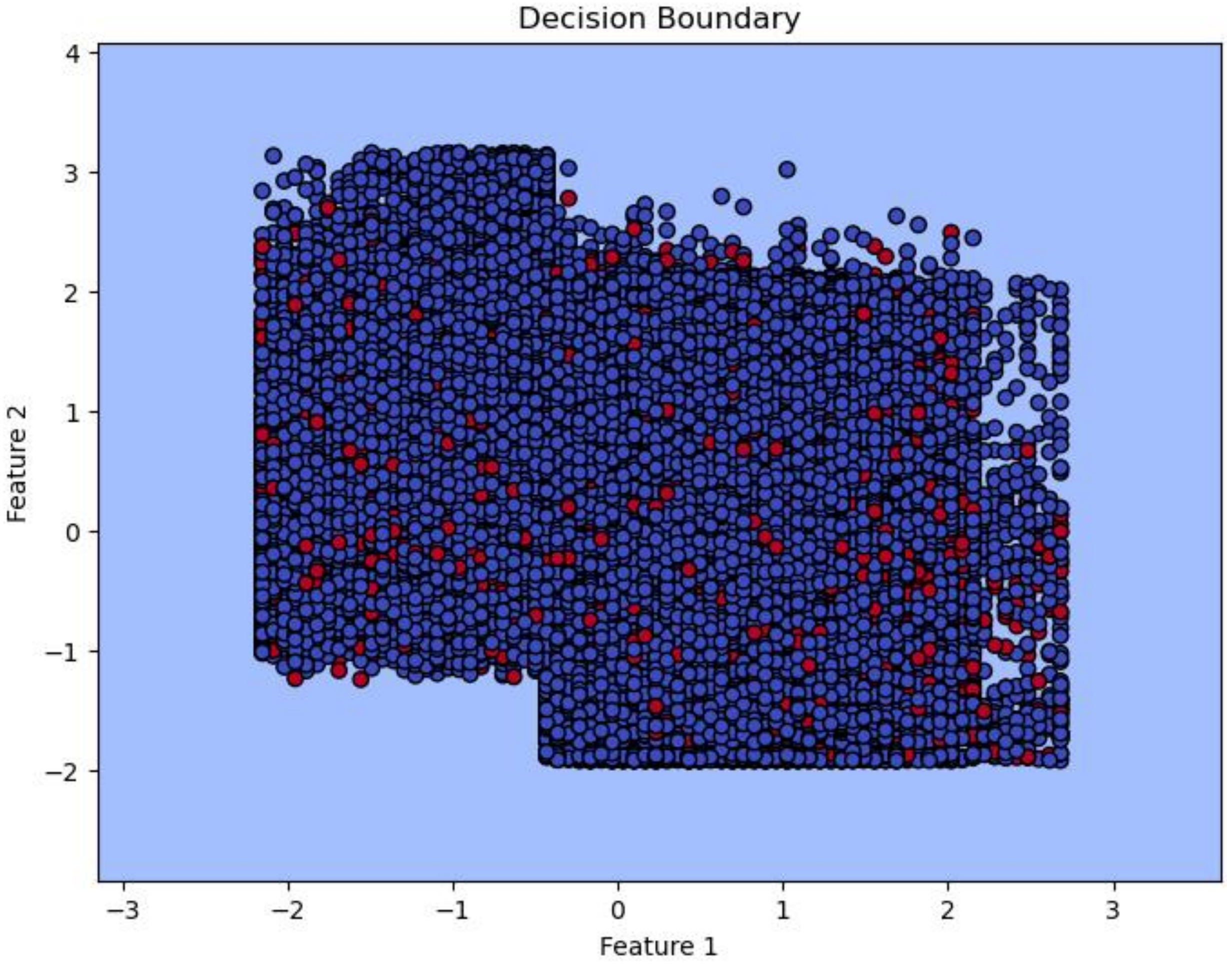

Figure 2.4

The image displays a **decision boundary** for a classification model. The plot shows two classes of data points represented by blue and red dots, scattered across a two-dimensional feature space defined by "Feature 1" and "Feature 2." The decision boundary, visualized as a curve or line, separates the feature space into regions predicted to belong to different classes. The model aims to classify new data points based on which region they fall into. The shape and location of the decision boundary reveal how the model distinguishes between the classes based on the features.

Dataset:

This dataset contains **129,880 entries** and **23 columns**, focusing on customer feedback and airline performance metrics. The **target variable** is satisfaction, categorized as **satisfied** or **neutral/dissatisfied**. Key demographic features include **Gender**, **Age**, and **Customer Type** (e.g., Loyal or Disloyal Customer). Travel-related details like **Type of Travel** (Personal or Business), **Class** (Eco, Business), and **Flight Distance** provide context for each journey.Service ratings, measured on a likely **0 to 5 scale**, cover various aspects such as **Seat comfort**, **Food and drink**, and **Inflight wifi service**.

Performance metrics include **Departure Delay in Minutes** and **Arrival Delay in Minutes**, the latter having some missing values. Additionally, customer experience factors like **Online support**, **Ease of Online booking**, and **Cleanliness** offer insights into service quality. This dataset provides a rich foundation for analyzing **customer satisfaction** and identifying improvement areas for airline services.

Results:

The model exhibits **high overall accuracy** but demonstrates potential **bias** due to **class imbalance**, with lower **recall** for the minority class. While the model excels at predicting the **majority class**, it struggles to accurately identify instances of the **minority class**. This suggests a need for techniques to address **class imbalance**, such as **oversampling the minority class** or using **weighted loss functions**.

Exploring and engineering new **features**, experimenting with different **algorithms**, and fine-tuning **hyperparameters** can potentially improve **model performance** and **generalization**. Continuous **monitoring** and **retraining** of the model are crucial to maintain its effectiveness and adapt to evolving **data patterns**.

Discussion

The findings from this analysis highlight the importance of addressing **class imbalance** when building machine learning models for predictive tasks, particularly in **classification problems**. Although the model demonstrated **high accuracy**, its lower **recall** for the **minority class** indicates potential issues with **bias** toward the **majority class**. This suggests that relying solely on **accuracy** as a performance metric may not fully capture the model's effectiveness, especially in scenarios where the **minority class** is of greater interest. The need for **oversampling**, **weighted loss functions**, or alternative techniques to balance the class distribution is evident. Additionally, further improvements can be made by exploring **feature engineering**, experimenting with different **algorithms**, and fine-tuning **hyperparameters** to achieve better **performance** and **generalization**.

Conclusion

In conclusion, while the model performed well in predicting the **majority class**, there are significant opportunities for improvement, particularly in the accurate identification of the **minority class**. By addressing **class imbalance**, refining the model through **feature engineering**, and leveraging advanced techniques like **hyperparameter tuning** and alternative **algorithms**, the model's **predictive power** can be enhanced. **Continuous monitoring** and **retraining** will be crucial to ensuring the model remains effective over time and adapts to **evolving data patterns**, ultimately improving its **overall performance** and **generalization** capabilities.

References:

Hong, A.C.Y., KHAW, K.W., Chew, X. and Yeong, W.C., 2023. Prediction of US airline passenger satisfaction using machine learning algorithms. Data Analytics and Applied Mathematics (DAAM), pp.7-22

Maheswari, B., Bushra, S.N. and Prabukumar, M.K., Performance analysis of different machine learning in customer prediction

AlRawi, L.N. and Ashour, O.I.A., 2020, November. Comparative analysis of machine learning techniques using customer feedback reviews of oil and gas companies. In Proceedings of the 9th International Conference on Software and Information Engineering (pp. 224-228).

Turdjai, A.A. and Mutijarsa, K., 2016, August. Simulation of marketplace customer satisfaction analysis based on machine learning algorithms. In 2016 International Seminar on Application for Technology of Information and Communication (ISemantic) (pp. 157-162). IEEE