Predicting Customer purchase method using machine learning

A Research Report

Submitted in partial fulfillment of the requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Submitted by

Shahrier Khan	**2019200000003**
Sadmanul Hoque	**2021100000021**
Md. Sajjatul Islam Jani	**2021100000029**

Supervised by

Nahid Hasan

Lecturer
Department of Computer Science and Engineering
Southeast University, Bangladesh

Department of Computer Science and Engineering
Southeast University, Bangladesh

Dhaka, Bangladesh

February, 2025

Letter of Transmittal

February, 2025

The Chairman,
Department of Computer Science and Engineering
Southeast University, Bangladesh
Tejgaon, Dhaka

Through: Supervisor, Nahid Hasan

Subject:

Dear Sir,

Enclosed please find our research paper titled "Predicting Customer purchase method using machine learning' Purchase Intention Based on Bangladeshi Branding Clothes Using Machine Learning." This paper explores the use of machine learning techniques to analyze factors influencing Bangladeshi university students' purchase intentions toward branded clothing.

We believe our findings contribute significantly to the fields of consumer behavior and machine learning, offering valuable insights for understanding purchasing intentions within a specific demographic.

Thank you for considering our work.

Sincerely Yours,

Shahrier Khan
2019200000003

Sadmanul Hoque
2021100000021

Md. Sajjatul Islam Jani
2021100000029

Supervisor:

Nahid Hasan
Lecturer & Supervisor
Department of Computer Science and Engineering
Southeast University, Bangladesh

CANDIDATE'S DECLARATION

We, hereby, declare that the thesis presented in this report is the outcome of the investigation performed by us under the supervision of Nahid Hasan, Lecturer, Department of Computer Science and Engineering, Southeast University, Bangladesh. The work was done through CSE459: Research Methodology course, in accordance with the course curriculum of the Department for the Bachelor of Science in Computer Science and Engineering program.

It is also declared that neither this research nor any part thereof has been submitted anywhere else for the award of any degree, diploma or other qualifications.

Shahrier Khan
2019200000003

Sadmanul Hoque
2021100000021

Md. Sajjatul Islam Jani
2021100000029

CERTIFICATION

This research titled, “**Predicting Customer purchase method using machine learning**”, submitted by the group as mentioned below has been accepted as satisfactory in partial fulfillment of the requirements for the degree B.Sc. in Computer Science and Engineering in February, 2025.

Group Members:

Shahrier Khan	**2019200000003**
Sadmanul Hoque	**2021100000021**
Md. Sajjatul Islam Jani	**2021100000029**

Supervisor:

Nahid Hasan
Lecturer & Supervisor
Department of Computer Science and Engineering
Southeast University, Bangladesh

Shahriar Manzoor
Associate Professor & Chairman
Department of Computer Science and Engineering
Southeast University, Bangladesh

ABSTRACT

In this paper, we used machine learning models to predict how customers make purchases: online or offline. We implemented a model using a dataset with customer demographics, purchase history, and other features, and we evaluated the performance of multiple models, including Random Forest, SVM, XGBoost, Logistic Regression, Decision Trees, and KNN. We applied one-hot encoding, label encoding, and SMOTE to process our dataset for the imbalanced class. The Random Forest model achieved the highest accuracy of 95.05% among all the models. We learn from our findings that machine learning can be trained on specific customer data to run appropriate targeted marketing on the assumption that the customer will purchase based on predicted behavior.

Keywords: Machine Learning, Customer Purchase Behavior, Random Forest, Classification, Fashion Industry.

Contents

Chapter 1

Introduction

1. Customer purchase behavior is critical to businesses because it helps stakeholders make better decisions. Knowing whether a customer will buy a product online or offline can give businesses insights into which type of advertisement will boost their sales. Such problems are a promising avenue for machine learning methods to solve, given that machine learning methods can detect complex patterns in data. In this paper, we predict customer purchase methods us- ing machine learning models. We analyze customer demo- graphics, purchase history, and other relevant features in the dataset. To this end, we evaluate a variety of algorithms, including Random Forest, Support Vector Machine (SVM), XGBoost, Logistic Regression, Decision Trees, and K-Nearest Neighbors (KNN). Techniques, including one-hot encoding, SMOTE (Synthetic Minority Oversampling Technique), and label encoding, are used to preprocess the data to handle class imbalance

Chapter 2

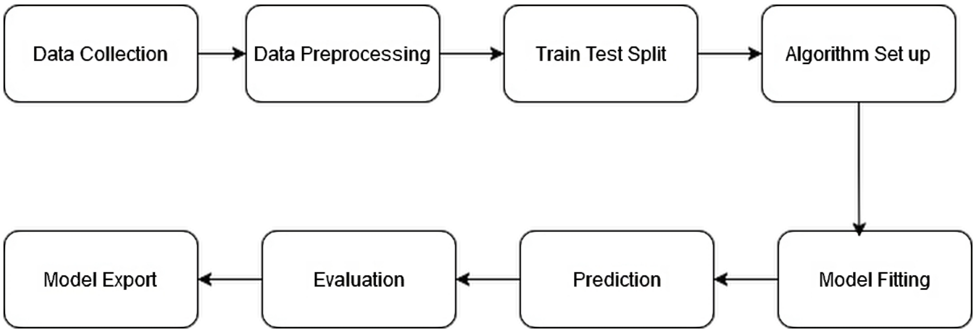
Objective

2.1 Objective

1. Predict Customer Purchase Method
2. Evaluate Machine Learning Models
3. Address Data Imbalance
4. Prepare Data
5. Compare Model Performance
6. Business Application

2.2 Material and Methods

Data collection and analysis have been performed for predicting the purchasing data. The research methodology of this study is as follows.



Chapter 3

Sample, Data Preprocessing, Analysis and Discussion

3.1 Sample

To carry out this research, the sample size of the data set of 9000 customers in Bangladesh was obtained. This study makes use of a dataset composed of customer demographics along with purchase history and other features like “Occasion”, “Customer Age Group” and “Selling Price”. ’Purchase Mode’ is the target variable that states if the purchase is done online or offline. The dataset contains both categorical and numerical features, therefore, this can be used with many different machine learning algorithms.

3.2 Data Preprocessing

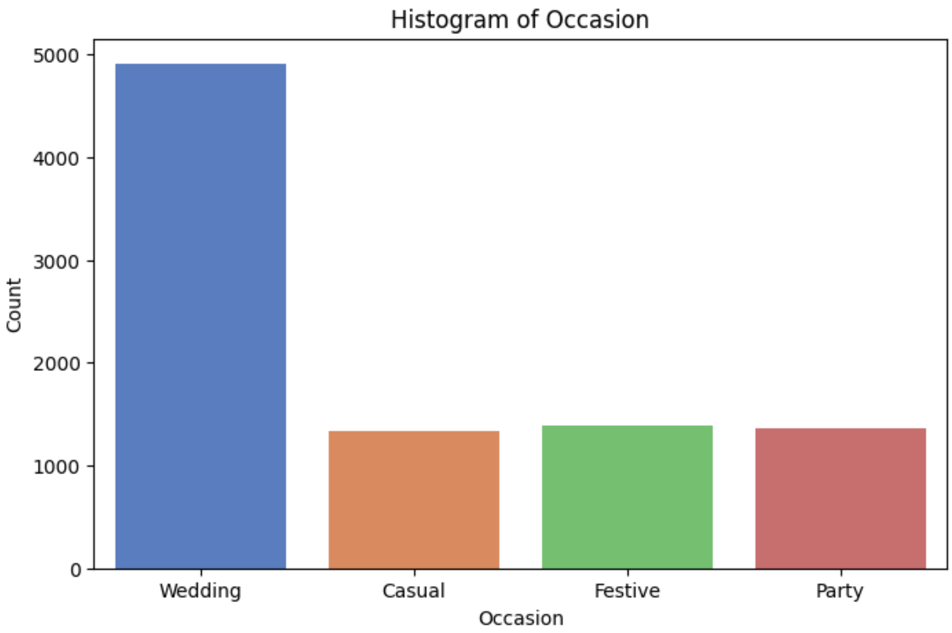
Each data set was preprocessed before training the models to be compatible with machine learning algorithms. The following steps were taken: Encoding Categorical Variables: We hot encoded categorical features like ’Occasion’ and ’Customer Age Group’, and we label encoded the target variable ’Purchase Mode’. Scaling Numerical Features: To make features with numeric values all on the same scale, features such as “selling price” were scaled using StandardScaler. Handling Class Imbalance: The dataset was imbalanced with one subclass (e.g., online purchases) having more instances than the other (e.g., offline purchases). To do that, we want to balance the data, so we applied SMOTE (Synthetic Minority Oversampling Technique).

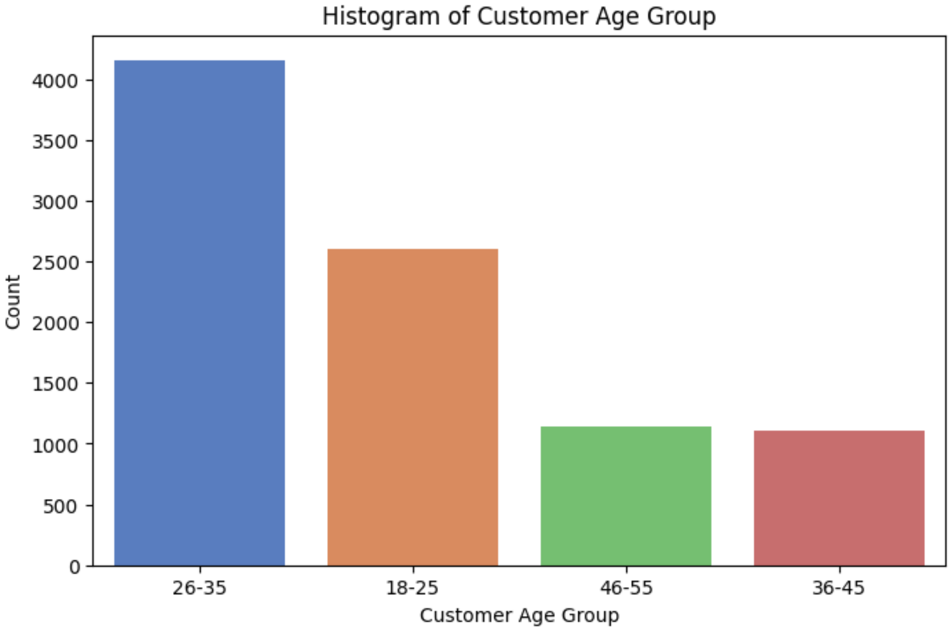
3.3 Analysis and Discussion

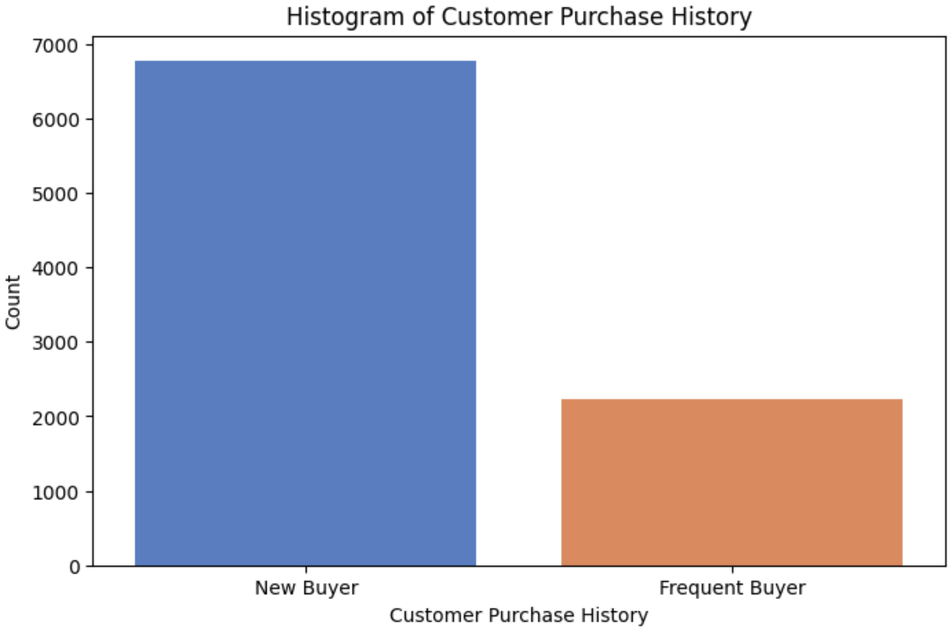
Regarding the original class distribution of purchase modes, around 8000 purchases were made online, while only about 1000 were offline, showing a clear imbalance with a strong preference for online shopping. This imbalance could skew predictive models if not addressed, potentially leading to biases towards predicting online purchases. After balancing the dataset, as shown in the balanced class distribution, both online and offline purchase counts were equalized to approximately 8000 each, using techniques like SMOTE. This balancing is crucial for training machine learning models to ensure fair and unbiased predictions.

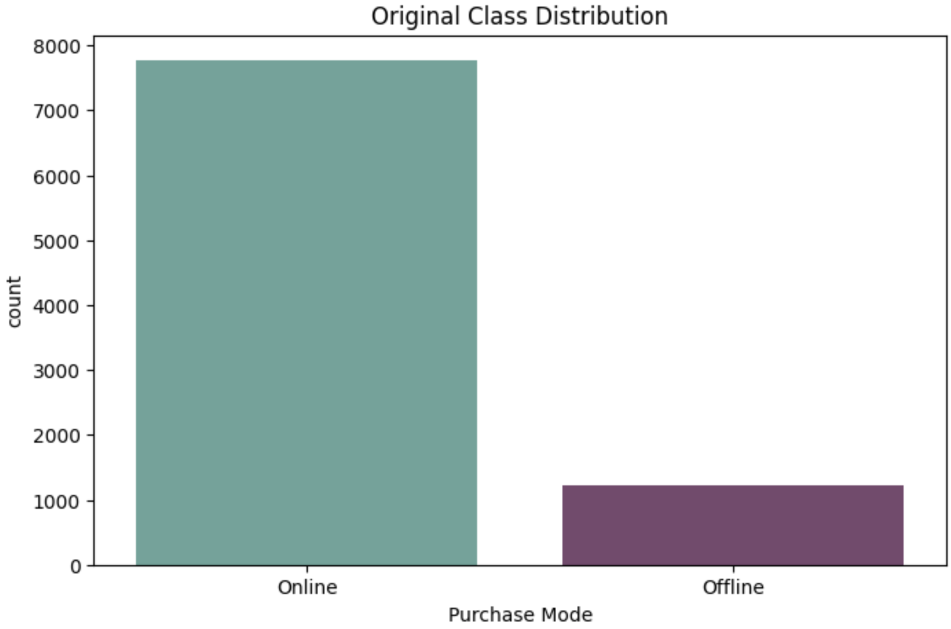
The histogram of occasions reveals that weddings are the most common reason for purchases, with around 5000 instances, suggesting that they are significant drivers of sales due to their associated high expenditure. Casual, festive, and party occasions each account for about 1000 purchases, indicating these are less frequent but still relevant for marketing. This insight suggests targeted marketing towards wedding-related products could be particularly effective, with additional seasonal promotions for other occasions.

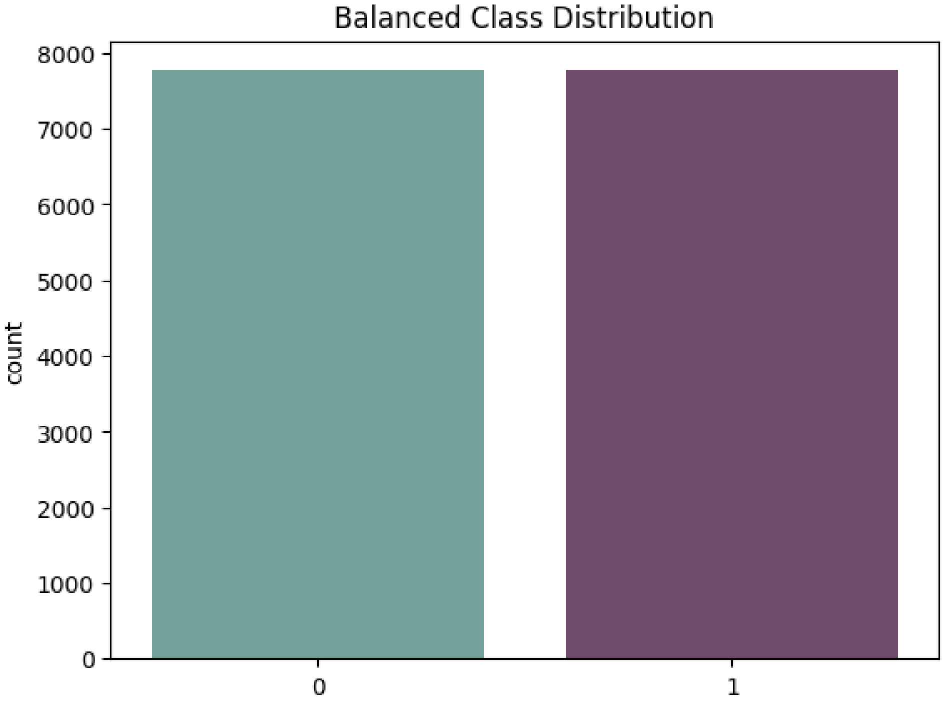
In terms of customer age groups, the 26-35 age range has the highest purchase count at around 4000, indicating this demographic is the most active in purchasing. The 18-25 group follows with about 3000 customers, showing a strong younger adult market. However, there's a noticeable drop in the 36-45 and 46-55 age groups, each with approximately 1000 customers, suggesting less engagement from these older demographics. Marketing efforts should focus on appealing to the young adult groups while exploring strategies to re-engage or better understand the purchasing behavior of older customers.




Histogram of Customer Age Group
Count
4000
3500
3000
2500
2000
1500
1000
500
0
26-35
18-25
46-55
36-45
Customer Age Group


Histogram of Customer Purchase History
Count
7000
6000
5000
4000
3000
2000
1000
0
New Buyer
Frequent Buyer
Customer Purchase History


Original Class Distribution
count
8000
7000
6000
5000
4000
3000
2000
1000
0
Online
Offline
Purchase Mode


Balanced Class Distribution
count
8000
7000
6000
5000
4000
3000
2000
1000
0
0
1

Chapter 4

Model Development and Training

4.1 Model Development and Training

After preprocessing the data, we moved on to the development and training of machine learning models to predict the 'Purchase Mode'. Various algorithms were considered to find the best fit for our dataset, considering both the balanced class distribution and the diverse features available. Several models were evaluated for this task, including:

1. Logistic Regression: As a baseline model due to its simplicity and interpretability.

2. Random Forest: For its ability to handle both categorical and numerical data and to provide feature importance.

3. Gradient Boosting: Specifically XGBoost, known for its performance in classification tasks.

4. Support Vector Machines (SVM): To leverage its effectiveness in high-dimensional spaces.

The choice of these models was driven by their known strengths in handling classification problems with mixed data types and their ability to manage class imbalance post-SMOTE application.

The training process involved the following steps:

- Splitting the Data: The dataset was split into training (80

- Hyperparameter Tuning: We performed hyperparameter tuning using techniques like Grid Search or Random Search to find the optimal parameters for each model. For instance, for Random Forest, we tuned parameters like the number of trees, max depth, and min samples split. For XGBoost, we adjusted learning rate, max depth, and number of estimators.

- Cross-Validation: To prevent overfitting and to get a robust estimate of model performance, k-fold cross-validation (with k=5 or 10) was used during the training phase.

- Model Evaluation: After initial training, models were evaluated using metrics like accuracy, precision, recall, F1-score, and the Area Under the ROC Curve (AUC-ROC) to ensure a comprehensive understanding of their performance across different aspects of classification.

Here are some insights from the training phase:

1. Logistic Regression: Provided a simple baseline, but struggled with capturing the complexity of the data due to its linear nature.

2. Random Forest: Showed good performance, benefited from the feature importance analysis which gave insights into which demographic and purchase history features were most predictive.

3. XGBoost: Outperformed other models in terms of accuracy and F1-score, highlighting the power of gradient boosting in handling complex relationships in the data.

4. SVM: Performed well, especially after tuning the kernel and C parameter, but was computationally intensive for large datasets.

Chapter 5

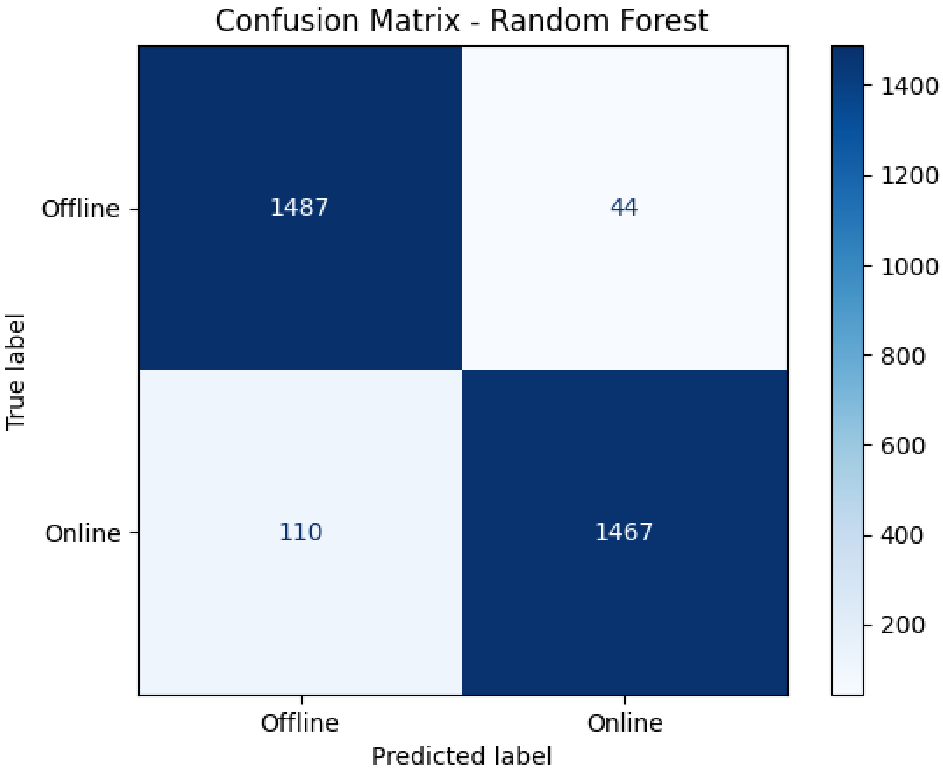
Model Evaluation

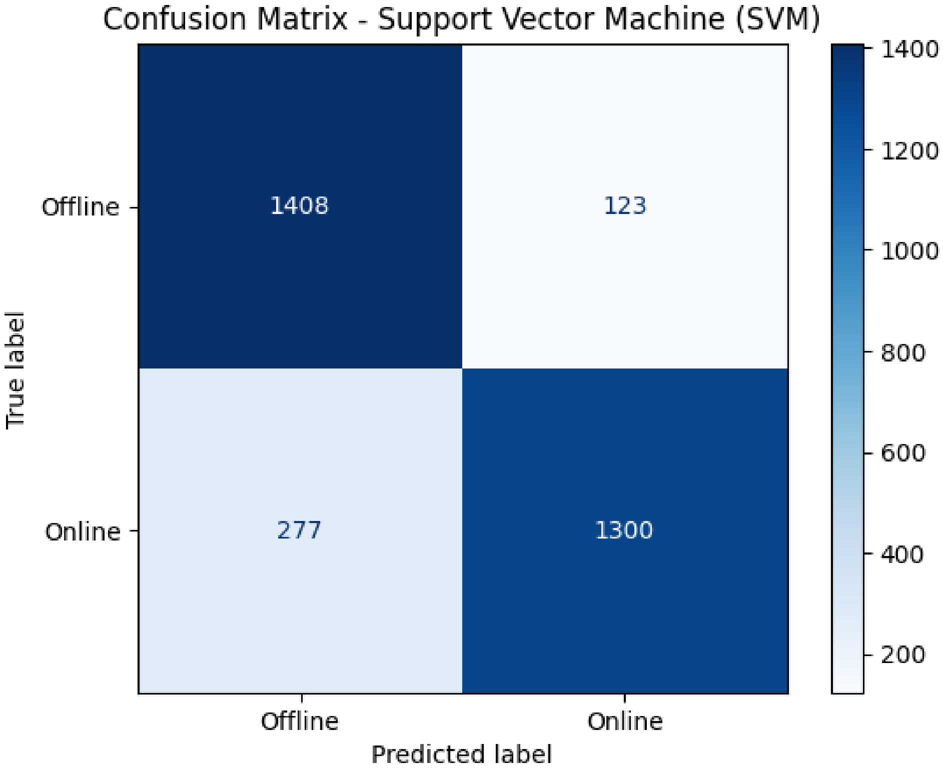
5.1 Model Evaluation

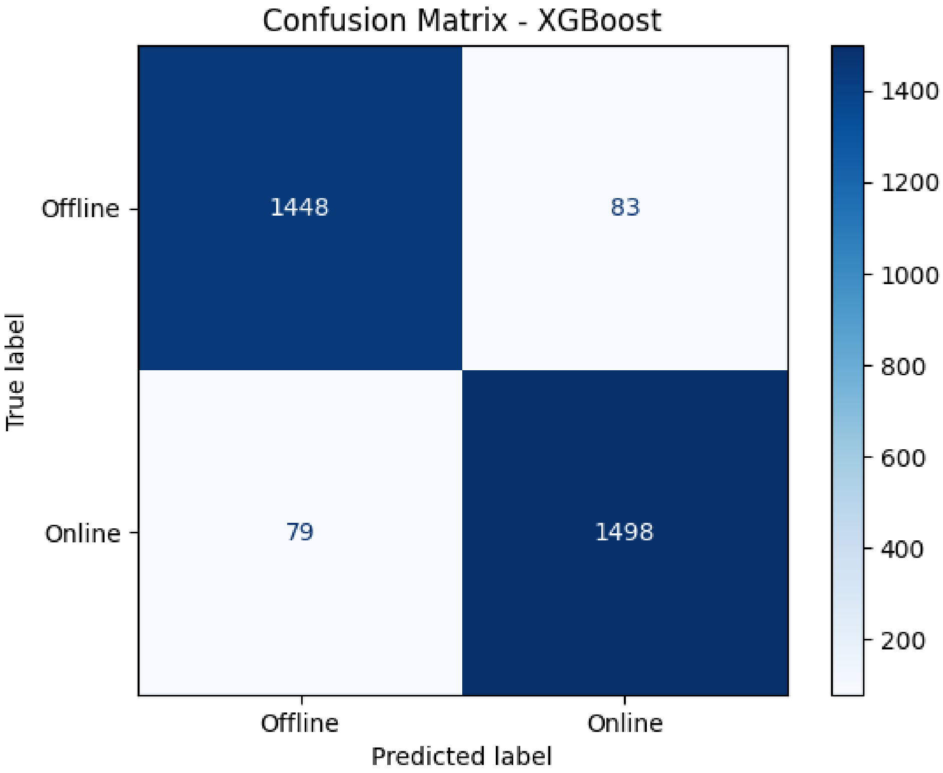
After preprocessing the data, we moved on to the development and training of machine learning models to predict the ’Purchase Mode’. Various algorithms were considered to find the best fit for our dataset, considering both the balanced class distribution and the diverse features available. Several models were evaluated for this task, including:

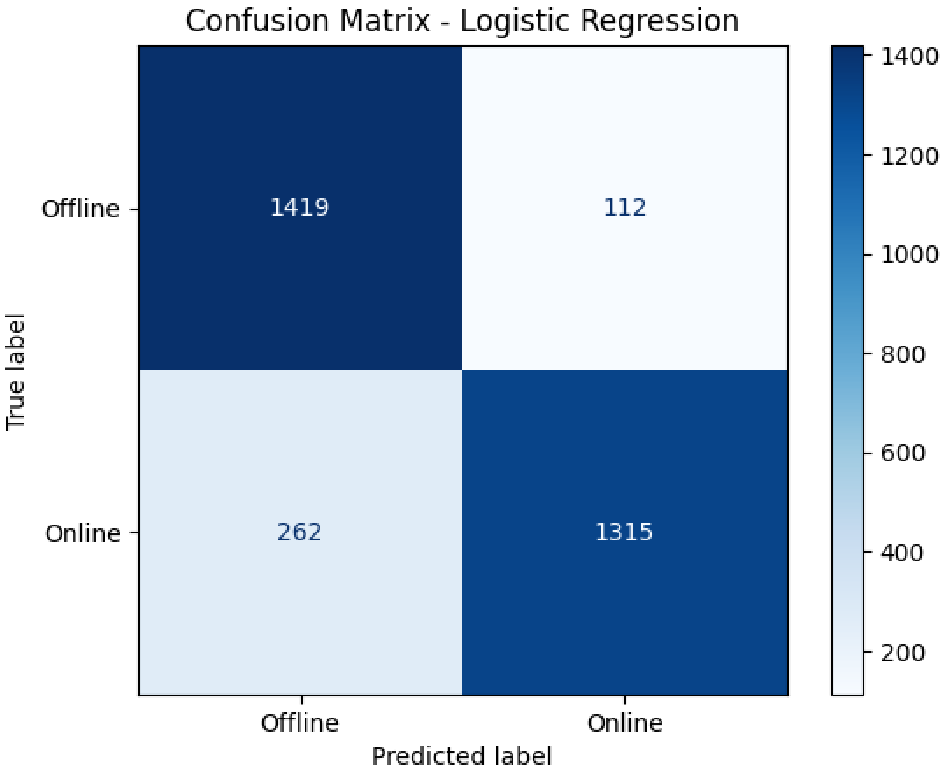
Table 5.1: Model Performance Comparison

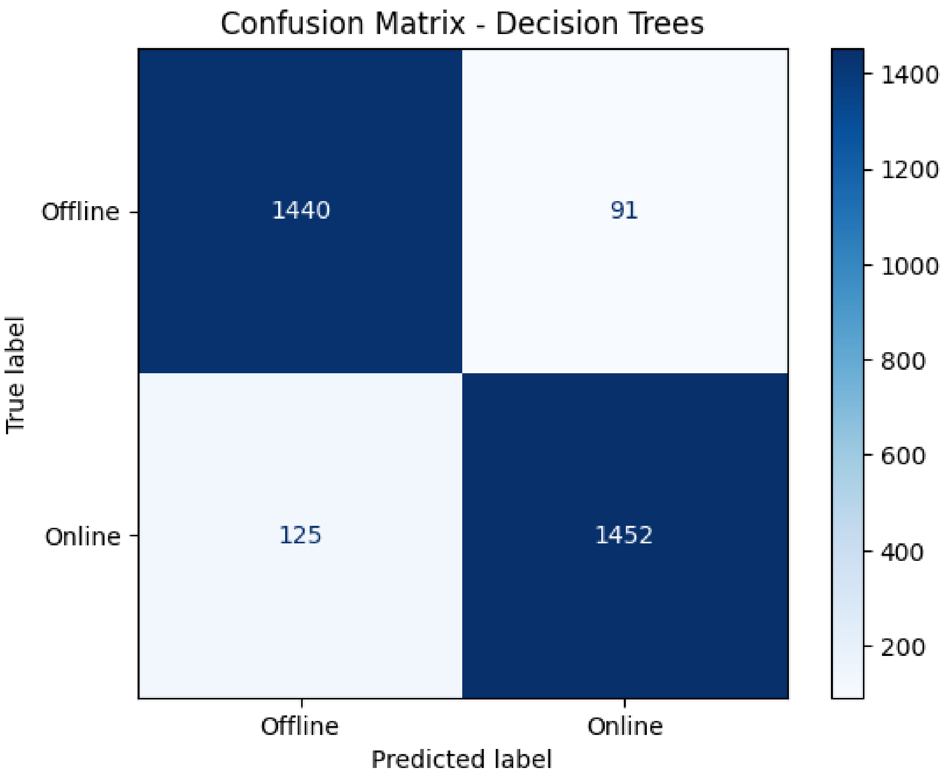
Model	Accuracy	Precision	Recall	F1-Score
Random Forest	95.05%	0.95	0.95	0.95
SVM	87.13%	0.88	0.87	0.87
XGBoost	94.79%	0.95	0.95	0.95
Logistic Regression	87.97%	0.88	0.88	0.88
Decision Trees	93.05%	0.93	0.93	0.93
KNN	84.36%	0.86	0.84	0.85

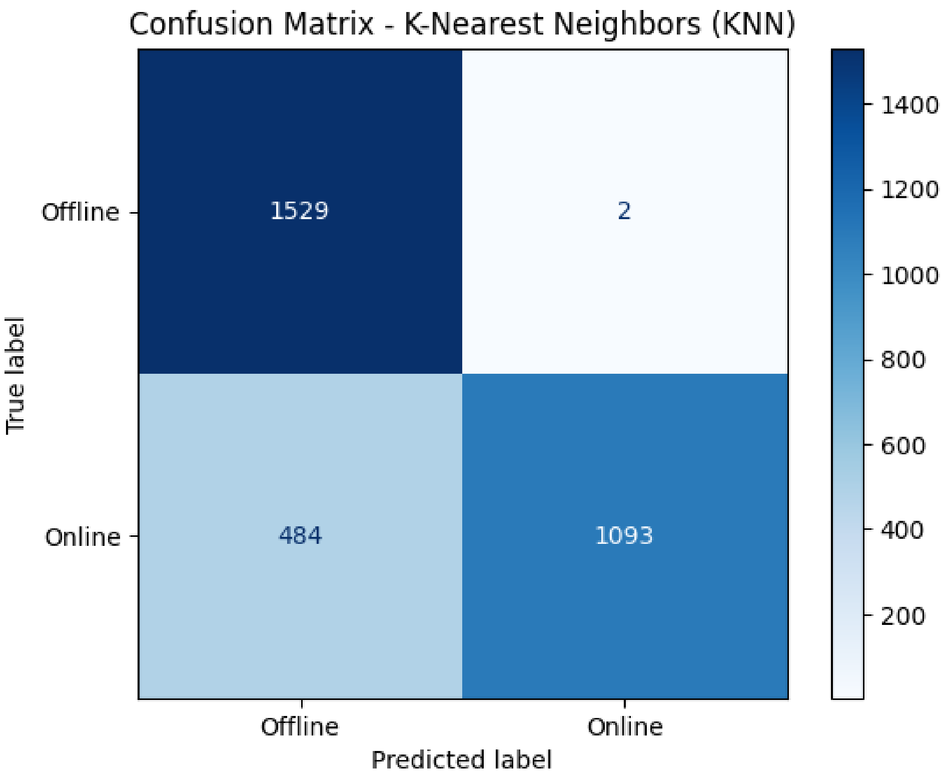

Confusion Matrix - Random Forest
True label
Offline
Online
1487
44
110
1467
Offline
Online
Predicted label
1400
1200
1000
800
600
400
200

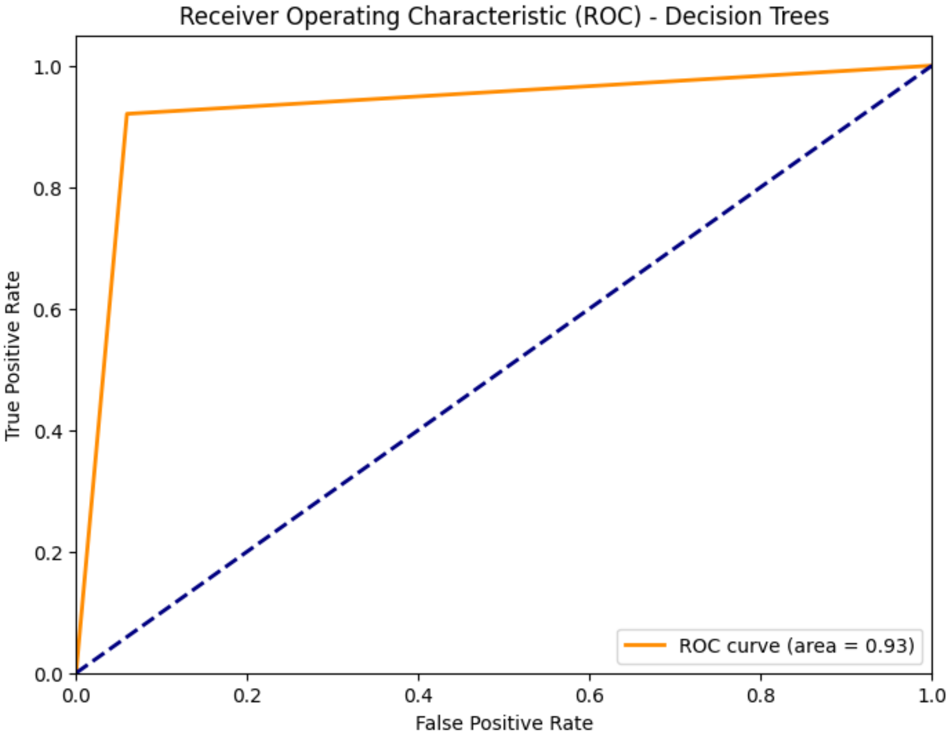

Confusion Matrix - Support Vector Machine (SVM)
True label
Offline
Online
1408
123
277
1300
Offline
Online
Predicted label
1400
1200
1000
800
600
400
200

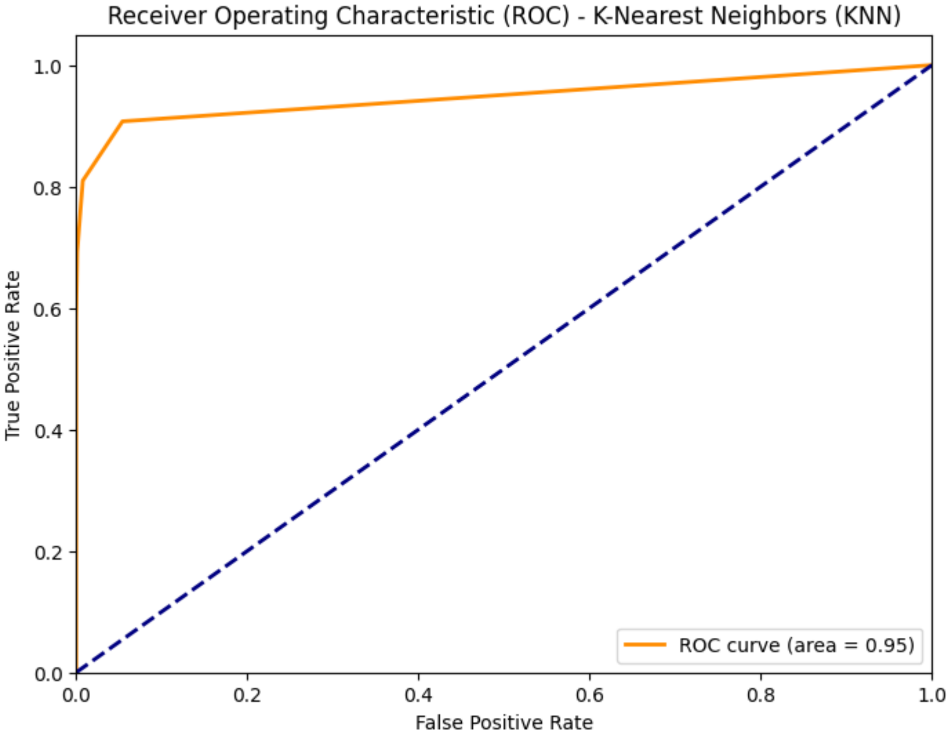

Confusion Matrix - XGBoost
True label
Predicted label
Offline
Online
Offline
Online
1448
83
79
1498
1400
1200
1000
800
600
400
200

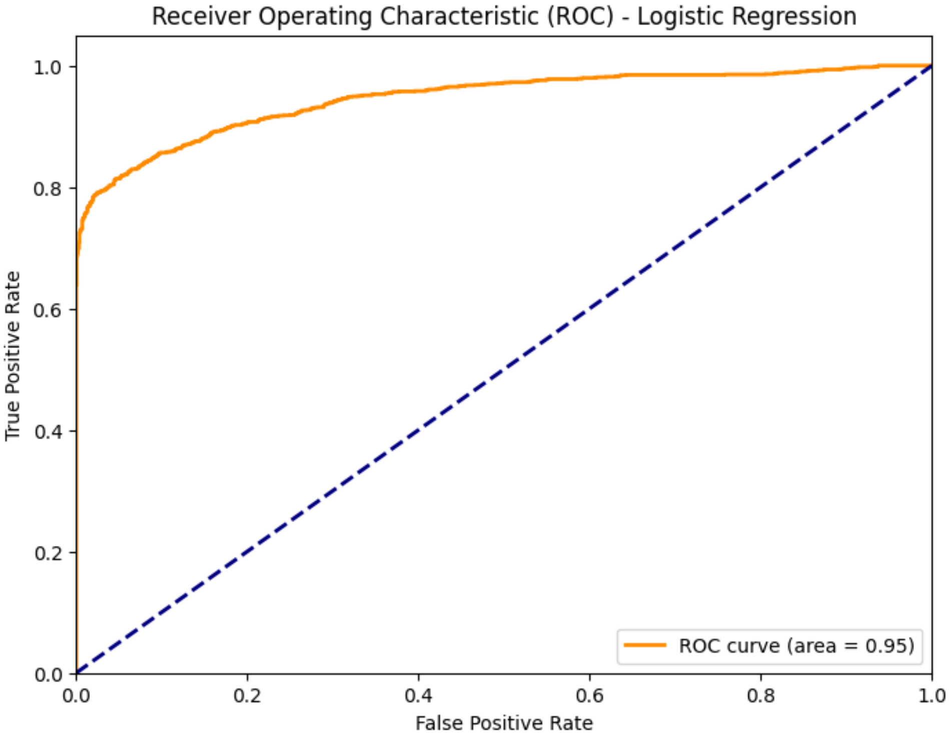

Confusion Matrix - Logistic Regression
True label
Predicted label
Offline
Online
Offline
Online
1419
112
262
1315
1400
1200
1000
800
600
400
200

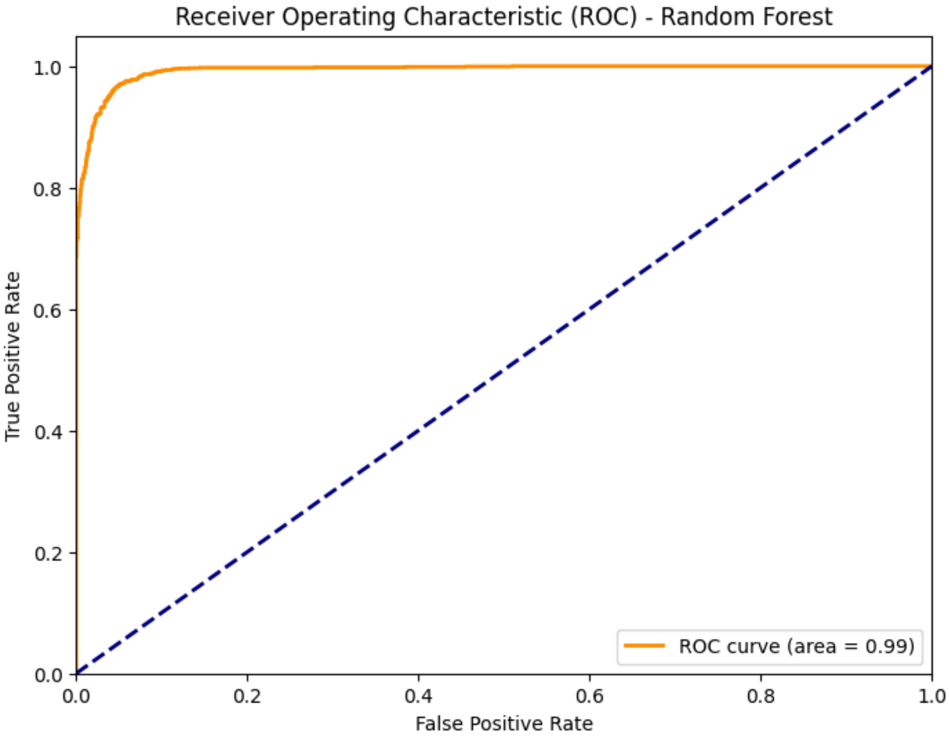

Confusion Matrix - Decision Trees
True label
Offline
Online
1440
91
125
1452
Offline
Online
Predicted label
1400
1200
1000
800
600
400
200

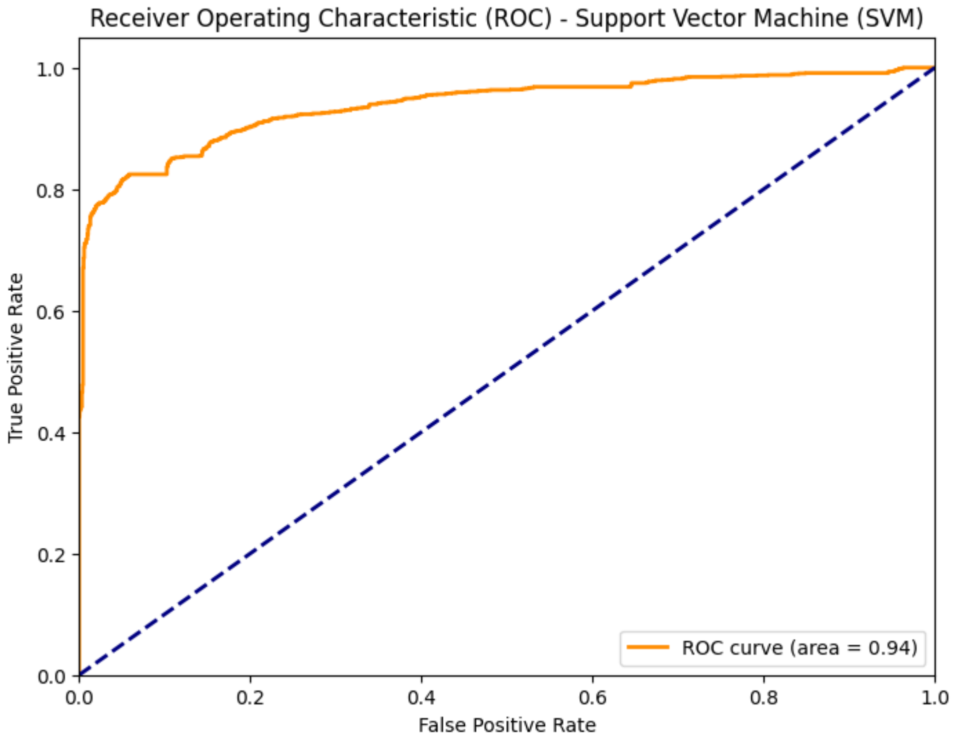

Confusion Matrix - K-Nearest Neighbors (KNN)
True label
Offline
Online
1529
2
484
1093
Offline
Online
Predicted label
1400
1200
1000
800
600
400
200

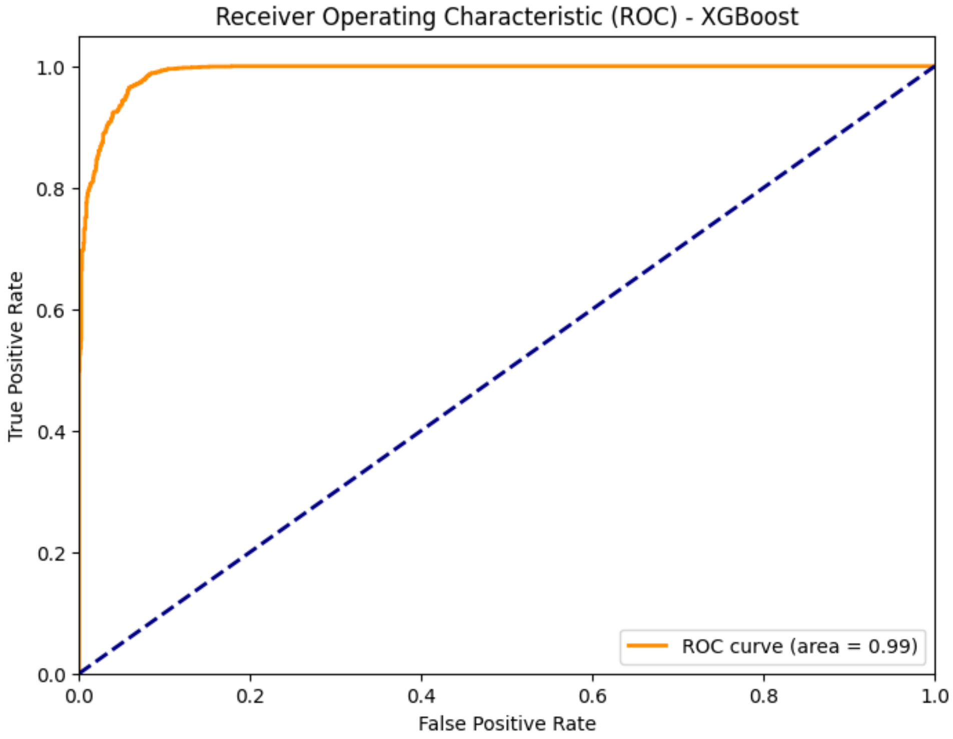

Receiver Operating Characteristic (ROC) - Decision Trees
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.93)


Receiver Operating Characteristic (ROC) - K-Nearest Neighbors (KNN)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.95)


Receiver Operating Characteristic (ROC) - Logistic Regression
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.95)


Receiver Operating Characteristic (ROC) - Random Forest
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.99)


Receiver Operating Characteristic (ROC) - Support Vector Machine (SVM)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.94)


Receiver Operating Characteristic (ROC) - XGBoost
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.99)

Chapter 6

Results, Discussion, Conclusion

6.1 Results

Random Forest achieved the highest accuracy at 95.05%, making it the best-performing model among those tested. XGBoost closely followed with an accuracy of 94.79%. Decision Trees also performed well with 93.05% accuracy. Logistic Regression and SVM had moderate performance with 87.97% and 87.13% accuracy respectively. KNN was the lowest performer with 84.36% accuracy.

6.2 Discussion

The results demonstrate that ensemble methods, particularly Random Forest and XGBoost, outperform other models in predicting customer purchase methods. Random Forest achieved the highest accuracy of **95.05%**, followed closely by XGBoost with **94.79%**. These models excel due to their ability to handle complex feature interactions and their robustness to overfitting. Their ensemble nature allows them to aggregate predictions from multiple decision trees, leading to higher accuracy and generalization.

In contrast, simpler models like Logistic Regression and Support Vector Machines (SVM) achieved moderate accuracy levels of **87.97%** and **87.13%** respectively. While these models are less accurate, they offer greater interpretability, making them suitable for scenarios where understanding the decision-making process is more important than raw predictive power. For example, Logistic Regression provides clear coefficients that indicate the influence of each feature on the target variable, which can be valuable for business stakeholders.

Decision Trees also performed well, achieving an accuracy of **93.05%**. However, they are prone to overfitting, especially when the tree depth is not properly controlled. This limitation makes them less reliable for generalizing to unseen data compared to ensemble methods like Random Forest and XGBoost.

The lowest-performing model was K-Nearest Neighbors (KNN), with an accuracy of **84.36%**. KNN's performance is hindered by its sensitivity to feature scaling and the curse of dimensionality. In datasets with many features, KNN struggles to identify meaningful patterns, as distances between data points become less informative. This result suggests that KNN may not be suitable for high-dimensional datasets like the one used in this study.

Overall, the findings highlight the importance of selecting the right model based on the specific requirements of the task. While ensemble methods like Random Forest and XGBoost offer superior accuracy, simpler models like Logistic Regression and SVM provide interpretability, which can be crucial for decision-making in business contexts. Future work could explore hybrid approaches that combine the strengths of both interpretable and high-performing models, as well as the integration of additional features such as customer reviews or social media activity to further improve prediction accuracy.

6.3 Conclusion

In this study, we demonstrated the effectiveness of machine learning in predicting customer purchase methods, with Random Forest achieving the highest accuracy of **95.05%**. XGBoost followed closely at **94.79%**. Simpler models like Logistic Regression and SVM provided interpretability with moderate accuracy. SMOTE was crucial for handling class imbalance, improving prediction reliability. These findings offer businesses actionable insights for tailored marketing, enhancing customer satisfaction and sales. Future research could explore additional data integration and hybrid model approaches to further refine predictions.

Chapter 7

Literature Review

7.1 Literature Review

7.1.1 Customer Purchase Prediction

Several researchers have analyzed factors affecting customer purchase decisions. For instance, [1] explored demographic and behavioral effects on online vs. offline shopping and identified age, income, and purchase history as key predictors.

7.1.2 Machine Learning in Customer Behavior Analysis

Machine learning has proven powerful for customer behavior analysis. For example, [2] used decision trees and logistic regression to predict customer churn. Ensemble methods like Random Forest and Gradient Boosting were shown to predict customer lifetime value effectively [3].

7.1.3 Handling Imbalanced Data

Customer purchase data often suffers from class imbalance (e.g., more online purchases than offline). [4] proposed SMOTE to improve model accuracy on minority classes, a key technique applied in this work.

7.1.4 Comparative Studies of Machine Learning Models

Studies comparing machine learning models often highlight the superior accuracy of ensemble methods. For example, [1] found Random Forest to outperform SVM and Logistic Regression in predicting customer preferences.

7.1.5 Gaps in Existing Research

Most research has focused on predicting customer churn or lifetime value, with limited studies addressing online vs. offline purchase prediction. This paper fills this gap by using SMOTE for class imbalance and comparing multiple models, as demonstrated in [5].

Bibliography

[1] E. Deniz and S. Çökekoglu Bulbül, "Predicting customer purchase behavior using machine learning models," *ADB: Information Technology in Economics and Business*, vol. 1, no. 1, pp. 1-6, 2024.

[2] Z. Lin, Z. Huang, H. He, J. Liang, X. Zheng, and S. Zhang, "Research on excessive medicine model based on machine learning," in *Advances in Artificial Intelligence, Big Data and Algorithms*, G. Grigoras and P. Lorenz, Eds. IOS Press, 2023.

[3] M. S. Azad, S. S. Khan, R. Hossain, R. Rahman, and S. Momen, "Predictive modeling of consumer purchase behavior on social media: Integrating theory of the model and machine learning for actionable insights," *PLOS ONE*, 2023.

[4] O. Ikechukwu Stasky, E. Osyeweuchi, A. E. Ikekwe, V. C. Orumake, E. P. Okemba, and O. C. Chimerie, "Predicting customer behavior on an online retail system using association and clustering machine learning opportunities and the model of the model and *r* languages," *IRE Journals*, vol. 7, no. 6, pp. 164-175, 2023.

[5] S. Fashion, "Dataset collected from Sanas Fashion," 2023, unpublished dataset.