PREDICTING CONSUMER BEHAVIOR IN ONLINE SHOPPING USING CLICKSTREAM DATA AND MACHINE LEARNING ALGORITHMS

YOUSIF AL-TAYEB

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

2076985

COMMITTEE

dr. Görkem Saygili
dr. Federico Zamberlan

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

January 22nd, 2024

WORD COUNT

8793

ACKNOWLEDGMENTS

I would like to express my profound gratitude to my family, who have always been beside me during my academic journey, granting me their invaluable and unwavering support, which has been crucial in reaching the point where I am today. Additionally, my deepest gratitude and appreciation go to my thesis supervisor, Dr. Gorkem Saygili, whose expertise and insightful guidance steered me through the challenging but rewarding journey of my academic endeavor. I also extend my heartfelt thanks to Dr. Federico Zamberlan, for his time dedicated to evaluating this thesis.

PREDICTING CONSUMER BEHAVIOR IN ONLINE SHOPPING USING CLICKSTREAM DATA AND MACHINE LEARNING ALGORITHMS

YOUSIF AL-TAYEB

CONTENTS

Abstract

The proliferation of e-commerce has transformed consumer behavior, making the prediction of online shopping patterns crucial but challenging for businesses. This thesis addresses the problem by exploring the efficacy of machine learning algorithms in predicting consumer behavior within the e-commerce domain, using clickstream and session-based data. Traditional algorithms such as Naive Bayes, Support Vector Machines (SVM), Random Forest, and XGBoost, alongside the advanced deep learning architecture TabNet, are employed and evaluated for their predictive capabilities. The research aims to determine the extent to which these conventional algorithms can accurately predict online consumer behavior and whether TabNet offers any improvement over these methods.

Utilizing a dataset from the UCI Machine Learning Repository, this study conducts an analysis involving feature importance and model performance assessment. The models are rigorously tested and compared using various metrics like precision, recall, F1-score, their weighted averages, confusion matrix analysis, and the Area Under the Receiver Operating Characteristic (ROC) Curve. Random Forest emerges as a standout performer among traditional models by demonstrating strong predictive power with high F1-scores and AUC values. Although TabNet shows promise, it does not substantially surpass the performance of traditional models namely XGBoost and Random Forest, highlighting the continued relevance of ensemble methods in the realm of e-commerce analytics.

The analysis identifies 'Page Values' and 'Exit Rates' as critical determinants in consumer purchasing decisions, offering actionable insights for e-commerce platforms. This research contributes to the understanding of consumer behavior in digital marketplaces and underscores the effectiveness of machine learning methods in e-commerce analytics.

1 SOURCE/CODE/ETHICS/TECHNOLOGY STATEMENT

The data has been acquired from the UCI Machine Learning Repository, a publicly accessible website. It includes detailed clickstream and session-based information specific to online shopping consumers. All figures presented in this thesis were created by the author. Portions of the code were adapted from publicly available resources found at GeeksforGeeks[1]. The code for this thesis could be accessed via GitHub[2]. ChatGPT, a generative language model, was employed as a debugging tool to assist in correcting programming errors and also provided assistance in language refinement, including paraphrasing, spell checking, and grammar correction. Typeset, an AI-powered tool, used to search for academic papers and get summaries from them (available at Typeset[3]). Additionally, a free online citation generator (Scribbr[4]) was utilized to ensure correct APA citation format. No other typesetting tools or services were used in the preparation of this thesis.

2 INTRODUCTION

In the rapidly evolving landscape of e-commerce, the ability to understand and predict consumer behavior has become a cornerstone for business success. The importance of e-commerce in the global economy is immense. As of 2023, global e-commerce sales have soared to an astounding $5.7 trillion, marking a clear shift in consumer preferences towards online shopping platforms (Statista, 2023). One of the factors in amplifying this shift was the COVID-19 pandemic, which has drastically changed shopping behaviors, pushing an even greater number of consumers to embrace online shopping (Organisation for Economic Co-operation and Development (OECD), 2021). With the advent of digitalization, not only have shopping habits been transformed, but vast amounts of data have also been generated. These data offer unprecedented opportunities to delve into consumer psychology and purchasing patterns, providing valuable insights for businesses.

Understanding online shopping behavior is crucial for enhancing consumers' experience and boosting sales, especially in the swiftly changing e-commerce sector. With advancements in technology, the ability to record and analyze session logs and behavioral traces of consumer groups on shopping websites has become increasingly feasible (Blasco-Arcas et al.,

[1] https://www.geeksforgeeks.org/machine-learning-with-python/
[2] https://github.com/YousifAl-Tayeb/Master-Thesis.git
[3] https://typeset.io/
[4] https://www.scribbr.com/citation/generator/

2022; Gao et al., 2022; Kukar-Kinney et al., 2022). Central to this analysis is the concept of a clickstream, which is a chronological series of web pages viewed during a user's session, providing a detailed outline of their browsing path. This clickstream data has proven effective in web usage analysis and in generating real-time predictions of online shopping patterns (Bucklin & Sismeiro, 2009). As a valuable resource, clickstream data offers deep insights into consumers' shopping preferences and how they interact with online platforms, including factors that capture their attention and influence their purchasing decisions. By leveraging clickstream data, consumers' purchasing behavior can be more effectively analyzed, offering a novel perspective and enriching the understanding of their decision-making processes (Z. J. Wen et al., 2023).

Although clickstream data may not provide every specific detail desired by researchers and practitioners, they offer considerably more information than the scanner panel data pivotal in developing and evaluating choice models in the early 1980s (Bucklin & Sismeiro, 2003). With their increased granularity, clickstream data bring a corresponding rise in complexity. Unlike scanner data, which primarily captured purchase decisions, clickstream data also traces the user's navigational journey leading to a purchase. This additional layer of information allows for a more comprehensive examination of search and purchase behaviors. However, the richness of these data also results to much larger datasets, posing challenges for researchers in effectively structuring user activities into coherent units for analysis (Bucklin & Sismeiro, 2009). Despite these challenges, the detailed nature of clickstream data makes them ideal for applying machine learning techniques to predict consumer behavior more accurately (Z. J. Wen et al., 2023).

3 PROBLEM STATEMENT & RESEARCH GOAL

This research aims to predict consumer behavior in the realm of online shopping using clickstream data and machine learning techniques. It focuses on understanding the factors that influence whether a website visitor will complete a transaction (make a purchase) during their online session. This research leverages the wealth of data available through clickstream information, which captures users' interactions with websites, and harnesses the power of machine learning to develop predictive models.

3.1 *Motivation*

The motivation for this project arises from the expanding e-commerce landscape, where an increasing number of businesses establish online

presence. Understanding online consumer behavior has become critical for e-commerce success in the digital transformation era. The widespread development of big data technology has led to the diversification and complexity of e-commerce consumer behavior (Gersen & Steckel, 2023). E-commerce platforms have greatly streamlined the shopping process, allowing individuals to conveniently make purchases online after their work hours. This advancement has notably diminished the time cost associated with consumer shopping activities (Lombart et al., 2020; Wu et al., 2019). Furthermore, in an era marked by digital transformation, businesses rely on actionable insights into online consumer behavior to remain competitive and responsive to evolving market trends.

3.2 *Societal & Scientific Relevance*

This research project holds substantial scientific and societal relevance in the ever-advancing field of e-commerce. Scientifically, it pushes the boundaries of data analytics by employing advanced machine learning techniques, specifically exploring the potential of TabNet, a deep learning architecture, against other conventional machine learning algorithms, to analyze clickstream data for predicting online consumer behavior. This approach not only contributes to the domain-specific research question of understanding digital consumer patterns but also proposes an innovative methodological framework that can be applied in broader data science contexts.

Societally, the project is poised to considerably impact the e-commerce sector by optimizing online platforms and boosting revenue (Lessmann et al., 2019). By leveraging the rich insights derived from clickstream data, it supports data-driven decision-making, enabling businesses to better cater to consumer needs (Pal et al., 2018). This approach not only assists companies in their growth and expansion but also creates more job opportunities, thereby contributing to the reduction of unemployment. As businesses grow, they often make improvements that benefit their employees, typically leading to increased incomes and enhanced living standards. Over time, these changes can demonstrate tangible improvements in societal conditions.

Furthermore, by bridging the gap between data science and consumer psychology, this research provides valuable insights into digital user behavior, thereby advancing e-commerce analytics and offering practical solutions that benefit both businesses and consumers. In essence, this thesis not only contributes to academic knowledge but also addresses real-world challenges, making it an important intersection of scientific innovation and societal advancement.

3.3 *Research Questions*

In the swiftly changing e-commerce landscape, characterized by intricate consumer behaviors, the role of machine learning in deciphering and predicting these actions becomes increasingly crucial for businesses seeking to adapt to this dynamic environment. Traditional algorithms such as Naive Bayes, XGBoost, SVM, and Random Forest, already proven in various domains, are being tested to see their effectiveness in predicting consumer behavior. For example, by accurately predicting which products a consumer is inclined to purchase, online shops can more effectively customize their product offerings, leading to more pertinent recommendations. This approach not only elevates the shopping experience but also heightens customer satisfaction. The necessity of evaluating the performance of these algorithms leads to formulating the first research question:

RQ1: *To what extent can conventional machine learning algorithms, including Naive Bayes, XGBoost, SVM, and Random Forest, predict online consumer behavior in the context of e-commerce?*

While traditional machine learning models have been the backbone of many predictive analytics applications, the advent of deep learning architectures like TabNet offers new horizons in data analysis. TabNet's ability to handle tabular data and its interpretative capabilities make it a potential tool in e-commerce analytics. However, its practical superiority over traditional models in predicting consumer behavior remains an open question. This gives rise to the formulation of the second research question:

RQ2: *To what extent can TabNet improve prediction accuracy compared to the best-performing traditional machine learning algorithm when predicting online consumer behavior in e-commerce?*

Understanding which features most considerably influence consumer behavior is paramount. This understanding not only enhances the predictive accuracy but also provides actionable insights for businesses. This necessity to identify and analyze the most impactful factors in consumer decisions motivates for sub-research question:

SQ: *What are the most influential features in predicting consumer behavior?*

4 RELATED WORK

Clickstream data, a vital resource in e-commerce, has the potential to improve user experiences and achieve business goals. Predicting online

shopping behavior and targeting real-time marketing interventions has revolutionized retail, cutting costs and boosting revenue. Various machine learning models have been built to handle clickstream data. Sufficient research papers have explored these models, offering valuable insights into e-commerce predictive analytics. Among these are the followings:

4.1 *Machine Learning in Consumer Behavior Prediction*

Koehn, Lessmann, and Schaal (2020) 's paper in Expert Systems with Applications centers on the application of deep learning to predict online shopping behavior using clickstream data. By introducing recurrent neural networks (RNNs), the authors tackle the challenge posed by the sequential structure of this data, a limitation often overlooked by traditional Supervised Machine Learning (SML) approaches. Their empirical evaluation demonstrates the superiority of RNN-based clickstream modeling in comparison to SML benchmarks. Furthermore, they highlight the feasibility of combining RNN-based and conventional classifiers within an ensemble, a strategy that consistently outperforms alternative models.

Other recent research papers, in the realm of e-commerce, have explored innovative avenues for predictive analytics. Wen et al. (2023) focused on predicting anonymous consumer purchase intent using their MBT-POP machine learning model, which leverages multi-behavioral trendiness (MBT) and product popularity (POP) from clickstream data, improving both accuracy and prediction speed. Sakar et al. (2019) introduced a real-time predictive system for online shoppers' purchasing intentions, employing a two-module approach. The first module integrates session-based features with clickstream data using classifiers like random forest, SVMs, and multilayer perceptron (MLP). The second module utilizes sequential clickstream data and a long short-term memory-based recurrent neural network to predict the likelihood of visitors leaving without completing a transaction.

Zhang and Wang (2021)'s study addresses e-commerce consumers' repurchase behavior, introducing an enhanced deep forest model that incorporates interactive behavior characteristics to enhance accuracy and reduce training time. Abdullah-All-Tanvir et al. (2023) presented a model designed to predict early purchase intentions on e-commerce websites, employing advanced feature selection and oversampling techniques. The model trains various supervised learning classifiers, including SVM, Random Forest, MLP, Decision Tree (DT), and XGBoost. Notably, the XGBoost classifier, enhanced with feature selection and oversampling, demonstrates superior performance.

There are also papers that focus mainly on specific types of machine learning algorithms, below are some of these papers:

4.1.1 *SVM Approach*

Maheswari and Priya (2017) utilized SVM to classify customers based on their purchasing patterns, emphasizing its role in analyzing customer behavior for business performance assessment and trend prediction. The study employed various SVM classifiers, including polynomial and radial basis functions, with a sigmoid kernel function.

Building on this, Tang et al. (2017) introduced a hybrid model combining SVM classification with the Firefly Algorithm (FA) for enhanced predictive accuracy in online purchasing behaviors. This model incorporated diverse factors such as online shopping cart usage, clickstream data, and previous purchase behaviors. Focused on an online furniture store, the FA-enhanced SVM model showed superior performance over traditional benchmarks.

X. Liu and Li (2016) further demonstrated SVM's application in analyzing behavioral data from Chinese E-commerce platforms, aiming to improve product recommendation accuracy and conversion rates. The study involved feature extraction and model training using Libsvm, an SVM-based software package. Lastly, Renuka (2023) acknowledged SVM's utility in customer behavior analysis, striving to create a precise predictive model by integrating the latest machine learning advancements and building on existing research. These studies collectively underscore SVM's notable role in e-commerce consumer behavior analysis and prediction.

4.1.2 *XGBoost & Gradient Boosting Approach*

Recent studies have emphasized the effectiveness of Extreme Gradient Boosting (XGBoost) and Gradient Boosting algorithms. Wang et al. (2023) introduced an XGBoost-based model for predicting user purchase behavior, excelling in accuracy, F1 score, and ROC value over traditional methods such as K-Nearest Neighbors (KNN), SVM, RF, and Back Propagation Neural Network (BPNN). This model, leveraging unique user value and tag features, demonstrates superior performance in predictive accuracy. In a similar vein, Gumber et al. (2021) proposed using XGBoost as an ensemble method for predicting customer behavior based on clickstream data. This method excels in achieving high accuracy and recall rates while effectively managing overfitting.

Complementing this, Renuka (2023) explored the Gradient Boosting algorithm, highlighting its capability in processing complex relationships and non-linear patterns, essential for predicting customer purchase likelihood.

This study focused on comprehensive feature engineering, model training, and evaluation, underlining the role of consumer behavior understanding in enhancing targeted marketing. Similarly, Cai and Rodavia (2023) utilized XGBoost to analyze consumer behavior, aiming to identify key predictive features for purchase intentions. This research contributed to personalized marketing strategies and tackled unbalanced data challenges, providing insights into consumer behavior patterns and product popularity trends over time.

4.1.3 *Random Forest Approach*

Prayogo and Karimah (2021) introduced a novel approach combining feature selection with Adaptive Synthetic Sampling (ADASYN) to enhance the prediction of online shopping intent. Utilizing Information Gain and Correlation for feature selection, the study effectively addressed class imbalance issues, demonstrating the Random Forest classifier's superior accuracy, precision, recall, and F1-score. This method's effectiveness was further validated through comprehensive evaluation techniques including confusion matrix and Mann-Whitney U test.

Ghosh and Banerjee (2020) developed a modified Random Forest model to predict customer purchase behavior in cloud services. The model incorporated variables such as advertisement click sequences and customer past behaviors, showing high accuracy in predicting future purchases, thereby underscoring the utility of Random Forest in customer behavior analysis in the cloud services sector.

Sang and Wu (2022) focused on real-time prediction of online shopper purchasing intent using Random Forest combined with oversampling techniques. This study achieved notable accuracy (86.78%) and an F1 Score of 0.6, proving the model's effectiveness in predicting online shopping behaviors right from the onset of a website visit. These studies collectively highlight the versatility and effectiveness of Random Forest algorithms in predicting customer purchase behavior across various e-commerce contexts.

4.1.4 *TabNet Approach*

In their method, Arık and Pfister (2021) introduced 'Interpretable Tabular Data Learning Using Sequential Sparse Attention', a method that employs TabNet, a deep tabular data learning architecture, on data processing hardware. This approach involves initially receiving a feature set and, through multiple sequential processing steps, utilizes a sparse mask within TabNet to select relevant features. These selected features are then processed by a TabNet feature transformer, generating decision-step outputs and information for subsequent processing steps in the sequence. Ultimately, the

method produces a final decision output by combining the decision step outputs from each processing step.

Houfani et al. (2022) applied TabNet in the medical field to predict Intensive Care Unit (ICU) admission needs for COVID-19 patients. This study highlighted TabNet's superior prediction accuracy over models like MLP, RF, LR (Logistic Regression), and KNN, particularly when integrated with the Synthetic Minority Oversampling Technique (SMOTE), showcasing its potential in optimizing healthcare resource allocation.

Joseph et al. (2022) demonstrated TabNet's versatility through the development of an interpretable model for early diabetes detection. By leveraging Bayesian optimization and TabNet's attention mechanism, the study achieved high accuracy in classifying various diabetes datasets, emphasizing the importance of model interpretability in enhancing trust in AI applications in healthcare.

Finally, Z. Liu (2023) extended the application of TabNet to geoscience, specifically for predicting porosity in subsurface fluid flow and reservoir evaluation. Comparing TabNet's performance with traditional machine learning and Long Short-Term Memory (LSTM) methods, the study found TabNet more effective, as indicated by its lower root mean square error (RMSE), thus offering an improved approach in reservoir evaluation.

4.2 *Class Imbalance*

Class imbalance is a prevalent challenge in machine learning, where certain classes are underrepresented in datasets, leading to biased models and inaccurate predictions (He & Garcia, 2009). This section delves into some of the strategies and methodologies developed to counteract this imbalance.

Because of this class imbalance, the necessity for specialized techniques arises to balance the training set. Resampling methods, such as oversampling the minority class and undersampling the majority class, are commonly employed to address this issue. Notably, the Synthetic Minority Over-sampling Technique (SMOTE) has been instrumental in generating synthetic data for the minority class, enhancing model training and performance (Chawla et al., 2002). Additionally, algorithmic adjustments, like cost-sensitive learning, have been explored to make algorithms more attentive to the minority class (Elkan, 2001).

Comparative studies reveal that the choice of technique largely depends on the specific context, such as the type of data and the extent of imbalance. Batista et al. (2004) and López et al. (2013) provided insights into the effectiveness of various techniques, suggesting that no one-size-fits-all solution exists. Practical applications in fields further underscore the importance of addressing class imbalance (Dal Pozzolo et al., 2015). Successful studies,

such as those by Flores et al. (2018) and Charte et al. (2015), demonstrate how effectively tackling class imbalance can improve model performance.

4.3 *Research Gap*

The field of machine learning and data analysis has made great advances in developing algorithms and methodologies for diverse challenges. While sufficient research has been conducted on consumer behavior prediction using algorithms like XGBoost, SVM, and RF, and studies such as Anh et al. (2023) have employed deep learning architecture including TabNet for analyzing clickstream data, the specific application of TabNet to clickstream and session-based data from online shopping platforms is comparatively under-explored. To the best of my knowledge, this observation holds true particularly at the time of initiating this study.

The complexity of clickstream data, crucial for decoding online consumer behavior, poses challenges deserving further exploration with advanced machine learning algorithms, specifically TabNet. Moreover, an important area of study is evaluating TabNet's performance with imbalanced datasets within e-commerce analytics, a prevalent issue, particularly in the context of techniques like SMOTE oversampling. Investigating the optimization of TabNet for analyzing consumer clickstream data could not only deepen our understanding of consumer behaviors and decision-making online but also provide insights into the efficacy of a deep learning method in this domain.

5 METHODOLOGY

This section delineates the methodology implemented in this research, which includes evaluating the effectiveness of conventional machine learning algorithms in predicting consumer behavior and comparing the best performer among them with the TabNet architecture. The approach is designed to address the specific research questions, systematically guiding the reader through the various stages of the study, from data preprocessing to model evaluation. This section also sheds light on the challenges encountered during the research and the strategies adopted to surmount them. By elaborating on these steps, the methodology provides a detailed roadmap to get the answers to the research questions.

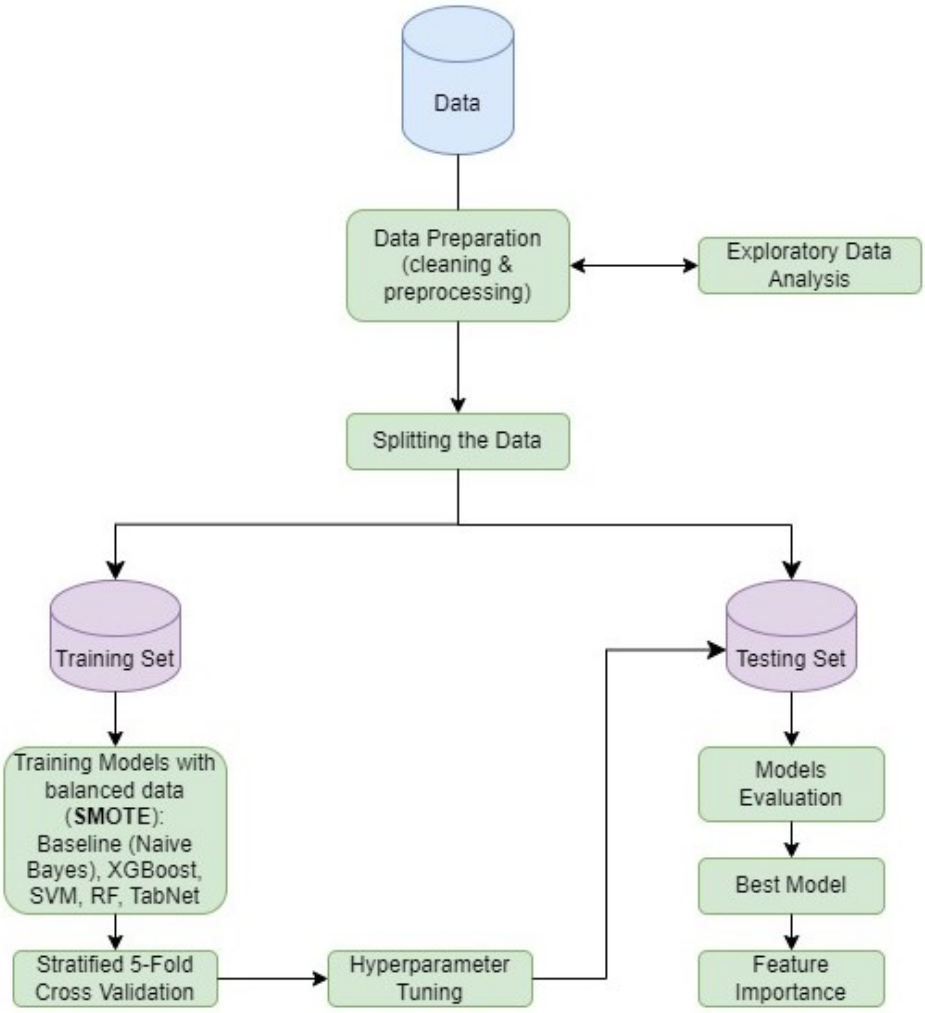

Figure 1: Methodology Flowchart

5.1 *Dataset Description*

The dataset, known as Online Shoppers Purchasing Intention (UCI Machine Learning Repository, 2018), comprises feature vectors from 12,330 distinct sessions[5]. It includes clickstream and session-based data and it was thoughtfully designed to ensure that each session corresponds to a different user within a one-year period. This dataset encompasses a total of 10 numerical attributes, as shown in Table 11 in Appendix A, and 8 categorical attributes shown in Table 12, with the 'Revenue' attribute serving

[5] A session refers to a user's single visit to a website, encompassing a sequence of interactions within a specific timeframe.

as the designated class label.

The dataset features track user engagement and site interactions. "Administrative", "Informational", and "Product Related" features, along with their respective durations, record the number and time spent on different page types, updated based on user navigation. Google Analytics metrics - "Bounce Rate", "Exit Rate", and "Page Value" - measure visitor engagement and page profitability. "Special Day" indicates the timing of visits relative to events like Mother's Day, affecting purchase likelihood. This value varies, peaking around specific dates (e.g., May 1-11 for Mother's Day). Additional data includes operating system, browser, location, traffic source, visitor type, weekend visit indicator, and visit month.

5.2 *Exploratory Data Analysis*

A fundamental step in machine learning and data science is Exploratory Data Analysis (EDA). It serves as a critical bridge between raw data and analytical modeling. As (Pearson, 2018; Tukey et al., 1977) highlighted, EDA is essential for understanding data characteristics, identifying patterns, and formulating hypotheses. This section presents the EDA conducted on the clickstream dataset, focusing on key methods and visualizations that pave the way for effective machine-learning applications.

Class label distribution: As illustrated in Figure 2, the target variable 'Revenue' in our dataset exhibits a notable disparity; a substantial majority of consumers, amounting to 10,422 or approximately 84.53% of the dataset, did not complete their online transactions. This indicates a pronounced imbalance, with a majority of consumers belonging to one class over the other, leading to an imbalanced dataset. Addressing this major imbalance is crucial to prevent inaccurate or biased predictions. The method employed to handle this issue is detailed in the preprocessing section.

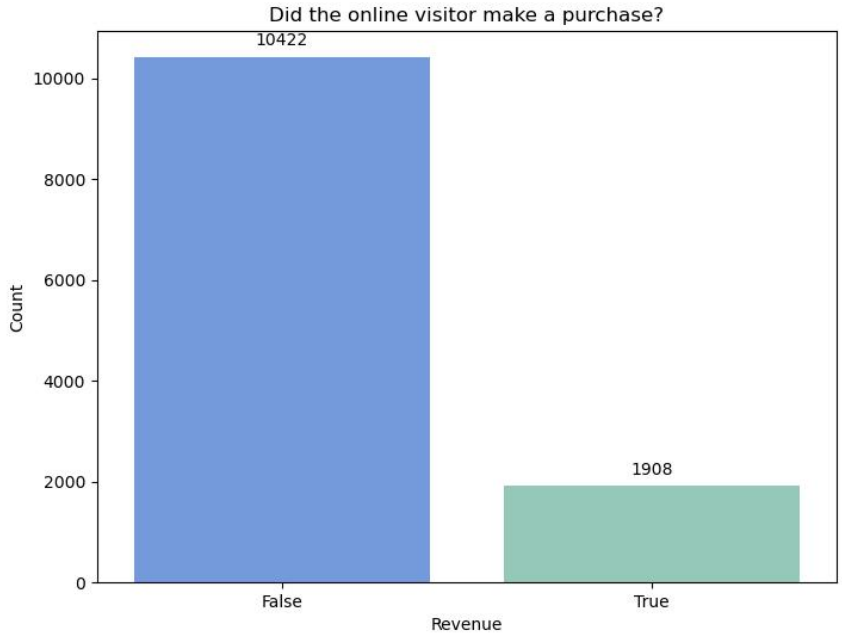

Figure 2: Distribution of Revenue

Correlations: The Pearson Correlation matrix, as depicted in Figure 3, encompasses all attributes within the dataset. Notably, it reveals a moderate positive correlation (0.49) between 'Page Value' and the target 'Revenue'. This suggests that pages with higher perceived value in e-commerce transactions may influence revenue generation. Conversely, the remaining features exhibit a lower correlation with the target. The analysis also highlights a notable correlation among administrative data. Furthermore, 'Information', 'Product Related', 'Bounce Rate', and 'Exit Rate' features display similar characteristics, as indicated by their correlation values.

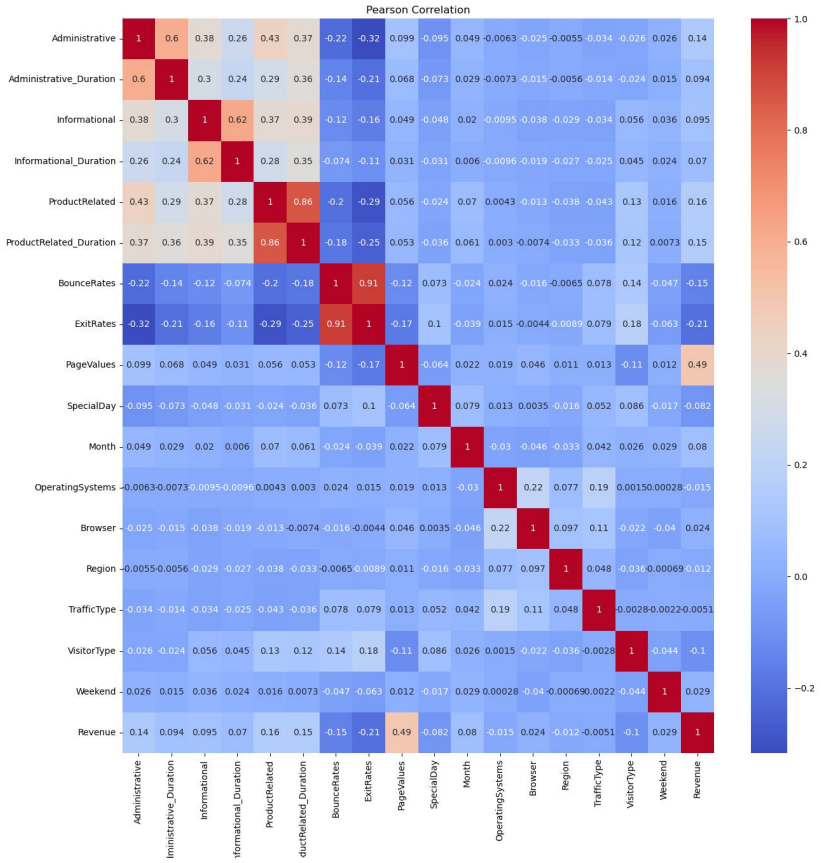

Figure 3: Pearson Correlation Matrix (Heatmap)

Outliers: Figure 4 below presents a grid of scatter plots for the numerical features in the dataset. It provides a clear view of the data and can be a valuable tool in preliminary data analysis. It allows to visually assess how values of a particular feature are distributed across the dataset. This can help identify patterns, trends, or anomalies such as outliers. The visualization shows that there are indeed outliers in some features. Dealing with this is mentioned in the data preparation section.

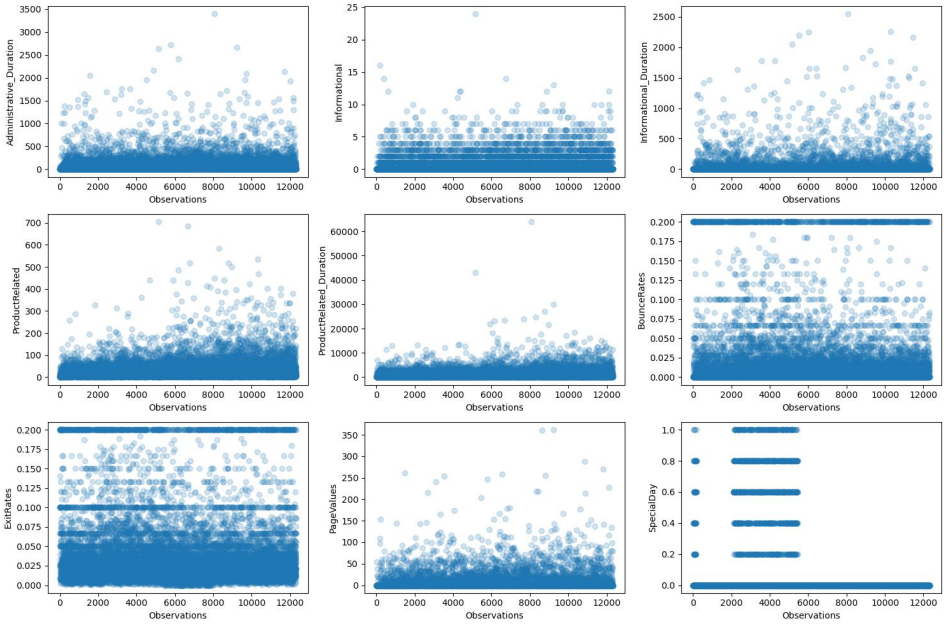

Figure 4: Scatter Plots Grid (Distribution of Observations)

(Weekdays & Weekends) and Revenue: Typically, weekends offer people more leisure time compared to weekdays. This additional time could be utilized for various activities, including browsing and shopping online. However, an intriguing pattern emerges from the data. As illustrated in Figure 5, the majority of online visits that resulted in a purchase surprisingly occurred during weekdays. This finding challenges the common assumption that weekends are the prime time for online shopping activities.

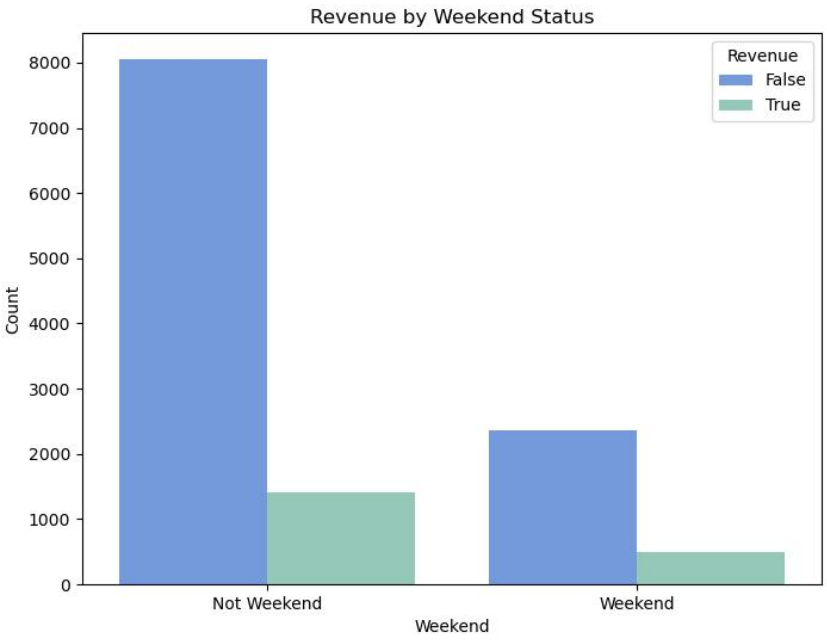

Figure 5: Comparison of Online Shopping Revenue Between Weekdays and Weekends

5.3 *Data Preparation*

According to Kotsiantis et al. (2006), effective data preparation is vital for improving data quality and interpretability. This step is also important in machine learning that influences the performance of predictive models. This section delves into the specific data preparation strategies employed in this study. These steps are crucial for ensuring that the dataset is optimally structured and ready for the subsequent stages of model development and analysis.

5.3.1 *Data Cleaning*

The data cleaning process involves handling missing values and outliers. In this dataset, there are no missing values. However, addressing outliers is explained in the subsequent section on Outlier Treatment.

Outliers Treatment: During the Exploratory Data Analysis in the earlier section, the data distribution was visualized (refer to Figure 4). This section focuses on the treatment of outliers present in the dataset. Outliers with extreme values were removed, as depicted in Figure 6. However, outliers without extreme values, demonstrating discernible patterns, were retained

(as seen in the same figure). It's essential to note that outliers can contain important information (Smiti, 2020), hence only the extreme outliers were eliminated. This decision was made to maintain data variability and retain potentially informative data points.

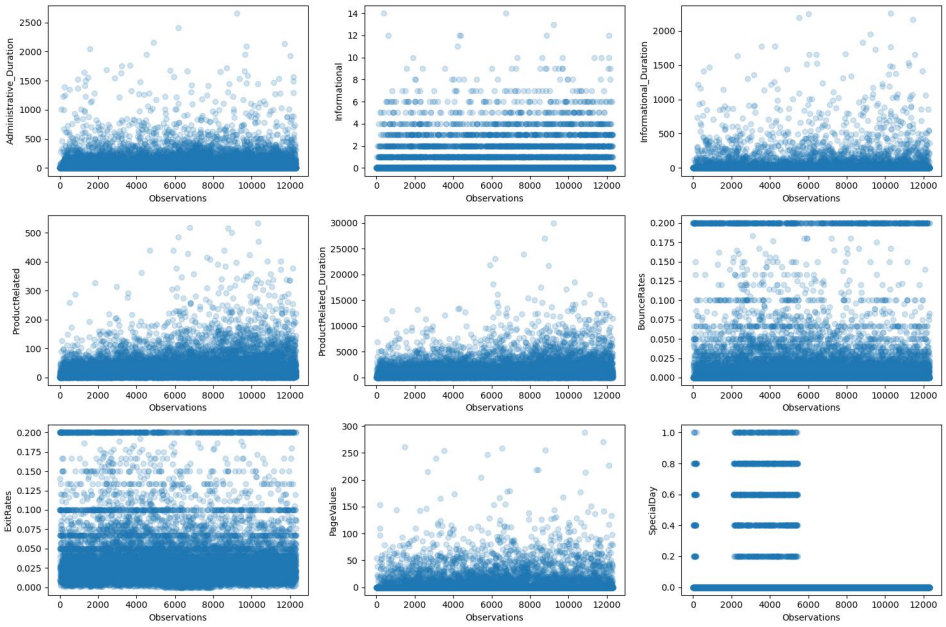

Figure 6: Outliers Removed

5.3.2 *Preprocessing*

Data preprocessing plays an important role in refining raw data, ensuring its compatibility for modeling, and enhancing the overall predictive performance of machine learning algorithms. Each subsection below details a specific preprocessing method adopted to address different challenges in data preparation.

A) One-Hot Encoding: The dataset comprises both numerical and categorical features. While numerical features can be directly used in many machine learning algorithms, categorical variables require a different approach. To address this, one-hot encoding is employed to convert these variables into a format suitable for machine learning models (Seger, 2018), thereby facilitating more accurate predictions. This technique involves transforming each category within a categorical variable into a distinct binary feature. Such a transformation is critical in ensuring that the model interprets these variables correctly. One-hot encoding is especially beneficial for nominal categories, which lack an inherent order, as it maintains the uniqueness of each category without implying any artificial hierarchy.

Consequently, the 'Month' and 'Visitor Type' features in the dataset were processed using one-hot encoding.

B) Label Encoding: Label encoding is an effective and straightforward approach for handling binary categorical features, such as 'Revenue' and 'Weekend' in the dataset. This method involves converting the two categories of each feature into numeric codes, typically 0 and 1. Such a conversion is not only efficient but also space-saving, as it avoids the unnecessary expansion of the feature space that would result from other encoding techniques.

C) SMOTE Oversampling: To address the class imbalance present in the dataset, methods such as oversampling and undersampling can be considered. While undersampling involves reducing the number of instances in the majority class, it may lead to a loss of valuable information. Conversely, oversampling techniques, particularly the Synthetic Minority Over-sampling Technique (SMOTE), are used to increase the number of instances in the minority class, thereby avoiding this loss of information (Akbani et al., 2004). Therefore, SMOTE oversampling has been utilized in this dataset to address the imbalance.

SMOTE oversampling approach was applied exclusively to the training set to maintain the validity and reliability of the model evaluation process, ensuring that the test set remains a true representation of real-world data and preventing data leakage. This approach generates synthetic samples from the minority class, effectively balancing the dataset. By doing so, it enhances the classifier's ability to detect patterns in the minority class without being biased towards the majority class. Such a step is crucial in avoiding model bias and ensuring a more robust and accurate representation of all classes in the data. Figure 10 illustrates the distribution of the minority class before and after the application of SMOTE oversampling.

D) Feature Scaling: Feature scaling, an essential preprocessing step, was implemented to normalize the dataset's features. In the dataset, the employed scaling technique is z-score standardization. This standardization helps in improving model accuracy and convergence speed (Han et al., 2012).

E) Stratification: In the presence of class imbalance within the dataset, stratification becomes an important step in the train-test split process. It guarantees that both training and test sets reflect the dataset's overall class distribution, preserving the class ratio and addressing imbalances. This step is important for avoiding model bias and improving its generalizability, thereby enhancing the robustness and accuracy of model evaluation.

5.4 *Models*

In this section, a range of machine learning models is explored to predict consumer behavior in the e-commerce domain. The focus is on Naive Bayes, XGBoost, SVM, Random Forest, and TabNet, each of which will be tested for their effectiveness in handling clickstream data.

5.4.1 *Naive Bayes*

The Naive Bayes classifier is a simple yet effective model for predictive modeling, particularly suitable for data with low-entropy distributions and specific feature dependencies (Rish et al., 2001). It offers several variants for different data types: Gaussian Naive Bayes, ideal for continuous data assuming normal distribution within each class; Multinomial Naive Bayes, more suitable for scenarios where features are represented as frequency vectors; and Bernoulli Naive Bayes, which is used in this study, is best for binary or Boolean data. As noted by Webb et al. (2010), its versatility in handling both categorical and numeric attributes, along with its stable performance across various data sizes under the principle of conditional independence, makes it particularly useful in text mining.

5.4.2 *SVM*

The support-vector network is a machine-learning method for two-group classification, transforming input vectors into a high-dimensional space to create a linear decision surface with notable generalization capabilities. It integrates optimal hyperplane techniques, dot product convolution, and soft margins, enhancing its adaptability from linear to nonlinear solutions and handling training set errors. Remarkably, its unique soft margin classifier solution ensures high generalization ability, even in infinite-dimensional spaces, underscoring its effectiveness in complex classification scenarios (Cortes & Vapnik, 1995).

5.4.3 *XGBoost*

XGBoost is a decision-tree-based ensemble algorithm that uses a gradient-boosting framework. It has gained popularity due to its speed and performance. Key to its efficacy is the introduction of a sparsity-aware algorithm tailored for sparse data and a weighted quantile sketch that enables efficient approximate tree learning. XGBoost further enhances its capability with advanced techniques in cache access, data compression, and sharding, allowing it to efficiently process billions of examples, thus positioning it as a highly effective tool for large-scale machine learning tasks. It excels in managing large and complex datasets, demonstrating notable robustness to

overfitting and optimizing the gradient boosting process (Chen & Guestrin, 2016).

5.4.4 *Random Forest*

Random Forest is a robust ensemble learning method that constructs multiple decision trees during training and determines the output class based on the mode of the classes predicted by individual trees. Renowned for its high accuracy and parallel processing capabilities, Random Forest is particularly resilient to overfitting, benefiting from the Law of Large Numbers and strategic randomness. This method's predictive strength and correlation, evaluated through out-of-bag estimation, competes well with the accuracy of arcing (Adaptive Resampling and Combining) algorithms. Unlike these algorithms, Random Forest maintains a consistent approach throughout, adeptly balancing bias and variance reduction. Additionally, its inherent feature selection ability helps identify the most key variables within large datasets, further enhancing its predictive power (Breiman, 2001).

5.4.5 *TabNet*

TabNet, a relatively recent architecture, utilizes neural networks specifically designed for tabular data. It combines deep learning with the decision-making logic of tree-based models. A unique feature of TabNet is its ability to perform feature selection. It uses sequential attention to choose which features to reason from at each decision step, thus enabling interpretable decision-making, a desirable quality in predictive analytics. The paper by Arık and Pfister (2021) notes that higher dimensional embedding can boost performance but may complicate interpretation. The study suggests balancing performance and complexity by adjusting TabNet hyperparameters N_d and N_a[6], cautioning that very high values might lead to overfitting and poor generalization.

5.4.6 *Hyperparameter Tuning*

This section explores the hyperparameter tuning process, which is critical for optimizing machine learning model performance. Unlike model parameters[7], hyperparameters are externally preset and notably influence the model's complexity, training speed, and ability to address overfitting or

[6] N_d and N_a are hyperparameters in the TabNet architecture that control the number of decision steps (N_d) and the number of attention heads (N_a).

[7] Model parameters in the context of machine learning are values that the model learns during the training process from the given data.

underfitting. Each model requires a unique set of hyperparameters, tuned for optimal accuracy, efficiency, and overall model robustness.

To effectively tune hyperparameters, defining a search space is essential. This search space outlines the possible range of values for each hyperparameter, which can be either discrete or continuous. Traditional methods like grid search, an approach that, while thorough, is often time-consuming. Alternatively, RandomSearch selects hyperparameters randomly, providing a method that is less systematic compared to grid search. This approach may not always reliably yield the most optimal hyperparameters. In response to these limitations, this study uses Optuna, which employs Bayesian optimization for more efficient hyperparameter exploration. Optuna strategically selects hyperparameters for testing, based on past trial performance, reducing trial numbers while potentially finding superior settings. This method enhances the search process, particularly in complex hyperparameter landscapes, by dynamically adjusting its strategy for targeted exploration.

The optimization process employs a 5-fold stratified cross-validation approach, using the StratifiedKFold method from scikit-learn. This technique, as noted by Kohavi et al. (1995), provides a more robust evaluation method, particularly beneficial for imbalanced datasets. The utilization of stratified cross-validation helps to minimize the risks of overfitting and enhances the generalizability and robustness of the model. The data is shuffled to ensure randomness in the selection process, and a consistent random state is set for reproducibility. The optimal hyperparameter configurations identified through this process are then assessed using the test dataset. To balance thoroughness in exploring the hyperparameter space with computational efficiency, the number of trials in Optuna was set to 40, taking into consideration constraints of time and computational resources.

A) Tuning Naive Bayes: The Naive Bayes classifier offers different variants for various data types, as outlined in Section 5.4.1. Gaussian Naive Bayes usually requires no hyperparameter tuning, reflecting the model's simplicity. The Multinomial and Bernoulli Naive Bayes variants, however, introduce an alpha hyperparameter for smoothing, allowing for some tuning. Despite their differences, these variants maintain the model's simplicity and adaptability. In this study, the alpha hyperparameter was kept at its default value of 1.0 after trials showed no notable improvements with other values.

B) Tuning SVM: For SVM, hyperparameter tuning focuses on hyperparameters like the ***kernel*** type, regularization hyperparameter ***C***, and kernel coefficient ***gamma***. More explanation of each hyperparameter is available in Appendix C 9.

Table 1: SVM Hyperparameters and their Values

Hyperparameter	Values
C	0.01 to 20
gamma	scale, auto
degree	2 to 5
coefo	0.0 to 10.0
class_weight	None, balanced
kernel	linear, rbf, poly, sigmoid

C) Tuning XGBoost: XGBoost also offers a range of hyperparameters, as shown in Table 2. For more detailed explanations of each hyperparameter, refer to Appendix C 9.

Table 2: XGBoost Hyperparameters and their Values Space

Hyperparameter	Values
n_estimators	100 to 300 (step 20)
max_depth	2 to 30
learning_rate	0.01, 0.1, 0.2
gamma	0 to 1 (step 0.01)
colsample_bytree	0.5 to 1
subsample	0.5 to 1
min_child_weight	1 to 3

D) Tuning Random Forest: In Random Forest, several key hyperparameters play an important role in model performance. These hyperparameters are shown in Table 3. Further explanation of each hyperparameter is available in Appendix C 9.

Table 3: Random Forest Hyperparameters and their Values Space

Hyperparameter	Values
n_estimators	100 to 300 (step 20)
max_depth	2 to 30
min_samples_split	2 to 10 (step 2)
min_samples_leaf	2 to 10 (step 2)
bootstrap	True, False

E) Tuning TabNet: TabNet was trained using a set of hyperparameters, as listed in Table 4. For further explanation of each hyperparameter, see Appendix C 9.

Table 4: Hyperparameters for the TabNet Architecture

Hyperparameter	Values Range
mask_type	entmax, sparsemax
max_epochs	50
patience	10
batch_size	64, 128, 256
lr	starting from 0.001 (logarithmically spaced)
n_d	8 to 64
n_a	8 to 64
virtual_batch_size	32, 64, 128

Following this detailed account of hyperparameter tuning across different models, it is important to note the distinction in their optimization processes. While Naive Bayes, Random Forest, and SVM do not use gradient descent-based optimizers, they do involve optimization in their training process, albeit in a different form than models like XGBoost and TabNet. XGBoost and TabNet use more explicit and iterative optimization processes, with TabNet utilizing a traditional optimizer like Adam.

5.5 *Evaluation Methods*

Due to the substantial class imbalance in the dataset, accuracy alone is not a reliable measure for model evaluation. Consequently, this section introduces a range of metrics more appropriate for this context. These evaluation metrics, crucial for assessing the models' performance in various aspects, include precision, recall, F1 score, their weighted averages, ROC (Receiver Operating Characteristic), and confusion matrix analysis. Together, they offer a comprehensive view of model performance. Detailed explanations of each metric are available in Appendix C 9, offering further insights into their relevance.

6 EXPERIMENTAL SETUP

This section details the settings, steps, and tools used in the study to answer the research questions. It outlines the software and packages utilized, the data sources and their features, the data preparation process, and the methods used for train-test splitting. This setup forms the foundation for the experimental work, ensuring reproducibility and reliability of the results.

6.1 *Data Source & Features*

The data source and a detailed description of its features, along with all related information, are mentioned in subsection 5.1.

6.2 *Data Preparation*

Data preparation involves several stages to ensure its suitability for running the experiment. Such meticulous preparation of the data is critical for ensuring the accuracy and reliability of subsequent predictions and analyses. The comprehensive details of each stage in the data preparation process are further elaborated in Section 5.3, providing an in-depth view of the methodologies employed.

6.3 *Software & Packages*

In this study, various software tools and packages were integral to setting up the experimental environment. Python, known for its versatility and extensive support in data analysis and machine learning, was the primary programming language (Van Rossum et al., 2007). Essential libraries imported included Pandas (McKinney et al., 2010) for managing dataframes and facilitating tasks like data import/export, and NumPy (Harris et al., 2020) for efficient matrix operations. Both were instrumental in the data preprocessing phase. Scikit-learn (Pedregosa et al., 2011) was used for implementing machine learning algorithms, while Matplotlib (Hunter, 2007) and Seaborn were key in data plotting and advanced visualization, respectively. The imbalanced-learn library supported the application of the SMOTE oversampling technique. XGBoost, a standalone library, was utilized for gradient boosting through its XGBClassifier. The Pytorch-tabnet library facilitated the implementation of the TabNet algorithm. Optuna was selected for hyperparameter optimization. These tools were chosen based on the need for accurate predictions, comprehensive feature analysis capabilities, their wide adoption in the research community, and their extensive feature sets, aligning with the project's requirements. Table 13 shows the version for each of the aforementioned software and libraries.

6.4 *Train-Test Splitting*

An essential aspect of the experimental setup was the splitting of the dataset into training and testing subsets. This process was critical for evaluating the performance of our machine learning models. In this study,

train_test_split function was employed from scikit-learn to divide the dataset. The split was configured to allocate 80% of the data to the training set and 20% to the testing set. The ***stratify*** parameter was set to ***target*** to ensure that both training and testing sets had a similar distribution of classes as the original dataset. This approach is particularly beneficial for handling imbalanced datasets, as it helps maintain class proportions.

Furthermore, Stratified K-Fold cross-validation was employed in conjunction with Optuna for hyperparameter tuning. Utilizing scikit-learn's StratifiedKFold with 5 splits, this method provided a reliable approach for assessing the model's performance across different subsets of the data during the hyperparameter optimization process. Optuna's integration into this methodology enabled a comprehensive and unbiased determination of the optimal hyperparameters, enhancing the robustness and generalizability of the experimental findings.

7 RESULTS

This section presents the findings of the machine learning models applied in this study — Naive Bayes, SVM, Random Forest, XGBoost, and TabNet — as detailed in Section 5.4. Integral to the efficacy of these models was the optimization of their hyperparameters, and Table 5 shows the best hyperparameters that yielded the results. In line with the research objectives outlined in Section 3.3, this section provides a thorough analysis of each model's ability to predict online consumer behavior. The outcomes are evaluated based on precision, recall, F1 score, their weighted averages, ROC, and the analysis of the confusion matrix. These metrics were identified as the most appropriate evaluation criteria for the class-imbalanced dataset, as discussed in Section 5.5.

Model	Optimal Hyperparameter Set
SVM	C = 0.0106, gamma = auto, kernel = poly, degree = 5, coefo = 9.64, class_weight = None
XGBoost	n_estimators = 300, max_depth = 27, learning_rate = 0.07586, gamma = 0.26, colsample_bytree = 0.5475, subsample = 0.8749, min_child_weight = 1
Random Forest	n_estimators = 280, max_depth = 24, min_samples_split = 2, min_samples_leaf = 2, bootstrap = False
TabNet	mask_type = entmax, n_d = 48, n_a = 18, batch_size = 256, virtual_batch_size = 128

Table 5: Optimal Hyperparameter Set for Each Model

7.1 *Models Performance Analysis*

The results of the models are presented below. Detailed confusion matrices, further interpretation of the results, and comparisons between models are included in Appendix D 9.

7.1.1 *Naive Bayes (Baseline)*

The Naive Bayes model, utilized as the baseline, offers an initial glimpse into the predictability of consumer behavior. Its performance is assessed by examining precision, recall, F1-score, and a confusion matrix.

Class	Precision	Recall	F1-Score	Support
No Revenue	0.943	0.857	0.898	2084
Revenue	0.478	0.717	0.574	381
Weighted Average	0.871	0.835	0.848	2465

Table 6: Naive Bayes Test Metrics

Performance Metrics:

- **Precision and Recall:**
  - **No Revenue Class (Non-Purchasers):** The model achieved a high precision of 94.3%, indicating its effectiveness in correctly identifying sessions that did not result in a purchase. However,

the recall of 85.7% suggests that while it is quite good at catching non-purchasing behavior, it misses some non-purchasing behavior.

- **Revenue Class (Purchasers):** The precision drops to 47.8%, indicating a substantial number of false positives in predicting actual purchases. Nonetheless, the recall is better at 71.7%, showing the model's reasonable ability to identify genuine purchasing sessions.

- **F1 Score:** The F1 score for the No Revenue class stands at 89.8%, indicating a strong balance between precision and recall. Conversely, the Revenue class has a lower F1 score of 57.4%, suggesting room for improvement in balancing false positives and false negatives in purchase prediction.

7.1.2 *SVM*

The performance of SVM model is examined next:

Class	Precision	Recall	F1-Score	Support
No Revenue	0.924	0.911	0.917	2084
Revenue	0.548	0.588	0.567	381
Weighted Average	0.866	0.861	0.863	2465

Table 7: SVM Test Metrics

Performance Metrics:

- **Precision and Recall:**
  - **No Revenue Class (Non-Purchasers):** Precision for non-purchasers is high at 92.4%, indicating a strong ability to correctly identify non-purchasing sessions. The recall is similarly high at 91.1%, suggesting that the model effectively recognizes most non-purchaser instances.
  - **Revenue Class (Purchasers):** The precision for predicting purchases is lower at 54.8%, pointing to a considerable rate of false positives. However, a recall of 58.8% indicates a moderate capability in identifying actual purchase sessions.

- **F1 Score:** The F1 score for the No Revenue class is a robust 91.7%, showing a well-balanced precision-recall tradeoff. For the Revenue class, the F1 score is 56.7%, highlighting an area for enhancement in accurately categorizing purchasing behavior.

7.1.3 *XGBoost*

The XGBoost model was evaluated as well for its effectiveness in prediction. Performance metrics are detailed below:

Class	Precision	Recall	F1-Score	Support
No Revenue	0.936	0.933	0.935	2084
Revenue	0.641	0.651	0.646	381
Weighted Average	0.890	0.890	0.890	2465

Table 8: XGBoost Test Metrics

Performance Metrics:

- **Precision and Recall:**
  - **No Revenue Class (Non-Purchasers):** The model achieved a high precision of 93.6% and a recall of 93.3%, indicating its strong capability in correctly identifying non-purchaser sessions.
  - **Revenue Class (Purchasers):** Precision and recall for purchasers were 64.1% and 65.1% respectively, showing reasonable effectiveness in identifying actual purchasing sessions, though with room for improvement.
- **F1 Score:** The No Revenue class's F1 score of 93.5% indicates a high balance between precision and recall. The Revenue class's F1 score of 64.6% suggests a decent balance.

7.1.4 *Random Forest*

The Performance metrics for Random Forest are detailed below:

Class	Precision	Recall	F1-Score	Support
No Revenue	0.939	0.935	0.937	2084
Revenue	0.651	0.667	0.659	381
Weighted Average	0.894	0.893	0.894	2465

Table 9: Random Forest Test Metrics

Performance Metrics:

- **Precision and Recall:**

  - **No Revenue Class (Non-Purchasers):** The model exhibits a high precision of 93.9% and recall of 93.5%, indicating its strong capability in correctly identifying non-purchaser sessions.
  - **Revenue Class (Purchasers):** For purchasers, the precision of 65.1% and recall of 66.7% are moderate. These figures show that while the model is reasonably effective in identifying actual purchasing sessions, it is not as proficient as it is with the No Revenue Class.

- **F1 Score:** The F1 score of 93.7% for the No Revenue class indicates excellent precision and recall balance, while the 65.9% F1 score for the Revenue class indicates fair balance between precision and recall.

7.1.5 *TabNet*

TabNet was also assessed with the following performance metrics:

Class	Precision	Recall	F1-Score	Support
No Revenue	0.939	0.901	0.919	2084
Revenue	0.555	0.677	0.610	381
Weighted Average	0.879	0.866	0.871	2465

Table 10: TabNet Test Metrics

Performance Metrics:

- **Precision and Recall:**
  - **No Revenue Class (Non-Purchasers):** TabNet shows high precision at 93.9% and recall at 90.1%, indicating effective identification of non-purchaser sessions.
  - **Revenue Class (Purchasers):** For purchasers, the model has a precision of 55.5% and a higher recall of 67.7%, suggesting its capability to identify actual purchases, albeit with a higher rate of false positives.

- **F1 Score:** The F1 score of 91.9% for the No Revenue class shows a strong balance between precision and recall. In the Revenue class, the F1 score is 61.0%, reflecting a reasonable balance but indicating room for improvement as well.

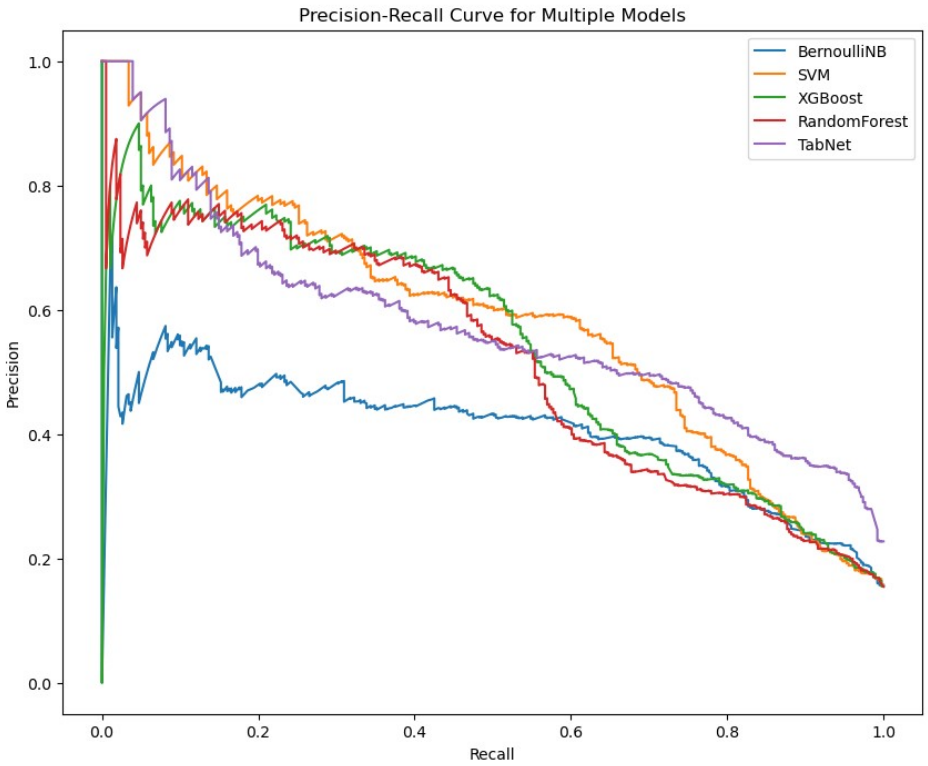

Figure 7: Precision-Recall Curves

7.1.6 *Performance Variation*

The Naive Bayes algorithm displayed the lowest F1 score (Table 15), while SVM recorded the lowest AUC. Conversely, Random Forest outperformed all the other models in all evaluated metrics. Moreover, the results revealed that all models were more effective in predicting the No Revenue class compared to the Revenue class (Table 16). This variation may be associated with the class imbalance in the dataset, addressed through SMOTE oversampling. Regarding computational efficiency, Naive Bayes demonstrated the fastest runtime. Random Forest and XGBoost exhibited comparable processing times, with SVM requiring longer. TabNet, however, necessitated the most extended processing period among the models tested.

7.2 *Comparative ROC Curve Analysis of the Predictive Models*

In assessing the performance of various predictive models for online shopping behavior, the ROC curve analysis presents an insightful look into the model performances, as illustrated in Figure 8. Random Forest emerges as the most proficient model with an AUC of 0.92, indicating its superior

ability to distinguish between purchasers and non-purchasers. XGBoost closely follows with an AUC of 0.91, reinforcing its efficacy in accurate classification. Interestingly, TabNet, with an AUC of 0.89, demonstrates commendable performance and highlights its effectiveness as well in distinguishing between purchasers and non-purchasers. In comparison, the Bernoulli Naive Bayes and SVM models exhibit lower discriminative power, with AUCs of 0.86 and 0.84, respectively. While still competent, these models suggest a marginally reduced capability in differentiating between the two classes.

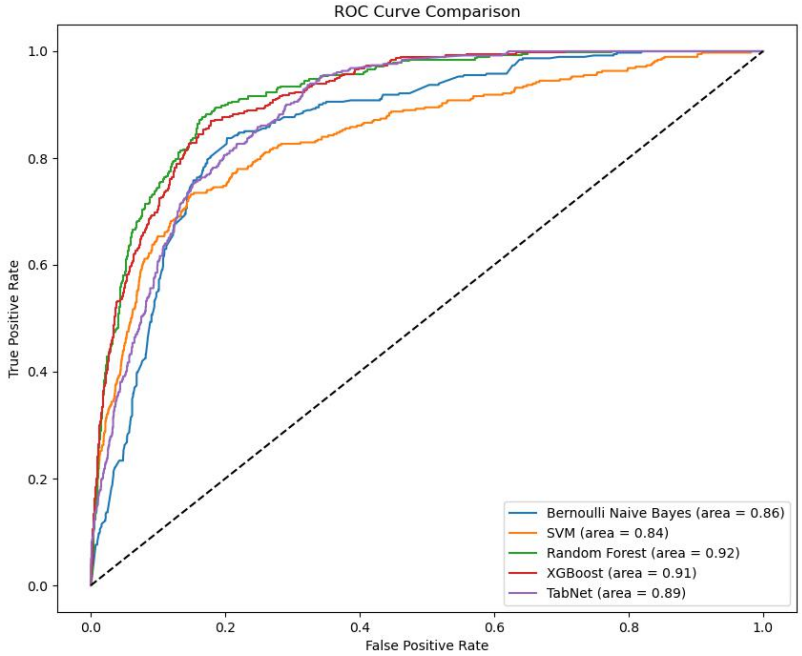

Figure 8: ROC Curve Comparison of Predictive Models

7.3 *Feature Importance*

To discern the most influential features in predicting online consumer behavior, a feature importance analysis was conducted using the Random Forest model, identified as the best performer. This analysis began with the training of the Random Forest classifier on the preprocessed dataset. Following the completion of the model training, the feature importance was extracted to understand which features most greatly influence consumer decisions to complete a transaction (make a purchase).

As illustrated in Figure 9, the most Influential predictor identified is ***Page Values***, highlighting the impact of the perceived value of a page on a consumer's purchasing decision to complete a transaction. ***Exit Rates*** and ***Product Related Duration*** follow, highlighting the importance of user engagement and interaction with product-related content. Additionally, ***Product Related***, ***Administrative Duration***, and ***Bounce Rates*** are notable predictors, emphasizing the roles of both the quantity and quality of user interactions with various site elements. Seasonal influences, as indicated by month-specific features, are also evident, suggesting shifts in consumer behavior at various times of the year. Collectively, these insights provide a multifaceted understanding of the drivers behind online consumer behavior. This also aligns with what was seen during feature correlations in the exploratory data analysis section.

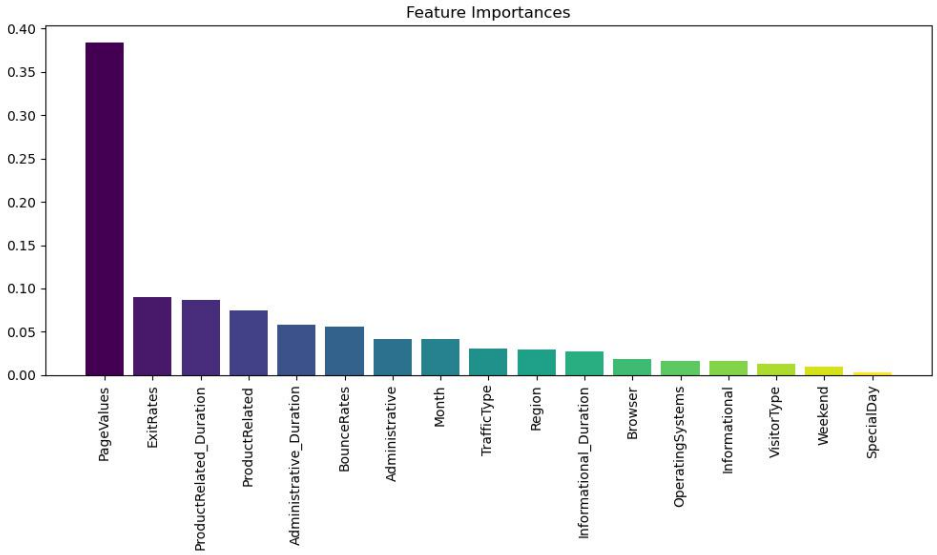

Figure 9: Feature Importance with Random Forest

8 DISCUSSION

8.1 *Summary and Discussion of the Results*

This research was initiated with the aim of deepening our comprehension of consumer behavior in the e-commerce sector, specifically through the application of machine learning models to clickstream data. The study was driven by two primary goals: firstly, to examine the capability of traditional machine learning algorithms, including Naive Bayes, SVM, Random Forest, and XGBoost, in accurately predicting consumer behavior in online settings;

and secondly, to explore how the TabNet architecture might enhance the predictive performance within this particular field. Subsequently, the research delved into analyzing which features exert the most influence on the completion of online transactions.

The findings, as summarized in Table 15 which presents the weighted average scores for all models, indicated that Random Forest emerged as the most effective among the traditional algorithms, effectively addressing the first research question. This model demonstrated a superior balance between precision and recall, evidenced by an F1-score of 89.4% and an AUC score of 0.92. XGBoost followed closely, displaying a slightly lower F1-score of 89.0% and an AUC score of 0.91. On the other hand, SVM and Naive Bayes (baseline), while still effective, lagged behind in their performance. SVM demonstrated a balanced outcome with an F1-score of 86.3% and an AUC score of 84.0%. Naive Bayes, with a slightly lower precision but higher recall, achieved an F1-score of 84.8% and an AUC score of 0.86.

This nuanced variation in model performances leads us to the second research question, where TabNet's role comes into focus. Despite its robust architecture, TabNet did not markedly surpass the traditional algorithms, especially Random Forest, in terms of overall effectiveness. However, it did outperform both SVM and Naive Bayes, indicating its competitive edge over these models. These results suggest that while advanced deep learning models like TabNet are promising, traditional machine learning algorithms, especially ensemble methods like Random Forest, remain highly competitive in handling the complexities and nuances of clickstream data in e-commerce.

In the process of addressing the research sub-question, the strengths of the Random Forest model (best performer) were leveraged to explore the importance of various features. This analysis was conducted with the aim of identifying the factors most influential in predicting consumer behavior. It was revealed that Page Values and Exit Rates held considerable influence over consumer decisions. These findings align well with the overarching goal of the study: the accurate prediction of transaction completions in online shopping environments. The impact of key features like Page Values and Exit Rates, once identified, enhances our understanding of which features could influence consumer behavior in these settings.

In light of the detailed analyses, it is evident that while all models are adept at predicting the No Revenue class, their ability to accurately classify the Revenue class (minority class) is relatively lower. This variation in performance can likely be attributed to the class imbalance in the dataset, with the use of SMOTE oversampling to address this imbalance potentially introducing additional complexities in model training and prediction.

The evaluation of models' performance also extended to their computational efficiency. It was observed that the Naive Bayes algorithm demonstrated the quickest runtime, significantly outpacing the other models in terms of speed. In contrast, Random Forest, and XGBoost exhibited relatively similar runtime durations, each striking a balance between computational demand and predictive performance. While SVM required more time than Random Forest and XGBoost, TabNet notably required the longest processing time among all the models evaluated. This extended runtime underscores TabNet's computational intensity, highlighting it as the most resource-demanding model in the study. Such a consideration is crucial in practical applications where computational resources and time constraints are of the essence.

8.2 *Comparison to the Literature*

The results of this study present both contrasts and similarities when compared to the existing literature, offering intriguing insights. For instance, the Random Forest model in this study achieved an F1-score of 89.4%, which is notably higher than the 60% reported by Sang and Wu (2022), as detailed in Table 14. This discrepancy might be attributed to various factors such as differences in dataset characteristics, feature engineering approaches, or data preprocessing methods. On the other hand, the XGBoost model's performance in this study, with an F1-score of 89.0%, is lower than the 97.6% reported by Wang et al. (2023). This discrepancy may be attributed to differences in dataset complexity, hyperparameter optimization processes, or potentially due to the dataset's imbalance and the application of SMOTE oversampling, which is known to generate synthetic data.

Furthermore, the performance of the TabNet model in our study, reflected by an F1-score of 87.1%, does not quite reach the 88.3% level reported in the study by Joseph et al. (2022). This somewhat lower performance of TabNet, compared to some traditional algorithms, offers a distinct perspective. This finding stands in partial contrast to previous studies, such as the one by Arık and Pfister (2021), which have underscored the superior performance of deep learning methods, with TabNet demonstrating notable advancements over traditional models. This divergence in performance could be attributed to specific architectural adjustments or training methods employed in the different studies, highlighting the nuanced nature of machine learning model efficacy across various contexts.

The Gradient Boost model used in the study by Renuka (2023) demonstrated a lower F1-score of 37.3% compared to the XGBoost model in this study. This outcome is somewhat unexpected and suggests that the effectiveness of gradient boosting algorithms may vary greatly, depending

on the specific characteristics of the dataset and the problem domain. In contrast, while the Random Forest model in this study was highly effective with an F1-score of 89.4% and a ROC-AUC of 92.0%, it did not quite match the performance of the XGBoost model reported in Wang et al. (2023)'s research. The disparity in their performances is particularly noteworthy: the XGBoost model in Wang et al. (2023)'s study achieved a remarkable F1-score of 97.6% and a ROC-AUC of 97.7%, outperforming our Random Forest model.

8.3 *Discussion of Scientific and Societal Impact*

This research contributes to e-commerce analytics by showcasing the effectiveness of machine learning models in predicting consumer behavior, particularly ensemble methods, to clickstream data. This approach not only enriches our understanding of consumer behavior but also transforms each online interaction into valuable business insights. The application of these models signifies a notable advancement in e-commerce personalization, allowing for the development of more tailored user experiences. This includes personalized user interfaces, targeted marketing strategies, and dynamic pricing mechanisms, all key to increasing user engagement and satisfaction.

Moreover, the insights from this study empower businesses to make well-informed decisions in critical areas like inventory management, marketing, and customer service, thereby improving strategies and market positioning. Additionally, this research aligns e-commerce offerings with consumer preferences, contributing to societal benefits such as enhanced shopping experiences, increased customer satisfaction, economic growth, and job creation. The impact is far-reaching, extending beyond individual consumers to benefit the broader community and economy.

8.4 *Limitations and Future Directions*

This research, while providing valuable insights into e-commerce analytics, recognizes certain limitations. One primary constraint is the study's reliance on a singular dataset, which may not fully capture the diverse spectrum of online consumer behaviors. This aspect, along with the class imbalance present in the dataset, could have influenced the performance of the models, especially the deep learning one like TabNet. Although TabNet performed commendably, its inability to notably surpass models like Random Forest and XGBoost in this context prompts further inquiry. This could be attributed to the specific nature of the clickstream data or the

tendency of deep learning models to require larger datasets for optimal performance.

Looking ahead, there are several promising avenues for future research. Testing these models on varied datasets from multiple e-commerce platforms could validate their effectiveness across different consumer behavior scenarios and confirm their adaptability in various e-commerce environments. Exploring advanced feature engineering and deepening the exploration of deep learning techniques might also yield improvements in predictive performance. Enriching the dataset with additional features such as user demographics could offer a more holistic view of consumer behavior.

Furthermore, with more computational resources, hyperparameter optimization could be expanded, enhancing model performance greatly. Additionally, applying cross-model learning, as evidenced by the success of XGBoost in Wang et al. (2023)'s study, could lead to improvements in models like Random Forest. This approach of integrating strengths from various models promises notable advancements in machine learning for e-commerce analytics.

9 CONCLUSION

Investigating the predictive power of machine learning in e-commerce, this thesis centered on how traditional algorithms (Naive Bayes, SVM, Random Forest, XGBoost) and the TabNet architecture perform on clickstream data. The key questions explored whether conventional algorithms can effectively predict consumer behavior and if TabNet could outperform these methods.

Findings reveal that traditional algorithms, particularly Random Forest, exhibit strong predictive capabilities, evidenced by their high F1-scores and AUC values. Although TabNet showcased promising results, it did not greatly outperform traditional methods, indicating the enduring relevance of ensemble methods like Random Forest in e-commerce analytics.

The insights gained from this research are valuable for e-commerce analytics, demonstrating the efficacy of various machine learning models in deciphering consumer behavior. The study's findings on feature importance are particularly beneficial for e-commerce entities in pinpointing key consumer decision influencers.

Moreover, the research enriches e-commerce by enhancing consumer behavior prediction. Improved predictive performance leads to more tailored services, elevating the shopping experience and potentially boosting customer loyalty and satisfaction. Such advancements contribute to more effective inventory management and targeted marketing, yielding notable economic benefits and eventually to societal overall welfare improvement.

While addressing its primary research questions, this thesis also lays a foundation for future research in this field. The insights and methodologies presented offer possibilities for enriching e-commerce analytics, underscoring the importance of machine learning in understanding and predicting consumer behavior in an increasingly digital marketplace.

REFERENCES

Abdullah-All-Tanvir, Iftakhar, A. K., Islam, M., Islam, S., & Shatabda, S. (2023). A gradient boosting classifier for purchase intention prediction of online shoppers. *Heliyon*, *9*(4), e15163. https://doi.org/10.1016/j.heliyon.2023.e15163

Akbani, R., Kwek, S., & Japkowicz, N. (2004). Applying support vector machines to imbalanced datasets. *Machine Learning: ECML 2004: 15th European Conference on Machine Learning, Pisa, Italy, September 20-24, 2004. Proceedings 15*, 39–50. https://doi.org/10.1007/978-3-540-30115-8_7

Anh, B. N., Giang, N. H., Hai, N. Q., Minh, T. N., Son, N. T., & Chien, B. D. (2023). An university student dropout detector based on academic data. *2023 IEEE Symposium on Industrial Electronics Applications (ISIEA)*, 1–8. https://doi.org/10.1109/ISIEA58478.2023.10212223

Arık, S. Ö., & Pfister, T. (2021, May). TABNET: Attentive Interpretable Tabular Learning. https://doi.org/10.1609/aaai.v35i8.16826

Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, *6*(1), 20–29. https://doi.org/10.1145/1007730.1007735

Blasco-Arcas, L., Lee, H. H., Kastanakis, M. N., Alcañiz, M., & Reyes-Menendez, A. (2022). The role of consumer data in marketing: A research agenda. *Journal of Business Research*, *146*, 436–452. https://doi.org/https://doi.org/10.1016/j.jbusres.2022.07.047

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32. https://doi.org/10.1023/a:1010933404324

Bucklin, R. E., & Sismeiro, C. (2003). A model of website browsing behavior estimated on clickstream data. *Journal of Marketing Research*, *40*(3), 249–267. https://doi.org/https://doi.org/10.1023/A:1020231107662

Bucklin, R. E., & Sismeiro, C. (2009). Click here for internet insight: Advances in clickstream data analysis in marketing. *Journal of Interactive Marketing*, *23*(1), 35–48. https://doi.org/https://doi.org/10.1016/j.intmar.2008.10.004

Cai, K., & Rodavia, M. R. (2023). Xgboost analysis based on consumer behavior. *Frontiers in Computing and Intelligent Systems*, *5*(2), 85–89. https://doi.org/10.54097/fcis.v5i2.12974

Charte, F., Rivera, A. J., del Jesus, M. J., & Herrera, F. (2015). Mlsmote: Approaching imbalanced multilabel learning through synthetic instance generation. *Knowledge-Based Systems*, *89*, 385–397. https://doi.org/https://doi.org/10.1016/j.knosys.2015.07.019

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*, 273–297. https://doi.org/10.1023/A:1022627411411

Dal Pozzolo, A., Caelen, O., Johnson, R. A., & Bontempi, G. (2015). Calibrating probability with undersampling for unbalanced classification. *2015 IEEE symposium series on computational intelligence*, 159–166. https://doi.org/10.1109/SSCI.2015.33

Elkan, C. (2001). The foundations of cost-sensitive learning. *International joint conference on artificial intelligence*, *17*(1), 973–978.

Flores, A. C., Icoy, R. I., Peña, C. F., & Gorro, K. D. (2018). An evaluation of svm and naive bayes with smote on sentiment analysis data set. *2018 International Conference on Engineering, Applied Sciences, and Technology (ICEAST)*, 1–4. https://doi.org/10.1109/ICEAST.2018.8434401

Gao, X. S., Currim, I. S., & Dewan, S. (2022). Validation of the information processing theory of consumer choice: Evidence from travel search engine clickstream data. *European Journal of Marketing*, *56*, 2250–2280. https://doi.org/https://doi.org/10.1108/EJM-07-2020-0503

Gersen, J., & Steckel, J. (2023). Understanding consumer behavior. In J. Gersen & J. Steckel (Eds.), *The cambridge handbook of marketing and the law* (pp. 7–102). Cambridge University Press. https://doi.org/10.1017/9781108699716.002

Ghosh, S., & Banerjee, C. (2020). A predictive analysis model of customer purchase behavior using modified random forest algorithm in cloud environment. *2020 IEEE 1st International conference for convergence in engineering (ICCE)*, 239–244. https://doi.org/10.1109/ICCE50343.2020.9290700

Gumber, M., Jain, A., & Amutha, A. (2021). Predicting customer behavior by analyzing clickstream data. *2021 5th International Conference on Computer, Communication and Signal Processing (ICCCSP)*, 1–6. https://doi.org/https://doi.org/10.1109/ICCCSP52374.2021.9465526

Han, J., Kamber, M., & Pei, J. (2012). Data mining concepts and techniques third edition. *University of Illinois at Urbana-Champaign Micheline Kamber Jian Pei Simon Fraser University*, *49*(06), 49–3305. https://doi.org/10.5860/choice.49-3305

Harris, C. R., Millman, K. J., Van Der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., et al. (2020). Array programming with numpy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Houfani, D., Slatnia, S., Kazar, O., Saouli, H., Eddine, H. S., Remadna, I., & Zouai, M. (2022). Tabnet based prediction model for icu admission in covid-19 patients. *2022 International Symposium on iNnovative Informatics of Biskra (ISNIB)*, 1–6. https://doi.org/10.1109/ISNIB57382.2022.10075767

Hsu, C.-W., Chang, C.-C., Lin, C.-J., et al. (2003). A practical guide to support vector classification.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/mcse.2007.55

Joseph, L., Joseph, E. A., & Prasad, R. (2022). Explainable diabetes classification using hybrid bayesian-optimized tabnet architecture. *Computers in Biology and Medicine*, *151*, 106178. https://doi.org/10.1016/j.compbiomed.2022.106178

Koehn, D., Lessmann, S., & Schaal, M. (2020). Predicting online shopping behaviour from clickstream data using deep learning. *Expert Systems with Applications*, *150*, 113342. https://doi.org/10.1016/j.eswa.2020.113342

Kohavi, R., et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Ijcai*, *14*(2), 1137–1145. http://ijcai.org/Proceedings/95-2/Papers/016.pdf

Kotsiantis, S. B., Kanellopoulos, D., & Pintelas, P. E. (2006). Data preprocessing for supervised leaning. *International journal of computer science*, *1*(2), 111–117. https://doi.org/10.5281/zenodo.1082415

Kukar-Kinney, M., Scheinbaum, A. C., Orimoloye, L. O., Carlson, J. R., & He, H. (2022). A model of online shopping cart abandonment: Evidence from e-tail clickstream data. *Journal of the Academy of Marketing Science*, *50*, 961–980. https://doi.org/https://doi.org/10.1007/s11747-022-00857-8

Lessmann, S., Haupt, J., Coussement, K., & De Bock, K. W. (2019). Targeting customers for profit: An ensemble learning framework to support marketing decision-making [online first]. *Information Sciences*. https://doi.org/10.1016/j.ins.2019.05.027

Liu, X., & Li, J. (2016). Using support vector machine for online purchase predication. *2016 International Conference on Logistics, Informatics and*

Service Sciences (LISS), 1–6. https://doi.org/10.1109/LISS.2016.7854334

Liu, Z. (2023). A new porosity prediction method based on deep learning of tabnet algorithm. *2023 IEEE 2nd International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA)*, 1681–1685. https://doi.org/10.1109/EEBDA56825.2023.10090680

Lombart, C., Millan, E., Normand, J.-M., Verhulst, A., Labbé-Pinlon, B., & Moreau, G. (2020). Effects of physical, non-immersive virtual, and immersive virtual store environments on consumers' perceptions and purchase behavior. *Computers in Human Behavior*, *110*, 106374. https://doi.org/10.1016/j.chb.2020.106374

López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information sciences*, *250*, 113–141. https://doi.org/https://doi.org/10.1016/j.ins.2013.07.007

Maheswari, K., & Priya, P. P. A. (2017). Predicting customer behavior in online shopping using svm classifier. *2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)*, 1–5. https://doi.org/10.1109/ITCOSP.2017.8303085

McKinney, W., et al. (2010). Data structures for statistical computing in python. *Proceedings of the 9th Python in Science Conference*, *445*(1), 51–56. https://doi.org/10.25080/majora-92bf1922-00a

Organisation for Economic Co-operation and Development (OECD). (2021). The COVID-19 pandemic and the acceleration of e-commerce [Accessed: 2023-11-11]. www.oecd.org/coronavirus/policy-responses/e-commerce-in-the-time-of-COVID-19-3a2b78e8

Oshiro, T. M., Perez, P. S., & Baranauskas, J. A. (2012). How many trees in a random forest? *Machine Learning and Data Mining in Pattern Recognition: 8th International Conference, MLDM 2012, Berlin, Germany, July 13-20, 2012. Proceedings 8*, 154–168. https://doi.org/10.1007/978-3-642-31537-4\{_}13

Pal, G., Li, G., & Atkinson, K. (2018). Big data real-time clickstream data ingestion paradigm for e-commerce analytics. *2018 4th International Conference for Convergence in Technology (I2CT)*, 1–5. https://doi.org/10.1109/I2CT42659.2018.9058112

Pearson, R. K. (2018). *Exploratory data analysis using r*. CRC Press. https://doi.org/10.1201/9781315382111

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine*

Learning research, *12*, 2825–2830. https://inria.hal.science/hal-00650905

Prayogo, R. D., & Karimah, S. A. (2021). Feature selection and adaptive synthetic sampling approach for optimizing online shopper purchase intent prediction. *2021 8th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, 1–5. https://doi.org/10.1109/ICAICTA53211.2021.9640270

Renuka, D. S. M. (2023). Predicting online customer purchase using gradient boost classifier. *International Journal For Science Technology And Engineering*, *11*(6), 3787–3791. https://doi.org/10.22214/ijraset.2023.54192

Rish, I., et al. (2001). An empirical study of the naive bayes classifier. *IJCAI 2001 workshop on empirical methods in artificial intelligence*, *3*(22), 41–46. https://www.researchgate.net/publication/228845263_An_Empirical_Study_of_the_Naive_Bayes_Classifier

Sakar, C. O., Polat, S. O., Katircioglu, M., & Kastro, Y. (2019). Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and lstm recurrent neural networks. *Neural Computing and Applications*, *31*, 6893–6908. https://doi.org/10.1007/s00521-018-3523-0

Sang, G., & Wu, S. (2022). Predicting the intention of online shoppers' purchasing. *2022 5th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE)*, 333–337. https://doi.org/10.1109/AEMCSE55572.2022.00074

Seger, C. (2018). An investigation of categorical variable encoding techniques in machine learning: Binary versus one-hot and feature hashing [URN: urn:nbn:se:kth:diva-237426, OAI: oai:DiVA.org:kth-237426, DiVA, id: diva2:1259073].

Smiti, A. (2020). A critical overview of outlier detection methods. *Computer Science Review*, *38*, 100306. https://doi.org/https://doi.org/10.1016/j.cosrev.2020.100306

Statista. (2023). Global retail e-commerce sales 2014-2026 [Accessed: August 4, 2023]. https://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales/

Tang, L., Wang, A., Xu, Z., & Li, J. (2017). Online-purchasing behavior forecasting with a firefly algorithm-based svm model considering shopping cart use. *Eurasia journal of mathematics, science and technology education*, *13*(12). https://doi.org/10.12973/ejmste/77906

Tukey, J. W., et al. (1977). *Exploratory data analysis* (Vol. 2). Reading, MA. https://doi.org/10.2307/2529486

UCI Machine Learning Repository. (2018). Online shoppers purchasing intention data set.

Van Rossum, G., et al. (2007). Python programming language. *USENIX annual technical conference*, *41*(1), 1–36. https://dblp.uni-trier.de/db/conf/usenix/usenix2007.html#Rossum07

Wang, W., Xiong, W., Wang, J., Tao, L., Li, S., Yi, Y., Zou, X., & Li, C. (2023). A user purchase behavior prediction method based on xgboost. *Electronics*, *12*(9), 2047. https://doi.org/10.3390/electronics12092047

Webb, G. I., Keogh, E., & Miikkulainen, R. (2010). Naıve bayes. *Encyclopedia of machine learning*, *15*(1), 713–714. https://doi.org/10.1007/978-1-4899-7502-7_581-1

Wen, Lin, W., Liu, H., & Zhanming. (2023). Machine-learning-based approach for anonymous online customer purchase intentions using clickstream data. *Systems*, *11*(5), 255. https://doi.org/10.3390/systems11050255

Wen, Z. J., Lin, W., & Liu, H. (2023). Machine-learning-based approach for anonymous online customer purchase intentions using clickstream data. *Systems*, *11*(5), 255. https://doi.org/10.3390/systems11050255

Wu, J., Lu, F., Zhang, J., Yang, J., & Xing, L. (2019). Choice of distribution channels for experience products using virtual reality. *IEEE Access*, *7*, 85319–85326. https://doi.org/10.1109/ACCESS.2019.2920308

Zhang, W., & Wang, M. (2021). An improved deep forest model for prediction of e-commerce consumers' repurchase behavior. *Plos one*, *16*(9), e0255906. https://doi.org/10.1371/journal.pone.0255906

APPENDIX A

Table 11: Description of Numerical Features

Feature	**Feature Description**	**Min Value**	**Max Value**
Administrative	Number of pages, which are visited by the visitor, related account management	0	27
Administrative duration	Spent time on account management (in seconds)	0	3398
Informational	Number of pages visited with generic information	0	24
Informational duration	Total duration (measured in seconds) that the visitor spent on informational pages.	0	2549
Product related	Number of product-related pages visited	0	705
Product related duration	Time spent by the visitor on product-related pages (seconds)	0	63973
Bounce Rate	The average bounce rate value of the navigated pages	0	0.2
Exit Rate	The average exit rate value of the navigated pages	0	0.2
Page Value	The average page value of the navigated pages	0	361
Special Day	Nearness (closeness) of the browsing session time to a special day	0	1.0

Table 12: Description of Categorical Features

Feature	**Feature Description**	**Levels**
Operating Systems	Operating system of the online shopper	8
Browser	Browser of the online shopper	13
Region	Location of the online shopper's browsing session	9
Traffic Type	Source of traffic leading to the shopping site	20
Visitor Type	Type of visitor as "new," "returning," "others"	3
Weekend	Showing if the visit occurred on a weekend or not	2
Month	The Month of the navigation (browsing)	12
Revenue	Showing if the visit resulted in a sale	2

Table 13: Software and Packages Version

Software/Library	Version
Python	3.9.16
Pandas	1.5.2
NumPy	1.23.5
SciPy	1.10.0
Matplotlib	3.6.2
Seaborn	0.12.2
Scikit-learn	1.0.2
Imbalanced-learn	0.10.1
XGBoost	1.7.5
PyTorch-TabNet	4.1.0
Optuna	3.4.0

Table 14: Some Results of Existing Studies

Study	**Model**	**Precision (%)**	**Recall (%)**	**F1 Score (%)**	**ROC (%)**
(Z. J. Wen et al., 2023)	MBT-POP	—	—	90.3	—
Zhang and Wang (2021)	IDF	88.6	84.2	86.3	—
Wang et al. (2023)	XGBoost	—	—	97.6	97.7
Renuka (2023)	Gradient Boost	67.6	59.2	37.3	—
Sakar et al. (2019)	MLP	—	—	86.0	—
Sang and Wu (2022)	RF	—	—	60.0	—
Joseph et al. (2022)	BO-TabNet	89.5	87.2	88.3	—

Table 15: Performance Comparison of the Employed Machine Learning Models (Weighted Average Scores)

Models	**Precision** (%)	**Recall** (%)	**F1-Score** (%)	**ROC-AUC** (%)
Naive Bayes (Baseline Model)	87.1	83.5	84.8	86.0
SVM	86.6	86.1	86.3	84.0
Random Forest	**89.4**	**89.3**	**89.4**	**92.0**
XGBoost	89.0	89.0	89.0	91.0
TabNet	87.9	86.6	87.1	89.0

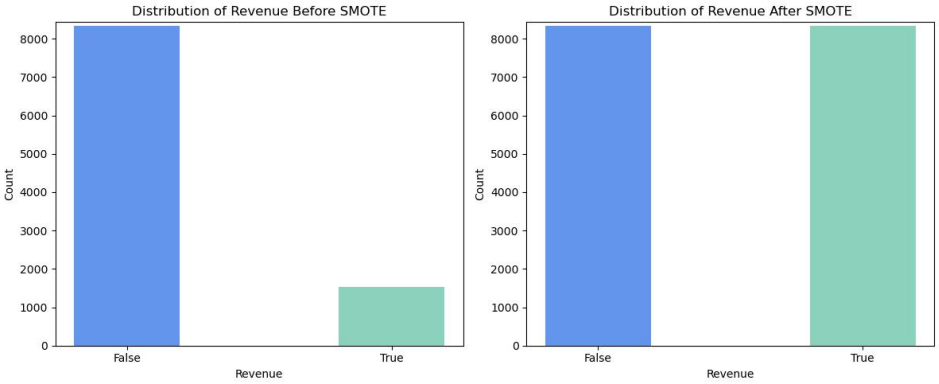

Figure 10: Distribution of Revenue Before and After Applying SMOTE

As shown in Figure 10, the application of SMOTE balance the classes in the revenue variable (target).

APPENDIX B

Further Exploratory Data Analysis

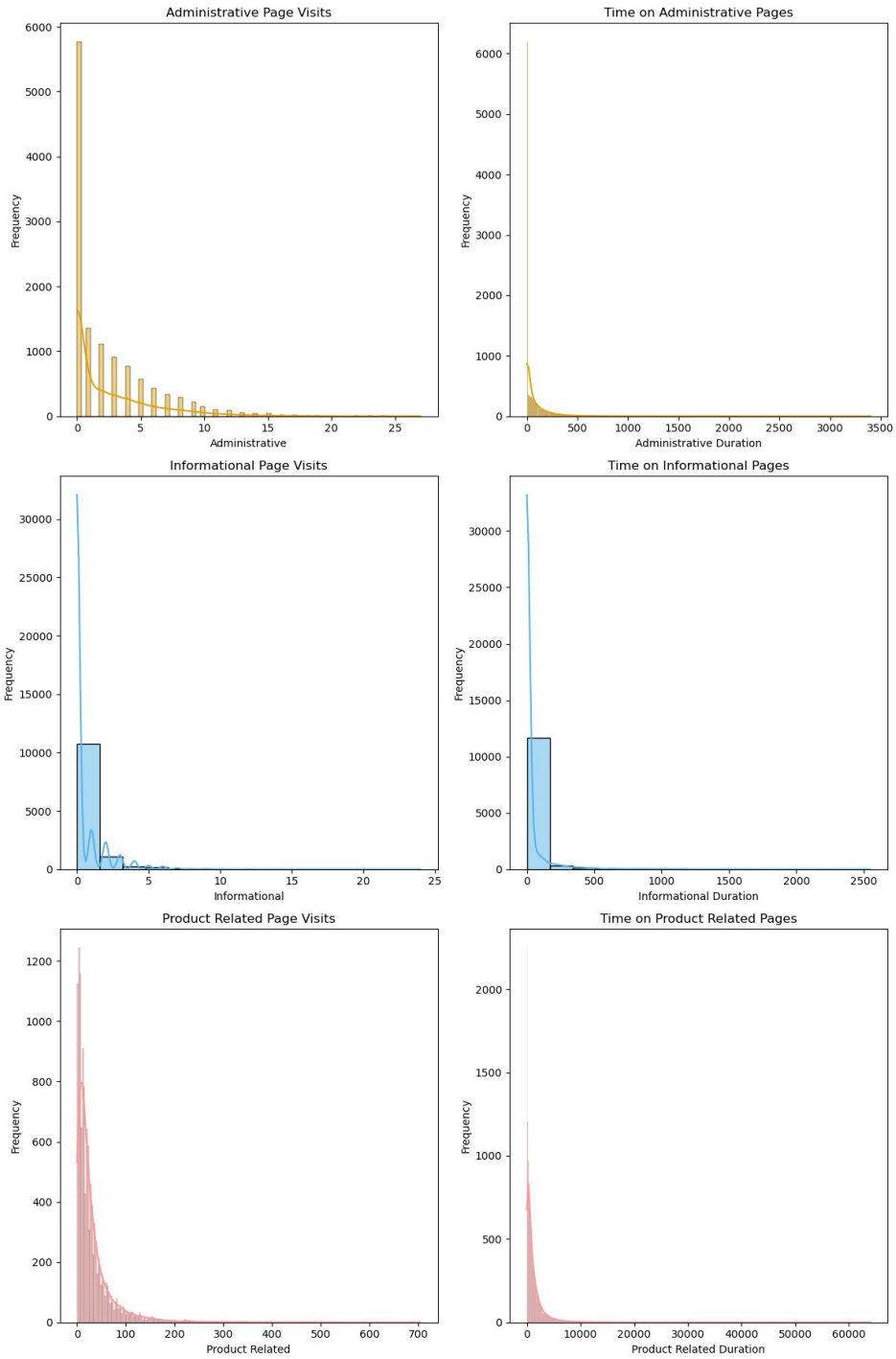

Figure 11: Distribution for Pages Related Features

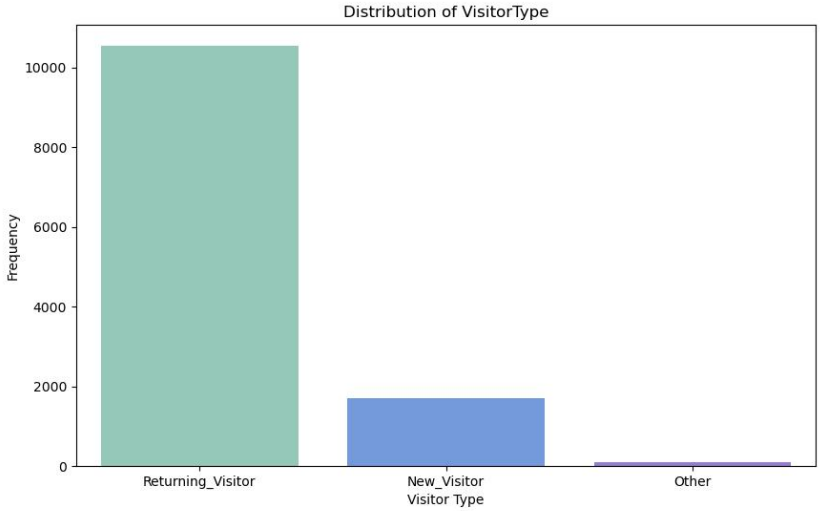

Figure 12: Distribution of Visitor Types

A bar chart was constructed to explore the marginal distribution of VisitorType, aiding in a better understanding of this variable. As illustrated in Figure 12, it's evident that returning visitors constitute the majority, followed by new visitors, while the 'Other' category of visitors is the least frequent.

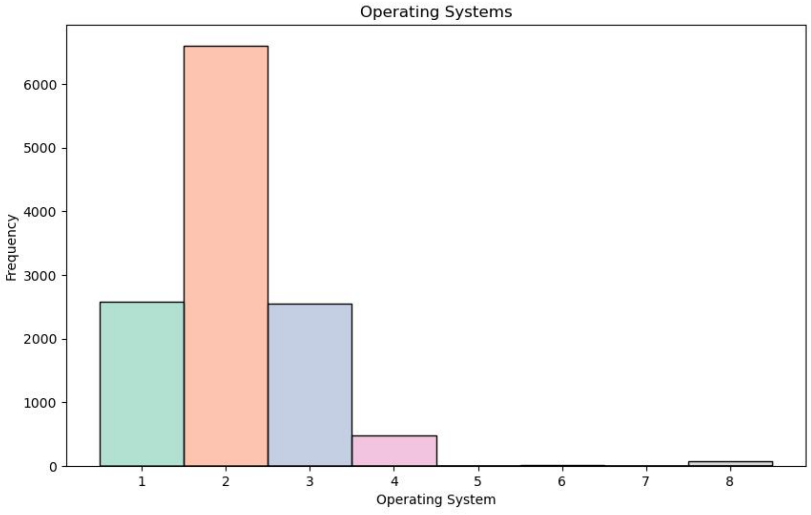

Figure 13: Operating Systems Distribution

In figure 13, various operating systems are observed, each identified by a numerical label. The predominant use of operating system 2 is noticeable.

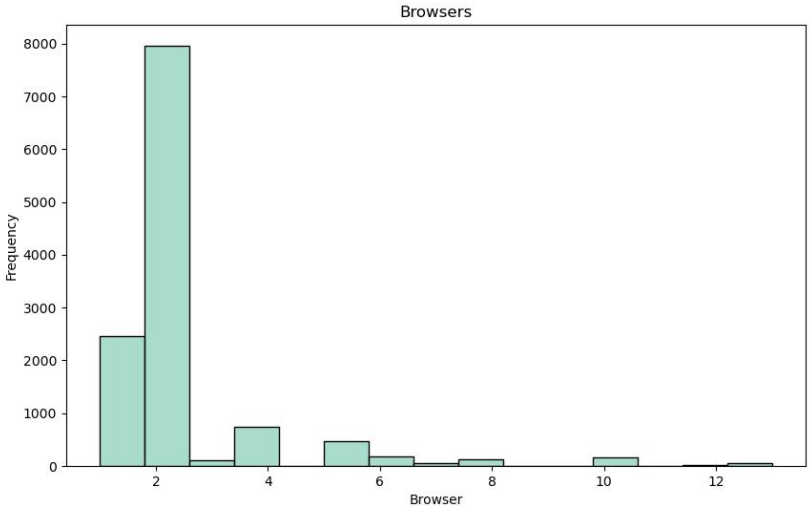

Figure 14: Browsers Distribution

Here in figure 14, it is observed that the vast majority of users prefer browser 2, followed by browser 1. Other browsers account for a smaller portion of the online users.

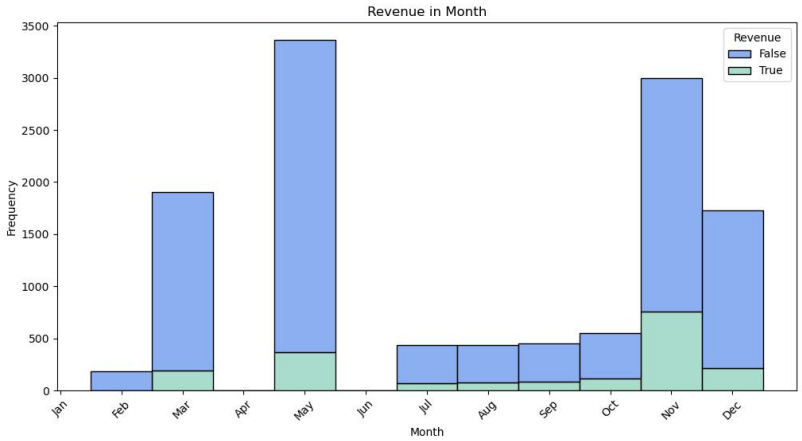

Figure 15: Revenue in Month

From figure 15, it is clear that most purchases occurred in March, May, November, and December.

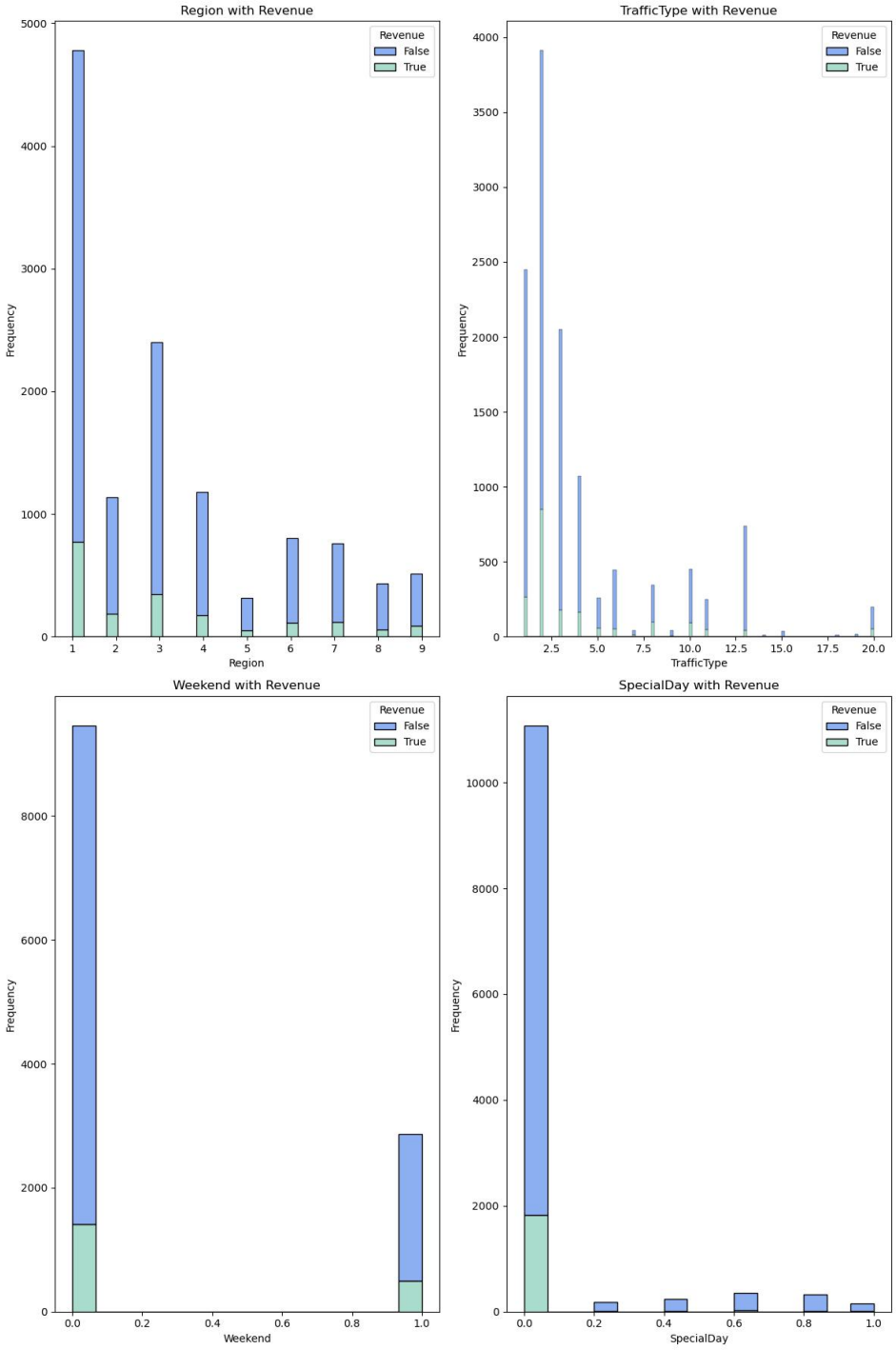

Figure 16: Other Features with Revenue

APPENDIX C

Hyperparameter Tuning

A) Tuning Naive Bayes: More details are mentioned in Section 5.4.6 (page 24).

B) Tuning SVM: For SVM, hyperparameter tuning focuses on hyperparameters like the kernel type, regularization hyperparameter ***C***, kernel coefficient ***gamma***, degree of the polynomial kernel ***degree***, independent term in the kernel function ***coefo***, and the class weighting ***class_weight***. The ***kernel*** type, which includes options such as ***linear***, ***polynomial***, or ***radial basis function (RBF)***, influences the decision boundary's shape (Hsu et al., 2003). The ***C*** hyperparameter controls the trade-off between achieving a low training error and a low testing error, crucial in preventing overfitting. ***Gamma*** in the RBF kernel defines the reach of a single training example, with low values indicating a wide reach and high values a close reach. The ***degree*** hyperparameter is important when using a polynomial kernel, determining the polynomial's complexity. ***Coefo*** is an independent term in polynomial and sigmoid kernels, influencing their flexibility in higher-dimensional space. Finally, ***class_weight***, which adjusts the weight of classes, is critical in datasets with class imbalance, aiding in balanced error penalization. Table 1 presents these hyperparameters along with their respective values.

C) Tuning XGBoost: XGBoost also offers a range of hyperparameters. ***N_estimators*** refers to the number of gradient boosted trees, and ***max_depth*** controls the maximum depth of each tree. These hyperparameters, similar to those in Random Forest, require careful tuning to balance the bias-variance trade-off. The ***learning_rate***, or ***eta***, affects the step size shrinkage used in model updates to prevent overfitting. Hyperparameters such as ***gamma***, ***colsample_bytree***, and ***subsample*** are crucial for the model's regularization, feature sampling, and training instance sampling, playing important roles in performance and generalization. Additionally, ***min_child_weight*** is important for controlling tree complexity in XGBoost, aiding in the prevention of overfitting (Chen & Guestrin, 2016). Table 2 shows these hyperparameters and their respective value space.

D) Tuning Random Forest: In Random Forest, several key hyperparameters play an important role in model performance. The ***n_estimators***, determining the number of trees in the forest, enhances model robustness. However, it's important to note the possibility of diminishing returns beyond a certain number of trees (Oshiro et al., 2012). The ***max_depth*** of each tree, crucial for capturing data complexities, needs careful calibration to avoid overfitting. ***min_samples_split*** and ***min_samples_leaf*** set the

minimum number of samples required to split an internal node and to be at a leaf node, respectively. The ***bootstrap*** hyperparameter decides whether to use different subsets of the data (bootstrap samples) or the entire dataset for training individual trees. These hyperparameters help prevent overfitting by ensuring sufficient data at each stage of decision-making. Balancing them is essential for an optimal Random Forest model. Table 3 details these hyperparameters and their values.

E) Training TabNet: For the TabNet model, the training process was designed to achieve a balance between training-related and architectural hyperparameters. While certain hyperparameters like n_d, n_a, and mask_type were chosen for their architectural importance, others related to the training process were set to optimize performance adaptively.

A starting learning rate ***lr*** of 0.001 was selected, with the Adam optimizer dynamically adjusting it during training. This adaptive learning rate approach ensures efficient convergence and model optimization. The ***batch_size*** options, ranging from [64, 128, 256], were included in the hyperparameter tuning using Optuna, suitable for the dataset of 12,330 instances. This flexibility in batch size aids in balancing training efficiency with model generalization, tailored to the dataset.

The training of TabNet was constrained with ***max_epochs*** set to 50, establishing an upper limit to training duration. To prevent overfitting, early stopping with a ***patience*** of 10 epochs was implemented, halting the training process if no improvement in model performance was noted over these epochs. The ***virtual_batch_size*** hyperparameter, unique to TabNet, was also optimized to manage efficient data processing and feature transformation.

The ***mask_type*** hyperparameter played an important role in the model's interpretability, influencing the selection and sparsity of features at each decision step. Together, these hyperparameters facilitated the effective training of the TabNet model, ensuring a harmonious balance between learning efficiency, model complexity, and generalization capability. Table 4 details these hyperparameters and their respective values.

A Detailed Explanation of the Used Evaluation Metrics

Precision

Precision is defined as the ratio of true positives (TP) to the sum of true positives and false positives (FP). A high precision indicates that the model has a lower rate of false positives, which is crucial in scenarios where the cost of false positives is high. Precision is especially important in

applications where the reliability of positive predictions is paramount. Its formula is:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall

Recall, also known as sensitivity or true positive rate, measures the proportion of actual positives correctly identified by the model. It is calculated as the ratio of true positives to the sum of true positives and false negatives. This metric is important in situations where missing a positive instance is more critical than incorrectly labeling a negative instance as positive. High recall is desirable in applications such as medical diagnoses, where failing to detect a condition could have serious consequences. The formula for recall is:

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1 Score

The F1 score is a harmonic mean of precision and recall, providing a balance between these two metrics. It is particularly useful when dealing with imbalanced datasets or when the costs of false positives and false negatives are roughly equivalent. The F1 score provides a comprehensive measure of a model's accuracy. Its formula is:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Weighted Average

The weighted average method addresses dataset imbalances by assigning a weight to the performance of each class, proportional to the number of true instances within that class. This approach is particularly informative for imbalanced datasets, such as the one used in this study. It offers a more accurate reflection of the model's performance by considering the prevalence of each class.

Confusion Matrix Analysis

The confusion matrix is an effective tool for visualizing the performance of a classification model. It presents the number of true positives, false positives, true negatives, and false negatives in a matrix format. Analysis of the confusion matrix offers insights into the types of errors made by the model and can guide efforts to improve its performance. This analysis is particularly beneficial for understanding the model's behavior across

different classes and for diagnosing performance issues related to specific classes. The components of the confusion matrix are as follows:

- True Positives (TP): Correct predictions of the positive class.
- False Positives (FP): Incorrect predictions of the positive class, or Type I errors.
- True Negatives (TN): Correct predictions of the negative class.
- False Negatives (FN): Incorrect predictions of the negative class, or Type II errors.

The matrix is typically represented as:

$$\begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix}$$

ROC Curve Analysis and AUC

The Receiver Operating Characteristic (ROC) curve is an important tool for assessing diagnostic tests. It plots the true positive rate (Sensitivity) against the false positive rate (1 - Specificity) at various thresholds. A ROC curve nearing the upper left corner indicates a high level of test accuracy, approaching 100% sensitivity and specificity. The area under the curve (AUC), ranging from 0 to 1, measures a test's ability to distinguish between two groups, such as diseased vs. normal or, in this study, revenue vs. no revenue. A higher AUC reflects better test performance, independent of the classification threshold. This makes the ROC and AUC particularly valuable for evaluating models in imbalanced datasets.

APPENDIX D

This Appendix presents the Confusion Matrices for all evaluated models and provides further insights and comparisons among them. The comparative analysis progresses sequentially, with each model being compared to its predecessor in the following order:

Model	**No Revenue**			**Revenue**		
	Precision	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
Naive Bayes (baseline)	0.943	0.857	0.898	0.478	0.717	0.574
SVM	0.924	0.911	0.917	0.548	0.588	0.567
Random Forest	0.939	0.935	0.937	0.651	0.667	0.659
XGBoost	0.936	0.933	0.935	0.641	0.651	0.646
TabNet	0.939	0.901	0.919	0.555	0.677	0.610

Table 16: Test Metrics for All Models

Naive Bayes

Contextual Interpretation:

By looking at Table 16, the Naive Bayes model demonstrates a notable ability in identifying non-purchasers, as indicated by its high precision and recall for the No Revenue class. This is advantageous for effectively filtering out less likely buyer sessions. However, the model exhibits moderate precision and recall in identifying actual purchasers (the Revenue class), which suggests a need for careful consideration to avoid overlooking potential buying opportunities. These findings underscore the importance of exploring more sophisticated models to better understand and predict the complex patterns of consumer purchasing behavior.

Confusion Matrix Insights:

The confusion matrix, as shown in Figure 17, provides further insights into the model's performance:

- **Correct Identifications:** The model accurately identified 1,786 out of 2,084 non-purchaser sessions and 273 out of 381 purchaser sessions.
- **Misclassifications:** It misclassified 298 sessions as non-purchases and incorrectly identified 108 purchasing sessions as non-purchasing.

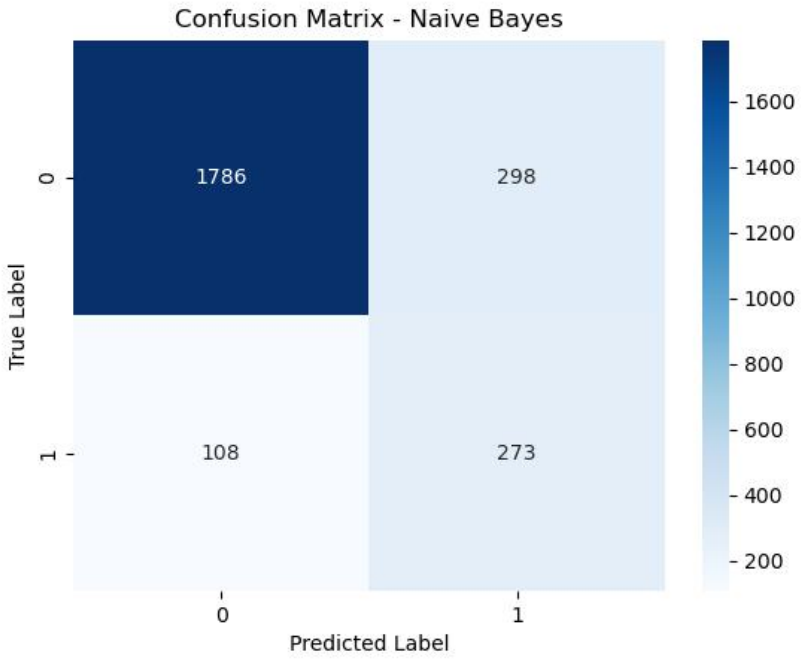

Figure 17: Naive Bayes Confusion Matrix

SVM

Contextual Interpretation:

The SVM model shows strong precision and recall, as presented in Table 16, in identifying non-purchasing behavior, indicating its effectiveness in this aspect. However, with more moderate precision and recall for purchasers, the model demonstrates challenges in accurately predicting actual purchasing sessions. This pattern reveals the model's strengths and areas for improvement, emphasizing the need for a balanced approach in e-commerce strategy formulation.

Comparison:

The SVM model shows a slight improvement in precision for the Revenue class (purchasers) with 54.8% compared to the Naive Bayes model's 47.8%. However, in terms of recall for the Revenue class, the SVM model at 58.8% is lower compared to Naive Bayes at 71.7%. This suggests that while SVM is slightly better at reducing incorrect purchase predictions, Naive Bayes is more effective in identifying actual purchasers, capturing a higher proportion of true buying sessions.

Confusion Matrix Insights:

As depicted in Figure 18, the confusion matrix for the SVM model presents the following observations:

- **Correct Identifications:**
  - Successfully identified 1,899 out of 2,084 non-purchaser sessions.
  - Accurately identified 224 out of 381 purchaser sessions.
- **Misclassifications:**
  - Incorrectly labeled 185 sessions as non-purchases that were actual purchases.
  - Missed identifying 157 purchasing sessions, classifying them as non-purchasing.

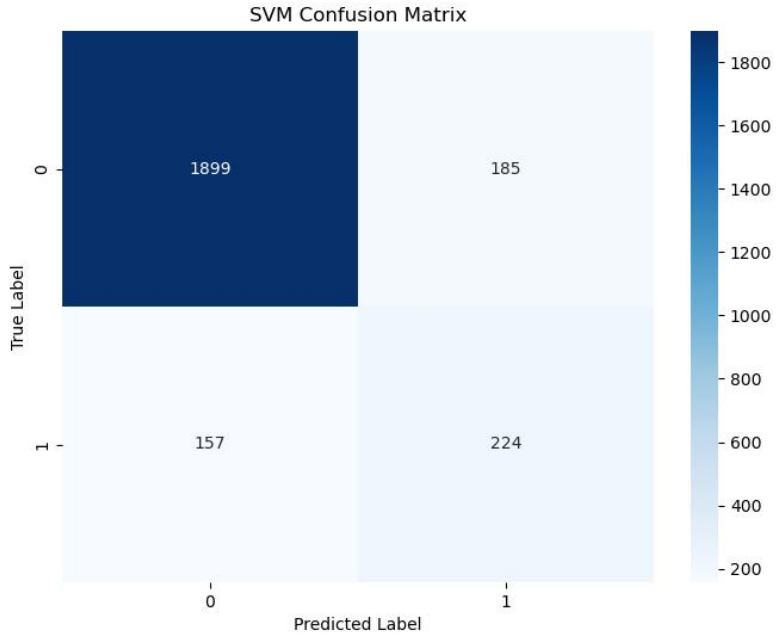

Figure 18: SVM Confusion Matrix

XGBoost

Contextual Interpretation:

XGBoost's strong performance in identifying non-purchasers (No Revenue) with high precision and recall is indicative of its potential in filtering

out less likely buyers efficiently. Table 16 shows the model's ability to predict actual purchases, while moderate, still indicates its usefulness in identifying potential purchasers. This balance in model performance highlights the intricate nature of consumer behavior.

Comparison:

XGBoost demonstrates an improvement in precision for the Revenue class over Naive Bayes, with a precision of 64.1% compared to Naive Bayes' 47.8%. It also gives better precision over SVM 54.8%. However, in terms of recall for the Revenue class, XGBoost's 65.1% is lower than Naive Bayes' 71.7% but higher than SVM's 58.8%. For the No Revenue class, XGBoost's precision and recall are 93.6% and 93.3% respectively, which are comparable to SVM's 92.4% precision and 91.1% recall. The F1 score for XGBoost in the No Revenue class is 93.5%, reflecting a high balance between precision and recall, similar to SVM's F1 score of 91.7%. Both models show a marked improvement in precision for non-purchasers over Naive Bayes. XGBoost achieves 64.6% for the Revenue class, which is an improvement over Naive Bayes' 57.4% and SVM's 56.7%. This suggests that XGBoost strikes a better balance between precision and recall for predicting actual purchasers compared to the other two models.

Confusion Matrix Insights:

Figure 19 displays the confusion matrix, revealing:

- **Correct Identifications:** The model accurately identified 1,945 out of 2,084 non-purchaser sessions and 248 out of 381 purchaser sessions.
- **Misclassifications:** There were 139 non-purchaser sessions falsely identified as purchases and 133 purchaser sessions incorrectly classified as non-purchases.

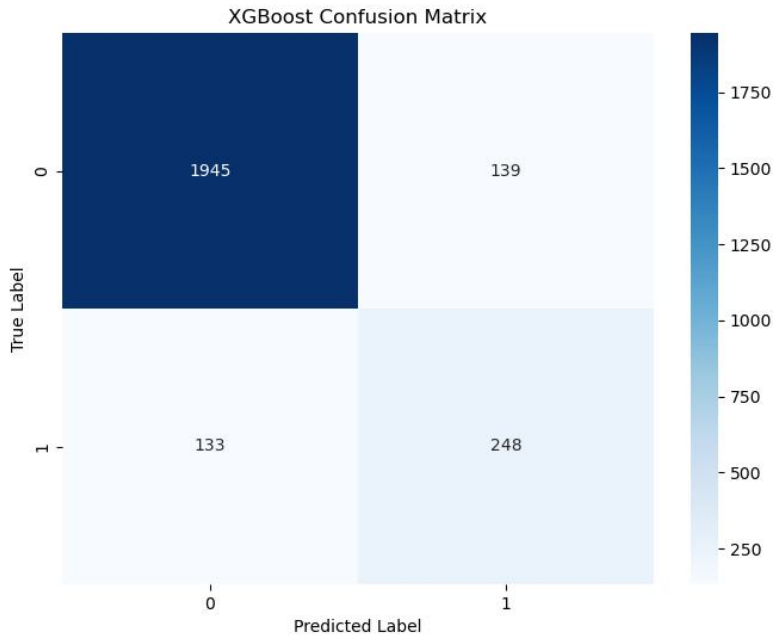

Figure 19: XGBoost Confusion Matrix

Random Forest

Contextual Interpretation:

As seen in Table 16, the Random Forest model shows strong performance, particularly in identifying non-purchasers, with high precision and recall. Its ability to predict actual purchases is also commendable. This model's F1-score, being the highest among the evaluated models, signifies its exceptional performance in balancing the intricacies of purchaser behavior prediction.

Comparison:

Compared to Naive Bayes and SVM, the Random Forest model demonstrates a better balance in precision and recall for both classes. It surpasses Naive Bayes in precision for the Revenue class and shows a higher recall than SVM for the same class, indicating a more effective identification of actual purchasers. For the No Revenue class, the Random Forest model's performance is on par with XGBoost and slightly better than SVM, demonstrating its effectiveness in correctly identifying non-purchaser sessions.

Although the model performs reasonably well for purchasers, the moderate precision and recall suggest complexities in the purchasing behavior that may not be fully captured by the Random Forest algorithm, or it might also be resulted from the nature of the imbalance data. This observation could point to the inherent variability and unpredictability in consumer decision-making processes in online shopping.

Confusion Matrix Insights:

Figure 20 displays the confusion matrix, revealing:

- **Correct Identifications:** Accurately identified 1,948 out of 2,084 non-purchaser sessions and 254 out of 381 purchaser sessions.
- **Misclassifications:** Misclassified 136 non-purchaser sessions as purchases and 127 purchaser sessions as non-purchases.

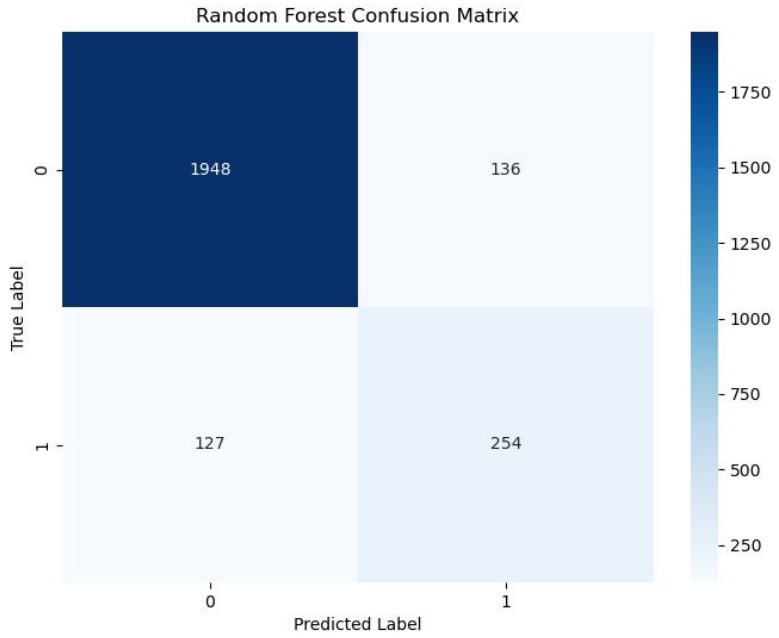

Figure 20: Random Forest Confusion Matrix

TabNet

Contextual Interpretation:

The high precision for non-purchasers and the high recall for purchasers suggest that the model is very reliable in identifying non-purchasers and also moderate at catching most of the purchasers (Table 16).

Comparison:

TabNet shows a higher recall than SVM in the Revenue class, indicating better identification of actual purchasers, with comparable precision. Compared to Naive Bayes, TabNet has higher precision but slightly lower recall, reflecting a more balanced detection of purchases. When compared with Random Forest and XGBoost, TabNet has a similar recall but lower precision for the Revenue class. For the No Revenue class, TabNet's precision and recall are competitive, closely aligning with those of Random Forest and XGBoost, showcasing its effectiveness in correctly identifying non-purchaser sessions.

Confusion Matrix Insights:

Figure 21 displays the confusion matrix, showing:

- **Correct Identifications:** TabNet correctly identified 1,877 out of 2,084 non-purchaser sessions and 258 out of 381 purchaser sessions.
- **Misclassifications:** There were 207 non-purchaser sessions classified as purchases and 123 purchaser sessions classified as non-purchases.

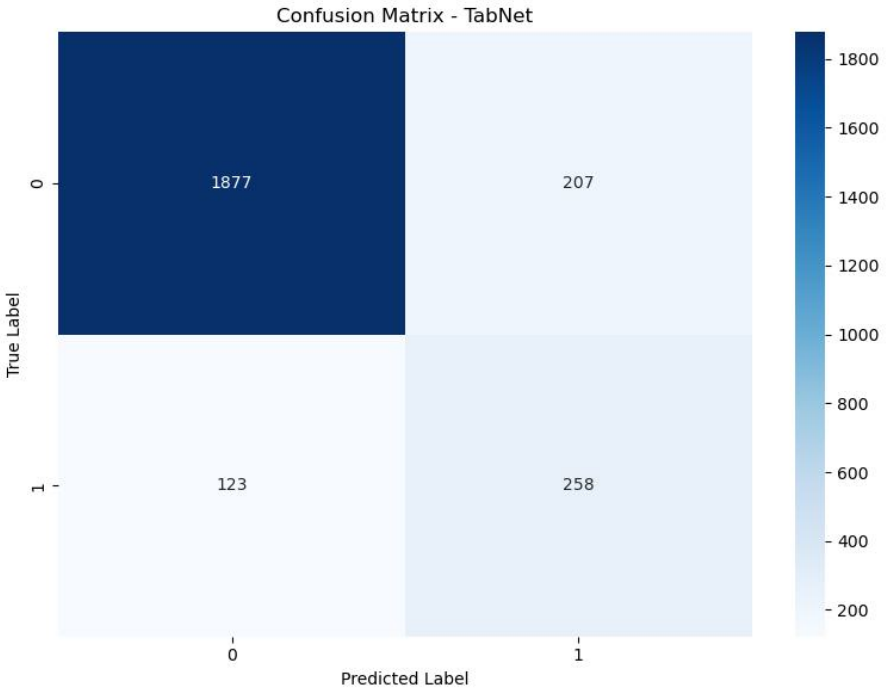

Figure 21: TabNet Confusion Matrix