National College of Ireland

Enhancing Purchase Predictions with Machine Learning: Customer Propensity Modelling through Predictive Analytics

MSc Research Project
Data Analytics

VINITH KUMAR GUDIBANDA PRASANNAKUMAR

Student ID: 22131248

School of Computing

National College of Ireland

Supervisor: Furqan Rustam

National College of

Ireland MSc Project

Submission Sheet

School of Computing

Student Name: ..... VINITH KUMAR GUDIBANDA PRASANNAKUMAR .....

Student ID: .....x22131248.....

Programme: .....Data Analytics..... **Year:** .....2023.....

Module: .....Research Project.....

Supervisor: .....Furqan Rustam.....

Submission Due Date: .....31/01/2024.....

Project Title: Enhancing Purchase Predictions with Machine Learning: Customer Propensity Modelling through Predictive Analytics

Word Count: .....8189..... **Page Count**.....24.....

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: ..... **VINITH KUMAR GUDIBANDA PRASANNAKUMAR**.....

Date: .....31/01/2024.....

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Contents

Abstract 1

1. Introduction 1

1.1 Research Question 2

2. Related Work 2

2.1. Evolution of Data Mining in Consumer Behavior Prediction 3

2.2. Machine Learning and Advanced Analytics in Understanding Consumer Behavior 3

2.3. Deep Learning and Neural Networks in Predicting Consumer Choices 3

2.4. Specific Applications and Contextual Studies in Consumer Behavior 3

2.5. Trust and Behavioral Aspects in Consumer Engagement 4

2.6. Novel Approaches and Emerging Trends in Predictive Modeling 4

3. Proposed Methodology 7

3.1 Data Cleaning 7

3.2 Statistical Details and Exploratory Data Analysis (EDA) 7

3.3 Data Transformation 7

3.4 Feature Selection Using OLS and RFECV 8

3.5 Data Normalization Using Min-Max Scaler 8

3.6 Data Balancing Using SMOTE and CTGAN 8

3.7 Model Building 9

3.8 Evaluation 9

4. Design Specification 9

5. Implementation 11

5.1 Exploratory Data Analysis 12

5.4 Data Balancing 14

5.3 Models Implemented 16

6. Evaluation 17

6.1 Basic Model without Balancing or Feature Selection 17

6.2 SMOTE for Balancing and OLS for Feature Selection 18

6.3 CT-GAN for Balancing and OLS for Feature Selection 18

6.4 CT-GAN for Balancing and RFECV for Feature Selection 18

6.5 Discussions 19

7. Conclusion and Future work 20

References 20

Abstract

In this research, the aim was to predict online customer purchasing behavior accurately. This issue is essential to businesses as they seek to optimize marketing efforts and improve customer engagement by determining the likelihood of purchase transactions. However, traditional analytics has limitations when dealing with imbalanced datasets and complex patterns in customer behavior, which necessitated this research. One of the limitations identified was that research indicates that the typical conversion rate in e-commerce is only about 2-3%. To address this issue, a dataset was sourced from the UCI repository, and a novel approach of using multiple algorithms such as Logistic Regression, Decision Tree classifier, GBM, and RNN was adopted. The model's ability to identify confirmed transactions (marked by the 'Revenue' class) was significantly improved by incorporating advanced data pre-processing techniques such as SMOTE, OLS, and RFECV. The technical approach included a detailed analysis of the dataset, pre-processing methods to improve data quality, and the use of GBM to model purchasing behavior. To ensure efficient performance across different data scenarios, GBM parameters were optimized using a rigorous cross-validation process, which is the novelty of this research. The results showed that when combined with our pre-processing strategy, the GBM outperforms standard prediction models. The accuracy of identifying 'Revenue' class transactions has significantly improved, providing businesses with actionable insights into customer purchasing patterns. The research identifies key factors that influence customer decisions, allowing for the development of more targeted and effective marketing strategies.

1. Introduction

In the current highly competitive business environment, having a comprehensive understanding of consumer behavior is essential for any organization striving for success. The ability to predict customer purchasing trends and identify potential new clients has become a crucial skill for companies looking to increase revenue, build long-lasting customer relationships, or minimize marketing attrition. With the evolution and complexity of data, traditional market analysis techniques are no longer sufficient. Online businesses face the challenge of converting a significant portion of their website traffic into paying customers. According to Sang and Wu (2022), many visitors do not return to websites and only a small fraction complete transaction. To address this challenge, consumer propensity modelling has emerged as a key strategy.

Research studies indicate that the average online conversion rate for e-commerce businesses ranges between 2-3%, which suggests that there is a significant opportunity to improve customer engagement and predict purchases more accurately. According to Suryadi's findings (2020), incorporating textual features from online customer reviews can help predict repurchase intentions and highlight the importance of qualitative data in understanding consumer behavior. Moreover, Xu et al. (2022) have shown that hybrid models are effective in predicting telecom customer intentions, which reflects the diversity and complexity of user behavior in different industries. The research highlights on growing need for advanced analytical models that can handle large and complex datasets and make accurate predictions.

Businesses are constantly trying to connect with their target customers as quickly and efficiently as possible. To do this, they need effective solutions that can provide targeted answers. They require models that can adapt to changing trends, handle a wide range of data collection techniques, provide real-time forecasts, and make accurate predictions. One such method is propensity modelling, which uses machine learning and predictive analytics to

anticipate customers' purchasing behavior. Through my work, I aim to provide businesses with an innovative technique that can not only predict new customers but also understand the factors that influence their purchasing decisions by creating a robust Customer Propensity Model.

My research provides a comprehensive approach to predicting and understanding consumer purchasing behavior using machine learning techniques. It highlights the significance of predictive modelling in the contemporary digital marketplace. Anticipating consumer behavior is crucial for business success, and this study underscores the importance of this approach.

1.1 Research Question

- How can machine learning-enhance customer propensity models be most effectively utilized to refine online marketing strategies, through insights into consumer behavior, and improve customer targeting?

This study employs four distinct models, with an emphasis on performing data balance using CT-GAN, which is considered the most advanced approach, to attain superior metrics. The enhancements demonstrated in this study are related to recent academic publications that have employed various methods for balancing data. An important advancement of this work is the implementation of CT-GAN, which has recently demonstrated its superior efficacy, distinguishing it from prior studies in the sector. The adoption of CT-GAN has resulted in improved performance across the models.

In order to conduct a comprehensive comparison with the most advanced approaches, I created four distinct programmes, each of which incorporates a distinct model. The codes were developed to showcase different methodologies, facilitating a thorough comparison between the outcomes of prior research and my implementation utilising CT-GAN. The state-of-the-art results were observed when CT-GAN was applied, the outstanding performance was seen in the Gradient Boosting Classifier rather than applying any other balancing techniques. So, these specific results which provide my state-of-the-art were carried out in experiment 3. My novelty for my research was applying the CT-GAN technique for data balancing which provided me a promising result for all the models.

In this research work, the introduction is followed by a literature review in Chapter 2 that establishes the academic context of the study. Chapter 3 details the methodology used, which includes data pre-processing and model development strategies. Chapters 4 and 5 focus on the research design and practical implementation, respectively. They outline the theoretical framework and the execution of the predictive models. Chapter 6 presents a critical analysis of the experiments conducted and their results, evaluating the effectiveness of the various models employed. Finally, Chapter 7 summarizes the findings of the research, discusses the limitations encountered, and proposes avenues for future work.

2. Related Work

This comprehensive review covers research papers on predictive models and methods for understanding and predicting consumer behavior, especially in the context of online shopping. The exploration of various datasets and methodologies provides a broad perspective on the current state of research in consumer behavior and intention prediction.

2.1. Evolution of Data Mining in Consumer Behavior Prediction

The research conducted by Alghanam et al. (2023) demonstrates the application of data mining methods, particularly K-means clustering and different classifiers, for forecasting client buying patterns in the field of e-commerce. The text emphasises the utilisation of the Apriori algorithm for item recommendation, specifically highlighting its notable accuracy and detailed methodology. Kareena and Kapoor (2019) provide a comprehensive examination of diverse data mining methodologies for forecasting consumer behaviour. They highlight the efficacy of Support Vector Machines (SVM) and underscore the significance of prior purchase records and customer reviews in the prediction process. In their study, Suryadi (2020) utilises machine learning techniques to forecast customer repurchase intention based on online reviews. The research specifically emphasises text analysis and feature selection, while also addressing the difficulties associated with contextually understanding textual data.

2.2. Machine Learning and Advanced Analytics in Understanding Consumer Behavior

In their study, Zhao and Keikhosrokiani (2023) use RFM analysis with machine learning algorithms such as XGBoost and Random Forest to forecast sales. They emphasise the shift from conventional sales models to B2C models in the context of the pandemic. The research conducted by Ahsain and Kbir (2023) examines different machine learning models, such as LightGBM and Random Forest, in order to forecast purchasing intentions. The study emphasises the effectiveness of these models in the field of e-commerce, as well as their limits when it comes to the specificity of the dataset. Valecha et al. (2018) and Zhang (2021) employ Random Forest and logistic regression models to forecast consumer behaviour and buy propensity. They highlight the models' precision and discuss the difficulties associated with managing intricate datasets.

2.3. Deep Learning and Neural Networks in Predicting Consumer Choices

Nisha and Singh (2023) explore the application of deep learning models, namely DNNs, in analysing customer behaviour during online purchases. They evaluate the potential of deep learning to enhance accuracy in this context. Zheng (2020) conducted a study that specifically examines the utilisation of Artificial Neural Networks (ANNs) in client Relationship Management (CRM) systems to forecast client buying patterns. The study places particular emphasis on the difficulties associated with implementing and interpreting ANN models. In their study, Ling, Zhang, and Chen (2019) utilise a deep learning architecture that incorporates FC-LSTM networks. Their objective is to forecast customer purchase intent in online multi-channel promotions. The study specifically focuses on the challenge of modelling nonlinear correlations in customer behaviour.

2.4. Specific Applications and Contextual Studies in Consumer Behavior

Matuszelan´ski and Kopczewska (2023) conducted a study on customer turnover in the Brazilian e-commerce industry. They utilised machine learning techniques and geographical analysis to investigate this phenomenon, highlighting the study's special focus on the Brazilian environment. The research conducted by Gupta et al. (2022) investigates the prediction of customer churn in the telecommunications industry through the utilisation of diverse data

mining methods. The study emphasises the distinct characteristics of the dataset employed and the difficulties encountered while using models in varying scenarios. The study conducted by Ravi, Sangaralingam, and Datta (2023) presents an innovative method for forecasting consumer brand preferences by using spatio-temporal data. The study also addresses the constraints associated with the cold start problem and data incompleteness.

2.5. Trust and Behavioral Aspects in Consumer Engagement

In their study, Scott B. Friend et al. (2018) examine the level of confidence that customers have in salespeople involved in B2B partnerships. They employ multilevel modelling to evaluate how salesperson attributes influence consumer trust. The research conducted by Cheung and To (2017) investigates the impact of trust on the attitudes of mobile users towards in-app adverts. The study use structural equation modelling in conjunction with the Theory of Planned Behaviour.

2.6. Novel Approaches and Emerging Trends in Predictive Modeling

In their study, Gumber et al. (2021) utilise XGBoost to forecast customer behaviour based on clickstream data. They specifically highlight the use of ensemble methods and the difficulties associated with handling intricate data. In Chen's (2022) study, a refined Bayesian algorithm is presented for forecasting customer purchase intentions. The study specifically emphasises its effectiveness in managing large datasets and addresses the constraints associated with the Bayesian approach. Xu et al. (2022) utilises a combined technique of selecting relevant features and merging multiple models to improve the accuracy of predicting client intentions in the telecom sector. Surendro (2019) presents a comprehensive examination of predictive analytics in forecasting customer behaviour, encompassing both theoretical foundations and the imperative for practical implementation specifics.

The reviewed studies show the increasing significance of machine learning and predictive models in understanding online consumer behavior. However, issues such as model complexity and moral dilemmas still exist. The results of the study provide a crucial foundation for further research and development as e-commerce continues to evolve.

Sl.No	Author/Year	Approach	Type/Method	Accuracy	Limitations
1	Zhao and Keikhosrokiani (2023)	Sales Prediction and Product Recommendation	RFM, XGBoost, Random Forest, Apriori	-	Potential Overfitting, Specific Context
2	Alghanam et al. (2023)	Customer Purchase Behavior Prediction	K-means, C4.5, J48, CS-MC4, MLR, Apriori	95.2%	Dataset Specificity, Model Complexity
3	Matuszelan´ski and Kopczewska (2023)	Customer Churn Prediction	XGBoost, Logistic Regression	-	Dataset Specificity, Model Complexity

4	Ahsain and Kbir (2023)	Client's Purchasing Intention Prediction	LightGBM, GBC, RFC	-	Dataset Specificity, Potential Overfitting
5	Kareena and Kapoor (2019)	Consumer Behavior Prediction Review	Various Data Mining Techniques	-	Lacks Practical Implementation Details
6	Nisha and Singh (2023)	Customer Behavior Prediction	Deep Neural Networks (DNNs)	-	Specific Dataset, Computational Complexity
7	Zheng (2020)	Customer Purchase Behavior Prediction	Artificial Neural Networks (ANN)	-	Specific to CRM Systems, Data Extensiveness
8	Ling, Zhang, and Chen (2019)	Customer Purchase Intent Prediction	FC-LSTM Networks	-	Complex Model, Extensive Data Requirement
9	Gupta et al. (2022)	Customer Churn Prediction	Random Forest, Logistic Regression, J48, Others	93.55%	Dataset Specificity, Model Complexity
10	Ravi, Sangaralingam, and Datta (2023)	Consumer Brand Preferences Prediction	ALS-WR based Matrix Factorization	-	Cold Start Problem, Incomplete Data
11	Gumber, Jain, and Amutha (202)	Customer Behavior Prediction	XGBoost	85.9%	Complex Clickstream Processing, Dataset Specificity
12	Suryadi (2020)	Repurchase Intention Prediction	Text Analysis, Fisher Score	-	Word Independence Assumption, Language Specific
13	Sang and Wu (2022)	Online Shoppers Purchasing Intention Prediction	Random Forest, SVM	-	Dataset Specificity, Potential Overfitting

14	Valecha et al. (2018)	Consumer Behaviour Prediction	Random Forest	-	Potential Overfitting, Dataset Specificity
15	Zhang (2021)	Customer Propensity Prediction	Logistic Regression, Random Forest	-	Dataset Specificity, Model Complexity
16	Asniar and Surendro (2019)	Predictive Analytics Review	Behavior Informatics and Analytics	-	Theoretical Nature, Generalizability Limitation
17	Scott B. Friend et al. (2018)	Trust in Salespeople Analysis	Multilevel Modeling	-	Industry Specific, Generalizability
18	Le Chen (2022)	Consumer Purchase Intention Prediction	Bayesian Network Classification	-	Hadoop Platform Dependency, Independence Assumption
19	Cheung and To (2017)	Mobile Users' Attitudes Toward Advertisements	Structural Equation Modeling	-	Demographic Specificity, Cultural Context
20	Xu, Sun, and Guo (2022)	User Intention Prediction	Hybrid Feature Selection, Stacking Ensemble	-	Model Complexity, Dataset Specificity

Table 1: Summary of the Related Work

Table 1 summarizes the existing literature which has revealed numerous important challenges and gaps that are crucial for guiding future research contributions and breakthroughs. A recurring constraint found in numerous research, including those conducted by Zhao and Keikhosrokiani (2023), Ahsain and Kbir (2023), and others, is the susceptibility of intricate models such as XGBoost, Random Forest, and deep learning algorithms to overfitting. The presence of overfitting gives rise to issues over the applicability of these models to diverse contexts and datasets. Moreover, the limited scope of the datasets employed, as emphasised in the research conducted by Alghanam et al. (2023) and Gupta et al. (2022), presents difficulties in generalising the results to wider contexts. Difficulties in implementing and understanding models, particularly in research that utilise sophisticated algorithms such as FC-LSTM networks and ANN, as demonstrated in Ling, Zhang, and Chen (2019) and Zheng (2020), have been identified as major obstacles to practical application. Moreover, the studies conducted by Kareena and Kapoor (2019) and Asniar and Surendro (2019) do not provide specific information on how their findings might be practically used, highlighting a disconnect between theoretical research and its practical implementation in real-world scenarios. The necessity for comprehensive research that takes into consideration various consumer demographics, as

demonstrated in the study conducted by Cheung and To (2017), implies the potential for conducting more internationally representative investigations. The stated limits and gaps emphasise important areas that need development and innovation. This creates an opportunity for research contributions that can solve these problems by using more reliable, scalable, and widely applicable approaches in predicting consumer behaviour.

3. Proposed Methodology

In this section, I will describe the comprehensive pre-processing and preparation steps taken to refine the dataset obtained from the UCI Machine Learning Repository. The dataset consists of 12,330 rows and 18 columns. The procedure is carefully designed to ensure that the representation is balanced, statistical insights are gained, efficient encoding is used, feature relevance is observed, normalization is applied, and high-quality data is obtained. Every stage is essential in providing a strong and optimally prepared dataset for later modelling and analysis. Each step in this process from Fig.1 plays a critical role in creating a robust and optimized outcome. This preparation lays a solid foundation for subsequent modelling and analysis phases, ensuring that the data is in its best possible form to yield accurate and meaningful insights.

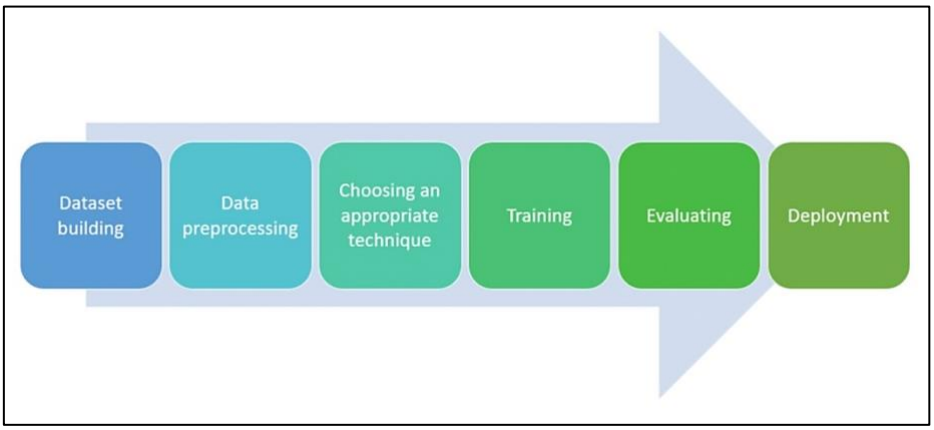

Fig. 1: Proposed research flow

3.1 Data Cleaning

Data cleaning is an essential step in the analysis of data that involves identifying and correcting errors or any missing data. In my case, I discovered that there were no null values in the dataset. Therefore, there was no need to use imputation strategies to handle missing data. Since my dataset was free of null values, I proceeded with the analysis without having to clean the data. This allowed me to concentrate on other data preparation and analysis tasks.

3.2 Statistical Details and Exploratory Data Analysis (EDA)

EDA techniques and statistical methods are utilized to gain a comprehensive understanding of the dataset. Statistical summaries, such as mean, standard deviation, minimum, maximum, and quartile values, are used to comprehend the distribution, dispersion, and central tendency of numerical variables. To ensure that the dataset is complete, the "info" function is also used to examine data types and non-null counts in detail. EDA techniques, including the visualization of class variable distributions, are used to identify class imbalance problems and to obtain a general overview of the dataset's properties.

3.3 Data Transformation

I transformed a dataset that contained both categorical and Boolean variables so that they could be effectively integrated into my models. To do this, I used a data transformation strategy called label encoding for both data types. This involved assigning distinct integers to each category,

typically based on their alphabetical order. By doing so, I converted categorical variables into numerical values that could be processed by many machine learning algorithms. Similarly, for Boolean variables in the dataset, label encoding proved to be a practical approach. I transformed True and False values into a binary format (1 and 0) so that the Boolean values were suitably formatted for algorithmic processing. This process helped to standardize the dataset, making it easier to analyse and train models.

When dealing with non-numeric data types like categorical and Boolean variables, a critical step in data pre-processing is to transform them into a format that is suitable for model training. This process usually involves techniques such as label encoding for categorical variables and appropriate conversions for Boolean values.

Label encoding is a technique used to convert categorical variables into a numerical format. It involves assigning a unique integer to each category. For Boolean variables, a simple transformation is usually applied, which involves converting 'True' and 'False' values into binary numeric formats, such as 1 and 0. These transformations of categorical and Boolean data types are important in preparing the dataset for analysis and training of machine learning models. They help in making the process more efficient.

3.4 Feature Selection Using OLS and RFECV

The purpose of feature selection is to improve model performance by reducing overfitting and increasing interpretability. Two popular techniques used for feature selection include OLS and RFECV:

1. *The Ordinary Least Squares (OLS)* is a statistical technique that is useful for feature selection. The main objective is to identify features that significantly contribute to explaining the variance in the target variable by fitting a model to the data.

2. *Recursive Feature Elimination with Cross-Validation, or RFECV,* is a technique used for selecting the most important features in a dataset. It does this by iteratively removing the least significant features while considering cross-validated performance. The process involves training the model, evaluating its performance, and recursively eliminating the least important features until the ideal subset is found.

Using these approaches ensures a thorough assessment of feature importance, as well as the robustness and interpretability of the model.

3.5 Data Normalization Using Min-Max Scaler

In my dataset, I utilized Min-Max Scaling to normalize the data. This technique is particularly effective for adjusting numerical features to a specific range, usually between 0 and 1. The primary advantage of using Min-Max Scaling in my dataset was to avoid any individual feature from dominating others due to variations in scale. By applying this scaling method, I ensured that each feature in the dataset contributed equally to the model training process.

3.6 Data Balancing Using SMOTE and CTGAN

Two techniques, Synthetic Minority Over-sampling Technique (SMOTE) and Conditional Generative Adversarial Network (CTGAN), have been utilized to handle imbalanced datasets with uneven class distribution. SMOTE creates artificial samples for the minority class, which successfully balances the distribution of classes. On the other hand, CTGAN generates synthetic data that closely mimics the statistical characteristics of the original dataset. Using both of these techniques together, biases caused by unequal class representation can be reduced.

3.7 Model Building

After completing a comprehensive pre-processing phase that included data transformation, feature selection, and addressing class imbalance, I divided the dataset into two distinct sets. To be specific, 70% of the data was provided for training purposes, while the remaining 30% was reserved for testing. This separation was critical for assessing the models' performance on unseen data.

My analysis consisted of four experiments using various models and techniques for data balancing and feature selection:

1. Basic Model without Balancing or Feature Selection
1. SMOTE for Balancing and OLS for Feature Selection
2. CT-GAN for Balancing and OLS for Feature Selection
3. CT-GAN for Balancing and RFECV for Feature Selection

3.8 Evaluation

After finishing the initial processing steps, which included transforming the data, selecting the appropriate features, and addressing any class imbalances, we evaluated the four experiments in detail using key performance metrics. These metrics are essential in determining the effectiveness of each model across different experimental setups:

1. Accuracy
2. Precision
3. Recall (Sensitivity)
4. F1 Score

4. Design Specification

The examined studies in consumer behavior prediction impact the choice of four models: Decision Tree, Gradient Boosting Machine, RNN, and Logistic Regression.

1. Logistic Regression: Zhang's (2021) research demonstrated the effectiveness of logistic regression in forecasting buyer propensity based on online browsing data. This study was inspired by Zhang's work and utilized regression analysis as its method. Logistic regression is a reliable model for capturing the complexities of consumer behavior, particularly in online environments. When dealing with regression or classification problems, a cost function is established to obtain the optimal model parameters by optimizing iterative methods, followed by verifying and testing the solution model's advantages and disadvantages. To begin, a suitable prediction function, typically expressed as the ‘h’ function, needs to be identified. This function is used to forecast the input data's judgment result, and it must be discovered as it is the classification function. This procedure necessitates knowledge of the prediction function's approximate form, such as a linear or nonlinear function. As the output of this function must represent two categories, the logistic function is used with the Sigmoid function. The resulting function has the following form:

$$g(z) = \frac{1}{1+e^{-z}} \tag{1}$$

We can create a cost function, also known as a loss function, to represent the difference between the predicted output (h) and the training data category (y). This difference can take the form of (h-y) or other variations:

$$Cost(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & if \quad y = 1 \\ -\log(1 - h_\theta(x)) & if \quad y = 0 \end{cases} \tag{2}$$

2. Decision Tree: Decision Trees have been incorporated into the design specification, taking inspiration from research such as Valecha et al. (2018), as they have shown potential in explaining the impact of various factors on consumer behaviour. This model is particularly good at identifying complex patterns that influence purchasing decisions. Here are two functions that can help determine the quality of the data.

Information gain can measure the impact of a feature on the classification outcome. Information entropy represents uncertainty. When the distribution is uniform, the uncertainty is at its highest, and hence, the entropy is also at its highest. When selecting a feature to classify a data set, the information entropy after classification will be smaller, and the reduced part is expressed as information gain.

The Gini index is a measuring parameter for data impurity, calculated using a specific formula:

$$Gini(D) = 1 - \sum_{i}^{c} p_i^2 \tag{1}$$

It is calculated using the proportion of samples in each category of the dataset. The Gini Index is calculated using the formula (c) where (c) represents the number of categories in the dataset, and (pi) represents the proportion of the number of samples (i) in a category to all samples. This formula suggests that a higher degree of data mixing in the dataset will result in a higher Gini Index.

For feature selection, the smallest split Gini Index is the one that needs to be selected. Additionally, the Gini exponent gain value can also be used to choose features of the decision tree. The formula for calculating the Gini exponent gain value is as follows:

$$\Delta Gini(A) = Gini(D) - Gini_A(D) \tag{2}$$

3. GBM (Gradient Boosting Machine): The study conducted by Ling, Zhang, and Chen (2019), which focused on predicting customer purchase intent in the context of online multi-channel promotions, inspired the inclusion of GBM in this current study. This study aims to understand how consumers react in dynamic promotional environments, and GBM is known to be effective in handling various features and capturing complex correlations.

4. RNN (Recurrent Neural Network): Xin (Robert) Luo, Ying Hua, Jose Benitez, Dongyi Wang, and others (2023) showcased the creative use of deep learning frameworks, which inspired the decision to select RNN. This was due to its ability to simulate client interactions with promotion channels. As RNNs shown in Fig 2 & Fig 3 are capable of capturing sequential dependencies in data, they are a valuable tool for analysing how customer behaviour changes over time.

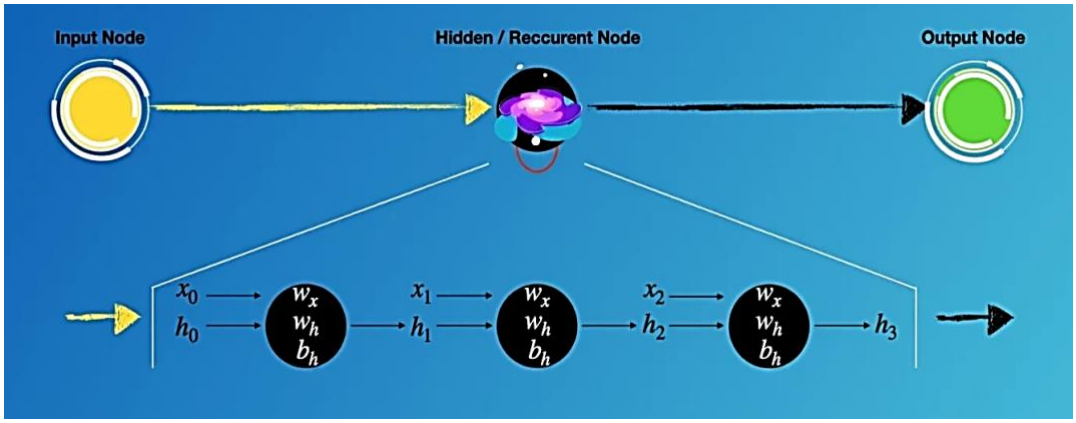

Fig 2: Layers in RNN Architecture

simple_rnn_input	input:	[(None, 17, 1)]
InputLayer	output:	[(None, 17, 1)]
simple_rnn	input:	(None, 17, 1)
SimpleRNN	output:	(None, 17, 64)
dropout	input:	(None, 17, 64)
Dropout	output:	(None, 17, 64)
simple_rnn_1	input:	(None, 17, 64)
SimpleRNN	output:	(None, 17, 64)
dropout_1	input:	(None, 17, 64)
Dropout	output:	(None, 17, 64)
simple_rnn_2	input:	(None, 17, 64)
SimpleRNN	output:	(None, 64)
dropout_2	input:	(None, 64)
Dropout	output:	(None, 64)
dense	input:	(None, 64)
Dense	output:	(None, 1)

Fig 3: Summary of Model Architecture

5. Implementation

Initially several preparatory steps were undertaken to set the groundwork. Using Google Colab's directory system, I sourced and loaded the dataset, ensuring correct import for analysis. After importing the data for analysis, I checked for missing or null values. This step is essential to identify any gaps in the data that could skew the analysis.

Subsequently, I proceeded to examining the information contained in each column of the dataset. This involved a thorough review of the data types, range of values, and any unique

attributes of the columns. Understanding the nature of each column was crucial for effective data pre-processing and for making informed decisions during feature selection.

Lastly, I explored the fundamental statistics of the columns, which provided me with a deeper understanding of the distribution, tendencies, and potential anomalies in the data. This statistical overview was a key component in guiding the subsequent steps in the EDA process, ensuring a data-driven approach to model development and analysis.

5.1 Exploratory Data Analysis

As I conduct my research, one crucial aspect I focus on is EDA, which involves conducting a preliminary investigation of the dataset, which helps to identify patterns, anomalies, and relationships among variables. This information is critical to make informed decisions during the later stages of modelling. In this chapter, I explore several topics such as the interaction between products, the impact of special days, the distribution of visitor types, monthly revenue trends, and the role of weekends in generating income. By acquiring this knowledge, I can ensure that our predictive models have a strong and reliable foundation, which results in greater accuracy and dependability.

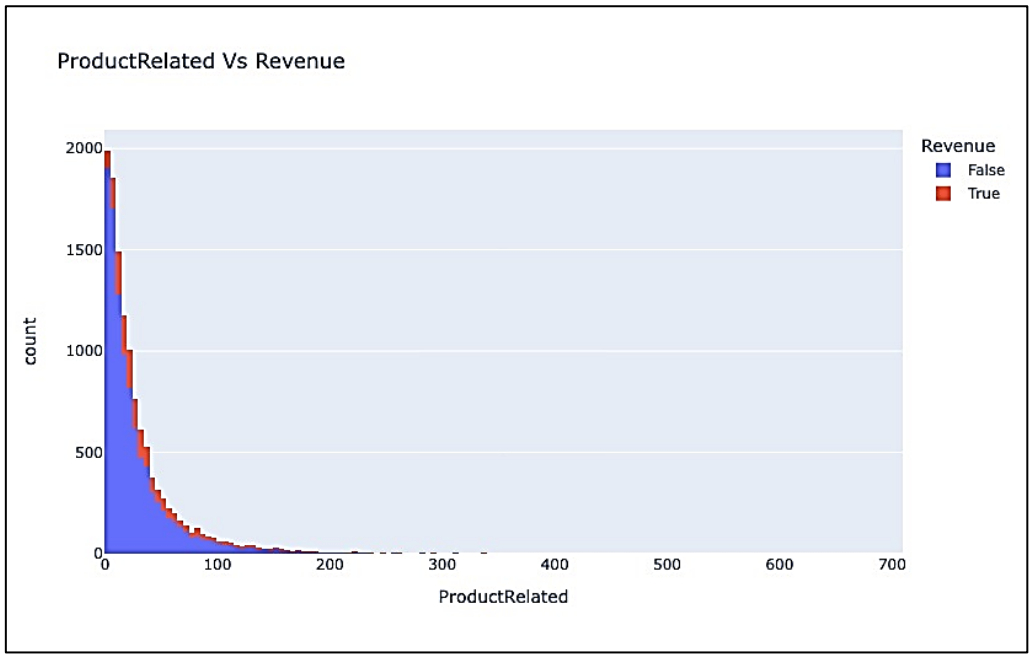

Fig. 4: Product-Related Engagement and Revenue Conversion EDA

Fig. 4 shows how user sessions are distributed based on the number of product-related pages visited and whether the session resulted in revenue. It provides important engagement metrics that can predict purchasing behavior.

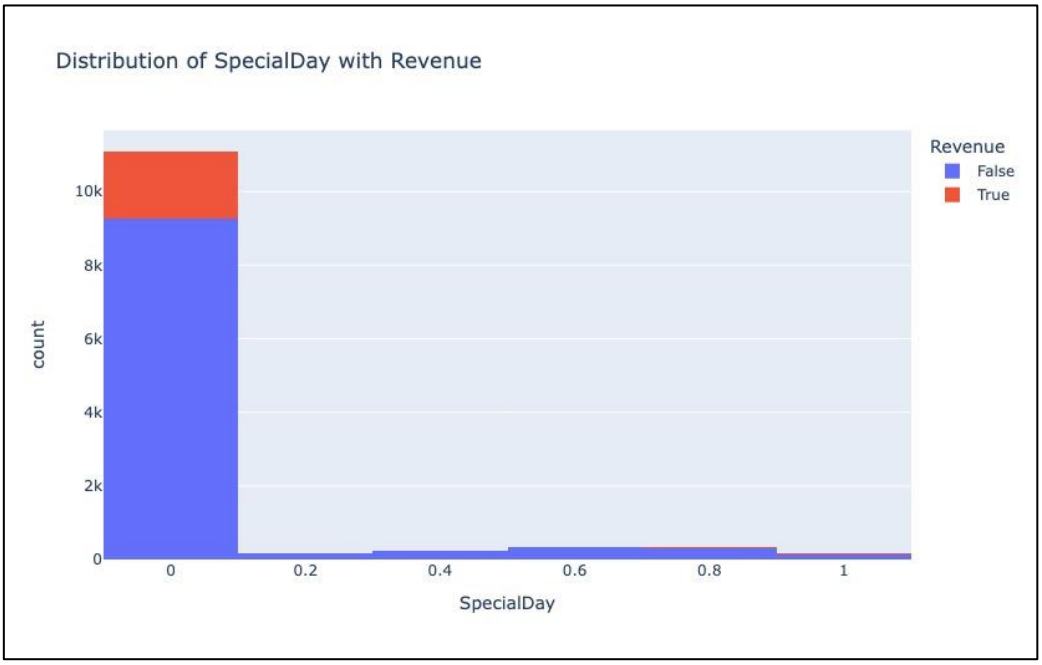

Fig. 5: Special Days Impact on Revenue EDA

Fig. 5 illustrates how special occasions impact revenue, indicating an increased likelihood of sales during these periods.

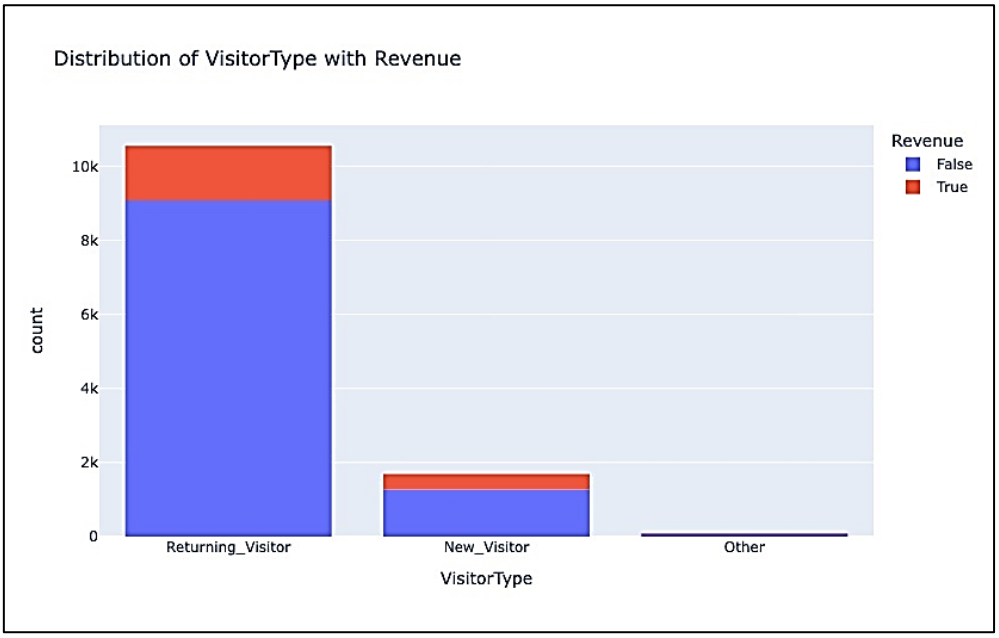

Fig. 6: Visitor Type Revenue Distribution EDA

Fig. 6 illustrates the revenue distribution across different types of visitors, which can help tailor user experience based on visitor familiarity with the site.

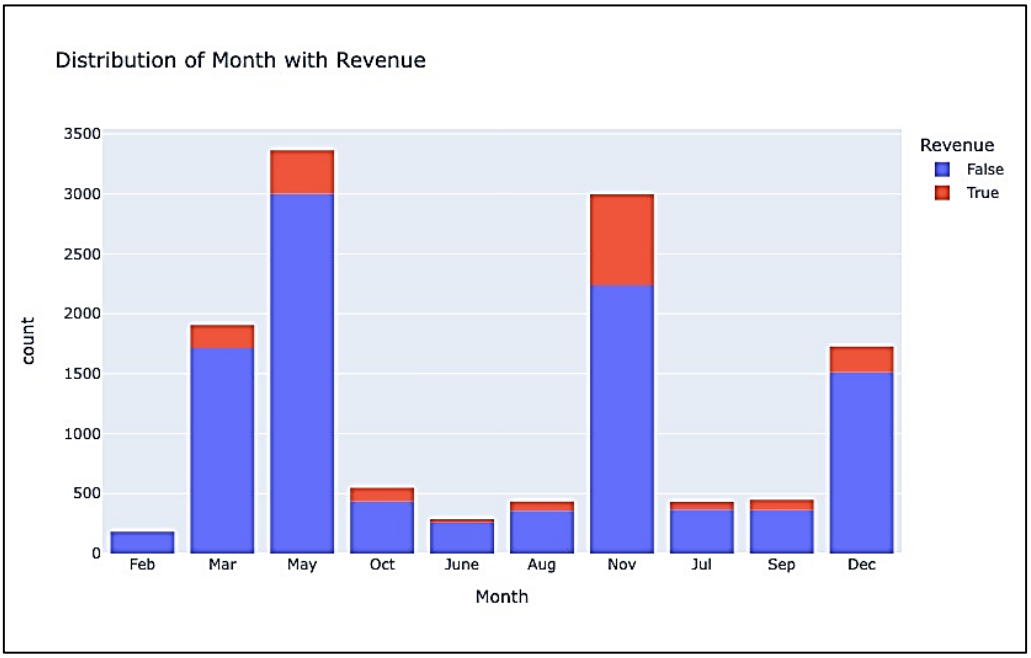

Fig. 7: Monthly Revenue Trends EDA

Fig. 7 examines the monthly revenue patterns to determine how seasonal fluctuations affect consumer buying behavior.

Fig. 8: Weekend Versus Weekday Revenue EDA

Fig. 8 provides insights into consumer behavior by comparing revenue generation on weekends versus weekdays, which can inform marketing and sales tactics.

5.4 Data Balancing

I came across the issue of class imbalance in my dataset, which is a common challenge in machine learning. It can lead to biased models that favour the majority class. To tackle this issue, I used techniques like the Synthetic Minority Over-sampling Technique (SMOTE) and Conditional Tabular Generative Adversarial Network (CT-GAN). These methods are particularly effective in balancing class representation, thereby improving the quality and diversity of the dataset. This approach was crucial in developing models that can generalize better and produce more accurate predictions while ensuring fairness.

1. SMOTE:

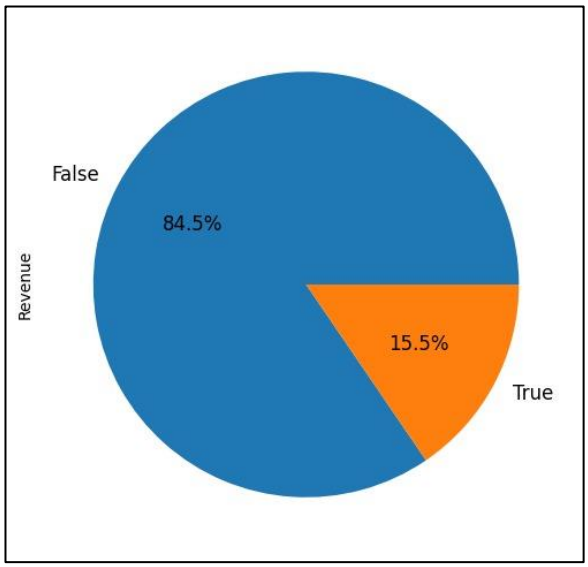

Fig. 9: Data Imbalance without SMOTE

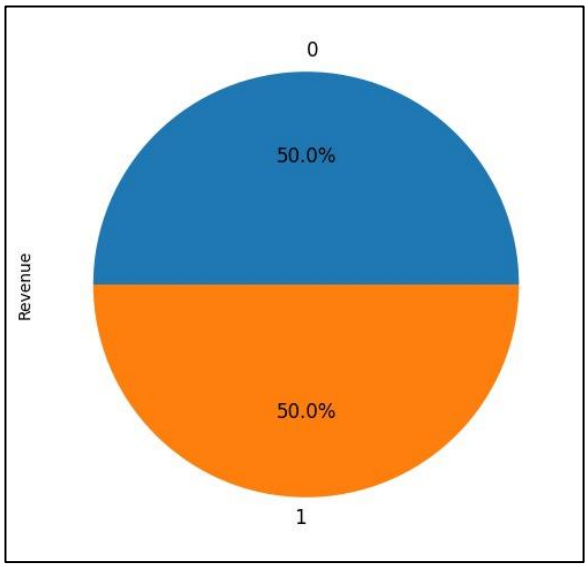

Fig. 10: Data Balanced after using SMOTE

Prior to using SMOTE, there was a significant class imbalance in the dataset, which is illustrated in Figure 9. The 'False' class, representing the majority, accounted for 84.5% of the data, while the 'True' class, representing the minority, made up only 15.5%. However, after applying SMOTE, the class distribution was equalized, with both classes '0' (previously 'False') and '1' (previously 'True') now comprising 50% of the data each, as shown in Figure 10. This balanced class distribution is crucial for enhancing the performance of classification algorithms.

2. CT-GAN:

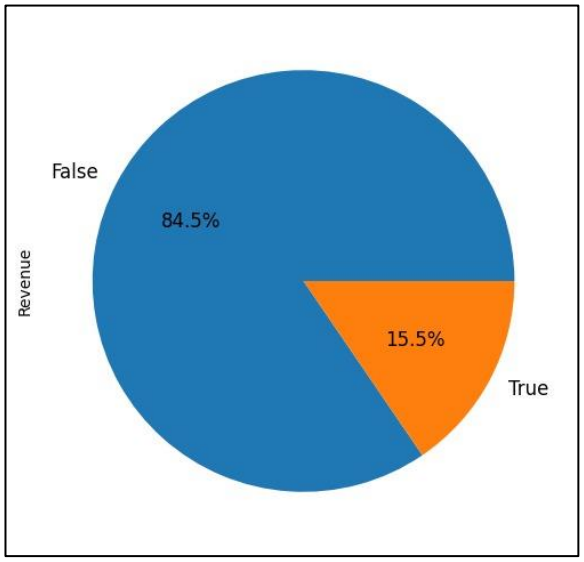

Fig. 11: Data Imbalance without CT-GAN

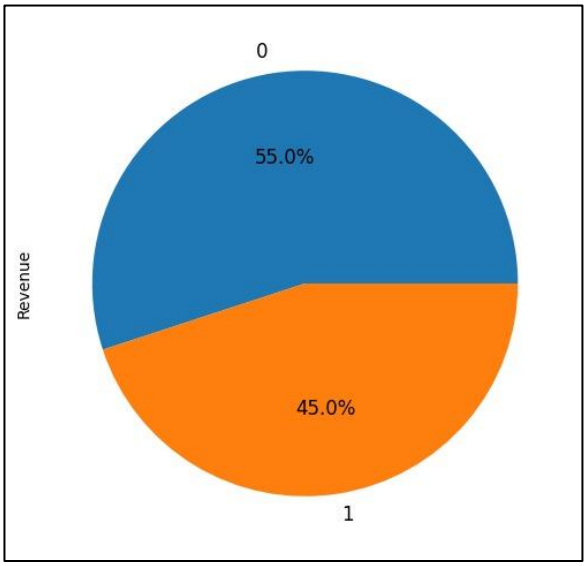

Fig. 12: Data Balanced after using CT-GAN

After applying CT-GAN, the distribution in Fig. 12 shows a shift towards 55% for "False" and 45% for "True". This is different from the situation after applying SMOTE, which resulted in a clean 50-50 split. This minor imbalance can better represent the fluctuations in real-world data. As a result, models can be trained on a balanced dataset that is still challenging and representative of real-world situations.

5.3 Models Implemented

After performing a comprehensive pre-processing phase on the dataset, which included transforming categorical and Boolean data types into a numerical format suitable for model training, feature selection, and addressing class imbalance, the dataset was partitioned into two segments. 70% of the data was allocated for training, and the remaining 30% was reserved for testing. This split was crucial in evaluating model performance on unseen data.

After the data split, I utilized Min-Max Scaling to normalize the numerical features, adjusting them to fall within a specific range, typically [0, 1]. Normalization was essential to prevent any one feature from dominating the model due to scale discrepancies. By ensuring each feature had an equal opportunity to influence model training, the model's accuracy was improved.

Following the data pre-processing, I carried out model implementation in various experiments that involved different forms of class balancing, feature selection, and algorithms. These experiments will be covered in more detail.

5.1.1 Experiment 1 - Basic Model without Balancing or Feature Selection

For the first experiment, a basic model was used without any feature selection or data balancing to establish a baseline. The hyperparameters for the algorithms were set as follows:

Decision Tree Classifier - To find the optimal max_depth parameter, a series of evaluations were conducted. The model was trained and validated on the dataset at various depth levels, incrementally increasing the depth of the tree. Performance metrics such as accuracy, precision, recall, and F1 score were observed to identify the max_depth that yielded the best balance between model complexity and predictive power. This approach helped to mitigate overfitting and ensured sufficient model depth to capture the underlying patterns in the data. The max_depth that corresponded to the highest cross-validated score was chosen as the appropriate depth for the Decision Tree Classifier in our experiments. The calculated max_depth was 4.

GBM- I have set the n_estimators parameter to 300 in our model. This allows the model to build 300 sequential trees, which we found to be a robust number for the dataset, without being overly time-consuming. Additionally, I have set the learning_rate to 0.05, which determines the contribution of each tree to the outcome and helps in preventing overfitting by allowing the model to learn gradually. To ensure consistency across multiple runs, a random_state of 42 has been used. Finally, I have set the max_features parameter to 6, which limits the number of features to consider when looking for the best split. This makes the training process faster and reduces overfitting.

RNN- During the configuration of the Recurrent Neural Network (RNN), I decided to train the model for 25 epochs. This number of epochs was chosen to ensure that the network has enough iterations to learn from the data without overfitting. In each epoch, the model processes the entire dataset once and adjusts the weights to minimize the loss function.

The batch size of 16 was chosen to balance computational efficiency with the benefits of stochastic gradient descent optimization. A smaller batch size ensures regular updates to the model, striking a good balance between the speed of each epoch and the granularity of weight updates. This can lead to better generalization of the model.

Every model selected (Decision Tree, RNN, GBM, and Logistic Regression) underwent evaluation in typical scenarios.

5.1.2 Experiment 2 - SMOTE for Balancing and OLS for Feature Selection

In the second experiment, it was recognized that balancing imbalanced datasets is crucial. To achieve this, SMOTE was used for data balancing and OLS was used for feature selection. In this experiment, it was determined that the max_depth for the Decision Tree Classifier should be 7. The parameters for the remaining models were kept the same as in Experiment 1. The goal of this experiment was to evaluate how these additional pre-processing techniques affected the models' predictive performance. This objective was also retained for experiments 3 and 4.

5.1.3 Experiment 3 - CT-GAN for Balancing and OLS for Feature Selection

The third experiment examined the use of Conditional Generative Adversarial Networks (CT-GAN) for data balancing and OLS for feature selection. Here too, max_depth was calculated at 7 for the Decision Tree Classifier, this experiment replicated the parameters of the other models from Experiment 1. The utilization of CT-GAN aimed to enrich the dataset with synthetic samples for the underrepresented classes, enhancing the balance and providing a more comprehensive learning experience for the models.

5.1.4 Experiment 4 - CT-GAN for Balancing and RFECV for Feature Selection

For the fourth experiment, the focus was on using CT-GAN to balance the data and Recursive Feature Elimination with Cross-Validation (RFECV) for feature selection. In this experiment, the max_depth for the Decision Tree Classifier was calculated to be 4 to reflect the dataset's new dynamics post CT-GAN balancing. The rest of the model parameters were consistent with those established in Experiment 1. RFECV's iterative process aimed to refine the feature set, improving model simplicity and interpretability by identifying the most impactful attributes.

In these experiments, only the max_depth parameter of the Decision Tree Classifier was altered, while all other model parameters remained as set in Experiment 1.

This implementation was inspired by studies that emphasised the importance of meticulous data pre-treatment in enhancing model performance. By carefully changing the approaches employed in each experiment, it aimed to shed light on the implications of different pre-processing procedures on predictive analytics in the context of consumer behaviour.

6. Evaluation

6.1 Basic Model without Balancing or Feature Selection

Algorithm	Accuracy	Precision	F1- score	Recall
Logistic Regression	0.87	0.86	0.87	0.84
Decision Tree	0.89	0.89	0.89	0.9
GBM	0.9	0.89	0.89	0.89
RNN	0.89	0.88	0.89	0.88

Table 2: Results for Experiment 1

The evaluation of Experiment 1 from Table 2 showed that each algorithm had its unique strengths and weaknesses without any additional pre-processing. Logistic Regression

demonstrated a good balance between precision and recall, whereas the Decision Tree slightly outperformed in classification accuracy. The GBM emerged as the top performer in terms of accuracy, while the RNN, although effective overall, compromised in precision. These outcomes established a fundamental baseline, highlighting the inherent capabilities of the models in their simplest form.

6.2 SMOTE for Balancing and OLS for Feature Selection

Algorithm	Accuracy	Precision	F1- score	Recall
Logistic regression	0.81	0.8	0.81	0.8
Decision Tree	0.891	0.89	0.89	0.89
GBM	0.9	0.89	0.9	0.89
RNN	0.88	0.87	0.88	0.87

Table 3: Results for Experiment 2

In Experiment 2, I have implemented two techniques to improve model performance: SMOTE for data balancing and OLS for feature selection. These changes had notable effects on the metrics of the models tested. Logistic Regression saw a marginal dip in its metrics, indicating that it is sensitive to class balance. The Decision Tree, on the other hand, demonstrated consistent robustness with slight accuracy gains. GBM continued to excel and was not affected by the changes, while RNN notably improved. The balanced data and refined feature set had a significant positive impact on the performance of RNN. These outcomes highlight the importance of class balance and feature selection in improving model performance. The outcome values of the experiment are projected in Table 3.

6.3 CT-GAN for Balancing and OLS for Feature Selection

Algorithm	Accuracy	Precision	F1- score	Recall
Logistic regression	0.83	0.83	0.83	0.83
Decision Tree	0.99	1	1	1
GBM	1	0.99	1	0.99
RNN	0.95	0.94	0.94	0.94

Table 4: Results for Experiment 3

In Experiment 3, I utilized CT-GAN for dataset balancing and OLS for feature selection, and it produced remarkable results. Logistic Regression demonstrated improved performance metrics, indicating its adaptability to the balanced dataset. The Decision Tree and GBM models achieved nearly perfect scores across all metrics, suggesting a significant enhancement in their predictive capabilities with the balanced data. The RNN also displayed substantial improvement, with its metrics approaching the high 0.90s, indicating that it greatly benefited from the combination of CT-GAN balancing and targeted feature selection. This experiment highlights the effectiveness of CT-GAN in optimizing model performance, particularly for complex algorithms. The outcome values of the experiment are projected in Table 4.

6.4 CT-GAN for Balancing and RFECV for Feature Selection

Algorithm	Accuracy	Precision	F1- score	Recall
Logistic regression	0.87	0.86	0.84	0.87
Decision Tree	0.88	0.88	0.88	0.89
GBM	0.9	0.89	0.89	0.89

RNN	0.89	0.87	0.88	0.88

Table 5: Results for Experiment 4

In Experiment 4, I used CT-GAN for data balancing and RFECV for feature selection showed impressive performances. Logistic Regression's metrics showed a notable increase, indicating its effective response to the more refined feature set. The Decision Tree displayed stable high performance, demonstrating its consistency. GBM maintained its lead, showing strong accuracy and balance across precision, F1-score, and recall. The RNN also performed exceptionally well, with its scores closely aligning with the top-performing models. This experiment highlighted the impact of combining CT-GAN's balancing with the precision of RFECV feature selection, enhancing overall model effectiveness. The outcome values of the experiment are projected in Table 3.

In summary, it was observed that Experiment 3 GBM outcome yielded superior results compared to the other experiments. This outcome established the approach of GBM in Experiment 3 as state-of-the-art, evidenced by the performance metrics. The methodologies employed in the other three experiments were found to be akin to previous works in this field. Furthermore, the application of k-fold validation to the GBM algorithm in Experiment 3, which resulted in a validation score of 0.99, has further validated these results, underscoring the robustness and reliability of the findings from this experiment-3 GBM.

6.5 Discussions

The study found that Decision Trees and GBM models performed well in the basic model scenario, even without any specific data treatments, indicating inherent robustness. However, the less impressive metrics of RNN and Logistic Regression showed a possible susceptibility to data instabilities. A slight decrease in Logistic Regression performance was observed when using SMOTE and OLS, indicating that it was vulnerable to variations in the distribution of the data. In contrast, GBM and Decision Trees showed resilience and continued to achieve high levels of accuracy.

The introduction of CT-GAN for data balancing, coupled with OLS for feature selection, significantly elevated the performance across all models. Decision Trees and GBM, in particular, achieved near-perfect scores in all metrics, solidifying the case for advanced data pre-processing methods. It's notable that the combination of CT-GAN and RFECV for feature selection further enhanced model outcomes, with RNN showing considerable improvement, suggesting that complex models like RNN might benefit more from sophisticated feature selection strategies.

In Experiment 3 GBM demonstrated superior outcomes in comparison to the other studies and Algorithms. The performance data clearly demonstrate that the GBM approach used in Experiment 3 is the most advanced and effective. The procedures utilised in the other three tests were discovered to be similar to earlier studies in this domain. In addition, the utilisation of k-fold validation on the GBM algorithm in Experiment 3 has reinforced the validity of these results, emphasising the strength and dependability of the findings from this experiment-3 GBM. The validation score of 0.99 further supports this conclusion.

This analysis underscores the importance of tailored data pre-processing to enhance model performance. Decision Trees and GBM consistently stood out, with GBM showing a slight edge in scenarios devoid of data augmentation. These insights advocate for a strategic approach

in predictive modelling, emphasizing the need for advanced techniques to address class imbalances and feature selection.

7. Conclusion and Future work

Experiment 3 repeatedly showed that the Gradient Boosting Machine (GBM) outperformed the other tests in handling complicated prediction tasks, highlighting its efficiency. A noteworthy observation was the model's consistent achievement of high scores in accuracy, precision, recall, and F1-score, often surpassing other models such as the Decision Tree, RNN, and Logistic Regression. Notably, GBM achieved nearly flawless or flawless scores in multiple tests, which was especially emphasised by its k-fold validation findings. Following a 10-fold cross-validation, the GBM model achieved an exceptional score of 0.99. The persistent brilliance demonstrated by GBM, particularly in different and hard data circumstances, strongly indicates that it is a more efficient model for predicting customer behaviour in our study. This finding is substantiated by the model's overall superior performance metrics in comparison to the other models that were examined.

Looking forward, it's essential to acknowledge the limitations of the current research and the potential areas for future exploration. One limitation lies in the dataset's scope; future studies could incorporate larger and more diverse datasets to capture a broader range of customer behaviors and preferences. Additionally, exploring other sophisticated machine learning algorithms and deep learning techniques could provide further insights into consumer behavior prediction. This expansion would not only offer a more granular understanding of customer propensity but also contribute to the development of more nuanced predictive marketing strategies. The continual evolution of machine learning technologies presents a fertile ground for further enhancing the accuracy and applicability of customer behavior prediction models.

References

1. Sang, G. and Wu, S., 2022, April. Predicting the Intention of Online Shoppers' Purchasing. In 2022 5th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE) (pp. 333-337). IEEE.
2. Suryadi, D., 2020, October. Predicting repurchase intention using textual features of online customer reviews. In 2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI) (pp. 1-6). IEEE.
3. Xu, Z., Sun, Y., Guo, Y., Zhou, Z., Cheng, Y. and Lin, L., 2022, December. User Intention Prediction Method Based on Hybrid Feature Selection and Stacking Multi-model Fusion. In 2022 IEEE 5th International Conference on Electronics and Communication Engineering (ICECE) (pp. 220-226). IEEE.
4. Ling, C., Zhang, T. and Chen, Y., 2019. Customer purchase intent prediction under online multi-channel promotion: A feature-combined deep learning framework. IEEE Access, 7, pp.112963-112976.
5. Chen, L., 2022, March. Reliability prediction of consumer purchase intention based on an improved Bayesian algorithm. In CIBDA 2022; 3rd International Conference on Computer Information and Big Data Applications (pp. 1-4). VDE.
6. Valecha, H., Varma, A., Khare, I., Sachdeva, A. and Goyal, M., 2018, November. Prediction of consumer behaviour using random forest algorithm. In 2018 5th IEEE Uttar Pradesh section international conference on electrical, electronics and computer engineering (UPCON) (pp. 1-6). IEEE.
7. Surendro, K., 2019, March. Predictive analytics for predicting customer behavior. In 2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT) (pp. 230-233). IEEE.
8. Maheswari, K. and Priya, P.P.A., 2017, March. Predicting customer behavior in online shopping using SVM classifier. In 2017 IEEE international conference on intelligent techniques in control, optimization and signal processing (INCOS) (pp. 1-5). IEEE.

9. Zheng, H., 2020, December. Customer Purchase Behavior Prediction and Analysis based on CRM Data Analysis Technology. In 2020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE) (pp. 1374-1378). IEEE.
10. Gumber, M., Jain, A. and Amutha, A.L., 2021, May. Predicting Customer Behavior by Analyzing Clickstream Data. In 2021 5th International Conference on Computer, Communication and Signal Processing (ICCCSP) (pp. 1-6). IEEE.
11. Kapoor, N., 2019, March. A Review on Consumer Behavior Prediction using Data Mining Techniques. In 2019 6th International Conference on Computing for Sustainable Global Development (INDIACom) (pp. 1089-1093). IEEE.
12. Yue Zhang, 2021. Prediction of Customer Propensity Based on Machine Learning. Asia-Pacific Conference on Communications Technology and Computer Science (ACCTCS), 10.1109/ACCTCS52002.2021.00009.
13. Nisha & Ajay Shanker Singh, 2023. Customer Behavior Prediction using Deep Learning Techniques for Online Purchasing. 2nd International Conference for Innovation in Technology (INOCON), 10.1109/INOCON57975.2023.10101102.
14. Mohammadhossein Ghahramani & MengChu Zhou, 2018. Retention analysis based on a logistic regression model: A case study. IEEE 15th International Conference on Networking, Sensing and Control (ICNSC), 10.1109/ICNSC.2018.8361375. Aravind Ravi, Kajanan Sangaralingam & Anindya Datta, 2018.
15. Predicting Consumer Level Brand Preferences Using Persistent Mobility Patterns. IEEE International Conference on Big Data (Big Data), 10.1109/BigData.2018.8622225. Kritarth Gupta, Atharva Hardikar, Devansh Gupta & Shweta Loonkar, 2022.
16. Forecasting Customer Churn in the Telecommunications Industry. IEEE Bombay Section Signature Conference (IBSSC), 10.1109/IBSSC56953.2022.10037334.
17. Dongyi Wang, Xin(Robert) Luo , Ying Hua & Jose Benitez, 2023.Customer's help-seeking propensity and decisions in brands' self-built live streaming E-Commerce: A mixed-methods and fsQCA investigation from a dual-process perspective. ScienceDirect, Journal of Business Research, 10.1016/j.jbusres.2022.113540.
18. Millissa F.Y. Cheung & W.M. To,2017. The influence of the propensity to trust on mobile users' attitudes toward in-app advertisements: An extension of the theory of planned behavior. ScienceDirect, Computers in Human Behaviour, 10.1016/j.chb.2017.07.011.
19. Scott B. Friend, Jeff S. Johnson & Ravipreet S. Sohi ,2018.Propensity to trust salespeople: A contingent multilevel-multisource examination. ScienceDirect, Journal of Business Research, 10.1016/j.jbusres.2017.09.048. Hamdullah Karamollaoğlu, İbrahim Yücedağ & İbrahim Alper Doğru, 2021.
20. Customer Churn Prediction Using Machine Learning Methods: A Comparative Analysis. 6th International Conference on Computer Science and Engineering (UBMK), 10.1109/UBMK52708.2021.9558876. RajaGopal Kesiraju VLN & P. Deeplakshmi , 2021.
21. Dynamic Churn Prediction using Machine Learning Algorithms - Predict your customer through customer behaviour. International Conference on Computer Communication and Informatics (ICCCI), 10.1109/ICCCI50826.2021.9402369.
22. X. Zhao and P. Keikhosrokiani, "Sales Prediction and Product Recommendation Model Through User Behavior Analytics," in Computers, Materials & Continua, vol. 70, no. 2, 2022.
23. O. A. Alghanam, S. N. Al-Khatib, and M. O. Hiari, "Data mining model for predicting customer purchase behavior in e-commerce context," in International Journal of Advanced Computer Science and Applications, vol. 13, no. 2, 2022.
24. K. Matuszelański and K. Kopczewska, "Customer Churn in Retail E-Commerce Business: Spatial and Machine Learning Approach," in Journal of Theoretical and Applied Electronic Commerce Research, vol. 17, no. 1, 2022, pp. 165-198.
25. S. Ahsain and M. A. Kbir, "Predicting the client's purchasing intention using Machine Learning models," in E3S Web of Conferences, vol. 351, 2022, p. 01070. EDP Sciences.