Enhancing Digital Marketing Strategies Using Predictive Analytics.

By

Gamaliel Akade

Student ID: 20001615

DECLARATION

I hereby state that this Project is my original work. All sources used has been cited. The findings, data and conclusions stated in this project are original. I accept total responsibility for the content of this project and contributions made by others have been properly cited. This project adheres strictly to standards and guidelines set by the institution.

Signed: Gamaliel Akade
Student ID: 20001615
Date: 20th May, 2024.

ACKNOWLEDGEMENT

I would like to extend my sincerest gratitude and thanks to all those who have supported and guided me throughout the entire course of this project. I would like to give all glory to God Almighty for the wisdom, strength and fortitude to embark on this project. My deepest appreciation goes to my parents, whose unending support, prayers and encouragements have been a great source of motivation. I am also grateful to my supervisor whose insights and constructive feedback have been most valuable to me during the course of the project.

TABLE OF CONTENT

Abstract

This research project investigates the topic of predictive analytics in the digital marketing domain, focusing on the forecasting of purchase intent following consumer exposure to advertisements. Leveraging the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology, the study precisely navigates through the iterative stages of business understanding, data understanding, data preparation, modelling, and evaluation. Two machine learning algorithms, namely Gaussian Naïve Bayes and Decision Tree classifiers, are employed to identify patterns within the dataset and predict purchase intent. Key pre-processing techniques, such as data engineering, balancing, and normalization, are systematically applied to improve data quality and optimize model performance. Through extensive experimentation, the research uncovers the profound impact of pre-processing steps on model accuracy, precision, and recall. Furthermore, comparative analysis with related studies underscores the importance of efficient pre-processing strategies in improving predictive analytics outcomes. The findings of this study offer practical insights for industry practitioners.

1.0 Introduction

With the introduction of the digital age, businesses now engage with their clientele in several ways. Digital marketing encompasses all forms of advertising that connect and engage with customers via internet channels. Digital marketing includes search engine optimisation (SEO), email marketing, social media marketing, content marketing, and online display ads.

As a result, executing a digital campaign by itself is no longer adequate. Success requires knowing who your target audience is and modifying your messaging accordingly. In this case, predictive analytics can be useful. This branch of data analysis makes use of machine learning and statistical techniques to predict future trends and make inferences from historical data. In the context of digital marketing, predictive analytics can help businesses anticipate the behaviours and preferences of their customers. However, how therefore can these predictions be made? These predictions can be achieved through the application of machine learning algorithms. These algorithms use statistical analysis to make inferences and learn from data. These algorithms will go through vast amounts of consumer data, looking for trends and connections that might help businesses anticipate customer behaviour in the future and enhance their marketing strategies.

In today's data-driven industry, an increasing number of companies are using predictive analytics as a tool to improve their advertising efforts and gain a competitive advantage. To identify trends and forecast future results, predictive analysis makes use of data, machine learning techniques, and mathematical models. Predictive analytics can be applied to digital marketing to minimise attrition, boost customer satisfaction, and optimise advertising efforts.

1.1 Aim and Objectives

1. Evaluate the accuracy of machine learning algorithms in predicting customer purchase intent: By analysing past customer behaviour and campaign interactions, we aim to assess how well these algorithms can identify individuals likely to make a purchase after exposure to a marketing campaign.
2. To assess the efficacy of machine learning algorithms in forecasting purchase intent after consumer exposure to advertisements.
3. To investigate the impact of pre-processing techniques, including data cleaning, balancing, and normalization, on the predictive accuracy of the models.
4. To compare and evaluate the performance of different machine learning algorithms in predicting purchase intent within the digital marketing domain.

1.2 Research Question

Marketers have a lot of opportunities because of the constantly expanding amount of client data available in the digital world. The purpose of this research project is to examine how digital marketing campaigns across several channels can be optimised and personalised by utilising machine learning algorithms in predictive analytics.

The research questions guiding this project are:

1. How well can machine learning algorithms accurately forecast purchase intent following consumer exposure to advertisements?
    a. How consumer actions (such as number of website visits) impact overall predictive accuracy of purchase intent following ad exposure?

b. How do consumer demographics (e.g., age, gender, location) and behaviour patterns impact the predictive accuracy of purchase intent following ad exposure?

1.3 Relevance and Significance

Predicting customer purchase intent and conversion rates are critical objectives for any digital marketing campaign. Understanding these factors allows marketers to:

1. Optimize campaign targeting: By focusing efforts on audiences with a higher predicted purchase intent, marketing spend becomes more efficient and effective.
2. Personalize marketing messages: Tailoring content and offers to individual customer needs based on predicted behaviour can significantly improve engagement and conversion rates.
3. Measure campaign ROI: Accurately predicting conversions allows for a clearer understanding of the return on investment (ROI) for marketing campaigns, enabling data-driven decisions for future strategies.

1.4 Research Problem

The process of critically evaluating research questions begins with a careful analysis of their clarity, specificity, and relevance to the project issue of improving digital marketing strategies using predictive analytics. The research problem is outlined below:
"How well can machine learning algorithms accurately forecast purchase intent following consumer exposure to advertisements?"

1.5 Clarity

The topic focuses on machine learning systems' capacity to predict purchase intent, specifically after consumers have seen adverts. It defines the aim variable (purchase intent) and the context (consumer exposure to commercials), indicating a clear direction for the study.

1.6 Specificity

The question focuses on the accuracy of machine learning systems in projecting purchase intent. This specialisation guarantees that the research focuses on a single component of digital marketing analytics, avoiding broad generalisations or ambiguities.

1.7 Relevance

The question is pertinent to the project topic because it directly addresses the primary goal of using predictive analytics to improve digital marketing campaigns. The study topic is strongly related to the project's overall goals because it focuses on purchase intent, which is an important aspect in consumer decision-making.

Subsequent sections in this research includes Literature Review, Research Methodology, Result & Analysis, and Discussion and Conclusions. The research problem highlights the research questions,

their clarity, specificity, and relevance to the overall research topic. The literature review provides a critical and analytical overview of significant literature related to the research topic. The research methodology explains the research procedure, techniques used, and data analysis methods. The results are thoroughly explained in the result and analysis section. Finally, the project is concluded with limitations faced and improvement suggestions for future works are detailed in the discussion and conclusions section.

2.0 Literature Review

In the realm of digital marketing, using predictive analytics is a game-changing tool that enables marketers to surpass traditional approaches. Businesses can now combine historical data trends with cutting-edge analytical approaches to estimate consumer behaviour with previously unachievable granularity. This ability to foresee allows marketers to efficiently modify their strategies and react in real-time to shifting consumer demands and wants, according to (Garcia et al., 2020). Because it can offer fresh perspectives on the intricate dynamics of consumer interactions online, predictive analytics is essential to digital marketing. It acts as a conduit for the discovery of intricate patterns hidden in massive data sets, giving advertisers insight into the underlying needs, drives, and decision-making processes of their target audience (Sayyad et al., 2020). Predictive analytics is a vital part of today's digital marketing toolkit because it combines data-driven insights, predictive modelling, and consumer-centric strategies in a way that helps businesses thrive in a fiercely competitive and ever-changing digital landscape. In the parts that follow, this research examines the various advantages, extensive applications, and challenges of integrating predictive analytics into digital marketing frameworks.

2.1 Traditional Marketing Strategies

Conventional marketing strategies have frequently relied on guesswork and intuition, which has produced less than ideal outcomes and missed chances (Sinha, 2018). A more scientific approach to marketing is provided by predictive analytics, which enables companies to make data-driven decisions that can spur expansion.

To reach target consumers, conventional digital marketing methods frequently depend on broad approaches and few data insights (Bist et al., 2022). Some of these strategies typically include:

- Mass email campaigns: this include sending generic emails devoid of any personalisation or targeted segmentation to a large audience.
- Static display ads: these are displayed on websites without the use of dynamic optimisation or user behaviour-based personalised targeting.
- Keyword-Centric SEO: Concentrating just on keywords without delving further into the intentions or actions of users.
- Last-Click Attribution: Ignoring the whole customer journey and just attributing credit for conversions to the final interaction before to a purchase.

2.2 Approaches and Their Limitations

It can be challenging to precisely evaluate marketing performance and pinpoint the most successful tactics when using conventional marketing strategies as they frequently rely on obsolete procedures. These restrictions may result in several difficulties, such as:

1. Ineffective Ad Campaigns: Conventional attribution methods, such last-click attribution, frequently fall short of capturing the complexities of the customer journey and the impact of several touchpoints, which results in the inefficient use of marketing budgets and unsuccessful advertising campaigns (Bist et al., 2022).
2. Lack of Personalised Client Engagement: Conventional marketing strategies frequently fall short in terms of personalisation, failing to consider the unique demands and preferences of each client.

Reduced conversions and a disengaged consumer base may result from this (Leeflang et al., 2014).

3. Difficulty in Identifying High-Value Customers: Conventional approaches have trouble identifying high-value clients at an early stage, which makes it difficult to concentrate marketing efforts on the market niches with the highest income potential (Kingsnorth, 2022).

2.3 Role of Data Analytics in Digital Marketing

2.3.1 Conversion/ Revenue Attribution

Businesses are always trying to maximise their Return on Ad Spend (ROAS) and optimise their techniques. This endeavour heavily relies on conversion attribution, which is the process of identifying and assigning value to the touchpoints that result in a customer's conversion. Conversely, traditional attribution techniques sometimes fail to capture the complex customer journey and the impact of multiple interactions (Bekmamedova and Shanks, 2014).

Every advertising network keeps track of interactions and conversions on its network, giving customers a fictitious return on investment. In digital marketing, this measurement is frequently referred to as the swim-lane measurement. Marketers use this phrase to characterise the process of using each marketing platform, assessing each of their campaigns as if they were independent analytical units apart from other marketing-related activities (Beck, 2005; Romero Leguina et al., 2020).

For instance, Company A uses Instagram and Snapchat to promote their website and increase visitors. After viewing the advertisement on Instagram, Person X does nothing. Afterwards, Person X visits the Company A website after clicking on the Snapchat advertisement. A view-

through conversion will be recorded by Instagram (since user X saw the advertisement on the platform). Snapchat will record a click-through conversion (because user X clicked on the Snapchat advertisement). Although there was only one conversion, two will be reported overall across the two platforms. With predictive analytics, this common swim-lane measurement/attribution example can be fixed.

By examining vast amounts of client data, data analytics may identify patterns, predict future behaviour, and provide more accurate insights into the customer journey. According to (Theodoridis and Gkikas, 2019) businesses who have used predictive analytics and artificial intelligence into their digital marketing strategies have shown improvements in engagement, conversions, and revenue growth.

2.3.2 Campaign Optimisations

Predictive analytics is a key component in revolutionising campaign optimisation in the field of digital marketing as it facilitates data-driven decision-making, improves targeting accuracy, and maximises campaign performance (Jeffery, 2010). Predictive analytics, at its foundation, gives marketers the ability to predict and anticipate future results using complex machine learning algorithms and trends found in past data. Predictive analytics helps with real-time campaign execution optimisation. With predictive models and real-time data streams, marketers can promptly modify campaign components including message, ad placements, and targeting criteria.

Statistical models that show how advertising affects consumer behaviour and company outcomes need to take into consideration hundreds of variables pertaining to competing activity, marketing strategies, and market conditions. These models are used each ad serving platform to

precisely assign the weight of each variable, optimise the advertising strategy, and optimise and reallocate campaign budget (Xu et al., 2015). Furthermore, the recognition of valuable audience segments and their habits is made easier by predictive analytics. Marketers may develop highly targeted advertisements that are suited to audience groups by examining a variety of datasets that include demographic data, historical behaviours, and interactions (D. Newton et al., 2013). This degree of customisation improves conversion rates and engagement, which results in more successful marketing outcomes.

2.4 Predictive Analytics in Digital Marketing

Predictive analytics involves the use of historical data, statistical algorithms, and machine learning techniques to forecast future outcomes and trends. Within the realm of digital marketing, predictive analytics holds significant promise for improving targeting, personalization, and overall marketing effectiveness. By analysing vast amounts of data generated from various digital touchpoints, businesses can gain insights into consumer behaviour, preferences, and purchasing patterns, enabling them to make data-driven decisions and tailor their marketing efforts more effectively. The digital marketing landscape is awash with data, presenting both a challenge and an opportunity. Traditional, intuition-based approaches are no longer sufficient. This literature review explores the intersection of predictive analytics and machine learning (ML) algorithms to optimize and personalize digital marketing strategies.

(Jeffery, 2010) demonstrated the applicability of predictive analytics techniques such as regression analysis and decision trees in analysing customer data and predicting future purchasing behaviour. Similarly, (Xu et al., 2015) emphasized the role of machine learning algorithms such as

random forest and neural networks in segmenting customers and personalizing marketing campaigns. These studies highlight the versatility and effectiveness of predictive analytics in optimizing digital marketing strategies.

As a result, a variety of advanced models are available to meet these various needs, including, but not limited to:

a. Recommender Systems: One particularly popular recommendation engine technique in the media and e-commerce industries is collaborative filtering. This method offers tailored recommendations for products or content by exploring the finer points of consumer behaviour and tastes, going beyond basic demographic segmentation. Collaborative filtering finds trends and similarities amongst users by examining large datasets that include user interactions, past purchases, ratings, and browsing habits. By utilising these insights, the system can make well-informed recommendations for goods or content that suit the tastes and interests of the user, increasing user engagement and increasing conversion rates. Collaborative filtering consistently improves its recommendations by adjusting to changing user preferences and market trends through iterative learning procedures and feedback loops. (Khodabandehlou, 2019; Kim and Kim, 2001)

b. Neural Networks: Reputed for their capacity to interpret information through interconnected layers of artificial neurons, Neural Networks constitute a potent class of machine learning models. Because of their skill at identifying complex relationships and patterns in data, these models can produce precise forecasts and projections based on past performance. Neural networks are highly effective across a wide range of

applications, providing priceless insights into consumer behaviour and market trends by utilising complex algorithms and vast amounts of data. Neural networks' ability to predict customer turnover is one of its main advantages; this is a crucial worry for companies in all sectors of the economy. Neural networks are able to detect minor signs and precursors of client attrition by analysing past customer data, including purchase patterns, frequency of interactions, and satisfaction levels. (Selvin et al., 2017)

c. Decision Trees: Using a hierarchical, tree-like structure, decision trees are a flexible and user-friendly machine learning model that analyses data and makes judgements depending on the characteristics of the input features. Decision trees provide insightful and useful information for a variety of applications in the field of digital marketing.
Decision trees work fundamentally by dividing the dataset into subsets according to the values of various attributes. This process eventually results in the construction of a tree structure, where each internal node denotes a decision made based on a feature and each leaf node in the tree indicates a class label or outcome. Decision Trees are very intriguing to marketing professionals who are looking for clear and simple decision-making procedures because of their natural interpretability. Based on past data, decision trees are excellent at projecting customer behaviour and making outcome predictions. Decision Trees can give useful insights into future customer behaviour, whether it's predicting if a client will make a purchase, take advantage of a deal, or cancel their subscription. (Bounsaythip and Rinta-Runsala, 2001; Olson and Chae, 2012).

d. Regression Analysis: One of the fundamental statistical methods used to investigate and measure the relationships between variables in a

dataset is regression analysis. It facilitates a deeper comprehension of cause-and-effect interactions in marketing situations by allowing marketers to explore the relationship between changes in one or more predictor factors and changes in a target variable. Regression analysis, for instance, can be used by marketers to investigate the connection between sales revenue and advertising expenses, providing insight into how well marketing efforts contribute to measurable business results. Regression analysis can show that some digital advertising platforms are more profitable than others for investments made in them, which would lead marketers to reallocate resources to get the most return on investment. Regression analysis also makes it easier to find possible areas in marketing plans where optimisation could be done. (Lemos, 2015).

2.5 Benefits of Predictive Analytics in Digital Marketing

The following are some of the main advantages of predictive analytics for improving digital marketing strategies:

- Improved Conversion Attribution: By precisely capturing the multi-touch aspect of the customer experience, predictive analytics can provide credit to various touchpoints according to their proportional effect (Kotane et al., 2019)
- Personalised Customer Engagement: By tailoring marketing efforts to the unique behaviours and interests of each customer, predictive analytics may create more pertinent and interesting interactions.
- Predictive analytics: By predicting which consumers are most likely to make purchases or interact with marketing campaigns, it is possible for organisations to concentrate their efforts on these high-value client categories (García et al., 2015).

- Optimisation of Marketing Campaigns: By determining the best ad locations, targeting tactics, and creative components, predictive analytics may assist companies in optimising their advertising campaigns.

3.0 Research Methodology

This study's approach follows a defined technique for investigating and resolving complicated business challenges using data analytics and mining. This technique, known as CRISP-DM (Cross-Industry Standard Process for Data Mining), consists of five vital steps, each with a specific function in the quest of explaining corporate objectives and using data-driven insights to achieve them (Azevedo and Santos, 2008):

- Business Understanding: The research begins with a thorough definition of the overall business objectives. During this stage, careful consideration is given to determining the role and possible value of data analytics and mining approaches in assisting the achievement of these goals.
- Data Understanding: This phase is defined by systematic collection and first exploration of the appropriate datasets. This primary examination, which is frequently helped by exploratory visualisations, serves to highlight the fundamental traits and nuances of the data.
- Data Preparation: The emphasis here is on implementing thorough data cleaning, transformation, and manipulation procedures. This phase is crucial for guaranteeing the dataset's quality and consistency, creating a solid foundation for future analytical efforts.
- Modelling: The modelling stage is a critical stage in which machine learning algorithms are deployed and fine-tuned to the dataset's unique characteristics. These algorithms are refined through repeated experimentation and validation on a variety of datasets to successfully discover patterns and relationships that support the business objectives under consideration.

- Evaluation: The Evaluation phase involves a critical assessment of the created models in relation to the predefined business objectives. By exposing the models to rigorous comparison analysis, the efficacy and performance of each model are evaluated, allowing for the selection of the most appropriate model for meeting the defined business goals.

A visual representation of the methodology as described, is represented in Figure 1.

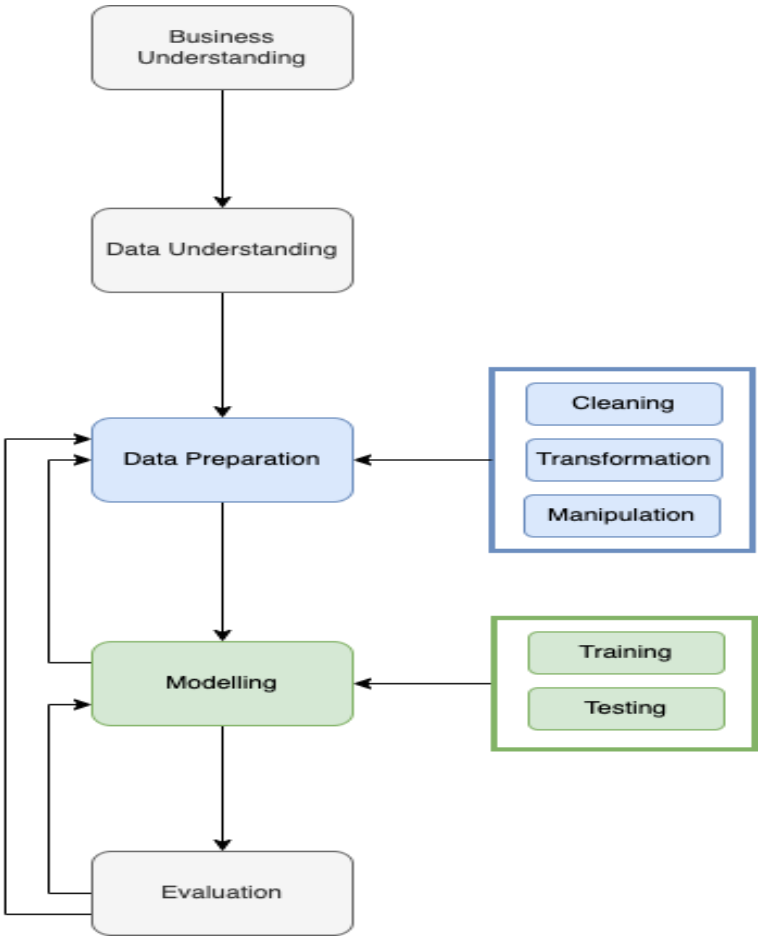

Figure 3.1: Data Methodology - CRISP-DM

3.1 Business Understanding:

The business understanding specifically necessitates a comprehensive comprehension of the commercial significance attached to determining consumer purchase intent subsequent to their interactions with advertisements. This has been properly described in the introduction and research problem section of this research. Ultimately, businesses seek the use of data and predictive analytics to understand consumer purchase intent and thus the effectiveness of their advertising campaigns. By aligning data mining objectives with overarching business objectives, organizations can ensure that data mining efforts are purposeful and aligned with strategic priorities. Ultimately, the business understanding step lays the groundwork for subsequent data mining activities, guiding the selection of appropriate techniques and methodologies to address specific business needs and maximize the value derived from data analysis.

3.2 Data Understanding:

Key activities in the data understanding step include data collection, data description, data exploration, and data quality assessment. Data collection involves gathering the relevant datasets from appropriate sources, ensuring they align with the project's objectives. Subsequently, the data is thoroughly described, documenting its format, types of variables, and any metadata available.

Data exploration involves visualizing and summarizing the data to understand its distribution, relationships between variables, and potential outliers or anomalies. Techniques such as statistical summaries, histograms, scatter plots, and correlation matrices are commonly employed during this phase to gain a comprehensive understanding of the dataset's characteristics.

The choice of dataset consists of data which is relevant to the research topic. It consists of over 2000 rows and 29 columns. The columns in the dataset are described below:

Table 1: Description of the dataset

SN	Column	Description	Data Type
1	ID	Row ID	Numeric
2	Year_birth	Birth year of the customer	Numeric
3	Education	Highest education level of the customer	Categorical
4	Marital_status	Marital status of the customer	Categorical
5	Income	Annual income of the customer	Numeric
6	Kidhome	Number of children of the customer below the age of 13	Numeric
7	Teenhome	Number of children of the customer over the age of 13	Numeric
8	Dt_customer		
9	Recency	Date of customer's most recent purchase	Date
10	MntWines	Amount spent on wines monthly	Numeric
11	MntFruits	Amount spent on fruits monthly	Numeric
12	MntMeatProducts	Amount spent on meat monthly	Numeric

13	MntFishProducts	Amount spent on fish monthly	Numeric
14	MntSweetProducts	Amount spent on sweets monthly	Numeric
15	MntGoldProds	Amount spent on premium products monthly	Numeric
16	NumDealsPurchases	Number of discounts purchases made monthly	Numeric
17	NumWebPurchases	Number of purchases made online	Numeric
18	NumCatalogPurchases	Number of purchases made through catalogs	Numeric
19	NumStorePurchases	Number of purchases made in store	Numeric
20	NumWebVisitsMonth	Number of monthly visits to the store's website	Numeric
21	AcceptCmp1	Purchase made after the first campaign?	Boolean
22	AcceptCmp2	Purchase made after the second campaign?	Boolean
23	AcceptCmp3	Purchase made after the third campaign?	Boolean
24	AcceptCmp4	Purchase made after the fourth campaign?	Boolean
25	AcceptCmp5	Purchase made after the fifth campaign?	Boolean
26	Complain	Customer made a complaint about campaigns?	Boolean

27	ZCostContact	Cost of contacting the customer	Numeric
28	ZRevenue	Revenue generated from the customer	Numeric
29	Response	Purchase made after the final campaign?	Boolean

3.3 Justification of Dataset Choice

The dataset chosen for this study is highly relevant to the current trends in digital marketing for several reasons:

1. **Comprehensive Customer Profiling**: The dataset includes detailed demographic information such as age, education level, marital status, and income. This allows for a thorough analysis of how these factors influence consumer behaviour and purchase intent, aligning with the trend of personalized marketing.
2. **Behavioural Insights**: It captures various aspects of customer behaviour, including the amount spent on different product categories (wines, fruits, meats, etc.), the number of purchases made online, through catalogues, and in stores, as well as the frequency of visits to the store's website. This aligns with the growing emphasis on omnichannel marketing and understanding customer journeys across multiple touchpoints.
3. **Campaign Response Tracking**: The dataset includes information on customer responses to different marketing campaigns, indicated by Boolean fields (e.g., AcceptCmp1, AcceptCmp2). This is crucial for evaluating the effectiveness of marketing strategies and tailoring future campaigns based on past responses, which is a current trend in data-driven marketing.

4. **Financial Metrics**: Inclusion of financial metrics such as revenue generated from the customer and the cost of contacting them (ZCostContact and ZRevenue) provides a direct link between marketing activities and financial outcomes, aligning with the trend towards ROI-focused marketing strategies.

3.4 Limitations of the Dataset

1. **Data Quality and Completeness**: The accuracy and completeness of the dataset are critical for reliable analysis. Missing or inaccurate data entries, particularly in key variables like income or purchase amounts, could lead to biased or incorrect conclusions. In addition, the size of the dataset is limited to just over 2000 rows, which is not representative for companies with much larger customer base.
2. **Time Frame**: The dataset does not cover a sufficiently long period to capture long-term trends and seasonal variations in consumer behaviour, which is crucial for developing robust predictive models.
3. **Limited Scope**: The dataset is focused on a specific set of products and marketing campaigns. It does not fully capture the diversity of marketing channels and customer interactions prevalent in the broader digital marketing landscape.
4. **Static Data**: The dataset provides a snapshot of customer behaviour and campaign responses at a given time. It may not reflect changes in consumer behaviour over time or the impact of external factors such as economic shifts or competitive actions.

3.5 Data Preparation:

Prior to data modelling, it is pertinent to prepare the data so as to ensure consistency and accuracy. For this research, data preparation processes such as cleaning the data and transforming the data were applied to ensure accuracy of the data for processing.

3.5.1 Data Cleaning:

This process involves the identification, and correcting of inconsistencies, errors, and anomalies in the dataset. It includes steps such as identifying and handling missing data, outlier detection and data normalisation – all of which was carried out in this project.

In this project, there were only 24 missing values from the “Income” column, hence those rows were dropped

3.5.2 Feature Engineering:

This involves the transforming of raw data into useful pieces of information that help machine learning models make accurate predictions. By selecting, combining, or creating new features from the data, we make it easier for the models to understand and solve the problem at hand. It's all about giving the models the right tools to do their job effectively.

For this project, several columns were transformed to create new columns which were used in the data modelling.

1. Age – the age column was derived by subtracting the birth year of each customer from the current year (2024).
2. childrenHome – this column representing the number of children in the household was derived by adding kidHome and TeenHome.
3. Marital Status – this column initially had 8 categories, however, some categories were collapsed, thus reducing the number of categories to 4 (“single”, “in a relationship”, “divorced”, and “widow”).

4. Education – similarly, the education column was also collapsed to include only 3 categories, namely, “Basic”, “Graduate”, “Post Grad”.
5. Success – this column was created as the new target column. It explains whether or not a purchase was made after any of the campaigns. This column was derived by first adding the values from AcceptedCmp1-5 and Response, then converting any value greater than 0 to 1 and every other value to 0. Thus making this column a Boolean value of 0 or 1 for each row.

3.5.3 Data Balancing:

In the dataset, just over 70% of the users did not purchase despite being exposed to adverts, while only 27% purchased.

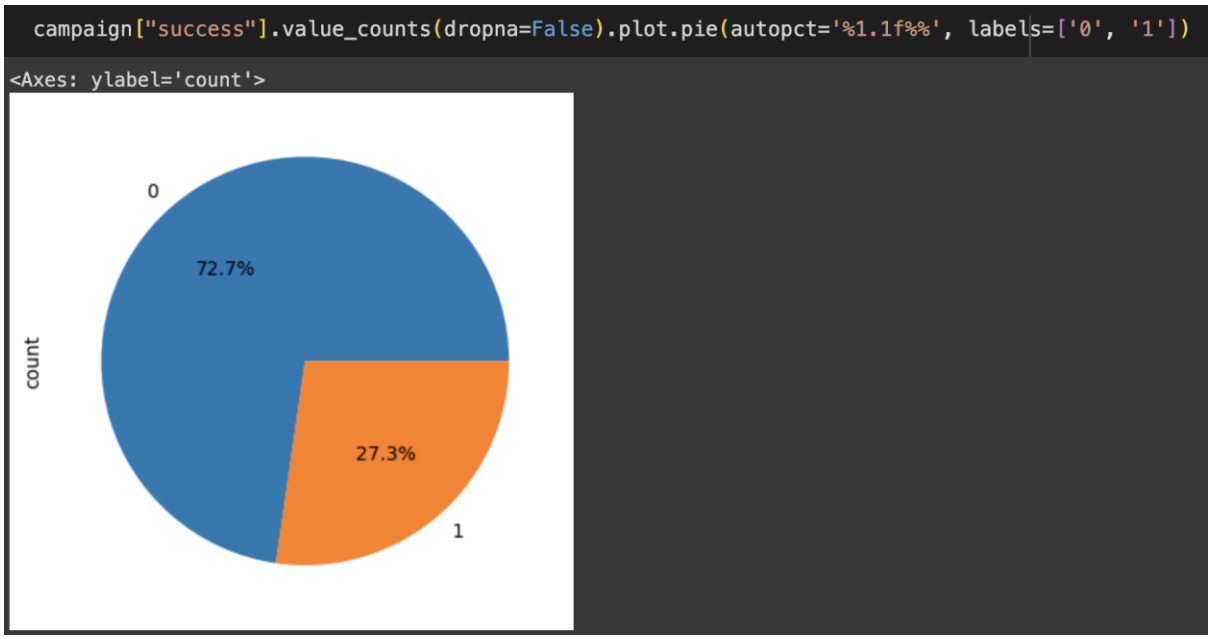

Figure 2: Imbalanced Dataset

The imbalance in the purchase intent ratio poses a significant challenge. When datasets are imbalanced, it leads to bias during modelling. Essentially, the model tends to favour the more prevalent value in the dataset, skewing predictions toward that dominant value.

Hence, it's crucial to address this bias by balancing the dataset. In this project, the Random Over Sampler technique is employed to achieve this. This method involves generating synthetic data for the less prevalent category in the dataset and increasing its representation to match that of the more prevalent category. Specifically, in this scenario, the number of "purchases" or "1" instances is augmented to align with the count of "non-purchases" or "0" instances. The effect of this technique is discussed further in the results and analysis.

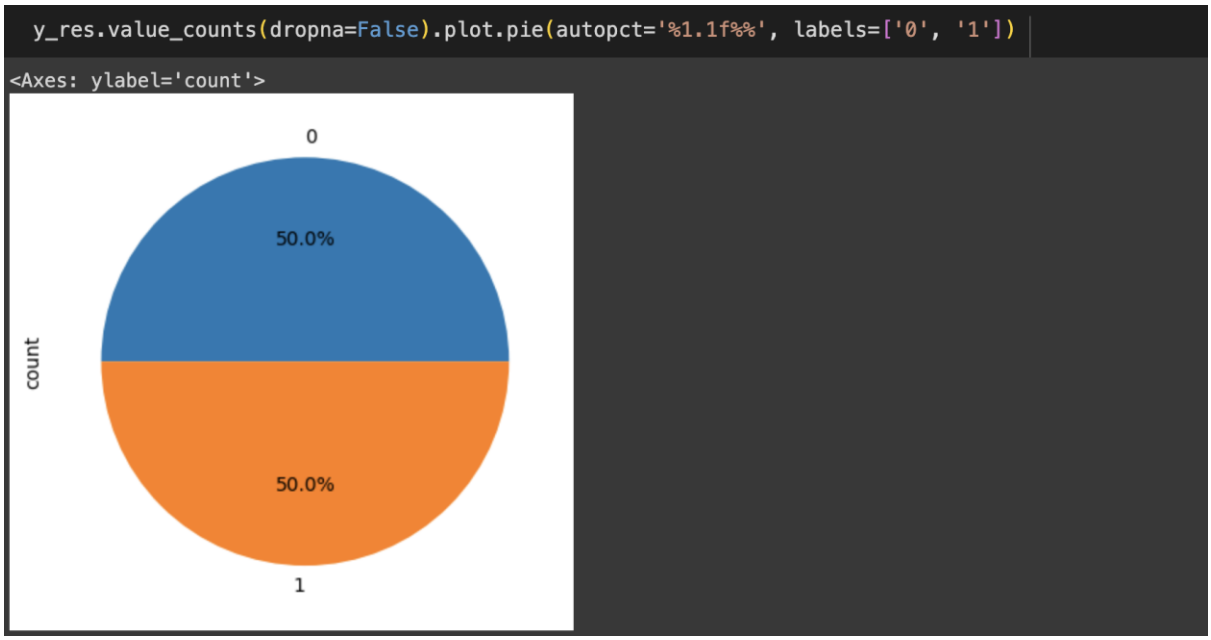

Figure 3: Balanced Dataset after Random Over Sampling

3.6 Data Modelling:

Data modelling is the creation and refining of mathematical representations or models based on the dataset at hand. These models are intended to identify patterns, connections, or trends in data that can be utilised to make predictions or acquire insights. Data modelling can be done using a variety of techniques, including machine learning algorithms, statistical models, and mathematical equations. The ultimate purpose of

data modelling is to provide accurate and actionable forecasts or insights that can help with decision-making and business results.
For this research, 3 experiments were conducted using 2 algorithms each, resulting in 6 models in total.

Algorithm 1:

Gaussian Naïve Bayes: Gaussian Naive Bayes is a popular algorithm in predictive analytics because of its simplicity and efficacy in categorization jobs. It is very useful when working with continuous data and is based on Bayes' theorem, which determines the likelihood of a hypothesis given the data. This technique is particularly useful in text classification and spam detection, where it can easily handle vast feature spaces. It is computationally inexpensive and can manage missing data effectively, making it suited for practical applications (Jahromi and Taheri, 2017).

Algorithm 2:

Decision Tree: The decision tree algorithm is a key tool in predictive analytics, valued for its clarity and flexibility. It works by recursively partitioning the data into subsets based on the attributes that best divide the dataset into homogeneous groups in terms of the target variable. This process continues until a stopping requirement is fulfilled, resulting in a tree-like structure with interior nodes representing feature-based decisions and leaf nodes representing anticipated outcomes. Decision trees have various advantages, including the capacity to handle both numerical and categorical data, require little data pre-processing, and provide visible and interpretable models. In conclusion, decision trees and their variants remain important in predictive analytics due to their simplicity, interpretability, and success across multiple domains (Du and Zhan, 2002).

Experiment 1 – Data Modelling Without Data Balancing:

This experiment seeks to compare the performance of several algorithms on an uneven dataset without using any data balancing strategies. The main goal is to determine the influence of modelling approaches on unbalanced data and how these algorithms manage skewed class distributions. The experiment aims to identify any biases, problems, or constraints related with modelling uneven data by applying several algorithms directly to the dataset without balancing. Furthermore, the experiment intends to provide best practices for dealing with class imbalance difficulties in predictive modelling tasks, as well as to assist future research on algorithm selection and performance optimisation for imbalanced data circumstances.

Experiment 2 – Data Modelling With Data Balancing:

This experiment is designed to investigate the impact of employing data balancing techniques on the performance of predictive modelling algorithms. In contrast to Experiment 1, where algorithms were applied directly to an unbalanced dataset, Experiment 2 introduces the Random Over Sampling strategy to mitigate class imbalance before modelling. By applying these data balancing techniques before modelling, Experiment 2 seeks to evaluate whether balanced datasets lead to more accurate, reliable, and generalizable predictive models compared to their unbalanced counterparts.

Experiment 3 – Data Modelling with Data Balancing & Scaling:

This experiment expands upon the preceding experiments by incorporating additional pre-processing steps to further enhance model performance. In this experiment, alongside data balancing techniques introduced in Experiment 2, MinMax scaling is applied to normalize feature values within a specific range. The primary objective of Experiment 3 is to investigate the combined effect of data balancing and MinMax scaling on the predictive modelling process. MinMax scaling transforms feature values to a predefined range (e.g., [0, 1]), ensuring that all features contribute equally to the model training process regardless of their original scale. By normalizing feature values, MinMax scaling mitigates the potential influence of feature magnitudes on model training and improves the stability and convergence of certain algorithms.

3.7 Evaluation

For each of the three experiments, the predictive models are evaluated using a variety of measures to determine their overall success. These measurements are accuracy, precision, recall, F1 score, and ROC AUC.

1. **Accuracy**: Accuracy measures the proportion of correctly predicted instances out of the total instances in the dataset. It provides an overall assessment of model performance but may be misleading in the presence of class imbalance (Vujović, 2021).
2. **Precision**: Precision quantifies the proportion of true positive predictions among all positive predictions made by the model. It indicates the model's ability to avoid false positive predictions and is particularly relevant in scenarios where false positives are costly (Vujović, 2021).

3. **Recall**: Recall, also known as sensitivity, measures the proportion of true positive predictions among all actual positive instances in the dataset. It reflects the model's ability to capture all relevant instances of the positive class and is crucial in applications where false negatives are detrimental (Vujović, 2021).
4. **ROC AUC**: The ROC AUC metric evaluates the trade-off between true positive rate (sensitivity) and false positive rate across different threshold values. It provides a comprehensive assessment of a model's ability to discriminate between classes, with higher values indicating better classification performance (Vujović, 2021).

4.0 Results and Analysis

4.1 Experiment 1 – Gaussian Naïve Bayes

This experiment applied the naïve bayes algorithm on an unbalanced dataset and was evaluated using the metrics earlier described. The confusion matrix was used to visualise the distribution of the true values versus the predicted values. It shows the true positive, true negative, false positive and false negative.

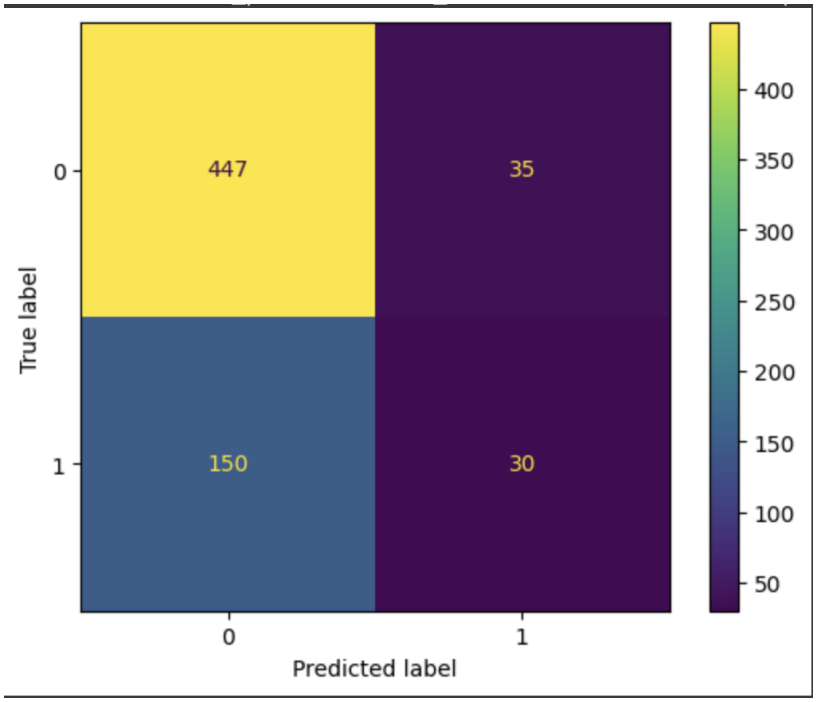

Figure 4.1: Confusion Matrix for Experiment 1 - Gaussian Naive Bayes

Given the confusion matrix: True positives (TP) = 30, True negative (TN) = 447, False negatives (FN) = 150 and False positive (FP) = 35

Accuracy is calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Accuracy = \frac{30 + 447}{30 + 447 + 35 + 150} \approx \mathbf{72.05\%}$$

Precision is calculated as:

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

$$Precision = \frac{30}{30 + 35} \approx \mathbf{46.15\%}$$

Recall also known as sensitivity or True Positive Rate (TPR), is calculated as the ratio of true positives to the sum of true positives and false negatives:

$$TPR = \frac{TP}{TP + FN}$$

$$TPR = \frac{30}{30 + 150} \approx \mathbf{16.67\%}$$

ROC AUC is calculated as:

$$AUC = \int_0^1 TPR(FPR^{-1})dFPR$$

$$AUC = \int_0^1 16.67(7.26^{-1})d7.26 \approx \mathbf{54.70\%}$$

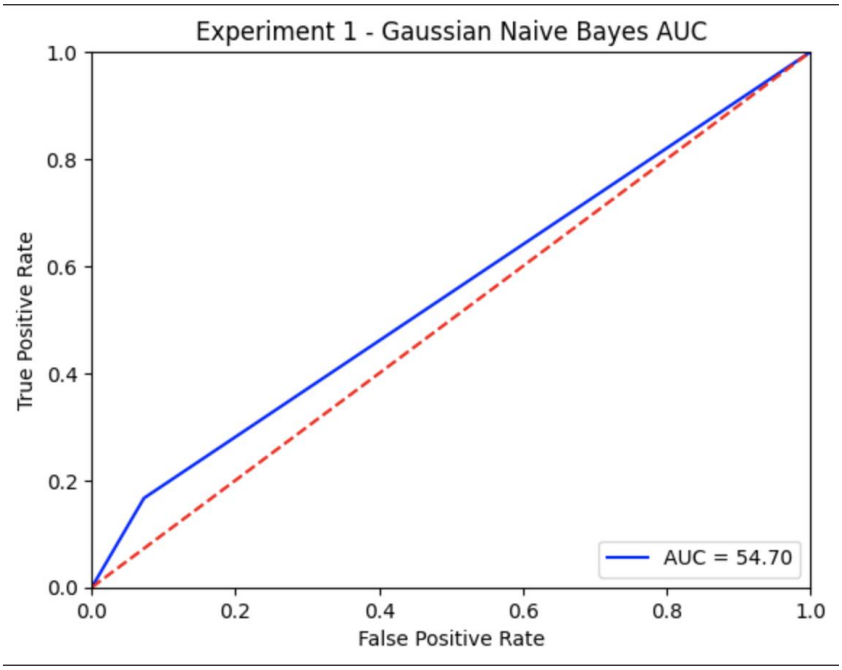

Figure 4.1.2: Experiment 1- AUC for Naive Bayes Classifier

Overall, the Gaussian Naïve Bayes algorithm demonstrated an accuracy of approximately 72.05%, indicating its effectiveness in classifying instances correctly. Precision, at about 46.15%, reflected the model's accuracy in identifying positive instances. The recall, or true positive rate, was approximately 16.67%, showcasing the model's poor capability to detect positive instances among all actual positives. The ROC AUC, around 54.70%, reflected the model's ability to differentiate between positive and negative instances across various thresholds. These metrics collectively offer valuable insights into the model's performance and aid informed decision-making in predictive analytics.

4.1.2 Experiment 1 – Decision Tree

This experiment applied the decision tree classifier on an unbalanced dataset and was evaluated using the metrics earlier described. The confusion matrix was also used to visualise the distribution of the true values versus the predicted values.

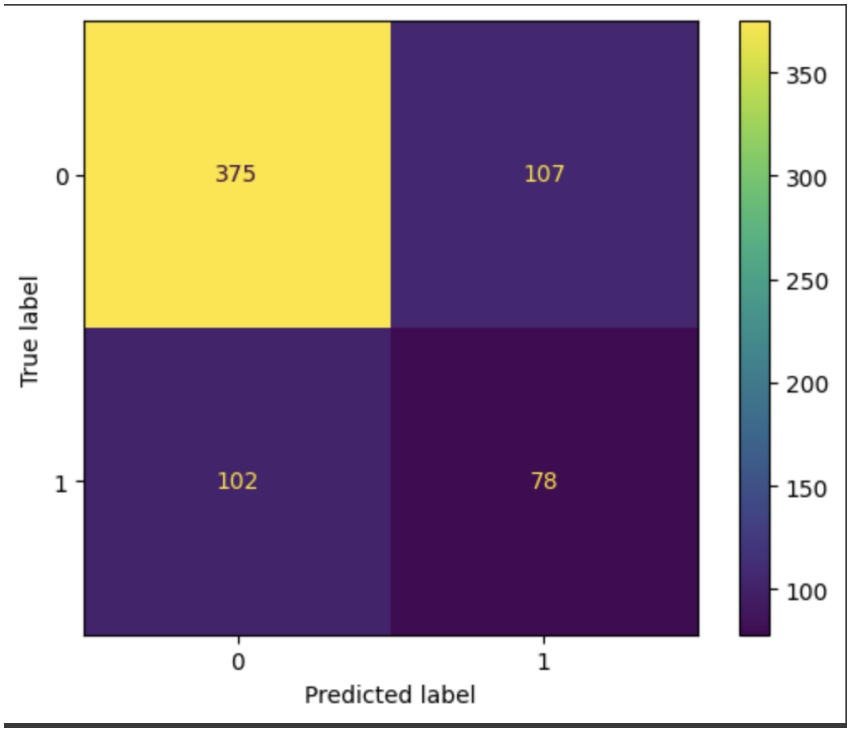

Figure 4.1.3: Confusion Matrix for Experiment 1 - Decision Tree Classifier

Given the confusion matrix: True positives (TP) = 78, True negative (TN) = 375, False negatives (FN) = 102 and False positive (FP) = 107

Accuracy is calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Accuracy = \frac{78 + 375}{78 + 375 + 107 + 102} \approx \mathbf{68.43\%}$$

Precision is calculated as:

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

$$Precision = \frac{78}{78 + 107} \approx \mathbf{42.16\%}$$

Recall also known as sensitivity or True Positive Rate (TPR), is calculated as the ratio of true positives to the sum of true positives and false negatives:

$$TPR = \frac{TP}{TP + FN}$$

$$TPR = \frac{78}{78 + 102} \approx \mathbf{43.33\%}$$

ROC AUC is calculated as:

$$AUC = \int_0^1 TPR(FPR^{-1})dFPR$$

$$AUC = \int_0^1 43.33(22.20^{-1})d22.20 \approx \mathbf{60.57\%}$$

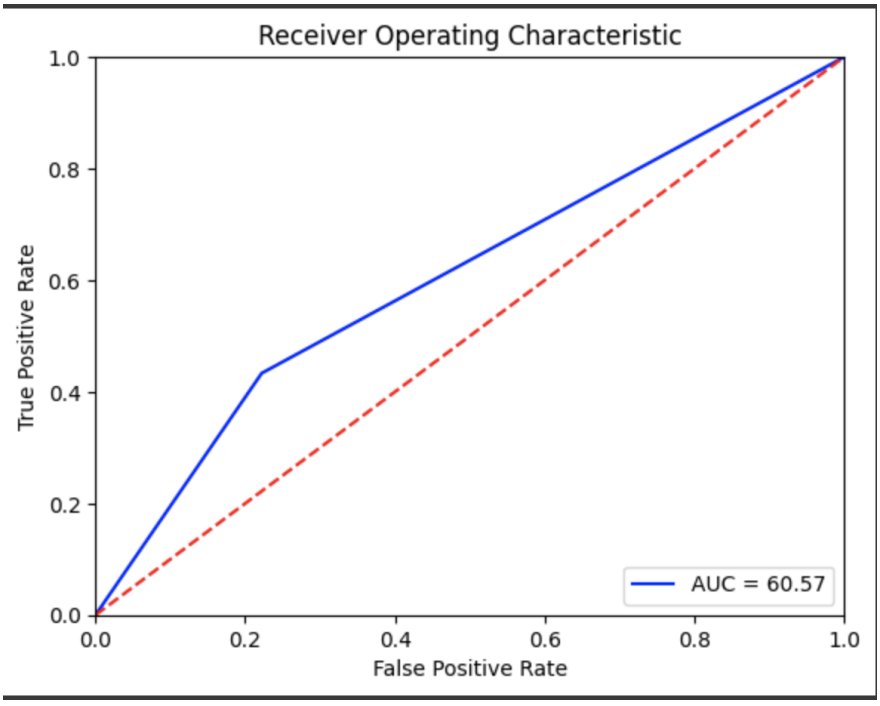

Figure 4.1.4: Experiment 1- AUC for Decision Tree Classifier

In this experiment, the Decision Tree Classifier was tested on an unbalanced dataset. The model showed an accuracy of about 68.43%, which means it correctly predicted around 68.43% of instances overall.

Precision, which measures how accurate positive predictions were, stood at about 42.16%. This means that out of all instances the model predicted as positive, only around 42.16% were truly positive. Recall (TPR), was approximately 43.33%. This means the model correctly identified about 43.33% of all actual positive instances. The ROC AUC, around 60.57%, indicates the model's ability to distinguish between positive and negative instances.

4.2 Experiment 2 – Gaussian Naïve Bayes:

This experiment applied the naïve bayes algorithm after the random over sampling technique was applied to balance the dataset.

Confusion Matrix

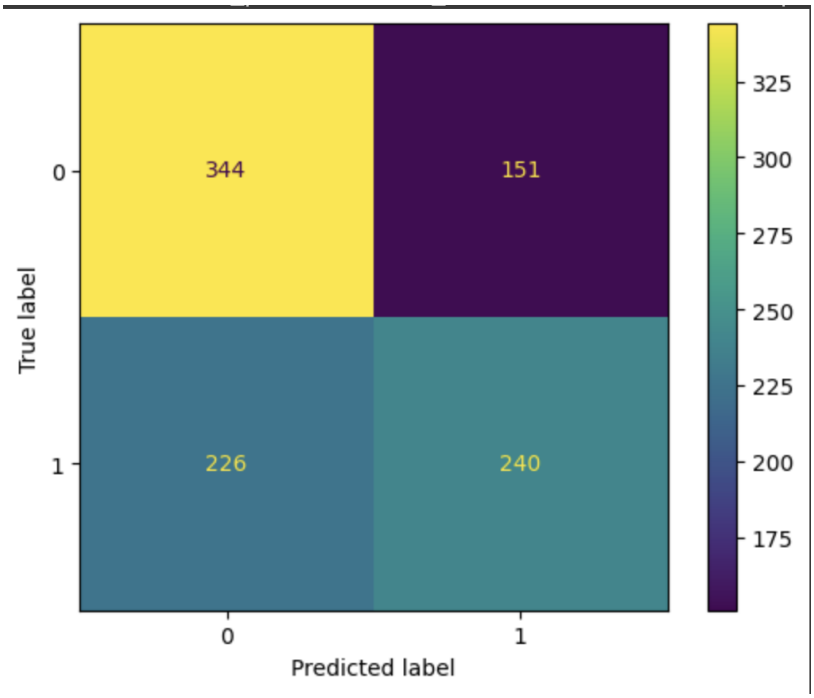

Figure 4.2.1: Confusion Matrix for Experiment 2 - Gaussian Naive Bayes

Given the confusion matrix: True positives (TP) = 240, True negative (TN) = 344, False negatives (FN) = 226 and False positive (FP) = 151

Accuracy is calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Accuracy = \frac{240 + 344}{240 + 344 + 151 + 226} \approx \mathbf{60.77\%}$$

Precision is calculated as:

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

$$Precision = \frac{240}{240 + 151} \approx \mathbf{61.38\%}$$

Recall also known as sensitivity or True Positive Rate (TPR), is calculated as the ratio of true positives to the sum of true positives and false negatives:

$$TPR = \frac{TP}{TP + FN}$$

$$TPR = \frac{240}{240 + 226} \approx \mathbf{51.50\%}$$

ROC AUC is calculated as:

$$AUC = \int_0^1 TPR(FPR^{-1}) dFPR$$

$$AUC = \int_0^1 51.50(30.51^{-1}) d30.51 \approx \mathbf{60.50\%}$$

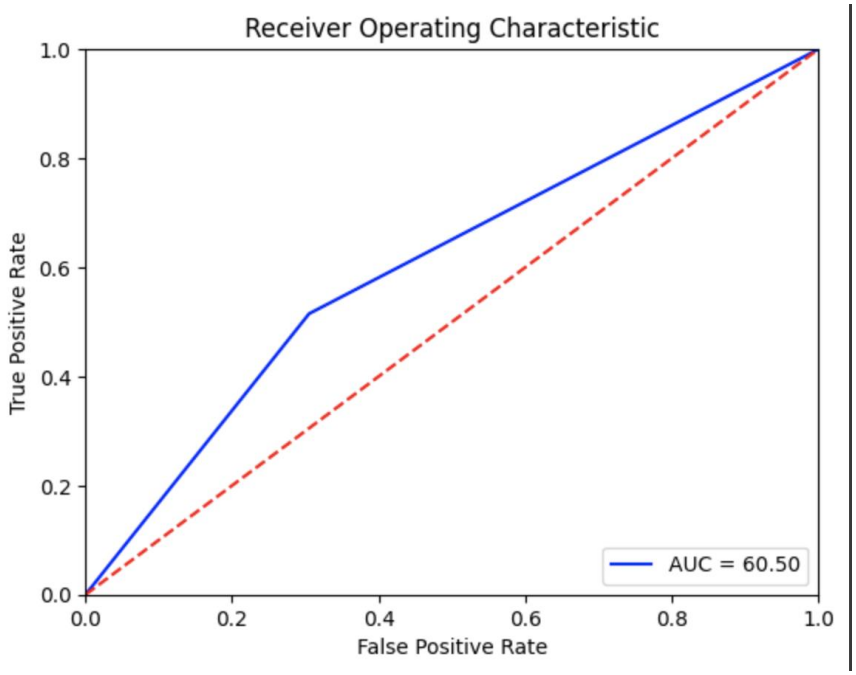

Figure 4.2.2: Experiment 2- AUC for Naive Bayes Classifier

4.2.1 Experiment 2 – Decision Tree:

This experiment applied the decision tree algorithm after the random over sampling technique was applied to balance the dataset. The results are shown below:

Confusion Matrix

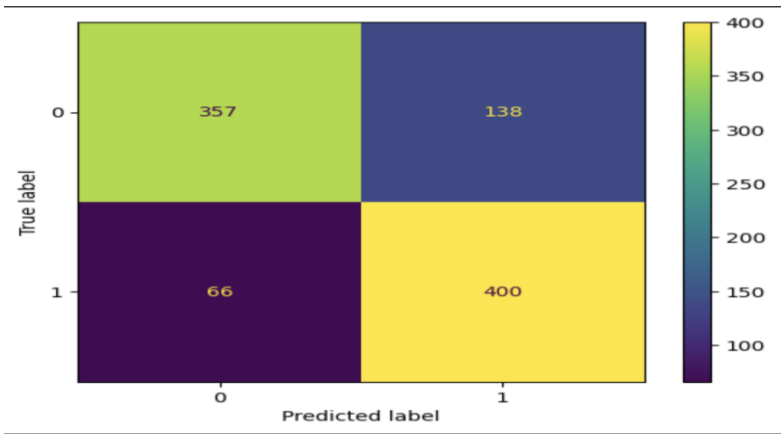

Figure 4.2.3: Confusion Matrix for Experiment 2 - Decision Tree Classifier

Given the confusion matrix: True positives (TP) = 400, True negative (TN) = 357, False negatives (FN) = 66 and False positive (FP) = 138

Accuracy is calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Accuracy = \frac{400 + 357}{400 + 357 + 138 + 66} \approx \mathbf{78.77\%}$$

Precision is calculated as:

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

$$Precision = \frac{400}{400 + 138} \approx \mathbf{74.35\%}$$

Recall also known as sensitivity or True Positive Rate (TPR), is calculated as the ratio of true positives to the sum of true positives and false negatives:

$$TPR = \frac{TP}{TP + FN}$$

$$TPR = \frac{400}{400 + 66} \approx \mathbf{85.84\%}$$

ROC AUC is calculated as:

$$AUC = \int_0^1 TPR(FPR^{-1}) dFPR$$

$$AUC = \int_0^1 85.84(27.88^{-1}) d27.88 \approx \mathbf{78.98\%}$$

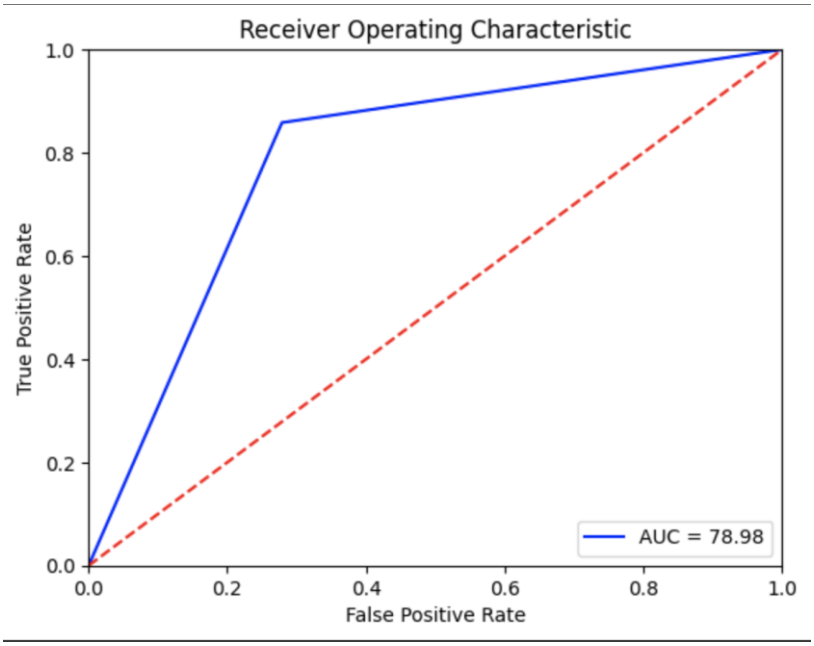

Figure 4.2.4: Experiment 2- AUC for Decision Tree

4.3 Experiment 3 – Gaussian Naïve Bayes:

This experiment applied the naïve bayes algorithm to after the random oversampling techniques was applied to balance the dataset and the MinMax scaler was applied to normalise the values in the dataset. The results are shown below:

Confusion Matrix

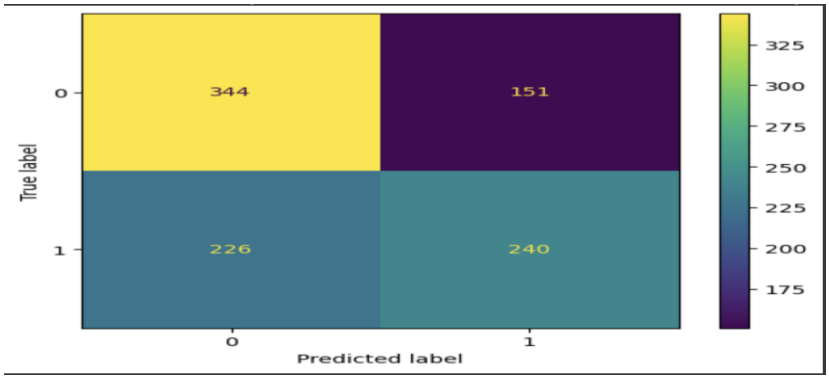

Figure 4.3.1: Confusion Matrix for Experiment 3 - Gaussian Naive Bayes

Given the confusion matrix: True positives (TP) = 240, True negative (TN) = 344, False negatives (FN) = 226 and False positive (FP) = 151

Accuracy is calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Accuracy = \frac{240 + 344}{240 + 344 + 151 + 226} \approx \mathbf{60.77\%}$$

Precision is calculated as:

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

$$Precision = \frac{240}{240 + 151} \approx \mathbf{61.38\%}$$

Recall also known as sensitivity or True Positive Rate (TPR), is calculated as the ratio of true positives to the sum of true positives and false negatives:

$$TPR = \frac{TP}{TP + FN}$$

$$TPR = \frac{240}{240 + 226} \approx \mathbf{51.50\%}$$

ROC AUC is calculated as:

$$AUC = \int_0^1 TPR(FPR^{-1}) dFPR$$

$$AUC = \int_0^1 51.50(30.51^{-1}) d30.51 \approx \mathbf{60.50\%}$$

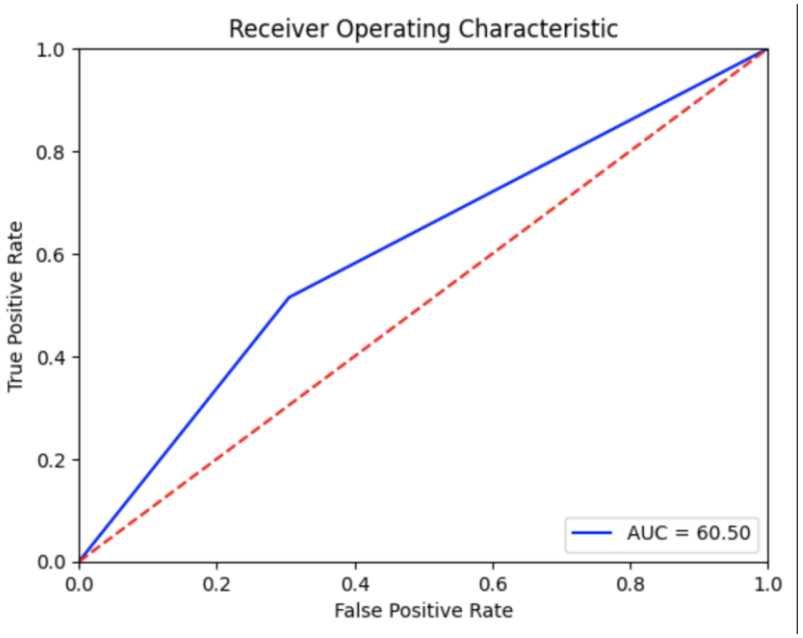

Figure 4.3.2: Confusion Matrix for Experiment 3 - Gaussian Naive Bayes

4.3.1 Experiment 3 – Decision Tree:

This experiment applied the decision tree algorithm to after the random oversampling techniques was applied to balance the dataset and the MinMax scaler was applied to normalise the values in the dataset. The results are shown below:

Confusion Matrix:

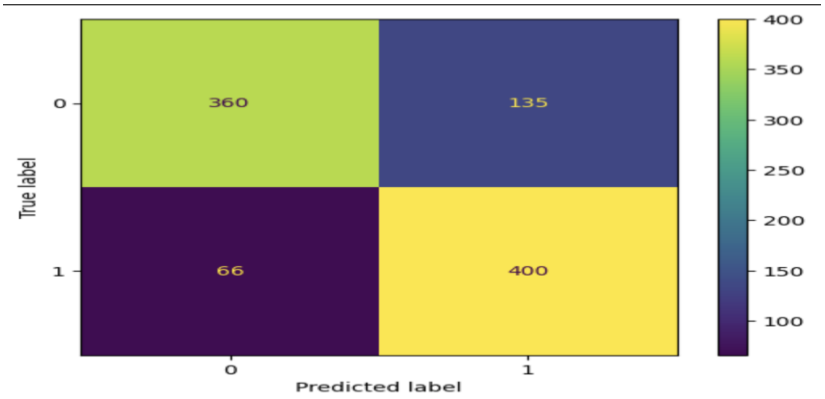

Figure 4.3.3: Confusion Matrix for Experiment 3 - Decision Tree Classifier

Given the confusion matrix: True positives (TP) = 400, True negative (TN) = 360, False negatives (FN) = 66 and False positive (FP) = 135

Accuracy is calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Accuracy = \frac{400 + 360}{400 + 360 + 135 + 66} \approx \mathbf{79.08\%}$$

Precision is calculated as:

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

$$Precision = \frac{400}{400 + 135} \approx \mathbf{74.77\%}$$

Recall also known as sensitivity or True Positive Rate (TPR), is calculated as the ratio of true positives to the sum of true positives and false negatives:

$$TPR = \frac{TP}{TP + FN}$$

$$TPR = \frac{400}{400 + 66} \approx \mathbf{85.84\%}$$

ROC AUC is calculated as:

$$AUC = \int_0^1 TPR(FPR^{-1}) dFPR$$

$$AUC = \int_0^1 85.84(27.27^{-1}) d27.27 \approx \mathbf{79.28\%}$$

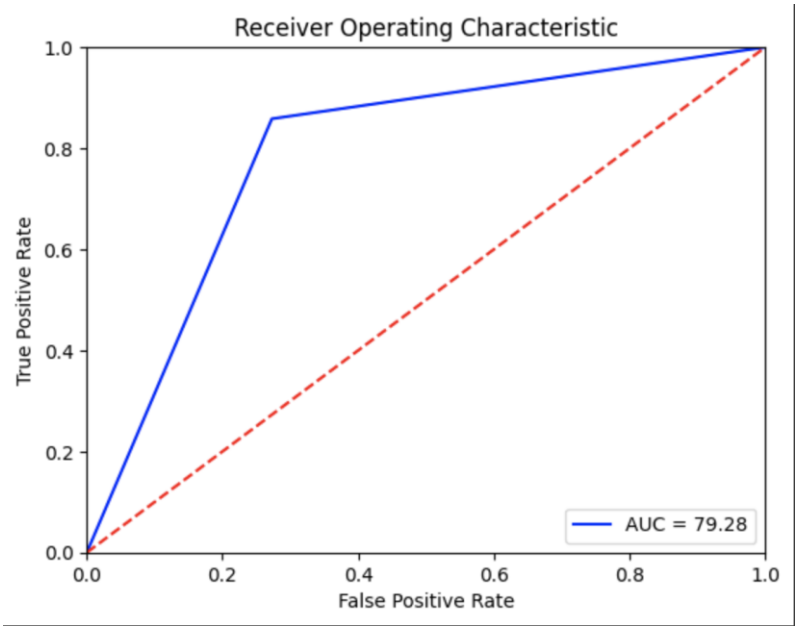

Figure 4.3.4: Experiment 3- AUC for Decision Trees

4.4 Research Analysis

The experiments conducted using Gaussian Naïve Bayes and Decision Tree classifiers across three different scenarios—using an unbalanced dataset, applying random oversampling, and further applying MinMax scaling—provided varied insights into their performance and implications for digital marketing strategies.

Experiment 1: Unbalanced Dataset

Gaussian Naïve Bayes

- Accuracy: 72.05%
- Precision: 46.15%
- Recall (TPR): 16.67%
- ROC AUC: 54.70%

Implications: The relatively high accuracy indicates the model's general effectiveness. However, the low recall suggests that the model struggles to identify true positives, meaning it fails to correctly predict a significant

portion of actual positive cases. In digital marketing, this translates to missed opportunities in targeting potential buyers, as the model cannot effectively identify customers who are likely to make a purchase after seeing an advertisement. The low ROC AUC further highlights its limited ability to differentiate between positive and negative instances.

Further Actions:

- **Enhanced Data Preprocessing**: Focus on balancing the dataset to improve recall and ensure that potential customers are not overlooked.
- **Segmentation and Personalization**: Utilise other methods to better segment and personalize campaigns, as this model alone may not be reliable for high-stakes decision-making.

Decision Tree

- Accuracy: 68.43%
- Precision: 42.16%
- Recall (TPR): 43.33%
- ROC AUC: 60.57%

Implications: The decision tree's recall is higher than that of the Gaussian Naïve Bayes, indicating better performance in identifying actual positives. This is crucial for ensuring potential customers are correctly identified, leading to more effective marketing campaigns. However, the accuracy and precision are moderate, suggesting some room for improvement in reducing false positives and negatives.

Further Actions:

- **Iterative Model Improvement**: Continuously refine the model using additional data and advanced techniques like pruning to improve accuracy and precision.

Experiment 2: Random Oversampling

Gaussian Naïve Bayes

- Accuracy: 60.77%
- Precision: 61.38%
- Recall (TPR): 51.50%
- ROC AUC: 60.50%

Implications: The random oversampling technique significantly improved the recall, indicating a better ability to detect actual positive cases. This suggests that the model is more capable of identifying customers likely to make a purchase, making it more useful for targeting purposes. However, the decrease in accuracy reflects the increased false positives meaning some customers who are unlikely to make are purchase are being predicted to make a purchase. The effect of high false positive on digital marketing would result in very poor Return On Ad Spend (ROAS) – this means companies would be increase their ad spend but realise very little conversion.

Actionable Strategies:

- **Cost-Benefit Analysis**: Conduct a cost-benefit analysis to weigh the benefits of reaching more potential customers against the cost of potentially increased false positives.

Decision Tree

- Accuracy: 78.77%
- Precision: 74.35%
- Recall (TPR): 85.84%
- ROC AUC: 78.98%

Implications: The decision tree model showed significant improvement across all metrics after applying random oversampling. The high recall and ROC AUC indicate the model's strong ability to identify true positives and effectively distinguish between positive and negative instances. This makes it highly effective for digital marketing strategies that rely on accurately predicting customer behaviour.

Actionable Strategies:

- **Highly Effective Campaigns**: Use the model to design highly effective and targeted marketing campaigns with a high likelihood of converting potential customers.
- **Resource Allocation**: Allocate marketing resources more efficiently by focusing on segments identified as high-potential by the model.

Experiment 3: MinMax Scaling and Random Oversampling

Gaussian Naïve Bayes

- Accuracy: 60.77%
- Precision: 61.38%
- Recall (TPR): 51.50%
- ROC AUC: 60.50%

Implications: Applying MinMax scaling did not significantly alter the results compared to Experiment 2. The model's performance metrics remained the same, indicating that scaling did not further enhance the model's predictive power in this case.

Further Actions:

- **Consistency Check**: Ensure the robustness and consistency of preprocessing steps. In this case, MinMax scaling may not add value but is essential for algorithms sensitive to data ranges.
- **Focus on Other Improvements**: Consider other enhancements like feature engineering or more advanced balancing techniques.

Decision Tree

- Accuracy: 79.08%
- Precision: 74.77%
- Recall (TPR): 85.84%
- ROC AUC: 79.28%

Implications: The decision tree classifier showed a slight improvement in accuracy and ROC AUC with MinMax scaling. This indicates that normalizing the data can marginally enhance the model's performance, further solidifying its reliability for predicting customer behaviour.

Actionable Strategies:

- **Scalable Solutions**: Implement the model in scalable marketing solutions, enabling automated, data-driven decisions for large-scale campaigns.

Overall, the decision tree classifier outperformed the Gaussian Naïve Bayes algorithm, especially after applying random oversampling and data normalization techniques. These findings suggest that for digital marketing strategies aiming to predict purchase intent, the decision tree model provides a more robust and reliable tool. It enables more precise targeting and resource allocation, ultimately enhancing the effectiveness of digital marketing campaigns. Future improvements could focus on further refining the decision tree model, exploring additional preprocessing techniques, and expanding the dataset to include more diverse customer interactions and external factors

4.5 Comparison of Results

The plot below shows a side-by-side comparison of the 6 models developed for the purpose of this research.

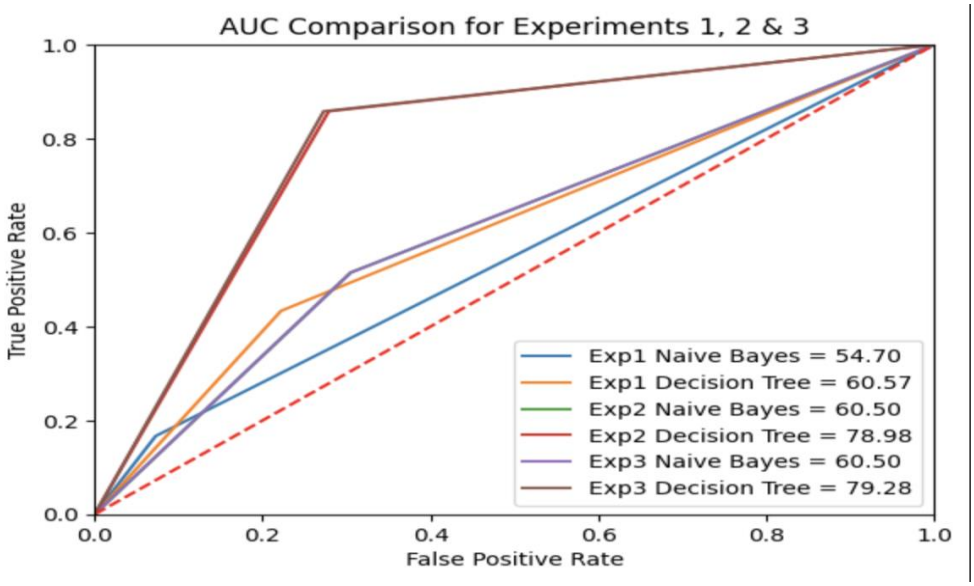

Figure 4.4: Comparison of AUC for Experiments 1-3

The effects of class imbalance on model performance was pronounced in Experiment 1, where despite achieving commendable accuracy rates, the models exhibited a notably poor recall (true positive rates). To address this limitation, Experiment 2 incorporated the random oversampling technique, aiming to rectify the imbalance by augmenting the minority class instances. Notably, both Naïve Bayes and Decision Tree algorithms in Experiment 2 demonstrated marked improvements in recall and ROC AUC values, indicative of their improved ability to correctly identify positive instances. However, it's noteworthy that the Decision Tree algorithm exhibited superior performance compared to Naïve Bayes in Experiment 2, suggesting its efficacy in handling imbalanced datasets.

Subsequently, Experiment 3 delved into the impact of data normalization by applying the MinMax scaler to standardize feature values within a predefined range. Interestingly, while Naïve Bayes yielded exact results

to those of Experiment 2, the Decision Tree algorithm exhibited marginal improvements in Experiment 3. These findings underscore the pivotal role of pre-processing techniques, such as data balancing and normalization, in improving the adverse effects of class imbalance and augmenting model efficacy in predictive analytics.

In a related study, (Narsimlu and Kumar, 2021) reported achieving an impressive accuracy of over 90%. However, it's noteworthy that their model was trained on a dataset five times larger than the one utilized in this study. While the expanded dataset contributed to improved results, it comes with computational burdens, necessitates access to larger data volumes, and extends the model runtime.

This research adds significant value to the existing body of knowledge by highlighting the efficacy of pre-processing steps in enhancing model performance. Through meticulous data cleaning, balancing, and normalization techniques, our study showcases notable improvements in model accuracy, precision, and recall. This underscores the importance of careful data pre-processing, particularly in addressing class imbalance issues prevalent in digital marketing datasets.

5.0 Discussion & Conclusion

In summary, the iterative nature of these experiments explains the multifaceted relationship between pre-processing methodologies and model performance. It emphasises the significance of techniques that mitigate challenges posed by imbalanced datasets. Such insights not only inform best practices in data-driven decision-making but also contribute to the advancement of predictive modelling techniques in digital marketing. Results from this research can be applied to the following areas:

1. Custom Bid Algorithms on Google and Meta: The trained model can be used to segment audiences based on predicted purchase intent, allowing marketers to identify high-intent users and tailor their campaigns more effectively. In addition, by targeting users who are more likely to convert, marketers can implement custom bidding strategies that adjust bids based on the likelihood of purchase. For example, higher bids can be set for users with a high predicted purchase intent, ensuring that ads are shown to the most valuable audience segments. This option of custom bid algorithm is available on most Ads Serving platforms such as Google's DV360 and Google Search Ads, and Meta's Facebook and Instagram.
2. Creative Rotations on Social Media Platforms: The models developed in this research can be effectively applied to creative rotations in social media ads by leveraging predicted purchase intent to optimize ad content delivery. By segmenting audiences based on their likelihood to purchase, marketers can dynamically adjust the creative elements of their ads to better resonate with different user groups. For instance, high-intent users can be shown ads that emphasize urgency and strong calls-to-action, while lower-intent users might receive ads focused on brand awareness and engagement. This targeted approach ensures that the most relevant

and compelling ad creatives are displayed to each segment, enhancing engagement and conversion rates across social media platforms.

Ultimately, the application of predictive analytics in digital marketing offers substantial benefits, such as improved targeting and personalization of advertisements, leading to enhanced consumer engagement and higher conversion rates. However, these advancements bring significant ethical considerations, particularly regarding consumer privacy and data protection.

Ethical Concerns

1. Data Collection and Consent: Predictive analytics relies heavily on the collection of vast amounts of consumer data, including personal information, browsing behaviour, purchase history, and demographic details. Ensuring that consumers are aware of what data is being collected and obtaining their explicit consent is crucial. Transparent data collection practices help build trust and maintain ethical standards.
2. Anonymisation and De-identification: To further protect consumer privacy, data used for predictive analytics should be anonymized or de-identified. This process involves removing or altering personal identifiers so that individuals cannot be readily identified from the data. Anonymization helps mitigate privacy risks while still allowing valuable insights to be derived from the data.
3. Avoiding Discrimination/ Biases: Predictive models must be carefully designed to avoid discriminatory practices. For example, biases in the data can lead to unfair targeting or exclusion of certain consumer groups. Ensuring diversity in training data and regularly

auditing models for bias can help mitigate these risks and promote fairness.

Limitations of the study

1. Data Availability and Quality: The effectiveness of machine learning algorithms heavily relies on the quality and quantity of data used for training. Consequently, the limited access to comprehensive customer data and data inconsistencies, due to privacy regulations such as GDPR and business sensitivity, presents a significant limitation to the scope of this research. Moreover, the dataset used does not capture all relevant factors influencing purchase intent, such as external market conditions or individual preferences not accounted for in the data.
2. Model Generalisation: Another limitation lies in the generalization of the developed models. While the study may achieve high accuracy within the confines of the dataset, there is uncertainty about the models' performance on unseen data or in different market contexts. There is need, ultimately, for the model to be test against new real-world data, different from what it was trained on.
3. Time Constraint: This was a significant limitation in this research. The duration of the study was limited, which restricted the ability to experiment with a wider range of machine learning algorithms and to conduct extensive hyperparameter tuning. Additionally, the time available for data preprocessing and feature engineering was limited, which may have affected the quality and completeness of the input data. A longer research period would have allowed for a more thorough exploration of different modelling approaches and refinement of the predictive models, potentially leading to better performance and more robust findings.

Improvement Suggestions

1. Enhanced Data Collection: Future research endeavours could benefit from enhanced data collection efforts, aiming to gather more comprehensive and diverse datasets. Collaboration with industry partners or access to larger datasets through data-sharing agreements could provide a richer source of information, enabling more robust model training and evaluation. Additionally, efforts to ensure data quality and consistency, such as thorough data cleaning and validation processes, can enhance the reliability of the findings.
2. Model Generalisation: To address concerns regarding model generalisation, future studies could incorporate validation on external datasets or conduct real-world experiments to assess model performance in varied market conditions. Employing techniques such as cross-validation or out-of-sample testing can help gauge the models' robustness and reliability across different contexts. Additionally, ensemble techniques or model stacking approaches could be explored to combine the strengths of multiple models and improve overall predictive performance.

By demonstrating the substantial impact of pre-processing steps on model outcomes, this research offers valuable insights for the digital marketing industry. These findings emphasize the significance of deploying optimized pre-processing pipelines to enhance predictive analytics in marketing campaigns. Moreover, this study underscores the potential for implementing efficient pre-processing strategies to maximize returns on marketing investments and improve overall campaign performance in the digital landscape.

References

Azevedo, A., Santos, M.F., 2008. KDD, SEMMA and CRISP-DM: a parallel overview. IADS-DM.

Beck, J.E., 2005. Engagement tracing: using response times to model student disengagement 8.

Bekmamedova, N., Shanks, G., 2014. Social media analytics and business value: a theoretical framework and case study, in: 2014 47th Hawaii International Conference on System Sciences. IEEE, pp. 3728–3737.

Bist, A.S., Agarwal, V., Aini, Q., Khofifah, N., 2022. Managing Digital Transformation in Marketing:" Fusion of Traditional Marketing and Digital Marketing". International Transactions on Artificial Intelligence 1, 18–27.

Bounsaythip, C., Rinta-Runsala, E., 2001. Overview of data mining for customer behavior modeling. VTT Information Technology Research Report, Version 1, 1–53.

D. Newton, J., J. Newton, F., Turk, T., T. Ewing, M., 2013. Ethical evaluation of audience segmentation in social marketing. European Journal of Marketing 47, 1421–1438.

Du, W., Zhan, Z., 2002. Building decision tree classifier on private data.

García, S., Luengo, J., Herrera, F., 2015. Data preprocessing in data mining. Springer.

Garcia, S., Parmisano, A., Erquiaga, M.J., 2020. IoT-23: A labeled dataset with malicious and benign IoT network traffic. https://doi.org/10.5281/ZENODO.4743746

Jahromi, A.H., Taheri, M., 2017. A non-parametric mixture of Gaussian naive Bayes classifiers based on local independent features, in: 2017 Artificial Intelligence and Signal Processing Conference (AISP). IEEE, pp. 209–212.

Jeffery, M., 2010. Data-driven marketing: the 15 metrics everyone in marketing should know. John Wiley & Sons.

Khodabandehlou, S., 2019. Designing an e-commerce recommender system based on collaborative filtering using a data mining approach. International Journal of Business Information Systems 31, 455–478.

Kim, B.-D., Kim, S.-O., 2001. A new recommender system to combine content-based and collaborative filtering systems. Journal of Database Marketing & Customer Strategy Management 8, 244–252.

Kingsnorth, S., 2022. Digital marketing strategy: an integrated approach to online marketing. Kogan Page Publishers.

Kotane, I., Znotina, D., Hushko, S., 2019. Assessment of trends in the application of digital marketing. Scientific Journal of Polonia University 33, 28–35.

Leeflang, P.S., Verhoef, P.C., Dahlström, P., Freundt, T., 2014. Challenges and solutions for marketing in a digital era. European management journal 32, 1–12.

Lemos, A.M.F., 2015. Optimizing multi-channel use in digital marketing campaigns. Universidade Catolica Portuguesa (Portugal).

Murgai, A., 2018. Transforming digital marketing with artificial intelligence. International Journal of Latest Technology in Engineering, Management & Applied Science 7, 259–262.

Narsimlu, M., Kumar, M.S., 2021. Futuristic Research on Digital Marketing Data to Identify Leads using Predictive Analytics. International Journal of Management (IJM) 12, 337–345.

Olson, D.L., Chae, B.K., 2012. Direct marketing decision support through predictive customer response modeling. Decision Support Systems 54, 443–451.

Romero Leguina, J., Cuevas Rumín, Á., Cuevas Rumín, R., 2020. Digital marketing attribution: Understanding the user path. Electronics 9, 1822.

Sayyad, S., Mohammed, A., Shaga, V., Kumar, A., Vengatesan, K., 2020. Digital Marketing Framework Strategies Through Big Data, in: Proceeding of the International Conference on Computer Networks, Big Data and IoT (ICCBI-2018). Springer, pp. 1065–1073.

Selvin, S., Vinayakumar, R., Gopalakrishnan, E.A., Menon, V.K., Soman, K.P., 2017. Stock price prediction using LSTM, RNN and CNN-sliding window model, in: 2017 International Conference on Advances in Computing, Communications and Informatics (Icacci). IEEE, pp. 1643–1647.

Sinha, R., 2018. A comparative analysis of traditional marketing vs digital marketing. Journal of Management Research and Analysis 5, 234–243.

Theodoridis, P.K., Gkikas, D.C., 2019. How artificial intelligence affects digital marketing, in: Strategic Innovative Marketing and Tourism: 7th ICSIMAT, Athenian Riviera, Greece, 2018. Springer, pp. 1319–1327.

Vujović, Ž., 2021. Classification model evaluation metrics. International Journal of Advanced Computer Science and Applications 12, 599–606.

Xu, J., Lee, K., Li, W., Qi, H., Lu, Q., 2015. Smart pacing for effective online ad campaign optimization, in: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 2217–2226.