**A black cover with white text

Description automatically generated**

# Abstract

This report utilises machine learning to explore customer behaviour in the multi-channel retail domain. It applies K-Means clustering for customer segmentation and evaluates these segments with silhouette scores for targeted marketing strategies.

Supervised learning models are used to predict marketing campaign effectiveness and customer churn. Models like Random Forest and ensemble methods are examined, emphasising the balance between accuracy and business applicability. A quintile-based Linear Regression model is emphasised for reliable churn prediction, compared to a less effective continuous model.

The report concludes with recommendations for future research, including granular data analysis, residual analysis, and advanced regularisation techniques, showcasing the role of machine learning in enhancing customer engagement and guiding retail strategies.

# Table of Contents

[Abstract 2](#_Toc156065307)

[Table of Contents 3](#_Toc156065308)

[List of Tables 5](#_Toc156065309)

[List of Figures 5](#_Toc156065310)

[Introduction 7](#_Toc156065311)

[Domain Description 8](#_Toc156065312)

[Relevance of the Dataset to the Domain 8](#_Toc156065313)

[Importance of Machine Learning in the Domain 8](#_Toc156065314)

[Problem Definition and Methodology 9](#_Toc156065315)

[1. Customer Segmentation (Unsupervised Learning) 9](#_Toc156065316)

[2. Campaign Effectiveness Prediction (Supervised Learning) 9](#_Toc156065317)

[3. Churn Prediction (Supervised Learning) 10](#_Toc156065318)

[Data Set Description 11](#_Toc156065319)

[Data Source 11](#_Toc156065320)

[Initial Data Assessment 11](#_Toc156065321)

[Characteristics of the Dataset 14](#_Toc156065322)

[Justification for Dataset Choice 14](#_Toc156065323)

[Data Set Exploration 15](#_Toc156065324)

[Preprocessing 15](#_Toc156065325)

[Exploratory Data Analysis (EDA) 15](#_Toc156065326)

[Correlation Matrix 15](#_Toc156065327)

[Outliers 16](#_Toc156065328)

[Density Plots for Key Features 17](#_Toc156065329)

[Grouped Bar Plot for Marketing Campaign Acceptance 17](#_Toc156065330)

[Experiments and Evaluations 18](#_Toc156065331)

[Customer Segmentation (Unsupervised Learning) 18](#_Toc156065332)

[Objective and Feature Engineering 18](#_Toc156065333)

[K-Means Clustering 19](#_Toc156065334)

[Evaluation Metrics 20](#_Toc156065335)

[Influence of PCA and Standard Scaler 20](#_Toc156065336)

[Campaign Effectiveness Prediction (Supervised Learning) 21](#_Toc156065337)

[Reversion to the Original Data 21](#_Toc156065338)

[Objective of Supervised Learning 21](#_Toc156065339)

[Data Preparation and Processing 21](#_Toc156065340)

[Model Training and Evaluation 21](#_Toc156065341)

[Handling Imbalanced Data 23](#_Toc156065342)

[Ensemble Learning 25](#_Toc156065343)

[Churn Prediction (Supervised Learning) 26](#_Toc156065344)

[Supervised Learning Objective and Feature Engineering 26](#_Toc156065345)

[Model Development and Evaluation 26](#_Toc156065346)

[Visualisation of Predictive Model Accuracy on Holdout Data 27](#_Toc156065347)

[Segment-wise Analysis of Predicted Versus Actual Churn Days 28](#_Toc156065348)

[Cross-Validation of Models 29](#_Toc156065349)

[Ridge and Lasso Regression Analysis 29](#_Toc156065350)

[Analysis and Results 30](#_Toc156065351)

[Unsupervised Learning – Customer Segmentation 30](#_Toc156065352)

[Supervised Learning - Campaign Effectiveness Prediction 30](#_Toc156065353)

[Supervised Learning - Churn Prediction 30](#_Toc156065354)

[Business Implications 30](#_Toc156065355)

[Conclusion 31](#_Toc156065356)

[Summary of Key Findings 31](#_Toc156065357)

[Implications of the Study 31](#_Toc156065358)

[Limitations of the Study 31](#_Toc156065359)

[Recommendations for Future Research 31](#_Toc156065360)

[Reference List 32](#_Toc156065361)

[Appendices 35](#_Toc156065362)

[Link to Full Code and Comments in Colaboratory 35](#_Toc156065363)

[Link to Kaggle Dataset 35](#_Toc156065364)

[Additional Figures 35](#_Toc156065365)

# List of Tables

Table 1. Data Set Description

# List of Figures

Fig. 1. Meta-Data Table

Fig. 2. Removal of Redundant Columns

Fig. 3. Duplicates

Fig. 4. Negative Values

Fig. 5. Cleaned Data Frame Dimensions

Fig. 6. Correlation Matrix

Fig. 7. Outliers

Fig. 8. Density Plots for Key Features

Fig. 9. Grouped Bar Plots for Marketing Campaign Acceptance

Fig. 10. Feature Engineering – Affluent Enthusiasts

Fig. 11. Feature Engineering – Family-Focused Budget Shoppers

Fig. 12. Feature Engineering – Digital Natives and Online Shoppers

Fig. 13. 2D Clusters

Fig. 14. 3D Clusters

Fig. 15. The Elbow Method

Fig. 16. Silhouette Score

Fig. 17. Campaign Effectiveness Prediction – Data Preparation and Processing

Fig. 18. Campaign Effectiveness Prediction – Model Training and Evaluation

Fig. 19. Classification Reports and Confusion Matrixes of Models

Fig. 20. Handling Imbalanced Data

Fig. 21. RandomOverSampler

Fig. 22. Classification Reports and Confusion Matrixes of Models on Oversampled Data

Fig. 23. The Ensemble Learning Approach

Fig. 24. Churn Prediction – RFM Feature Engineering

Fig. 25. Churn Prediction – RFM Feature Engineering and Model Logic

Fig. 26. Quantile-Based and Continuous Models Output

Fig. 27. Visualisation of Predictive Model Accuracy on Holdout Data

Fig. 28. Segment-wise Analysis of Predicted Versus Actual Churn Days

Fig. 29. Cross-Validation Models

Fig. 30. Ridge and Lasso

# Introduction

Artificial Intelligence (AI) is increasingly crucial in business, offering revolutionary changes in operations. It enhances automation, productivity, and decision-making through advanced data analysis, leading to improved customer experiences and operational efficiency based on outputs from cognitive technologies (Zharovskikh, 2022). As of 2023, the AI market is valued between $136.55 billion (Thormundsson, 2023) to $207.9 billion, with projections of significant growth to $3,636 billion by 2033 (Grand View Research, 2022). In the UK, approximately 15% of businesses have adopted AI technologies, a trend expected to continue rising (GOV.UK, 2022).

The multi-channel retail environment, integral to the digital economy, combines various sales and communication channels for a unique shopping experience. Customers engage and shop through multiple platforms like physical stores, websites, and mobile apps (Crawley, 2021). Accelerated by the Covid-19 pandemic, this sector has seen rapid growth and transformation, influencing customer habits and innovating business models. The integrated retail approach has significantly altered shopping behaviours and transformed business-customer interactions (Briedis et al., 2023).

The aims and objectives of this report are centred around employing advanced analytical techniques to examine and interpret large-scale, multi-layered data from various retail channels using Python programming. Using both supervised and unsupervised machine learning techniques, this report aims to provide detailed insights into customer behaviour preferences and interactions across different retail channels. The goal of this report is to leverage these insights to provide the business with the ability to make informed, data-driven decisions, providing strategic vision and information that can shape their future in an ever-evolving multi-channel retail landscape.

# Domain Description

This report focuses on the multi-channel retail domain, where businesses interact with customers through various channels, each offering unique customer experiences and engagement opportunities. This domain includes traditional retail stores, online marketplaces, social media platforms and mobile applications. It emphasises the importance of seamless shopping experience across all channels, catering to the diverse preferences and behaviours of customer (Iglesias-Pradas and Acquila-Natale, 2023).

## Relevance of the Dataset to the Domain

The dataset pertinent to this study encapsulates a broad spectrum of customer engagement across multiple retail channels. It includes detailed information ranging from in-store interactions to online purchasing patterns, covering various aspects such as customer demographics, transaction histories, and responses to marketing efforts across different channels. This rich dataset is crucial in understanding the complex and evolving nature of customer behaviour in a multi-channel retail setting.

## Importance of Machine Learning in the Domain

Machine learning is fundamental in the multi-channel retail environment, enabling analysis of complex datasets for actionable insights and enhancing customer experience and operational efficiency (Mitra et al., 2022). It also enables retailers to identify patterns in customer behaviour, facilitating personalised marketing strategies and improved customer experiences across different platforms. Techniques such as predictive analytics for customer behaviour, segmentation algorithms for personalised marketing, and churn prediction models are invaluable in this domain.

# Problem Definition and Methodology

This section tackles three machine learning problems using algorithms tailored to each challenge. The selection was based on the learning type required and the unique aspects of each problem, ensuring an accurate application of machine learning to derive insightful solutions.

## 1. Customer Segmentation (Unsupervised Learning)

**Objective:** The overall goal of this unsupervised learning model is to effectively segment the customer base into meaningful groups based on their purchasing behaviours and demographic characteristics.

**Target Variable:**

* ‘Affluent Enthusiasts’ = High 'Income', 'MntWines', 'MntMeatProducts', 'MntFishProducts', 'MntSweetProducts', 'MntGoldProds', and 'MntTotal'; Low 'NumDealsPurchases'
* ‘Family-Focused Budget Shoppers’ = High 'Kidhome', 'Teenhome', 'NumDealsPurchases'; Moderate to low 'MntWines', 'MntFruits', 'MntMeatProducts', 'MntFishProducts', 'MntSweetProducts', 'MntGoldProds'
* Digital Natives and Online Shoppers = High 'NumWebPurchases', 'NumWebVisitsMonth', 'AcceptedCmp3', 'AcceptedCmp4', 'AcceptedCmp5'; Potentially younger age

**Model Justification:**

K-Means Clustering: To identify homogeneous groups within the customer base by finding the optimal partitioning of data (Melanie, 2023).

Silhouette Score: To evaluate the quality of clustering and ensure that each segment is distinct and internally cohesive (Scikit-learn, 2019).

Elbow Method: To determine the optimal number of clusters, thus balancing the granularity of segmentation with practical applicability (Saji, 2021).

Principal Component Analysis (PCA): To reduce the dimensionality of while retaining most of the significant variability in the data (Jaadi, 2019).

Standard Scaler: To normalise the feature set, ensuring each attribute contributes equally to the analysis (Scikit-Learn, 2019).

## 2. Campaign Effectiveness Prediction (Supervised Learning)

**Objective:** The primary objective is to predict whether a customer will accept an offer in a future marketing campaign. To achieve this, a large set of customer attributes and historical interaction data have been leveraged, serving as a foundation for identifying patterns and trends. Supervised learning has been chosen for this task as it excels in situations where the goal is to predict future occurrences based on past data (Jiang, Gradus, and Rosellini, 2020).

**Target Variable:** ‘Response’ is the target variable indicating whether a customer accepted an offer in past marketing campaigns. This binary variable provides a clear outcome for each customer where the goal is to predict a specific result based on input features.

**Model Justification:**

Logistic Regression: known for its simplicity and interpretability, served as a baseline, offering quick insights with reasonable accuracy (Jain, 2023).

Decision Trees and Random Forests: crucial for capturing non-linear relationships, with the latter providing enhanced generalisation through its ensemble approach (McClarren, 2021).

Gradient Boosting and AdaBoost: both boosting techniques, sequentially improved weak learners, focusing on misclassified instances in prior iterations, thereby enhancing the model's accuracy (Kumar, 2020).

K-Nearest Neighbours (KNN): leveraged similarity metrics to make predictions, offering a different perspective based on proximity in the feature space (IBM, 2023b).

Gaussian Naive Bayes: with its assumption of feature independence, provided a fast, baseline probabilistic approach (Oleszak, 2023).

Support Vector Machine (SVM): with its kernel trick, was adept at handling high-dimensional space (S, 2021).

XGBoost: was included for its efficiency and effectiveness in handling diverse datasets (XGBoost developers, 2022).

Voting Classifier: capitalised on their collective strengths, mitigating individual weaknesses, and offering a robust, well-rounded predictive tool (Medium, 2023).

This ensemble approach, especially after addressing class imbalance, proved highly effective in accurately predicting customer responses to future marketing campaigns.

## 3. Churn Prediction (Supervised Learning)

**Objective:** To forecast the number of days until customers are likely to end their association with the company (churn). By pinpointing those at risk of churning, pre-emptive measures to retain them can be implemented in the business.

**Target Variable:** The target variable in this study is 'Days Until Churn', which is predicted based on the RFM (Recency, Frequency, Monetary Value) model. 'Recency' is directly obtained from the dataset, 'Frequency' is engineered from the sum of NumWebPurchases, NumCatalogPurchases, NumDealsPurchases and NumStorePurchases, and 'Monetary' is derived from the 'MntTotal' feature.

**Model Justification:**

Linear Regression: To understand the relationship between the target variable ‘Days Until Churn’ and independent variables. Serves as foundation model (Mali, 2021).

Quantile-based Model: This approach divides the data into equal-sized segments based on rank-ordered values, making it particularly useful in handling skewed distributions and identifying patterns within specific segments of a dataset.

Continuous Model: Uses the actual numeric values of the data, aiming to leverage the full spectrum of information available in each variable.

Cross-Validation: Ensures the model’s robustness and generalisability across different data subsets.

Ridge Regression: Addresses overfitting by penalising large coefficients (L2 regularisation) (Scikit-learn, 2023).

Lasso Regression: Enhances model simplicity and interpretation by reducing some coefficients to zero (L1 regularisation) (Scikit-learn, 2023b).

# Data Set Description

## Data Source

The dataset analysed in this study was obtained from a publicly available source, Kaggle, a popular platform for data science and machine learning projects. This dataset is specifically tailored for marketing analytics, making it invaluable resource for understanding customer behaviour and preferences in a multi-channel retail context.

## Initial Data Assessment

The dataset contains 2,205 instances (records) and 39 features (attributes). It is notable that the features in the dataset do not align with those described in the Kaggle description card (Fig. 1), suggesting either updates or variations in the dataset’s version. The dataset's actual features are listed below (Table 1), along with the presumptive descriptions derived from the column names. These are presented in the exact same order as they appear in the CSV file.

A list of products with text

Description automatically generated with medium confidence

A list of products with text

Description automatically generated with medium confidenceA close-up of a form

Description automatically generated

A close-up of a form

Description automatically generated

*Table 1. Data Set Description*

## Characteristics of the Dataset

The dataset includes a wide range of attributes, from more specific metrics related to purchasing behaviour across multiple product categories to more fundamental demographic data such as age and marital status. Data on website visits, customer complaints, and customer interactions with marketing efforts are also included. This volume of information offers a thorough understanding of customer profiles and how they interact with the brand and its marketing initiatives. The dataset’s combination of Booleans, float and integers allows for a multidimensional analysis method that considers both the categorical and quantitative aspects of customer behaviour, enabling the use of several machine learning techniques.

## Justification for Dataset Choice

The dataset was chosen for its relevance to the selected domain of multi-channel retail and its stability for addressing the three identified machine learning problems – Campaign Effectiveness Prediction, Customer Segmentation, and Churn Prediction. The variety of features allows for an in-depth analysis of customer behaviour and the application of both supervised and unsupervised machine learning techniques.

# Data Set Exploration

## Preprocessing

Data preprocessing is a crucial stage in data analysis, focusing on cleaning, transforming, and streamlining data. It involves rectifying errors, inconsistencies, and inaccuracies to ensure data accuracy, completeness, and consistency. This process is vital for reliable decision-making, as it enhances data manageability and accuracy (Miller, 2019). This report outlines a preprocessing approach that ensures the integrity and quality of the data set for subsequent analyses.

Key reprocessing steps included the removal of redundant columns, ‘Z\_Cost Contact’ and ‘Z\_Revenue’, reducing the feature set to 37 (Fig. 2). This refinement was crucial in maintaining focus on the most relevant data. Furthermore, 184 duplicate entries, which can skew analysis, were identified, and removed (Fig. 3). This action refined the dataset from 2205 to 2021 instances, enhancing its authenticity.

The detection and exclusion of records with negative values in the 'MntRegularProds' column was another critical step. This measure addressed minor inconsistencies in the dataset, ensuring its consistency and validity (Fig. 4). Post these rigorous steps, the dataset was reduced to 2018 instances, each accurately representing a customer profile without redundancy or erroneous data (Fig. 5).

The entire preprocessing stage transformed the raw dataset into 'cleaned\_data.csv', a polished resource ready for insightful analysis.

## Exploratory Data Analysis (EDA)

EDA is integral in uncovering underlying patterns and relationships, which are essential for informed decision-making in marketing strategies (IBM, 2023a).

### Correlation Matrix

The foundation of this analysis was the creation of a correlation matrix (Fig. 6) visualised using the “Mako” colour palette. This large-scale, detailed heatmap provided a comprehensive view of the interdependencies between features and highlighted potential collinearities and valuable relationships that are important for predictive modelling. The diagonal running from the top left to the bottom right consists of 1s, which represented the correlation of each variable with itself (perfect correlation). This matrix was used to identify potential relationships between variables that might warrant further analysis.

A blue and white chart

Description automatically generated with medium confidence

*Fig. 6. Correlation Matrix*

### Outliers

Further, a series of boxplots for each feature played a pivotal role in the identification of outliers. In the context of marketing data, these outliers were deemed crucial as they represented actual customer behaviour. The strategic decision to retain these outliers ensured that all market segments, including atypical behaviours, were adequately considered in future campaign strategies (Fig. 7).

A group of blue squares

Description automatically generated

*Fig. 7. Outliers*

### Density Plots for Key Features

Density plots for key features like 'Customer\_Days', 'Age', 'Income', 'MntTotal', and 'Complain' offered deeper insights into the distribution of these variables. Using the same 'mako' colour palette, these plots helped in understanding the variability and skewness in the data, revealing the demographic and spending patterns of the customer base (Fig. 8).

A graph of a diagram

Description automatically generated with medium confidence

*Fig. 8. Density Plots for Key Features*

### Grouped Bar Plot for Marketing Campaign Acceptance

Analysis of campaign acceptance revealed significant variations in success across different campaigns. Data were compiled into a new DataFrame to calculate the customer acceptance for each campaign, resulting in a grouped bar chart. The chart efficiently showed the varying impact of campaigns, highlighting Campaign 4 as particularly successful. While the overall campaign success was positive, individual response rates were lower than the cumulative campaign performance (Fig. 9).

A graph showing a bar graph

Description automatically generated with medium confidence

*Fig. 9. Grouped Bar Plots for Marketing Campaign Acceptance*

# Experiments and Evaluations

## Customer Segmentation (Unsupervised Learning)

### Objective and Feature Engineering

The objective of the unsupervised learning model detailed in this report was to segment the customer base into meaningful groups. These groups, or clusters, were based on purchasing behaviours and demographic characteristics. This segmentation aimed to identify distinct types of customers, such as 'Affluent Enthusiasts', 'Family-Focused Budget Shoppers', and 'Digital Natives and Online Shoppers'. Feature engineering (Fig. 10, 11, and 12) crafted new variables to reflect each segment's spending habits and preferences, such as weighting spending levels and online engagement while considering demographics like age.

A screenshot of a computer program

Description automatically generated

*Fig. 10. Feature Engineering – Affluent Enthusiasts*

A screenshot of a computer screen

Description automatically generated

*Fig. 11. Feature Engineering – Family-Focused Budget Shoppers*

A screen shot of a computer program

Description automatically generated

*Fig. 12. Feature Engineering – Digital Natives and Online Shoppers*

### K-Means Clustering

The K-Means clustering algorithm was selected for its efficacy in identifying distinct groups within a dataset (Mannor et al., 2011). For this model, three clusters were predetermined to reflect the three customer segments of interest, despite the Elbow Method suggesting an optimal cluster number of two. This decision was informed by domain knowledge and the business context, which necessitated a tripartite segmentation.

Clusters were visualised through both 2D (Fig. 13) and 3D (Fig. 14) scatter plots, with a custom colour scale to distinguish between them. While 2D plots offer simplicity, they lack the capacity to represent the complexity of multi-dimensional data. The 3D plots, on the other hand, provided a more nuanced view of the customer segments in the context of the engineered features (Tian et al., 2021).

A graph with blue dots

Description automatically generated

*Fig. 13. 2D Clusters*

A graph showing a diagram

Description automatically generated with medium confidence

*Fig. 14. 3D Clusters*

### Evaluation Metrics

The Sum of Squared Errors (SSE) was calculated across a range of cluster numbers to assess the quality of clustering, with the Elbow Method guiding the selection of an optimal cluster count (Fig. 15). However, the second derivative of the SSE indicated a preference for two clusters, which was ultimately overruled due to the business requirement of three distinct customer profiles.

A graph with a line

Description automatically generated

*Fig. 15. The Elbow Method*

The silhouette score, a metric used to evaluate the coherence of clusters (Yadav, 2023), initially suggests a moderately strong clustering structure with a score of 0.55 (Fig. 16). This score represented a balance between cohesion with clusters and separation between them.

### Influence of PCA and Standard Scaler

An experiment to assess the influence of PCA and Standard Scaler on clustering revealed a decrease in the silhouette score from 0.55 to 0.39. This suggested that while PCA is a powerful tool for dimensionality reduction, its application in this context may have led to the loss of significant variance or distortion of cluster structures that are crucial for meaningful customer segmentation.

## Campaign Effectiveness Prediction (Supervised Learning)

### Reversion to the Original Data

In preparation for supervised learning, this analysis reverted to the original, cleaned DataFrame. The dimensions of the DataFrame were reaffirmed, ensuring that the data structure was intact for model training and evaluation.

### Objective of Supervised Learning

The primary goal of this supervised learning approach was to predict customer acceptance of future marketing campaign offers. This predictive task is crucial for optimising marketing strategies and resource allocation.

### Data Preparation and Processing

Missing values were imputed with feature means, and the dataset was divided into features and the 'Response' target variable, followed by a train-test split for model validation. Feature scaling standardised the data. No missing values remained, but correlations exceeding 0.8 indicated possible multicollinearity, potentially impacting models like logistic regression that require feature independence (Fig. 17).

### Model Training and Evaluation

A suite of models, including Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, AdaBoost, KNeighbors, GaussianNB, SVC, and XGBoost, were trained and evaluated. The performance of each model was assessed based on accuracy, with Logistic Regression, Gradient Boosting, and AdaBoost performing best, achieving an accuracy of approximately 87.6% (Fig. 18).

Classification reports and confusion matrices were generated for each model to provide deeper insights into their performance. The reports revealed varying levels of precision, recall, and F1-score across models, indicating differences in their ability to predict customer responses accurately (Fig. 19).

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated  
A screenshot of a computer

Description automatically generated

*Fig. 19. Classification Reports and Confusion Matrixes of Models*

### Handling Imbalanced Data

Recognising the potential bias introduced by class imbalance (Fig. 20), RandomOverSampler (Fig. 21) was employed to balance the classes in the training data (Brownlee, 2020). This technique improved the minority class representation, potentially enhancing model performance for detecting positive responses.

A graph showing a number of classes

Description automatically generated

*Fig. 20. Handling Imbalanced Data*

A graph showing a number of classes

Description automatically generated

*Fig. 21. RandomOverSampler*

Subsequent model training on the oversampled data revealed changes in accuracy, with the Random Forest classifier showing the best performance at approximately 86.1% (Fig. 22). ROC and AUC were used to further analyse the results. However, the improvement in accuracy must be weighed against the increased likelihood of overfitting due to the synthetic oversampling of the minority class.

A screen shot of a computer

Description automatically generatedA screenshot of a computer screen

Description automatically generatedA screenshot of a computer screen

Description automatically generated  
A screenshot of a computer screen

Description automatically generated

*Fig. 22. Classification Reports and Confusion Matrixes of Models on Oversampled Data*

### Ensemble Learning

An ensemble approach (Fig. 23) was adopted by creating a Voting Classifier comprising all individual models. This ensemble method aimed to combine the predictions from each model to improve overall accuracy and robustness. The voting classifier's performance report and confusion matrix indicated a balanced trade-off between precision and recall, achieving an accuracy of 84%.

A screen shot of a computer program

Description automatically generated

*Fig. 23. The Ensemble Learning Approach*

This analysis emphasises the importance of aligning model evaluation with business objectives. While accuracy is important, the ultimate choice of model should consider the trade-offs between all performance metrics and their implications on campaign success and customer experience.

## Churn Prediction (Supervised Learning)

### Supervised Learning Objective and Feature Engineering

The objective of the supervised learning model was to predict the likelihood of customers discontinuing their relationship with the company - a critical insight for implementing proactive engagement strategies. For feature engineering, the RFM (Recency, Frequency, Monetary) model was utilised (Fig. 24). The 'DaysUntilChurn' feature was introduced with the logic that lower recency indicates a higher chance of churn (Fig. 25).

A graph of a graph of a graph

Description automatically generated with medium confidence

*Fig. 24. Churn Prediction – RFM Feature Engineering*

*A black screen with colorful text

Description automatically generated*

*Fig. 25. Churn Prediction – RFM Feature Engineering and Model Logic*

### Model Development and Evaluation

After splitting the dataset into training and holdout sets, quintiles were calculated for each RFM metric. Two models were developed: one based on RFM quintiles and another using continuous RFM values. Both models underwent standard feature scaling to ensure model input standardisation.

Two metrics were used in evaluating the models:

* Mean Squared Error (MSE): A measure of prediction accuracy which assesses the average squared difference between the estimated values and the actual value (Allwright, 2022).
* R2 Score: A statistical measure of how close the data are to the fitted regression line (Turney, 2022).

The quintile-based model produced an MSE of 30.53 and an R2 score of 0.965 on the test set. For the continuous model, the MSE was astonishingly close to zero, and the R2 score was a perfect 1.0 on both test and holdout sets. These results were highly unusual, suggesting an overfitting problem in the continuous model, as perfect prediction is improbable in real-world scenarios (Fig. 26).

A screenshot of a computer

Description automatically generated

*Fig. 26. Quantile-Based and Continuous Models Output*

### Visualisation of Predictive Model Accuracy on Holdout Data

The quintile-based model's accuracy was visualised against the actual data (Fig. 27). The plots showed varying degrees of predictive accuracy across different customer segments based on the RFM metrics. For instance, higher quintiles in the Monetary value plot correlated with lower predicted 'Days Until Churn,' indicating a lesser likelihood of churning for customers with higher spend.

A graph with a red line

Description automatically generated

*Fig. 27. Visualisation of Predictive Model Accuracy on Holdout Data*

### Segment-wise Analysis of Predicted Versus Actual Churn Days

The analysis revealed that different RFM segments had varying churn predictions, with the actual vs. predicted 'Days Until Churn' scatter plots indicating the quintile-based model's performance across segments. This segment-wise analysis provided actionable insights into the model's prediction accuracy, suggesting where improvements could be made (Fig. 28).

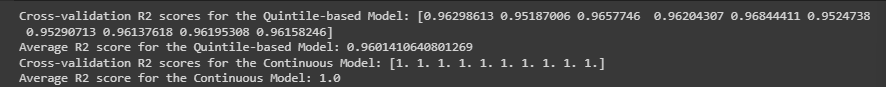
A graph on a white sheet

Description automatically generated

*Fig. 28. Segment-wise Analysis of Predicted Versus Actual Churn Days*

### Cross-Validation of Models

The quintile-based model was cross-validated with an average R2 score of 0.960, confirming its strong performance. In contrast, the continuous model achieved a cross-validation R2 score of 1.0, which again suggested overfitting (Fig. 29).



*Fig. 29. Cross-Validation Models*

### Ridge and Lasso Regression Analysis

Both Ridge and Lasso regression techniques were applied to the models to combat overfitting. The quintile-based model maintained high R2 scores of 0.965 and 0.965, respectively, for Ridge and Lasso regression, which is consistent with a strong and reliable model. The continuous model saw marginal improvements in MSE and maintained a near-perfect R2 score, which still indicated potential overfitting (Fig. 30).

The quintile-based model emerged as more reliable, as its performance metrics were strong without being suspiciously perfect. This model was recommended for practical application, as its predictive power and generalisability are more aligned with real-world scenarios. The continuous model's perfect scores are a cause for concern and warrant further investigation into potential data leakage or the complexity of the model. Implementing regularisation techniques like Ridge or Lasso regression provided insights into improving the models' robustness and highlighted the quintile-based model's reliability.

# Analysis and Results

This section consolidates the findings from various experiments across the unsupervised and supervised learning problems tackled in this study, synthesising them into coherent insights that align with the domain context and the initial objectives.

## Unsupervised Learning – Customer Segmentation

The unsupervised learning model successfully identified customer segments, aiding in creating personalised marketing campaigns and resource allocation. The distinct spending patterns and demographic features of 'Affluent Enthusiasts' suggest targeted high-value product offerings, while the 'Family-Focused Budget Shoppers' could be engaged through discount-focused campaigns. Feature engineering was crucial for the effective use of K-Means clustering. The silhouette scores indicated moderate to strong cluster cohesion, substantiating the chosen segmentation despite alternatives presented by the Elbow Method. The 3D visualisations enhanced the understanding of the multi-dimensional nature of customer behaviours, solidifying the model's relevance to the domain.

## Supervised Learning - Campaign Effectiveness Prediction

The supervised learning models addressed the key objective of predicting the success of marketing campaigns. The accuracy scores of Logistic Regression, Gradient Boosting, and AdaBoost, which hovered around 87.6%, were signifying a strong predictive capability. The ensemble approach provided a nuanced balance of precision and recall, guiding the deployment of marketing resources more efficiently. However, the challenge of class imbalance was notable. The application of RandomOverSampler improved representation and predictive performance, especially for the Random Forest classifier, which emerged as the standout model. It is important to note that while improved accuracy is desirable, it is the trade-off between various metrics that would ultimately guide the practical application of these models in real-world scenarios.

## Supervised Learning - Churn Prediction

Churn prediction is critical for pre-emptive customer engagement. The quintile-based and continuous models demonstrated stark contrasts in performance, with the latter showing suspiciously perfect metrics indicative of overfitting. Consequently, the quintile-based model, with its realistic performance measures, stands out as the more applicable model for business use. The visualisation and segment-wise analysis of churn predictions highlighted the quintile-based model's capacity to differentiate between customer groups effectively. These insights are actionable, directing attention to segments at higher risk of churn. The model's ability to inform retention strategies is of high value, emphasising machine learning's role in enhancing customer retention in the multi-channel retail space.

## Business Implications

Integrating these machine learning insights into the multi-channel retail domain could lead to significant advancements in customer experience and operational efficiency. The identification of customer segments can lead to more effective inventory management, tailored marketing strategies, and optimised channel engagement. The predictive models for campaign effectiveness and churn can inform customer relationship management systems, aiding in the design of personalised customer retention plans and optimising marketing budgets for maximum impact.

# Conclusion

## Summary of Key Findings

This report's analyses highlighted key insights into customer behaviours within the multi-channel retail domain. The application of K-Means clustering revealed distinct customer segments. The supervised learning models for campaign effectiveness and churn prediction highlighted Logistic Regression, Gradient Boosting, AdaBoost, and Linear Regression as robust performers. The quintile-based model was notably effective in churn prediction without overfitting.

## Implications of the Study

The findings from this study underscore the significance of domain expertise in guiding machine learning approaches and the importance of evaluating models not just on accuracy but also on their alignment with business objectives. In marketing strategies, where missing out on potential acceptances could entail significant opportunity costs, models that balance precision and recall - especially those with higher recall for positive responses - may offer greater business value.

## Limitations of the Study

Despite the insights gained, the study has limitations:

* Model Generalisability: The models may not perform equally well across different datasets or retail environments.
* Data Quality and Availability: The conclusions are as good as the data quality and may not account for uncollected or unobservable factors influencing customer behaviour.
* Complexity of Multi-Channel Data: The nature of multi-channel retail data, involving various platforms and customer touchpoints, poses a challenge for comprehensive and accurate analysis. The models might not fully capture the complexity of customer interactions across different channels.

## Recommendations for Future Research

* Granular Data Analysis: Delving into more detailed customer data could reveal nuanced behavioural patterns that broad segmentation may miss.
* Residual Analysis for Insights: Examining the residuals of regression models could illuminate systematic prediction errors, informing more accurate and reliable predictive models.
* Additional Regularisation Techniques: Exploring regularisation methods such as Elastic Net may provide a nuanced regularisation approach, potentially enhancing model robustness and performance.

In conclusion, the strategic application of machine learning has the potential to transform multi-channel retail operations by providing deep customer insights and enabling data-driven decision-making. Future research should continue to refine these models, ensuring they remain relevant and valuable in a rapidly evolving retail landscape.

# Reference List

Allwright, S. (2022). *What is a good MSE value? (simply explained)*. [online] Stephen Allwright. Available at: https://stephenallwright.com/good-mse-value/ [Accessed 28 Dec. 2023].

Briedis, H., Kronschnabl, A., Rodriguez, A. and Ungerman, K. (2023). *Adapting to the next normal in retail | McKinsey*. [online] www.mckinsey.com. Available at: https://mckinsey.com/industries/retail/our-insights/adapting-to-the-next-normal-in-retail-the-customer-experience-imperative [Accessed 11 Dec. 2023].

Brownlee, J. (2020). *Random Oversampling and Undersampling for Imbalanced Classification*. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/random-oversampling-and-undersampling-for-imbalanced-classification/ [Accessed 27 Dec. 2023].

Crawley, C. (2021). *Council Post: Omnichannel Versus Multichannel Marketing: Which Is Right For Your Brand?* [online] Forbes. Available at: https://www.forbes.com/sites/forbescommunicationscouncil/2021/05/18/omnichannel-versus-multichannel-marketing-which-is-right-for-your-brand/ [Accessed 11 Dec. 2023].

GOV.UK (2022). *AI activity in UK businesses: Executive Summary*. [online] GOV.UK. Available at: https://www.gov.uk/government/publications/ai-activity-in-uk-businesses/ai-activity-in-uk-businesses-executive-summary [Accessed 11 Dec. 2023].

Grand View Research (2022). *Artificial Intelligence Market Size, Share | AI Industry Report, 2025*. [online] Grand View Research. Available at: https://www.grandviewresearch.com/industry-analysis/artificial-intelligence-ai-market [Accessed 8 Jan. 2024].

IBM (2023a). *A Complete Guide on Exploratory Data Analysis*. [online] www.odinschool.com. Available at: https://www.odinschool.com/blog/a-complete-guide-on-exploratory-data-analysis [Accessed 15 Dec. 2023].

IBM (2023b). *What is the k-nearest neighbors algorithm? | IBM*. [online] www.ibm.com. Available at: https://www.ibm.com/topics/knn [Accessed 28 Dec. 2023].

Iglesias-Pradas, S. and Acquila-Natale, E. (2023). The Future of E-Commerce: Overview and Prospects of Multichannel and Omnichannel Retail. *Journal of Theoretical and Applied Electronic Commerce Research*, [online] 18(1), pp.656–667. doi:https://doi.org/10.3390/jtaer18010033.

Jaadi, Z. (2019). *A Step by Step Explanation of Principal Component Analysis*. [online] Built In. Available at: https://builtin.com/data-science/step-step-explanation-principal-component-analysis [Accessed 28 Dec. 2023].

Jain, O. (2023). *Logistic Regression: A Comprehensive Introduction*. [online] Medium. Available at: https://osheenjain.medium.com/unlock-the-power-of-logistic-regression-a-comprehensive-introduction-e0e8ba98917d [Accessed 28 Dec. 2023].

Jiang, T., Gradus, J.L. and Rosellini, A.J. (2020). Supervised Machine Learning: A Brief Primer. *Behavior Therapy*, 51(5), pp.675–687. doi:https://doi.org/10.1016/j.beth.2020.05.002.

Kumar, R. (2020). *A Comparitive Study Between AdaBoost and Gradient Boost ML Algorithm*. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2020/10/adaboost-and-gradient-boost-comparitive-study-between-2-popular-ensemble-model-techniques/ [Accessed 28 Dec. 2023].

Lemos, A.R. (2022). *Cross-Validation*. [online] Medium. Available at: https://towardsdatascience.com/cross-validation-705644663568 [Accessed 15 Dec. 2023].

Mali, K. (2021). *Linear Regression | Everything you need to Know about Linear Regression*. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2021/10/everything-you-need-to-know-about-linear-regression/ [Accessed 28 Dec. 2023].

Mannor, S., Jin, X., Han, J., Jin, X., Han, J., Jin, X., Han, J. and Zhang, X. (2011). K-Means Clustering. *Encyclopedia of Machine Learning*, pp.563–564. doi:https://doi.org/10.1007/978-0-387-30164-8\_425.

McClarren, R.G. (2021). Decision Trees and Random Forests for Regression and Classification. *Machine Learning for Engineers*. doi:https://doi.org/10.1007/978-3-030-70388-2\_3.

Medium (2023). *Improving Accuracy and Performance with Voting Classifiers in Ensemble Learning*. [online] Medium. Available at: https://blog.tdg.international/improving-accuracy-and-performance-with-voting-classifiers-in-ensemble-learning-ab5bd69e631f [Accessed 28 Dec. 2023].

Melanie (2023). *K-Means Clustering in Machine Learning: A Deep Dive*. [online] Data Science Courses | DataScientest. Available at: https://datascientest.com/en/k-means-clustering-in-machine-learning-a-deep-dive [Accessed 28 Dec. 2023].

Miller, R. (2019). *Data Preprocessing: what is it and why is important*. [online] CEOWORLD magazine. Available at: https://ceoworld.biz/2019/12/13/data-preprocessing-what-is-it-and-why-is-important/ [Accessed 15 Dec. 2023].

Mitra, A., Jain, A., Kishore, A. and Kumar, P. (2022). A Comparative Study of Demand Forecasting Models for a Multi-Channel Retail Company: A Novel Hybrid Machine Learning Approach. *Operations Research Forum*, [online] 3(4). doi:https://doi.org/10.1007/s43069-022-00166-4.

Oleszak, M. (2023). *Not-so-naive Bayes*. [online] Medium. Available at: https://towardsdatascience.com/not-so-naive-bayes-eb0936fa8b4a [Accessed 28 Dec. 2023].

S, P. (2021). *The A-Z guide to Support Vector Machine*. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2021/06/support-vector-machine-better-understanding/#:~:text=Support%20Vector%20Machine%20(SVM)%20%E2%80%93%20(Interval%20block)%3A [Accessed 28 Dec. 2023].

Saji, B. (2021). *K Means Clustering | K Means Clustering Algorithm in Machine Learning*. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2021/01/in-depth-intuition-of-k-means-clustering-algorithm-in-machine-learning/ [Accessed 28 Dec. 2023].

Scikit-Learn (2019). *sklearn.preprocessing.StandardScaler — scikit-learn 0.21.2 documentation*. [online] Scikit-learn.org. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html [Accessed 28 Dec. 2023].

Scikit-learn (2019). *sklearn.metrics.silhouette\_score — scikit-learn 0.21.3 documentation*. [online] Scikit-learn.org. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette\_score.html [Accessed 28 Dec. 2023].

Scikit-learn (2023a). *Ridge coefficients as a function of the L2 Regularization*. [online] scikit-learn. Available at: https://scikit-learn.org/stable/auto\_examples/linear\_model/plot\_ridge\_coeffs.html#:~:text=To%20address%20overfitting%2C%20Ridge%20regularization [Accessed 28 Dec. 2023].

Scikit-learn (2023b). *sklearn.linear\_model.Lasso*. [online] scikit-learn. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.linear\_model.Lasso.html [Accessed 28 Dec. 2023].

Thormundsson, B. (2023). *Artificial Intelligence Market Size 2030*. [online] Statista. Available at: https://www.statista.com/statistics/1365145/artificial-intelligence-market-size/ [Accessed 6 Jan. 2024].

Tian, Z., Zhai, X., van Steenpaal, G., Yu, L., Dimara, E., Espadoto, M. and Telea, A. (2021). Quantitative and Qualitative Comparison of 2D and 3D Projection Techniques for High-Dimensional Data. *Information*, 12(6), p.239. doi:https://doi.org/10.3390/info12060239.

Turney, S. (2022). *Coefficient of Determination (R2) | Calculation & Interpretation*. [online] Scribbr. Available at: https://www.scribbr.com/statistics/coefficient-of-determination/ [Accessed 28 Dec. 2023].

Xgboost developers (2022). *XGBoost Documentation — xgboost 1.5.1 documentation*. [online] xgboost.readthedocs.io. Available at: https://xgboost.readthedocs.io/en/stable/ [Accessed 28 Dec. 2023].

Yadav, S. (2023). *Silhouette Coefficient Explained with a Practical Example: Assessing Cluster Fit”*. [online] Medium. Available at: https://medium.com/@Suraj\_Yadav/silhouette-coefficient-explained-with-a-practical-example-assessing-cluster-fit-c0bb3fdef719 [Accessed 27 Dec. 2023].

Zharovskikh, A. (2022). *Importance of AI in Modern Corporate World*. [online] InData Labs. Available at: https://indatalabs.com/blog/importance-of-ai-in-corporate-world [Accessed 11 Dec. 2023].

# Appendices

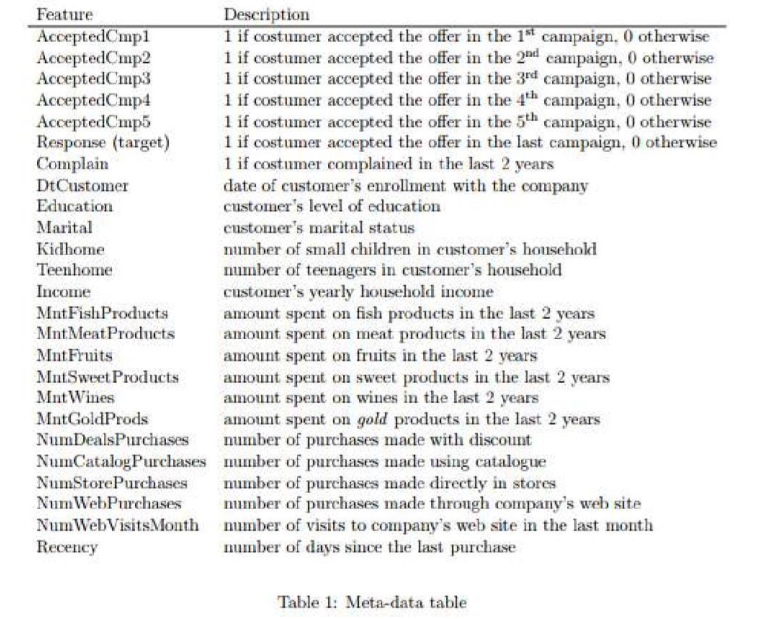
## Link to Full Code and Comments in Colaboratory

<https://colab.research.google.com/drive/1gv_-VaORPibntroro-x6UH6R1S3Hhrgi?usp=sharing>

## Link to Kaggle Dataset

<https://www.kaggle.com/datasets/jackdaoud/marketing-data>

## Additional Figures

*****Fig. 1. Meta-Data Table*

A screen shot of a computer code

Description automatically generated

*Fig. 2. Removal of Redundant Columns*

A screenshot of a computer

Description automatically generated

*Fig. 3. Duplicates*

A screenshot of a computer

Description automatically generated

*Fig. 4. Negative Values*

A screen shot of a computer

Description automatically generated

*Fig. 5. Cleaned Data Frame Dimensions*

A computer screen with text

Description automatically generated

*Fig. 16. Silhouette Score*

A screen shot of a computer program

Description automatically generated

*Fig. 17. Campaign Effectiveness Prediction – Data Preparation and Processing*

A screen shot of a computer program

Description automatically generated

*Fig. 18. Campaign Effectiveness Prediction – Model Training and Evaluation*

A screenshot of a computer program

Description automatically generated

*Fig. 30. Ridge and Lasso*