12-2022

Understanding Customer Purchasing Decisions using RFM and Machine Learning

Majed Alfalasi
maa5749@rit.edu

Understanding Customer Purchasing Decisions using RFM and Machine Learning

by

Majed Alfalasi

A Capstone Submitted in Partial Fulfilment of the Requirements for the Degree of Master of Science in Professional Studies: Data Analytics

Department of Graduate Programs & Research

Rochester Institute of Technology

RIT Dubai

2022 - December

RIT

Master of Science in Professional Studies:

Data Analytics

Graduate Capstone Approval

Student Name: Majed Alfalasi

Graduate Capstone Title: Understanding Customer Purchasing Decisions using RFM and Machine Learning

Graduate Capstone Committee:

Name: **Dr. Sanjay Modak** **Date:**

Chair of committee

Name: **Dr. Hammou Messatfa** **Date:**

Member of committee

Acknowledgments

In this project, I would firstly like to convey my sincere appreciation to my mentor Dr. Hammou as well as the chairman Dr. Sanjay for the constant support and guidance throughout the course of the research. Without their support, I would not have been able to work on this and learn as much as I did today and I would be thankful to them for that.

I would also like to thank my parents and friends for their constant support throughout the graduation process who also provided me moral support to progress through this course with ease.

Everyone responsible for this project, and related to the same who have extended their sincere support and mentoring for my progress, I would like to offer my appreciation and blessings.

Abstract

In this report, we plan to use commonly available datasets from Kaggle (PATEL, 2021), which is basically customer's records collected from a grocery firm's database. In the course of the reporting process, we would be implementing RFM analysis (Correia, 2016) to be able to segment customers based on their buying patterns. This would help us understand the types of customers based on their historical purchase behaviour. Once that is done, we will also use classification models to be able to factor in the segmented customers and then target the customers with high potential or propensity to accept marketing offers. These customers would then be the ideal customers who would have higher conversion rates for any promotions sent to them.

Even before that, we would be going through the details of the dataset through some exploratory data analyses, and then cleaning the data for inconsistency and then finally performing RFM segmentation and then classification based on the information and data that we obtain to predict the response to the offers. It becomes a starting point to building relationships with consumers, who in turn become reliable with respect to generating businesses and provide improved conversion rates for the products offered by the businesses (Kim, 2006).

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Problem Statement

Present times have observed an increase in customer demands and buying patterns, they have become smarter when it comes to choosing a service or product for purchase. This is one of the key problem areas wherein businesses are trying to identify patterns as well as similarities among users to be able to promote their product and services better. A typical example would be the user base who visit websites like Noon or Amazon on a daily basis. These companies would need to understand the needs of their users to be able to roll out the appropriate offers and products in the long run.

In addition, it is crucial to use Data Analytics and modelling techniques to understand and cohort customers, to have a holistic understanding of their behaviour and responsiveness to marketing campaigns. There are many questions to be answered by using such methods like -

- What are the attributes of customers in different clusters?
- What are the demographics traits of customers belonging to each cohort like what is the salary, what is the household size, what is the recency of the customer enrolment with the store etc.?
- How do customers from different segments respond to marketing campaigns?

Since we do not have open-source platforms or websites to cluster customer datasets, we would be using Data Analytics and Machine Learning models to understand these customer details in depth. This would help us answer the above questions in depth and create appropriate customer cohorts/segments for future works.

1.2 Introduction

Modelling techniques implemented in data analyses programming languages involve a lot of pre-requisites and know-hows about the dataset. Knowing these details helps in implementing the correct techniques to the correct set of data. There are different types of datasets, as well as variables within a dataset. The top-level bifurcation for any dataset or field is being either qualitative or quantitative, i.e., either categorical or numerical. As shown in the diagram below, qualitative and quantitative can further be divided into sub-categories to identify datasets and their respective fields. This is very crucial when we are trying to understand the data.

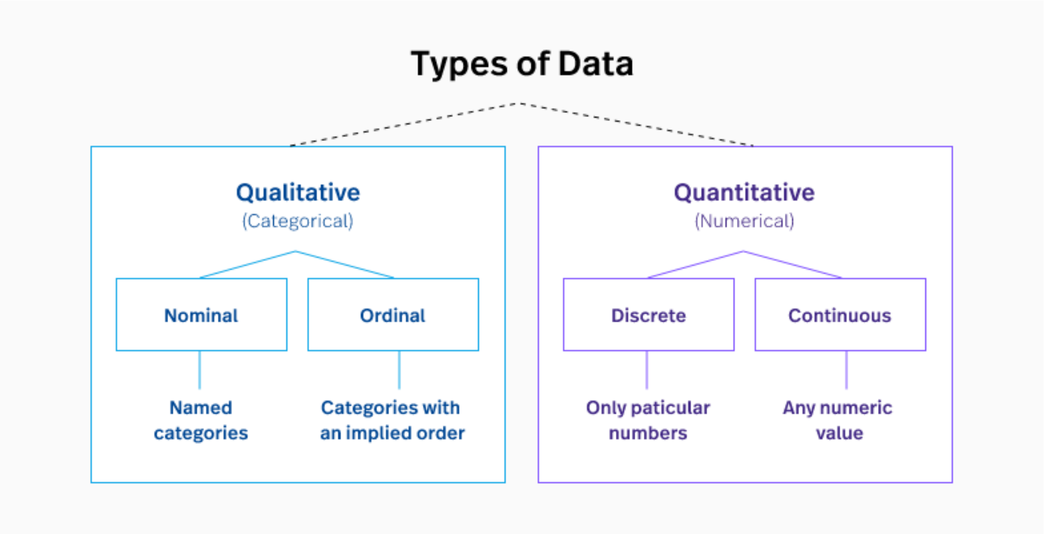
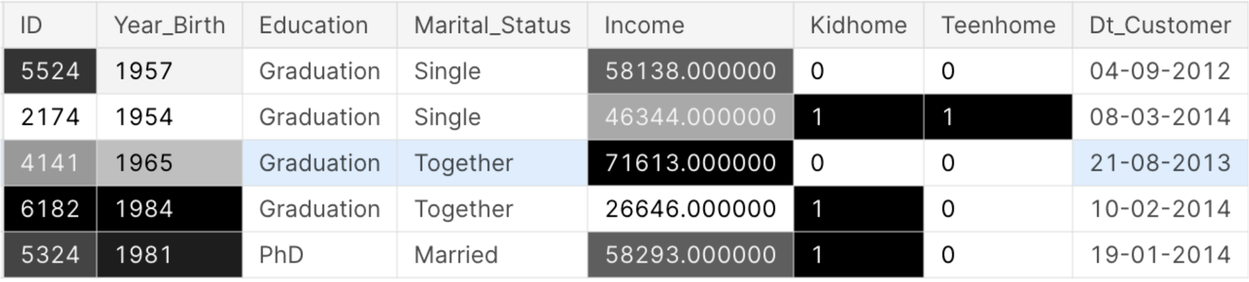


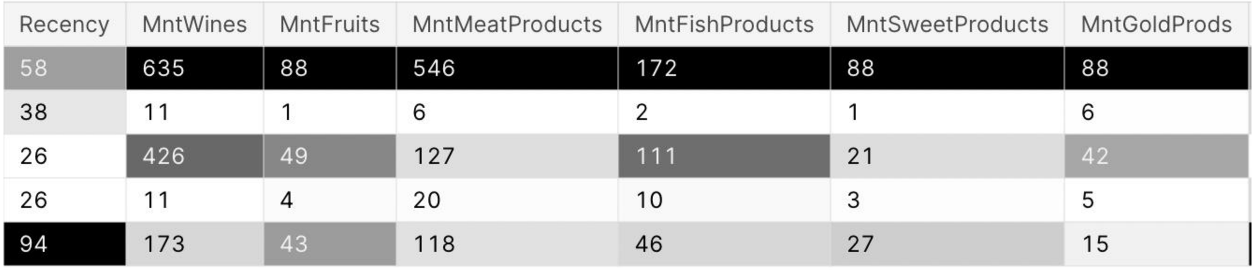
Fig.1.1. Data types and their visual representation

Customer segmentation is the process of dividing users into groups of similar interests based on a number of factors like buying pattern, income bracket, interests, number of people in household and many others. Clustering facilitates the smooth flow of rolling out marketing strategies or products for their user base, based on their interests and common traits. This ensures higher conversion rates and enables better targeting strategies and use of marketing budgets and efforts. (Shopify, 2020) Many companies rely on customer profile dataset, along with other data like transaction history, browsing history as well as clickstream dataset to understand user journeys on different websites and platforms along with their interests in specific items or articles. It is also imperative to understand how predictive modelling comes into effect for such campaigns in terms of marketing. When different marketing campaigns are rolled out to users, it is also kept in mind to analyse such data and understand responsiveness to such campaigns in order to differentiate between customer clusters and their responsiveness towards each marketing campaign. These companies use the data to understand that Marketing #1 worked well on cluster #1 of customers compared to cluster #2, and these patterns help them plan future marketing campaigns accordingly.

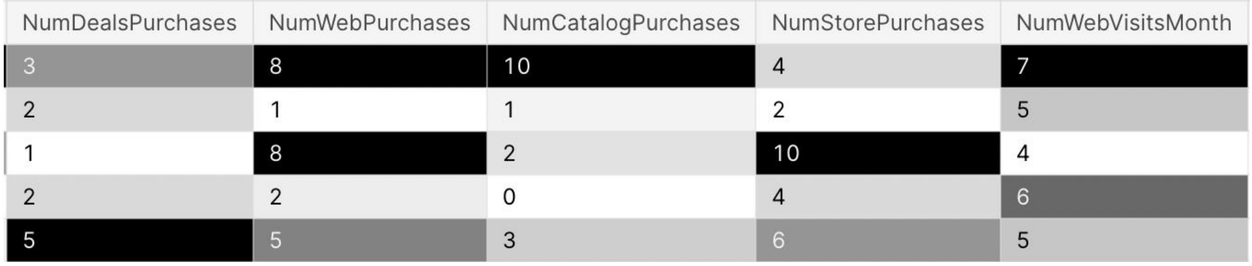
Different models behave differently to different types of target variables, which can be nominal, ordinal, discrete or continuous. This is the key differentiator for different fields in the dataset. In our report and the dataset that we plan to use, the data contains 2240 data points with 29 attributes/columns. Now, the columns can further be categorised in the following manner –



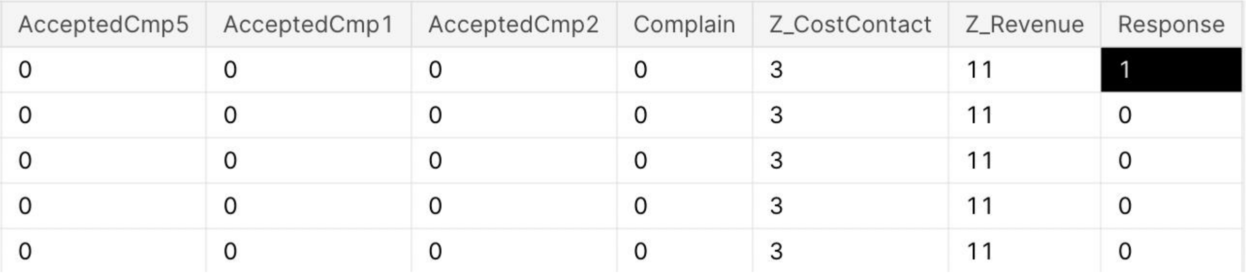
ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer
5524	1957	Graduation	Single	58138.000000	0	0	04-09-2012
2174	1954	Graduation	Single	46344.000000	1	1	08-03-2014
4141	1965	Graduation	Together	71613.000000	0	0	21-08-2013
6182	1984	Graduation	Together	26646.000000	1	0	10-02-2014
5324	1981	PhD	Married	58293.000000	1	0	19-01-2014



Recency	MntWines	MntFruits	MntMeatProducts	MntFishProducts	MntSweetProducts	MntGoldProds
58	635	88	546	172	88	88
38	11	1	6	2	1	6
26	426	49	127	111	21	42
26	11	4	20	10	3	5
94	173	43	118	46	27	15



NumDealsPurchases	NumWebPurchases	NumCatalogPurchases	NumStorePurchases	NumWebVisitsMonth
3	8	10	4	7
2	1	1	2	5
1	8	2	10	4
2	2	0	4	6
5	5	3	6	5



AcceptedCmp5	AcceptedCmp1	AcceptedCmp2	Complain	Z_CostContact	Z_Revenue	Response
0	0	0	0	3	11	1
0	0	0	0	3	11	0
0	0	0	0	3	11	0
0	0	0	0	3	11	0
0	0	0	0	3	11	0

Customer Information

- ID
- Year Birth
- Education
- Marital Status
- Income
- Kidhome
- Teenhome
- Dt Customer
- Recency

Products (Amounts spent on different products in last 2 years)

- MntWines
- MntFruits
- MntMeatProducts
- MntFishProducts
- MntSweetProducts
- MntGoldProds

Place

- NumWebPurchases
- NumCatalogPurchases
- NumStorePurchases
- NumWebVisitsMonth

Promotion

- NumDealsPurchases
- AcceptedCmp1
- AcceptedCmp2
- AcceptedCmp3
- AcceptedCmp4
- AcceptedCmp5
- Response

Clustering dataset involves multiple steps, before even forming the clusters. As shown in the figure given below, the first step is the data acquisition phase wherein we would need to obtain the appropriate dataset for our entire exercise. The next step involves the pre-processing step wherein the dataset is cleaned and is made ready for the unsupervised machine learning algorithm to be able to identify the features appropriately. Messy data is always bad for machine learning algorithms because it can lead to errors and biases, and to avoid such problems the dataset needs to be cleaned and normalised.

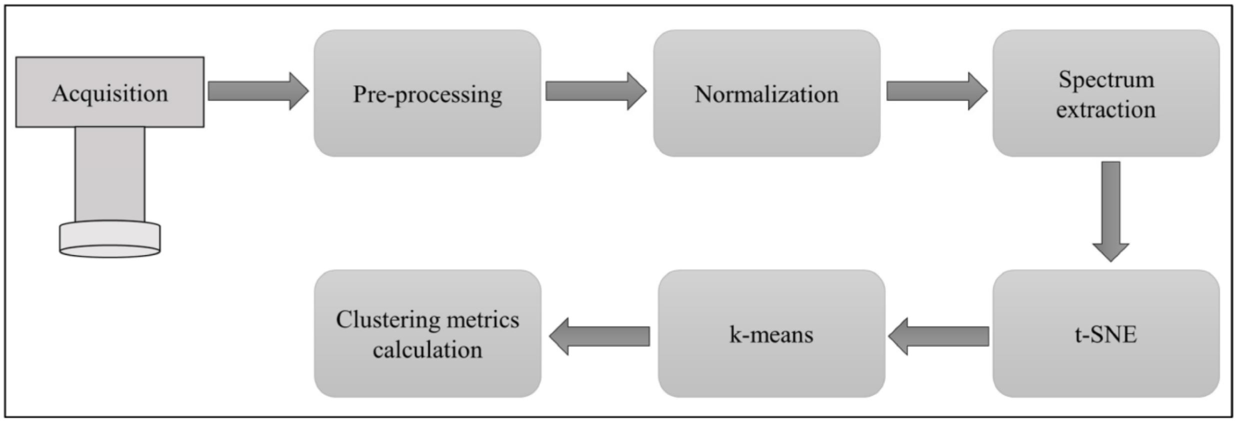


Fig.1.2. Clustering data pipeline setup end-to-end

Once the process of cleaning and normalisation is in place, we can move ahead to creating the cluster using the k-means method. Once the clusters are formed, we move ahead to identify the different matrices of the different clusters. In a typical clustering technique, the method to identify the appropriate number of clusters is done by different techniques like Within Sum of Squares (WSS) or Silhouettes coefficient technique. This ensures that we consider an optimal value for the number of clusters even before moving to creating the model.

With the help of the clustered dataset, we can use the relevant predictive model to understand the effectiveness of a certain marketing campaign and interpret the responsiveness towards the campaigns accordingly.

1.3 Project goals

The scope of our project involves collecting the appropriate dataset related to customer profiles as well as different traits that explain the customer characteristics. In our case, we will use a grocery company's customer database and then identify different clusters as well as performing data analysis and different visualisations on these clusters. This would help us in understanding the properties of each cluster even after performing the clustering technique, and post-hoc analysis is very important in any clustering technique. Not only that, but we will also use analysis to understand the response variable, i.e., the response to a marketing campaign to understand the customer patterns to respond to different marketing campaigns.

Our objective of the project is to identify different clusters from the customer database obtained from a grocery company, and once we obtain the clusters, we can use the information to understand different properties for each of the clusters based on data analysis techniques. For example, we can understand the average income range of each of the clusters, or the average age of different people contained in different clusters and many other traits (Hamka, 2014). On an ad-hoc basis we could also perform dimensionality reduction techniques and perform an agglomerative clustering algorithm to obtain the required clusters.

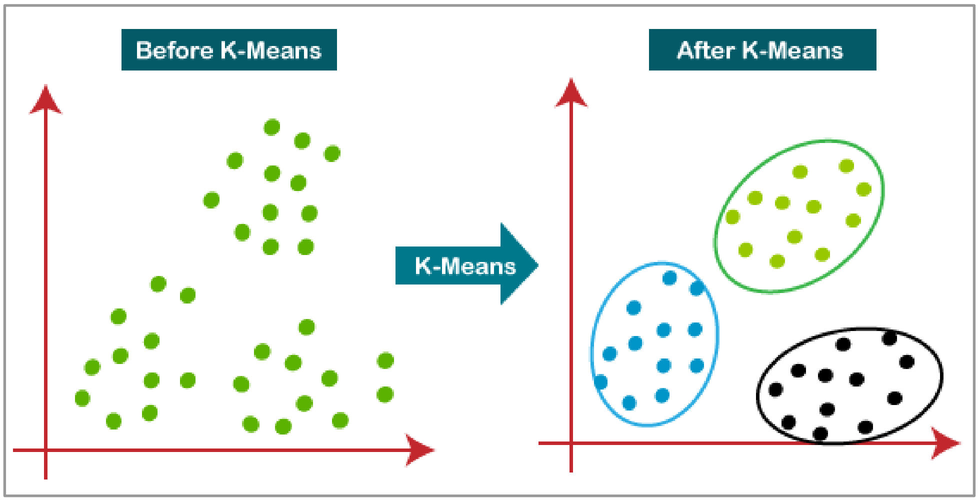


Fig.1.3. K-means clustering results before-after

The idea behind any clustering algorithm is to group data points based on similarities, as shown in figure 2, clusters are obtained on the right which are based on similarities in different attributes from the dataset. This is the main goal that we try to achieve from the dataset shown above as well as our explanation on the subject. In this project, we will use K-Means clustering technique which is a form of vector quantization method which uses the process of partitioning n observations into k different clusters and each of the observations are sent to the cluster with the nearest mean. A target number k refers to the total number of centroids that are needed in the dataset. The K-Means clustering algorithm focuses on identifying k number of centroids and then the data points are allocated eventually to the nearest cluster, in the same time trying to keep the centroid very small. In a typical process of the algorithm functions, the first group of randomly selected centroids are used in the data mining process, and these are in turn used as the beginning points of each cluster. Now, after the centroids are determined, iterative calculations are used to optimise each of the positions of the centroid. When it comes to K-Means clustering, there are different distance methods used to achieve the distance of the data points from the centroids, they are -

1. Euclidean Distance
2. Manhattan Distance
3. Pearson Correlation Distance
4. Eisen Cosine Correlation Distance

5. Spearman Correlation Distance
6. Kendall Correlation Distance

To explain a bit more about the type of distance method and its usage implications, if we want to identify clusters of observations along with the overall profiles (without the need to determine their magnitudes), then we would need correlation-based distance as a dissimilarity measure. Let's say, in the case of marketing if we want to use the data to identify the customers of the same group with respect to preference of items, regardless of the volume of items they bought. In the case of Euclidean distance, the observations with high value of features are often clustered together which is sometimes not ideal. The same is true for observations which have low values of features. In our project, we plan to use correlation-based similarity methods so that a holistic approach is considered when clustering the observations.

1.4 Aims and Objectives

The aim of this report is to understand consumer behaviour through the use of grocery purchase data from a store. We plan to use analytics and modelling tools like SPSS Statistics and Modeller to shape and prepare the data along with deriving statistical inferences from the same. Subsequently, we will use modelling and clustering to determine the different clusters and understand the customer behaviour in different clusters. This will help us distinguish between customers with high, medium, and low purchase propensity.

1.5 Research Questions

During the phase of the coming section, we will perform some univariate and bivariate analysis to understand more about the data points and their distribution. We will understand some key factors like -

- Does the age of the customer have a relationship with the amount they tend to spend or their income?
- The distribution of customers in the database based on their marital status, education etc.
- The distribution of different data points like income of the customers

The above hypothesis, when answered with the help of data, can help us understand the relationship of the features that we have in the dataset which we can then use for the modelling purpose.

Chapter 2 Literature Review

2.1 Literature Review

In their paper, (Wu, 2005) has leveraged credit card consumption data for the model-building samples to present a modelling framework. This is then used to build a segment-level predictive model which can utilise a pattern-based approach through clustering technique. Having obtained the clustered dataset along with the customer profiles, Wu goes ahead and devises a monetary matrix along with the addition of a fluctuate-rate matrix to study the action of different modes. By clustering on the above-mentioned matrixes, different customer characteristics are uncovered and after utilising these same characteristics, two-dimension consumption-based customer clustering models are built at the end of the process. This is one way of leveraging customer Clustering techniques to obtain cohorts of customers and Wu and her team used this technique for credit card consumption data to be able to obtain actionable insights and usable data at the end.

The more the marketing sphere develops, the more important the long-term relationships with customers become important in the day-to-day life of the businesses. CRM looks for long-term relationships with the customers who are profitable. It is very important to understand the customer relationship and loyalty patterns in order to maintain a healthy marketing infrastructure deployed towards targeted customers and profitable customers. The success of any corporation or service/product depends on their relationships with the customers, which is why it is very crucial to build redefined plans and implementations to build long-term customer relationships. (Kim, 2006) Having said that, using advanced unsupervised machine learning techniques like k-Means and hierarchical clustering techniques can result in tremendous business advantages for organisations wherein business potentials are unlocked, after having built strong customer bonds. Customer Clustering techniques enable businesses to understand their customer habits in terms of purchasing, interests etc. and understanding these can help companies roll out appropriate and tailored products/services to the customers. This is where customers become loyal in treating the business and sustainable services are developed.

During the global pandemic, a lot has changed over the globe with respect to businesses and other factors. The influence of customers on the market and purchasing habits have even changed along with this, and this is still not clear to researchers as to how they changed and what exactly changed. In their paper, Akar carried out a survey on 520 online users from Turkey which has been aimed at investigating pandemic related customer purchase habits. They tested their hypotheses with partial least squares, and through the results it was indicated that the customers' pandemic concerns had an impact on their intentions behind purchasing anything online (Akar, 2021). It is understood that during the global pandemic, a lot of the patterns and purchasing intentions were impacted and this was aptly proven with the help of the above research that Akar and his team performed. Through proper analyses and machine learning techniques, appropriate insights can be obtained regarding customer behaviour and buying patterns.

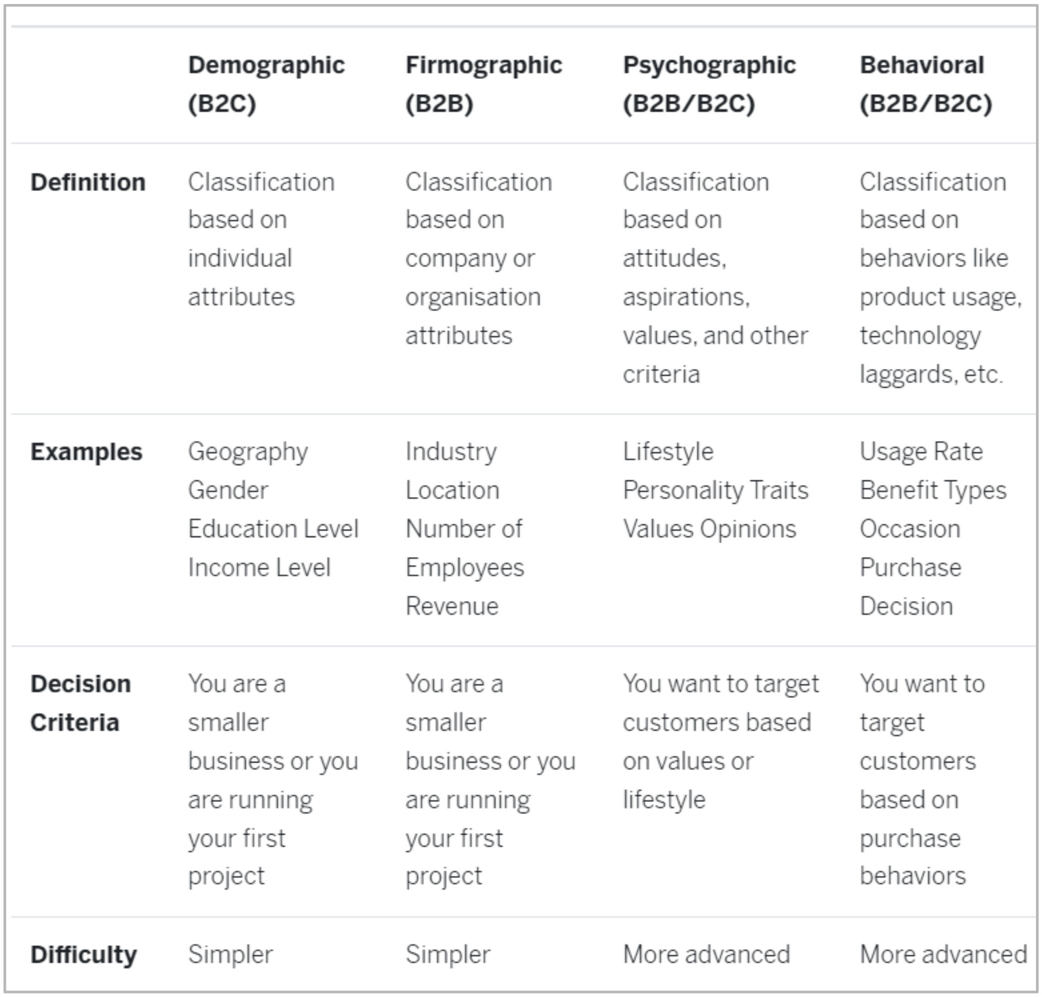
E-Commerce has been an age-old booming business and many people rely on these platforms to make their daily purchases when it comes to clothes, merchandise, food, grocery etc. Now, these platforms tend to overload customers with their never-ending inventory and assortments exposed to these users who get too overwhelmed with information displayed (Sari, 2016). Hence, when companies market, the personalisation techniques that are implemented also play a big role in bagging customers for long term sales. This is achieved through customer Clustering of market Clustering techniques. In their paper, Sari discusses customer Clustering techniques using data, and the customer Clustering data were divided into both internal and external datasets. Ultimately, different methods are implemented to process the data like Business rule, Magento, Customer Profiling, Quantile Membership etc. These methods are discussed in depth in the paper which helps people understand the business objective behind these techniques and their performance implications. There are firms like Migros Turk, who construct novel and very inventive ways of using Clustering strategies. There are different types of clustering analysis, mostly descriptive and predictive wherein predictive clustering include cluster regression and CHAID (Cooil, 2008).

Naturally, it is always simpler to make assumptions and utilise "gut instincts" to set the rules that would categorise customers into logical categories, such as those who purchased a specific product or service, originated from a specific source, or resided in a specific place. These broad classifications, meanwhile, rarely produce the desired outcomes. It goes without saying that certain clients will spend more money with a business than others. The top clients will spend a lot over a long period of time. Good clients will either spend a lot in a short amount of time or modestly over a lengthy period. Others will refrain from making large purchases or from remaining too long. To address each cohort in a way which will most likely maximise that potential, or lifetime, worth, the proper way to segment consumers is based on estimates of their total future value to the business. (Optimove, 2022)

If you sell seasonal products, geographical segmentation is critical. Don't irritate your customers by sending them offers that are out of season or do not correspond to the climatic conditions in their area. Consider the location of all your customers when creating a promotion for the best results. Lifestyles and cultures differ from one location to the next. Understanding them ensures that you respect their beliefs and tailor your marketing campaigns accordingly. using the information about your customers, such as purchasing habits, to offer products, services or offers that match their interest and preferred times. (Raghavan)

These technologies allow brands to continuously acquire and analyse client data. Companies can personalise offerings and messages to specific consumer segments using the same ML-driven platform. Starbucks, with 30,000 locations and 25 million active rewards members, wanted to be the most personalised brand internationally. AI-powered solutions are used to segment customers based on purchases and interactions. Starbucks' AI and ML skills have helped them understand what customers want and generate personalised loyalty packages. 10x marketing operations execution speed and 3x tailored marketing sales lift. AI automates micro-segmentation more successfully than humans since it recognises unique client traits. It improves offers and experiences for major company segments, new vs. lapsed clients. This technique eliminates the need for loyalty and digital marketers to manually alter programs. Machine learning-powered optimisation solutions run in the background so businesses and loyalty marketers may focus on strategic program decisions that strengthen consumer emotional connections. Formation helps segmentation. Formation's Dynamic Offer Platform helps Fortune 500 brands create customer relationships. Formation uses businesses' first-party data to build each customer's best action and reward offer, then handles deployment, measurement, and fulfilment across the marketing stack. Segmenting customers is changing. Personalisation within customer segmentation drives 8X revenue growth, 60% savings in reward expenditure efficiency, increased engagement, and 40% category exploration. (Customer segmentation models there's a better approach for 2022, 2022)

Your business model revolves around customers. Your company can't survive without (profitable) consumers. To better serve your clients, organise them by needs, jobs-to-be-done, behaviours, or other factors. Your business strategy may have huge or tiny Customer Segments. Choose which parts to serve and which to ignore. Once you decide, you may construct your business model around customer wants and jobs-to-be-done. Customer groups are independent segments if: - Their needs necessitate a distinct Value Proposition - They are reached through different Distribution Channels - They require different types of interactions - Their profitability is significantly different - They'll pay for different Value Proposition features (Business Model Canvas, 2019).



	Demographic (B2C)	Firmographic (B2B)	Psychographic (B2B/B2C)	Behavioral (B2B/B2C)
Definition	Classification based on individual attributes	Classification based on company or organisation attributes	Classification based on attitudes, aspirations, values, and other criteria	Classification based on behaviors like product usage, technology laggards, etc.
Examples	Geography Gender Education Level Income Level	Industry Location Number of Employees Revenue	Lifestyle Personality Traits Values Opinions	Usage Rate Benefit Types Occasion Purchase Decision
Decision Criteria	You are a smaller business or you are running your first project	You are a smaller business or you are running your first project	You want to target customers based on values or lifestyle	You want to target customers based on purchase behaviors
Difficulty	Simpler	Simpler	More advanced	More advanced

Fig.2.1. The basics of segmentation table

The Figure above simplifies the four basics of segmentation. One of the main goals of market segmentation and targeting is to anticipate consumer behaviour, such as product purchases. It is common practice to use a predictive model in research, which allows for the easy categorisation of respondents into predetermined groups according to their responses to survey questions. (Oliver, 2022).

Big data is a big amount of structured and unstructured data from various sources such as the web, business organisations, and others that arrives at a high rate, making processing difficult using typical database management methods. The torrent is increasing. Big data processing challenges include storage, search, distribution, transport, analysis, and visualisation. Previously, "Analytics" meant the examination of existing data to explore prospective patterns and assess the effects of decisions or occurrences to generate valuable insights for business intelligence. Today's largest challenge is finding all the hidden information in massive amounts of data from many sources. Customer analytics predicts buyer behaviour, boosting sales, market optimisation, inventory planning, fraud detection, and other uses. Decision trees for classification are a popular method for consumer analytics. (Khade, Researchgate, 2016)

Focus on the client needs to grow your brand and business. Identifying those needs is difficult. Customer surveys, which fill inboxes and websites with pop-ups, have minimal impact. Due to overloaded inboxes and pop-up ads, many consumer surveys get unaddressed, misrepresenting results. Customer analysis can help businesses learn what customers want and improve by knowing and analysing client behaviour.

Lower Customer acquisition costs

Small businesses and startups often lose money on customer acquisition. Advertising, targeted marketing, and underground brand-building efforts are expensive ways to attract new clients. Customer acquisition could be lost without good customer analysis. Finding clients in the wrong areas is unlikely to succeed. Quality customer analysis can help you uncover and convert hidden buyers into brand advocates.

Better Customer Retention

Retention costs less than acquisition. The appropriate analysis may improve retention, raise profitability, and reduce expenses by helping firms understand their customers. Customer retention is key to corporate success. Your current clients are your most important assets, therefore keeping them is good for business and brand.

Efficient Customer Service

The appropriate analysis can help firms with customer service. Knowing your customers may simplify customer service, improve efficiency, and boost effectiveness. On-going consumer analysis can also discover gaps in your existing customer service processes, so you can beef up your offerings and establish a better brand. If customer data shows that a lot of clients are from a different time zone, the business could add a second shift or expand customer service hours.

Increased Sales and Improved Profits

Customers buy anything you sell. Without those customers, sales and profitability will swiftly drop to zero. Customer quality determines your business's profitability and sales growth. By researching your clients and finding out what they have in common, you can better sell your items and increase earnings and sales. (Correia, 2016)

Customer analysis goes beyond product sales. The appropriate analysis can help you create new products and services that clients may not even realise they need. This way, your new product lines could boost sales and earnings, helping you grow your business. (The Business Benefits of Customer Analysis)

RFM. This is a popular customer segmentation technique which is implemented based on the historical purchase behaviour of the customers. Based on multiple factors like Recency, Frequency and Monetary value the segments are generated and we can interpret different KPIs based on their segments

Random Trees. This is a classification model wherein multiple decision trees are considered during the feature split phase, and the best tree is considered for the prediction values. The number and significance of each tree split is based on numerous parameters like mtry, gain and many others which are determined during the modelling process.

C5. This node in SPSS Modeler uses either decision tree or rule set to derive a predictive approach. Here, the approach of splitting the tree again and again is followed in an iterative manner (IBM, 2022)

Logistic Regression. This modelling technique is popular and can be used for classifying records based on the input field values. It is like linear regression, with the difference in the process that it considers categorical value in the target field instead of numerical values. (Michael, 2008)

Neural Network. These are ANNs (Artificial Neural Networks) which mimic the human brain with the set of input, output and hidden layers to decide on the probable outcome from the provided input features (Wang, 2003)

Tree-AS. This module of SPSS Modeler allows the construction of decision trees by using either CHAID or Exhaustive CHAID model (IBM, 2022)

ROC (Receiver Operating Characteristic Curve). This process is done using a graph to show the performance of different classification models with different thresholds for these classification steps. The two main components of the plot are True Positive Rate (TP) and the False Positive Rate (FP). (Narkhede, 2018)

AUC (Area Under the Curve). This is another model evaluation KPI (key performance indicator) which provides an aggregation of performance measures across all classification thresholds. There are multiple ways of interpreting AUC and one of them is that the probability of the model ranking random positive examples is higher than random negative examples.

Gain. In this process, the reduction in entropy is calculated which might occur from modifying the dataset. This process is used for making decision trees primarily in the train set wherein the information for each variable is evaluated. With the help of this factor, the splits are performed in the tree and for each split, the relevant features are then selected. This process can also be used for feature selection steps which helps in determining the best features for the modelling process. (Google, 2022)

2.2 Takeaways from Literature Review

From our literature survey, we observe that there is a lot done in the customer Clustering domain in the last couple of years. Some of the key takeaways are -

- Businesses need to adapt to the changing market scenarios and be able to profile their customers better, for increased revenue and conversions
- There have been tremendous developments in terms of machine learning models and resources which can enhance the accuracy of customer profiling and cohort creation
- For long term business relationships with customers, it is very crucial to understand different customer needs and profile which can be obtained using customer Clustering techniques

With the above summarisation of our research on different papers and resources, it is understood that customer Clustering plays an important function in the modern world. Through this project, we plan to implement some of the most important techniques to understand these customer models and behaviour. Moreover, the above research has helped us with our project steps in the following ways -

- To be able to determine the appropriate features that would be needed in our modelling process, which can be called feature engineering or feature determination. The process of reducing dimensions helps the model in making accurate judgement, if we are able to reduce noise
- Based on the above research study, it is understood that clustering helps people identify the inherent traits of the segments in terms of demographics, behaviour, nature etc. With that in mind, we would be able to explain our own clusters that would be formed using the necessary method
- Clustering can be used in many industries, based on the research done. They are used in marketing, retail, e-commerce, banking etc. Different use cases arise from different industries for the clustering technique which helps us in identifying customer profiles for tailoring the right experience to them

Chapter 3 Research Methodology

3.1 Methodology

In this project, we start with obtaining the appropriate dataset of customer profiles from a grocery company. The first step to any data analytics project is preparing the dataset or collecting the data for further analyses. Now, when we have the dataset ready the next step is to prepare the data, and this can be through cleaning, manipulation or even aggregations as per our need. There might be a lot of inconsistencies in the dataset like missing values, outliers, data type errors etc. which may deter us from creating the models appropriately. When we treat the dataset, it is a clean dataset which can further be used for aggregations and manipulations. Once we have the dataset ready, we can use it to build our model or in the case of the project, we can perform the clustering technique using k-means method. Before building any clustering model, it is very crucial that the optimal number of clusters are determined before doing it.

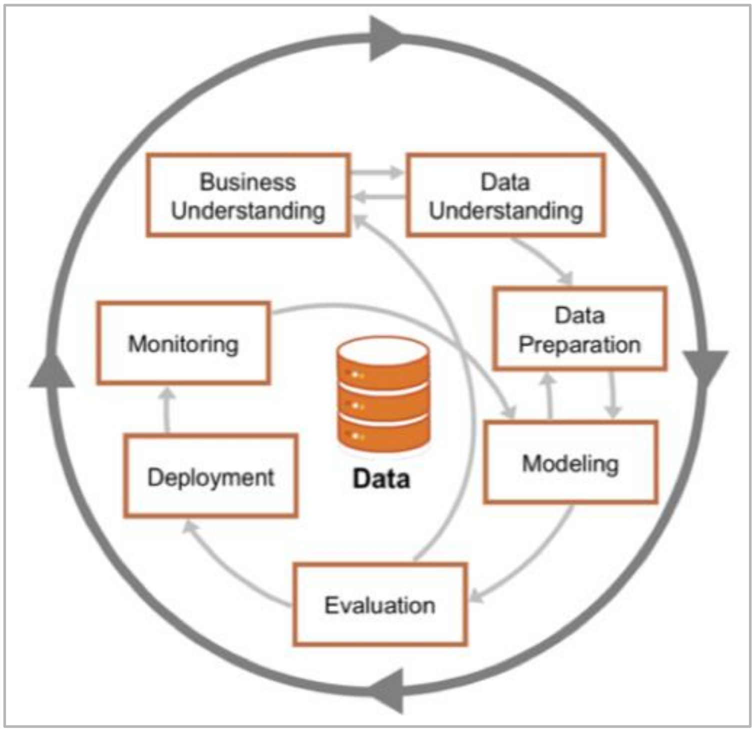


Fig.3.1. Data Analytics project lifecycle, CRISP-DM

The following steps are followed when it comes to a data analytics project lifecycle, especially while using the CRISP-DM structure to move ahead with solving the problem.

Business Understanding

We must start by understanding our industry and reading about use cases in the customer segmentation and clustering problem space, so that we get a good knowledge about the topic and are able to move ahead with good solution approaches while working on the same.

Data Understanding

In this step, the dataset is obtained and used to perform exploratory data analysis and different statistical models to understand in-depth about the data. We would start with exploring or getting an overview of the dataset like the mean prices, standard deviation of the predictor and other continuous features.

Data Preparation

In this step of the process, the dataset must be cleaned of inconsistencies and missing values after an overview exploration for missing values and data types for all features. This helps us perform the analyses better, for example the response column might have some missing values due to typing errors and this might result in biased results. Hence, we need to clean the data off these problems.

Data Modelling

Data Modelling is the process of using the clean and prepared dataset to make a machine learning model, train the data and be able to predict the outcome values in the final step. Once we understand the features in our dataset, we can move ahead to prepare our train and test set splitting. The split of the dataset allows us to train it, as well as test the dataset for accurate clustering results.

Evaluation

Now, this is the final step where we would need to validate the constructed model and check if the clusters formed are good from the model or not. We would need to see if the values in the clusters make sense to us, so that we do not derive the wrong information from the clusters.

Following the above-mentioned steps from the CRISP-DM lifecycle enables a structured flow to solving the entire problem and have accurate results for our stakeholders as well as recommendations. Hence, using such a process ensures a good project process and we plan to use the same to achieve our results and final output.

This helps in generating a very accurate model at the end with appropriate results with respect to cluster KPIs and different metrics. We finally move ahead to operationalise the model and present our findings and results to the relevant stakeholders. This is the end-to-end project implementation lifecycle that we plan to use in our project, with clear goals of having customer cohorts with appropriate traits as insights for each of these clusters. This will help businesses in identifying the correct promotion or products to market, as this would help in better conversion rates and profits in the long term.

3.2 Deliverables

In this project, we will use some of the most advanced and complex tools for analytics and Machine Learning. The tools have been developed by SPSS and now acquired by IBM, and they are used by marketing survey teams, as well as researchers and data scientists to derive insights and information from dataset. In the following section, we discuss a few more details about the same.

- SPSS Statistics - is a statistical and data manipulation tool which has been developed by SPSS and we will use this tool to perform statistical descriptive analysis to understand the different aspects of the dataset and determine the factor significance from the dataset. Not only that, but we would also be performing data cleaning and manipulation to shape the data well in shape
- SPSS Modeller - is used for building pipelines for Machine Learning models, along with data cleaning and manipulation techniques. This tool is used by researchers and statisticians to prepare data in the required format for predictions and other pipeline

For the scope of this report and research we will only be using the above tools as they are sufficient for the process. Statistics can be used to perform tasks like typecasting features, detecting, and removing outliers, and factor analysis to choose important features. Modellers on the other hand have the capabilities to shape and modify data using pipelines and then create advanced models along with comparative analyses.

Chapter 4 Business Understanding

4.1 Business Section

RFM analysis is the process of segmenting customers by using data and the full form of RFM is Recency, Frequency and Monetary value. All these properties define the customer segments and their propensity towards purchasing any article or responding towards marketing campaigns. RFM analysis is used extensively in businesses to monetise customers based on their purchase patterns and determining their preferences. Based on the customer segmentation, users are provided promotions and products based on their interests. For example, if the cluster algorithm determines three different clusters based on income like High, Medium and Low companies can provide promotions based on the income like providing expensive products to High income and providing offers and cheap products to the Low-income customers which ensures higher rate of conversion at the end. The advantages of using RFM analysis are manyfold, and below are some of the key advantages -

- Provides companies to send personalised offers and highly relevant products based on the customer's interests and this ensures that customers keep coming back for more through these high interest product offerings
- Ensures high conversion rate because of relevant customer product offerings and this ensures recurring business and revenue for the organisation in the long run

For the RFM analysis, we mentioned the three components, i.e., Recency, Frequency and Monetary and we would like to elaborate on the same below -

- Recency (R): The recency of customers based on their purchase. This identifies the time spent by the customer between each purchase thereby indicating the customer activity with the business
- Frequency (F): The frequency of customers with their purchase, i.e., this helps us to bifurcate users based on the purchase frequency and determine highly active users from the dormant users on the platform
- Monetary (M): The amount of money that the customer paid for their purchase. This helps us understand important KPIs like average basket value, basket economics etc. and the customer's value with the business

Chapter 5 Data Understanding

5.1 Data Description

Before we proceed with the data exploration and modelling, we want to explain about the dataset in an overview about where we obtained the same and what is the significance of the dataset. This dataset has been obtained from Kaggle which is a grocery store purchase dataset and tracks the customer's daily purchase information. The below is the description of each of the features that we have in the dataset. The objective behind creating the dataset was to determine who will respond to a marketing campaign based on the historical trend behind each customer. In the below section, we define each of the features as available in the dataset.

AcceptedCmp1 - 1 if customer accepted the offer in the 1st campaign, 0 otherwise

AcceptedCmp2 - 1 if customer accepted the offer in the 2nd campaign, 0 otherwise

AcceptedCmp3 - 1 if customer accepted the offer in the 3rd campaign, 0 otherwise

AcceptedCmp4 - 1 if customer accepted the offer in the 4th campaign, 0 otherwise

AcceptedCmp5 - 1 if customer accepted the offer in the 5th campaign, 0 otherwise

Response (target) - 1 if customer accepted the offer in the last campaign, 0 otherwise

DtCustomer - date of customer's enrolment with the company

Education - customer's level of education

Marital - customer's marital status

Kidhome - number of small children in customer's household

Teenhome - number of teenagers in customer's household

Income - customer's yearly household income

MntFishProducts - amount spent on fish products in the last 2 years

MntMeatProducts - amount spent on meat products in the last 2 years

MntFruits - amount spent on fruits products in the last 2 years

MntSweetProducts - amount spent on sweet products in the last 2 years

MntWines - amount spent on wine products in the last 2 years

MntGoldProds - amount spent on gold products in the last 2 years

NumDealsPurchases - number of purchases made with discount

NumCatalogPurchases - number of purchases made using catalogue

NumStorePurchases - number of purchases made directly in stores

NumWebPurchases - number of purchases made through company's website

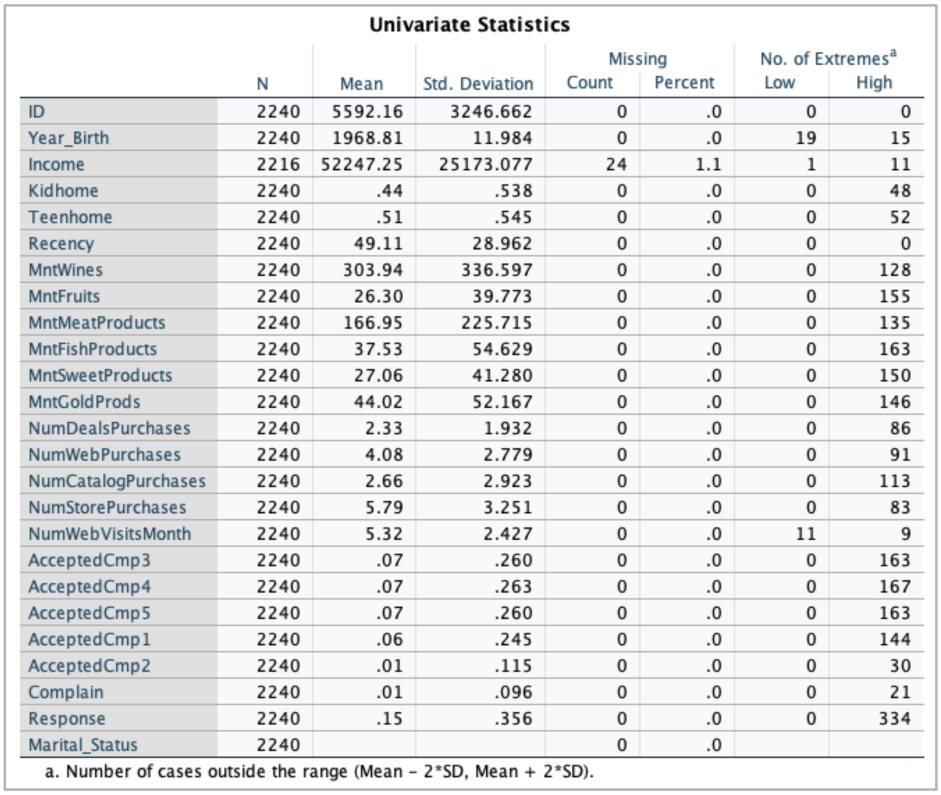
NumWebVisitsMonth - number of visits to company's web site in the last month

Recency - number of days since the last purchase

The dataset contains 24 features and 2240 variables in total, which we will be using to analyse and understand the problem statement better.

5.2 Data Cleaning

In this section, we will explore the dataset for inconsistencies and fix them for the next phase of our reporting. Based on the table below, we observe that the data has no missing value except for the feature Income which only has 1.1% missing value of all the rows. The outliers (with low extremes) are also not present for 95% of our features except for three which are Year_Birth, Income and NumWebVisitsMonth.



Univariate Statistics

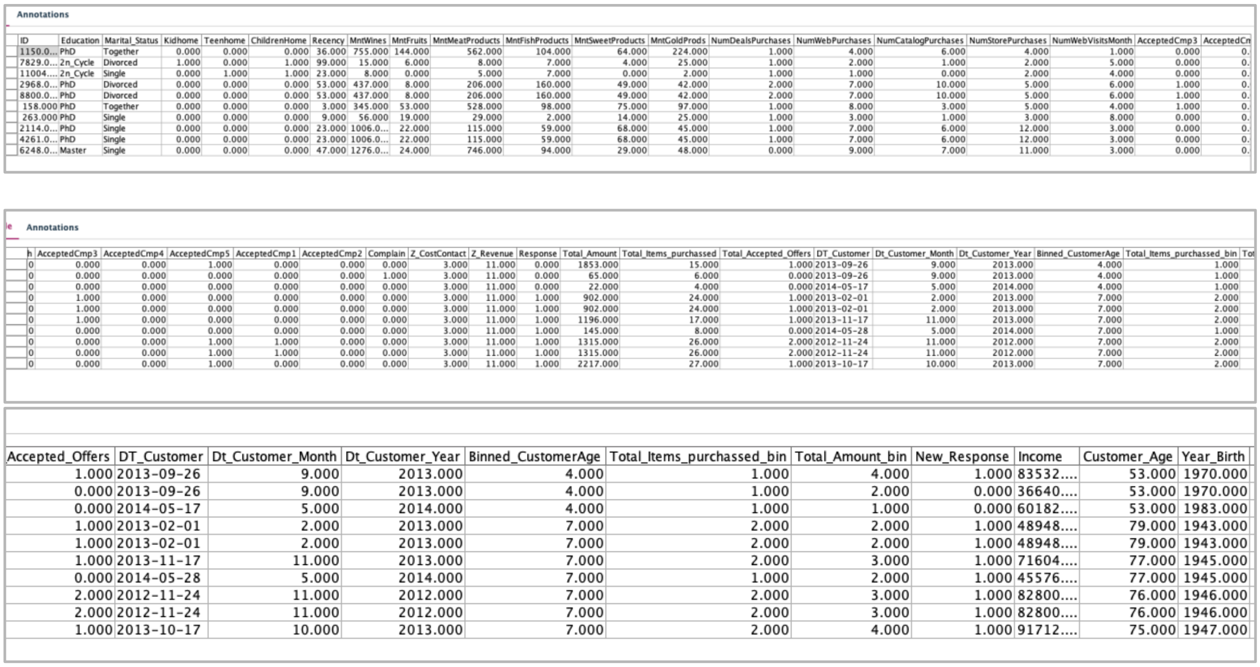
	N	Mean	Std. Deviation	Missing		No. of Extremes[a]	
				Count	Percent	Low	High
ID	2240	5592.16	3246.662	0	.0	0	0
Year_Birth	2240	1968.81	11.984	0	.0	19	15
Income	2216	52247.25	25173.077	24	1.1	1	11
Kidhome	2240	.44	.538	0	.0	0	48
Teenhome	2240	.51	.545	0	.0	0	52
Recency	2240	49.11	28.962	0	.0	0	0
MntWines	2240	303.94	336.597	0	.0	0	128
MntFruits	2240	26.30	39.773	0	.0	0	155
MntMeatProducts	2240	166.95	225.715	0	.0	0	135
MntFishProducts	2240	37.53	54.629	0	.0	0	163
MntSweetProducts	2240	27.06	41.280	0	.0	0	150
MntGoldProds	2240	44.02	52.167	0	.0	0	146
NumDealsPurchases	2240	2.33	1.932	0	.0	0	86
NumWebPurchases	2240	4.08	2.779	0	.0	0	91
NumCatalogPurchases	2240	2.66	2.923	0	.0	0	113
NumStorePurchases	2240	5.79	3.251	0	.0	0	83
NumWebVisitsMonth	2240	5.32	2.427	0	.0	11	9
AcceptedCmp3	2240	.07	.260	0	.0	0	163
AcceptedCmp4	2240	.07	.263	0	.0	0	167
AcceptedCmp5	2240	.07	.260	0	.0	0	163
AcceptedCmp1	2240	.06	.245	0	.0	0	144
AcceptedCmp2	2240	.01	.115	0	.0	0	30
Complain	2240	.01	.096	0	.0	0	21
Response	2240	.15	.356	0	.0	0	334
Marital_Status	2240			0	.0		

a. Number of cases outside the range (Mean – 2*SD, Mean + 2*SD).

Table. 5.1. Univariate Statistics of factors in the dataset

5.3 Data Preparation

In this section of the report, we want to understand the dataset by exploring a bit and then also performing some data preparation methods to have a better view of the data.



Annotations

ID	Education	Marital_Status	Kidhome	Teenhome	ChildrenHome	Recency	MntWines	MntFruits	MntMeatProducts	MntFishProducts	MntSweetProducts	MntGoldProds	NumDealsPurchases	NumWebPurchases	NumCatalogPurchases	NumStorePurchases	NumWebVisitsMonth	AcceptedCmp3	AcceptedCm
1150.0...	PhD	Together	0.000	0.000	0.000	36.000	755.000	144.000	562.000	104.000	64.000	224.000	1.000	4.000	6.000	4.000	1.000	0.000	0.
7829.0...	2n_Cycle	Divorced	1.000	0.000	1.000	99.000	15.000	6.000	8.000	7.000	4.000	25.000	1.000	2.000	1.000	2.000	5.000	0.000	0.
11004....	2n_Cycle	Single	0.000	1.000	1.000	23.000	8.000	0.000	5.000	7.000	0.000	2.000	1.000	1.000	0.000	2.000	4.000	0.000	0.
2968.0...	PhD	Divorced	0.000	0.000	0.000	53.000	437.000	8.000	206.000	160.000	49.000	42.000	2.000	7.000	10.000	5.000	6.000	1.000	0.
8800.0...	PhD	Divorced	0.000	0.000	0.000	53.000	437.000	8.000	206.000	160.000	49.000	42.000	2.000	7.000	10.000	5.000	6.000	1.000	0.
158.000	PhD	Together	0.000	0.000	0.000	3.000	345.000	53.000	528.000	98.000	75.000	97.000	1.000	8.000	3.000	5.000	4.000	1.000	0.
263.000	PhD	Single	0.000	0.000	0.000	9.000	56.000	19.000	29.000	2.000	14.000	25.000	1.000	3.000	1.000	3.000	8.000	0.000	0.
2114.0...	PhD	Single	0.000	0.000	0.000	23.000	1006.0...	22.000	115.000	59.000	68.000	45.000	1.000	7.000	6.000	12.000	3.000	0.000	0.
4261.0...	PhD	Single	0.000	0.000	0.000	23.000	1006.0...	22.000	115.000	59.000	68.000	45.000	1.000	7.000	6.000	12.000	3.000	0.000	0.
6248.0...	Master	Single	0.000	0.000	0.000	47.000	1276.0...	24.000	746.000	94.000	29.000	48.000	0.000	9.000	7.000	11.000	3.000	0.000	0.

Annotations

h	AcceptedCmp3	AcceptedCmp4	AcceptedCmp5	AcceptedCmp1	AcceptedCmp2	Complain	Z_CostContact	Z_Revenue	Response	Total_Amount	Total_Items_purchassed	Total_Accepted_Offers	DT_Customer	Dt_Customer_Month	Dt_Customer_Year	Binned_CustomerAge	Total_Items_purchassed_bin	Tot
0	0.000	0.000	1.000	0.000	0.000	0.000	3.000	11.000	0.000	1853.000	15.000	1.000	2013-09-26	9.000	2013.000	4.000	1.000	
0	0.000	0.000	0.000	0.000	0.000	1.000	3.000	11.000	0.000	65.000	6.000	0.000	2013-09-26	9.000	2013.000	4.000	1.000	
0	0.000	0.000	0.000	0.000	0.000	0.000	3.000	11.000	0.000	22.000	4.000	0.000	2014-05-17	5.000	2014.000	4.000	1.000	
0	1.000	0.000	0.000	0.000	0.000	0.000	3.000	11.000	1.000	902.000	24.000	1.000	2013-02-01	2.000	2013.000	7.000	2.000	
0	1.000	0.000	0.000	0.000	0.000	0.000	3.000	11.000	1.000	902.000	24.000	1.000	2013-02-01	2.000	2013.000	7.000	2.000	
0	1.000	0.000	0.000	0.000	0.000	0.000	3.000	11.000	1.000	1196.000	17.000	1.000	2013-11-17	11.000	2013.000	7.000	2.000	
0	0.000	0.000	0.000	0.000	0.000	0.000	3.000	11.000	1.000	145.000	8.000	0.000	2014-05-28	5.000	2014.000	7.000	1.000	
0	0.000	0.000	1.000	1.000	0.000	0.000	3.000	11.000	1.000	1315.000	26.000	2.000	2012-11-24	11.000	2012.000	7.000	2.000	
0	0.000	0.000	1.000	1.000	0.000	0.000	3.000	11.000	1.000	1315.000	26.000	2.000	2012-11-24	11.000	2012.000	7.000	2.000	
0	0.000	0.000	1.000	0.000	0.000	0.000	3.000	11.000	1.000	2217.000	27.000	1.000	2013-10-17	10.000	2013.000	7.000	2.000	

Accepted_Offers	DT_Customer	Dt_Customer_Month	Dt_Customer_Year	Binned_CustomerAge	Total_Items_purchassed_bin	Total_Amount_bin	New_Response	Income	Customer_Age	Year_Birth
1.000	2013-09-26	9.000	2013.000	4.000	1.000	4.000	1.000	83532....	53.000	1970.000
0.000	2013-09-26	9.000	2013.000	4.000	1.000	2.000	0.000	36640....	53.000	1970.000
0.000	2014-05-17	5.000	2014.000	4.000	1.000	1.000	0.000	60182....	53.000	1983.000
1.000	2013-02-01	2.000	2013.000	7.000	2.000	2.000	1.000	48948....	79.000	1943.000
1.000	2013-02-01	2.000	2013.000	7.000	2.000	2.000	1.000	48948....	79.000	1943.000
1.000	2013-11-17	11.000	2013.000	7.000	2.000	3.000	1.000	71604....	77.000	1945.000
0.000	2014-05-28	5.000	2014.000	7.000	1.000	2.000	1.000	45576....	77.000	1945.000
2.000	2012-11-24	11.000	2012.000	7.000	2.000	3.000	1.000	82800....	76.000	1946.000
2.000	2012-11-24	11.000	2012.000	7.000	2.000	3.000	1.000	82800....	76.000	1946.000
1.000	2013-10-17	10.000	2013.000	7.000	2.000	4.000	1.000	91712....	75.000	1947.000

Table.5.2. Data view for the top n rows before cleaning

Below is a snapshot of the dataset in SPSS Statistics which shows all the rows that we have of the raw dataset obtained from Kaggle. Each row identifies a customer who responded or did not respond to a marketing campaign that was rolled out. Next, we perform the factor analysis in SPSS Statistics along with determining the Mahalanobis Distance for each variable. Then the p-value is estimated for each value based on the Mahalanobis Distance. If the p-value is less than 0.001 then the row is termed an outlier, otherwise not.

Mahalanobis Distance: it represents the distance between two points in multivariate, used to detect the outliers.

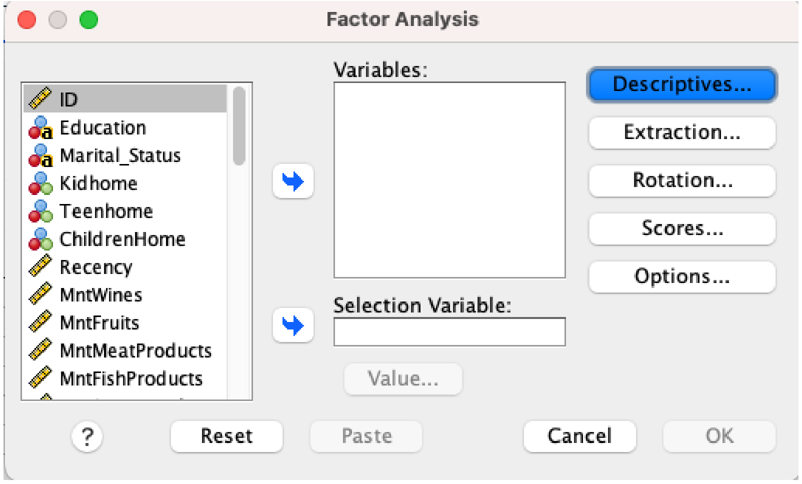
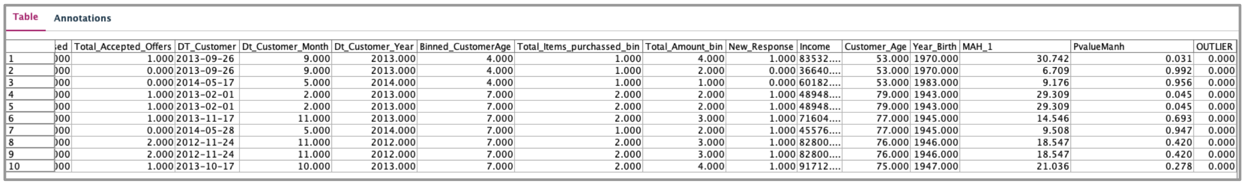


Fig.5.1. Factor analysis screenshot from SPSS Statistics

These processes have helped us prepare the dataset in a required format which can then be used for the modelling phase in the next step. Having a clean and standardised dataset is imperative to data analytics and modelling to prevent any form of biases in the output results. These are mainly caused due to extreme values, missing data points and some other factors which need to be checked during the statistical analysis and data cleaning.



	;ed	Total_Accepted_Offers	DT_Customer	Dt_Customer_Month	Dt_Customer_Year	Binned_CustomerAge	Total_Items_purchassed_bin	Total_Amount_bin	New_Response	Income	Customer_Age	Year_Birth	MAH_1	PvalueManh	OUTLIER
1	)00	1.000	2013-09-26	9.000	2013.000	4.000	1.000	4.000	1.000	83532....	53.000	1970.000	30.742	0.031	0.000
2	)00	0.000	2013-09-26	9.000	2013.000	4.000	1.000	2.000	0.000	36640....	53.000	1970.000	6.709	0.992	0.000
3	)00	0.000	2014-05-17	5.000	2014.000	4.000	1.000	1.000	0.000	60182....	53.000	1983.000	9.176	0.956	0.000
4	)00	1.000	2013-02-01	2.000	2013.000	7.000	2.000	2.000	1.000	48948....	79.000	1943.000	29.309	0.045	0.000
5	)00	1.000	2013-02-01	2.000	2013.000	7.000	2.000	2.000	1.000	48948....	79.000	1943.000	29.309	0.045	0.000
6	)00	1.000	2013-11-17	11.000	2013.000	7.000	2.000	3.000	1.000	71604....	77.000	1945.000	14.546	0.693	0.000
7	)00	0.000	2014-05-28	5.000	2014.000	7.000	1.000	2.000	1.000	45576....	77.000	1945.000	9.508	0.947	0.000
8	)00	2.000	2012-11-24	11.000	2012.000	7.000	2.000	3.000	1.000	82800....	76.000	1946.000	18.547	0.420	0.000
9	)00	2.000	2012-11-24	11.000	2012.000	7.000	2.000	3.000	1.000	82800....	76.000	1946.000	18.547	0.420	0.000
10	)00	1.000	2013-10-17	10.000	2013.000	7.000	2.000	4.000	1.000	91712....	75.000	1947.000	21.036	0.278	0.000

Table.5.3. Data view for top n rows after cleaning

In the above screenshot, we have the view of the data after we have determined the Mahalanobis Distance, along with the p-value and outlier flag for the dataset. Hence, we are free to remove rows which have the Outlier flag as 1 indicating that the row is an outlier for the analysis and modelling phase in the next steps.

Chapter 6 - Data Analytics

6.1 Exploratory Data Analysis

Exploratory data analysis is used to understand the data distribution along with the relationship of features with one another. This section is dedicated to visualisations and insights generated from the EDA phase which helps us understand more about the data so that we can make informed decisions for the modelling phase.

In the below figure (Fig.1.), we see that the income is right skewed in terms of the distribution with the average income being 52247 and the standard deviation being 25037. This clearly indicates that we have a lot of data points with outlier incomes, and maybe we can filter them to have an even distribution.

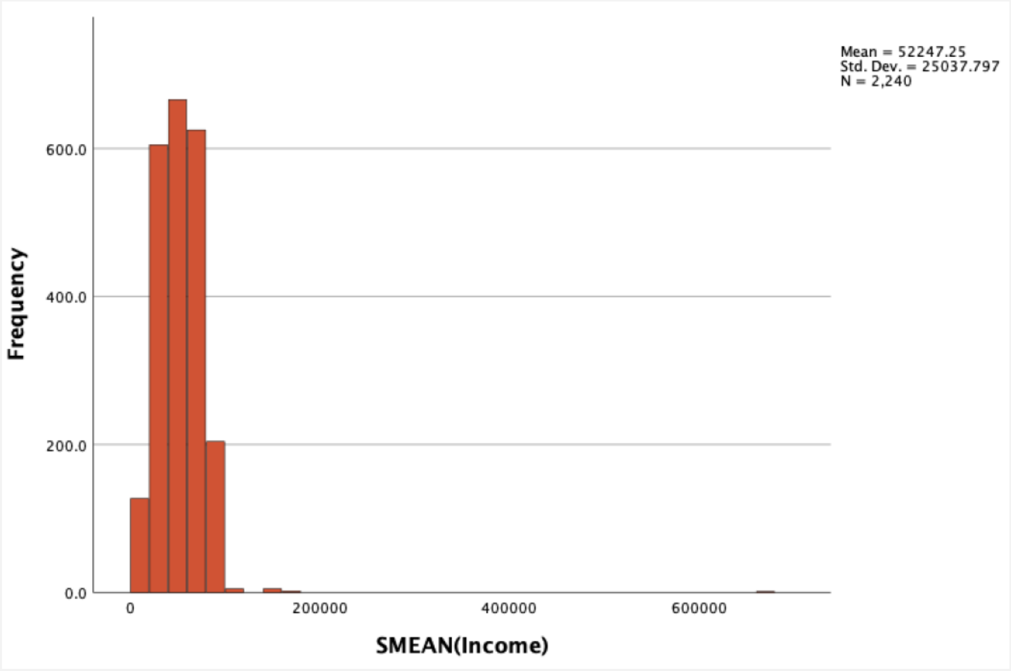


Fig.6.1. Distribution of the Income variable in the dataset

Now, we remove the outlier values of the income feature and are left with the histogram distribution below. With the data evenly spread, we understand that the average income is 51640 with the standard deviation being 20601 with a total of 8 outlier values removed.

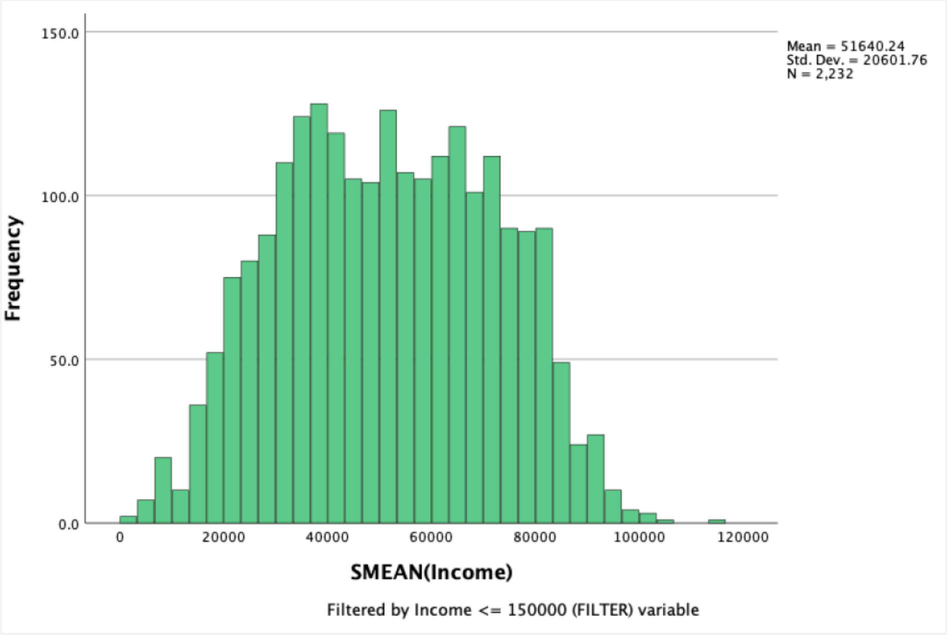


Fig.6.2. Distribution of the Income variable with outliers removed

To understand a bit more about the distribution of our customer base based on different demographics, we will plot the same below. In the following section, we want to understand customer distribution based on different factors like education, marital status, income bucket as well as spend bucket. This will help us understand the types of customers that engage with the supermarket.

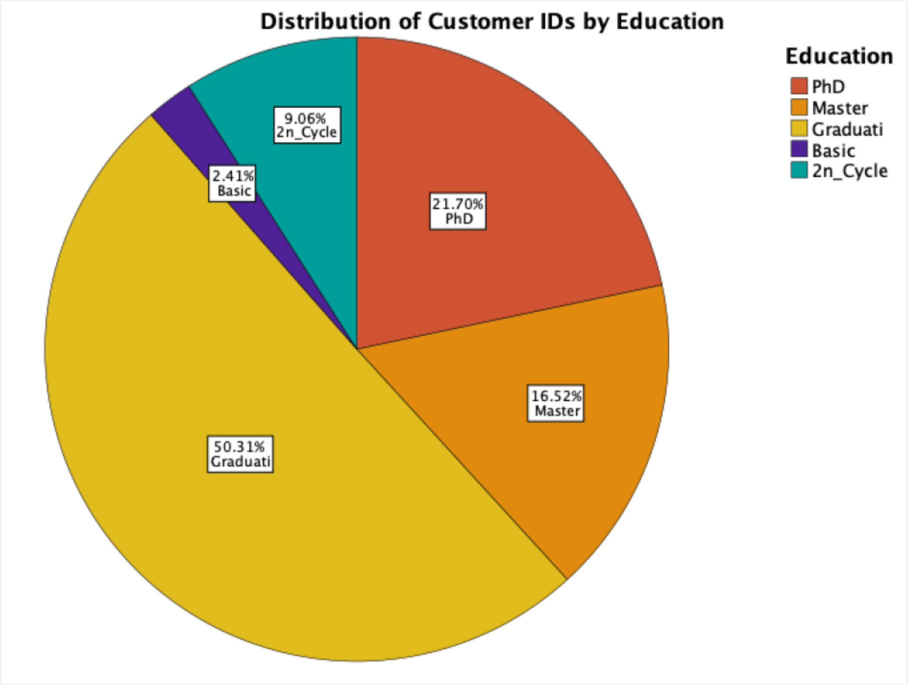


Fig.6.3. Distribution of customers by Education

The above distribution of our customer base shows that 50% of them are graduated with some degree. While the other major groups are the PhDs or master's customers. This helps us understand the distribution of the customers based on their qualifications and prepare and roll out offers accordingly. Stores often use these types of understanding to roll out personalised content or marketing for their users.

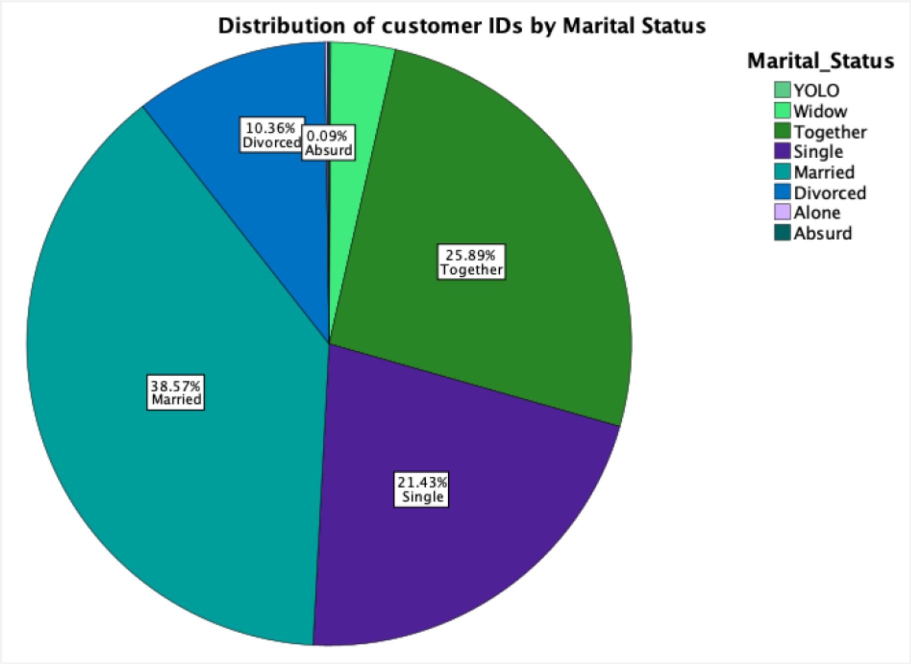


Fig.6.4. Distribution of customers by Marital Status

Majority of our customer base are either married or together with a partner (38% and 25.8% respectively). Since they are partnered some way or the other, it would be worthwhile to showcase promotions suited for families or couples. These are some ways of tailoring promotions based on the customer demographics and their distribution in the store.

In the following section, we plot the age distribution of the customers and observe that the average age is 53 for all our customers, with most of the customers being between 40 and 60. This makes sense because it is adults who mostly shop with stores and register themselves instead of their children or someone else.

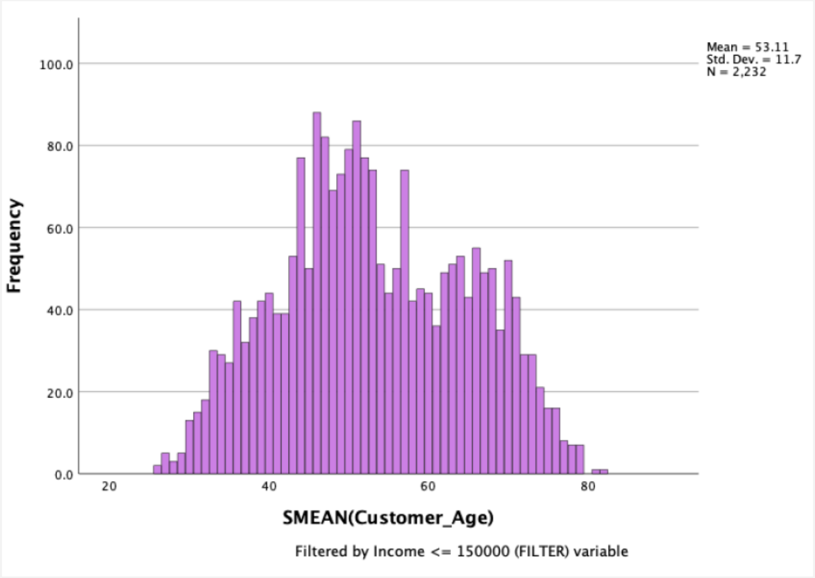
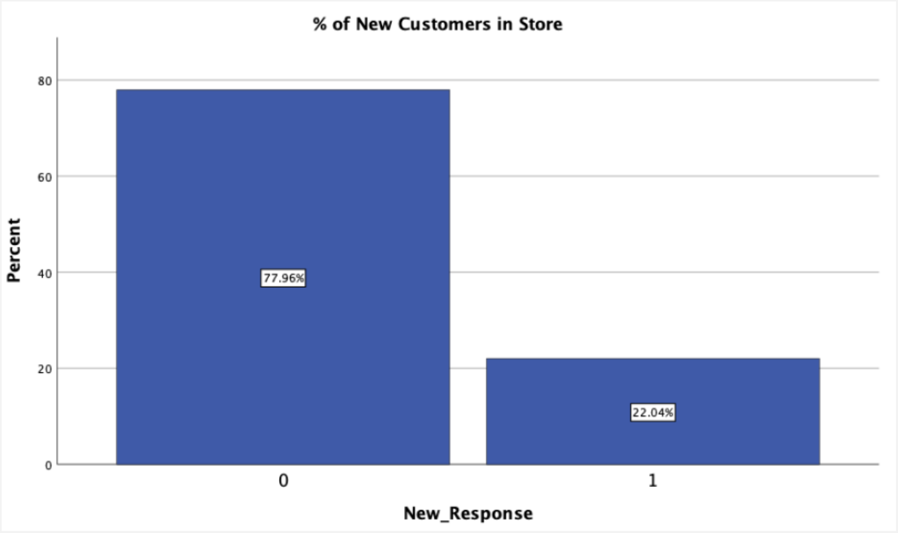


Fig.6.5. Customer Age distribution



We would also like to understand the total new responses that we have received in the store database, which means whether the customers that responded to the promotion are new or existing customers. We notice that 22% of the customers are new respondents while the rest of the 77.9% are old ones. This helps us understand the new customer base vs the existing customers and plan promotions for our new customers to keep them coming back and convert to long term customers with the store.

In Fig.6., we plot the average spends of the customers by their educational qualification and see that customers with PhDs tend to spend more on an average compared to the rest. This makes sense because PhD qualified customers would have higher income bands compared to the rest of the segments.

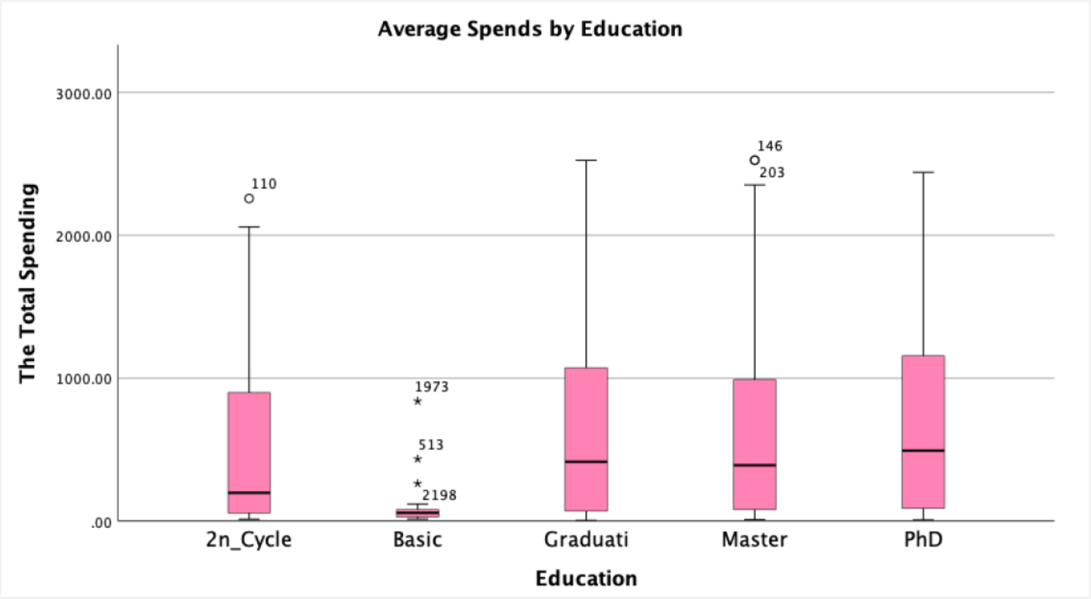


Fig.6.6. Box plot of spends by educational qualification for customers

We do the same plot for average items purchased distributed across different age groups and notice that the average items purchased is highest for customers with 74+ age while second in line is 58-66. This tells us the average basket size of each customer age buckets and helps us understand the bucket with the highest basket average.

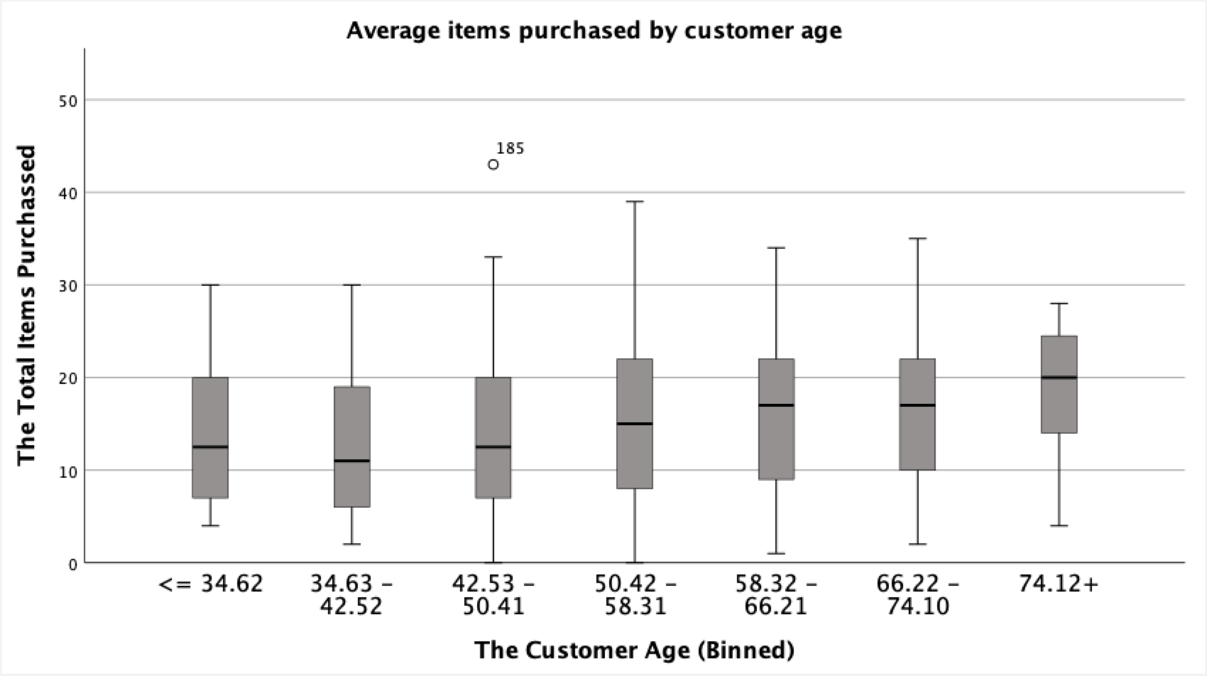


Fig.6.7. Average items purchased by customer age buckets

Finally, in Fig.8., we see the relationship between total spends by total items purchased by the customers and see that a relationship exists between these two features. The total spending increases as the total items within a basket increase in the data. With these linear dependencies, we can understand that the regression models would be explained well with these types of dependencies and collinearity.

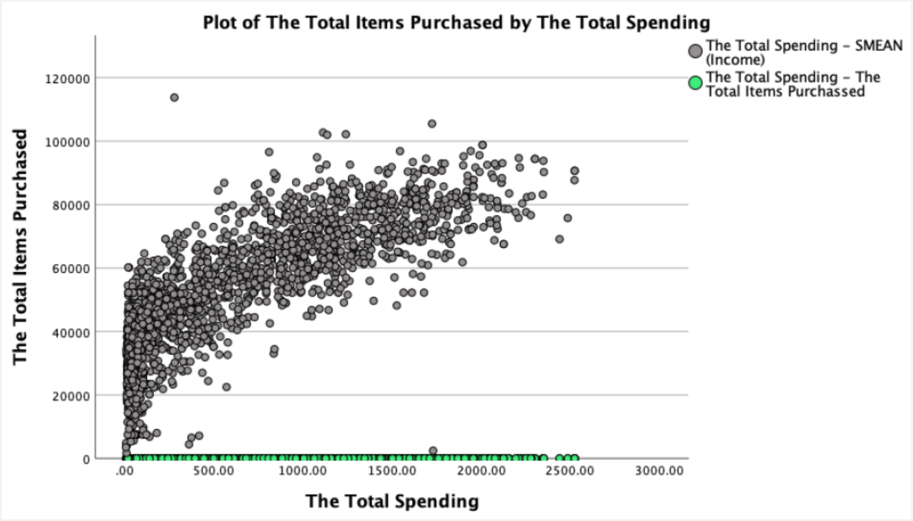


Fig.6.8. Scatter plot of total spends by total items purchased

6.2 Modelling

In this section of the report, we want to implement various statistical analyses and modelling techniques to be able to solve our problem statement. We will use the cleaned dataset in a tool called SPSS Modeler to perform the modelling part of the experiment. Our modelling would consist of factor significance determination along with implementing multiple models to perform a comparative study of the models based on their performances.

6.2.1 RFM Analysis

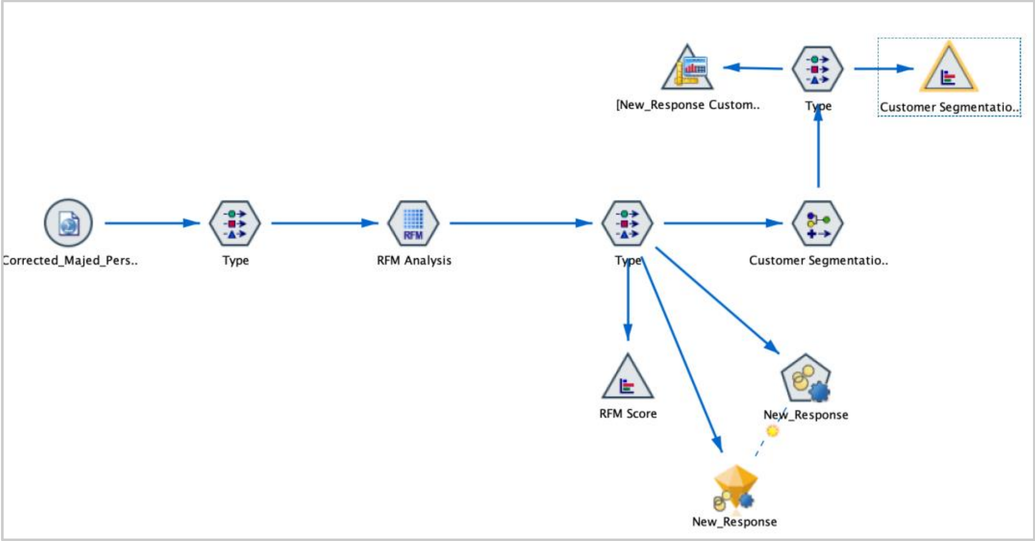
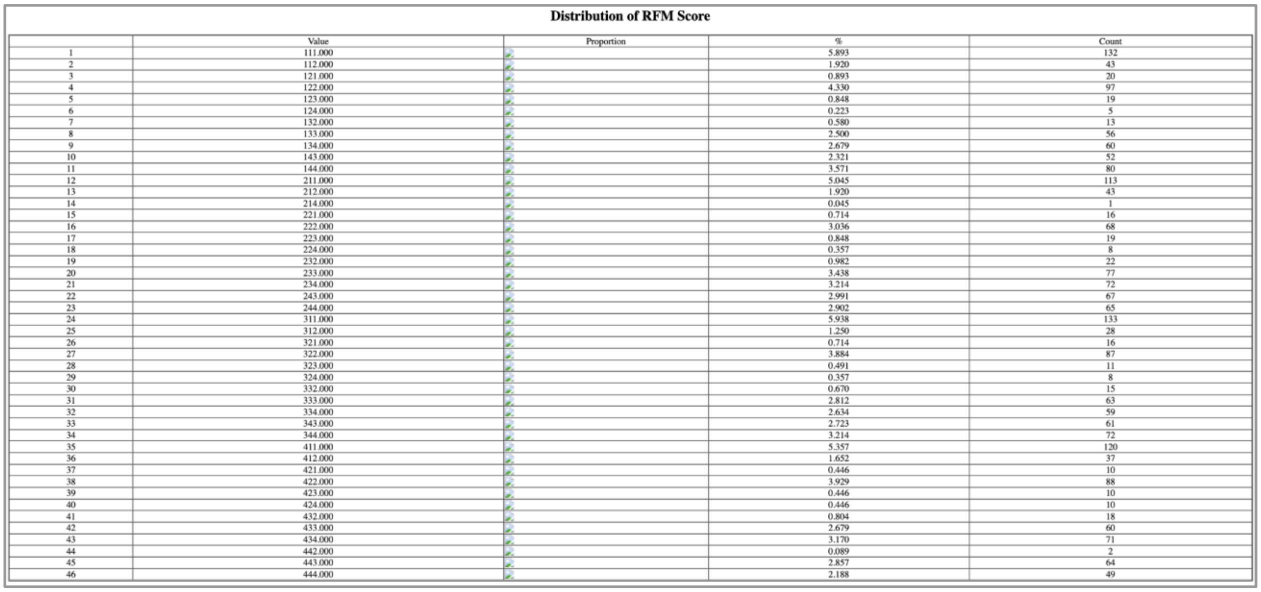


Fig.6.2.1. SPSS Modeller pipeline for RFM Analysis and Clustering

customers loyal and continue business for the company.



Distribution of RFM Score

	Value	Proportion	%	Count
1	111.000		5.893	132
2	112.000		1.920	43
3	121.000		0.893	20
4	122.000		4.330	97
5	123.000		0.848	19
6	124.000		0.223	5
7	132.000		0.580	13
8	133.000		2.500	56
9	134.000		2.679	60
10	143.000		2.321	52
11	144.000		3.571	80
12	211.000		5.045	113
13	212.000		1.920	43
14	214.000		0.045	1
15	221.000		0.714	16
16	222.000		3.036	68
17	223.000		0.848	19
18	224.000		0.357	8
19	232.000		0.982	22
20	233.000		3.438	77
21	234.000		3.214	72
22	243.000		2.991	67
23	244.000		2.902	65
24	311.000		5.938	133
25	312.000		1.250	28
26	321.000		0.714	16
27	322.000		3.884	87
28	323.000		0.491	11
29	324.000		0.357	8
30	332.000		0.670	15
31	333.000		2.812	63
32	334.000		2.634	59
33	343.000		2.723	61
34	344.000		3.214	72
35	411.000		5.357	120
36	412.000		1.652	37
37	421.000		0.446	10
38	422.000		3.929	88
39	423.000		0.446	10
40	424.000		0.446	10
41	432.000		0.804	18
42	433.000		2.679	60
43	434.000		3.170	71
44	442.000		0.089	2
45	443.000		2.857	64
46	444.000		2.188	49

Figure 6.2.2 Distribution of RFM Scores

In the above table, we also have the distribution of the RFM scores obtained from the above steps which is indicative of the number of records we have within each score. The corresponding RFM value is also shown in the table below which helps us define each segment based on the same value.

Segment	**RFM**	**Description**	**Marketing**
Best Customers	444	Who tend to buy the most	No discounts required
Loyal Customers	X4X	Who purchased very recently	Implementing R and M for detailed segments
Big Spenders	XX4	Who spend the most	Market expensive and luxury items
Almost Lost	244	Didn't purchase anything for a long time	Send many discount programs
Lost Customers	144	Didn't purchase anything for a long time	Send many discount programs
Lost Cheap Customers	111	Long time purchase gap and spends very less	Not required to spend on them

Table 6.1. RFM Scores

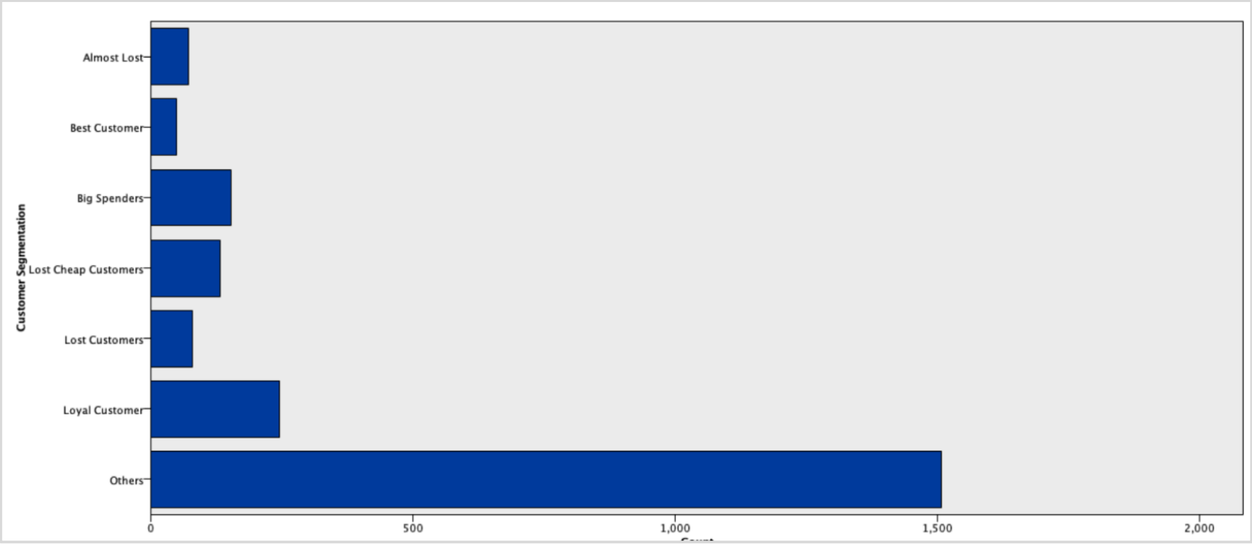
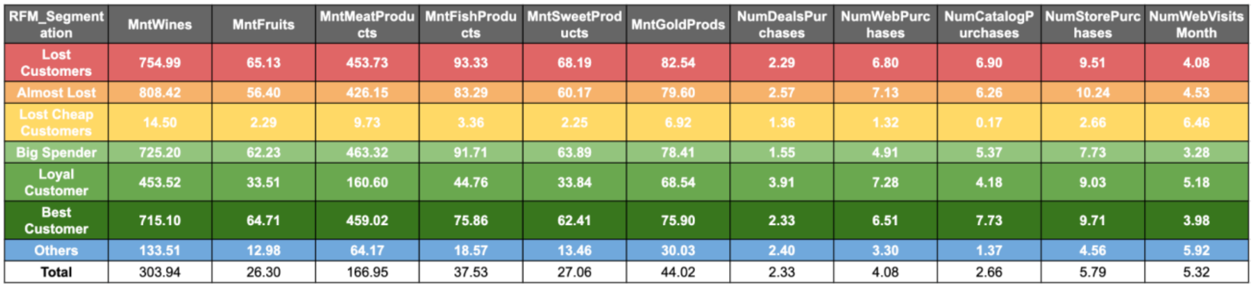


Fig.6.2.3. Customer Segments from RFM

From the RFM scoring and customer segmentation, we obtain the following customer types and their distribution in the dataset. While most of the customers have been segmented as "Others", the first majority are "Loyal Customers", after which we have the "Big Spenders" and then the "Lost Cheap

Customers”. A brief understanding from the chart is that we would need to increase the number of Loyal, Big Spenders, Best Customers while we would need to decrease the number of Lost customers or Cheap Customers in our customer base as they can prove detrimental for the business.



RFM_Segmentation	MntWines	MntFruits	MntMeatProducts	MntFishProducts	MntSweetProducts	MntGoldProds	NumDealsPurchases	NumWebPurchases	NumCatalogPurchases	NumStorePurchases	NumWebVisitsMonth
Lost Customers	754.99	65.13	453.73	93.33	68.19	82.54	2.29	6.80	6.90	9.51	4.08
Almost Lost	808.42	56.40	426.15	83.29	60.17	79.60	2.57	7.13	6.26	10.24	4.53
Lost Cheap Customers	14.50	2.29	9.73	3.36	2.25	6.92	1.36	1.32	0.17	2.66	6.46
Big Spender	725.20	62.23	463.32	91.71	63.89	78.41	1.55	4.91	5.37	7.73	3.28
Loyal Customer	453.52	33.51	160.60	44.76	33.84	68.54	3.91	7.28	4.18	9.03	5.18
Best Customer	715.10	64.71	459.02	75.86	62.41	75.90	2.33	6.51	7.73	9.71	3.98
Others	133.51	12.98	64.17	18.57	13.46	30.03	2.40	3.30	1.37	4.56	5.92
Total	303.94	26.30	166.95	37.53	27.06	44.02	2.33	4.08	2.66	5.79	5.32

Fig.6.2.4. Customer Buying Behaviour from RFM segmentation

In the above table, we have obtained the different Segmentations of customers based on the RFM scoring and have been able to determine the type of customers based on their buying behaviour. The corresponding average for different products and activities are also obtained in Figure 6.2.3. which indicated the average propensity that the customer in each segment might purchase. From the table we have obtained two segments of customers which are bad for any business, i.e., Lost Customers and Almost Lost customers. These types of customers are about to be lost by the company and any form of customer loss is loss of business for the same. On the other hand, we also obtain the profitable segment of customers from the above table which are the Big Spenders, Loyal Customers and Best Customers which need to be retained in the long run for continued business profitability. Various offers and promotions should be directed towards both the above customer sects to retain them. One of the common techniques that companies use to reactivate churned or almost lost customers is by sending notifications with great offers and discounts which helps the customers come back to shop with the sellers again. These numerous tactics can be used to keep

6.2.2 Classification

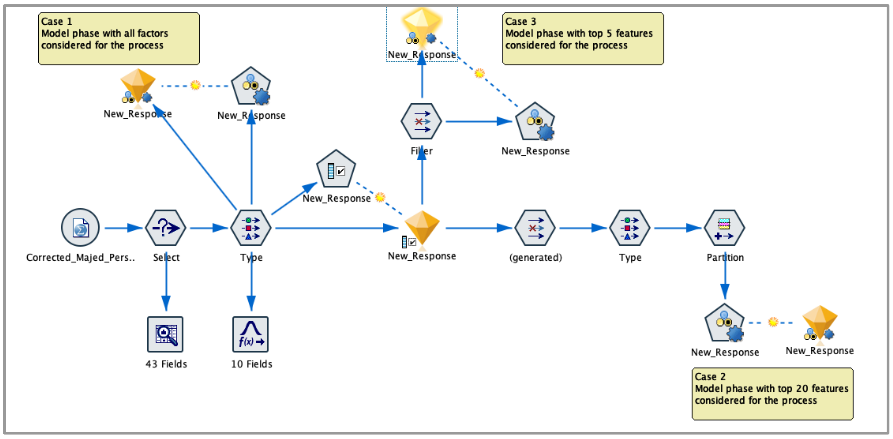
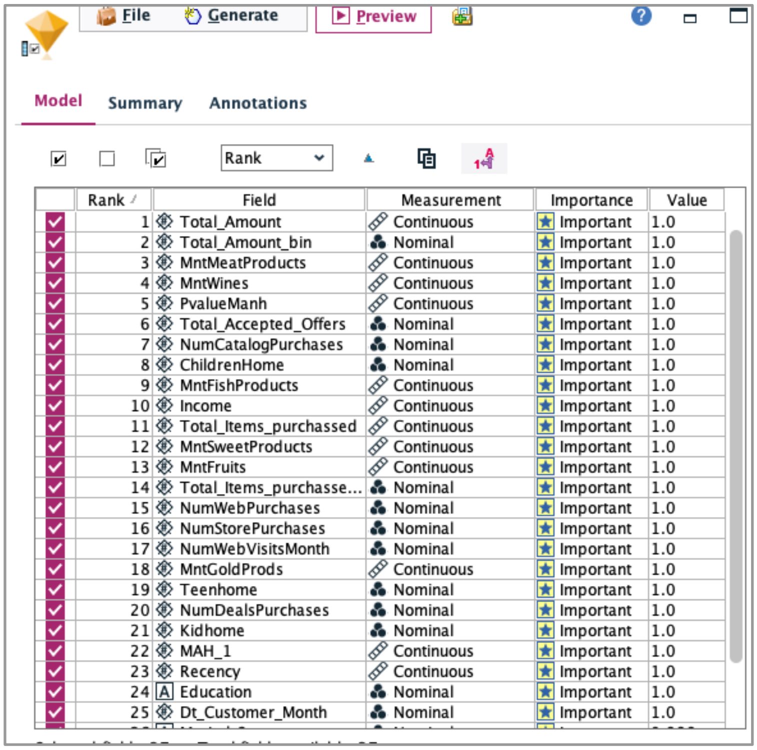


Fig.6.9. Model pipeline steps in SPSS Modeler

In the above pipeline setup, we have performed the following steps to draw conclusions at the end of the process. They are -

- We imported the cleaned dataset which contains the factor analysis and outlier detection steps from the SPSS Statistics tool
- In the next step we typecast the dataset and explored a bit about the distribution of the features, which gives us an overview of the data distribution
- Here we perform a part of the modelling called feature importance scoring technique. This step determines the important features that impact the target feature based on correlation and other factors
- In the next step we pick the important features (with values above 0.98) and then partition the dataset into train and test split (which is a 70-30 split for the train and test sets)
- Finally, we implement the ML models to determine the comparative scores of the different models based on various parameters which we will discuss in detail in the subsequent sections



	Rank	Field	Measurement	Importance	Value
☑	1	Total_Amount	Continuous	Important	1.0
☑	2	Total_Amount_bin	Nominal	Important	1.0
☑	3	MntMeatProducts	Continuous	Important	1.0
☑	4	MntWines	Continuous	Important	1.0
☑	5	PvalueManh	Continuous	Important	1.0
☑	6	Total_Accepted_Offers	Nominal	Important	1.0
☑	7	NumCatalogPurchases	Nominal	Important	1.0
☑	8	ChildrenHome	Nominal	Important	1.0
☑	9	MntFishProducts	Continuous	Important	1.0
☑	10	Income	Continuous	Important	1.0
☑	11	Total_Items_purchassed	Continuous	Important	1.0
☑	12	MntSweetProducts	Continuous	Important	1.0
☑	13	MntFruits	Continuous	Important	1.0
☑	14	Total_Items_purchasse...	Nominal	Important	1.0
☑	15	NumWebPurchases	Nominal	Important	1.0
☑	16	NumStorePurchases	Nominal	Important	1.0
☑	17	NumWebVisitsMonth	Nominal	Important	1.0
☑	18	MntGoldProds	Continuous	Important	1.0
☑	19	Teenhome	Nominal	Important	1.0
☑	20	NumDealsPurchases	Nominal	Important	1.0
☑	21	Kidhome	Nominal	Important	1.0
☑	22	MAH_1	Continuous	Important	1.0
☑	23	Recency	Continuous	Important	1.0
☑	24	Education	Nominal	Important	1.0
☑	25	Dt_Customer_Month	Nominal	Important	1.0

Table.6.2. Table for Factor Significance

From the factor significance scoring, the above features have been identified to be the most important for the modelling process. Some of the top few are the Total Amount, Total Amount bin, Mnt Meat Products, Mnt Wines and PvalueManh. In the subsequent steps we will try to model based on different scenarios from the findings from our previous steps. Below are some of the scenarios we will try to explore using the SPSS Modeler tool -

1. Using all features to build our baseline comparative model study to determine the best fit model based on various parameters
2. Using only top 20 important features to implement in the modelling process and determine the model performance parameters
3. Using only top 5 important features and determine the model performance results based on comparative study

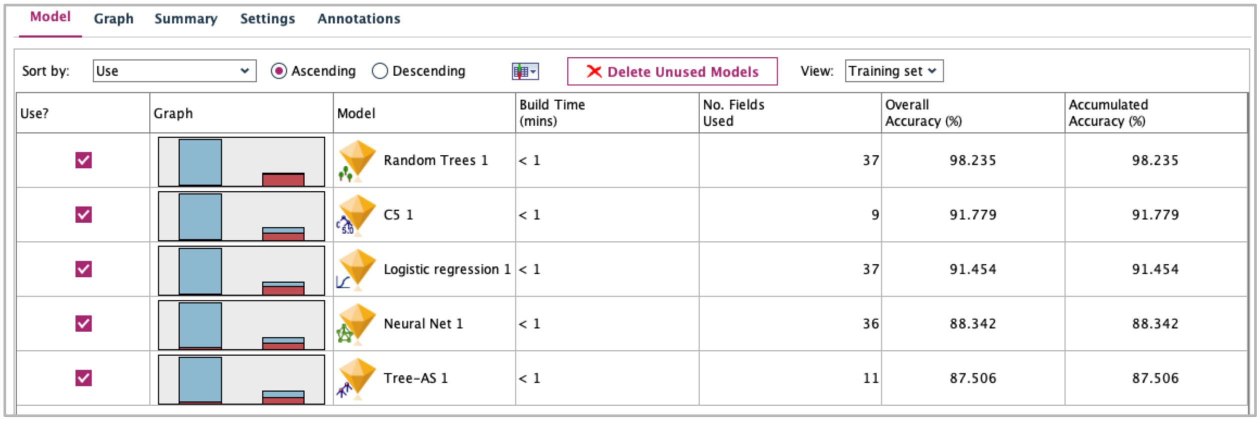
Moreover, the list of models that we tend to use in the below sections are -

- Random Trees. This is a process of ensemble learning which can be used for solving classification, regression problems by using a collection of decision trees
- C5. This node in SPSS Modeler uses either decision tree or rule set to derive a predictive approach. Here, the approach of splitting the tree again and again is followed in an iterative manner

- Logistic Regression. This model estimates the event and its probability of occurrence based on the set of independent variables in the dataset. The outcome value should be between 1 and 0 since it is either of the above
- Neural Network. These are ANNs which mimic the human brain with the set of input, output and hidden layers to decide on the probable outcome from the provided input features
- Tree-AS. This module of SPSS Modeler allows the construction of decision trees by using either CHAID or Exhaustive CHAID model

Scenario 1

In the first case of the model implementation, we use all the features from the dataset to build a comparative model study. In the below figure, we see that Random Trees, C5, Logistic Regression, Neural Network and Tree-AS have been implemented. Out of all, Random Trees has the best performance with an overall accuracy of 98.2% and the second best being 91.7%. This can be treated as the ground for a comparative study with the other cases now.



Use?	Graph	Model	Build Time (mins)	No. Fields Used	Overall Accuracy (%)	Accumulated Accuracy (%)
☑		Random Trees 1	< 1	37	98.235	98.235
☑		C5 1	< 1	9	91.779	91.779
☑		Logistic regression 1	< 1	37	91.454	91.454
☑		Neural Net 1	< 1	36	88.342	88.342
☑		Tree-AS 1	< 1	11	87.506	87.506

Table.6.3. Model performance summary from scenario 1

Having chosen the baseline with the above scenario, considering all features and the models we can now move on to the next scenario wherein we will choose only selected features based on our model interpretations of factor significance.

We now plot the ROC curve for the best performing model from the first scenario and observe that the Sensitivity or the True Positive Rate of the model is well above 0.95 and peaks around 0.98. While the False Positive Rate, i.e., the (1-Specificity) is also around 0.14 which is a good indication of a model score performance. We also have the Gain plot which shows that the value peak towards 98 and this is an indicator for a good model performance as it allows us to understand the predictive capability of the model through the curves in the chart.

ROC: Stands for receiver operating characteristic curve, it is a graph which presents the performance of a classification model at different classification thresholds, it has 2 parameters

- *True Positive Rate.*
- *False Positive Rate.*

AUC: Stands for area under the ROC curve, its purpose is to measure the entire two-dimensional areas under the ROC curve

Gain: used to assess the effectiveness of the classification model. They evaluate how much better one can expect to do when using a predictive model as opposed to not using one

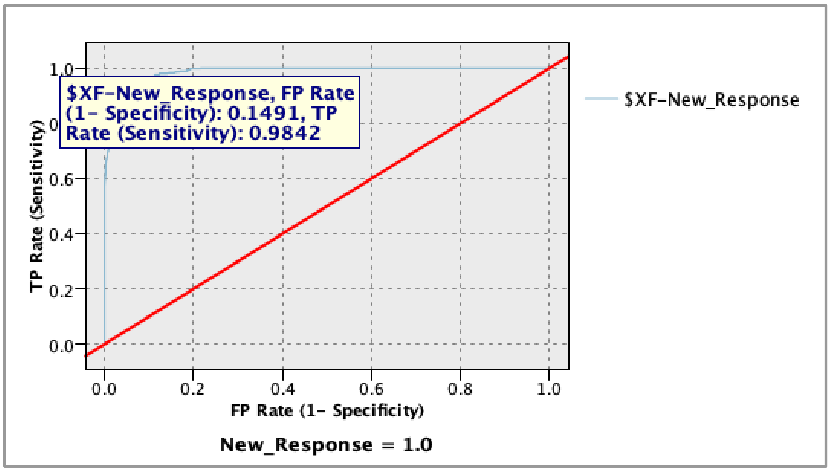


Fig.6.10. ROC Curve for best fit model from scenario 1

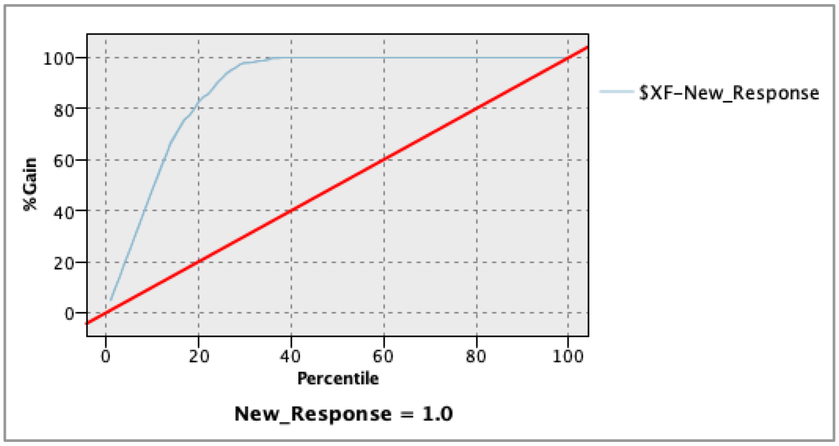
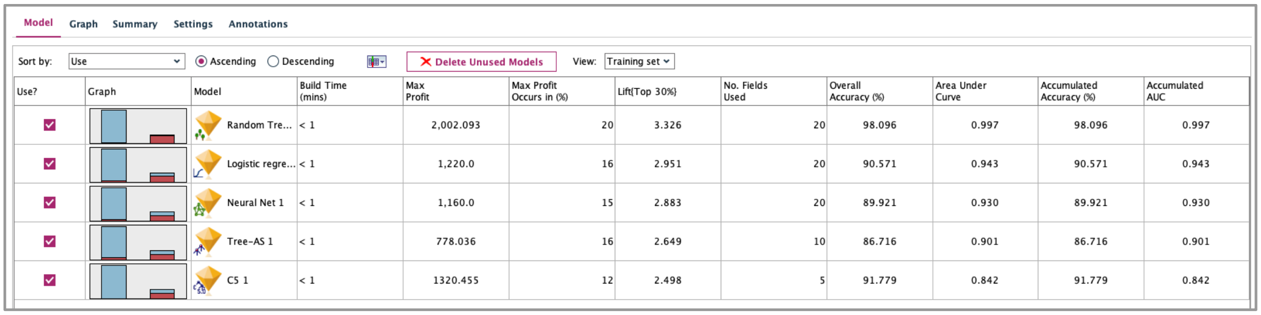


Fig.6.10. Gain for best fit model from scenario 1

Scenario 2

In the next scenario, we would like to use only the top 20 features recommended from the factor significance (as shown in Table.6.1.). As shown in the table below, Random Trees, C5 and Logistic

Regression have been identified as the top performing models with the accuracies being 98%, 91.7% and 90.5% respectively with the other parameters like AUC, Accumulated AUC also being well for the model.



Use?	Graph	Model	Build Time (mins)	Max Profit	Max Profit Occurs in (%)	Lift{Top 30%}	No. Fields Used	Overall Accuracy (%)	Area Under Curve	Accumulated Accuracy (%)	Accumulated AUC
☑		Random Tre...	< 1	2,002.093	20	3.326	20	98.096	0.997	98.096	0.997
☑		Logistic regre...	< 1	1,220.0	16	2.951	20	90.571	0.943	90.571	0.943
☑		Neural Net 1	< 1	1,160.0	15	2.883	20	89.921	0.930	89.921	0.930
☑		Tree-AS 1	< 1	778.036	16	2.649	10	86.716	0.901	86.716	0.901
☑		C5 1	< 1	1320.455	12	2.498	5	91.779	0.842	91.779	0.842

Table.6.4. Model performance summary from Scenario 2

For the second scenario of our model implementation wherein we chose 20 features based on the feature significance scores, we also obtain the ROC curve to determine the sensitivity and specificity of the model. We see that the TP Rate (or the Sensitivity) of the model peaks around 0.98 and then flattens at 1, which is an indication of good model performance. While the FP Rate or the difference between 1 and Specificity for both the train and test dataset indicates a good value in the resultant curves shown below.

We also have the gain % for both the train and test data which allows us to understand the model performance again, we see that the gain % peaks towards 95 which validates the model performance both on training and testing set.

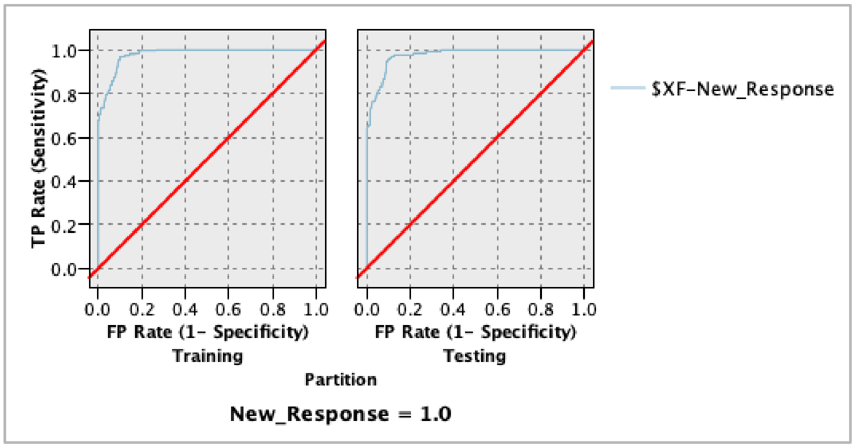


Fig.6.11. ROC Curve for best fit model from Scenario 2

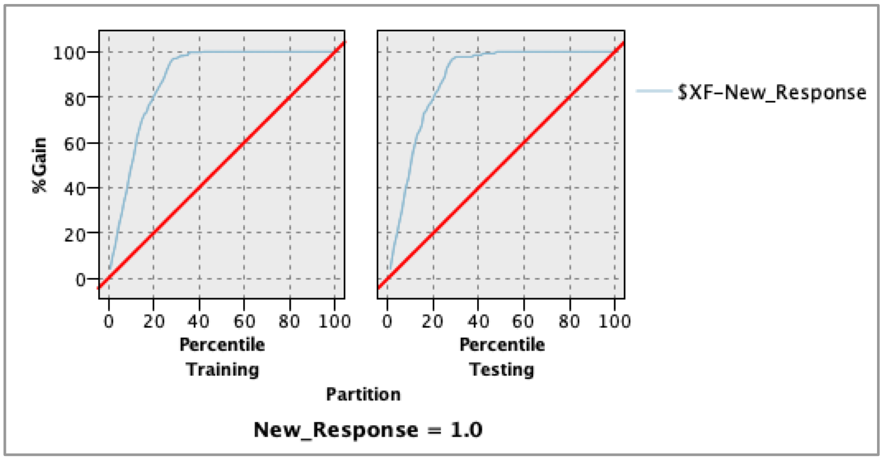
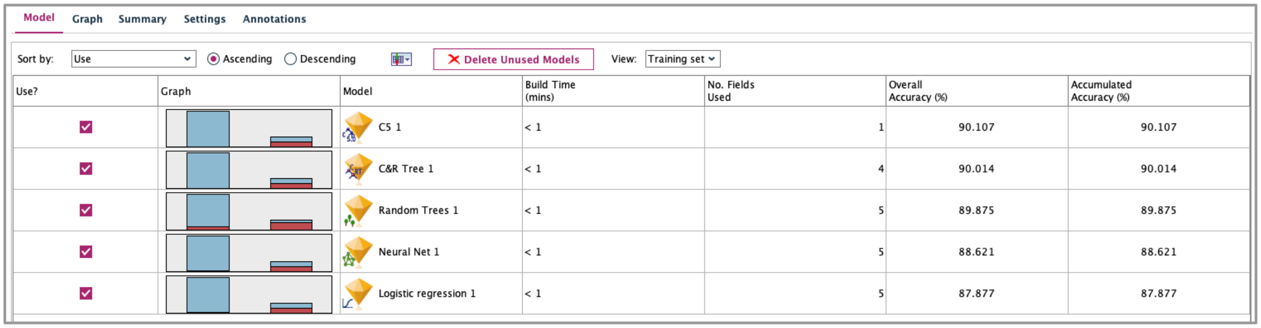


Fig.6.11. Gains curve for best fit model from scenario 2

scenario 3

In the third scenario, we only pick the top 5 features from the factor significance step to determine the comparative model study and pick the top model for our prediction based on the accuracy. It is observed in this scenario that C5 performs the best followed by C&R Tree and Random Forest. The accuracies are 90.1%, 90% and 89.8% respectively for the above-mentioned models.



Sort by: Use ● Ascending ○ Descending View: Training set

Use?	Graph	Model	Build Time (mins)	No. Fields Used	Overall Accuracy (%)	Accumulated Accuracy (%)
☑		C5 1	< 1	1	90.107	90.107
☑		C&R Tree 1	< 1	4	90.014	90.014
☑		Random Trees 1	< 1	5	89.875	89.875
☑		Neural Net 1	< 1	5	88.621	88.621
☑		Logistic regression 1	< 1	5	87.877	87.877

Table.6.5. Model performance summary from scenario 3

In the third scenario, wherein we chose the top 5 features to use in our model we also obtain the ROC curve for the same and observe that the Sensitivity peaks around 0.90 with the FP Rate at 0.2. Relatively the curve is not great compared to the other two scenarios but still fits our threshold of good performance in our comparative study. Based on the above performances and derivations we will discuss our conclusion of the best fit model in the coming section to conclude.

In this scenario of the modelling, we also have the gain % of the model which peaks around 85 and is comparatively lower than the above cases but still a good score to proceed ahead with the model outputs for our considerations.

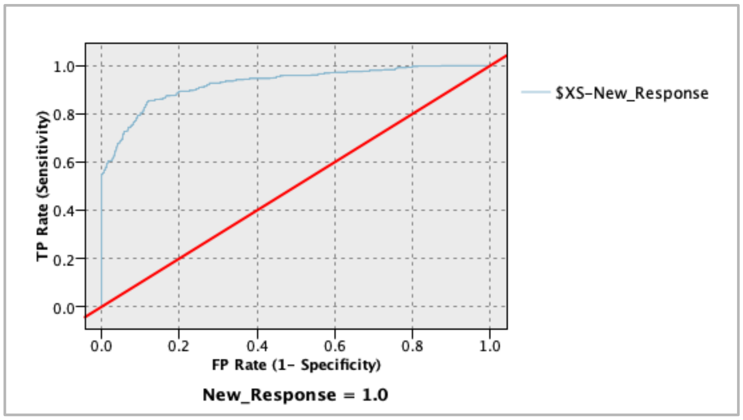


Fig.6.12. ROC Curve for best fit model from scenario 3

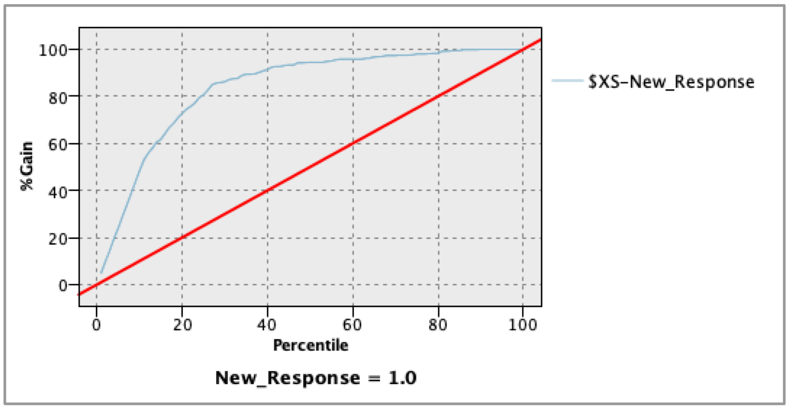


Fig.6.12. Gain Curve for best fit model from scenario 3

From the above analyses and comparative study for modelling to determine the best fit model we considered different scenarios of selecting features with combinations of different scenarios. This helped us identify the best fit model for prediction based on several factors. From the study Random Trees has been identified to be the best model in two scenarios of 3 (66% pick rate from three cases). In the first and second scenario, the Random Trees accuracy was 98.2% and 98% respectively. The second-best model that can be considered from the above study is C5 which was second in the first case, and first in the third case with accuracy of 91.7% and 90% respectively for both the scenarios. Hence, we would like to recommend Random Trees to be the best fit model for the above problem statement and accordingly recommend the same to our readers based on the accuracy and performance evaluation criteria.

6.3 Evaluation

In this section, we would like to perform an evaluation study of the different scenarios that we implemented in our modelling phase and account for the different performance criteria and model considered from each one.

Scenario considered	Algorithm	Number of features	Accuracy	AUC	Gain %
Scenario 1	Random Trees	37	98.2%	0.991	98
Scenario 2	Random Trees	20	98%	0.987	95
Scenario 3	C5	1	90.1%	0.982	85

Table.6.6. Evaluation Table

We can observe above in the table that we implemented different model sets using SPSS Modeler which provided us certain evaluation metrics to identify the best fit model of all. It was identified that Random Trees was the best performing model for Scenario 1 and Scenario 2 with different sets of features used. The accuracy was consistently at 98% in terms of its predictive capability along with the AUC being above 0.98. Also, the random trees are more explainable from a decision point of view. The significant features evaluated were determined to be the Total Amount, Amount of Meat and Wine product purchases done by the customers. Because of the high interpretability of the Random Trees model (which is a collection of Decision Trees) and the other parameters aligning with the performance KPIs, we were able to move ahead with the same model for better results in the long run.

Chapter 6 - Conclusion

6.1 Conclusion

Through the entire course of the report, we have determined a problem statement and then determined the appropriate dataset to derive the optimal solution for the same. We used one of the most popular data science problem solution approaches called the CRISP-DM which helps us define and understand a problem statement and tackle the problem in a streamline manner. In this report, we have also used one of the most complex and sophisticated tools called SPSS Statistics and Modeler to perform data manipulation and preparation along with Statistical analysis and modelling. For the Statistical part we perform factor analysis and dimension reduction, RFM Analysis along with factor significance analyses and for the modelling part we implemented various Machine Learning models to perform a comparative study to obtain the best model of all based on the predictive capability of these models. The evaluation metric was picked to be Accuracy of the model along with AIC, Accumulated AIC and Accuracy etc.

Throughout the course of the study we identified several insights and based on these we would like to recommend the following.

- We see that the RFM analysis was able to identify different cohorts of customers based on their behaviour with the superstore. Based on that the Big Spenders, Loyal Customers and Best Customers were identified as the best cohort customers that we have in the data.
- For the above-mentioned customer segments, we should ensure that they are always willing to purchase with the store and then roll out lucrative offers and customer loyalty programs for the same. This would ensure that they are valued, and they will continue to do business with us, and they are retained in the long run
- For the customers belonging to the detrimental cohort like lost Cheap Customers, will not need any further marketing since spending on marketing for them will not justify the cost
- For the customers belonging to the detrimental cohorts like lost customers and almost lost customers, should be discussed with the management to understand why these type of high spending customers are moving away from the store and try to lure them back using special discounts and also using loyalty special discount cards or subscriptions along with special online discounts for the customers who tend to purchase through the website.
- From the modelling it was identified that Random Trees performed the best in terms of correctly identifying the correct responses to marketing campaigns. The model performed well in two of

the three scenarios that we have created while having 98% accuracy in both, and also maintain a high ROC score greater than 95% along with the AUC and a high gain score. Hence it is imperative that this model be used for predicting if a customer responds to a marketing campaign or not

- To summarise from our entire exercise, we were able to identify the different customer segments and their responsiveness towards marketing campaigns. This should be used to leverage and target the right audience with the appropriate marketing techniques so that we have higher conversion rates with our business.

6.2 Future Work

There was a lot of learning within the due course of the entire exercise. We performed various data preparation, cleaning, and analysis to obtain the best data form for our task. We then implemented various Machine Learning techniques and were able to perform a comparative study of these models to pick the best one of all. For future work, there is scope for many other tasks like extending the dataset to contain more features describing the demographic factors and other marketing related KPIs that define the marketing dataset better. With an exhaustive view of the dataset with more features, the model has even more factors to explain the results better and the accuracy can be further improved.

Bibliography

[1] Customer segmentation definition - what is Customer Segmentation. Shopify. (n.d.). from https://www.shopify.com/encyclopedia/customer-segmentation#:~:text=Customer%20segmentation%20is%20the%20process

[2] Kim, S.-Y., Jung, T.-S., Suh, E.-H., & Hwang, H.-S. (2006). Customer segmentation and strategy development based on Customer Lifetime Value: A case study. *Expert Systems with Applications*, *31*(1), 101–107. https://doi.org/10.1016/j.eswa.2005.09.004

[3] Wu, J., & Lin, Z. (2005). Research on customer segmentation models by clustering. *Proceedings of the 7th International Conference on Electronic Commerce - ICEC '05*. https://doi.org/10.1145/1089551.1089610

[4] Akar, E. (2021). Customers' online purchase intentions and customer segmentation during the period of COVID-19 pandemic. *Journal of Internet Commerce*, *20*(3), 371–401. https://doi.org/10.1080/15332861.2021.1927435

[5] Sari, J. N., Nugroho, L. E., Ferdiana, R., & Santosa, P. I. (2016). Review on Customer Segmentation Technique on Ecommerce. *Advanced Science Letters*, *22*(10), 3018–3022. https://doi.org/10.1166/asl.2016.7985

[6] Ma, H. (2015). A study on customer segmentation for e-commerce using the Generalised Association rules and Decision Tree. *American Journal of Industrial and Business Management*, *05*(12), 813–818. https://doi.org/10.4236/ajibm.2015.512078

[7] Silversteinc, B.M. (1997) Beyond Market Basket: Generalizing Association Rules to Correlations. Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data (SIGMOD97), Tucson, Date, 265-276

[8] Cooil, B., Aksoy, L., & Keiningham, T. L. (2008). Approaches to customer segmentation. *Journal of Relationship Marketing*, *6*(3-4), 9–39. https://doi.org/10.1300/j366v06n03_02

[9] Hamka, F., Bouwman, H., de Reuver, M., & Kroesen, M. (2014). Mobile customer segmentation based on smartphone measurement. *Telematics and Informatics*, *31*(2), 220–227. https://doi.org/10.1016/j.tele.2013.08.006

[10] Hosseini, M., & Shabani, M. (2015). New approach to customer segmentation based on changes in customer value. *Journal of Marketing Analytics*, *3*(3), 110–121. https://doi.org/10.1057/jma.2015.10

[11] Business Model Canvas. (2019). Retrieved from Strategyzer: https://www.strategyzer.com/business-model-canvas/customer-segments

[12] customer segmentation models there's a better approach for 2022. (2022, April 25). Retrieved from formation: https://formation.ai/blog/customer-segmentation-models-theres-a-better-approach-for-2022/

[13] Oliver. (2022). Market segmentation: Definition, types and best practices. Retrieved from qualtrics: https://www.qualtrics.com/uk/experience-management/brand/market-segmentation/

[14] Optimove. (2022, August). 1. Retrieved from Optimove: https://www.optimove.com/resources/learning-center/customer-segmentation

[15] P., J. (2021). geographic segmentation. Retrieved from qualtrics: https://www.qualtrics.com/uk/experience-management/brand/geographic-segmentation/?rid=ip&prevsite=en&newsite=uk&geo=AE&geomatch=uk

[16] Raghavan, R. (n.d.). customer segmentation. Retrieved from acowebs: https://acowebs.com/ecommerce-customer-segmentation/#1_The_Cart_Abandoners

[17] Serrano S. (2022, July). RFM Analysis with predictive segmentation examples. https://www.barilliance.com/rfm-analysis/#:~:text=RFM%20analysis%20is%20a%20data,much%20they've%20spent%20overall

[18] Khade, A. A. (2016). Retrieved from Researchgate: https://www.researchgate.net/publication/300080057_Performing_Customer_Behavior_Analysis_using_Big_Data_Analytics

[19] *The Business Benefits of Customer Analysis.* (n.d.). Retrieved from Spencersaving: https://www.spencersavings.com/the-business-benefits-of-customer-analysis/

[20] Joao Correia, How to segment your customers and increase sales with RFM Analysis, 12 July 2016, Insights, https://joaocorreia.io/blog/rfm-analysis-increase-sales-by-segmenting-your-customers.html

[21] Michael P. LaValley, Logistic Regression, Aha Journals, 6 May 2008, https://www.ahajournals.org/doi/full/10.1161/CIRCULATIONAHA.106.682658

[22] Wang, SC. (2003). Artificial Neural Network. In: Interdisciplinary Computing in Java Programming. The Springer International Series in Engineering and Computer Science, vol 743. Springer, Boston, MA. https://doi.org/10.1007/978-1-4615-0377-4_5

[23] Sarang Narkhede (2018). Understanding AUC - ROC Curve, 26 June 2018. https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

[24] Correia, J. (2016, July 12). *How to segment your customers and increase sales with RFM*
PATEL, A. (2021). *Customer Personality Analysis*. Retrieved from Kaggle: https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis