# Predicting the Next Purchase Date for an Individual Customer using Machine Learning

by

Marli Droomer




Thesis presented in partial fulfilment of the requirements for the degree of Master of Engineering (Industrial Engineering) in the Faculty of Engineering at Stellenbosch University

Supervisor: Prof JF Bekker

December 2020

# Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: 2020/08/27

# Abstract

We live in a world that is rapidly changing when it comes to technology. Gathering a customer's information becomes easier as companies have loyalty programs that track the customer's purchasing behaviour. We live in an era where search engines suggest your next word, online shopping is no longer scary, and people order a ride by means of an application. The fact is that technology is evolving, and gathering information from customers is becoming easier. Given this change, the questions, however, are: How do companies use this information to gain a competitive advantage? Do they use this information to benefit the customer? How can a company use customer information to give each individual a unique experience?

A research study was conducted to determine if an individual customer's next purchase date for specific products can be predicted by means of machine learning. The focus was on fast-moving consumer goods in retail. This next purchase date can then be used to individualise marketing to customers, which benefits the company and the customer. In this study, the customer's purchase history is used to train machine learning models. These models are then used to predict the next purchase date for a customer-product pair. The different machine learning models that are used are recurrent neural networks, linear regression, extreme gradient boosting and an artificial neural network. Combination approaches are also investigated, and the models are compared by the absolute error, in days, that the model predicts from the target variable.

The artificial neural network model performed the best, predicting 31.8% of the dataset with an absolute error of less than one day, and 55% of the dataset with an absolute error of less than three days. The application of the artificial neural network as the Next Purchase Date Predictor is also demonstrated and shows how individualised marketing can be done using the Next Purchase Date Predictor.

The encouraging results of the Next Purchase Date Predictor showed that machine learning could be used to predict the next purchase date for an individual customer.

# Opsomming

Vandag se wêreld is besig om baie vinnig te verander as dit kom by tegnologie. Dit raak al hoe makliker om kliënte se koopgewoontes vas te vang met lojaliteitskaarte wat beskikbaar is by meeste winkels. Hierdie inligting maak dit makliker om kliënte se koopgewoontes te analiseer. Ons bly ook in 'n wêreld waar al hoe meer mense aanlyn aankope maak, waar ons toepassings gebruik om kos af te lewer of selfs 'n toepassing gebruik om 'n rit lughawe toe te bespreek. Tegnologie ontwikkel en om inligting van kliënte te versamel raak makliker. Gegewe hierdie veranderinge, laat dit 'n paar vrae: Hoe word hierdie beskikbare inligting gebruik deur besighede om bo hulle mededingers uit te troon? Gebruik besighede hierdie inligting tot die voordeel van hulle kliënte? Hoe kan 'n besigheid hierdie inligting gebruik om vir elke kliënt 'n meer individuele koopervaring te gee?

Om hierdie vrae te ondersoek is 'n navorsingstudie gedoen wat ondersoek of masjienleer gebruik kan word om te voorspel wanneer 'n kliënt 'n spesifieke produk gaan aankoop. Die fokus was op vinnig-vloeiende verbruikersitems in die kleinhandel. As hierdie voorspelling gemaak kan word kan dit gebruik word om vir die kliënt spesifieke advertensies te skep op die spesifieke tyd wat die kliënt die produk nodig het. Historiese kooptransaksies van kliënte word in hierdie studie gebruik om masjienleermodelle te skep. Hierdie modelle word dan gebruik om die voorspelling te maak vir 'n kliënt-produk paar. Die verskillende masjienleermodelle wat geskep is sluit in: Herhalende Neurale Netwerke, lineêre regressie, uiterste gradientverhoging en kunsmatige neurale netwerke. Om die modelle met mekaar te vergelyk was die absolute fout (in dae) tussen die voorspelde waarde en die regte waarde, van al die modelle met mekaar vergelyk. Kombinasies van verskillende modelle was ook getoets om te kyk of dit die voorspelling kan verbeter.

Die kunsmatige neurale netwerk model het die beste gevaar om die voorspelling te maak en kan 31.8% van die datastel met 'n absolute fout van minder as een dag voorspel. Verder kan dit ook 55% van die datastel met 'n absolute fout van minder as drie dae voorspel. Die kunsmatige neurale netwerk was gekies om die voorspeller te wees en 'n toepassing van die model word gebruik om te demonstreer hoe individuale advertensies vir kliënte gegenereer kan word.

# Acknowledgements

I would like to acknowledge the following people for their contribution towards the completion of this work:

- Prof. Bekker for all his help, guidance and insights throughout this project.
- My parents, Len and Joni Droomer, for all their love and support.
- Johan, for all his help and patience during the past years.

# Contents

**Nomenclature** xiii

**1 Research Proposal** 1
- 1.1 Research Background . . . 1
- 1.2 Research Scope and Statement . . . 3
- 1.3 Research Objectives . . . 4
- 1.4 Deliverable Envisaged . . . 4
- 1.5 Research Methodology . . . 5
- 1.6 Structure of Thesis . . . 7
- 1.7 Conclusion: Introduction . . . 8

**2 Customer Behaviour and Marketing Strategies** 9
- 2.1 Customer Behaviour . . . 9
  - 2.1.1 Defining Customers . . . 9
  - 2.1.2 Customer Relationship Management . . . 9
  - 2.1.3 Implementing Customer Relationship Management . . . 12
- 2.2 Marketing . . . 14
  - 2.2.1 Marketing in General . . . 14
  - 2.2.2 Customer Attraction . . . 18
  - 2.2.3 Customer Retention . . . 20
  - 2.2.4 Customer Development . . . 22
    - 2.2.4.1 Customer Lifetime Value . . . 22
    - 2.2.4.2 Market Basket Analysis . . . 24
    - 2.2.4.3 Sequential Pattern Analysis . . . 25
    - 2.2.4.4 Up-selling and Cross-selling . . . 26
- 2.3 Customer Profiling and Customer Segmentation . . . 28
- 2.4 Conclusion: Chapter 2 . . . 31

**3 Data Analytics, Machine Learning and Future Event Prediction** 32
- 3.1 Data Analytics . . . 32
  - 3.1.1 Data Analytics Processes . . . 34
    - 3.1.1.1 KDD Process . . . 34
    - 3.1.1.2 SEMMA Process . . . 35
    - 3.1.1.3 CRISP-DM Process . . . 36

3.1.2 Comparison of Data Analytics Processes . . . 37
3.2 Machine Learning . . . 38
3.2.1 History of Machine Learning . . . 39
3.2.2 Data Preprocessing and Transformation . . . 40
3.2.2.1 Dimensionality Reduction Techniques: PCA . . . 41
3.2.2.2 Dimensionality Reduction Techniques: LDA . . . 41
3.2.2.3 Comparison of LDA and PCA . . . 42
3.2.3 Creating a Machine Learning model . . . 43
3.3 Predictive Analytics . . . 51
3.3.1 Machine Learning and Predictive Analytics . . . 54
3.4 Using Machine Learning to predict future events . . . 55
3.4.1 Linear Regression . . . 55
3.4.2 Artificial Neural Networks . . . 57
3.4.2.1 How Artificial Neural Networks work . . . 58
3.4.2.2 Activation Functions . . . 60
3.4.2.3 Adjusting the weights and biases of the ANN . . . 60
3.4.2.4 Recurrent Neural Networks . . . 65
3.4.3 Extreme Gradient Boosting . . . 68
3.5 Conclusion: Chapter 3 . . . 70

**4 NPD requirements specification, dataset selection and data understanding 71**
4.1 Requirements for the NPD Predictor . . . 71
4.2 Requirements of a dataset needed to develop the NPD Predictor . . . 71
4.3 Comparing online datasets and dataset selection . . . 73
4.4 Dataset Selection . . . 74
4.5 Data Understanding . . . 74
4.5.1 Relational structure of the dataset . . . 75
4.5.2 Insights from the data tables . . . 79
4.5.2.1 Products most often purchased . . . 79
4.5.2.2 Customers and their orders . . . 81
4.5.2.3 Times that customers make orders . . . 82
4.5.3 Findings on the reorder rate . . . 84
4.6 Conclusion: Chapter 4 . . . 85

**5 Data Preparation** **86**
5.1 Sequence-based Features . . . 86
5.1.1 Feature 1: days between orders per product . . . 86
5.1.2 Feature 2: days since prior order per product . . . 88
5.2 Non-sequence-based Features . . . 89
5.3 Conclusion: Chapter 5 . . . 92

**6 Next Purchase Date Predictor Modelling** **93**
6.1 Recurrent Neural Network . . . 93
6.2 Linear Regression . . . 95
6.3 Extreme Gradient Boosting . . . 95
6.4 Neural Network . . . 98
6.5 Conclusion: Chapter 6 . . . 99

**7 Results of the NPD Predictor (Evaluation)** **100**
7.1 Comparing the models . . . 100
7.2 Combination Methods . . . 104
7.3 Conclusion: Chapter 7 . . . 105

**8 Application of the NPD Predictor** **106**
8.1 Customer Segmentation . . . 107
8.1.1 Generating RFM Features . . . 107
8.1.2 Clustering the customers based on their RFM scores . . . 109
8.2 Application of the NPD Predictor . . . 112
8.3 Generating Individualised Advertisements . . . 115
8.4 Conclusion: Chapter 8 . . . 119

**9 Summary, Conclusion and Recommendations** **120**
9.1 Summary of the Project . . . 120
9.2 Key Findings . . . 121
9.3 Recommendations and Future work . . . 121
9.4 Personal Reflection . . . 122
9.5 Concluding remarks . . . 122

**References** **124**

**A Hyperparameter tuning for XGBoost and Artificial Neural Network** **137**

**B Results of the NPD Predictor for Cluster 3** **141**

# List of Figures

1.1 Illustrating how retailers can use the NPD Predictor . . . 2
1.2 Research Methodology . . . 6

2.1 Concept of Customer Relationship Management . . . 11
2.2 Customer Relationship Management Model . . . 12
2.3 Three dimensions of CRM . . . 13
2.4 The Marketing Process . . . 15
2.5 Four Ps of the marketing mix . . . 16
2.6 Communication with customers . . . 17
2.7 Mass Marketing vs Direct Marketing . . . 17
2.8 Direct Marketing Process . . . 19
2.9 Customer Lifetime Value . . . 23
2.10 Up-selling vs Cross-selling . . . 27
2.11 The difference between Customer Profiling and Customer Segmentation . . . 29

3.1 The value of Machine Learning through different types of analysis . . . 33
3.2 An overview of the steps of the KDD process . . . 34
3.3 Overview of the SEMMA process . . . 35
3.4 CRISP-DM Process for data mining. . . . 37
3.5 PCA transformation in two dimensions . . . 42
3.6 Linear Discriminant Analysis with two classes . . . 42
3.7 PCA versus LDA . . . 43
3.8 Machine learning algorithms . . . 46
3.9 Machine Learning and Predictive Analytics . . . 55
3.10 Linear Regression . . . 56
3.11 Neural Network architecture . . . 57
3.12 Neural Network Regression model . . . 58
3.13 Neural Network eight matrix . . . 59
3.14 RNN Architecture . . . 66
3.15 Example of Boosting . . . 69

4.1 Relational data structure of the online FMCG dataset . . . 75
4.2 Frequency of purchases made from each department . . . 80
4.3 Number of orders per customer . . . 81
4.4 Number of products per order . . . 82

4.5 Number of orders per day of the week . . . 82
4.6 Number of orders per hour of the day . . . 83
4.7 Frequency of the days between orders . . . 83
4.8 Highest reorder rate for products . . . 85

5.1 Example of "days_between_orders_per_product" feature creation . . . 87
5.2 Example of "days_since_prior_order_per_product" feature creation . . . 88
5.3 Train and test set splitting . . . 90

6.1 RNN implementation with one feature . . . 93
6.2 RNN implementation with two features . . . 94
6.3 Example of how Linear Regression was implemented for a sequence . . . 95
6.4 Results of hyperparameter random search configurations . . . 96
6.5 Feature importance determined by XGBoost . . . 97
6.6 Results of hyperparameter random search configurations . . . 99

7.1 Absolute error for all user-product pair instances . . . 100
7.2 Absolute error for first 40 000 user-product pair instances . . . 101
7.3 Number of user-product pairs predicted per absolute error in days . . . 101
7.4 Percentage of the dataset that each technique predicted in absolute error days 102
7.5 Venn diagram of ANN, RNN and XGBoost best performers . . . 103
7.6 Venn diagram user-product pairs predicted with an error of more than 40 days 104
7.7 Number of user-product pairs predicted per absolute error in days . . . 105

8.1 Proposed analysis structure for application of the NPD Predictor . . . 106
8.2 RFM distributions for all customers in the dataset . . . 108
8.3 Recency and Monetary distributions log transformed . . . 108
8.4 Distortion score elbow for K-Means clustering . . . 110
8.5 Data Clustered based on Recency, Frequency and Monetary values . . . 110
8.6 Average RFM values per cluster . . . 111
8.7 ANN implementation . . . 114
8.8 Individual advertisements for user 50, proposed to send in 8 to 9 days . . . 118

# List of Tables

2.1 Direct Marketing Campaigns . . . 18
2.2 Advantages and Disadvantages of RFM . . . 24
2.3 Example of Market Basket Analysis . . . 24
2.4 Association Rule Measures . . . 25
2.5 Advantages of SPA . . . 26
2.6 Examples of Segmentation, Profiling and both techniques used together . . . 29

3.1 Correspondence between KDD, SEMMA and CRISP-DM . . . 38
3.2 History of Machine Learning . . . 40
3.3 Clustering Techniques . . . 44
3.4 Regression Techniques . . . 47
3.5 Classification Techniques . . . 48
3.6 Clustering Techniques . . . 50
3.7 How Predictive Analytics are used in industry . . . 52
3.8 Activation Functions . . . 61
3.9 Advantages and Disadvantages of the Optimisers . . . 63
3.10 Examples of sequential data . . . 67
3.11 Examples of different types of RNN models . . . 68

4.1 Attributes needed to develop the NPD predictor . . . 72
4.2 Attributes that would be nice to have when developing the NPD Predictor . . . 73
4.3 Comparison of different online consumer goods datasets . . . 74
4.4 `Products` table first five entries . . . 76
4.5 All departments in lookup table `departments` . . . 76
4.6 Sample of the `aisles` table . . . 77
4.7 Sample of the `orders` table . . . 77
4.8 Sample of the `order_products` table . . . 78
4.9 Online FMCG data merged table . . . 78
4.10 Online FMCG data merged table continued . . . 79
4.11 Online FMCG data merged table continued . . . 79
4.12 Most purchased products from the “Produce” department . . . 80
4.13 Most purchased products from the “Dairy Eggs” department . . . 81
4.14 Percentage reorders . . . 84

5.1 Example of product order detail for a customer . . . 87

5.2 days_between_orders_per_product for all the products that Customer $X$ purchased 87
5.3 days_since_prior_order_per_product for all the products that Customer $X$ purchased 89
5.4 Dataset created with two sequence features . . . . . 89
5.5 Features (non-sequence-based) . . . . . 91
5.6 Example of the final form of the training datasets . . . . . 91
5.7 Example of the final form of the test datasets . . . . . 92

6.1 Hyperparameter search space for XGBoost . . . . . 96
6.2 Best performing parameter configuration for XGBoost . . . . . 97
6.3 Hyperparameter search space for the Neural Network model . . . . . 98
6.4 Best performing parameter configuration for the Neural Network . . . . . 99

7.1 Predictions of the ANN model . . . . . 103

8.1 Summary of RFM features after scaling . . . . . 109
8.2 Sample of the dataset for the chosen cluster . . . . . 113
8.3 Predictions for Cluster 3 . . . . . 115
8.4 Associations that can be made with the predicted products . . . . . 116
8.5 Associations with product names . . . . . 117
8.6 Up-sell oportunities for “Reduced Fat Milk”. . . . . 118

A.1 Hyperparameters for XGBoost . . . . . 137
A.2 Hyperparameter search space for the Neural Network model . . . . . 138
A.3 Results of hyperparameter random search cross validation tests for XGBoost . 139
A.4 Results of hyperparameter random search cross validation tests for Neural Network 140

B.1 NPD Predictions . . . . . 141

# Nomenclature

### Acronyms

| | |
|---|---|
| Adam | Adaptive Moment Estimation |
| AI | Artificial Intelligence |
| ANN | Artificial Neural Network |
| CLV | Customer Lifetime Value |
| CRISP-DM | Cross Industry Standard Process for Data mining |
| CRM | Customer Relationship Management |
| FMCG | Fast-Moving Consumer Goods |
| KDD | Knowledge Discovered in Database |
| LDA | Linear Discriminant Analysis |
| MBA | Market Basket Analysis |
| ML | Machine Learning |
| NMAE | Negative Mean Absolute Error |
| NPD | Next Purchase Date |
| PA | Predictive Analytics |
| PCA | Principal Component Analysis |
| PDO | Personalised Discount Offers |
| ReLU | Rectified Linear Unit |
| RFM | Recency, Frequency and Monetary |
| SEMMA | Sample, Explore, Modify, Model and Access |
| SGD | Stochastic Gradient Descent |
| SPA | Sequential Pattern Analysis |

| | |
|---|---|
| TanH | Hyperbolic Tangent |
| XGBoost | Extreme Gradient Boosting |

# Chapter 1

# Research Proposal

This chapter provides the research proposal, which commences with a brief background to understand why this study was conducted, followed by the research statement and the research objectives. The scope of the project is given, which identifies the boundaries of the research, followed by a discussion of the research methodology, and lastly, the document layout is given.

## 1.1 Research Background

We live in a world that is rapidly changing when it comes to technology. The use of paper is becoming redundant; for instance, more forms are being filled in online, newspapers can be read online, and students' assignments can be submitted electronically. Signing up for a new application is the new norm, including applications to track sleeping patterns or fitness levels or even just an application to play a game. Gathering a customer's information becomes easier as companies have loyalty programs that track their purchasing behaviour. We live in an era where search engines suggest your next word, online shopping is no longer scary, and people order a ride by means of an application. The fact is that technology is evolving and gathering information from customers is becoming easier. Given this change, the questions, however, are: How do companies use this information to gain a competitive advantage? Do they use this information to benefit the customer? How can a company use customer information to give each individual a unique experience?

Marketing has shifted from being product-oriented to being customer-oriented. In this information-rich era, customer behaviour can help marketing managers to choose the most effective marketing strategy for their customers (Hosseini & Shabani, 2015). If a company knows exactly what a customer wants to purchase at a specific time, the company can market according to the customer's needs and potentially gain a competitive advantage. Predicting a customer's purchasing behaviour opens doors for companies for marketing to a specific individual at the right time. This will have a different impact on a customer as compared with the traditional pamphlet in the newspaper.

The purpose of this research is to conduct a study to develop a predictor that predicts when an individual will purchase a specific product again, from now on referred to as the Next Purchase Date (NPD) *Predictor*. This will be done by analysing the past purchasing behaviour of individual customers.

Figure 1.1 illustrates how a retailer can use the predictor to market to the customer. The customer buys items from a retail store. The retail store captures the customer's purchase

history and provides this sales data to the NPD *Predictor*. The NPD *Predictor* uses machine learning techniques on the given sales data to determine when the customer will purchase a specific product again. This information is then provided to the marketing department. The marketing team can then use this information to advertise to a particular individual at the appropriate time.

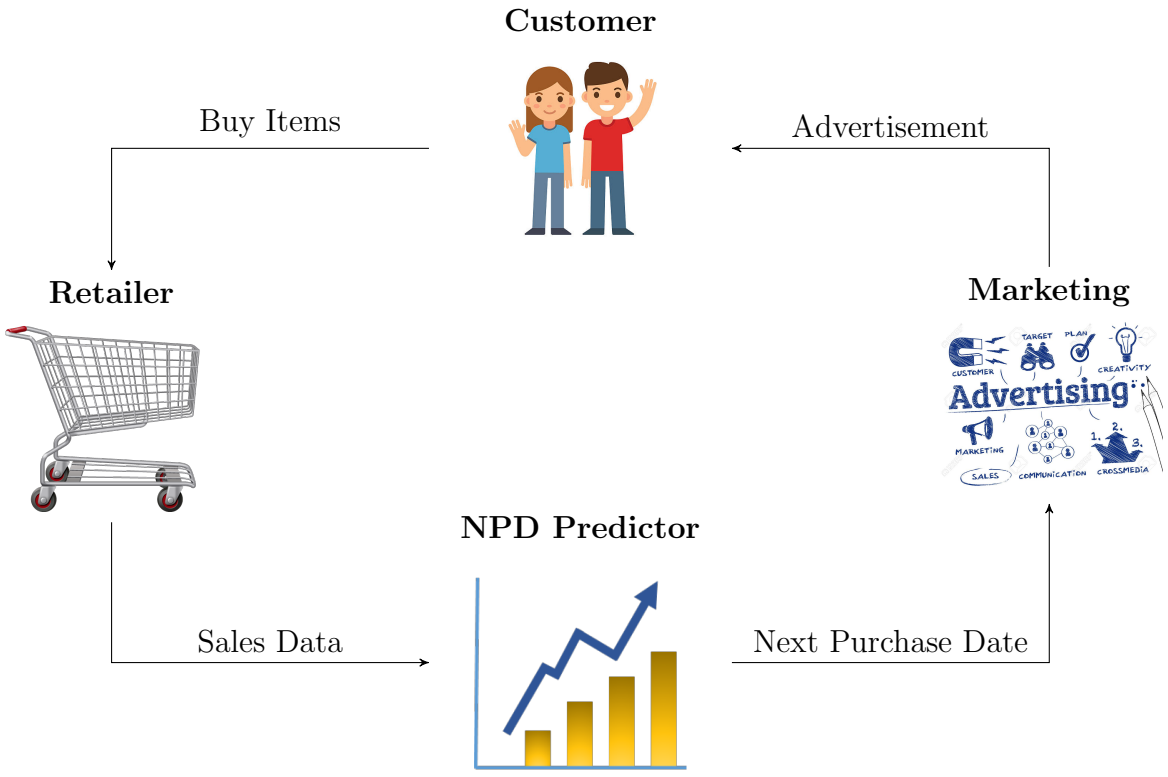



Figure 1.1: Illustrating how retailers can use the NPD Predictor

Several studies have been done to predict what customers will buy, notably in the fast-moving consumer goods (FMCG) sector. These studies include market basket analysis, predicting customer shopping lists and predicting a next purchase date to cross- and up-sell items to customers. The main idea of market basket analysis is to know which products customers buy at the same time, to improve marketing. Say, for example, a customer often buys Product A with Product B, then a discount can be offered on the combination or on one of the products, provided that the other product is bought. It is used to increase sales and maintain inventory by focusing on the point of sale transaction data (Berry & Linoff, 2004). It is also used by retailers to determine the store layout and to ensure that products that are frequently purchased together are near each other in the store (Raorane et al., 2012). A different study related to customer purchasing behaviour was done by Cumby et al. (2005), predicting a shopping list for an individual customer. This list was displayed to the customer when entering

the store to remind them to buy certain products while in the store. The business case for this study was that if a realistic shopping list could be predicted for a customer, they would be reminded of items that would otherwise be forgotten. This means that the suggestions are translated into recovered revenue for the store, which might otherwise be transferred to a competitor, or even foregone as the customer will only buy the item when they visit the store again (Cumby et al., 2004). Another study includes the development of a personalised discount offer (PDO) system done by Els (2019). The system proposes a personalised discount offer to a customer when they are visiting the store. These offers are only valid for a specific customer at a specific time. In this system, the customer is subject to cross- and up-selling. Say, for instance, a customer enters the store and wants to buy shampoo: cross-selling an item means that the system suggests to the customer that they also buy conditioner, to try to sell items that are frequently bought together. Up-selling, on the other hand, would be if the same customer wanting shampoo, gets an offer for a different, higher-priced shampoo than the one that the customer originally wanted to buy. This creates opportunities for alternative revenue streams. Els (2019) also attempted to use analytical techniques to predict when a customer would buy certain products in the fast-moving consumer goods domain, but these methods need improvement. The present work will seek improved methods to predict when a customer will buy a certain product.

This can be used by retailers to gain competitive advantage. Marketing managers can determine how they want to market to an individual, given that they know when that individual is likely to need a specific product. This shifts the marketing strategy from being product-oriented to being customer-oriented.

## 1.2 Research Scope and Statement

As mentioned, it is important for companies to gain a competitive advantage. This means that companies must keep up with the rapid change in technology and use it to their advantage. This project aims to *predict an individual customer's next purchase date of a specific item*, to be able to market to that customer according to their needs at that specific time.

To formulate the research statement, the scope of the study is first outlined. This study aims to predict the next date on which an individual customer will purchase specific items. The items that will be studied therefore need to be products that are purchased periodically by customers; thus the study will focus on fast-moving consumer goods. This includes products such as food, beverages, toiletries and cleaning products. The utilisation of retailer data is thus implied. To prove the concept, an appropriate dataset must be chosen.

The predictor will use historical purchasing behaviour data of customers to predict the next purchase date, which can be used to individualise customer marketing. Thus, the research statement can be formulated as follows:

| Develop a Next Purchase Date Predictor for individuals of fast-moving consumer goods, using Machine Learning (ML) Algorithms |
|---|

## 1.3 Research Objectives

To complete the research the following objectives will be pursued:

1. Conduct a sufficient literature study related to:
   (a) Customer behaviour and Customer Relationship Management,
   (b) Marketing Strategies,
   (c) Data Analytics,
   (d) Appropriate Machine Learning algorithms, and
   (e) Machine learning and future events prediction.
2. Design and Develop a Next Purchase Date (NPD) *Predictor*, using one or more Machine Learning algorithms.
3. Evaluate and test the NPD *Predictor*.
4. Analyse and communicate the results of the NPD *Predictor*.

The literature study will start with readings on customer behaviour and marketing strategies to get acquainted with these fields, which are not studied in the industrial engineering undergraduate programme. Since the research statement requires the use of machine learning, appropriate machine learning algorithms must be identified and studied to assess their suitability for the problem.

## 1.4 Deliverable Envisaged

This study will contribute towards personalised marketing strategies, by developing an NPD Predictor, by using machine learning techniques.

## 1.5 Research Methodology

To achieve the research objectives stated in Section 1.3, a research methodology must be followed. The proposed methodology is shown in Figure 1.2. First, the research problem must be formulated. This is done based on the research background, the research scope and statement and the objectives as documented in Sections 1.1 – 1.3. The second step in the methodology is to carry out foundational research, which will be done via a literature study on various related topics. This relates to Objective 1 in Section 1.3. The purpose of the literature study is to gain an understanding of how the NPD *Predictor* should be designed and developed as well as how the NPD *Predictor* can be applied to market to a specific customer. The structure of the literature review will be as follows:

1. Customer Behaviour and Customer Relationship Management:
    - Defining a customer.
    - Explaining what Customer Relationship Management is.
    - Explaining how Customer Relationship Management is implemented.
2. Marketing Strategies:
    - General overview of marketing and marketing strategies.
    - Explaining customer attraction, retention and development.
    - Customer profiling versus customer segmentation.
3. Data Analytics:
    - Investigate data analytics processes.
4. Appropriate Machine Learning algorithms:
    - What is Machine Learning.
    - Different Machine Learning algorithms and how they are used.
    - How Machine Learning algorithms can be used to predict customer behaviour.
5. Machine Learning and future events prediction:
    - Different methods used to predict future events.
    - Methods used to predict continuous outputs.

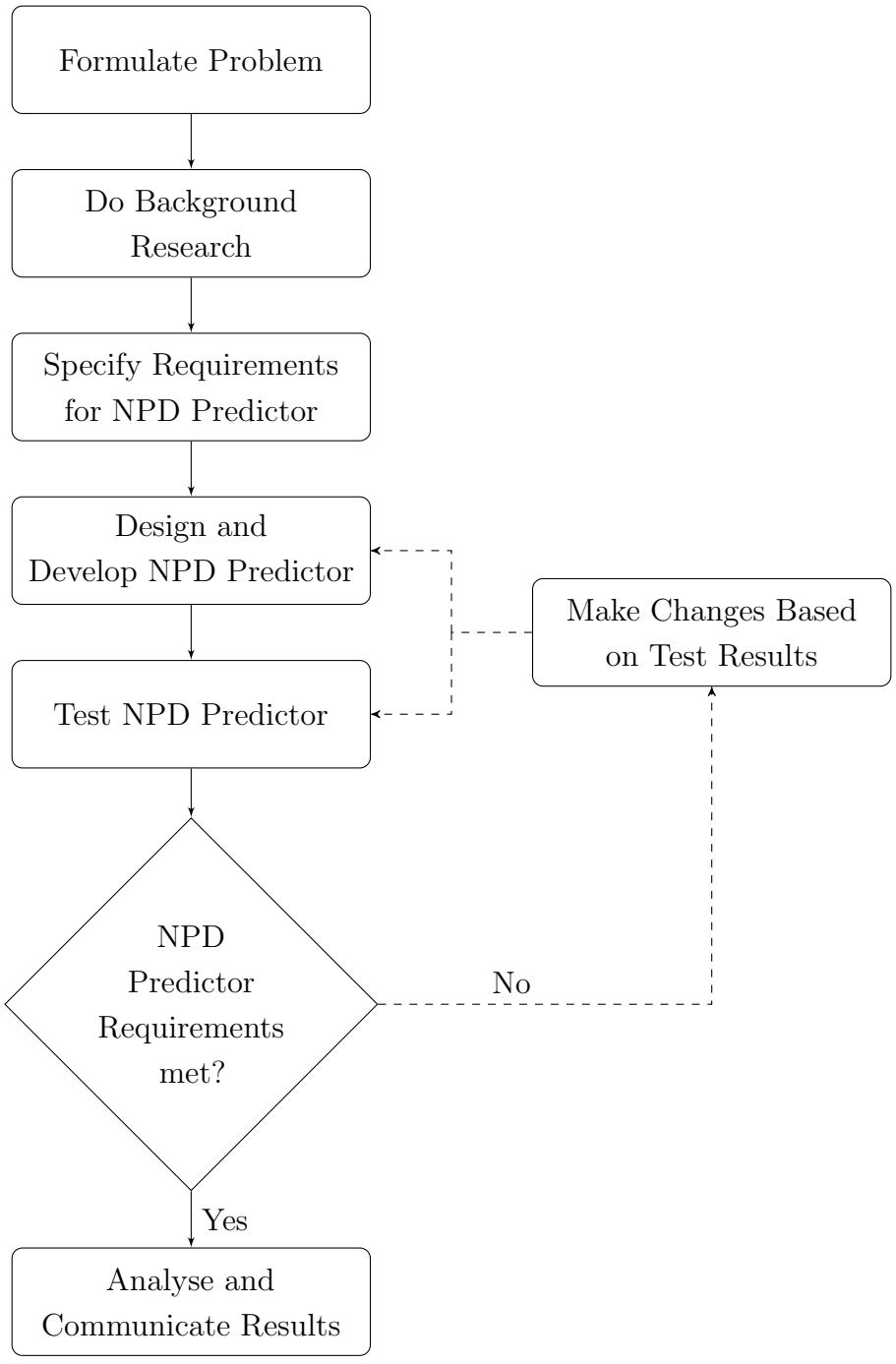



Figure 1.2: Research Methodology

The third step is to specify the requirements of the NPD *Predictor*, after which the NPD *Predictor* must be designed and developed. This corresponds to Objective 2 in Section 1.3. After the NPD *Predictor* has been developed, appropriate tests must be conducted, to see if the NPD *Predictor* meets the specified requirements. This corresponds to Objective 3 in

Section 1.3. If the requirements are only partially met, changes must be made based on the test results. If the NPD *Predictor* meets the specified requirements the results must be analysed and communicated, which corresponds to Objective 4 in Section 1.3.

## 1.6 Structure of Thesis

The thesis document takes on the following structure:

**Chapter 1: Research Proposal**
In this chapter the background for the problem is given, followed by the research problem, the research objective, the scope of the research, the deliverables envisaged as well as the methodology that will be followed.

**Chapter 2: Customer Behaviour and Marketing Strategies**
In this chapter Customer Behaviour is discussed by defining a customer and investigating the Customer Relationship Management cycle and how CRM is implemented. This is followed by a study of Marketing, which includes customer attraction, retention and development. Finally, customer profiling and customer segmentation are discussed.

**Chapter 3: Data Analytics, Machine Learning and Future Event Prediction**
This chapter includes a study of data analytics and some well-known data analytics processes. After these processes have been discussed and compared, machine learning and different machine learning algorithms are discussed. This chapter also includes machine learning techniques used to predict continuous outputs and future events.

**Chapter 4: NPD requirements specification, dataset selection and data understanding**
In this chapter requirements for the NPD predictor and the dataset that will be used to develop the NPD predictor are specified. Different online datasets are compared and a dataset is selected. The selected dataset is then explored and cleaned.

**Chapter 5: Data Preparation**
In this chapter the dataset selected in the previous chapter is modified to create the desired features to train an NPD predictor. Sequence based and non-sequence based features are created and a dataset that can be used to develop the NPD predictor is created.

**Chapter 6: Next Purchase Date Predictor Modelling**
In this chapter different machine learning models are trained using the sequence based and non-sequence based features created in the previous chapter.

**Chapter 7: Results of the NPD Predictor (Evaluation)**
In this chapter the results of the models in the previous chapter are compared and the best model is selected as the NPD Predictor.

**Chapter 8: Application of the NPD Predictor**
This chapter demonstrates how the NPD Predictor can be used in practice. The chapter starts with clustering an entire dataset based on its Recency, Frequency and Monetary scores to identify a segment on which to do NPD predictions. Then the NPD Predictor is used on the desired cluster and advertisements are generated for up-sell and cross-sell opportunities based on market basket analysis and the NPD Prediction for a user-product pair.

**Chapter 9: Summary, Conclusion and Recommendations**
This chapter summarises the project, gives key findings, identifies future work suggested by this project and gives a personal reflection.

## 1.7 Conclusion: Introduction

This chapter gave a background to the research, to explain the purpose of the study. It also demonstrated how the NPD *Predictor* coordinates with the customer, the retailer and marketing. After giving the background to the research, the research statement was given, followed by the objectives that had to be met for the study to be successful. The scope of the study was then provided, indicating the boundaries of the study, followed by the deliverable envisaged. The research methodology was provided as a guideline of how the study would be executed. Lastly, the structure of the thesis was provided, to provide clarity on how the study would be documented.

# Chapter 2

# Customer Behaviour and Marketing Strategies

Chapter 1 provided a research proposal. The methodology that will be followed is given in Section 1.5. The next step in the methodology is to conduct a literature study. The literature study consists of two parts, the first part relating to customer behaviour and marketing strategies and the second part covering literature related to machine learning and predictive analytics. This chapter relates to the first part of the literature study.

In this chapter research relating to customers and the management of relationships between customers and companies will be reviewed. As the customer is the main variable it is crucial for companies to understand their customers and their behaviour. Customer Relationship Management (CRM) is thus an important aspect of this study, as this study focuses on how to market to customers. This chapter includes a study on CRM and the various elements in the CRM process are discussed.

## 2.1 Customer Behaviour

To be able to understand customer relationship management it is important to first answer the question: What is a customer? If this is well defined within the context of this study further research can be done on CRM.

### 2.1.1 Defining Customers

The Oxford English Dictionary (1989) defines a customer as "A person who buys goods or services from a shop or business". Furthermore, Investopedia (2019) defines a customer as "an individual or business that purchases the goods or services produces by a business". Combining these two definitions it can be said that a customer is an individual or a business or any group that purchases goods or services from a business. Leading from these definitions, the customer can be defined for the purposes of this study, as an individual that purchases fast-moving consumer goods from a specific store.

### 2.1.2 Customer Relationship Management

Although the topic of CRM is slanted more towards the business environment it is important to study it to understand which engineering tools can be applied to satisfy customers and

build better relationships with them.

Parvatiyar & Sheth (2001) defined customer relationship management as, "a comprehensive strategy and process of acquiring, retaining, and partnering with selective customers to create superior value for the company and the customer. It involves the integration of marketing, sales, customer service, and the supply-chain functions of the organization to achieve greater efficiencies and effectiveness in delivering customer value."

Goldenberg (2000) defined customer relationship management by saying "We believe that CRM is not merely technology applications for marketing, sales and service, but rather, when fully and successfully implemented, a cross-functional, customer-driven, technology-integrated business process management strategy that maximizes relationships and encompasses the entire organization".

These definitions show that when CRM is successfully implemented the customer and the organisation will benefit. CRM has four components, (i) customer identification, (ii) customer attraction, (iii) customer retention and (iv) customer development (Au & Chan, 2003; Kracklauer et al., 2004; Ngai et al., 2009). These four components and how they follow consecutively one after the other can be seen in Figure 2.1. This figure also shows the elements within each component. These components form a closed circle of a customer management system (Ngai et al., 2009).

i *Customer Identification:* This phase of CRM involves targeting the group of customers that are most likely to become most profitable to the company. It analyses customers that are most likely to migrate to the competition or have migrated to the competition and how these customers can be won back (Kracklauer et al., 2004). Target customer analysis and customer segmentation are two elements that are used for customer identification. Target customer analysis, investigates customers' underlying characteristics to seek their profitable segments. Customer segmentation, on the other hand, divides the entire customer base into smaller customer groups based on similar behaviour (Alelyani et al., 2018).

ii *Customer Attraction:* This phase follows the customer identification phase. If the segments and other customer behaviour are identified the organisation can launch an effort to attract the target customers or customer segments. Direct marketing is an element of customer attraction (Ngai et al., 2009). Direct marketing, for instance, could take the form of personalised emails or coupon distribution. This will be discussed in more length in section 2.2.2.

iii *Customer Retention:* The central concern for CRM is customer retention. To retain customers it is essential for them to be satisfied; this refers to the comparison of the

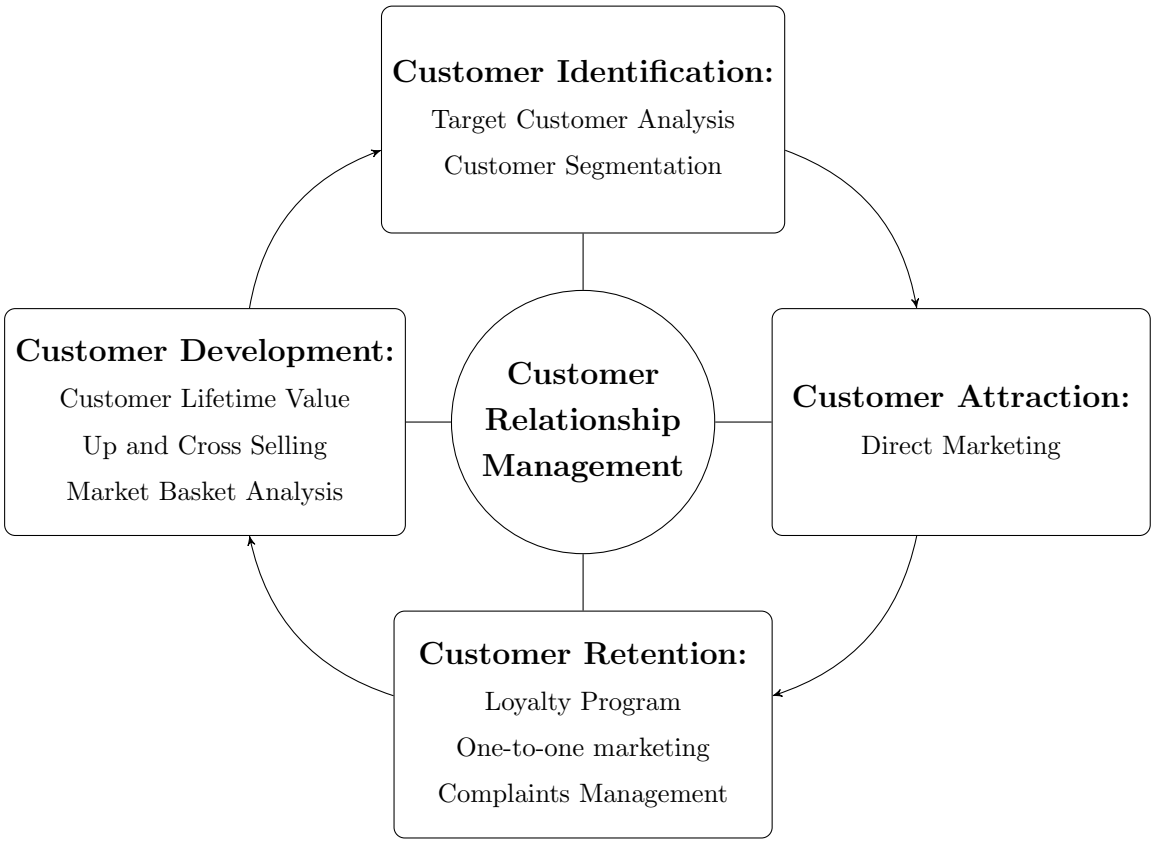



Figure 2.1: Concept of Customer Relationship Management (modified from Kracklauer et al. (2004) and Ngai et al. (2009))

customers' expectations and their perception of being satisfied (Kracklauer et al., 2004). Loyalty programmes, one-to-one marketing and complaints management are elements of customer retention.

iv *Customer Development:* This involves individual customer profitability, transactional value and the consistent expansion of customers' transaction intensity. The elements involved with customer development are customer lifetime value, up- and cross-selling and market basket analysis. Customer lifetime value refers to the total net income that a company can expect from that customer (Etzion et al., 2004). Up- and Cross-selling refer to promotional activities within a company that offer a customer a different product of the same range or a different product that complements the original product that they purchased (Prinzie & Van den Poel, 2006). These elements are discussed in more detail in Section 2.2.4. Market basket analysis aims at maximising the transaction intensity and value of a customer by analysing regularities in customers' purchase behaviour(Chen et al., 2005). THis is also discussed in more detail in Section 2.2.4.

These dimensions are used for the two main objectives of CRM (1) customer retention through customer satisfaction and (2) customer development though customer insight (Tsiptsis & Chorianopoulos, 2011). It is important to understand how CRM is implemented in companies. This will be discussed in the next subsection.

### 2.1.3 Implementing Customer Relationship Management

Winer (2001) proposed a seven step model for implementing CRM in a company. This model can be seen in Figure 2.2. The seven components that need to be considered in order to develop a CRM solution are discussed to give a general idea of the process. When following these steps all four components of CRM are touched on, but they focus more on practical implementation.

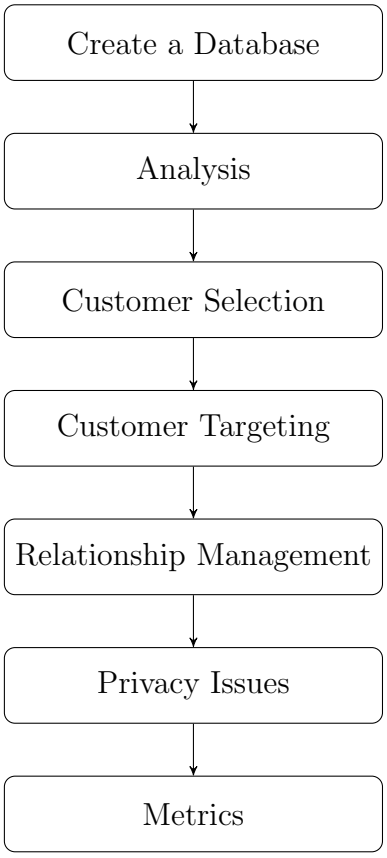



Figure 2.2: Customer Relationship Management Model (Winer, 2001).

1. *Creating a database:* This database contains information about customer activities and forms the foundation of the CRM solution.

2. *Analysing the data:* This step uses the database to segment customers or to draw customer profiles.

3. *Deciding which customers to target:* This step uses the results from the analysis stage to determine which customers to target with marketing activities.

4. *Tools to target customers*: During this step the appropriate tools and methods to target the customers identified in the previous step, are determined. These tools will be used to market the product/service to the customers.

5. *Building relationships with the targeted customers:* Determine the programmes that should be used to build and maintain relationships with customers.

6. *Privacy Issues:* Customers' information must be protected and the trade-off between improving relationships and the amount of information needed must be kept in mind.

7. *Measuring the success by set metrics:* Customer-centric metrics must be developed to determine if the CRM solution is successful. This should be able to give managers a better idea of the CRM programmes and policies and how they work.

When looking at CRM it is important to realise that it is not only about the customer or the technology, but it has a three-dimensional aspect (Chen & Popovich, 2003). The three dimensions are people, processes and technology and these three interact as seen in Figure 2.3. Implementing CRM solutions requires a company-wide, cross-functional and customer-focused approach if it is to be successfully implemented in a company (Goldenberg, 2000). It is thus important to integrate the technology, the people and the processes.

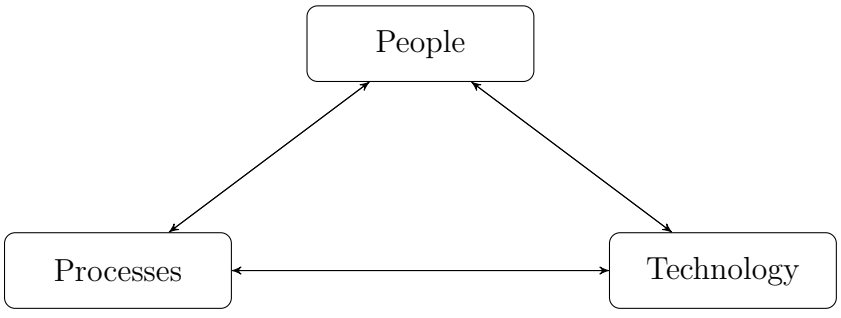



Figure 2.3: Three dimensions of CRM (Chen & Popovich, 2003)

From the concept of CRM discussed in this chapter it is clear that customer attraction, customer retention and customer development are important for this study. The elements of these CRM components will be discussed in the next subsections. As this study focuses on

marketing to individuals, the element of customer attraction, which is direct marketing, will be discussed first, to establish ways to market to customers.

## 2.2 Marketing

Information and technology have changed the way companies market to their customers. Twenty years ago it was hard to imagine that there would be 2.7 billion smartphone users in 2019 (Deyan, 2019). Having this statistic in mind, companies should adjust accordingly. More people are connected to the internet than there are people who buy newspapers. Thus, companies should acknowledge this change and implement marketing strategies according to their target market and the technology they use.

When a millennial is looking for a restaurant to have dinner with their friends, they simply search for one on the internet and make their decision based on the search results. Most companies have established websites over the last few years. A lot of companies have also started to sell products online, in addition to the stores that they own. As well as signboards in the city and advertisements in local newspapers, companies also use modern marketing techniques such as social media marketing, marketing by email and internet advertisements. Take YouTube for example; before watching the desired video YouTube first plays an advertisement, usually associated with the content of the video, or related to an individual's search results.

What is marketing? Marketing has various definitions from very broad to very specific. BusinessDictionary.com (2019b) defines marketing as the management process that moves goods and services from a concept to the customer. Kotler et al. (2017), defines marketing as the delivery of customer satisfaction at a profit. All companies have different ways of marketing to their customers. Marketing to customers has changed and become customer-focused in the last few years, with companies focusing more on serving and satisfying the needs of their customers (Kotler et al., 2017). The appropriate marketing strategy depends on the industry that a company is in as well as the business strategy that the company chooses.

### 2.2.1 Marketing in General

Marketing is a process that does not consist simply of selling and advertising goods. The process consists of analysing marketing opportunities, selecting target markets, developing the marketing mix and managing the marketing effort. This process is depicted in Figure 2.4.

This process shows the importance of building customer relationships and creating value for the customer (first four steps) to be able to capture the value from the customer in return (last step). The first step in the process is to understand the customer's needs as well as the marketplace. To do this the needs, wants and demands of the customer must be established.

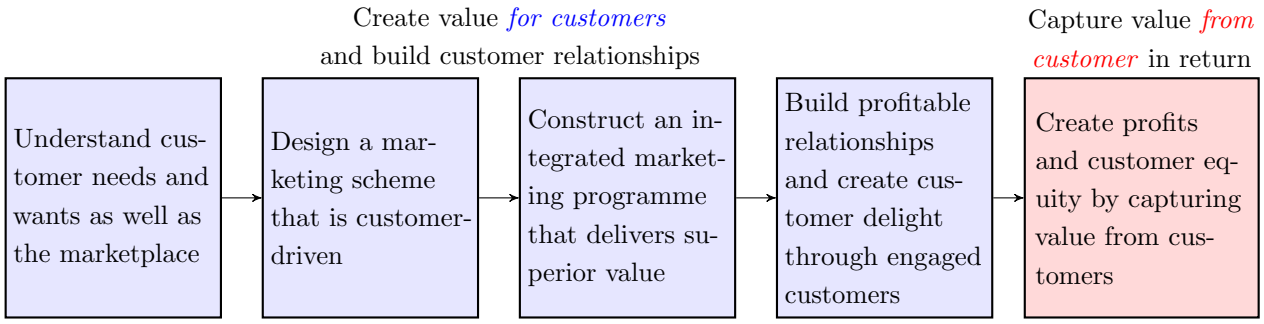



Figure 2.4: The Marketing Process (Kotler et al., 2017).

After these have been established the next step is to design a marketing scheme that is customer-driven. This can be done by deciding which customers the company will serve and how the company will best serve the chosen target market (Kotler et al., 2017). The next step is to construct a marketing programme that will deliver superior value. The marketing mix framework (set of tools used to transform marketing planning into practice (Goi, 2009)) is used to choose a mixture of marketing elements to form the marketing plan. The marketing mix which originally had 12 marketing elements, proposed by Borden (1964) to translate marketing strategies into action, were reduced by EJ McCarthy into four elements known as the 4Ps: product, price, place and promotion (Goi, 2009). Figure 2.5 shows the 4Ps and the marketing tools associated with each of them.

Some literature adds three additional Ps: people, process and physical evidence. Goi (2009) did research on the 4P framework and alternative frameworks and concluded that the 4P framework is the one most widely used in literature, but he mentioned that the founders of the framework said that the number of possible strategies of the marketing mix are infinite. Because the customer and customer relationships are so important, this study will also include people in the marketing mix. Including people in the marketing mix strives toward creating a habit of thinking in terms of people inside and outside the company who are responsible for marketing strategies, sales and general business activities.

After the marketing programme has been constructed, the next step in the marketing process is to build customer relationships. Building customer relationships is achieved through various methods of communication. These include mass marketing, segment-based marketing, direct marketing or one-to-one marketing, each of which will be explained briefly. In Figure 2.6, the number of people targeted for each strategy is represented by the size of the area. The number of people targeted decreases from mass marketing to one-to-one marketing.

- Mass marketing: Is to market a product or a service on a large scale to customers or potential customers. Usually one offer is sent to all of the customers, with the idea that

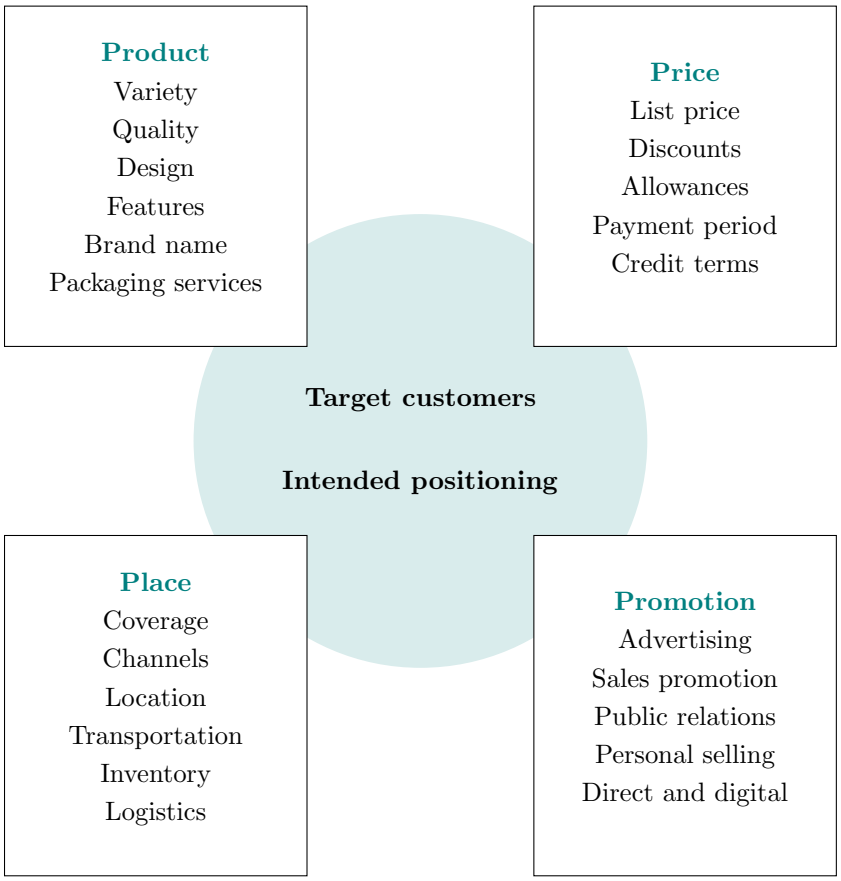



Figure 2.5: Four Ps of the marketing mix (Kotler et al., 2017).

the maximum number of customers will get the message. This is usually done by radio, television or newspaper advertisements (BusinessDictionary.com, 2019a).

- Segment-based marketing: This strategy uses existing relationships to divide the customers into segments. These segments are then used to market a product to a particular segment that shows interest in the product. These segments can be based on similar demographics, physiological features or similar behaviour (Dickson & Ginter, 1987).

- Direct marketing: This means of marketing communicates directly with the customer and requires a direct response back from the customer. This happens through telephone selling or direct mail (McFadden, 2019).

- One-to-one marketing: This marketing strategy analyses customer data to deliver individualised messages to current and prospective customers. Customer preferences are used to suggest specific products to the customer (Peppers et al., 1999).

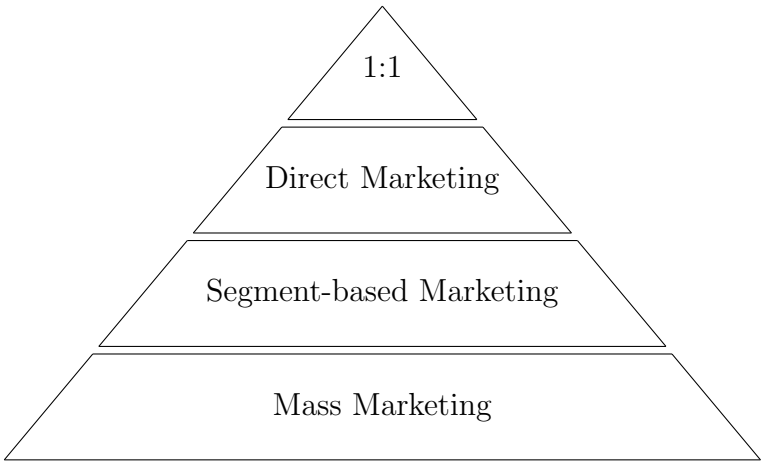



Figure 2.6: Communication with customers (Andersen et al., 1999)

As mass-marketing is very expensive and the return on investment is frequently questioned, marketing strategies tend to shift to a more individual effort. If the company knows the customer's behaviour or when a customer will churn, the company can act accordingly and only market to those customers (Bounsaythip & Rinta-Runsala, 2001).

The conceptual difference between mass marketing and direct marketing can be seen in Figure 2.7. This image visually displays the market and how it is divided into the groups just discussed.

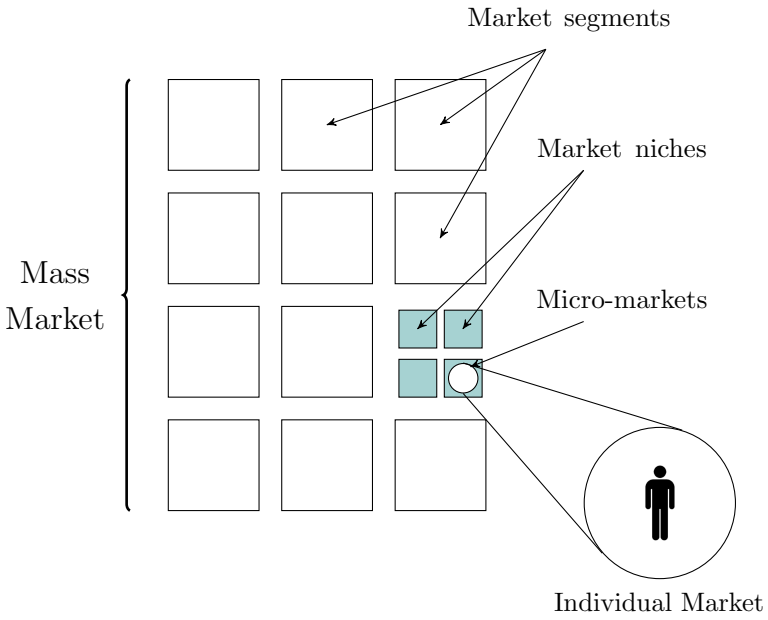



Figure 2.7: Mass Marketing vs Direct Marketing (Thomas et al., 2007)

As seen in the figure mass marketing consists of all customers; these are then segmented for segment-based marketing which is illustrated by grouping the entire market into smaller blocks. The smaller blocks can then further be grouped for direct marketing and finally individuals are identified to create an individual market.

The final step in the marketing process in Figure 2.4, involves capturing value from the customer in return. This is done by creating loyal customers so that they can be retained. The first four steps ensure that the customers are highly satisfied and thus stay loyal and buy more, which increases the lifetime value of a customer (Kotler et al., 2017).

The focus of this project is to target individual customers and therefore the next section will focus on direct marketing strategies, related to customer attraction in the concept of CRM in Figure 2.1.

### 2.2.2 Customer Attraction

Direct marketing is the main element for customer attraction and it is used for three different reasons that are summarised in Table 2.1 which shows the campaign along with the goal to be achieved through that specific campaign (Tsiptsis & Chorianopoulos, 2011). If customers that are suitable for a campaign can be identified this can lead to more customers engaging in the campaign, as customers that are more likely to act on such a campaign are targeted.

Table 2.1: Direct Marketing Campaigns (Tsiptsis & Chorianopoulos, 2011)

| **Campaign** | **Goal** |
|---|---|
| Acquisition | Draw customers that are potentially valuable to the company away from the company's competition. |
| Cross-selling and up-selling | Used to sell additional products to a customer, to sell more of the same product to a customer or to sell products that are more profitable to the customer. |
| Retention | Preventing valuable customers from terminating their current relationship with a company. |

To be able to use direct marketing a few steps must be followed. Thomas et al. (2007) propose a twelve-step plan as seen in Figure 2.8. This process has four main marketing aspects and each aspect has three questions that should be answered. The steps must also be followed from top to bottom for the process to work effectively. Each of the direct marketing steps will be discussed (Thomas et al., 2007).

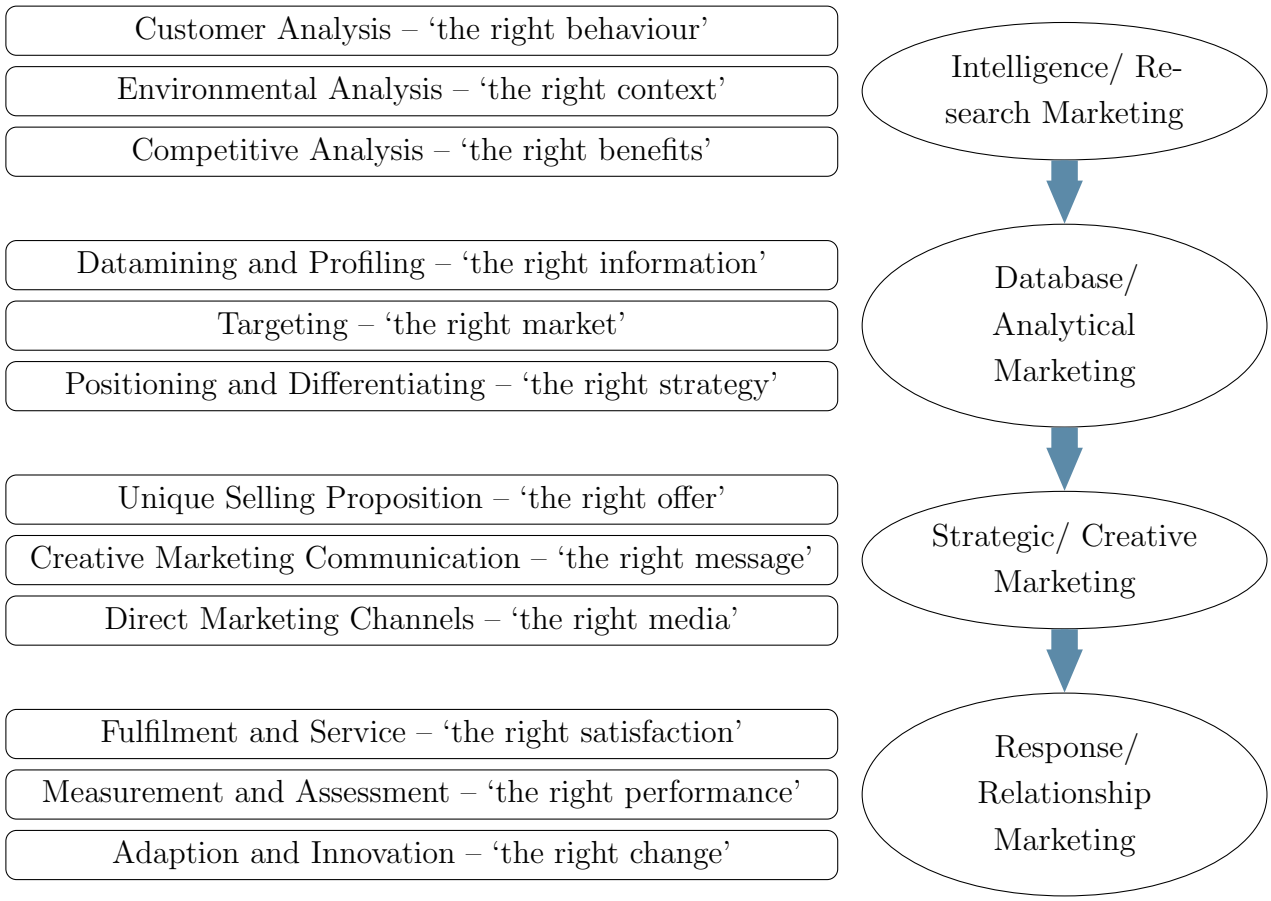



Figure 2.8: Direct Marketing Process (Thomas et al., 2007)

**Intelligence/ Research marketing:**

*Customer analysis* – Identify what customers buy and why they buy it. If this can be answered, the customer's needs, motivations and buying profiles can be mapped.

*Environmental analysis* – Companies need to anticipate their competitors' next move or evaluate new markets that can potentially emerge.

*Competitive analysis* – Evaluate what the competition is doing right and what they are doing wrong. If so, the company can develop their own unique message.

**Database/ Analytical marketing:**

Data mining and profiling – Develop a database of prospects; this is then used to extract and analyse as much information as possible from the customers so that a clear picture of the audience can be drawn.

Targeting – Use the database developed in the previous step and further refine it to get the best possible prospects.

Positioning and Differentiating – Develop an offer or central selling point, by using a three-step process: (1) identify which attributes of an offer make it unique from the competitor's, (2) describe the benefits that the customers will receive when they accept the offer and lastly, (3) make claims that include the promised benefits for taking advantage of the offer.

**Strategic/ Creative marketing:**
*Unique Selling Proposition* – Make a statement of perceived value to the customer such as, 'the product will make you healthier, happier'.
*Creative Marketing Communication* – Determine how the statement/message in the previous step will be shaped. The components of the direct mail campaign are involved in this step, such as the tone or the type-style.
*Direct Marketing Channels* – The means by which the message is delivered to the customer is decided in this step. Direct marketing campaigns such as phone calls, email or newsletters can be used. This is decided based on the best way to get to the customer's heart.

**Response/ Relationship Marketing:**
*Fulfilment and Service* – Once the campaign has worked and the customer takes up the offer, this step ensures that the customer has a good experience redeeming the coupon you offered or service that the company promised them.
*Measurement and Assessment* – Track the results so that the company knows what they did right and what they did wrong. The campaign worked if the gap between the customer and the company was closed in a cost-effective manner.
*Adaption and Innovation* – The last step is a step that revises, refines and relaunches. Reflect on the campaign and if not satisfied, the campaign can be altered to present a different message, a different communication channel or any other campaign element.

This twelve-step marketing process has a lot of similarities with the CRM model discussed in Subsection 2.1.3. These two models take different angles but, in the end, satisfying the customer and identifying customers that are most likely to be of high-value to the company is important, so that these customers can be targeted and retained. This brings us to Customer Retention which is also a component of the CRM concept seen in Figure 2.1.

### 2.2.3 Customer Retention

For direct marketing to work effectively, another strategy must align with the direct marketing campaign. This is known as relationship marketing. Relationship marketing has a customer-

centric focus, which ensures long-term relationships with customers. Relationship marketing is defined by Paley (2007) as, "the practice of building long-term satisfying relations with key parties – customers, suppliers and distributors – in order to retain long term preference and business". These types of relationships are part of the customer retention component of CRM.

The main relationship traditionally focused on was that between suppliers and consumers. One relationship that has become increasingly important and can no longer be ignored is the relationship between customer and product (Paas et al., 2005). This shift, along with the importance of customer retention has shifted marketing attention to loyalty-based programmes, one-to-one marketing, and up-selling and cross-selling opportunities.

The relationship between a customer and a company is based on three dimensions, (1) the duration of the relationship, (2) the balance of interest and (3) the intensity and direction of communication between the parties (Paas et al., 2005). As the length of the relationship is one of the dimensions, transactions that a customer makes cannot be seen as discrete events that do not contain any value. Companies roll out loyalty programmes so that they can track their customers' transactional history, so it can be analysed.

These relationships between customers and companies are nurtured by the way companies use the data obtained from their loyalty programmes to market to their customers. Companies used personalised marketing campaigns to target specific customers (Kallier, 2017; Khodakarami & Chan, 2014). Direct marketing goes deeper into the marketing segments and focuses on the micro-markets as well as on the customer as an individual (Thomas et al., 2007). When targeting the individual with a marketing strategy it is known as *one-to-one marketing* or *targeted marketing*. A key concept of one-to-one marketing is to customise an offer presented to an individual based on their needs. It is possible to pair this strategy with current technological innovations. This means that individualised messages can be sent to customers through the method that they would be most likely to respond to.

Two main approaches of personalisation were identified by Changchien et al. 2004 and Dyche (2002). The first approach is *rule-based* personalisation while the second approach is *adaptive* personalisation. In rule-based personalisation, rules are established to dictate the personalisation. The similarity of products that have been purchased by customers are measured (Changchien et al., 2004). These established rules are normally hard-coded software and difficult to maintain. Adaptive personalisation, on the other hand, learns as time passes by using the behaviour of customers that act similarly, or it tries to find similarities between customers' behaviour. This type of personalisation will recommend items to customers based on similar behaviour of another customer. These two approaches are used in the e-commerce industry as recommender systems (Dean, 2014; Erl et al., 2015; Kamber et al., 2012).

The needs of an individual customer can be analysed by using customer data and can then be used for a direct marketing campaign. Customer profiles and customer segments are used for this purpose and are discussed in more detail in Section 2.3.

To personalise marketing campaigns, customer behaviour must be analysed and predicted to be able to run one-to-one marketing campaigns (Chen et al., 2005; Jiao et al., 2006). To enhance customer retention, one-to-one marketing is used alongside relationship marketing. Targeted marketing is very important for this study as individuals are targeted and thus a one-to-one marketing strategy should be used to ensure customer retention.

The next component of the CRM model, seen in Figure 2.1, is Customer Development. The elements of this component will be discussed in the next subsection.

### 2.2.4 Customer Development

The elements of customer development are *customer lifetime value*, *up-selling* and *cross-selling* and *market basket analysis*. These techniques will be discussed in this subsection as this study aims at improving targeted marketing for individuals based on their needs.

#### 2.2.4.1 Customer Lifetime Value

Customer lifetime value (CLV) refers to the future expected revenue that a company will gain from a specific customer based on the relationship between the company and the customer. This is based on tangible or intangible benefits, causing the customer to have value to the company (Krishna & Ravi, 2016). Different CLV models can be used to know where to allocate marketing resources for customer acquisition, customer retention and for cross-selling opportunities (Gupta et al., 2006).

Figure 2.9 shows how a firm's marketing actions influence the behaviour of the customer (acquisition, retention and expansion). This then affects the customer's CLV or in other words their profitability to the firm. The CLV of customers (current and future), often referred to as Customer Equity, eventually forms a proxy for firm value (Gupta et al., 2006).

*Recency, Frequency and Monetary* (RFM) is a model that is often used to determine the lifetime value of a customer. RFM is a data mining technique that is used to discover the nature of a customer by employing Recency, Frequency and Monetary indicators, in an attempt to identify the most profit-generating customers (Dursun & Caber, 2016). This technique defines customers that simultaneously have high recency, frequency and monetary values (Hu & Yeh, 2014) and is used to understand customers' purchasing behaviour. It is used in the retail industry to change marketing strategies as changes in the behaviour of customers are detected.

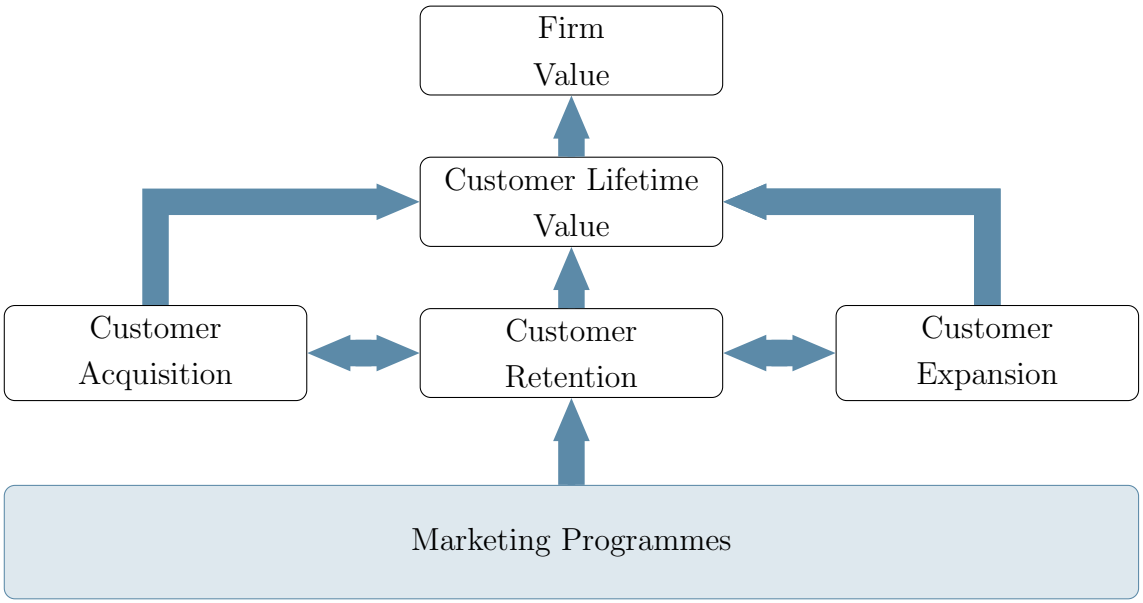



Figure 2.9: Customer Lifetime Value (Gupta et al., 2006).

The terms have the following meanings (Dean, 2014; Tsiptsis & Chorianopoulos, 2011):

*Recency*: This refers to the recency of a purchase or the time that has elapsed since a customer's most recent transaction.

*Frequency*: This is a measure that indicates how frequently the customer engages in transactions over a certain time period. This is denoted as the average number of purchases per unit of time.

*Monetary*: This indicates the average amount that a customer spends on a purchase.

Table 2.2 summarises the advantages and disadvantages of this model. CLV and RFM are mainly used on customer segments, discussed in more detail in Section 2.3. This study, however, focuses more on the analysis of individual behaviour, but CLV can still be used to gain more insight into different market segments, before focusing on the individual. The next element in the CRM component of Customer Development is Market Basket Analysis and this will be discussed in the next section.

Table 2.2: Advantages and Disadvantages of RFM (Dursun & Caber, 2016).

| Advantages | Disadvantages |
|---|---|
| It is a powerful tool used to assess customer lifetime value and can be combined with other frequent pattern mining techniques (Hu & Yeh, 2014). | It is insufficient for generating marketing campaigns based on only the three RFM indicators (Fitzpatrick, 2001). |
| It is effective for predicting response and for boosting a company's profit in a short period of time (Baecke & Van den Poel, 2011). | Frequency and Monetary values have a high correlation (Olson et al., 2009). |
| It is a good base technique that can be followed by other techniques to improve customer segmentation (Elsner et al., 2003). | The RFM indicators have different levels of importance in different industries (Bacila et al., 2012). |
| | RFM models are scoring models that do not provide a money-based value for a customer (Gupta et al., 2006). |

#### 2.2.4.2 Market Basket Analysis

Market Basket Analysis (MBA), also known as association rule learning, is a data mining technique that is used to identify which products are frequently purchased together by customers. This technique can be used in various fields, but it is commonly used in marketing.

This technique is used on large datasets to discover interesting relationships between products. To illustrate this technique, consider a simple example of a transaction given in Table 2.3 with items purchased for each transaction.

Table 2.3: Example of Market Basket Analysis (Kaur & Kang, 2016)

| Transaction ID | Items in cart |
|---|---|
| 1 | Burger, Cheese, Butter |
| 2 | Milk, Butter, Cheese |
| 3 | Butter, Milk |

An interesting relationship that can be discovered in the form of an association rule is:

$$\text{Milk} \rightarrow \text{Butter}.$$

This rule indicates that there is a strong relationship between milk and butter. This rule is used to maximise the transactional intensity of a customer (Ngai et al., 2009). These association

rules set up through MBA can be used for cross-selling opportunities, as it is known which items are frequently purchased together.

Association rules consist of two measures, *support* and *confidence*, which are explained in Table 2.4, using the following example:

**When someone buys milk, 20% of the time they also buy butter, with a support of 2%.**

Table 2.4: Association Rule Measures (Bounsaythip & Rinta-Runsala, 2001; Kamber et al., 2012)

| Measure | Explanation | Example |
|---|---|---|
| Support | The support indicates the frequency of the pattern that occurs. This indicates how many times items are purchased together. Minimum support for an association is needed for the association to be of business value. Calculating Support: $s(\mathbf{A} \Rightarrow \mathbf{B}) = P(\mathbf{A} \cup \mathbf{B})$ | Milk and Butter appear together in 2% of all the transactions that are analysed. |
| Confidence | The confidence indicates the predictability and strength of an association. This indicates the likelihood of the successor given the predecessor, or how much one item depends on the other. Calculating confidence: $c(\mathbf{A} \Rightarrow \mathbf{B}) = P(\mathbf{B}\|\mathbf{A})$ | Given that milk is purchased there is 20% confidence that Butter will also be purchased. |

This means that product **A** and product **B** only appear together in 2% of all transactions, but when product **A** appears in a transaction, there is a 20% chance that product **B** will also appear in that transaction (Kamber et al., 2012).

MBA is an appropriate approach to create customer profiles for this study, because it is based on analysing customers' transactional data. This technique is suitable for identifying cross-selling and up-selling opportunities which will be discussed in Subsection 2.2.4.4, but first Sequential Pattern Analysis will be discussed as this technique adds a time element to MBA, and for the purpose of this study it should be investigated.

#### 2.2.4.3 Sequential Pattern Analysis

Mooney & Roddick (2013), defined *Sequential Pattern Analysis* (SPA) as "Given a database of sequences, where each sequence consists of a list of transactions ordered by transaction time and each transaction is a set of items, sequential pattern mining is to discover all sequential patterns with a user-specified minimum support, where the support of a pattern is the number

of data-sequences that contain the pattern". In other words SPA exists when adding a time factor to market basket analysis, which creates an analysis of associations over time, identifying a series of events or patterns in a specific sequence. The provided definition focuses on transactional data, but the field has since expanded to applications in other industries such as health care, telecommunication and biotechnology, and the definition has since also been reformulated by Mooney & Roddick (2013).

SPA just like MBA can be used for marketing strategies such as cross-selling and up-selling, as this analysis can possibly predict the products that a customer is most likely to buy next (Dyche, 2002). More advantages of SPA are summarised in Table 2.5.

Table 2.5: Advantages of SPA (Bounsaythip & Rinta-Runsala, 2001)

| Advantage | Explanation |
|---|---|
| Cross-marketing and timing | SPA is useful for marketing new products at the right time based on the sequential association rules. |
| Coupons and discounting | Discount can be offered on products that are frequently bought after each other or on products that are frequently bought together. |
| Product Placement | Products with strong relationships based on SPA can be placed together or close to each other to take advantage of the natural correlation between products. |

As said MBA and SPA can both be used for up-selling and cross-selling, up-selling and cross-selling will be discussed in the next subsection.

#### 2.2.4.4 Up-selling and Cross-selling

Up-selling and cross-selling marketing techniques aim at raising the value of a single sale transaction made by a customer. They are used to reduce the risk of a customer migrating to a competitor (Kubiak & Weichbroth, 2010).

Up-selling means to move 'up' to a higher-grade (and more expensive) item, for an item that they planned to purchase. For example refer to Figure 2.10. Up-selling is when a customer wants to buy a cellphone and the sales person persuades the customer to buy a higher end one. Schiffman (2005) defines up-selling as "it is what happens when you take initiative to ask someone who already has purchased something and you offer to purchase more of it – or more of something else". The main reason for this promotional technique is to gain better unit profit, for a more expensive product (Kubiak & Weichbroth, 2010).

Cross-selling is a technique used to sell additional products to a customer. This usually involves selling items that complement the original product that the customer purchased. In Figure 2.10, it is shown that when a customer is already buying a cellphone, an offer of headphones can be made to them. This attempts to capture a larger share of the customer market through individualised customer needs being met.

Identifying cross-selling and up-selling opportunities is important for customer development as it can increase the intensity of customer transactions and ultimately gain a more profitable customer.


Figure 2.10: Up-selling vs Cross-selling

To effectively use up-selling and cross-selling companies must understand their customers, so that the company can offer their customers the right product at the right time. Looking at the opposite by offering customers products that they do not need can have a negative effect on customer relationships (Paas & Molenaar, 2005).

Identifying up-selling and cross-selling opportunities requires three main objectives (Salazar et al., 2007):

1. Identifying and understanding the acquisition pattern of a customer.
2. Identification of factors that influence the decision to repurchase by a customer.
3. Forecasting the time that the customer will possibly repurchase.

The third objective relates directly to the objectives of this study, i.e. to predict when a customer will purchase a certain item. The next section will look more into profiling and

segmenting customers. When successfully profiled or segmented, up-selling and cross-selling opportunities can be identified.

## 2.3 Customer Profiling and Customer Segmentation

Customer segmentation and profiling is used by companies to target their most important customers. If this is used the focus is entirely on the customer and their importance. It is used to analyse and understand customer behaviour which can be used to retain and develop customers.

Customer profiling is used to create a portrait or profile of a customer. This portrait includes both transactional (behavioural) data as well as personal (factual) data (Adomavicius & Tuzhilin, 2001; Upadhyay et al., 2016). These details are then used to describe what the customer does and who the customer is. Customer profiles are created to individualise customers to understand their needs, which are largely unknown to a company. Following profiling, customers are better understood and can be targeted individually. A profile is then used to decide which strategy to use when marketing to the customer (Shaw et al., 2001), and this is also used by marketers to present an offer to the customer which they are most likely to respond positively to (Lanjewar & Yadav, 2013).

Customer profiles are used to predict the behaviour of a customer, by discovering patterns in the collected behavioural data. In this study, customer transactional data can be seen as the behavioural data. Profiling is used to discover knowledge from the personal and or transactional data of a customer that was not previously known (Bounsaythip & Rinta-Runsala, 2001; Lanjewar & Yadav, 2013). Understanding the customer and his behaviour is very important in the context of this study as an individual will be investigated. Customer profiles and how they are developed will be discussed later in this section.

Customer segmentation happens when customers are divided into homogenous groups (called segments), based on characteristics or habits shared between them (Krishna & Ravi, 2016). Some very trivial customer segments such as gender groups, age groups or geographical location can sometimes expose behavioural attributes that would not have been picked up when just looking at customer data, hence it can help marketers achieve better performance (Upadhyay et al., 2016). This is similar to customer profiles as they are used to identify unknown needs, in this case of a segment of customers. When customers are segmented, a marketing mix can be developed to best reflect the needs of each segment. This could yield a higher return on investment as less effort is spent when designing the marketing mix as the company does not have to focus on the needs of customers that would never have been satisfied anyway (Dolnicar et al., 2018).

Customer profiling and customer segmentation are often used interchangeably, but they do have different meanings. The difference between customer segmentation and customer profiling is visually depicted in Figure 2.11. In (a) customer profiling can be seen and in (b) customer segmentation can be seen. For the purpose of this project the two terms will be used as described in this section.

(a) Customer Profiling (b) Customer Segmentation

Figure 2.11: The difference between Customer Profiling and Customer Segmentation

Table 2.6 gives examples of where customer profiling and customer segmentation have been used. Some references argue that customer segmentation should happen before customer profiling, but literature does not agree as to whether there should be an order in which to perform these techniques. Because this study focuses on the individual customer, profiling could work, but if little to no information is available about a customer (say a new customer enters the system or has made only one transaction), it could be necessary to use customer segmentation.

Table 2.6: Examples of Segmentation, Profiling and both techniques used together

| **Examples of segmentation** | **References** |
|---|---|
| Customers from an online store in Romania were segmented based on the Recency, Frequency and Monetary (RFM) criteria, and found that a small number of customers contribute to more than 50% of the company's total sales. | Pater et al., 2019 |

Table 2.6 continues on next page

| Examples of segmentation | References |
|---|---|
| A grocery chain operating in Turkey segmented customers into five different groups: high contribution loyal customers, low contribution loyal customers, uncertain customers, high spending lost customers and low spending lost customers. For each of these groups unique CRM and marketing strategies were recommended. | Peker et al., 2017 |
| An automobile company in Taiwan segmented their customers based on customer transaction behaviour and customer satisfaction variables. Based on two clustering methods, k-means and expectation maximisation they grouped their customers into four clusters: loyal, potential, VIP and churn. Based on these clusters they identified unique/customised marketing strategies for each segment. | Tsai et al., 2015 |
| **Examples of Customer profiling** | **References** |
| Individual customer behaviour was modelled by constructing conjunctive rules. An example of such a profile is that "John Doe usually buys lemon juice at RiteAid, more specifically, in 95% of the cases when he buys lemon juice, he buys it at RiteAid". These profiles were constructed using data mining techniques. These profiles are used to personalise marketing. | Adomavicius & Tuzhilin, 2001 |
| YouTube social media data was used to create user profiles in real time. To advance these user profiles, different social media platforms were sourced to generate even richer profiles. This was done to create corporate insight into competitive marketing. | An et al., 2017 |
| *Target* tracks customer behaviour and profiles their customers. One known application of their customer profiling is identifying pregnant women and personalising marketing to these women. This is done in general to improve customer satisfaction, identify promotional strategies and to increase revenue. | Corrigan et al., 2014 |
| **Examples of Customer profiling and customer segmentation** | **References** |
| To manage existing credit card customers in a bank a proposed integrated data mining and behavioural scoring model was implemented. Groups of customers were grouped and identified by using a self-organising map neural network, based on repayment behaviour and RFM scoring. Three groups were obtained from this segmentation. From these segments customers were profiled by customer features determined by using association rules. The study shows the use of segmentation and profiling for marketing strategy development. | Hsieh, 2004 |
| Vodafone customer profiles were created and customers were segmented, based on the profiles created. The profiles were based on age, gender, income and lifestyle features. No practical application was performed based on this study, but this can be used for personalised marketing campaigns. | Jansen, 2007 |

End of Table 2.6

Profiling is done by using collected information of a customer to build a model of the customers' behaviour (Bounsaythip & Rinta-Runsala, 2001). Customer profiling, unlike customer segmentation, cannot be done if little information is available about the customer, thus a complete set of customer data must be available to be able to profile a customer. Features used for profiling are dictated by the availability of the data and the development techniques. A factual profile can be derived from transactional data as well as demographical data. The behavioural profile can in turn be derived from customer transactional data (Adomavicius & Tuzhilin, 2001; Bounsaythip & Rinta-Runsala, 2001). This transactional data will be used for this study. Other sources of behavioural data can be social media data and online web usage data.

Shaw et al. (2001) provides a list of transactional characteristics that help with marketing decisions. This list consists of:

- Size of purchase,
- the recency of purchases,
- the frequency of purchases,
- identification of typical customer groups,
- computing the lifetime value of customers,
- information regarding prospective customers,
- successful or failed marketing programmes used in the past.

Some of these characteristics will be needed for this study, to gain better knowledge of the customer. Techniques such as CLV and MBA discussed in Section 2.2.4 can be used in this study to gain more insight about the customer.

## 2.4 Conclusion: Chapter 2

In this chapter customer behaviour and the management thereof was discussed along with an overview of marketing and marketing strategies. The customer relationship management concept was discussed and each of its elements was discussed in more detail. Different techniques for each of the elements shows that machine learning can be used in the CRM process. This leads into the next chapter which will study data analytics, machine learning, predictive analytics, along with machine learning methods to predict future events. It is important for this study to identify the necessary techniques to be able to develop the NPD Predictor. CRM is a large field and it is only touched on in the context of what must be achieved in this study.

# Chapter 3

# Data Analytics, Machine Learning and Future Event Prediction

The previous chapter discussed customer behaviour and customer relationship management. The CRM process shows a lot of potential for machine learning, and this study will investigate if machine learning can be used for the Next Purchase Date (NPD) Predictor. In this chapter the data analytics process will be discussed along with machine learning and predictive analytics. It will also look at how future events can be predicted.

## 3.1 Data Analytics

Machine learning cannot be excluded from the bigger process called *data analytics*. Machine learning contains techniques used in the process of data analytics, and therefore the entire process will be discussed to see how machine learning fits into the bigger picture. Data analytics is the use of data to gain actionable knowledge and insights, by formulating a hypothesis that needs to be evaluated through data (Rajaraman, 2016). Data analytics can be grouped into four main types of analytics which include: (i) descriptive analytics, (ii) diagnostic analytics, (iii) predictive analytics and (iv) prescriptive analytics (Rajaraman, 2016). Figure 3.1 shows how these types of analytics increase in computational difficulty, while also increasing the value added to a company. It also displays the question that is answered by each type of analysis. The four types of analytics include:

i *Descriptive analytics* is the preliminary stage in data processing where a summary of historical data is created to yield useful results in the form of information. This looks at historical data and organises the data so that trends and relationships can be identified that would otherwise not be visible (TechTarget, 2015). An example of a question that can be answered by using descriptive analytics is: What was the sales volume of a specific product over a given period of time?

ii *Diagnostic analytics* takes a deeper look at the data to try to understand why a specific event occurred. It aims at determining the cause of a specific phenomenon that occurred in the past and focuses on the reasons for the events that occurred (Erl et al., 2015). This form of analysis requires a more advanced skill set than descriptive analytics, but provides more value, as seen in Figure 3.1. An example of a question that can be answered by diagnostic analysis is: Why were the sales of Product A less than the sales of Product B?

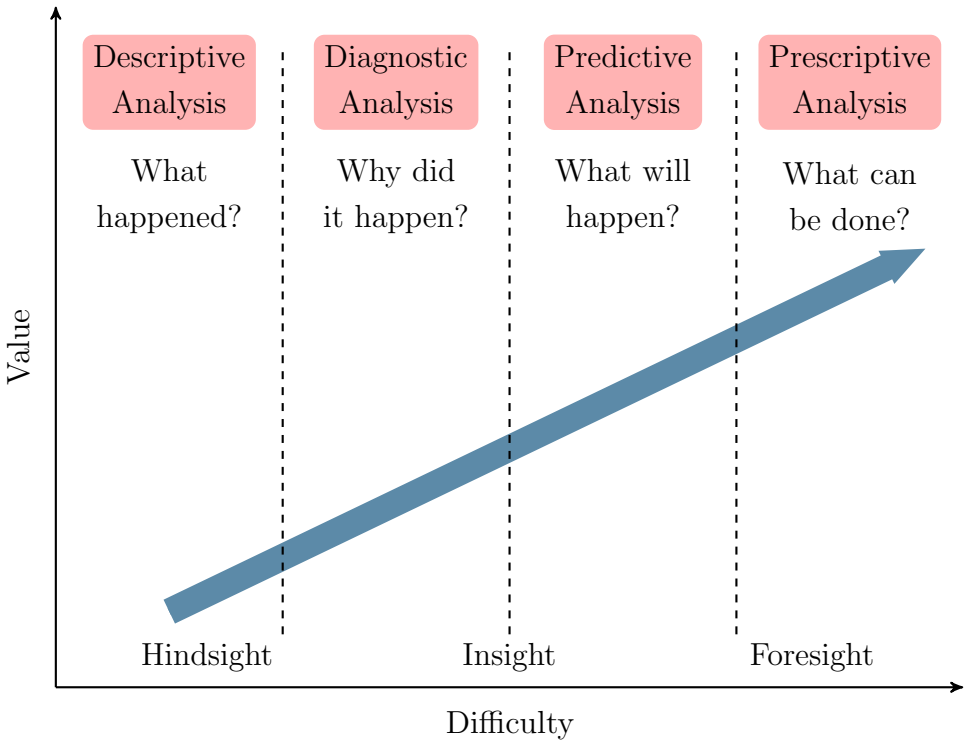



Figure 3.1: The value of Machine Learning through different types of analysis (Poh, 2019)

iii *Predictive analytics* aims at being able to know what will happen in the near future, by utilising available data. Future predictions are made based on events that happened in the past. This answers the question of what will happen in the future. An example of a question that can be answered by using predictive analytics is, if a customer has purchased both Product A and Product B, what are the chances that the customer will also buy Product C? (Erl et al., 2015). Predictive analysis is widely used in marketing to predict customer behaviour (Erevelles et al., 2016). Tools that are used to perform this type of analysis include machine learning algorithms and time series analyses using statistical methods.

iv *Prescriptive analytics* builds upon the results obtained from predictive analytics, by prescribing actions that can be taken. It answers the question of which option is best to follow as well as why this option should be followed. Through this analysis, decision suggestions are made to take best advantage of future opportunities or to mitigate potential future risk (TechTarget, 2015). This type of analysis yields the most value to a company, but is also the most difficult to perform as seen in Figure 3.1. An example of a question that can be answered through this type of analysis is: When would be the best time to trade a particular stock? (Erl et al., 2015).

Asking the right questions and knowing how to answer them, can lead to very useful answers and insights for companies. Each of these analysis types is useful, but answers completely different questions and therefore it is important to understand the question that should be answered, before trying to analyse it.

### 3.1.1 Data Analytics Processes

Formal processes for data analytics exist, and three of these are subsequently discussed. These three processes are: Knowledge Discovered in Databases (KDD); Sample, Explore, Modify, Model and Access (SEMMA), and Cross Industry Standard Process for Data Mining (CRISP-DM), as they are popular data analytics processes.

#### 3.1.1.1 KDD Process

The Knowledge Discovery in Databases (KDD) process was defined by Fayyad et al. (1996) as, "the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data". This means that through data mining methods, low-level data can be transformed to high-level knowledge. The process is depicted in Figure 3.2 and consists of five main steps (Azevedo & Santos, 2008):

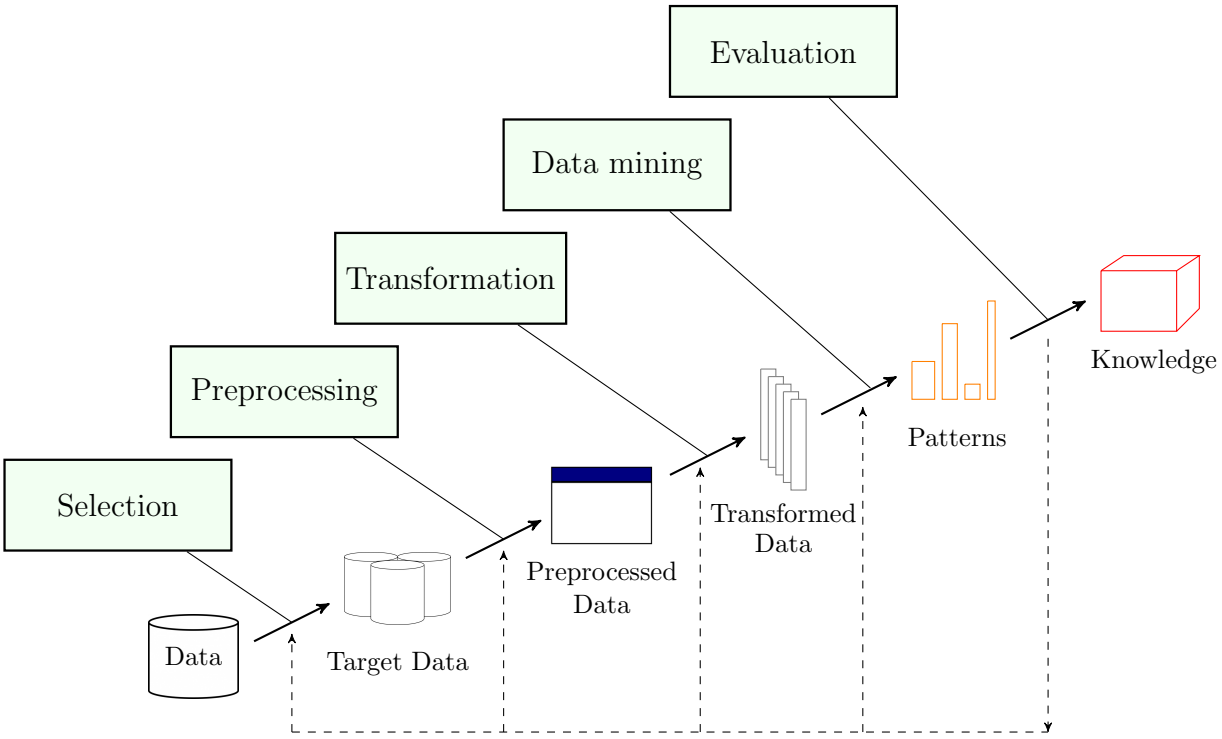



Figure 3.2: An overview of the steps that compose the KDD process (Fayyad et al., 1996)

1. *Selection:* Creating a target dataset, or focus on a subset of variables or data samples that is explored for discovery.

2. *Preprocessing:* Cleaning the data as well as preprocessing the data to obtain consistent data.

3. *Transformation:* Transform the data by using dimensionality reduction techniques or transformation models.

4. *Data mining:* Search for patterns that are of interest in a particular representational form.

5. *Interpretation/Evaluation:* Interpreting and evaluating the mined patterns.

Each step in the process transforms the data to a more useful state, closer to the end goal. This process is an interactive and iterative process, as indicated by the arrows in Figure 3.2 and consists of numerous steps. These steps require decision-making by the user to ultimately arrive at a useful solution. This process is application driven, and must be used with clear knowledge of the goal that the end user has (Fayyad et al., 1996).

#### 3.1.1.2 SEMMA Process

The Sample, Explore, Modify, Model and Assess (SEMMA) process was developed by the *SAS Institute* and also forms part of their *Enterprise Miner* data mining support tool. This process focuses mainly on the model development aspect of data mining (Mariscal et al., 2010). The SEMMA process can be seen in Figure 3.3.


Figure 3.3: Overview of the SEMMA process (Mariscal et al., 2010)

As seen from Figure 3.3, the SEMMA process consists of the following five stages (Azevedo & Santos, 2008):

1. *Sample:* This is an optional stage where a portion of the large dataset is extracted. The extracted dataset should be large enough to contain significant information, but small enough to manipulate easily.

2. *Explore:* During this stage the data is explored by searching for unanticipated trends and anomalies to gain an understanding of the data.

3. *Modify:* During this stage of the process the data is modified by creating, selecting and transforming variables in preparation for data modelling. This step includes both data cleaning and data transformation.

4. *Model:* During this stage a model is built that allows the software to search automatically for combinations and trends in the data to reliably predict a desired outcome.

5. *Assess:* In this stage the usefulness and reliability of the results from the previous step are evaluated.

#### 3.1.1.3 CRISP-DM Process

The third and final data mining process that will be discussed is Cross Industry Standard Process for Data Mining (CRISP-DM). The process can be seen in Figure 3.4. The arrows indicate the dependencies between the stages in the process, and the outer circle indicates that the process has a cyclical nature and that the lessons learnt during the data mining process and the deployment solution can trigger new, sometimes more focused, business questions (Chapman et al., 2000). This process was described by Shearer (2000), as a comprehensive data mining methodology and process that could be used by all from novice users to data mining experts, because it provides a complete blueprint for conducting a data mining project. This process consist of six stages (Azevedo & Santos, 2008; Shearer, 2000):

1. *Business understanding:* During this stage the project objectives are defined from a business perspective. This knowledge is then translated to a data mining problem, and a preliminary plan is designed to meet the objectives.

2. *Data understanding:* During this stage data is collected, the quality of the data is checked and the data is explored to get insight so that hypotheses can be formulated.

3. *Data preparation:* During this stage of the process the final dataset is selected and prepared for use. This stage covers all activities that are performed to construct the final dataset, from the initial raw data.

4. *Modelling:* During this stage various modelling techniques are selected and applied to the data. The parameters are then calibrated to optimal values.

5. *Evaluation:* During this stage the obtained models are evaluated. The models' construction is evaluated to be certain that the business objectives are achieved, and that the results obtained can be used.

6. *Deployment:* During this stage the knowledge gained through the model must be organised and presented in a way that makes it possible for the customer to use it. Depending on the requirements, the deployment stage can be as complex as implementing a repeatable data mining process across an entire company or as simple as compiling a report.

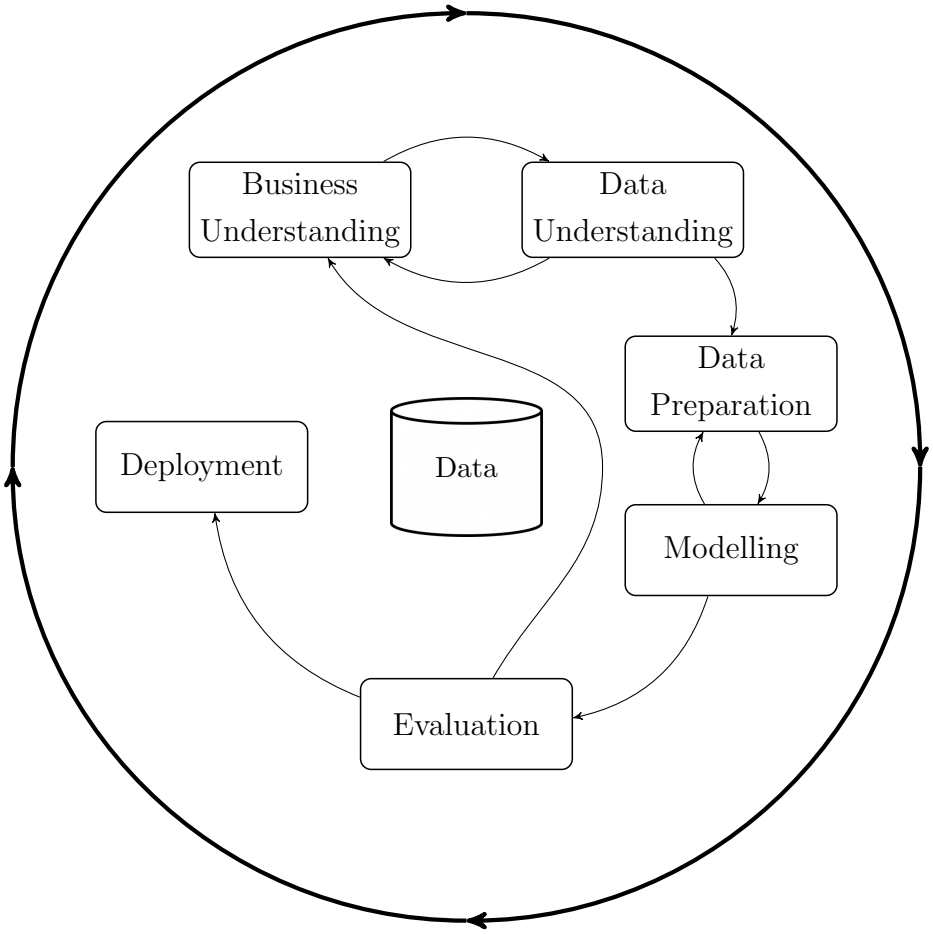



Figure 3.4: CRISP-DM Process for data mining (Chapman et al., 2000)

### 3.1.2 Comparison of Data Analytics Processes

When comparing the three processes it is clear that they are very similar. When comparing KDD and SEMMA the processes are almost identical as Selection can be associated with Sample, Preprocessing can be associated with Explore, Transformation can be associated with modify, Data Mining can be associated with Model and Evaluation can be associated with Assess.

When comparing the KDD process with the CRISP-DM process, it also shows similarities, but they are not as straightforward as with the previous comparison (Azevedo & Santos, 2008).

The CRISP-DM process has an additional stage, Business Understanding as the first step in the process. None of the steps from the KDD process correspond to this step. This step precedes the Selection step in the KDD process. The CRISP-DM process also has a Deployment stage that happens after evaluation of the KDD process when knowledge is incorporated into the system. Furthermore, the Data Understanding phase can be identified as the combination of the Selection and the Preprocessing stages of the KDD process. The Data Preparation stage can be identified as the Transformation stage, the Modelling stage can be identified as the Data Mining stage and the Evaluation stage are the same for both processes. A summary of this comparison can be seen in Table 3.1.

Table 3.1: Correspondence between KDD, SEMMA and CRISP-DM (Azevedo & Santos 2008; Shafique & Qaiser 2014).

| Data Analytics Process | **KDD** | **SEMMA** | **CRISP-DM** |
|---|---|---|---|
| Number of Steps | 5 | 5 | 6 |
| Name of Step | – | – | Business Understanding |
| | Selection | Sample | Data Understanding |
| | Preprocessing | Explore | |
| | Transformation | Modify | Data preparation |
| | Data mining | Model | Modelling |
| | Evaluation | Assess | Evaluation |
| | – | – | Deployment |

SEMMA and CRISP-DM can be viewed as an implementation of the KDD process (Azevedo & Santos, 2008). The comparison shows that CRISP-DM is more complete than SEMMA, as it starts with an understanding of what should be investigated and ends with deployment of the solution or the knowledge gained through the process. The above-mentioned processes are the most widely used processes, but there are more processes available in the academic domain as well as in practice that can be used.

## 3.2 Machine Learning

Machine learning will be used to predict the Next Purchase Date (NPD) of a customer and therefore it is important to understand what machine learning is and how it can be used for predictive analysis. Machine learning has become a popular concept in recent years and there are various definitions for it. Mitchell (1997), defined machine learning as, “A computer

program is said to learn from experience $\mathcal{E}$ with respect to some class of tasks $\mathcal{T}$ and performance measure $\mathcal{P}$, if its performance at task $\mathcal{T}$, as measured by $\mathcal{P}$, improves with experience $\mathcal{E}$". This definition helps one to think about the data that should be collected (E), the decisions that the software should make (T), and how the result will be evaluated (P). Take playing Checkers for example: E would be the experience of playing many games of Checkers, T would be the task of actually playing Checkers and P would be the probability that the program will win the next game.

Samuel (1959) described machine learning in a more informal way claiming that "A computer can be programmed so that it will learn to play a better game of Checkers than can be played by the person who wrote the program", meaning that computers can learn, without being explicitly programmed to do so. For computers to learn, concepts and results from many fields such as statistics, information theory, philosophy, cognitive science, control theory, artificial intelligence and computational complexity are used (Mitchell, 1997). Machine learning in the context of this project will be investigated to discover patterns in transaction history of customers, to predict future purchasing behaviour.

### 3.2.1 History of Machine Learning

Machine learning is widely used today in various industries, with a lot of different applications. Some of the applications include self-driving cars, smart assistants and filtered social media feeds. However, this was not all discovered in one day. Some algorithms used in machine learning were already discovered in the 1800s. Table 3.2 gives a brief overview of what has happened in the past few decades that has contributed towards modern machine learning.

As seen from Table 3.2 over the years increasingly more work has been done in the field of machine learning. Regarding game playing, in the last decade world Champion Go player Lee Sedol was beaten by Google DeepMind's AlphaGo (BBC, 2016); some of the world best Jeopardy! players were beaten by IBM's Watson (Ferrucci, 2012) and the world's best Dota2 players were beaten by OpenAI's bot (OpenAI Five, 2019). But machine learning, as already mentioned, is used for far more than playing games. It is used in various industries, such as healthcare, logistics planning and fraud detection (Russel & Norvig, 2010), and has become very popular in the last decade as collecting information has become easier.

Table 3.2: History of Machine Learning (reproduced from BBC Academy (2019))

| | |
|---|---|
| <1940s | • Thomas Bayes did breakthrough work in the eighteenth century that led to Bayes' Theorem discovered by Pierre-Simon Laplace (1812). |
| | • Adrien-Marie Legendre developed the Least Square Method for data fitting (1805). |
| | • Andrey Markov invented Markov Chains (1913). |
| 1940s | • Stored-program computers (computers that can hold instructions) were developed. This started a modern computing revolution. |
| 1950s | • Pioneering machine learning research was conducted using simple algorithms. |
| | • The first artificial neural network, a computer-based simulation of the way the brain works, was built by Marvin Minsky and Dean Edmonds (1951). |
| 1960s | • Bayesian methods were introduced for probabilistic inference in machine learning. |
| 1970s | • The first 'Artificial Intelligence (AI) Winter' as failure of machine translation and overselling AI's capabilities led to reduced funding. |
| | • Different names such as informatics, machine learning and computational intelligence came about, to replace AI. |
| 1980s | • Rediscovery of back-propagation causes a resurgence in machine learning research. |
| 1990s | • IBM's computer, Deep Blue, beat world chess champion Garry Kasparov; this increased public awareness of machine learning. |
| | • Machine learning shifted from knowledge-driven to data-driven, analysing large amounts of data to draw conclusions. |
| 2000s | • Back-propagation, Support Vector Clustering and other methods became more widespread. |
| 2010s | • Machine learning became used in many software services and applications. |

### 3.2.2 Data Preprocessing and Transformation

Before a machine learning model can be trained, the data must first be preprocessed and transformed. These steps are done so that the model can be more accurate as cleaner data is used for the model. During the preprocessing stage the data must be explored to gain a deeper understanding of it. After the data has been explored, and the appropriate features have been created, the datasets can be split into training, validation and test sets. If needed, dimensionality reduction techniques can be performed on the data. These techniques transform data in high-dimensional space to a space of fewer dimensions. This is advantageous as it reduces the storage space as well as the time required for processing. It is also easier to visualise

data with less dimensions. *Principal Component Analysis* (PCA) and *Linear Discriminant Analysis* (LDA) are two well-known dimensionality reduction techniques and they will be discussed briefly.

#### 3.2.2.1 Dimensionality Reduction Techniques: Principal Component Analysis

PCA is a technique used to reduce a large set of variables into a smaller set. The smaller set still contains most of the information included in the large set. The technique projects the data along the direction where the data varies the most.

The changes in one variable are associated with change in another variable and is described by the covariance of the two variables. The covariance matrix

$$\begin{bmatrix} \text{cov(x,x)} & \text{cov(x,y)} \\ \text{cov(y,x)} & \text{cov(y,y)} \end{bmatrix},$$

expresses the covariance for all the variables of a dataset. The principal components (direction of most variance) are obtained by performing eigenvalue decomposition on the covariance matrix. The eigenvector that corresponds to the largest eigenvalue gives the direction of most variance. The secondary axis is obtained by the second biggest eigenvalue (Jolliffe, 2011).

Some variables capture more of the values' variation than others, thus the principal components with the largest values correspond to the dimensions with the most variance. In many cases with high dimensional data, most of the variance is captured within a small subspace of the data. Figure 3.5 shows a two-dimensional dataset, captured in the original data space with axis $x_1$ and $x_2$. The original data is best captured by the rotated data space, where $\mathcal{X}_1$ is the first principal and $\mathcal{X}_2$ is the second principal component (Sorzano et al., 2014).

#### 3.2.2.2 Dimensionality Reduction Techniques: Linear Discriminant Analysis

Linear discriminant analysis (LDA), just like PCA, reduces the number of dimensions of a dataset. PCA gives the axis that has the biggest variance of the dataset, while ignoring the class to which the data belongs. For LDA the class labels are known, so this technique obtains the axis with maximum separation between classes (Raschka, 2015), as seen in Figure 3.6. LDA maximises the the Fisher criterion. This criterion maximises the distance between the means while minimising the scatter between points within the same class (Mikat et al., 1999).

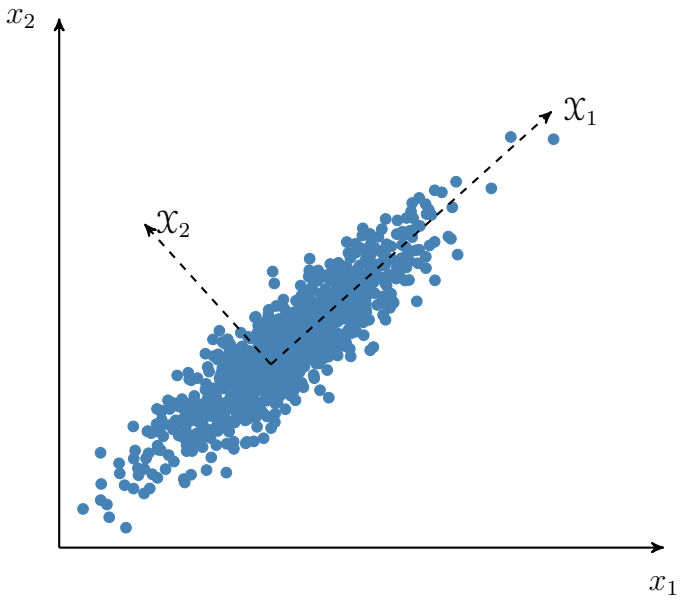



Figure 3.5: PCA transformation in two dimensions (Sorzano et al., 2014).

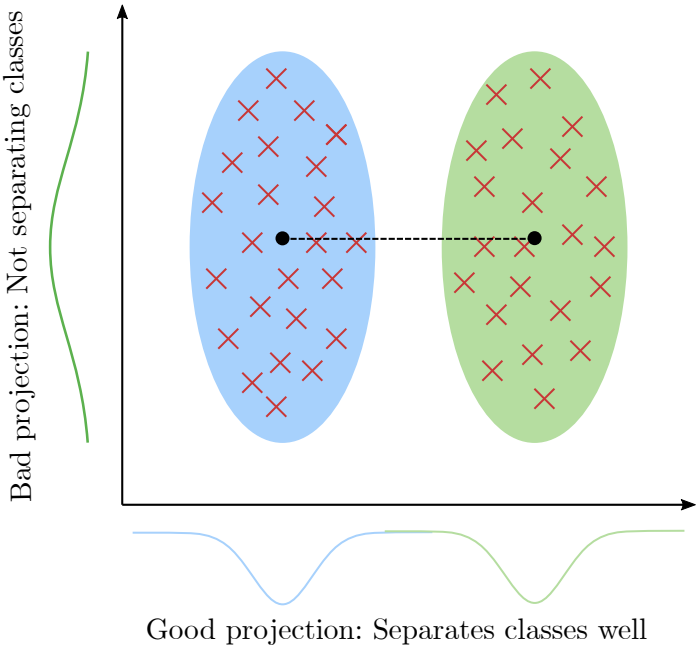



Figure 3.6: Linear Discriminant Analysis with two classes

#### 3.2.2.3 Comparison of LDA and PCA

A comparison of the axis projection for PCA and LDA can be seen in Figure 3.7. Both PCA and LDA are represented on this graph. It can be seen that in this case LDA separates the classes better than LDA. The performance of LDA compared to PCA is dependent on the

specific application that it is used and both can be tested to compare what works best for a specific application.

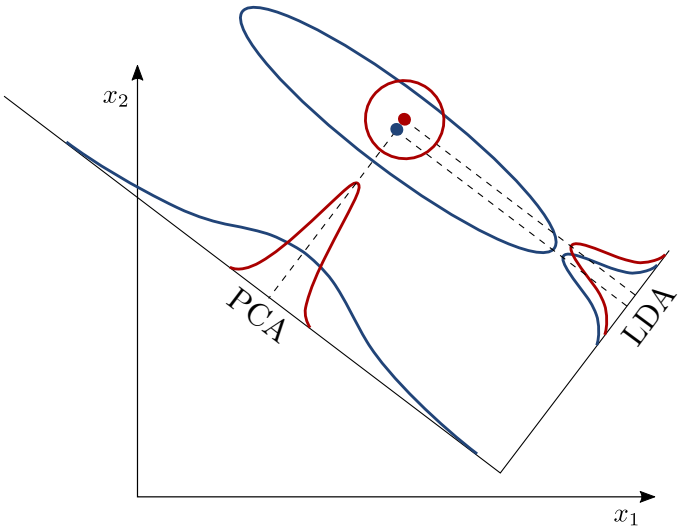



Figure 3.7: PCA versus LDA (Guo, 2017).

After features are created and if necessary dimensionality reduction has been performed, a machine learning model can be built by using various machine learning algorithms. These are discussed in the next section.

### 3.2.3 Creating a Machine Learning model

Machine learning algorithms are used in the modelling stage of the data analytics process. Different machine learning algorithms exist, that differ in the training data that is available, and how this data is used as input to obtain the specified output (Mohri et al., 2012). These algorithms are used to perform different tasks as different algorithms have unique strengths and weaknesses. These algorithms are categorised in three main categories; namely, (i) supervised learning, (ii) unsupervised learning and (iii) reinforcement learning (Lison, 2012).

i *Supervised learning (SL):* The desired output is developed by analysing datasets that are provided. The algorithm trains a model based on previous data that is labelled. Supervised learning is further categorised into classification and regression problems. Classification problems are problems that have discrete outputs, whereas regression problems have continuous output functions. For example, a supervised learning classification problem could be predicting if a customer will purchase a specific item on a specific day given the person's historic purchasing behaviour, whereas a supervised learning regression problem would be predicting the price of an item, when provided with the previous prices of that item.

ii *Unsupervised learning (UL):* In contrast to supervised learning, UL is used when little or no information is available about what the output or result should look like. This technique mostly uses clustering techniques to find relationships between variables in the data, for example, clustering customers together based on purchasing behaviour.

iii *Reinforcement learning (RL):* This technique uses trial-and-error interactions and then determines what the desired behaviour or output is. This is done in a dynamic environment. The learning algorithm has a reward that should be optimised. The technique uses agents and for each trial the agents are either penalised or rewarded based on their performance. The behaviour is then updated for the next trial, until their actions yield the optimal reward. Reinforcement learning is typically used to train bots that can play computer games or other games such as chess and robot soccer.

To perform machine learning one or more techniques can be used. Regression, classification and clustering are the three main techniques:

i Regression: It is mostly used to explore relationships that occur within the dataset, between a dependent variable and independent variables.

ii Classification: It is used to divide the dataset into predefined classes with associated class labels.

iii Clustering: It is used to cluster the data based on certain characteristics in the data.

Table 3.3 shows the machine learning techniques and the algorithms that they are associated with. The table also provides selected sources that explain how the techniques work as well as applications of the techniques. It is not possible to give an exhaustive literature reference so the reader is provided with a succinct set of pointers. A brief description of each technique is given in Tables 3.4 – 3.6 along with applications of the techniques. The relationships between the learning algorithms and the techniques are depicted in Figure 3.8.

Table 3.3: Machine learning techniques (USMA, 2017)

| **Techniques** | **Algorithms** | | | **Typical Sources** |
|---|---|---|---|---|
| | SL | UL | RL | |
| **Regression** | | | | |
| Linear Regression | ✓ | | | Gera & Goel (2015)<br>Dean (2014)<br>Salkind (2007)<br>Yang et al. (2017)<br>Shalev-Shwartz & Ben-David (2014) |

Table 3.3 continues on next page

| Techniques | Algorithms | | | Typical Sources |
|---|---|---|---|---|
| | SL | UL | RL | |
| Non-linear Regression | ✓ | | | Gera & Goel (2015)<br>Chatterjee & Hadi (2015);<br>Riffenburgh (2012)<br>Ruckstuhl (2010)<br>Tellis & Ambler (2007) |
| Logistic Regression | ✓ | | | Caruana & Niculescu-Mizil (2006);<br>Chatterjee & Hadi (2015)<br>Salkind (2007);<br>Riffenburgh (2012);<br>Karp (1999);<br>Montgomery et al. (2013)<br>Abdou & Pointon (2011) |
| **Classification** | | | | |
| Neural Networks | ✓ | ✓ | | Dean (2014)<br>Kamber et al. (2012) |
| Kernel Estimator | ✓ | | | Salkind (2007)<br>Kamber et al. (2012);<br>Larose & Larose (2014) |
| Decision Trees | ✓ | | | Caruana & Niculescu-Mizil (2006)<br>Larose & Larose (2014) |
| Support Vector Machines | ✓ | | | Caruana & Niculescu-Mizil (2006)<br>Dean (2014)<br>Jansen (2007) |
| Naive Bayes | ✓ | | | Kamber et al. (2012) |
| **Clustering** | | | | |
| Partitioning methods | | | ✓ | Maimon & Rokach (2010)<br>Kamber et al. (2012)<br>Berkhin (2006)<br>Äyrämö & Kärkkäinen (2006) |
| Hierarchical methods | | | ✓ | Maimon & Rokach (2010)<br>Berkhin (2006)<br>Pierson (2015) |
| Density-based Methods | | | ✓ | Maimon & Rokach (2010)<br>Kamber et al. (2012)<br>Berkhin (2006)<br>Ester et al. (1996) |
| Grid-based Methods | | | ✓ | Kamber et al. (2012)<br>Berkhin (2006)<br>Bounsaythip & Rinta-Runsala (2001) |

Table 3.3 continues on next page

| Techniques | Algorithms | | | Typical Sources |
|---|---|---|---|---|
| | SL | UL | RL | |
| Self-organising maps | | | ✓ | Bounsaythip & Rinta-Runsala (2001)<br>Tsiptsis & Chorianopoulos (2011)<br>Wu & Chow (2004) |
| | | | | End of Table 3.3 |

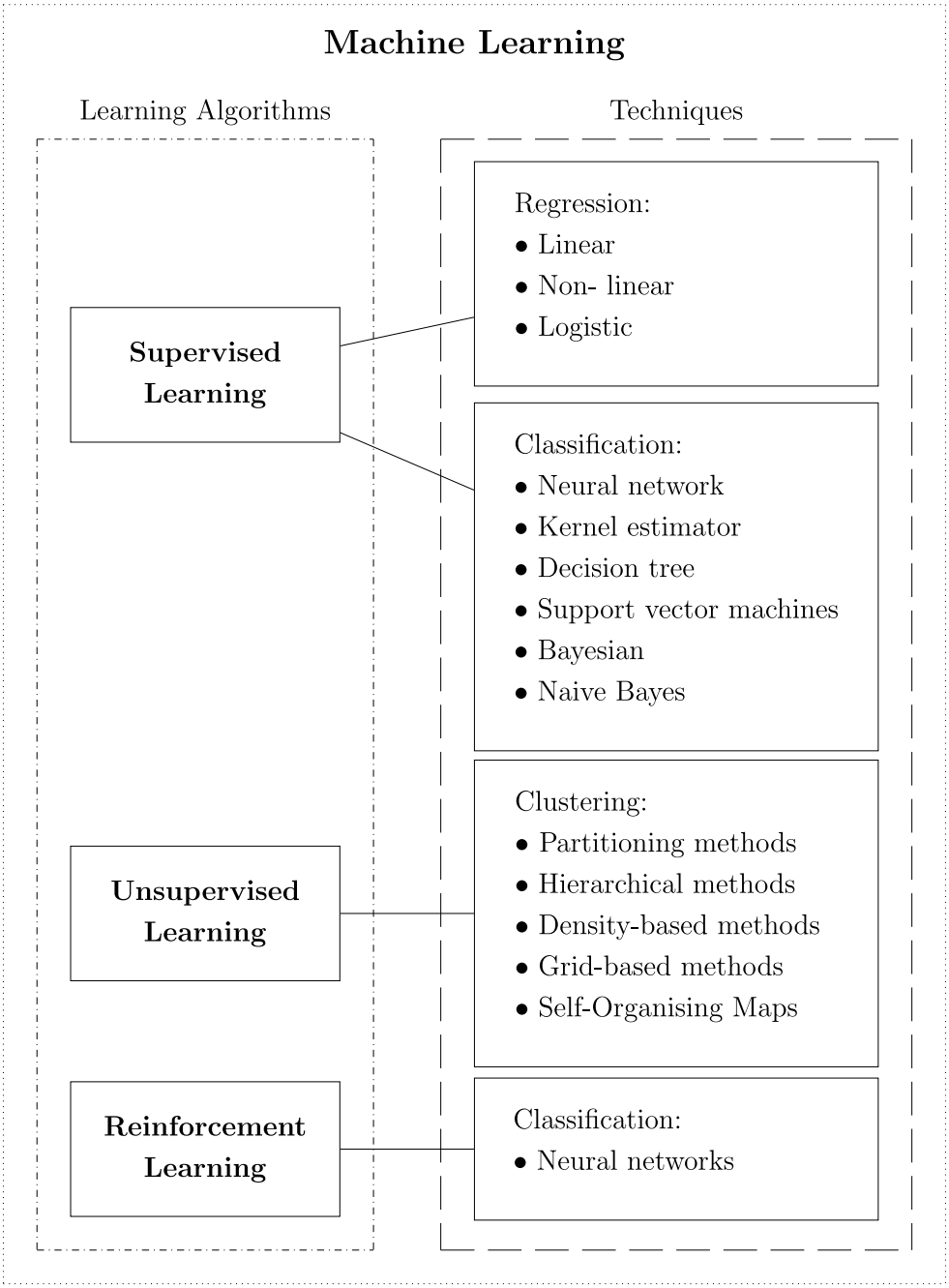



Figure 3.8: Machine learning algorithms (modified from USMA (2017))

Table 3.4: Regression Techniques

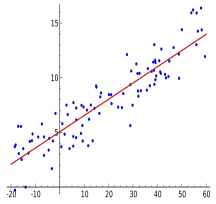
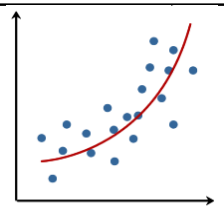
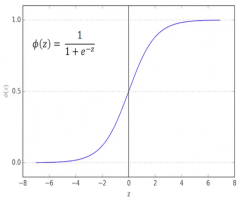
| Regression Techniques | Description | Application | Diagram |
|---|---|---|---|
| **Linear Regression** | A statistical method that attempts to model the relationship between two or more variables, by fitting a straight line on the data. This indicates the impact that one variable has on the other variables. | Assess Insurance/ Financial Risk<br>Forecasting<br>Trend evaluation<br>Evaluating marketing effectiveness<br>Customer lifetime value estimation |  |
| **Non-Linear Regression** | A statistical method that models the relationship between variables, by fitting a non-linear function on the data. This indicates the relationship between dependent and independent variables through a non-linear function. | Assess the effectiveness of advertisements on different age groups. |  |
| **Logistic Regression** | A binary classifier that uses the logistic sigmoid function to calculate the probability that a data point will belong to a specific class. | Fraud detection<br>Segmentation<br>Direct marketing<br>Credit scoring |  |

Table 3.5: Classification Techniques

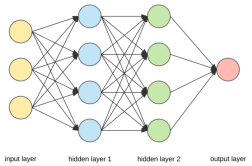
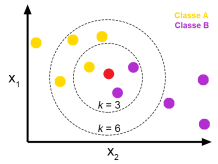
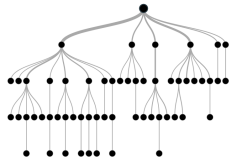
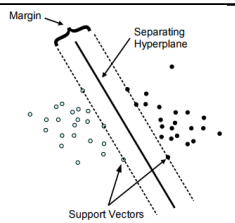
| Classification Techniques | Description | Application | Diagram |
|---|---|---|---|
| **Neural Network** | Linear combinations from the input data are extracted to derive attributes, the target is then modelled as a non-linear function of the derived attributes. | Pattern Recognition<br>Face identification<br>Decision-making<br>Spam filtering<br>Segmentation<br>Direct marketing |  |
| **Kernel Estimator**<br>k-Nearest Neighbour | Classifies an object based on the position of the object relative to its neighbours. The class assigned to the object, is the class that is most common among the object's k nearest neighbours. | Loyalty programmes<br>Outlier detection<br>Concept search<br>Recommendation systems<br>Anomaly detection |  |
| **Decision Trees**<br>Random Forest<br>CART<br>C4.5 | Based on a specific variable's value, data is recursively partitioned into discrete subcategories. It is designed for problems where the learning set uses preclassified variables. | Customer identification<br>Direct marketing<br>One-to-one marketing<br>Loyalty programmes<br>Target customer analysis |  |
| **Support Vector Machines** | A hyperplane in an n-dimensional space is constructed, to distinctly classify the data points. The dimension of the hyperplane depends on the number of features (n). The hyperplane is a decision boundary that helps classify the data points, based on their position relative to the hyperplane. | One-to-one marketing<br>Pattern recognition<br>Bio-informatics<br>Text and hypertext categorisation |  |

Table 3.5 continues on next page

| Classification Techniques | Description | Application | Diagram |
|---|---|---|---|
| **Bayesian** Naive Bayes | Objects are classified based on Bayes's theorem, meaning that probability rules are used to classify the objects. Naive Bayes assumes independence between the features in the dataset. | Pattern recognition Spam filtering Customer lifetime value Direct marketing | $P(A\|B) = \frac{P(A)P(B\|A)}{P(B)}$ |

End of Table 3.5

Table 3.6: Clustering Techniques

| Clustering Techniques | Description | Application |
|---|---|---|
| **Partitioning Methods** | Instances are relocated by moving them from one cluster to another, starting from the initial partition. A preselected number of clusters must be chosen. The points are then relocated iteratively between the chosen number of clusters to attempt optimisation. | Used on small- or medium-sized datasets, as the algorithm creates single sets of clusters. |
| **Hierarchical Clustering** | A hierarchy method is used to group the dataset into clusters. Two approaches can be used for this method a bottom-up (agglomerative) method or a top-down (divisive) method. | Decision support for large scale R&D projects. Analysis of market entry strategies. |
| **Density-based Method** | The technique clusters together sets of points that are closely packed together (based on density conditions). Points that do not have nearby neighbours are identified as outliers. | Outlier detection. Discovering noise in a spatial database. |
| **Grid-based Method** | Clustering operations are performed by partitioning the space into a finite number of cells to form a grid structure. | Used on datasets that are not manageable due to size or dimensionality. |
| **Self-Organising Maps** | High dimensional data is represented in a one- or two-dimensional grid, displaying the topological relationship of the higher dimensional data. A neural network is trained by using a neighbourhood function to preserve the topological properties of the input space. | Segmentation<br>Complaint management<br>Target customer analysis |

After the model has been built, it must be evaluated, to make sure that the initial question is answered and to make sure that the results obtained can be used. This step must be done to ensure that the results obtained from the model are reliable and useful. This is the last step in the data analytics process.

## 3.3 Predictive Analytics

Predictive analytics (PA) was defined by University of Columbia economist Eric Siegel as: "Technology that learns from experience (data) to predict the future behaviour of individuals in order to drive better decisions" (Siegel, 2013). In other words, predictive analytics is the art of building models that can make predictions based on patterns that are extracted from historical data (Kelleher et al., 2015). These predictions are then used to help make decisions.

One of the earliest efforts to predict human behaviour was done by MIT mathematician and philosopher, Norbert Wiener. He tried to predict the behaviour of German pilots in the 1940s. His goal was to shoot them from the sky by taking the input trajectory of the aeroplane from its observed motion and considering the most likely evasive manoeuvres that the pilot would make and then predict where the aeroplane would be in the near future. He managed to predict only one second ahead of the aeroplane's actual motion, which was not enough, as 20 seconds of future trajectory were needed to shoot down an aeroplane (Siegel, 2013). Since then, predictive analytics has evolved so much that it has become mainstream and relevant to each individual as it impacts people when shopping, driving, communicating and even when they are seeing the doctor.

More data, more storage and more computing power have become available and although companies were slow to adopt predictive technologies this has changed in the past few years.

Companies that use predictive analytics usually answer some questions, such as how many supporters will be at a specific football match, so that they can use the answer for a particular reason; in this case so that vendors can know exactly how much stock they will need on the day of the match. Table 3.7 gives examples of how predictive analytics is used in practice. The table is grouped into the following industries: marketing and advertising, healthcare, fraud detection and law enforcement, fault detection, workforce, education and lastly a few examples are given on how predictive analytics is used in people's personal lives.

Table 3.7: How Predictive Analytics are used in industry (Siegel, 2016).

| Organisation | Example of how predictive analytics are used |
|---|---|
| **Marketing and Advertising** | |
| **Harbor Sweets** | Targeted lapsed customers, winning the customers back at a 40% response rate, while contacting specific customers to stay within the marketing budget (DeBevois, 2008). |
| **Elie Tahari** | Predicts fashion trends and the demand for women's fashion line products. PA is also used to predict the right mix of products and sizes for every location (Mitchell, 2011). |
| **Sprint Communications** | Used PA to predict customer churn in the Telecommunications Industry. Identified customers that were three times more likely than average to cancel their subscriptions (Lu, 2002). |
| **Google** | Predicts which web pages will meet the user's standards if shown as search results, improving search functionality (Levy, 2011).<br>Predicts which ads will get bounced (when a user clicks on the ad and then immediately clicks the back button), to warn paying advertisers (Sculley et al., 2009). |
| **Researchers** | Uses PA to predict which Hollywood films will be blockbusters. They also predict which songs will be hits (Asur & Huberman, 2010). |
| **Healthcare** | |
| **Stanford University** | Used PA to derive an innovative method to diagnose breast cancer better than doctors. This method considers a greater number of factors in a tissue sample to predict the cancer (Beck et al., 2011). |
| **Brigham Young University and University of Utah** | Used PA to predict premature births, based on peptide biomarkers found in blood exams from as early as 24 weeks of pregnancy (Esplin et al., 2011). |
| **Health Insurance Company** | Predicts the likelihood that an elderly policy holder will pass away within the next 18 months. This is used to trigger end-of-life counselling (Siegel, 2013). |
| **University of Pittsburgh Medical Center** | Uses PA to predict the risk that a patient will be readmitted within the next 30 days; this helps with the decisions to release a patient (Zaino, 2015). |
| **Riskprediction.org.uk** | Predicts the risk of death during surgery based on personal aspects of the patient as well as the patient's condition (Smith & Tekkis, 2012). |

Table 3.7 continues on next page

| Organisation | Example of how predictive analytics are used |
|---|---|
| **Fraud Detection and Law Enforcement** | |
| **Life insurance companies** | Use PA to predict the age of death of their customers in order to decide whether to approve a policy application and to decide upon the price of the policy. |
| **Israel Institute of Technology** | Researchers at this institute created prediction models that predict 51% of riots with 91% accuracy, contributing towards the prediction of civil unrest (Radinsky & Horvitz, 2013). |
| **Chicago Police Department** | Predicts whether a murder will be solved based on characteristics of the homicide as well as characteristics of the victim (Alderden & Lavery, 2007). |
| **Fault detection** | |
| **Argonne** | Uses PA to model nuclear reactor failures, specifically cracks in cooling pipes (Mohanty et al., 2012). |
| **US National Institute of Standards and Technology** | Uses PA to predict faults in manufacturing equipment, as preventing failures can provide significant savings. Predictive modelling is also used to model maintenance, to achieve cost savings during maintenance (Lechevalier et al., 2014). |
| **Fortune 500 Companies** | Predicts failure of components in electrical equipment such as hard drives and printers to know if this equipment is in need of replacement in order to preload repair dispatch trucks (Abbot, 2012). |
| **Workforce** | |
| **Wikipedia** | Predicts which of their volunteer editors will discontinue their work (Qin et al., 2015).<br>Predicts how many edits a Wikipedia editor will make within the next five months (Zhang, 2011). |
| **Researchers** | Used PA to demonstrate that Facebook profiles can be used to predict job performance. Curiosity, conscientiousness and agreeability are all attributes that can be derived from Facebook, this is then correlated with job performance evaluations (Kluemper et al., 2012). |
| **CareerBuilder** | Used PA to predict which job advertisements to send to their clients based on previous applications, work history and demographic information (CareerBuilder.com, 2012). |

Table 3.7 continues on next page

| Organisation | Example of how predictive analytics are used |
|---|---|
| **Education** | |
| **Oklahoma State University** | Use PA to predict which students are at risk of dropping out of University so that they could be assisted in the hope of retaining them (Delen, 2011). |
| **Hewlett Foundation** | PA was used to develop a system that grades student-written essays. Compared to human graders the system grades just as accurately (Shermis & Hamner, 2013). |
| **University of Phoenix** | PA is used to predict which students are at risk of failing a subject, so that intervention methods could be applied to possibly help them (Barber & Sharkey, 2012). |
| **Personal life** | |
| **Target** | Used PA to detect from their shopping behaviour if their clients are pregnant, in order to send them offers related to the needs of new parents. |
| **Airbnb** | Used PA to predict the acceptance of a booking request by a customer, so that Airbnb matches customers who are more likely to be accepted by hosts with these hosts. Increased booking conversations by nearly 4% (Geron, 2013). |
| **Microsoft** | Used PA to determine, based on GPS data, one's location up to multiple years in advance (Sadilek & Krumm, 2012). |

End of Table 3.7

From these examples it is clear that firstly some sort of model was created from historical data, to predict something that will happen in the future to help with decision-making. To build these models machine learning was used. Most of these models, however, lack the time variable for the predictions, which is very important for this study. Thus further investigation on using machine learning while not ignoring the time domain will be investigated in Section 3.4.

### 3.3.1 Machine Learning and Predictive Analytics

Figure 3.9 shows the predictive model and how machine learning contributes towards this model. The model considers various features of an individual to derive a single predictive score for that individual. This score is then used by organisations to guide decision-making. The predictive model is shown in the top part of Figure 3.9, but before such a model can be used, it should be built and machine learning is used for this predictive model (Kelleher et al., 2015). Machine learning was described earlier in this chapter, but in this context machine

learning refers to building a model from scratch and predictive analytics then uses the power of machine learning to be able to make predictions (Siegel, 2016).

For the purpose of this study Figure 3.9 can be interpreted, as the data is the purchasing history of an individual, the data is analysed and learnt from using machine learning, to be able to develop a predictive model that can predict what an individual will purchase in the future and when this individual will make the purchase.

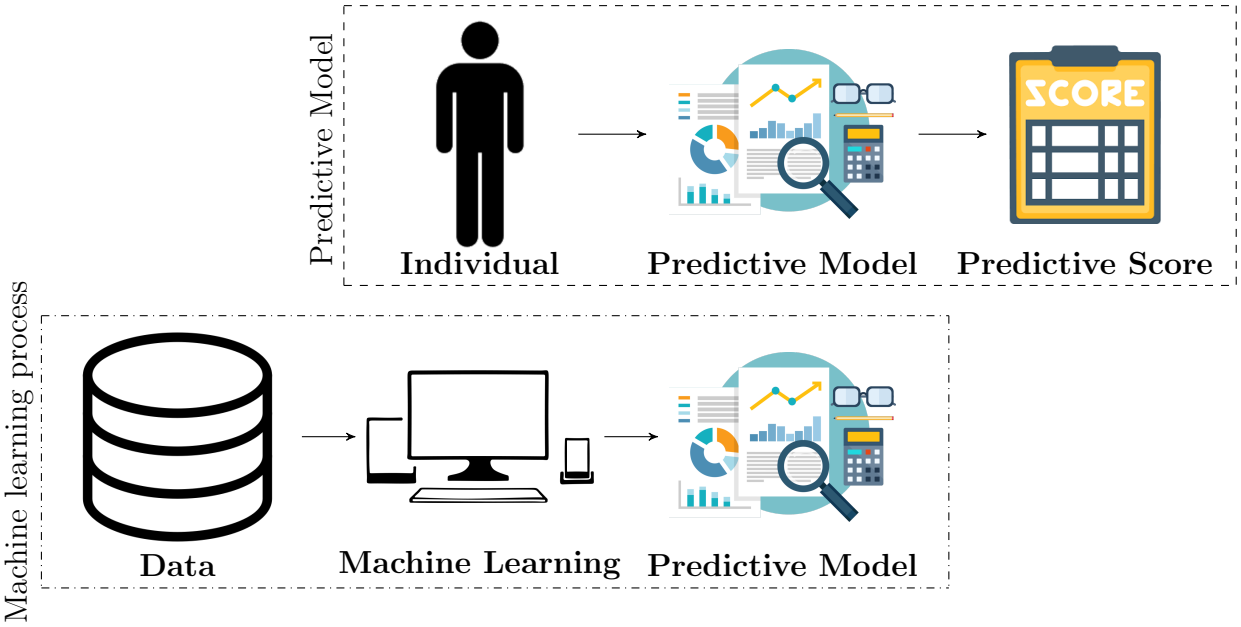



Figure 3.9: Machine Learning and Predictive Analytics (modified from Siegel (2016))

## 3.4 Using Machine Learning to predict future events

In the previous section a few examples were provided on uses of predictive analytics in practice. This section will look at techniques that are used to predict continuous values, taking into account a time variable or sequence of events, which is what the proposed NPD Predictor will aim to do.

### 3.4.1 Linear Regression

Linear regression is a statistical approach to modelling the relationship between a dependent variable and one or more independent variables. For linear regression a line is fitted through the data according to a specific mathematical criterion. This line can, for example, be fitted to minimise the sum of squared distances between the data and the line. This allows an

estimation of the dependent variables (Duda et al., 2012). This technique is widely used mostly for prediction or forecasting.

Linear regression is commonly modelled as $Y_i = f(X_i, B) + e_i$, where $Y_i$ represents the dependent variable, $X_i$ represents the independent variable, B represents an unknown parameter and $e_i$ represents the error term. The goal is then to estimate the function $f(X_i, B)$, that best fits the data. The parameter B is estimated by using various tools provided by regression analysis, such as the least squares method which finds the value of B that minimises the squared error between the line and the data. After this value is estimated the data can then be fitted for prediction.

See Figure 3.10 for modelling $n$ data points. There is a single independent variable and two unknown parameters; the line is represented by the following equation,

$$y_i = B_0 + B_1 x_i + e_i,$$

where $i = 1, \ldots, n$. So, given a random sample from the population, the estimated population parameters are obtained, and the sample linear regression model can be represented as:

$$\hat{y}_i = \hat{B}_0 + \hat{B}_i x_i.$$

The error $e_i$ is the difference between the dependent value and the predicted dependent value. Minimisation of an error function is used to yield the parameter estimators. The topic of linear regression is discussed in full by Montgomery et al. (2013).

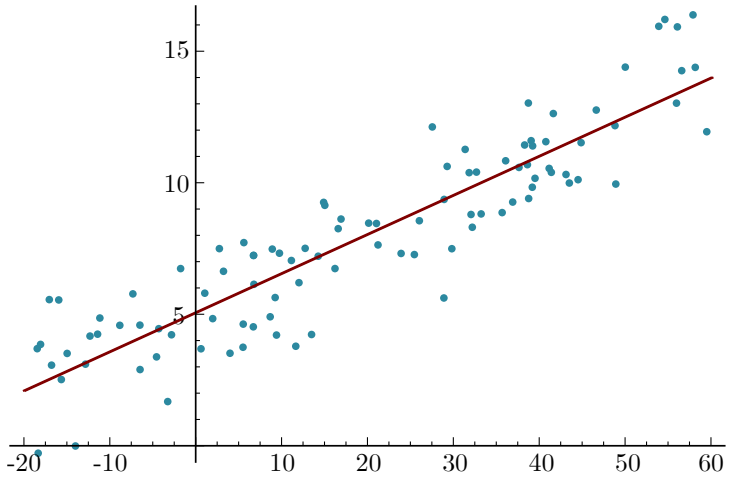



Figure 3.10: Linear Regression

### 3.4.2 Artificial Neural Networks

An Artificial Neural Network (ANN) is a set of connected neurons that are organised in layers. As seen in Figure 3.11 there are three types of layers:

i Input layer - this layer brings the data into the system to be further processed by the subsequent layers.

ii Hidden layer(s) - this layer(s) lies between the input and output layers, where artificial neurons take a set of inputs that are weighted and produces an output through an activation function.

iii Output layer - this layer is the last layer of neurons (there could be one or more neurons), that produces a given output for the system.

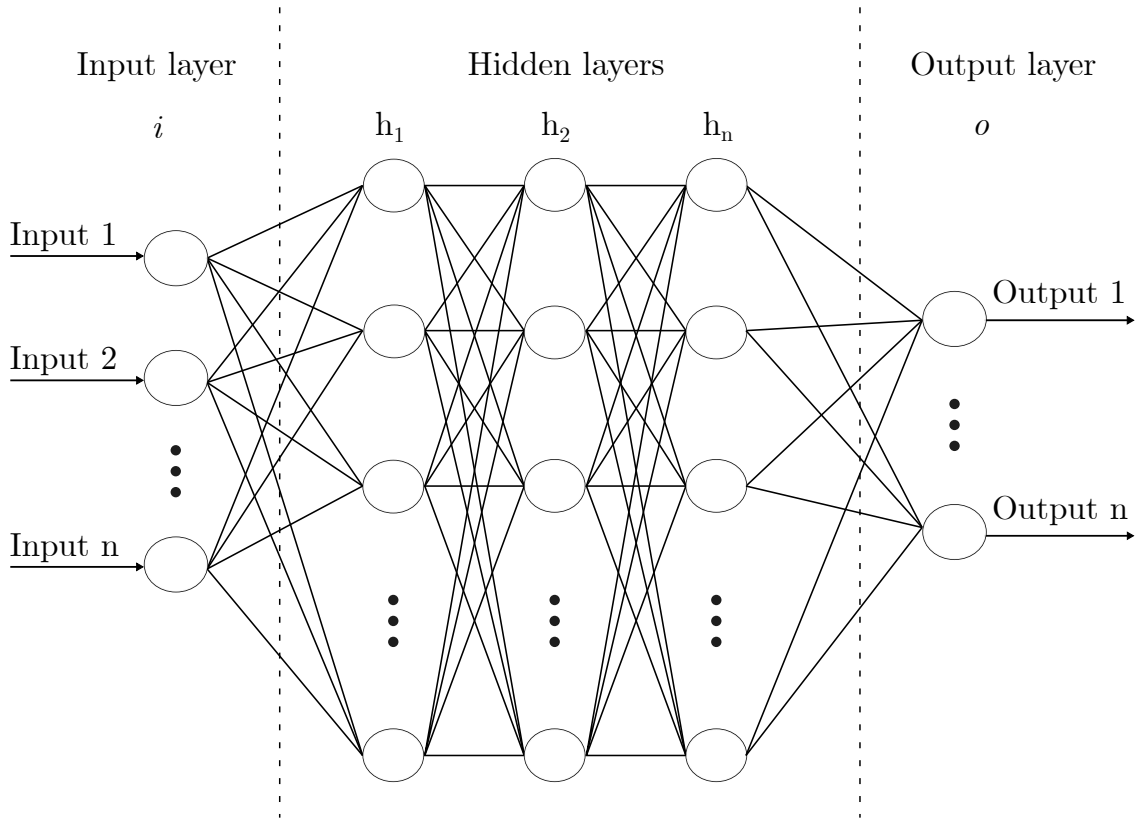



Figure 3.11: Neural Network Architecture

In Section 3.2.3 neural networks are described as a supervised learning classification algorithm. Neural networks are mostly used for classification such as number predictions (from 0 - 9), but neural networks can also be used for regression problems. This will be explained through a very simple neural network model with one input neuron, one hidden neuron and

one output neuron. See Figure 3.12. The input neuron, consisting of several input parameters multiplied by their weights, are run through a sigmoid function and a unit step function. This closely represents the logistic regression function, with an error term. To find the coefficients that fit the data backpropogation is performed by using a gradient method. This will be explained later in this section.

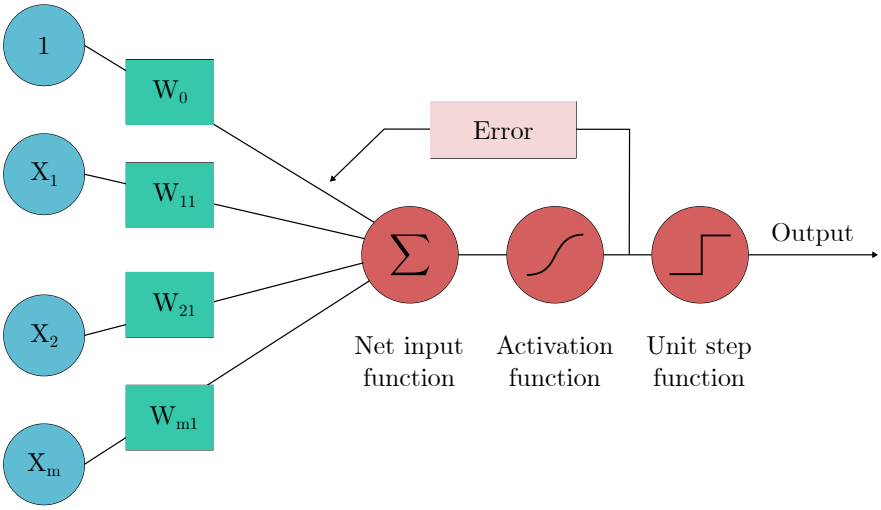



Figure 3.12: Neural Network Regression model

### 3.4.2.1 How Artificial Neural Networks work

The connections between the neurons in different layers of the neural network are represented by weights, which form the basis of the decision-making in a neural network. As seen in Figure 3.11, each neuron is connected to all the neurons in the next layer; this is called a fully-connected neural network. A weight is a connection between the neurons that carries a value. The higher the value, the larger the weight that it carries and more importance is attached to the neuron on the input side of the weight. Weights are viewed in a matrix format illustrated by Figure 3.13. As seen in the figure, the input layer has three neurons, and the next layer, which is the first hidden layer, has four neurons. This can be represented by a weights matrix $W$, also expanded in the figure, with dimensions of three rows and four columns. If the ANN has more layers, each layer will have a weight matrix. In general if a layer $h$ has $n$ neurons and the next layer $h+1$ has $m$ neurons, the weight matrix will be an $n$ by $m$ matrix.

A bias, as seen in Figure 3.12 is added to cater for unforeseen and non-observed factors. The bias also has a weight associated with it. The weight matrix for the bias has one column. Thus, if a layer has $n$ connections to the next layer, the bias weight matrix will be a one by $n$ matrix.

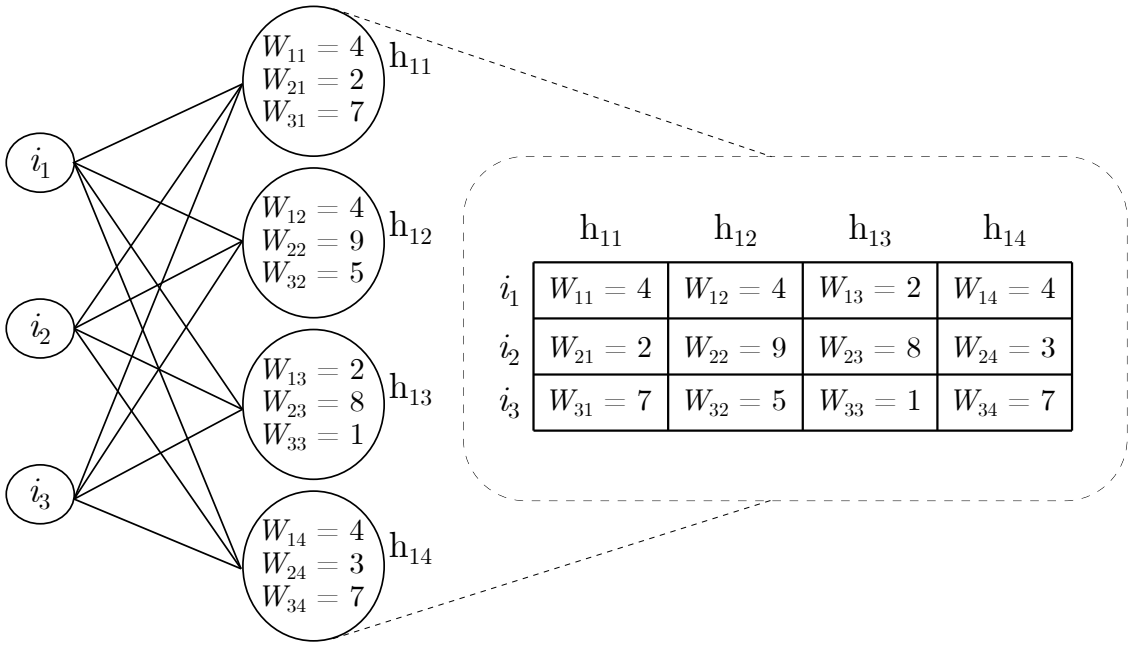



Figure 3.13: Neural Network weight matrix (Babs, 2018)

To calculate the outputs of a layer by using matrix manipulation will be explained by the following steps (from the input layer to the first hidden layer)(Babs, 2018):

**Step 1:** Create an $n$ by 1 matrix, $I$, for the input layer.

**Step 2:** Create a weight matrix, $W$ from the input layer as described, with $n$ by $m$ dimensions.

**Step 3:** Create a bias weight matrix, $b$, with $m$ by 1 dimensions (a vector).

**Step 4:** Transpose the weight matrix, to create an $m$ by $n$ matrix.

**Step 5:** Calculate the dot product of the transposed weights and the input matrix,

$$W^\intercal \cdot I,$$

this will result in an $m$ by 1 matrix.

**Step 6:** Add the matrix obtained in Step 5 to the bias weight matrix, $b$.

The matrix that is obtained from Step 6, corresponds to the values of the neurons in the next layer of the neural network. These values must then be run through the chosen activation function. Different activation functions can be used, which will determine the output of the neural network. If a linear activation function, summarised in Table 3.8, is used, the neural network turns into a linear regression model. Some activation functions will be discussed in the next subsection.

#### 3.4.2.2 Activation Functions

The output of a neural network is determined by mathematical equations called activation functions. An activation function is attached to all the neurons in the network, and the function determines if the neuron should be activated or not, based on the relevance of the neuron's input for the model's prediction. Another feature of the activation function is that it normalises the output of the neurons to a range between 0 and 1 or -1 and 1 depending on the activation function chosen.

Table 3.8 summarises some activation functions used for neural networks. In this table, the equation, a graphical representation, advantages and disadvantages of each activation function are summarised.

#### 3.4.2.3 Adjusting the weights and biases of the ANN

A forward pass has been completed when all the steps to calculate the next layer in the neural network have been performed, and an output is obtained. After each forward pass, a backward pass through the network to adjust the model's weights and biases must be completed to train the model. As a neural network is a supervised learning technique, the desired output value is known. An error metric called a loss function is used to give an indication of how much precision is lost if the real output is replaced by the output generated by the trained neural network.

The backpropagation algorithm is used to compute the gradient of the loss function with respect to the weights of the network for a single input-output example. The backpropagation algorithm efficiency makes it feasible to use gradient methods for training multilayer networks. These gradient methods are also called optimisers and different optimisers will be discussed later in this section (Doshi, 2019; Hansen, 2019). Advantages and disadvantages of the discussed optimisers are summarised in Table 3.9.

The backpropagation algorithm computes the gradient of the loss function with respect to each weight by using the chain rule. It computes the gradient one layer at a time and iterates backwards from the last layer to avoid redundant calculations of the intermediate terms in the chain rule. This algorithm is an example of dynamic programming.

Table 3.8: Activation Functions (Sharma & Athaiya, 2020)

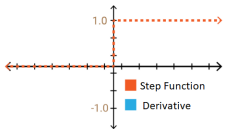
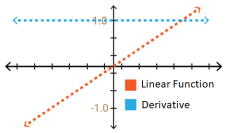
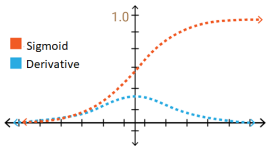
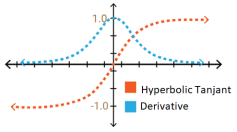
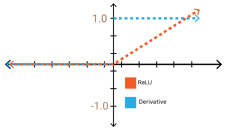
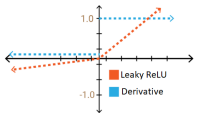
| Function | Equation | Diagram | Advantages | Disadvantages |
|---|---|---|---|---|
| **Binary Step** | $f(x) = \begin{cases} 0 & x < 0 \\ 1 & x \geq 0 \end{cases}$<br>Range: $\{0, 1\}$ |  | Good with binary classification | Gradient of function is zero, thus; not possible to use backpropagation.<br>Cannot be used for multi-class classification. |
| **Linear** | $f(x) = x$<br>Range: $(-\infty, \infty)$ |  | Allows multiple outputs | Not possible to use back-propagation.<br>Last layer will be a linear function of the first. |
| **Sigmoid/ Logistic** | $f(x) = \sigma(x) = \frac{1}{1+e^{-x}}$<br>Range:$(0, 1)$ |  | Bound output values<br>Output of each neuron is normalised<br>Has a smooth gradient | Vanishing gradient problems<br>Output not zero centred<br>Computationally expensive |
| **TanH (Hyperbolic Tangent)** | $f(x) = \tanh(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$<br>Range:$(-1, 1)$ |  | Zero centered<br>Bound output values<br>Normalised outputs<br>Smooth gradient | Vanishing gradient problems<br>Computationally expensive |

Table 3.8 continues on next page

| Function | Equation | Diagram | Advantages | Disadvantages |
|---|---|---|---|---|
| **ReLU (Rectified Linear Unit)** | $f(x) = \begin{cases} 0 & x < 0 \\ x & x \geq 0 \end{cases}$<br>Range: $[0, \infty)$ |  | Computational efficient<br>Non-linear<br>Allows backpropagation | Dying ReLU problem, when input approach 0, or negative inputs, the gradient of the function becomes 0, thus network cannot perform backpropagation and cannot learn. |
| **Leaky (ReLU** | $f(x) = \begin{cases} 0.01 & x < 0 \\ x & x \geq 0 \end{cases}$<br>Range: $(-\infty, \infty)$ |  | Prevents dying ReLU problem, as it has a small positive slope in the negative area.<br>Allows backpropagation even for negative values | Does not provide consistent predictions for negative inputs. |

End of Table 3.8

Table 3.9: Advantages and Disadvantages of the Optimisers (Doshi, 2019)

| Advantages | Disadvantages |
|---|---|
| **Gradient Descent** | |
| Easy to implement<br>Easy to understand<br>Easy to compute | Requires large memory to calculate the gradient on entire dataset.<br>Can get trapped at local minima.<br>Can take very long to execute. |
| **Stochastic Gradient Descent** | |
| Converge in less time<br>Requires less memory<br>May get new minima | High variance in the model parameters.<br>Can overshoot even after arriving at the global minima. |
| **Momentum** | |
| Reduce oscillations and high variance of the parameters<br>Converge faster than gradient descent | A hyper-parameter which must be selected manually are added to the equation. |
| **Adam** | |
| Very fast and converge rapidly<br>Rectifies vanishing learning rate | Computationally costly. |

### Gradient Descent

Gradient descent is a very basic optimisation algorithm. It is a first-order optimisation algorithm, which means it is dependent on the first-order derivative of the loss function. It calculates which way the weights should be adjusted for the loss function to reach a minimum. By using the backpropagation algorithm, the loss is transferred from one layer to another, and the weights are modified depending on the losses, to minimise the loss function. The equation used for gradient descent is

$$\theta = \theta - \alpha \cdot \nabla_\theta J(\theta),$$

where $\theta$ is a parameter, for example, the activation function, weights and biases, $\alpha$ is the learning rate, $\nabla$ is the gradient taken of $J$, and $J$ is the loss function.

### Stochastic Gradient Descent

Stochastic Gradient Descent (SGD), is a variant of Gradient Descent. This algorithm updates the model's parameters more frequently. The model's parameters are updated after the computation of loss on each training example. This means, for example, if the dataset consisted of 100 training examples, for SGD the model parameters would be updated 100 times in one cycle of the dataset, and not only once like Gradient Descent. The equation used to update the parameters of the neural network using SGD is

$$\theta = \theta - \alpha \cdot \nabla_\theta J(\theta;\, x,\, y),$$

thus each parameter $\theta$ is updated by subtracting the learning rate $\alpha$ times the ratio of change $\nabla J(\theta)$, from the original parameter $\theta$, where $J(\theta;\, x,\, y)$ means that the parameter $\theta$ is input along with a training example, $x$, and a label, $y$.

### Momentum

The Momentum optimiser was invented to reduce the high variance in SGD, and it softens the convergence. This optimiser reduces the fluctuation to the irrelevant direction and accelerates the convergence toward the relevant one, which helps to get a local minimum faster. For this, a temporal element is added to the equation for updating the neural network parameters. Adding this time element increases the momentum to descend more quickly. The function for momentum is almost the same as that of SGD, but an additional term is added:

$$\theta = \theta - \alpha \nabla J(\theta) + \gamma v_t,$$

where $\gamma$ is a constant term called the momentum and the previous update, $v_t$ is multiplied by the momentum constant and $v_t$ is calculated by

$$v_t = \alpha \nabla J(\theta_{t-1}) + v_{t-1},$$

which essentially stores the calculation of the gradients to be used in all the next updates to a parameter $\theta$.

### Adam

Adaptive Moment Estimation (Adam) (Kingma & Ba, 2014), uses Adaptive Learning Rates and Momentum to converge faster. The intuition for adaptive learning rates is that it starts off by taking big steps toward the minimum and finishes with small steps. This means that initially fast progress is made and as the learning rate decays smaller steps are taken so as not

to overshoot the local minimum. To do this, Adam stores an exponentially decaying average of the past gradients and squared gradients. To update the parameter using Adam, the following equation is used:

$$\theta_{t+1} = \theta_t - \frac{\alpha}{\sqrt{\hat{v}_t} + \epsilon}\hat{m}_t,$$

where,

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t},$$
$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t},$$

and

$$m_t = (1 - \beta_1)g_t + \beta_1 m_{t-1},$$
$$v_t = (1 - \beta_2)g_t^2 + \beta_2 v_{t-1},$$

and

$$g = \nabla J(\theta_{t,i}).$$

$\epsilon$ is just a small term to prevent dividing by zero, and $\beta_1$ and $\beta_2$ are two exponential decay rates. These two terms are close to the $\gamma$ term in Momentum optimisation, but instead of having one term, this optimiser has two terms. The values of $\beta_1$ and $\beta_2$ are 0.9 and 0.999 respectively, thus with the time exponent, say $t = 5$, then $\beta_1^{t=5} = 0.9^5 = 0.59049$ . This method is computationally costly but converges rapidly.

This concludes the discussion on optimisers used in the backprapagation algorithm.

#### 3.4.2.4 Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are neural networks that are good at modelling sequential data (Wen et al., 2013). This technique is used for sequential signals such as speech recognition, stock prediction and language translation. It is considered in this thesis because the data of customer purchases follow a sequence over time.

As with a feed-forward neural network, an RNN also has an input layer, a hidden layer and an output layer. However, an extra loop is added that can pass information forward. Figure 3.14, shows a rolled-out version of an RNN and will be used to explain how it works. It is possible to process a sequence of vectors $\mathbf{x}$ by applying a recurrent formula at every time step.

At time step $t$ a pattern is followed where the model reads an input, updates the hidden state and then produces an output, the function form of the recurrence relation can be represented by

$$h_t = f_W(h_{t-1}, x_t),$$

where $h_t$ is the new hidden state, $f_W$ is some function with a parameter $W$, which corresponds to weight, $h_{t-1}$ is the old state and $x_t$ is the input vector at some time step. The updated hidden state $h_t$ will then be passed into the same function when the next input $x_{t+1}$ is read in. The same function and the same set of parameters are used at every time step, meaning that the same weights matrix is used at every time step.

The simplest functional form can be represented by

$$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t),$$

$$y_t = (W_{hy}h_t),$$

where $W_{hh}$ is the weight matrix of the previous hidden state and $W_{xh}$ is the weight matrix for the input, $W_{hy}$ is the weight corresponding to the output layer. The tanh function can be replaced by other activation functions that allow backpropagation as explained through Table 3.8, to introduce non-linearity to the system.

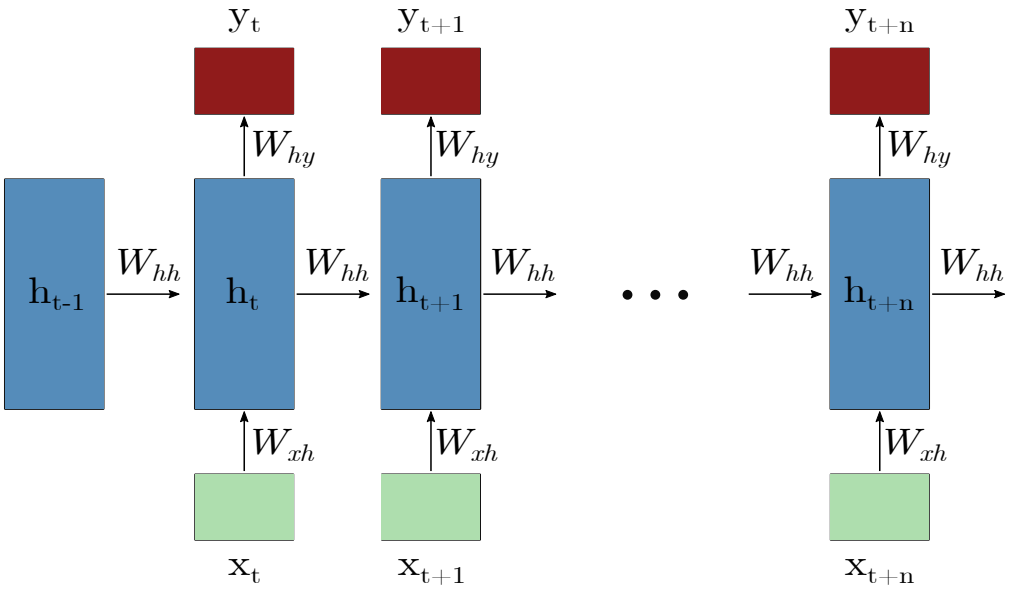


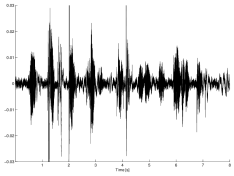

Figure 3.14: Recurrent Neural Network Architecture (modified from Gupta (2017))

The loss can be calculated by taking the true output value and comparing it to the output value at each time step $y_t$, with some loss function. Thus the loss for the entire training step will be the sum of all the individual losses.

To train the model, the same backpropagation algorithm is used for the RNN as for the ANN, except that RNN has multiple time steps. Each time step computes a local gradient on the weights, which are summed to give the final gradient. This gradient is computed to minimise the loss function.

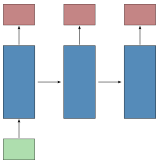
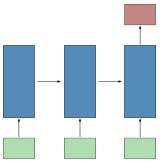
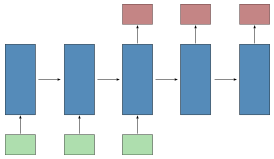
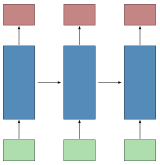
Different applications of sequential data analysis can be seen in Table 3.10. The table gives an example of the input and output of each application. As seen from the examples, the samples cannot be isolated, and the sequence of events is important.

Table 3.10: Examples of sequential data (Lipton et al., 2015)

| **Application** | **Input** | **Output** |
|---|---|---|
| Sentiment Analysis | "There is nothing to like in this movie." |  |
| Speech Recognition |  | "The quick brown fox jumped over the lazy dog." |
| Machine Translation | Jabulela usuku lwakho | Have a great day. |
| Image activity recognition |  | Running |

The RNN explained is an example of a many-to-many RNN as it takes multiple inputs and produces multiple outputs. There are different types of RNN models which model different input data. The types of RNN models are summarised in Table 3.11. The table gives the RNN type, an illustration of the model as well as the application where the specific model is used. These different types of RNN models give more flexibility in the type of data that models can process.

Table 3.11: Examples of different types of RNN models (Lipton et al., 2015)

| RNN type | Illustration | Application |
|---|---|---|
| one-to-many |  | Image captioning |
| many-to-one |  | Sentiment classification |
| many-to-many (output length not necessarily equal to input length) |  | Machine translation |
| many-to-many (output same length as input length) |  | Video classification on frame level |

### 3.4.3 Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting was developed on the framework of gradient boosting (Chen & Guestrin, 2016). Boosting, also known as the "sequential ensemble method", creates a sequence of models in which the models attempt to correct the mistakes that were made in the previous model in the sequence. The first model is built based on the training data, then the second model attempts to improve on the first model, after which the third model attempts to improve on the second model, and so on.

Figure 3.15 shows that the original data is passed to the first classifier. The yellow area represents the predicted blue hyphen and the blue area the predicted red cross. Thus, in this first attempt, the classifier misclassified the three circled instances. After this, the weights of these incorrectly classified instances are adjusted and sent to the second classifier. The second classifier then correctly predicts the three instances incorrectly predicted by the first classifier, but incorrectly predicts three different instances. This process is then repeated until the specified number of iterations is reached, or a certain threshold is reached by the classifier.

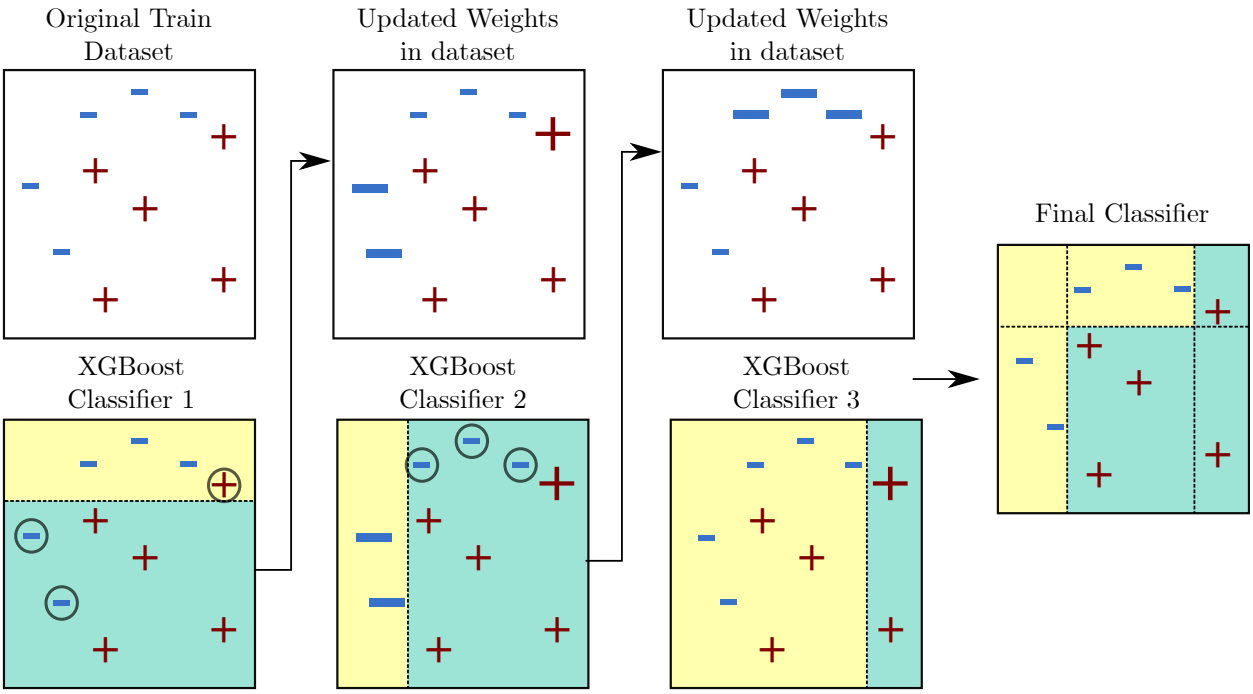



Figure 3.15: Example of Boosting (Kasturi, 2019)

Gradient boosting uses an approach where a new model is created that predicts residuals (errors) of the prior models, which when added together make the final prediction. Suppose there are $K$ trees (weak models), then the model can be represented by

$$\sum_{k=1}^{K} f_k,$$

where $f_k$ is the prediction from the decision three, and the model represents a collection of decision trees. The prediction is then made based on all the decision trees by

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i),$$

where $x_i$ is the feature vector for the $i$-th data point. Thus the prediction for the $t$-th step can be defined as

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i).$$

A loss function must be minimised to train the model. Say, for example, a Root Mean Squared Error loss function is used, the function will have the form

$$L = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2,$$

where $y_i$ is the real value (label) which is known from the training dataset and $\hat{y}_i$, is the predicted value. This loss function can be used for regression. Gradient Boosting uses gradient descent to minimise the loss function (the difference between the predicted value and the actual value). The objective function of the XGBoost model can be calculated by adding the loss function with a regularisation component,

$$Obj = L + \Omega,$$

where $\Omega$ is the regularisation term. This means that the loss function has predictive power and the regularisation term controls the simplicity and the overfitting of the model.

XGBoost can be used for both regression and classification problems with a loss function such as Root Mean Squared Error as explained above for regression versus a loss function such as log loss for binary classification. One major advantage of XGBoost is that it parallelises the tree-building components of the boosting algorithms; it is thus very fast to train and test.

## 3.5 Conclusion: Chapter 3

In this chapter several data analytics processes and the different steps within them were compared. Then each of the steps in the data analytics processes were discussed including preprocessing and creating a data analytics model. A brief history of machine learning was given followed by a discussion of different machine learning techniques. This chapter also gave examples of predictive analytics and how it is used in practice. This chapter ended with a study of how machine learning is used to predict continuous outputs, which is what the proposed NPD Predictor will aim to do. The next chapter will discuss requirements for the NPD Predictor along with requirements for the dataset that is needed to develop the NPD Predictor, and give insights into the dataset.

# Chapter 4

# NPD requirements specification, dataset selection and data understanding

In Chapter 3, data analytics processes were discussed. These processes will be followed to design and develop the Next Purchase Date (NPD) Predictor. The first phase in the CRISP-DM cycle is a business understanding step. The business case was already established in Section 1.1 and this was translated to a data mining problem though the research statement in Section 1.2, stating that a NPD Predictor must be developed for individuals for products in the fast-moving consumer goods sector, using machine learning. Before the rest of the data-mining process can be followed an appropriate dataset must be identified, and requirements for the NPD Predictor established. In this chapter, specifications for the dataset will be established, and a dataset will be chosen for the development of the NPD Predictor.

## 4.1 Requirements for the NPD Predictor

Requirements for the NPD Predictor must be determined based on the use of the NPD Predictor. The NPD Predictor has the following requirements:

- Must make NPD Prediction for a user-product pair,
- Must make predictions in days to next purchase,
- Must be able to use the NPD Prediction of a user-product pair to generate individualised advertisements.

These requirements will be used with further analysis of the research statement from Chapter 1 to choose an appropriate dataset to develop the NPD Predictor.

## 4.2 Requirements of a dataset needed to develop the NPD Predictor

Taking the research statement from Chapter 1 and highlighting the primary variables, it is possible to arrive at four main dependencies when designing features for the NPD Predictor.

The research statement can be translated to, predicting when a specific customer will purchase a specific product, in the fast-moving consumer goods sector. These indicate:

**when:** the predictor is time-dependent

**specific customer:** the predictor is user specific

**specific item:** the predictor is product specific

**fast-moving consumer goods:** the products purchased by a customer must be from this sector.

Looking at the research statement as a whole it should also be noted that user-product pairs must be identifiable as the prediction will not be made only for a specific user or a specific product but for a specific user-product pair. The dataset chosen to develop the NPD Predictor must at least consist of transactions with the attributes summarised in Table 4.1. The table lists the attributes and gives a reason why these attributes are needed.

Table 4.1: Attributes needed to develop the NPD predictor

| **Attribute** | **Reason** |
|---|---|
| User_ID | The Prediction must be made for a specific user; thus, users must be distinguishable. |
| Product_ID | The prediction must be made for a specific product, thus the products that a user purchases must be identifiable. |
| Date identifier | The NPD Predictor must predict when a user will purchase an item; thus, the time of purchase must be specified. The date must either be specified or be possible to calculate. Most importantly, the days between a purchase for a user-product pair must either be available or must be calculated. |
| User-product pair identifier | It must be possible to link a user and a product as the NPD Predictor should predict the NPD for a specific user-product pair. |
| Fast-moving Consumer Goods | The predictor must predict products in the fast-moving consumer goods sector; thus, the dataset must have products from this sector. |

There are also a few attributes that are not absolutely necessary for the development of the NPD, but which it will be nice to have. These are summarised in Table 4.2.

Table 4.2: Attributes that would be nice to have when developing the NPD Predictor

| Attribute | Reason |
|---|---|
| The amount spent by a user | The amount spent by a user could help in identifying possible clusters of important customers to target with marketing strategies by using the NPD predictor. |
| The quantity of a product purchased | The quantity of products purchased per transaction could influence the NPD for the product. |
| The time that a transaction was made | This could further refine the time that a user makes purchases. |

Furthermore, things to take into account when choosing a dataset to develop the NPD Predictor includes:

- number of users,
- number of transactions made, and
- number of products.

These specifications should again not be the deciding factors in choosing a dataset but should be considered if datasets perform the same on other criteria.

## 4.3 Comparing online datasets and dataset selection

A dataset that satisfies the requirements mentioned above must be chosen so that the NPD Predictor can be developed. In Table 4.3, three online datasets are compared to indicate which dataset has attributes that will best suit this research problem. The datasets compared are:

i Instacart Online Grocery Shopping Dataset of 2017,

ii Online Retail II Data Set of 2017, and

iii Brazilian E-Commerce Public Dataset by Olist of 2018.

A tick in the appropriate column of the table indicates whether the attribute is available in the dataset or if the attribute can be calculated from other information available in the dataset. The symbol "–" indicates that the attribute is not available in the dataset.

Table 4.3: Comparison of different online consumer goods datasets

| **Dataset** | **Instacart** | | **Online Retail II** | | **Olist** | |
|---|---|---|---|---|---|---|
| **Attributes** | **Available** | **Calculate** | **Available** | **Calculate** | **Available** | **Calculate** |
| **Products sold** | FMCG | | Giftware | | Department store | |
| **User ID** | ✓ | | ✓ | | ✓ | |
| **Product ID** | ✓ | | ✓ | | ✓ | |
| **Date identifier** | | ✓ | ✓ | | ✓ | |
| **Time of purchase** | ✓ | | ✓ | | ✓ | |
| **Quantity purchased** | – | – | ✓ | | – | – |
| **Monetary value of purchase** | – | – | ✓ | | – | – |
| **Number of records** | > 3 million | | > 1 million | | > 90 000 | |
| **Number of users** | > 200 000 | | | | > 90 000 | |
| **Country** | USA | | UK | | Brazil | |
| **Source** | Instacart, 2017 | | Dua & Graff, 2017 | | Olist, 2018 | |

## 4.4 Dataset Selection

From Table 4.3, it can be seen that the Instacart dataset is the most suitable for this project. This dataset conforms to the specification of user_id, product_id, which can be associated with each other. The time between purchases is available or can be calculated, and the products purchased by users in this dataset are in the fast-moving consumer goods spectrum. The other datasets are not from the fast-moving consumer goods sector and are thus not selected. The Instacart dataset will be used to develop the NPD Predictor and can now be explored. This dataset will also be referred to as the online FMCG dataset.

## 4.5 Data Understanding

This subsection will give a summary of the dataset chosen in the previous section, which will serve as the Explore step in the SEMMA process discussed in Section 3.1, and the Data Understanding step in the CRISP-DM process discussed in the same chapter. Data preparation will be done after the summary is given.

The dataset contains over 3 million grocery orders from more than 200 000 users of the online store, Instacart, each customer having between four and 100 orders. The data was anonymised by the provider. This chapter will look into the dataset and give a broad summary to understand the relational structure of the data, along with a few insights from the tables.

### 4.5.1 Relational structure of the dataset

The dataset has a relational structure that can be seen in Figure 4.1 and consists of the following five tables:

- orders
- departments
- products
- aisle
- order_products

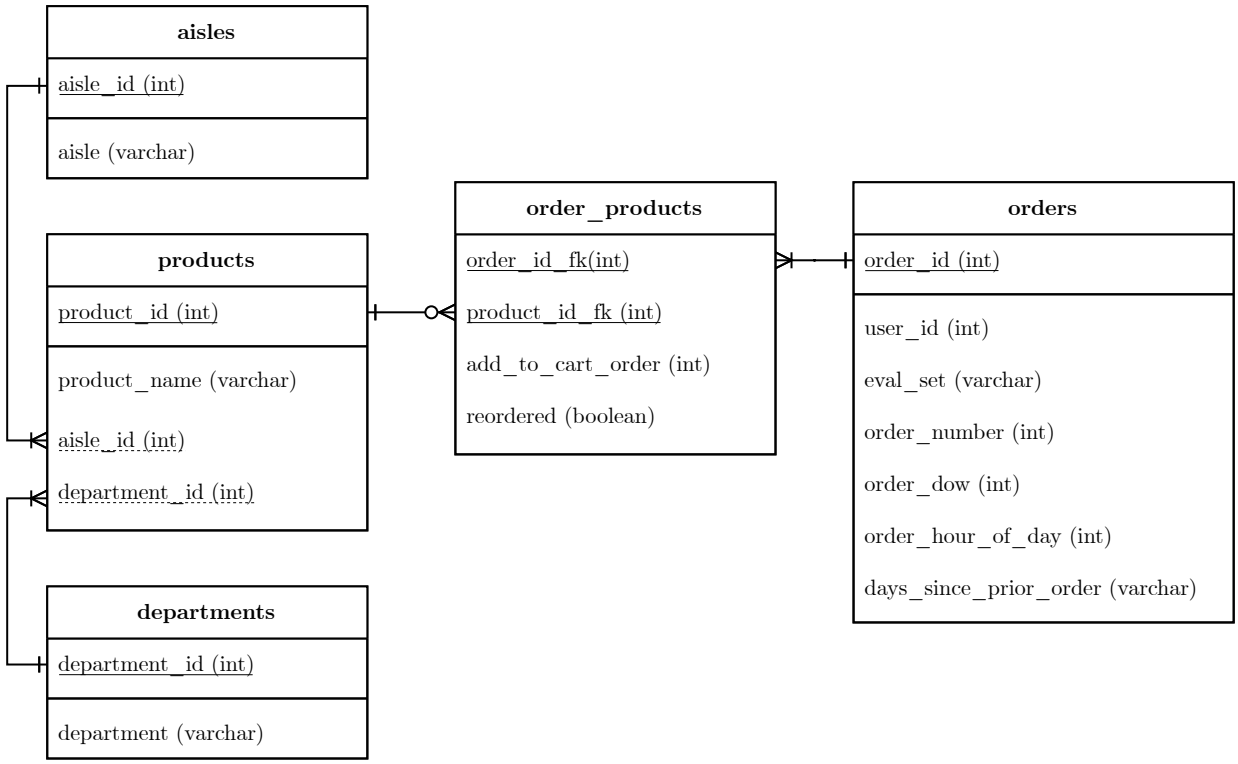



Figure 4.1: Relational data structure of the online FMCG dataset

Each of these tables has features as seen in Figure 4.1. Tables 4.4 – 4.8, show entries from the tables in Figure 4.1 and each table is briefly discussed.

As an example of the data, the first five entries of the `products` table are presented in Table 4.4. There are 49 688 unique product_ids in the `products` table. This means that the dataset consists of 49 688 products each having their own product_name and associated with an aisle_id and a department_id. Of these products, 10 products were never purchased, and

7 165 products were purchased less than 10 times. These products will not be able to give a sufficient trend in purchase behaviour and this must be taken into consideration when working on a solution to predict the Next Purchase Date.

Table 4.4: `Products` table first five entries

| product_id | product_name | aisle_id | department_id |
|---|---|---|---|
| 1 | Chocolate Sandwich Cookies | 61 | 19 |
| 2 | All-Seasons Salt | 104 | 13 |
| 3 | Robust Golden Unsweetened Oolong Tea | 94 | 7 |
| 4 | Smart Ones Classic Favorites Mini Rigatoni With Vodka Cream Sauce | 38 | 1 |
| 5 | Green Chile Anytime Sauce | 5 | 13 |

The `departments` lookup table can be seen in Table 4.5. This table displays all the departments in the dataset. There are 21 departments, which include a produce department, household department and even a babies' department.

Table 4.5: All departments in lookup table `departments`

| department_id | department | department_id | department |
|---|---|---|---|
| 1 | frozen | 12 | meat seafood |
| 2 | other | 13 | pantry |
| 3 | bakery | 14 | breakfast |
| 4 | produce | 15 | canned goods |
| 5 | alcohol | 16 | dairy eggs |
| 6 | international | 17 | household |
| 7 | beverages | 18 | babies |
| 8 | pets | 19 | snacks |
| 9 | dry goods pasta | 20 | deli |
| 10 | bulk | 21 | missing |
| 11 | personal care | | |

The first five entries of the `aisles` lookup table can be seen in Table 4.6. There are 134 unique aisles in total.

Table 4.6: Sample of the `aisles` table

| aisle_id | aisle |
|---|---|
| 1 | prepared soups salads |
| 2 | specialty cheeses |
| 3 | energy granola bars |
| 4 | instant foods |
| 5 | marinades meat preparation |

Table 4.7 contains all the data about the customer's order such as the order_id, the user associated with the order, the order number, the day of the week that the order was placed (order_dow) along with the hour of the day (order_hour_of_day) that the order was placed. In the last column, the number of days since the previous order was made, is given. Each first entry for a customer displays NaN (not a number), as the customer has never made a purchase before. It can be seen that User 1 has multiple orders with different order_ids.

Table 4.7: Sample of the `orders` table

| order_id | user_id | order_number | order_dow | order_hour_of_day | days_since_prior_order |
|---|---|---|---|---|---|
| 2539329 | 1 | 1 | 2 | 8 | NaN |
| 2398795 | 1 | 2 | 3 | 7 | 15 |
| 473747 | 1 | 3 | 3 | 12 | 21 |
| 2254736 | 1 | 4 | 4 | 7 | 29 |
| 431534 | 1 | 5 | 4 | 15 | 28 |

The last table to be discussed from Figure 4.1, is the `order_products` table. This table contains an order_id associated with the orders table, a product_id associated with a product in the products table, an add_to_cart_order which indicates the order in which the item was added to the cart and a reordered column, which indicates if the product was included in the previous order that the customer made. Table 4.8 displays the first few rows of the table. Through this table and the `orders` table, the user-product pairs can be associated with each other as the product_id is linked to an order_id, which is in turn linked to a user_id in the `orders` table.

Table 4.8: Sample of the `order_products` table

| order_id | product_id | add_to_cart_order | reordered |
|---|---|---|---|
| 2 | 33120 | 1 | 1 |
| 2 | 28985 | 2 | 1 |
| 2 | 9327 | 3 | 0 |
| 2 | 45918 | 4 | 1 |
| 2 | 30035 | 5 | 0 |
| 2 | 17794 | 6 | 1 |
| 2 | 40141 | 7 | 1 |
| 2 | 1819 | 8 | 1 |
| 2 | 43668 | 9 | 0 |
| 3 | 33754 | 1 | 1 |

All the tables were merged to construct a user view, which can be seen in Tables 4.9 – 4.11. It is displayed over multiple tables as the view was too large to display in one table. This view is grouped by order_id and user_id.

Table 4.9: Online FMCG data merged table

| order_id | user_id | order_number | order_dow | order_hour_of_day | days_since_prior_order |
|---|---|---|---|---|---|
| 2539329 | 1 | 1 | 2 | 8 | NaN |
| 2539329 | 1 | 1 | 2 | 8 | NaN |
| 2539329 | 1 | 1 | 2 | 8 | NaN |
| 2539329 | 1 | 1 | 2 | 8 | NaN |
| 2539329 | 1 | 1 | 2 | 8 | NaN |
| 2398795 | 1 | 2 | 3 | 7 | 15 |
| 2398795 | 1 | 2 | 3 | 7 | 15 |
| 2398795 | 1 | 2 | 3 | 7 | 15 |
| 2398795 | 1 | 2 | 3 | 7 | 15 |
| 2398795 | 1 | 2 | 3 | 7 | 15 |

Table 4.10: Online FMCG data merged table continued

| order_id | department_id | department | product_id | product_name |
|---|---|---|---|---|
| 2539329 | 7 | beverages | 196 | Soda |
| 2539329 | 16 | dairy eggs | 14084 | Organic Unsweetened Vanilla Almond Milk |
| 2539329 | 17 | household | 26405 | XL Pick-A-Size Paper Towel Rolls |
| 2539329 | 19 | snacks | 12427 | Original Beef Jerky |
| 2539329 | 19 | snacks | 26088 | Aged White Cheddar Popcorn |
| 2398795 | 4 | produce | 13176 | Bag of Organic Bananas |
| 2398795 | 7 | beverages | 196 | Soda |
| 2398795 | 14 | breakfast | 13032 | Cinnamon Toast Crunch |
| 2398795 | 19 | snacks | 10258 | Pistachios |
| 2398795 | 19 | snacks | 12427 | Original Beef Jerky |

Table 4.11: Online FMCG data merged table continued

| order_id | aisle_id | aisle | add_to_cart_order | reordered |
|---|---|---|---|---|
| 2539329 | 77 | soft drinks | 1 | 0 |
| 2539329 | 91 | soy lactosefree | 2 | 0 |
| 2539329 | 54 | paper goods | 5 | 0 |
| 2539329 | 23 | popcorn jerky | 3 | 0 |
| 2539329 | 23 | popcorn jerky | 4 | 0 |
| 2398795 | 24 | fresh fruits | 4 | 0 |
| 2398795 | 77 | soft drinks | 1 | 1 |
| 2398795 | 121 | cereal | 6 | 0 |
| 2398795 | 117 | nuts seeds dried fruit | 2 | 0 |
| 2398795 | 23 | popcorn jerky | 3 | 1 |

### 4.5.2 Insights from the data tables

#### 4.5.2.1 Products most often purchased

In Figure 4.2, the frequency of purchases from each department can be seen. From the figure, it is clear that the products most ordered are from the "produce" department and the fewest products are purchased from the "bulk" department.

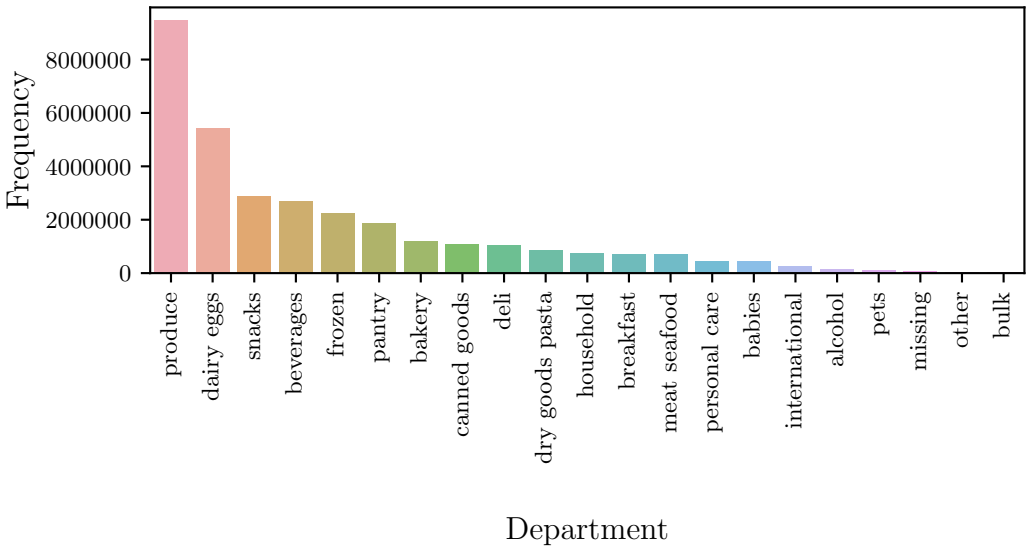



Figure 4.2: Frequency of purchases made from each department

Looking into the "Produce" department, Table 4.12 displays the 10 most frequently purchased products from this department. It can be observed that these are mainly organic items and fruits. Interestingly the top nine products are also the most purchased products in the entire dataset, with the tenth entry replaced by "Organic Whole Milk", with a total of 137 905 purchases. This product is the most purchased product from the "Dairy Eggs" department.

Table 4.12: Most purchased products from the "Produce" department

| product_name | Times purchased |
|---|---|
| Banana | 472 565 |
| Bag of Organic Bananas | 379 450 |
| Organic Strawberries | 264 683 |
| Organic Baby Spinach | 241 921 |
| Organic Hass Avocado | 213 584 |
| Organic Avocado | 176 815 |
| Large Lemon | 152 657 |
| Strawberries | 142 951 |
| Limes | 140 627 |
| Organic Raspberries | 137 057 |

The most frequently purchased products from the "Dairy Eggs" department can be seen in Table 4.13, these products mainly being milk, almond milk, cheese and eggs.

Table 4.13: Most purchased products from the "Dairy Eggs" department

| product_name | Times purchased |
|---|---|
| Organic Whole Milk | 137 905 |
| Organic Half & Half | 76 360 |
| Half & Half | 69 217 |
| Organic Whole String Cheese | 59 676 |
| Organic Unsweetened Almond Milk | 57 895 |
| Unsweetened Almondmilk | 49 569 |
| Organic Reduced Fat 2% Milk | 47 839 |
| Grated Parmesan | 45 238 |
| Large Alfresco Eggs | 40 376 |
| Organic Grade A Free Range Large Brown Eggs | 40 045 |

#### 4.5.2.2 Customers and their orders

Figure 4.3 shows the maximum number of orders that customers made as found in the dataset. It can be seen that the lowest number of orders was four and the highest number of orders was 100, with a spike at 100 orders. This is most probably because the number of orders was capped at 100, so all customers having more than 100 orders were binned together, and their orders after the hundredth one were deleted. The figure is right-tailed, indicating that the number of customers decreases as the number of orders per customer increases.

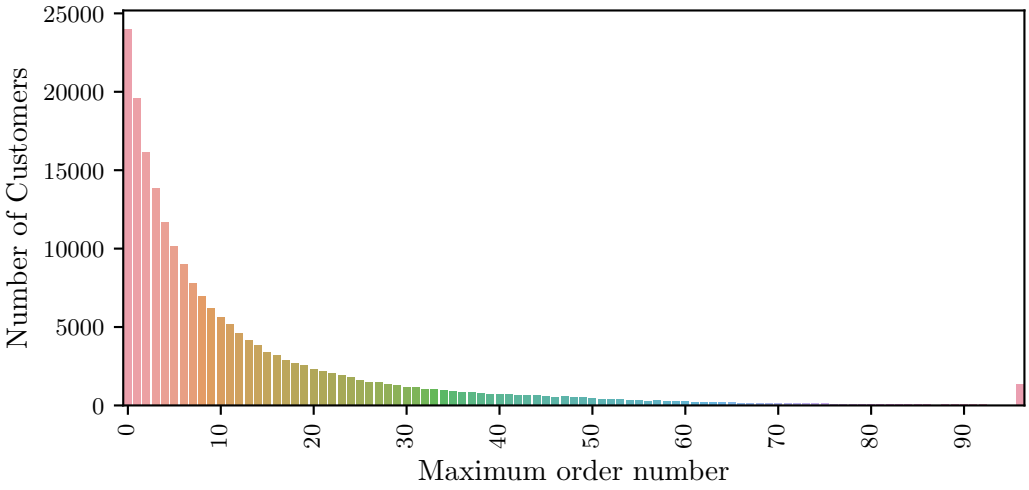



Figure 4.3: Number of orders per customer

In Figure 4.4 the number of products per order can be seen. Again a right-tailed distribution can be observed, with the most frequent products per order being five.

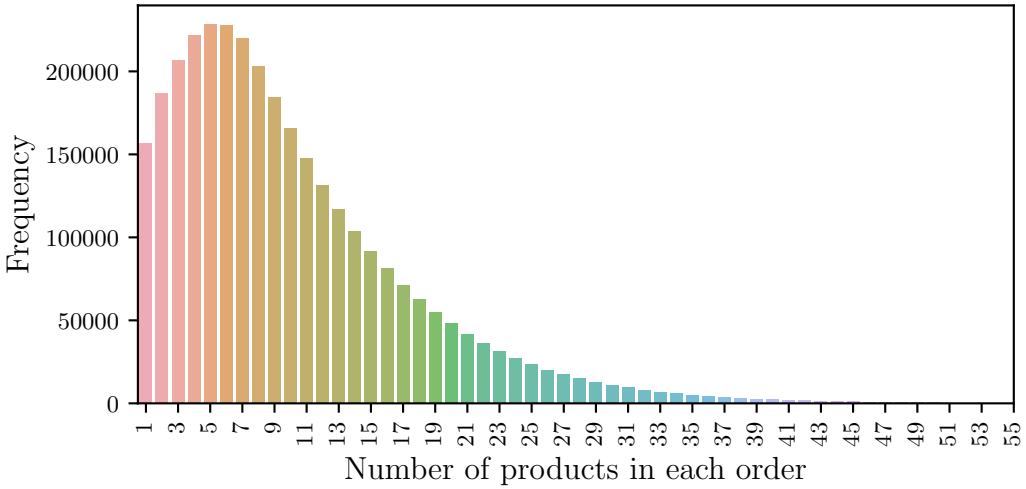



Figure 4.4: Number of products per order

#### 4.5.2.3 Times that customers make orders

Figure 4.5 shows the number of orders per day of the week. From the data, it is not possible to say which day of the week corresponds to the number specified in the dataset. It is, however, apparent that more orders are made on Day 1 and 2.

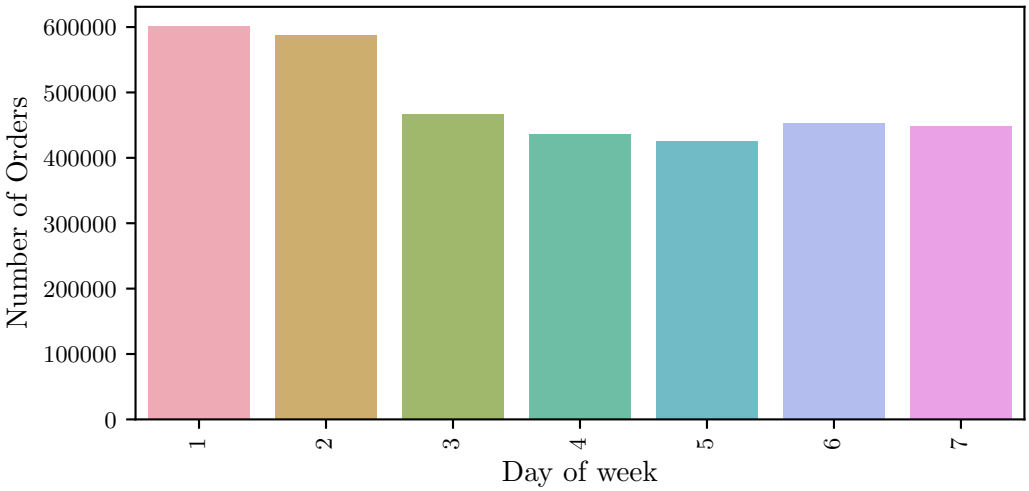



Figure 4.5: Number of orders per day of the week

Figure 4.6 shows the number of orders made per hour of the day. It can be seen that very few orders are made between 00h00 and 05h00, which would be expected as customers mostly

make purchases during the day. Purchases increase through the morning to a maximum at 10h00, with a slight dip over lunch hours and then gradually decreases as the day progresses, with few purchases late in the evening.

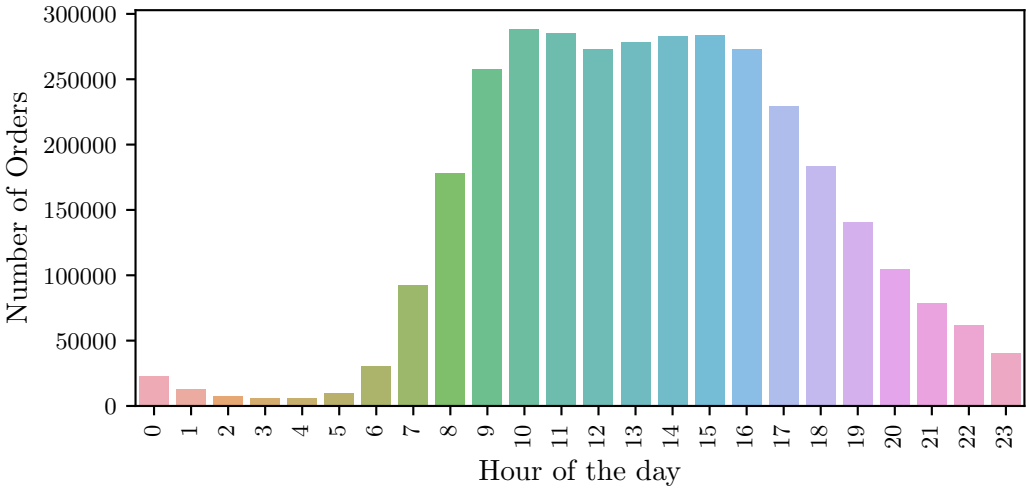



Figure 4.6: Number of orders per hour of the day

Figure 4.7 shows the days between orders for all the orders in the dataset. It can be seen that there are spikes at seven days between orders and 30 days between orders. The seven days between orders could possibly be a weekly trend, whereas the 30 days between orders could be because the data was censored at 30 days between orders. Smaller peaks can also be observed at 14 days, 21 days and 28 days. This could be because of a weekly pattern (that a customer needs a product every second or third week, for example).

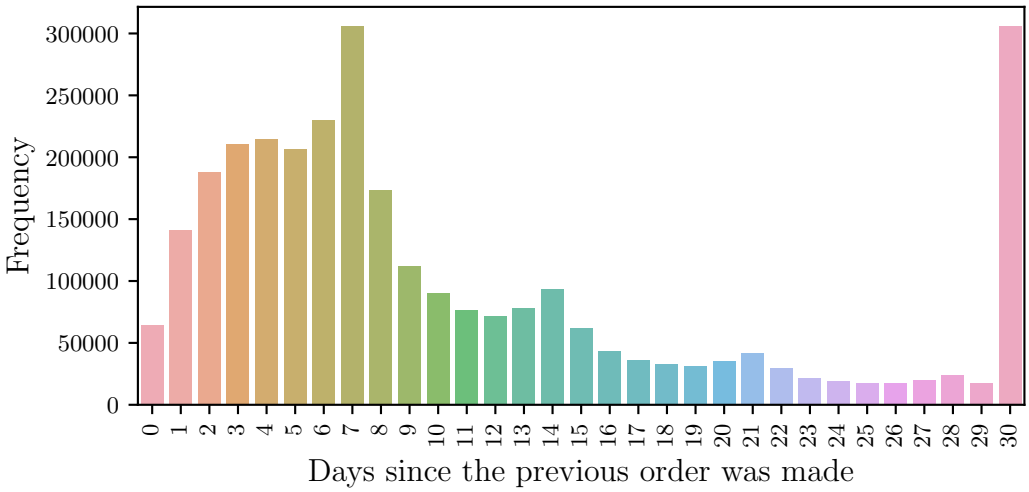



Figure 4.7: Frequency of the days between orders

### 4.5.3 Findings on the reorder rate

In the order_product table, there is a feature "reordered". This is a Boolean value, with 1 indicating the product was purchased in the previous order and 0 indicating that the item was not purchased in the previous order. Every time a customer starts using the service for the first time, the values are 0, because it is the first time that a customer purchases certain products. Table 4.14 shows the percentage of reorders versus the percentage of products that were not reordered. The reorder percentage is higher than the non-reorder percentage, which indicates that customers often purchase the same items.

Table 4.14: Percentage reorders

| **reorder** | **counts** | **percentage** |
|---|---|---|
| reordered | 19126536 | 58.97% |
| non-reorder | 13307953 | 41.03% |

In Figure 4.8, the reorder rate per product for the products with the highest reorder rate can be seen. This plot shows the products which are most often reordered by customers. These products are from different departments, but a product such as the "Fragrance Free Clay with Natural Odor Eliminator Cat Litter" can be expected to be reordered if the customer owns a cat and needs clean cat litter and is mostly satisfied with the product.

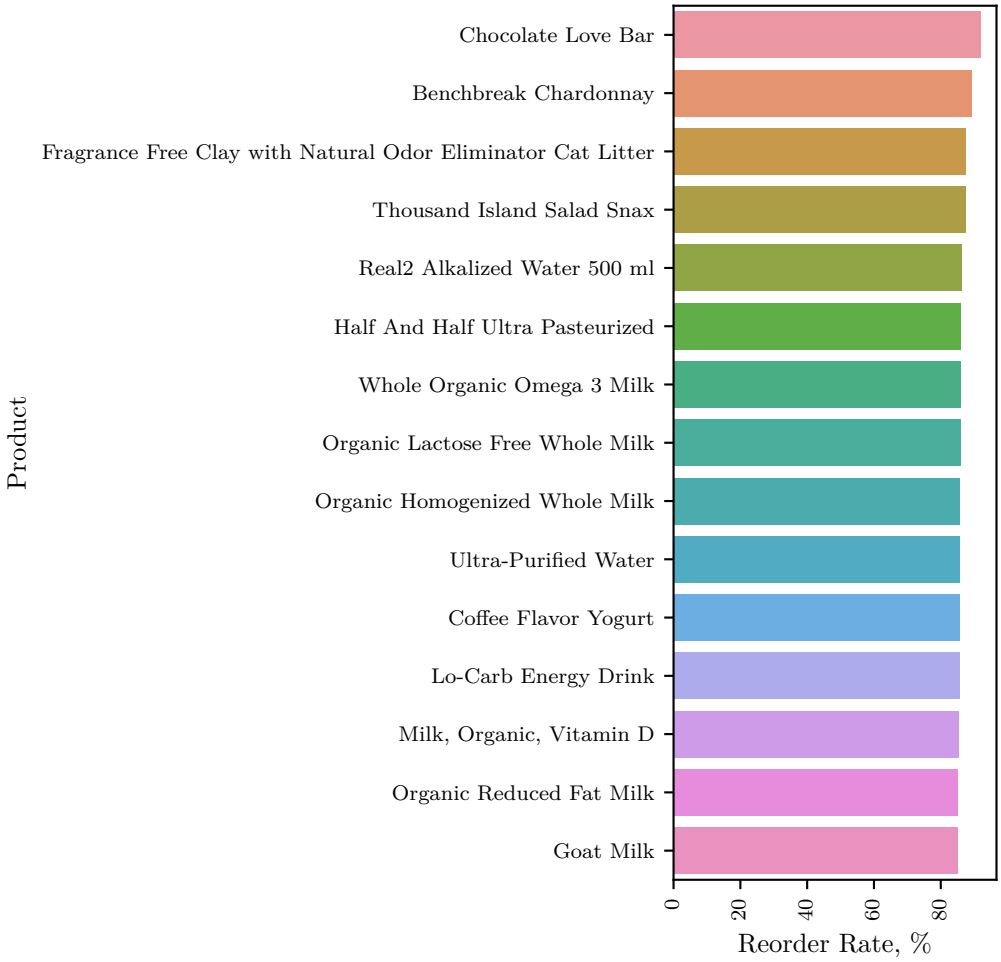



Figure 4.8: Highest reorder rate for products

This concludes the discussion of aspects of the online FMCG dataset.

## 4.6 Conclusion: Chapter 4

In this chapter the dataset that will be used to create the NPD Predictor was selected and explored. A broad overview of the dataset was given, starting with the relational structure of the data. Insights from the tables were explored throughout the rest of the chapter. This concludes the Explore step in the SEMMA process and the Data Understanding step in the CRISP-DM process. Understanding the data is vital before data preparation can commence. The next step in the CRISP-DM cycle is the Data Preparation step and the Modify step in the SEMMA cycle. This will commence in the next chapter.

# Chapter 5

# Data Preparation

In the previous section a summary of the dataset was given corresponding to the Data Understanding step in the CRISP-DM model. In this section the data will be prepared so that a model can be built to predict the Next Purchase Date. This corresponds with the third step in the CRISP-DM and SEMMA models, discussed in Section 3.1.

The Next Purchase Date (NPD) per user-product pair is the target feature that must be predicted. This feature is not explicitly available in the dataset; thus, this feature must be derived. However, the feature that is available that does describe the time between orders for a user is the "days_since_prior_order" feature, available in the `orders` table described in Chapter 4. This feature does not capture the days between which a specific product was ordered. Thus a new feature must be created to capture the target variable.

The features that were created can be described as sequence-based and non-sequence-based features. Sequence-based features were created as the purchasing of items by a customer forms a sequence, and the Next Purchase Date in the sequence should be predicted.

## 5.1 Sequence-based Features

The sequence-based features that were created will be discussed in this section.

### 5.1.1 Feature 1: days between orders per product

The dataset only specifies days between orders for a user. This is a helpful feature, but as it is desired to predict the NPD for a user-product pair it is necessary to transform the "days_since_prior_order" feature into a "days_between_orders_per_product" feature, as this feature will capture the desired value to be predicted. This was done by constructing a table for each user, that consists of the order number for the user with a corresponding 0 or 1, which indicates that the user purchased the product in that specific shopping instance. The "days_between_orders" feature is then used to calculate the "days_between_orders_per_product" by looking at the instances when the product was purchased. Table 5.1 shows an example for a user. For instance, this user purchased "Organic Fuji Apples", "Organic String Cheese" and "Original Beef Jerky" in their fifth shopping instance.

An example of how the "days_between_orders_per_product" is calculated, can be seen in Figure 5.1 for the product "Original Beef Jerky". The symbol " | " indicates the product was purchased with that order, and an O indicates that the product was not purchased with

Table 5.1: Example of product order detail for a customer

| order_id | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Organic Fuji Apples | 0 | 0 | 0 | 0 | 1 |
| Organic String Cheese | 0 | 0 | 1 | 1 | 1 |
| Organic Unsweetened Vanilla Almond Milk | 1 | 0 | 0 | 0 | 0 |
| Original Beef Jerky | 1 | 1 | 0 | 1 | 1 |
| days_since_prior_order | 0 | 15 | 21 | 29 | 28 |

that order. As seen, the user did not purchase the item with order 3. The corresponding days between product purchases can also be seen in the figure. The resulting sequence for "days_between_orders_per_product" is thus,

$$[15, (21 \text{ days} + 29 \text{ days}) = 50 \text{ days}, 28].$$

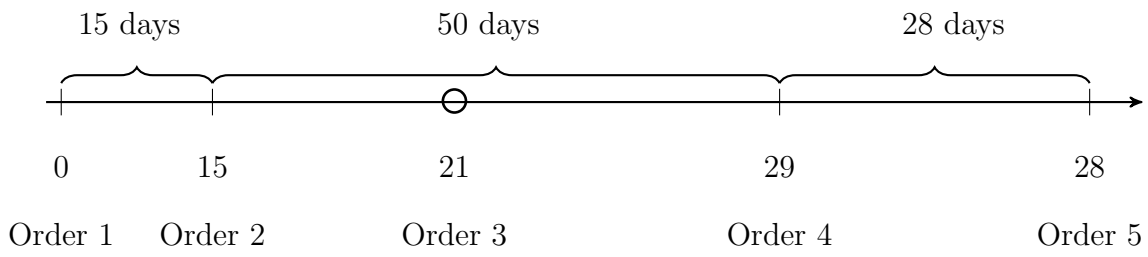



Figure 5.1: Example of "days_between_orders_per_product" feature creation for "Original Beef Jerky"

Table 5.2 shows the feature for all the products that this customer, called Customer $X$, purchased. It can be observed that the sequence for "Organic Unsweetened Vanilla Almond Milk" is empty. This is because the product was only purchased in the first order; thus, no prior knowledge is available.

Table 5.2: days_between_orders_per_product for all the products that Customer $X$ purchased

| customer | product | days_between_orders_per_product |
|---|---|---|
| $X$ | Organic Fuji Apples | [93] |
| $X$ | Organic String Cheese | [36,29,28] |
| $X$ | Organic Unsweetened Vanilla Almond Milk | [] |
| $X$ | Original Beef Jerky | [15,50,28] |

The last value of the sequence created in feature 1 is the target variable that must be predicted.

### 5.1.2 Feature 2: days since prior order per product

The "days_since_prior_order_per_product" feature was created to capture some of the user's behaviour, along with the user-product behaviour. An example of how this feature is created again for "Original Beef Jerky" can be seen in Figure 5.2. This feature evaluates, given that the customer purchased the product, how many days ago the user made a purchase (not necessary the product that is being evaluated but any product). As seen in the example, the product was purchased with Order 2; thus the customer previously ordered 15 days ago. This means the first entry for the sequence will be 15, but in Order 3 this product was not purchased, so no record for this instance will be taken as the product was not ordered. The product was again ordered with Order 4; therefore, an entry will be made, and the previous order was made 29 days ago. Thus, the complete sequence in this example for the "days_since_prior_order_per_product" will be

$$[15, 29, 28].$$

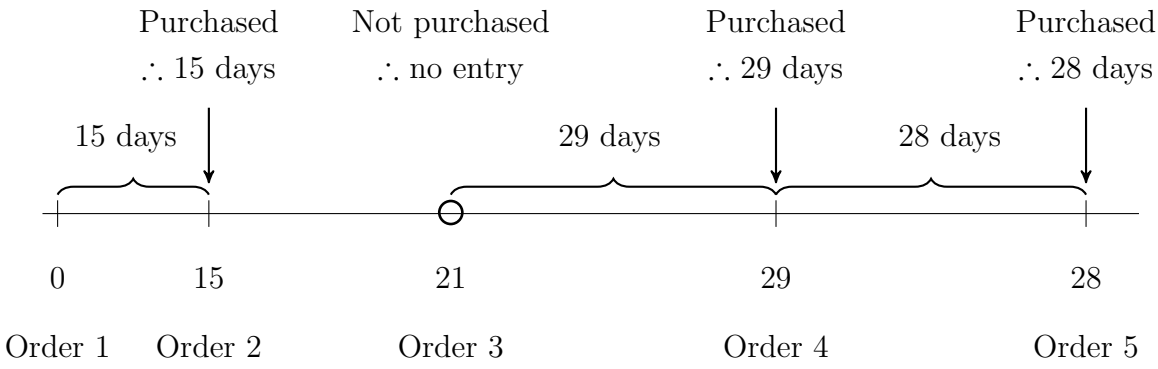



Figure 5.2: Example of "days_since_prior_order_per_product" feature creation for "Original Beef Jerky".

Table 5.3 shows this feature for all the products that customer $X$, purchased. Again it can be seen that "Organic Unsweetened Vanilla Almond Milk" has an empty sequence, and as before, the feature is dependent on prior history and for this product, there is no purchase history.

This feature combines the customer purchase behaviour with the customer product purchase behaviour, whereas "days_between_orders_per_product" only takes into account the user's behaviour for the specific product.

Table 5.3: days_since_prior_order_per_product for all the products that Customer $X$ purchased

| customer | product | days_between_orders_per_product |
|---|---|---|
| $X$ | Organic Fuji Apples | [28] |
| $X$ | Organic String Cheese | [21,29,28] |
| $X$ | Organic Unsweetened Vanilla Almond Milk | [] |
| $X$ | Original Beef Jerky | [15,29,28] |

These features were created for all user-product pairs in the dataset described in Chapter 4. They were stored and used to derive non-sequence-based features from these sequences, which will be described in the next section. These sequences are also used in the modelling phase. The structure of this newly created dataset with its features can be seen in Table 5.4.

Table 5.4: Dataset created with two sequence features

| user_id | product_id | days_between_orders_per_product | days_since_prior_order_per_product |
|---|---|---|---|
| 136235 | 24010 | [100.0, 15.0, 105.0, 70.0] | [6.0, 4.0, 4.0, 3.0, 30.0] |
| 91009 | 5764 | [10.0, 27.0, 20.0, 62.0, 50.0] | [30.0, 10.0, 23.0, 20.0, 9.0, 11.0] |
| 182083 | 24177 | [69.0, 210.0, 9.0, 8.0] | [8.0, 8.0, 8.0, 9.0, 8.0] |
| 83817 | 3479 | [40.0, 43.0, 41.0, 19.0, 29.0, 30.0, 19.0, 22.... | [30.0, 30.0, 26.0, 30.0, 14.0, 29.0, 30.0, 19.... |
| 193273 | 16797 | [104.0, 8.0, 77.0, 76.0, 5.0] | [7.0, 9.0, 7.0, 5.0, 6.0, 5.0] |

## 5.2 Non-sequence-based Features

Some features were created to describe the sequence but do not form a sequence themselves. These features will be described in this section and are created to look at the bigger picture (all the data) and not just at the sequence of one user-product pair. Table 5.5 gives a list of all the non-sequence-based features that were generated with a description to explain how the features were derived. To understand these features, it is important to explain how the entire dataset was constructed. Figure 5.3 shows how the dataset for the non-sequence-based features was created. Each user-product pair's last value of the sequence "days_between_orders_per_product", is the target variable t_0. The sequences were then split into windows with a size of six, each window having five training examples and a target value to predict. The last window was separated into a test set. The other windows were flattened, which means that there is more than one user-product pair, and this created the training set.

The sets were then further expanded with features derived from the five variables in the training window. These features are listed and explained in Table 5.5. Each of these variables was created based on the five values in the training window. For example, if the "days_between_orders_per_product" sequence looked as follows:

$$[23, 19, 16, 24, 23, 28, 30, 12, 19, 27, 15, 21, 29],$$

the target variable t_0 = 29, the last window = 30, 12, 19, 27, 15, 21, (days) with 21 the target variable for this window, the second last window = 23, 19, 16, 24, 23, 28, (days) with 28 the target variable for this window.

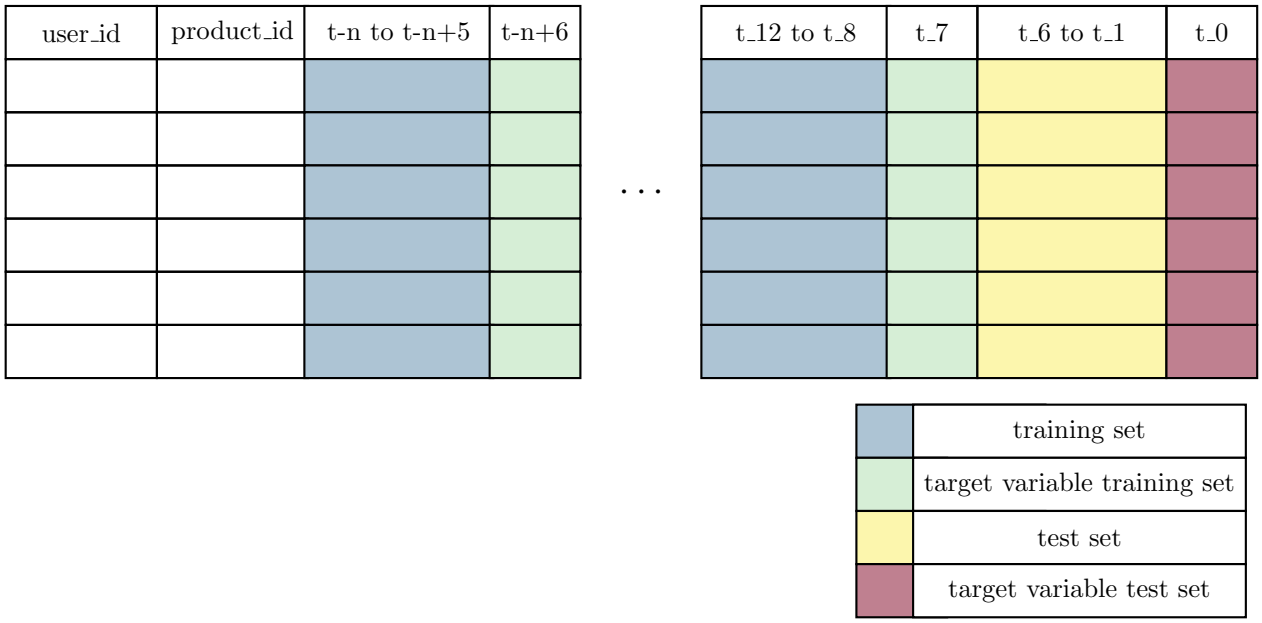



Figure 5.3: Train and test set splitting

The features seen in Table 5.5, are for each window in the dataset created as explained through Figure 5.3. This means that the "Max days" feature for the last window would be 30 days, and so on.

Table 5.6 shows the final form of the training dataset with the target variable (value to be predicted) t_0. This value of t_0 is the Next Purchase Date of that entry. As it can be seen there are multiple entries for User 17 for both products 7350 and 18534. This is because the sequence was split into windows of size six and flattened. The user-product pair (User:17, product:7350), has a sequence length of at least 24 entries.

Table 5.5: Features (non-sequence-based)

| Feature | Description |
|---|---|
| User id | Unique user identifier. |
| Product id | Unique product identifier. |
| Max days | Maximum days that the user went without the product. |
| Min days | Minimum days that the user went without the product. |
| Variance | The variance of the sequence created in Feature 1. |
| Average days | The average of the sequence created in Feature 1. |
| Days since | The second last entry of the sequence created in Feature 2, as the last entry contains information about the target variable. |
| t_5, t_4, t_3, t_2, t_1 | The value in the sequence that corresponds with the window. |

Table 5.6: Example of the final form of the training datasets

| User id | Product id | Max days | Min days | Avg days | Days since | variance | t_5 | t_4 | t_3 | t_2 | t_1 | t_0 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 17 | 7350 | 30 | 0 | 8.16 | 8 | 54.47 | 5 | 4 | 3 | 5 | 4 | 8 |
| 17 | 7350 | 30 | 0 | 9.22 | 9 | 66.95 | 5 | 6 | 0 | 9 | 6 | 9 |
| 17 | 7350 | 30 | 4 | 11.42 | 3 | 82.07 | 4 | 4 | 5 | 30 | 12 | 3 |
| 17 | 7350 | 30 | 4 | 11.00 | 16 | 77.66 | 7 | 4 | 9 | 30 | 5 | 16 |
| 17 | 17762 | 36 | 0 | 11.50 | 3 | 113.37 | 6 | 0 | 9 | 6 | 18 | 3 |
| 17 | 17762 | 36 | 3 | 14.00 | 5 | 146.80 | 20 | 9 | 36 | 6 | 3 | 5 |
| 17 | 18534 | 39 | 3 | 9.71 | 6 | 78.20 | 9 | 3 | 5 | 4 | 8 | 6 |

Table 5.7 shows the final form for of the test dataset. It can be seen in this dataset there is only one entry for each of the user-product pairs. This is because the last window for each pair was separated into the test set. The t_0 value for this dataset is the Next Purchase Date for each user product pair and this is the value that must be predicted by the NPD Predictor.

Table 5.7: Example of the final form of the test datasets

| User id | Product id | Max days | Min days | Avg days | Days since | variance | t_5 | t_4 | t_3 | t_2 | t_1 | t_0 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 17 | 7350 | 30 | 0 | 8.30 | 4 | 50.47 | 6 | 21 | 4 | 4 | 10 | 4 |
| 17 | 17762 | 36 | 0 | 10.86 | 5 | 98.39 | 5 | 18 | 21 | 4 | 4 | 5 |
| 17 | 18534 | 39 | 3 | 9.37 | 5 | 63.05 | 13 | 8 | 8 | 10 | 4 | 5 |
| 21 | 23729 | 28 | 2 | 11.44 | 0 | 48.58 | 5 | 9 | 6 | 16 | 2 | 0 |
| 27 | 1194 | 38 | 1 | 14.85 | 1 | 106.03 | 15 | 30 | 22 | 18 | 22 | 52 |

## 5.3 Conclusion: Chapter 5

This chapter explained the creation of the datasets that will be used in the modelling phase of the CRISP-DM cycle and the model phase in the SEMMA cycle. It also explains the features that were created along with the difference between a sequence-based approach and a non-sequence-based approach. The next chapter will explain how the models were built to predict the Next Purchase Date for a user-product pair using a sequence-based approach and a non-sequence-based approach.

# Chapter 6

# Next Purchase Date Predictor Modelling

The previous chapter explained how the data was prepared. In this section, the prepared data will be used to build models that will predict the Next Purchase Date (NPD) for a user-product pair. This corresponds to the fourth step in the CRISP-DM and SEMMA processes.

Various machine learning techniques will be discussed and modelled to predict the NPD. The first approach followed was to predict the NPD by using only the sequence-based features. This was trained using a Recurrent Neural Network (RNN) and a Regression model as explained in Section 3.4. The second approach used the non-sequence-based features to train an Extreme Gradient Boosting (XGBoost) model and an Artificial Neural Network (ANN) model also explained in Section 3.4.

## 6.1 Recurrent Neural Network

The RNN was trained using only the sequence-based features discussed in Section 5.1. The first implementation took the entire sequence of feature 1, except the last entry of the sequence to train the model and the last entry in the sequence as the test variable, with which the prediction can be compared. This is visually displayed in Figure 6.1, with $f1_1$ representing the first variable in the feature 1 sequence and $f1_n$ representing the last variable in the feature 1 sequence, excluding the test feature. The NPD prediction is represented by $y_n$ and $h_{t-1}$ is initialised with 0.

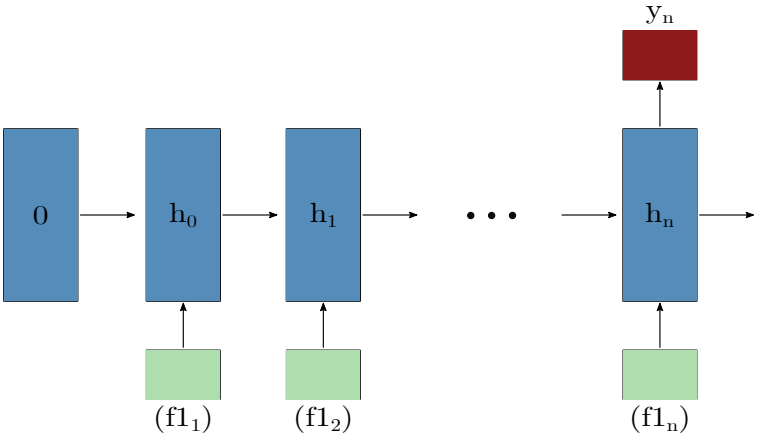



Figure 6.1: RNN implementation with one feature

The second implementation of an RNN model took both the sequences of feature 1 and feature 2. The target variable remained the same (the last entry in the feature 1 sequence), and the last entry of both sequences was removed from the training. Thus, the training set size is equal to the user-product pair sequence size without the last variable. Figure 6.2 displays the implementation of the second RNN model. Two features f1 and f2 are used to train the model at each time step, and the Next Purchase Date prediction is represented by $y_n$.

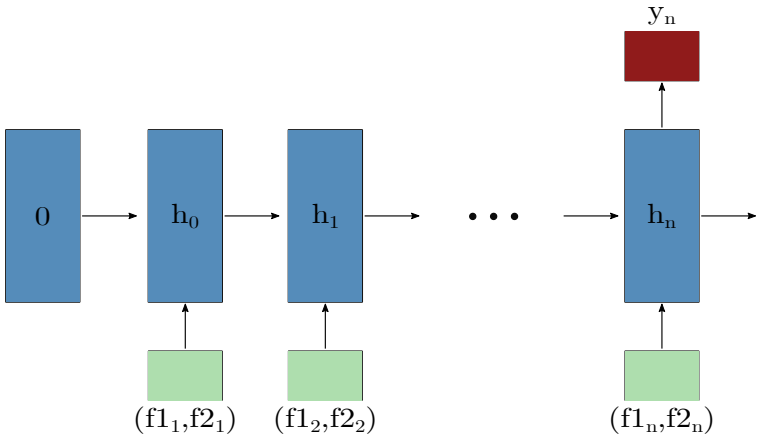



Figure 6.2: RNN implementation with two features

Before the RNN models were trained the sequence data was scaled to range between 0 and 1 by using,

$$\text{scaled}(\text{f1}_i) = \frac{\text{f1}_i}{\text{f1\_max}},$$

for the first RNN implementation, where f1_max is the largest value in the sequence f1, and $\text{f1}_i$ represents the $i$-th entry in the f1 sequence. The prediction is then scaled back by multiplying with f1_max. For the second implementation of the RNN, with two features, both features were scaled by,

$$\text{scaled}(\text{f1}_i, \text{f2}_i) = \frac{(\text{f1}_i, \text{f2}_i)}{\max(\text{f1\_max}, \text{f2\_max})}.$$

The parameters set were the optimiser, learning rate, number of hidden layers, activation function and the criterion. The Adam optimiser was used (discussed in Section 3.4.2.3), with a learning rate of 0.01 and 10 hidden layers. A rectified linear unit (ReLU) activation function (discussed in Section 3.4.2.2) was used with a mean squared error loss criterion. Because sufficient data is needed to train the model, the user-product pair was filtered to have at least 20 entries, thus leaving a dataset with 112 796 user-product pairs to predict the NPDs.

## 6.2 Linear Regression

The next sequence-based approach was implemented by using linear regression. Figure 6.3 shows how the sequence data was represented for linear regression, with the shopping instance on the $x$-axis and the sequence value (days between purchases) on the $y$-axis. For this implementation the next value of the shopping instance, in this case instance 46, was predicted.

Before training the model, the data was scaled. This was done by using the Standard scaler function of Scikit-learn, a software machine learning library for the Python programming language (Pedregosa et al., 2011). The standard score of a sample $F$ is calculated as:

$$\mathrm{z} = \frac{(\mathrm{F} - \mu)}{\mathrm{s}},$$

where $\mu$ is the mean of the training sample and s is the standard deviation of the training sample. Centring and scaling happen independently on each feature (if the dataset has more than one feature) by comparing the relevant statistics on the sample in the training set. The mean and the standard deviation for the feature are then stored and used to transform the predictions made by the model.

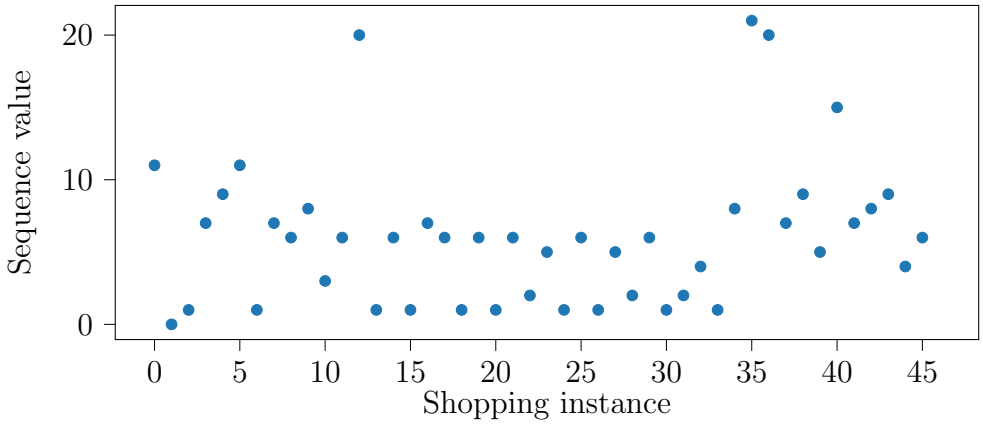



Figure 6.3: Example of how Linear Regression was implemented for a sequence

## 6.3 Extreme Gradient Boosting

An approach to not use only sequence-based features was explored, by training an Extreme Gradient Boosting (XGBoost) model, with the training set seen in Table 5.6. This was done to see if there is value in training data from other users as well.

Unlike the RNN and regression techniques discussed, XGBoost does not require scaled data, but XGBoost has hyperparameters that must be set before training the model. It is important to distinguish between hyperparameters and model parameters. Hyperparameters

are parameters that are set before the machine learning process begins. Model parameters are parameters that are learnt through the process, such as the weights. The hyperparameters were tuned using the training set seen in Table 5.6 and a K-fold cross-validation approach. With K-fold cross-validation the dataset is randomly divided into K subsets, known as folds. During hyperparameter tuning, for each new configuration, the model is trained on the folds $Y = K \setminus \{a\}$, and the resultant model is tested by using fold $a$. This is then repeated for $a = 1, 2, \ldots, K$. The final model performance is then calculated as the average performance over all $K$ iterations (Stone, 1974). The parameters' search space can be seen in Table 6.1, where the parameter is listed along with the parameter default value and the range that was specified as a solutions search space.

Table 6.1: Hyperparameter search space for XGBoost

| **Parameter** | **Default** | **Specified search space** |
|---|---|---|
| n_estimator | 100 | [10,100,500,800,1000,1200,1500] |
| max_depth | 3 | [2,3,4,5,8,10,15] |
| booster | gbtree | [gbtree,gblinear] |
| learning_rate | 0.1 | [0.01,0.05,0.1,0.15,0.2] |
| min_child_weight | 1 | [1,2,3,4] |
| base_score | 0.5 | [0.25,0.5,0.75,1] |

Fifty iterations of randomly selected parameters from the specified lists in Table 6.1 were performed and the results of the tests can be seen in Appendix A. The results of the top-ranked 28 iterations are shown in Table A.3. The performance metric used is the negative mean absolute error (NMAE), and the results for all 50 configurations are plotted in Figure 6.4.

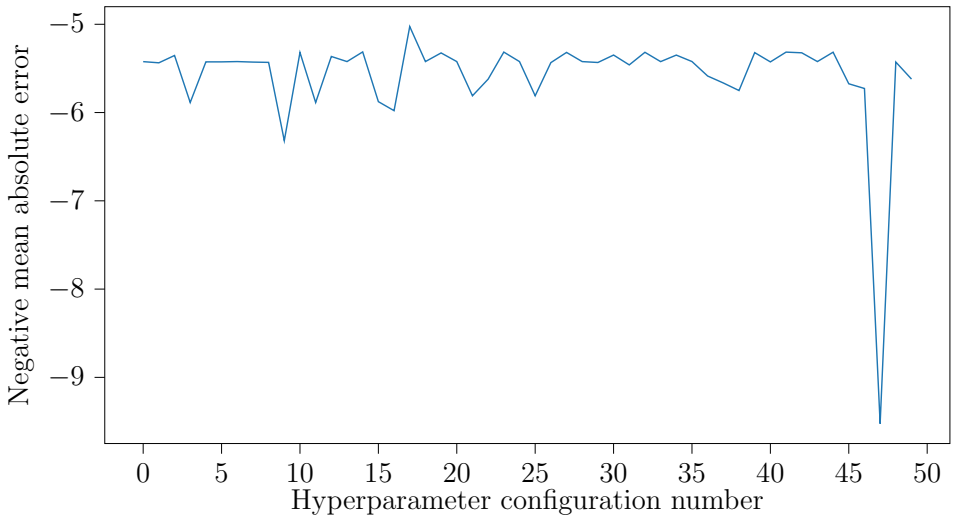



Figure 6.4: Results of hyperparameter random search configurations

In this plot, it can be seen that most of the iterations resulted in an NMAE (in days) between -5 and -7 days, with one of the configurations having an NMAE over -9 days. The best configuration had an NMAE of -5.03, with parameters as seen in Table 6.2.

Table 6.2: Best performing parameter configuration for XGBoost

| Parameter | Value |
|---|---|
| n_estimators | 10 |
| min_child_weigh | 2 |
| max_depth | 15 |
| learning_rate | 0.2 |
| booster | gbtree |
| base_score | 0.5 |

After training the model with the specified hyperparameters, the feature importance can be seen in Figure 6.5. As seen from this figure, it is important for the model to know for which user and which product it is making a prediction, followed by the variance of the last five purchases. It can also be noted that t_5 to t_1 carry the same importance, with min_days and days_since being the least important features.

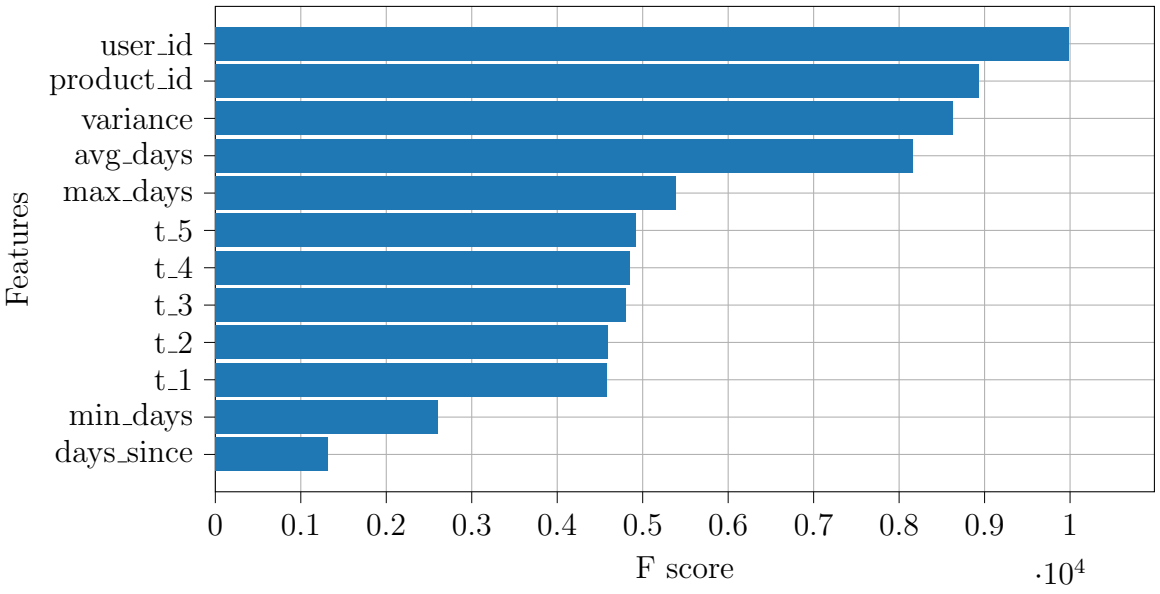



Figure 6.5: Feature importance determined by XGBoost

## 6.4 Neural Network

An Artificial Neural Network (ANN), (from here on also referred to as Neural Network) model was trained with the same training set as the XGBoost model. Just like the RNN and Regression models, the data must first be scaled before the neural network can be trained. This was done with the same Standard scaler function of Scikit-learn used for the regression implementation.

After the data was scaled, hyperparameters had to be set for the neural network as well, and the same process of random search k-fold cross-validation used for the XGBoost model was followed. The hyperparameter search space for the neural network can be seen in Table 6.3. The search space includes neural networks with up to three hidden layers, two activation functions and two different batch sizes.

Table 6.3: Hyperparameter search space for the Neural Network model

| **Parameter** | **Specified search space** |
|---|---|
| layers | 1 layer : neurons[[10],[20]] |
| | 2 layers: neurons[[10,5],[20,10]] |
| | 3 layers: neurons[[45,30,15]] |
| activation functions | [relu, sigmoid] |
| batch size | [50,100] |

Again, the NMAE performance metric was used, and the results of the 20 configurations can be seen in Figure 6.6. Notably, this graph has a much smaller range for the NMAE values, and the NMAE are much smaller than those of the XGBoost model. This is because the data is scaled as explained. Once the Neural Network model has been trained with the hyperparameters that perform best and the data is scaled back, the two models can be compared.

When looking at Figure 6.6 it can be seen that some of the configurations perform better than others. There is no noticeable parameter causing this difference in performance, and it is thus due to the combination of parameters. The hyperparameters that performed best according to the random search cross-validation can be seen in Table 6.4 and this configuration had an NMAE of -0.6633 with the scaled data. The results of all 20 configurations can be seen in Table A.4.

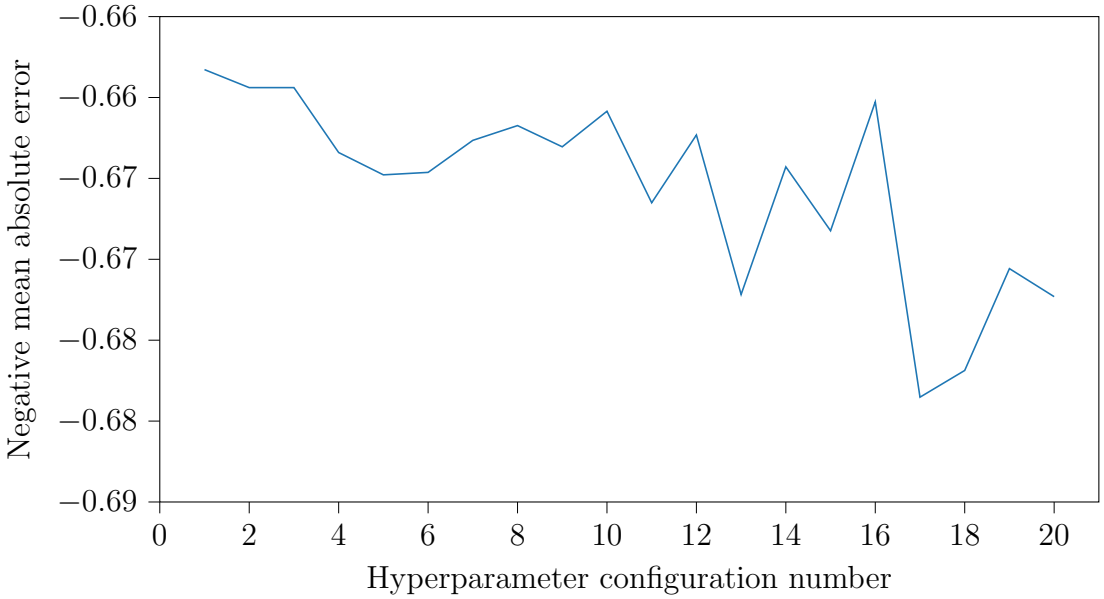



Figure 6.6: Results of hyperparameter random search configurations

Table 6.4: Best performing parameter configuration for the Neural Network

| Parameter | Value |
|---|---|
| layers | 2 layers: [10,5] |
| activation function | relu |
| batch size | 100 |

The Neural Network was trained with 12 input features (corresponding to 12 neurons in the input layer), two hidden layers with 10 and 5 neurons respectively and the output layer with one neuron. A rectified linear unit (relu) activation function was used with a mean squared error loss criterion with an Adam optimiser for the final implementation.

## 6.5 Conclusion: Chapter 6

This chapter explained how all the models (RNN, Regression, XGBoost and Neural Network) were created[1]. It also explained the hyperparameter optimisation for the XGBoost model as well as for the Neural Network model. In the next chapter, the results of all the models on the test set will be explained and the best model will be chosen for the NPD Predictor.

[1]The sourcecode is available at https://github.com/MarliDroomer/NPD-predictor

# Chapter 7

# Results of the NPD Predictor (Evaluation)

The previous chapter explained the various approaches to modelling the NPD Predictor. This chapter will explain the results from the models and a final selection of the appropriate model for the NPD Predictor will be made. This corresponds to the Evaluation step of the CRISP-DM process and the Assess step in the SEMMA cycle of data mining.

## 7.1 Comparing the models

To evaluate the models, the absolute error in days was calculated for each user-product pair in the test set, for each technique discussed in Chapter 6. The absolute errors were sorted from small to large for each technique and are plotted as shown in Figures 7.1 - 7.2. Figure 7.1 shows all the user-product pairs' absolute prediction errors. The Neural Network implementation performed best as it has the smallest absolute error over the number of user-product pair instances. In Figure 7.2 only the first 40 000 user-product pair instances are plotted. This shows that the Neural Network has a much smaller gradient than the other techniques. This also shows that the Neural Network can predict at least 40 000 user-product pairs with an absolute error of less than one-and-a-half days.

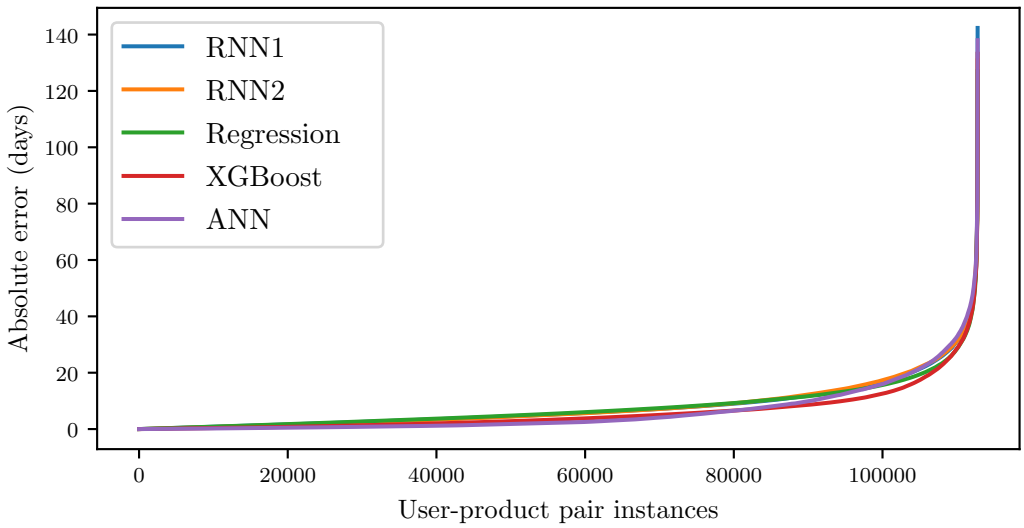



Figure 7.1: Absolute error for all user-product pair instances

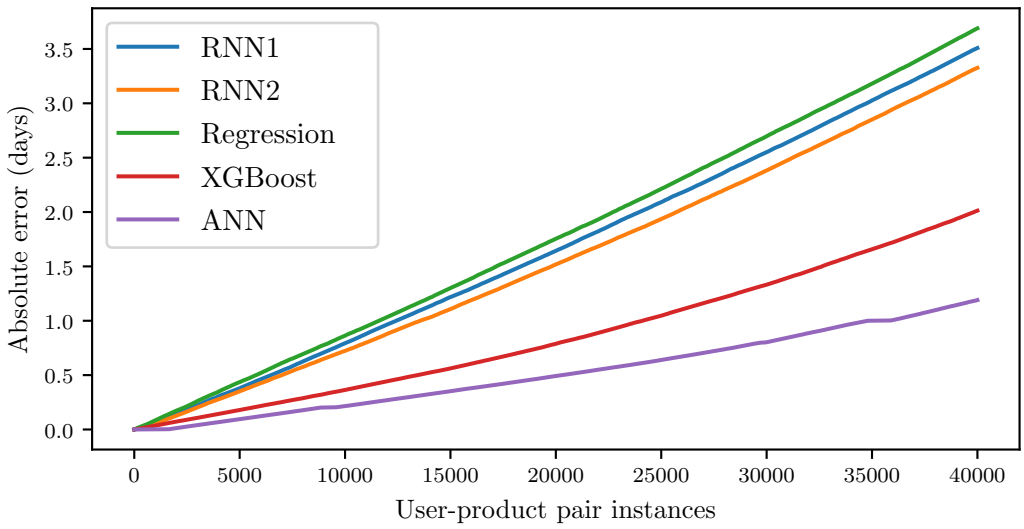



Figure 7.2: Absolute error for first 40 000 user-product pair instances

Figure 7.3 shows the number of user-product pairs per absolute error. It shows that the Neural Network predicts far more NPDs with smaller errors than the rest of the algorithms, while XGBoost also outperforms the sequence-based approaches. Thus, it can be said that the non-sequence-based algorithm outperforms the other algorithms in this instance. This also shows that looking at the bigger picture and not only looking at an individual user-product pair can be advantageous.

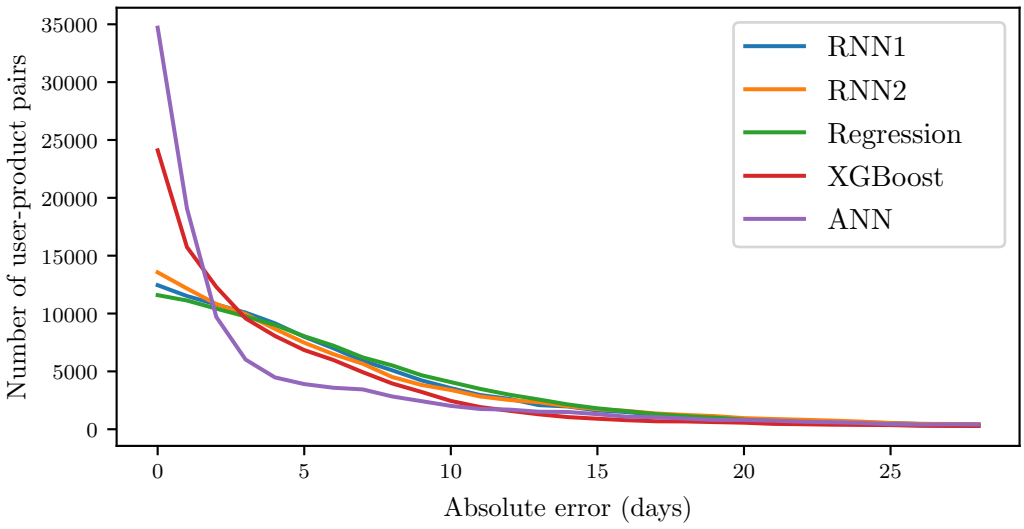



Figure 7.3: Number of user-product pairs predicted per absolute error in days

Figure 7.4 shows the percentage of the dataset that each algorithm predicted per category.

The Neural Network algorithm can predict the NPD with an error of less than one day for 31.8% of the dataset, and another 16.8% with an error of between one and two days. This outperforms all the other algorithms with the next best algorithm, XGBoost, predicting 19.3% with an error of less than one day. This also shows that the Neural Network can predict more than 55% of the user-product pairs with an error of fewer than three days.

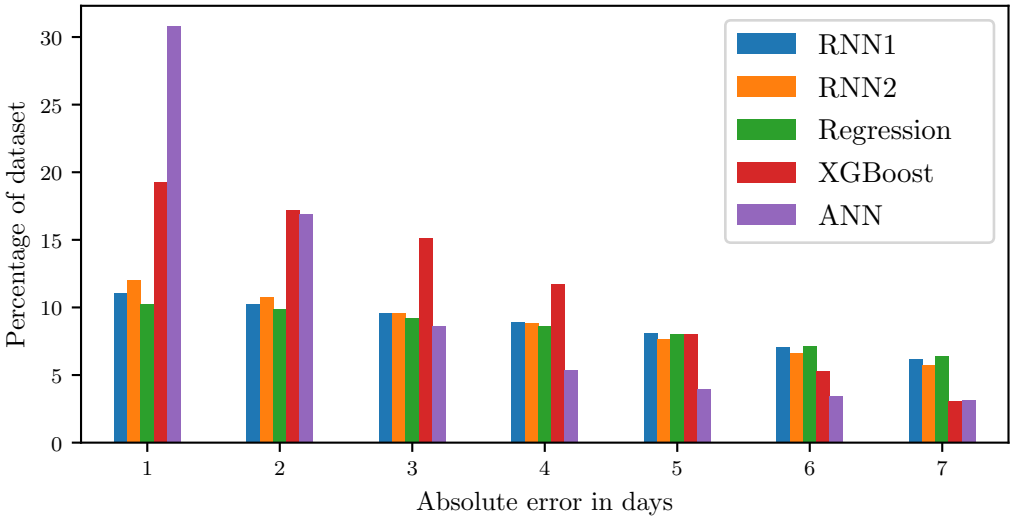



Figure 7.4: Percentage of the dataset that each technique predicted in absolute error days

To see if the algorithms performed well on the same user-product pairs, the user-product pairs from RNN2, XGBoost and ANN with an error of less than one day were plotted with a Venn-diagram which can be seen in Figure 7.5. This shows that only 3 702 of the total 52 760 (7%) pairs were predicted with an error less than one day by all three algorithms. Thus, a combined approach could possibly increase the quality of the NPD predictions.

It can also be noted that some products can be predicted with more accuracy than other products. Table 7.1 shows the number of times a sample of products was predicted correctly by the ANN model. These products were sampled based on the number of times the prediction was made correctly by the ANN model, for the product. The table gives the product_id, the product_name and the ANN_count, (which indicates the number of times the ANN predictor predicted the NPD for this product with an error of less than one day). The num_times_testset indicates the number of times the product appears in the test set, and the last column in the table, percentage_ANN gives the percentage that the ANN NPD model predicted the product with an error of less than one day. For example, the last entry in the table, Banana, the NPD was predicted with an error of less than one day 2 095 times, the product appeared 5 251 times in the test set; thus the percentage the ANN predicted the NPD with an error of less than one day is 39.9%. This table also shows that although Bananas have the most times that the

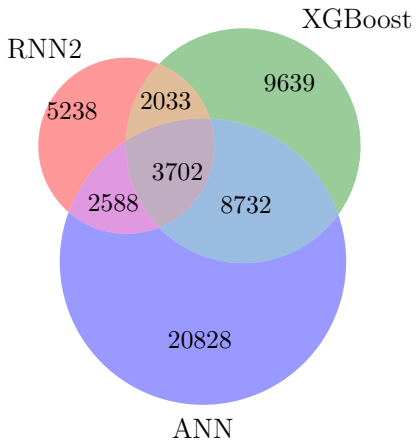



Figure 7.5: Venn diagram of user-product pairs predicted with error less than one day

NPD is predicted with an error of less than one day, Spring Water has the highest percentage correctly predicted from this list.

Table 7.1: Predictions of the ANN model

| product_id | product_name | ANN_count | num_times_testset | percentage_ANN |
|---|---|---|---|---|
| 44632 | Sparkling Water Grapefruit | 221 | 585 | 37.77 |
| 19660 | Spring Water | 248 | 574 | 43.20 |
| 27966 | Organic Raspberries | 331 | 1 031 | 32.10 |
| 47766 | Organic Avocado | 404 | 1 158 | 34.88 |
| 21903 | Organic Baby Spinach | 455 | 1 577 | 28.85 |
| 21137 | Organic Strawberries | 492 | 2 010 | 24.47 |
| 47209 | Organic Hass Avocado | 550 | 1 794 | 30.65 |
| 27845 | Organic Whole Milk | 564 | 1 555 | 36.27 |
| 13176 | Bag of Organic Bananas | 1 415 | 4 071 | 34.76 |
| 24852 | Banana | 2 095 | 5 251 | 39.89 |

When sorting this table by percentage_ANN, 1 420 products were predicted with a 100% percentage_ANN, which means that the ANN model predicted those NPD with an error of less than one day every time the product was predicted.

It should also be noted that there are some user-product pairs that are not so predictable. In Figure 7.6 the prediction with errors greater than 40 days are shown on a Venn-diagram for the ANN, XGBoost and RRN2 models. It can be seen that the same user-product pairs

usually perform poorly, with 726 out of the total 1 687 (43%) predictions predicted by all three models.

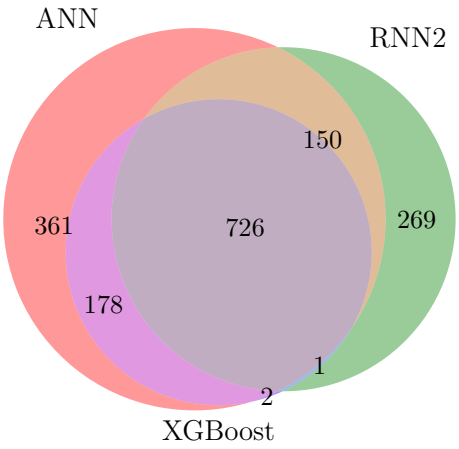



Figure 7.6: Venn diagram of user-product pairs predicted with an error of more than 40 days

To see if a combination approach increased the accuracy of the NPD Predictor, combination approaches were explored and are discussed in the next section.

## 7.2 Combination Methods

A sequence, sequence and sequence, non-sequence combination approach was followed for this analysis. The ANN, XGBoost and RNN1 models were used and the combinations tested were:

i ANN with XGBoost,

ii ANN with RNN, and

iii XGBoost with RNN.

To model the combination, each individual model (ANN, XGBoost and RNN) was trained using the same approach as described in Chapter 6, but one time-step back, meaning that the training data was up to t_2 and t_1 was the target variable. These predictions were then compared to the real t_1, and the model that performed best out of the two models was then selected to predict t_0.

The results of the combination methods can be seen in Figure 7.7. Interestingly none of the combination methods performs better than the ANN method. From the combination

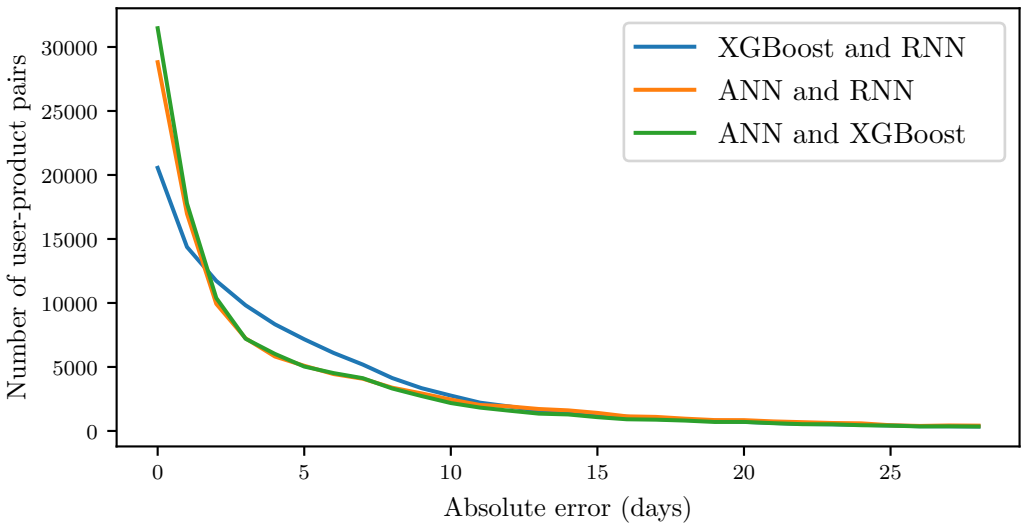



Figure 7.7: Number of user-product pairs predicted per absolute error in days

approaches, the combination between the ANN model and the XGBoost model performs the best, which can be expected as these two models also perform the best on their own.

Thus, none of the combination approaches will be chosen as the NPD Predictor, as a single model approach, the ANN, outperforms all the combination approaches. This means that the ANN model is chosen as the NPD Predictor. This model conforms to the first two requirements set for the NPD Predictor in Chapter 4, as the NPD Predictions are made for a user-product pair and the predictions are made in days to next purchase. The third requirement will be addressed in the next chapter.

## 7.3 Conclusion: Chapter 7

In this chapter all the models developed were tested on a test dataset. The model that performs the best was identified, after trying a combination model approach to see if the NPD Predictor should be modelled with a combination of models. The ANN model is chosen as the NPD Predictor as it can predict 31.8% of the user-product pairs in the test set with an error of less than one day. This chapter also includes an analysis of which products are mostly predicted with an error of less than one day. The ANN model can now be used as the NPD Predictor to personalise marketing by proposing advertisements to the customer for the specific products that they will need, by the predicted date.

# Chapter 8

# Application of the NPD Predictor

In the previous chapter it was shown that the Artificial Neural Network (ANN) model is to be chosen as the NPD Predictor. In this chapter, the proposed application of the NPD Predictor will be demonstrated. This corresponds to the first part of the deployment step of the CRISP-DM cycle, as it will present the knowledge gained through the model and will be presented for possible use to a customer. As explained in Chapter 1, a retailer can use the NPD Predictor to market to their customers and guidelines will now be provided.

Figure 8.1 shows the proposed structure in which the NPD Predictor is to be used. In this chapter, the elements of this structure will be explained to show how they can support the NPD Predictor for use in industry. The structure uses retail data and segments the customers so that specific marketing strategies can be used for the different customer segments. This also reduces the processing time for the NPD Predictor. Once the most valuable customer segments have been identified the NPD Predictor can predict the next purchase date for a customer. This next purchase date can then be used to generate individual advertisements and create up-selling and cross-selling opportunities.

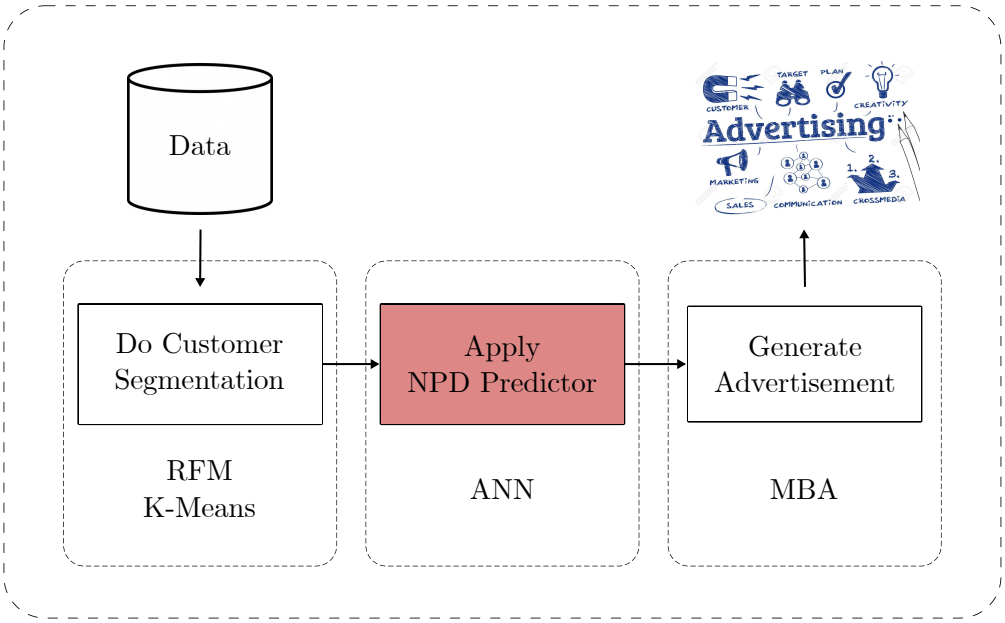



Figure 8.1: Proposed analysis structure for application of the NPD Predictor

Each element of the proposed structure requires some analysis steps, and these will be discussed in the sections that follow. The dataset discussed in Section 4.5 will be used for this discussion.

## 8.1 Customer Segmentation

As discussed in Section 4.5, there are over 3 million transactions in the online FMCG dataset and it contains over 200 000 customers. To reduce the number of customers in the dataset, customer segmentation is performed on it to identify customers that will be the most valuable to target with individual marketing strategies. Recency, Frequency and Monetary (RFM) is a data mining technique that attempts to identify the most profit-generating customers as explained in Section 2.2.4, and will be used along with a clustering technique to segment the customers.

### 8.1.1 Generating RFM Features

The first step in customer segmentation is to generate RFM features for the dataset. These features will then be used to segment the customers by using a clustering technique based on their RFM scores. Each customer has RFM features, explained as follows:

**R** The average days between orders per customer, measured in days.

**F** The total number of purchases that a customer has made; a count measure.

**M** The average number of products that a customer purchased, measured in number of products.

The dataset does not contain a unit price or a timestamp. Thus, it is assumed that the monetary value of a customer can be derived from the average number of products in their cart. This assumption is based on the monetary value of a customer growing with each product they purchase. The recency feature is created based on the assumption that the average time between purchases shows some evidence of how recently the customer made a purchase. For a dataset with a unit price, the monetary value for a customer can be calculated by the average of the total price per purchase. If a dataset has a timestamp per purchase, the recency feature can be calculated as the time that has elapsed since the most recent transaction that the customer made.

The RFM features were created for all customers in the dataset. The distributions for the three features can be seen in Figure 8.2. To cluster the customers based on their RFM scores, a hierarchical clustering method, K-Means will be used. This is a suitable clustering technique as it is good with big datasets, as explained in Table 3.6.

As seen in the figure, both the Frequency and Monetary features are positively skewed. K-Means clustering performs best when the features are not skewed. Thus, to change this, the logarithm of the two features was taken, and less skewed distributions were obtained, as shown in Figure 8.3.

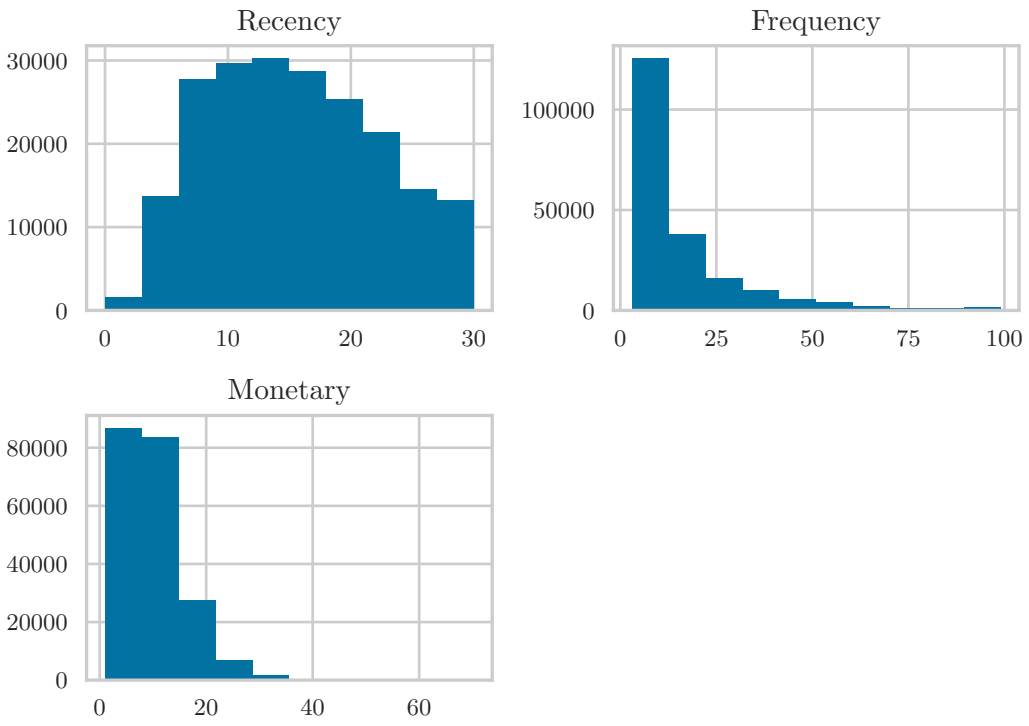



Figure 8.2: The distributions of the RFM features created from all the customers in the dataset

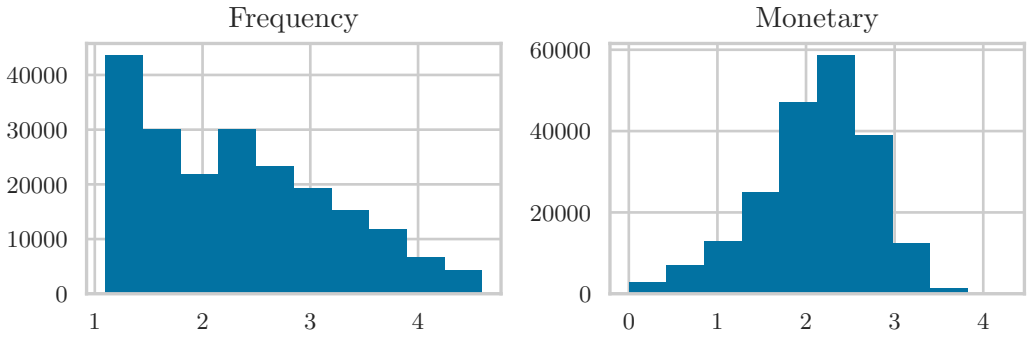



Figure 8.3: Recency and Monetary distributions log transformed

In addition to the skewness, the K-Means algorithm also performs best when the features are scaled to have a mean of approximately zero and a standard deviation of approximately one. The RFM features were scaled to fit the mentioned criteria. A summary of the features is given in Table 8.1 and it can be seen that the means are centred around zero and the standard deviations are one.

Table 8.1: Summary of RFM features after scaling

| | Recency | Frequency | Monetary |
|---|---|---|---|
| count | 206 209 | 206 209 | 206 209 |
| mean | -0.00 | 0.00 | 0.00 |
| std | 1.00 | 1.00 | 1.00 |
| min | -2.23 | -1.40 | -3.29 |
| 25% | -0.81 | -0.82 | -0.57 |
| 50% | -0.06 | -0.15 | 0.12 |
| 75% | 0.75 | 0.70 | 0.70 |
| max | 2.10 | 2.57 | 3.32 |

This completes the preparation for clustering, which can now be performed.

### 8.1.2 Clustering the customers based on their RFM scores

As mentioned, the K-Means algorithm will be used to cluster the customers into segments. This technique is used to find a fixed number of clusters $k$ in the data. The K-Means clustering technique uses $k$ centroids (the point at the centre of a cluster), and assigns all the data points to the cluster with the closest centroid. Firstly the K-Means algorithm defines $k$ centroids, randomly. After this is initialised, two steps are iteratively performed:

**Step 1:** Assign each data point to the centroid that has the closest euclidean distance to the point.

**Step 2:** For each of the $k$ centroids, calculate the mean of the values of all the points that belong to the centroid. This calculated mean becomes the new value of the centroid.

These two steps are repeated until there is no change in the mean value for the centroids, which means that the data was correctly grouped. The process can also be stopped by specifying the number of iterations.

To determine the optimal number of clusters, $k$, the elbow method is used. The "elbow" or "knee of a curve" method is often used to determine a cut-off point for selecting the number of clusters. This heuristic is used to choose a point where diminishing returns are not worth the additional cost. For this method, the K-Means model is fitted with a range of values for $k$. The elbow method was implemented, and this can be seen in Figure 8.4. This figure shows the distortion score over the number of clusters, with the clusters ranging from 2 to 14. The distortion score should be minimised as it calculates the average of the squared distances from the cluster centres of the respective clusters, using the euclidean distance metric. The

intuition is that when increasing the number of clusters, the fit will improve since there are more clusters to use, which naturally at some point will overfit the data, which is reflected by the elbow. Figure 8.4 also shows the time that it takes to cluster the data into $k$ clusters. The optimal number of clusters, as seen from the plot, is six.

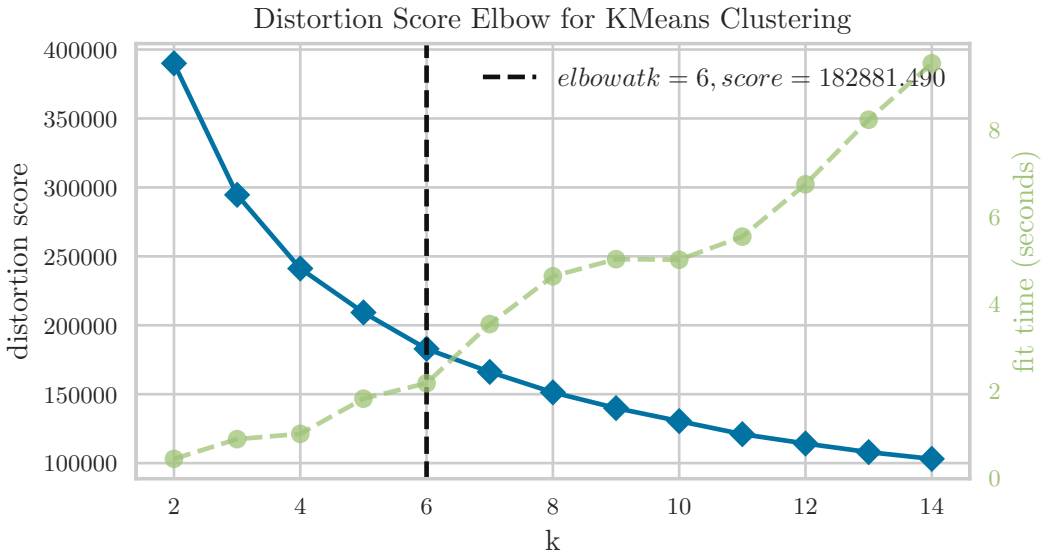



Figure 8.4: Distortion score elbow for K-Means clustering

The customers were clustered into six clusters, based on their RFM scores. Figure 8.5 shows the recency, frequency and monetary values on three axes, with the colour representing the cluster of the data point.

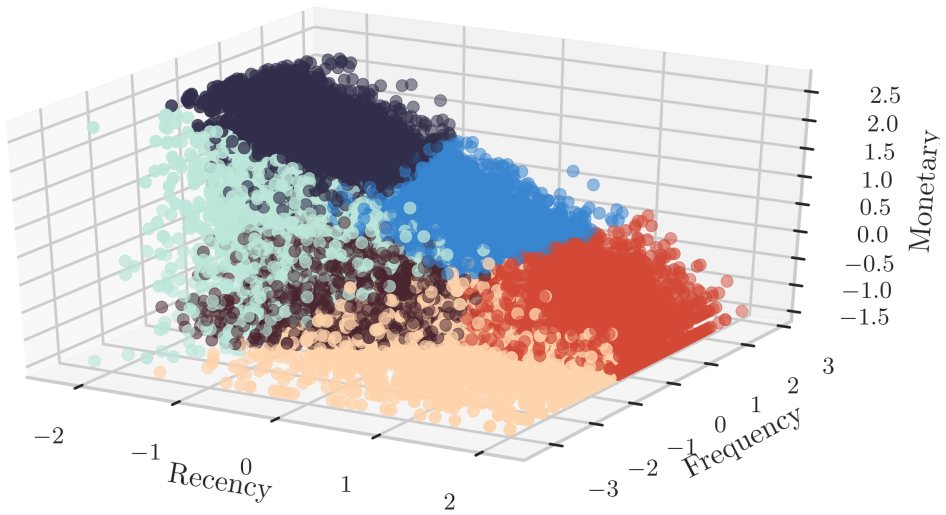



Figure 8.5: Data Clustered based on Recency, Frequency and Monetary values

Figure 8.6 shows the average recency, frequency and monetary score for all the clusters. These can be used to identify different marketing techniques for different segments.

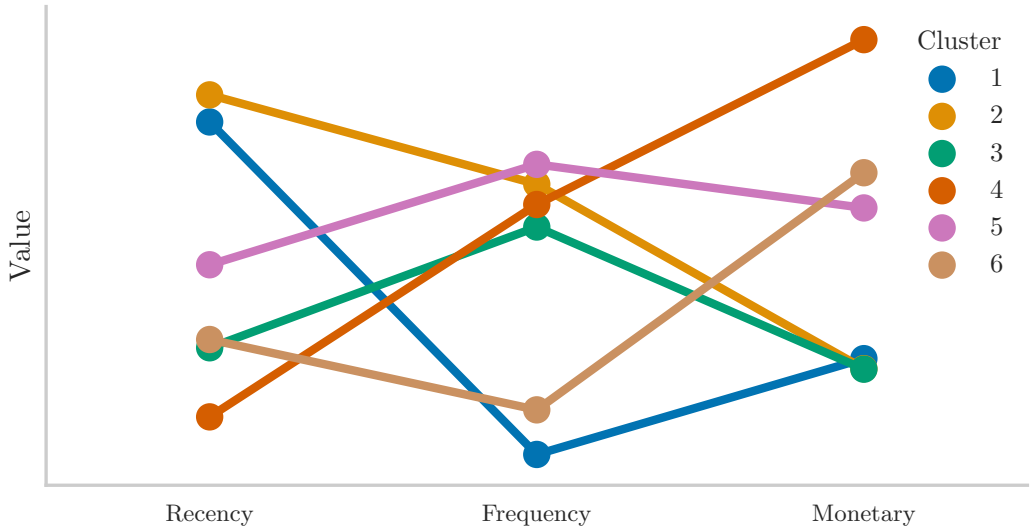



Figure 8.6: Average RFM values per cluster

Different marketing strategies can be implemented for the different clusters as members of these clusters have different needs and exhibit different buying behaviour. For example, a cluster with a good recency score, good frequency score and a bad monetary score can be targeted to increase their monetary value. The NPD Predictor, along with market basket analysis, can be used to create cross-selling opportunities to these customers. From the way that the recency, frequency and monetary values were calculated, *low recency, high frequency and high monetary values* are desirable. The different clusters have the following potential marketing opportunities:

**Cluster 1:** High recency, low frequency and mid-low monetary values. These customers do not use the store often, only purchased a small number of items and have not been to the store recently. These customers have a lot of room for improvement and marketing to these customers could improve the frequency and recency of their usage.

**Cluster 2:** High recency, mid-range frequency and low monetary values. These customers do not use the store often. These customers are more or less the same as those in Cluster 1 but they make smaller purchases. Advertising to these customers can improve their recency and frequency.

**Cluster 3:** Low recency, mid-range frequency and low monetary values. These customers used the store recently, have a fairly good frequency score, but make small purchases. Using

cross-selling or up-selling advertisements on these customers can improve the monetary value of these customers.

**Cluster 4:** Low recency, high frequency and high monetary value. These are good customers as they make frequent purchases with high monetary value and were at the store recently. These customers can be targeted with up-sell products to further increase their monetary value.

**Cluster 5:** Mid-range recency, mid to high frequency with mid-range monetary value. These customers use the store fairly often and make reasonable sized purchases. When marketing to these customers it can be to focus on decreasing their recency value.

**Cluster 6:** Low recency, low frequency and high monetary values. These customers have tried the store recently, and do not make frequent purchases, but when they do the purchases have high monetary value. When marketing to these customers, it must be focused on improving their frequency score.

From these observations, the cluster chosen to use in the next step of the proposed environment is Cluster 3. These customers should be encouraged to make more frequent purchases with higher monetary value. This can be done by marketing to these customers, as they will be reminded to use the company. Up-sell and cross-sell opportunities can also be presented to these customers to increase their monetary value. Predicting their Next Purchase Date for the user-product pairs will help to generate individualised advertisements for these customers. The Next Purchase Date for the user-product pairs from the users in this cluster will be predicted and explained in the next section. It is important to note that when using a different dataset the "elbow" point will be different and the customer segments will have to be evaluated based on their average RFM values per cluster.

## 8.2 Application of the NPD Predictor

Now that Cluster 3 has been identified as the targeted cluster, the NPD Predictor will be used to predict the Next Purchase Date for each user-product pair for the customers in this cluster.

The dataset that will be used to train the NPD Predictor to predict the Next Purchase Date is the same as the non-sequence-based dataset explained in Chapter 5, but for all the customers in Cluster 3. A sample of the dataset can be seen in Table 8.2.

Table 8.2: Sample of the dataset for the chosen cluster

| user_ id | product_ id | max_ days | min_ days | avg_ days | days_ since | variance | t_5 | t_4 | t_3 | t_2 | t_1 | t_0 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $f_1$ | $f_2$ | $f_3$ | $f_4$ | $f_5$ | $f_6$ | $f_7$ | $f_8$ | $f_9$ | $f_{10}$ | $f_{11}$ | $f_{12}$ | |
| 17 | 7350 | 30 | 0 | 8.16 | 8 | 54.47 | 5 | 4 | 3 | 5 | 4 | 8 |
| 17 | 7350 | 30 | 0 | 9.22 | 9 | 66.95 | 5 | 6 | 0 | 9 | 6 | 9 |
| 17 | 7350 | 30 | 4 | 11.41 | 3 | 82.08 | 4 | 4 | 5 | 30 | 12 | 3 |
| 17 | 7350 | 30 | 4 | 11.00 | 16 | 77.67 | 7 | 4 | 9 | 30 | 5 | 16 |
| 17 | 17762 | 36 | 0 | 11.50 | 3 | 113.38 | 6 | 0 | 9 | 6 | 18 | 3 |
| 17 | 17762 | 36 | 3 | 14.00 | 5 | 146.80 | 20 | 9 | 36 | 6 | 3 | 5 |
| 17 | 18534 | 39 | 3 | 9.71 | 6 | 78.20 | 9 | 3 | 5 | 4 | 8 | 6 |
| 17 | 18534 | 39 | 4 | 11.07 | 9 | 100.73 | 15 | 5 | 6 | 9 | 6 | 9 |
| 17 | 18534 | 30 | 4 | 9.56 | 30 | 66.91 | 5 | 30 | 5 | 16 | 4 | 39 |
| 50 | 5612 | 22 | 0 | 8.96 | 4 | 25.88 | 5 | 13 | 7 | 13 | 18 | 9 |
| 50 | 5612 | 18 | 0 | 7.68 | 8 | 16.43 | 11 | 11 | 9 | 6 | 0 | 22 |
| 50 | 5612 | 18 | 2 | 7.77 | 4 | 17.41 | 4 | 2 | 6 | 6 | 14 | 8 |
| 50 | 5612 | 18 | 5 | 8.71 | 3 | 18.20 | 10 | 5 | 5 | 6 | 7 | 8 |
| 50 | 6182 | 33 | 0 | 6.89 | 7 | 37.92 | 6 | 5 | 4 | 3 | 3 | 7 |
| 50 | 6182 | 33 | 0 | 7.26 | 5 | 42.61 | 4 | 3 | 10 | 11 | 11 | 6 |

The ANN trained to predict the Next Purchase Date can be seen in Figure 8.7. Here it can be seen that the ANN has 12 input features with two hidden layers consisting of 10 and five neurons respectively and one output layer, which is the NPD prediction. The 12 features are the features $f_1 - f_{12}$ shown in Table 8.2, with t_0 used to train the model. The relu activation function was used for all the layers. For backpropagation, an MSE loss function was used with an Adam optimiser. The batch size was set to 100.

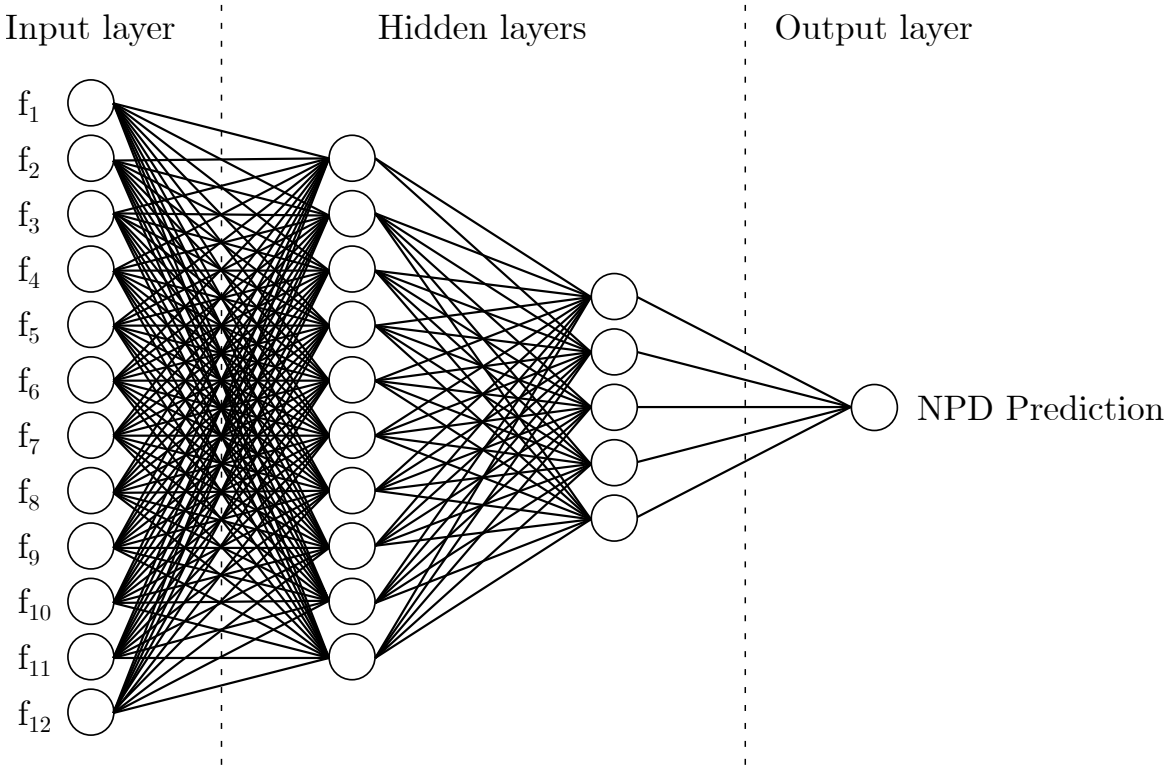



Figure 8.7: ANN implementation

The predictions for some of the user-product pairs for Cluster 3 can be seen in Table 8.3. More predictions can be seen in Appendix B. In this table, it can be seen that user 17 will need products 7 350 and 18 534 approximately at the same time, in 8-9 days. This prediction can therefore be used to send an advertisement to this customer. The same can be said for user 210, for products 4 799, 4 920 and 10 292, in approximately eight days and products 23 909 and 24 852 in approximately 4-5 days. The process of generating advertisements will be explained in the next section.

Table 8.3: Predictions for Cluster 3

| user_id | product_id | NPD prediction | user_id | product_id | NPD prediction |
|---|---|---|---|---|---|
| 17 | 7350 | 8.39 | 210 | 23909 | 4.33 |
| 17 | 17762 | 10.41 | 210 | 24852 | 4.76 |
| 17 | 18534 | 8.98 | 210 | 29388 | 10.66 |
| 50 | 5612 | 9.53 | 210 | 32030 | 6.38 |
| 50 | 6182 | 8.18 | 210 | 33043 | 10.76 |
| 50 | 23165 | 9.96 | 210 | 35547 | 6.26 |
| 140 | 432 | 10.15 | 210 | 40198 | 7.02 |
| 140 | 5973 | 3.90 | 210 | 41065 | 7.04 |
| 140 | 19057 | 13.10 | 210 | 42585 | 10.74 |
| 140 | 19894 | 8.73 | 210 | 47209 | 12.79 |
| 140 | 21616 | 13.75 | 223 | 5250 | 12.39 |
| 140 | 34262 | 13.86 | 223 | 8309 | 11.88 |
| 140 | 36865 | 12.74 | 223 | 27086 | 13.48 |
| 140 | 47610 | 11.28 | 223 | 40396 | 11.19 |
| 187 | 29436 | 4.76 | 236 | 19660 | 8.68 |
| 210 | 4799 | 8.57 | 243 | 27104 | 18.18 |
| 210 | 4920 | 8.85 | 290 | 23405 | 10.11 |
| 210 | 10292 | 8.77 | 290 | 30251 | 12.58 |
| 210 | 21903 | 9.19 | 290 | 31981 | 13.00 |
| 210 | 23375 | 9.56 | 290 | 34355 | 8.95 |

## 8.3 Generating Individualised Advertisements

Having segmented the customer dataset and after application of the NPD Predictor, individualised advertisements can now be created, which is the ultimate objective of the proposed analysis structure. For cross-sell and up-sell opportunities explained in Section 2.2.4.4, it is necessary to do further analysis on items frequently purchased together. For cross-selling, products that are identified as frequently purchased together can be suggested to a customer. For up-selling, the unit price of a product is important as a similar product with higher value should be identified. The online FMCG dataset used throughout this project does not have a unit price per product, but a company can run through its product list and search similar products with a higher value. The rest of this section will look at generating advertisements for cross-selling opportunities and suggest how up-sell opportunities can be made.

Association rule mining, or market basket analysis, discussed in Section 2.2.4.2, can be used to identify cross-selling opportunities. Market basket analysis will be performed on the entire

dataset described in Chapter 4, this will help to identify the products that all the customers frequently purchase together, so that these associations can be used on individual customers at the time predicted by the NPD Predictor.

After performing market basket analysis on the entire dataset, the NPD Predictions for the user-product pair can be associated with products frequently purchased along with these products. This can be used for cross-selling opportunities. The associations are shown in Table 8.4, which gives the user_id, product_id, NPD Prediction and the products that were identified as associations through market basket analysis.

Table 8.4: Associations that can be made with the predicted products

| user_id | product_id | NPD prediction | Associated product_ids |
|---|---|---|---|
| 17 | 7350 | 8.39 | [7026 24852] |
| 17 | 17762 | 10.41 | None |
| 17 | 18534 | 8.98 | None |
| 50 | 5612 | 9.53 | [47626 22935 26209 47766 21903 47209 21137 24852 13176] |
| 50 | 6182 | 8.18 | [13176] |
| 50 | 23165 | 9.96 | [39812 9092 39984 34126 48745 47734 19048 10246 16759 18531 44359 17794 19678...] |
| 140 | 432 | 10.15 | [40545 1158 33120 20842 30450 24489 46676 28849 46802 9076 25890 41844 43961...] |
| 140 | 5973 | 3.90 | None |
| 140 | 19057 | 13.10 | [18090 46969 22825 8174 43122 11941 2825 19706 38777 1244 24221 38159 5134...] |
| 140 | 19894 | 8.73 | [20794 6287 30776 48559 45603 47209 21903 21137 13176 24852] |
| 140 | 21616 | 13.75 | [33401 41593 1468 40604 21405 248 12845 13984 43789 39984 15937 16145 1025...] |
| 140 | 34262 | 13.86 | [30489 40706 42265 44632 8277 5876 22935 21137 21903 24964 47626 45007 26209...] |
| 140 | 36865 | 12.74 | [13263 11448 8309 28465 24799 43662 26131 8490 14870 47912 30489 22035 45066...] |
| 140 | 47610 | 11.28 | None |
| 187 | 29436 | 4.76 | None |
| 210 | 4799 | 8.57 | [18370 29987 12206 2314 40604 38273 39619 32734 31343 35108 42828 7021 45535...] |

When replacing the product_id with the name, cross-selling opportunities are more intuitive. This can be seen in Table 8.5. For example product_id 5612 corresponds to "Reduced Fat Milk" and can be associated with the products seen in the table. Thus, a prediction is made that user 50 will purchase "Reduced Fat Milk" in approximately 9 to 10 days, "Natural Artesian Bottled Water" in approximately 8 days and "Organic Leek" in approximately 10 days. Thus, an individual advertisement can be sent to this user, offering cross-selling opportunities. These opportunities can be identified from the Associated products column in Table 8.5.

Table 8.5: Associations with product names

| user_id | product_id | Product name | Associated products |
|---|---|---|---|
| 17 | 7350 | Natural Lime Flavor Sparkling Mineral Water | Sparkling Mineral Water, Natural Lemon Flavor<br>Banana |
| 50 | 5612 | Reduced Fat Milk | Large Lemon<br>Organic Yellow Onion<br>Limes<br>Organic Avocado<br>Organic Baby Spinach<br>Organic Hass Avocado<br>Organic Strawberries<br>Banana<br>Bag of Organic Bananas |
| 50 | 6182 | Natural Artesian Bottled Water | Bag of Organic Bananas |
| 50 | 23165 | Organic Leek | Organic Thyme<br>Organic Vegetable Broth<br>Organic Dill<br>Organic Italian Parsley Bunch<br>Globe Eggplant |

For an up-selling opportunity for user 50 a more valuable product than "Reduced fat milk" can be identified by searching for a product that is more or less the same but has a higher monetary value. A list containing a search for organic milk in the original dataset can be seen in Table 8.6. If the dataset contained unit prices, then it would be easy to identify a product with a higher value. But say for example "Reduced Fat Milk" has a unit price of $4-00 and "0% Fat Free Organic Milk" has a unit price of $6-00, then selling the "0% Fat Free Organic

Milk" with 20 % discount will result in the customer paying $4.80 for the milk. This means that the customer feels that they are saving but are still paying more than they would have for their original choice of milk, namely $4-00.

Table 8.6: Up-sell oportunities for "Reduced Fat Milk".

| product_id | product_name |
|---|---|
| 1463 | Organic Milk |
| 6533 | 1% Lowfat Organic Milk |
| 11821 | Organic Milk Whole |
| 13166 | Organic Milk Reduced Fat, 2% Milkfat |
| 16234 | Vitamin D Organic Milk |
| 45257 | Lactose-Free Vitamin D Organic Milk |
| 47983 | Organic Milkmaid Tea |
| 49517 | 0% Fat Free Organic Milk |

An example of an individualised advertisement for user 50 can be seen in Figure 8.8. Both an up-selling and a cross-selling opportunity are presented. This advertisement can be sent to this user in eight days' time, as the prediction was made that this user will purchase milk, water and organic leek in the next 8-9 days.


(a) Cross-selling offer


(b) Up-selling offer

Figure 8.8: Individual advertisements for user 50, proposed to send in 8 to 9 days

These advertisements were created using reasoning and can be automated in a future study, to generate advertisements for all the users based on the predicted NPDs for their products.

This successfully illustrates the use of the NPD Predictor to generate individualised advertisements. This shows that the NPD Predictor conforms to the third requirement for the NPD Predictor set in Chapter 4, as the NPD Prediction can be used to generate individualised advertisements.

The techniques used to cluster the entire dataset as well as the market basket analysis are for illustrative purposes and can be refined with a deeper study of these fields. These techniques were used to show how the NPD Predictions made by the NPD Predictor can be used to individualise marketing. Note that the conclusions in the analysis depend on the dataset provided, and for a different dataset, a data scientist will need to find their target clusters through analysis and reasoning.

## 8.4 Conclusion: Chapter 8

In this chapter, the use of the NPD Predictor was demonstrated with other marketing tools, all contained in a structure which follows specified steps. The NPD Predictor was used to make predictions for each user-product pair in a cluster of customers formed with K-Means, and this cluster serves to provide a narrowed-down marketing target audience. Market basket analysis was performed on the entire dataset to get product associations for cross-selling opportunities. Up-selling opportunities could be presented if the unit price was available. An example of an individualised advertisement was presented by using the associations along with the NPD Prediction.

# Chapter 9

# Summary, Conclusion and Recommendations

In this chapter a project summary will be given, explaining what was accomplished through the project, based on the knowledge gained through the literature study and the development of the Next Purchase Date (NPD) Predictor. The summary includes a discussion of how the objectives that were set in Chapter 1 were met. Recommendations for future work will be listed and a personal reflection will be given.

## 9.1 Summary of the Project

The project started with a Research proposal in Chapter 1, which gave background information on the problem, stating the objectives that were to be met through the project and gave a methodology of how the project would be executed.

In Chapter 2 and Chapter 3, the first objective was met by conducting a literature study on Customer Behaviour Management, Marketing Strategies, Data Analytics, Machine Learning techniques and how Machine Learning can be used to predict future events. This established a good background for the project and was followed by selecting an appropriate dataset for the problem along with giving requirements for the NPD Predictor in Chapter 4. Furthermore, in this chapter, the dataset was explored, and visual insight into the dataset was provided. This gave a good understanding of the dataset so that the NPD Predictor could be designed and developed. The CRISP-DM and SEMMA data analytics processes discussed in Chapter 3 were followed to design and develop the NPD Predictor.

The second objective, to design and develop an NPD Predictor, was achieved in Chapters 5 - 6. In Chapter 5, the dataset was prepared to isolate the target variable, the Next Purchase Date. Features that describe the target variable were also created and a training set and a test set were created for both sequence-based and non-sequence-based approaches. In Chapter 6 the datasets created in Chapter 5 were used to build machine learning models. The sequence-based models that were built were Recurrent Neural Network models with both one and two features as well as a linear regression model. The non-sequence-based models that were created were an XGBoost model and an Artificial Neural Network model. These models were used to predict the Next Purchase Date for a customer-product pair in the test set. Chapter 7 evaluated and communicated the results of the machine learning model predictions, which achieved the third and fourth objectives. The chapter also explored whether or not combinations of these models could increase the accuracy of the predictions. The ANN model was chosen as the

NPD Predictor as it predicted the Next Purchase Date of a user-product pair with an absolute error of less than one day for 31.8% of the test dataset and an absolute error of fewer than three days for 55% of the test dataset. The NPD Predictor successfully predicted the Next Purchase Date for a user-product pair, made the prediction in days to next purchase and can be used to generate individualised advertisements, achieving the requirements that were set for the NPD Predictor in Chapter 4.

In Chapter 8, an application of the NPD Predictor was demonstrated. The chapter showed how the online FMCG dataset can be clustered to reduced pre-processing, how the NPD Prediction can be made for a chosen cluster and how these predictions can then be used along with market basket analysis to generate individualised advertisements. This chapter showed that the NPD Predictor can be used to personalise marketing strategies.

An article that reports on this work was published illustrating the development of an NPD Predictor (Droomer & Bekker, 2020).

## 9.2 Key Findings

Machine learning can be used to predict the Next Purchase Date of a customer-product pair. Some key findings and observations of developing a NPD Predictor are as follows:

- Preparing the dataset and isolating the target variable takes much longer than training a machine learning model.
- Training the non-sequence-based models was less computationally expensive than training the sequence-based models, as it was only necessary to train one model, instead of a model for each user-product pair sequence.
- Further to the above point the non-sequence-based models performed much better as they took into account all the user-product pairs' data and not just the data of one user-product pair sequence.

## 9.3 Recommendations and Future work

This study identifies opportunities for further work that will further the research, including:

i Extending the analysis to other datasets.

ii Deployment of the NPD Predictor at a retail chain using the ANN algorithm.

iii Explore different ways to combine the models developed in Chapter 6.

iv Refine the marketing strategies proposed in Chapter 8, to target specific users, using the NPD Predictor.

v Generating NPD Predictions in real time and automating the individualised advertisements, including developing a graphical user interface. This will make the analysis and reasoning required faster and easier for the data scientist.

## 9.4 Personal Reflection

Throughout this project various skills were developed. A few lessons were learnt through the development of the NPD Predictor. These lessons include academic and personal lessons:

- Following the data analytics processes is very important. It is better to understand the problem and the data before diving in and trying to solve the problem.
- To develop a tool takes trial and error, which can be time-consuming.
- Machine learning is a very big field and it takes a lot of time to understand some of it.
- Self-isolating while conducting this research did not pose a problem.

Resulting from these lessons, analytical and problem-solving skills were improved and developed. Through this project interest in the fields of marketing and machine learning were deepened.

Data is the new oil, and hence a resource that must be carefully utilised and managed. It is thus natural for industrial engineers to develop building blocks of systems to support business operations and decision-making with machine learning. In this study, it was shown how machine learning could be used to predict the Next Purchase Date for an individual retail customer. A retailer can use the results to develop a system that can offer individual customers personalised discounts on specific products, on specific days, as illustrated in Chapter 8. This concept differs from the typical loyalty programmes available since the offers are personalised based on the individual customer's buying history, and the discount offers are thus personalised.

## 9.5 Concluding remarks

Predicting a user-product pair's Next Purchase Date is possible by using machine learning. It can be seen through this project that using a technique that takes all the data into account including data from other user-product pairs performs better than the approach of predicting the NPD by only using the data of one user-product pair. The ANN outperforms the

other algorithms, with XGBoost performing second best; better than all the sequence-based approaches.

# References

Abbot, D. A. A. (2012). Inductive Business-Rule Discovery in Text Mining. Available: http://abbottanalytics.com/index.php [15 November 2019]. 53

Abdou, H. & Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria: a review of literature. *Intelligent Systems in Accounting, Finance & Management*, 59–88. 45

Adomavicius, G. & Tuzhilin, A. (2001). Using data mining methods to build customer profiles. *Computer*, *34*(2), 74–82. 28, 30, 31

Alderden, M. A. & Lavery, T. A. (2007). Predicting homicide clearances in Chicago: Investigating disparities in predictors across different types of homicide. *Homicide Studies*, *11*(2), 115–132. 53

Alelyani, S., Tang, J., & Liu, H. (2018). Feature selection for clustering: A review. In *Data Clustering* (pp. 29–60). Chapman and Hall/CRC. 10

An, J., Kwak, H., & Jansen, B. J. (2017). Automatic generation of personas using youtube social media data. In *Proceedings of the 50th Hawaii International Conference on System Sciences (HICSS)*. 30

Andersen, H., Andreasen, M., & Jacobsen, P. (1999). The CRM handbook: From group to multi-individual. *Norhaven: PricewaterhouseCoopers*. 17

Asur, S. & Huberman, B. A. (2010). Predicting the future with social media. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01*, (pp. 492–499). IEEE Computer Society. 52

Au, W.-H. & Chan, K. C. (2003). Mining fuzzy association rules in a bank-account database. *IEEE Transactions on Fuzzy Systems*, *11*(2), 238–248. 10

Äyrämö, S. & Kärkkäinen, T. (2006). Introduction to partitioning-based clustering methods with a robust example. *Reports of the Department of Mathematical Information Technology. Series C, Software engineering and computational intelligence*, (1/2006). 45

Azevedo, A. & Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: A parallel overview. Technical report. 34, 35, 36, 37, 38

Babs, T. (2018). The Mathematics of Neural Networks. Available: https://medium.com/coinmonks/the-mathematics-of-neural-network-60a112dd3e05 [28 July 2020]. 59

Bacila, M.-F., Radulescu, A., & Marar, I. L. (2012). RFM based segmentation: An analysis of a telecom company's customers. In *The Proceedings of the International Conference "Marketing-from Information to Decision"*, (pp.52̃). Babes Bolyai University. 24

Baecke, P. & Van den Poel, D. (2011). Data augmentation by predicting spending pleasure using commercially available external data. *Journal of Intelligent Information Systems*, *36*(3), 367–383. 24

Barber, R. & Sharkey, M. (2012). Course correction: Using analytics to predict course success. In *Proceedings of the 2nd international conference on learning analytics and knowledge*, (pp. 259–262). ACM. 54

BBC (2016). Artificial intelligence: Go master Lee Sedol wins against AlphaGo program. Available: https://www.bbc.com/news/technology-35797102 [23 April 2019]. 39

BBC Academy (2019). The history of machine learning. Available: https://www.bbc.co.uk/teach/ai-15-key-moments-in-the-story-of-artificial-intelligence/zh77cqt [9 April 2019]. 40

Beck, A. H., Sangoi, A. R., Leung, S., Marinelli, R. J., Nielsen, T. O., van de Vijver, M. J., West, R. B., van de Rijn, M., & Koller, D. (2011). Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Science Translational Medicine*, *3*(108), 108–113. 52

Berkhin, P. (2006). A survey of clustering data mining techniques. In *Grouping multidimensional data* (pp. 25–71). Springer. 45

Berry, M. J. A. & Linoff, G. (2004). *Data mining techniques : for marketing, sales, and customer relationship management*. John Wiley & Sons. 2

Borden, N. H. (1964). The concept of the marketing mix. *Journal of advertising research*, *4*(2), 2–7. 15

Bounsaythip, C. & Rinta-Runsala, E. (2001). Overview of data mining for customer behavior modeling. *VTT Information Technology Research Report, Version*, *1*, 1–53. 17, 25, 26, 28, 31, 45, 46

BusinessDictionary.com (2019a). Mass marketing. Available: http://www.businessdictionary.com/definition/mass-marketing.html [19 June 2019]. 16

BusinessDictionary.com (2019b). What is marketing? Definition and meaning. Available: http://www.businessdictionary.com/definition/marketing.html [10 June 2019]. 14

CareerBuilder.com (2012). Predict Which Jobs Users Will Apply To. Available: https://www.kaggle.com/c/job-recommendation [23 August 2019]. 53

Caruana, R. & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, (pp. 161–168)., New York, NY, USA. ACM. 45

Changchien, S. W., Lee, C.-F., & Hsu, Y.-J. (2004). On-line personalized sales promotion in electronic commerce. *Expert Systems with Applications*, *27*(1), 35–52. 21

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. Technical report, DaimlerChrysler, SPSS, OHRA and NCR. 36, 37

Chatterjee, S. & Hadi, A. S. (2015). *Regression analysis by example*. John Wiley & Sons. 45

Chen, I. J. & Popovich, K. (2003). Understanding customer relationship management (CRM) People, Process and Technology. *Business process management journal*, *9*(5), 672–688. 13

Chen, T. & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International conference on knowledge discovery and data mining*, (pp. 785–794). 68

Chen, Y. L., Tang, K., Shen, R.-J., & Hu, Y.-H. (2005). Market basket analysis in a multiple store environment. *Decision support systems*, *40*(2), 339–354. 11, 22

Corrigan, H. B., Craciun, G., & Powell, A. M. (2014). How does Target know so much about its customers? Utilizing customer analytics to make marketing decisions. *Marketing Education Review*, *24*(2), 159–166. 30

Cumby, C., Fano, A., Ghani, R., & Krema, M. (2004). Predicting customer shopping lists from point-of-sale purchase data. In *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '04*, (pp. 402)., Seattle, Washington, USA. ACM Press. 3

Cumby, C., Fano, A., Ghani, R., & Krema, M. (2005). Building intelligent shopping assistants using individual consumer models. In *Proceedings of the 10th international conference on Intelligent user interfaces*, (pp. 323)., San Diego, California, USA. 2

Dean, J. (2014). *Big data, data mining and machine learning: value creation for business leaders and practitioners*. Wiley. 21, 23, 44, 45

DeBevois, K. (2008). 'Harbor Sweets' Billie Phillips on Driving Off-season Sales. Available: https://www.targetmarketingmag.com/article/harbor-sweets-billie-phillips-driving-off-season-sales-111219/all/ [3 June 2019]. 52

Delen, D. (2011). Predicting student attrition with data mining methods. *Journal of College Student Retention: Research, Theory & Practice*, *13*(1), 17–35. 54

Deyan, G. (2019). 67+ Revealing Statistics about Smarthpone Usage in 2019 - Techjury. Available: https://techjury.net/stats-about/smartphone-usage/ [10 June 2019]. 14

Dickson, P. R. & Ginter, J. L. (1987). Market segmentation, product differentiation, and marketing strategy. *Journal of Marketing*, *51*(2), 1–10. 16

Dolnicar, S., Grün, B., & Leisch, F. (2018). Market segmentation. In *Market Segmentation Analysis* (pp. 3–9). Springer. 28

Doshi, S. (2019). Various Optimization Algorithms For Training Neural Network. Available: https://towardsdatascience.com/optimizers-for-training-neural-network-59450d71caf6 [ 8 June 2020]. 60, 63

Droomer, M. & Bekker, J. (2020). Using machine learning to predict the next purchase date for an individual retail customer. *The South African Journal of Industrial Engineering*, *31*(3), 69–82. 121

Dua, D. & Graff, C. (2017). UCI machine learning repository. Available: https://archive.ics.uci.edu/ml/datasets/Online+Retail+II [28 July 2020]. 74

Duda, R. O., Hart, P. E., & Stork, D. G. (2012). *Pattern classification*. John Wiley & Sons. 56

Dursun, A. & Caber, M. (2016). Using data mining techniques for profiling profitable hotel customers: An application of RFM analysis. *Tourism management perspectives*, *18*, 153–160. 22, 24

Dyche, J. (2002). *The CRM handbook: A business guide to customer relationship management*. Addison-Wesley Professional. 21, 26

Els, Z. (2019). Development of a data analytics-driven information system for instant, temporary personalised discount offers. Master's thesis, Stellenbosch University. 3

Elsner, R., Krafft, M., & Huchzermeier, A. (2003). Optimizing Rhenania's mail-order business through dynamic multilevel modeling (DMLM). *Interfaces*, *33*(1), 50–66. 24

Erevelles, S., Fukawa, N., & Swayne, L. (2016). Big Data consumer analytics and the transformation of marketing. *Journal of Business Research*, *69*(2), 897–904. 33

Erl, T., Khattak, W., & Buhler, P. (2015). *Big Data Fundamentals*. Prentice Hall. 21, 32, 33

Esplin, M. S., Merrell, K., Goldenberg, R., Lai, Y., Iams, J. D., Mercer, B., Spong, C. Y., Miodovnik, M., Simhan, H. N., van Dorsten, P., et al. (2011). Proteomic identification of serum peptides predicting subsequent spontaneous preterm birth. *American journal of obstetrics and gynecology*, *204*(5), 391.e1 – 391.e8. 52

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, volume 96, (pp. 226–231). 45

Etzion, O., Fisher, A., & Wasserkrug, S. (2004). E-CLV: a modelling approach for customer lifetime evaluation in e-commerce domains, with an application and case study for online auctions. In *IEEE International Conference on e-Technology, e-Commerce and e-Service*, (pp. 149–156). IEEE. 11

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, *17*(3), 37–37. 34, 35

Ferrucci, D. A. (2012). Introduction to "This is Watson". *IBM Journal of Research and Development*, *56*(3.4), 1:1–1:15. 39

Fitzpatrick, M. (2001). Statistical analysis for direct marketers–in plain English. *Direct Marketing*, *64*(4), 54–56. 24

Gera, M. & Goel, S. (2015). Data Mining-Techniques, Methods and Algorithms: A Review on Tools and their Validity. Technical Report 18. 44, 45

Geron, T. (2013). Airbnb and the unstoppable rise of the share economy. *Forbes*, *11*(Feb). 54

Goi, C. L. (2009). A Review of Marketing Mix: 4Ps or More? *International Journal of Marketing Studies*, *1*(1). 15

Goldenberg, B. (2000). What is CRM? What is an e-customer? Why you need them now. In *Proceedings of DCI customer relationship management conference, Boston, MA*, (pp. 27–29). 10, 13

Guo, A. (2017). PCA (4) : LDA (Linear Discriminant Analysis). Available: https://algorithmsdatascience.quora.com/PCA-4-LDA-Linear-Discriminant-Analysis [16 May 2019]. 43

Gupta, D. S. (2017). Fundamentals of Deep Learning – Introduction to Recurrent Neural Networks. Available: https://www.analyticsvidhya.com/blog/2017/12/introduction-to-recurrent-neural-networks/ [12 June 2020]. 66

Gupta, S., Hanssens, D., Hardie, B., Kahn, W., Kumar, V., Lin, N., Ravishanker, N., & Sriram, S. (2006). Modeling customer lifetime value. *Journal of service research*, *9*(2), 139–155. 22, 23, 24

Hansen, C. (2019). Optimizers Explained - Adam, Momentum and Stochastic Gradient Descent. Available: https://mlfromscratch.com/optimizers-explained/{#}/ [12 June 2020]. 60

Hosseini, M. & Shabani, M. (2015). New approach to customer segmentation based on changes in customer value. *Journal of Marketing Analytics*, *3*(3), 110–121. 1

Hsieh, N.-C. (2004). An integrated data mining and behavioral scoring model for analyzing bank customers. *Expert systems with applications*, *27*(4), 623–633. 30

Hu, Y.-H. & Yeh, T.-W. (2014). Discovering valuable frequent patterns based on RFM analysis without customer identification information. *Knowledge-Based Systems*, *61*, 76–88. 22, 24

Instacart (2017). The Instacart Online Grocery Shopping Dataset 2017. Available: https://www.instacart.com/datasets/grocery-shopping-2017 [19 May 2020]. 74

Investopedia (2019). What is a Customer. Available: https://www.investopedia.com/terms/c/customer.asp [17 September 2019]. 9

Jansen, S. M. H. (2007). Customer segmentation and customer profiling for a mobile telecommunications company based on usage behavior. Master's thesis, University of Maastricht. 30, 45

Jiao, J. R., Zhang, Y., & Helander, M. (2006). A Kansei mining system for effective design. *Expert Systems with Applications*, *30*(4), 658–673. 22

Jolliffe, I. (2011). *Principal Component Analysis*, (pp. 1094–1096). Berlin, Heidelberg: Springer. 41

Kallier, S. M. (2017). The influence of real-time marketing campaigns of retailers on consumer purchase behavior. *International Review of Management and Marketing*, *7*(3), 126–133. 21

Kamber, M., Han, J., & Pei, J. (2012). *Data Mining Concepts and Techniques.* (3rd ed.). Morgan Kaufmann Publishers. 21, 25, 45

Karp, A. H. (1999). Using logistic regression to predict customer retention. Technical report. 45

Kasturi, S. N. (2019). Boosting and Bagging explained with examples. Available: https://medium.com/swlh/boosting-and-bagging-explained-with-examples-5353a36eb78d [11 July 2020]. 69

Kaur, M. & Kang, S. (2016). Market Basket Analysis: Identify the changing trends of market data using association rule mining. *Procedia computer science*, *85*, 78–85. 24

Kelleher, J. D., Mac Namee, B., & D'arcy, A. (2015). *Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies.* MIT Press. 51, 54

Khodakarami, F. & Chan, Y. E. (2014). Exploring the role of customer relationship management (CRM) systems in customer knowledge creation. *Information & Management*, *51*(1), 27–42. 21

Kingma, D. P. & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980.* 64

Kluemper, D. H., Rosen, P. A., & Mossholder, K. W. (2012). Social Networking Websites, Personality Ratings, and the Organizational Context: More Than Meets the Eye? *Journal of Applied Social Psychology*, *42*(5), 1143–1172. 53

Kotler, P., Kotler, P., Armstrong, G., & Opresnik, M. (2017). *Principles of marketing.* Pearson. 14, 15, 16, 18

Kracklauer, A. H., Mills, D. Q., & Seifert, D. (2004). Customer management as the origin of collaborative customer relationship management. In *Collaborative Customer Relationship Management* (pp. 3–6). Springer. 10, 11

Krishna, G. J. & Ravi, V. (2016). Evolutionary computing applied to customer relationship management: A survey. *Engineering Applications of Artificial Intelligence*, *56*, 30–59. 22, 28

Kubiak, B. & Weichbroth, P. (2010). Cross- and Up-Selling Techniques in E-commerce Activities. *Journal of Internet Banking and Commerce*, *15*(3), 217–225. 26

Lanjewar, R. & Yadav, O. P. (2013). Understanding of Customer Profiling and Segmentation Using K-Means Clustering Method for Raipur Sahkari Dugdh Sangh Milk Products. *International Journal of Research in Computer and Communication Technology*, *2*(3), 103–107. 28

Larose, D. T. & Larose, C. D. (2014). *Discovering knowledge in data: an introduction to data mining*. John Wiley & Sons. 45

Lechevalier, D., Narayanan, A., & Rachuri, S. (2014). Towards a domain-specific framework for predictive analytics in manufacturing. In *2014 IEEE International Conference on Big Data (Big Data)*, (pp. 987–995). IEEE. 53

Levy, S. (2011). TED 2011: The 'Panda' that hates farms: A Q & A with Google's top search engineers. Available: https://www.wired.com/2011/03/the-panda-that-hates-farms/ [28 July 2019]. 52

Lipton, Z. C., Berkowitz, J., & Elkan, C. (2015). A critical review of recurrent neural networks for sequence learning. *arXiv preprint arXiv:1506.00019*. 67, 68

Lison, P. (2012). An introduction to machine learning. Technical report, University of Oslo. 43

Lu, J. (2002). Predicting Customer Churn in the Telecommunications Industry — An Application of Survival Analysis Modeling Using SAS. *Data Mining Techniques*, *27*, 1–6. 52

Maimon, O. & Rokach, L. (2010). *Data mining and knowledge discovery handbook*. Springer. 45

Mariscal, G., Marban, O., & Fernandez, C. (2010). A survey of data mining and knowledge discovery process models and methodologies. *The Knowledge Engineering Review*, *25*(2), 137–166. 35

McFadden, R. (2019). Lester Wunderman, Father of Direct Marketing, Dies at 98 - The New York Times. Available: https://www.nytimes.com/2019/01/14/business/lester-wunderman-dead.html[19 June 2019]. 16

Mikat, S., Weston, J., Scholkopft, B., & Mullert, K.-R. (1999). Fisher Discriminant Analysis with Kernels. Technical report. 41

Mitchell, R. L. (2011). The Art and Science of Fashion. Available: https://www.computerworld.com/article/2550035/enterprise-applications-the-art-science-of-fashion.html [16 May 2019]. 52

Mitchell, T. M. (1997). *Machine Learning* (1 ed.). New York, NY, USA: McGraw-Hill, Inc. 38, 39

Mohanty, S., Majumdar, S., & Natesan, K. (2012). A review of stress corrosion cracking/fatigue modeling for light water reactor cooling system components. *Argonne, IL: Nuclear Engineering Division Argonne National Laboratory*. 53

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). *Foundations of Machine Learning*. The MIT Press. 43

Montgomery, D., Peck, E., & Vining, G. (2013). *Introduction to Linear Regression Analysis*. Wiley Series in Probability and Statistics. Wiley. 45, 56

Mooney, C. H. & Roddick, J. F. (2013). Sequential pattern mining–approaches and algorithms. *ACM Computing Surveys (CSUR)*, *45*(2), 19. 25, 26

Ngai, E. W., Xiu, L., & Chau, D. C. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert systems with applications*, *36*(2), 2592–2602. 10, 11, 24

Olist (2018). Brazilian E-Commerce Public Dataset by Olist. Available: https://www.kaggle.com/olistbr/brazilian-ecommerce [19 May 2020]. 74

Olson, D. L., Cao, Q., Gu, C., & Lee, D. (2009). Comparison of customer response models. *Service Business*, *3*(2), 117–130. 24

OpenAI Five (2019). Watch our AI system play against the reigning Dota 2 world champions, OG, at the OpenAI Five Finals on Saturday, April 13th. Available: https://openai.com/five/ [23 April 2019]. 39

Oxford English Dictionary (1989). Oxford English dictionary. *Simpson, JA & Weiner, ESC*. 9

Paas, L. J., Kuijlen, A. A., & Poiesz, T. B. (2005). Acquisition pattern analysis for relationship marketing: a conceptual and methodological redefinition. *The Service Industries Journal*, *25*(5), 661–673. 21

Paas, L. J. & Molenaar, I. W. (2005). Analysis of acquisition patterns: A theoretical and empirical evaluation of alternative methods. *International Journal of Research in Marketing*, *22*(1), 87–100. 27

Paley, N. (2007). *The marketing strategy desktop guide*. Thorogood Publishing. 21

Parvatiyar, A. & Sheth, J. N. (2001). Customer relationship management: Emerging practice, process, and discipline. *Journal of Economic & Social Research*, *3*(2). 10

Pater, A.-M., Vári-Kakas, S., Poszet, O., & Pintea, I. G. (2019). Segmenting users of an online store using data mining techniques. In *2019 15th International Conference on Engineering of Modern Electric Systems (EMES)*, (pp. 205–208). IEEE. 29

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Édouard Duchesnay (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*(85), 2825–2830. 95

Peker, S., Kocyigit, A., & Eren, P. E. (2017). LRFMP model for customer segmentation in the grocery retail industry: A case study. *Marketing Intelligence & Planning*, *35*(4), 544–559. 30

Peppers, D., Rogers, M., & Dorf, B. (1999). Is Your Company Ready for One-to-One Marketing? Available: https://hbr.org/1999/01/is-your-company-ready-for-one-to-one-marketing [19 July 2019]. 16

Pierson, L. (2015). *Data science for dummies*. John Wiley & Sons. 45

Poh, J. (2019). Implementing an Effective Data Science Blueprint with the Modern Stack. In *BI&Analytics Summit 2019 Being Intelligent About Data*, Johannesburg South Africa. 33

Prinzie, A. & Van den Poel, D. (2006). Investigating purchasing-sequence patterns for financial services using markov, mtd and mtdg models. *European Journal of Operational Research*, *170*(3), 710–734. 11

Qin, X., Greene, D., & Cunningham, P. (2015). A latent space analysis of editor lifecycles in wikipedia. In *Big Data Analytics in the Social and Ubiquitous Context* (pp. 46–69). Springer. 53

Radinsky, K. & Horvitz, E. (2013). Mining the web to predict future events. In *Proceedings of the sixth ACM international conference on Web search and data mining*, (pp. 255–264). ACM. 53

Rajaraman, V. (2016). Big Data Analytics. Technical report. 32

Raorane, A., Kulkarni, R., & Jitkar, B. (2012). Association Rule-Extracting Knowledge Using Market Basket Analysis. *Research Journal of Recent Sciences*, *1*(2), 19–27. 2

Raschka, S. (2015). Principal Component Analysis in 3 Simple Steps. Available: https://sebastianraschka.com/Articles/2015_pca_in_3_steps.html [23 July 2019]. 41

Riffenburgh, R. (2012). *Statistics in Medicine* (Third Edition ed.). Elsevier Science. 45

Ruckstuhl, A. (2010). Introduction to Nonlinear Regression. *IDP Institut für Datenanalyse und Prozessdesign. ZHAW*. 45

Russel, S. & Norvig, P. (2010). *Artificial Intelligence: A Modern Approach* (3rd ed.). Prentice Hall. 39

Sadilek, A. & Krumm, J. (2012). Far out: Predicting long-term human mobility. In *Twenty-sixth AAAI conference on artificial intelligence*. 54

Salazar, M. T., Harrison, T., & Ansell, J. (2007). An approach for the identification of cross-sell and up-sell opportunities using a financial services customer database. *Journal of Financial Services Marketing*, *12*(2), 115–131. 27

Salkind, N. (2007). *Encyclopedia of measurement and statistics*. Number v. 1 in Encyclopedia of Measurement and Statistics. SAGE Publications. 44, 45

Samuel, A. L. (1959). Some studies in machine learning using the game of Checkers. *IBM Journal of Research and Development*, *3*(3), 210–229. 39

Schiffman, S. (2005). *Upselling Techniques: That Really Work!* Simon and Schuster. 26

Sculley, D., Malkin, R. G., Basu, S., & Bayardo, R. J. (2009). Predicting bounce rates in sponsored search advertisements. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, (pp. 1325–1334). ACM. 52

Shafique, U. & Qaiser, H. (2014). A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA). Technical Report 1. 38

Shalev-Shwartz, S. & Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press. 44

Sharma, S. & Athaiya, A. (2020). Activation functions in neural networks. *International Journal of Engineering Applied Sciences and Technology*, *4*(12), 310–316. 61

Shaw, M. J., Subramaniam, C., Tan, G. W., & Welge, M. E. (2001). Knowledge management and data mining for marketing. *Decision support systems*, *31*(1), 127–137. 28, 31

Shearer, C. (2000). The CRISP-DM Model: The New Blueprint for Data Mining. *Journal of Data Warehousing*, *5*. 36

Shermis, M. D. & Hamner, B. (2013). 19 contrasting state-of-the-art automated scoring of essays. *Handbook of automated essay evaluation: Current applications and new directions*, 313. 54

Siegel, E. (2013). *Predictive Analytics : The Power To Predict Who Will Click, Buy, Lie, Or Die*. Wiley. 51, 52

Siegel, E. (2016). *Predictive Analytics : The Power To Predict Who Will Click, Buy, Lie, Or Die*. Wiley. 52, 55

Smith, J. & Tekkis, P. (2012). Risk prediction in surgery. *RiskPrediction.org.uk*. 52

Sorzano, C. O. S., Vargas, J., & Montano, A. P. (2014). A survey of dimensionality reduction techniques. *arXiv preprint arXiv:1403.2877*. 41, 42

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, *36*(2), 111–133. 96

TechTarget (2015). What is descriptive analytics? - Definition from WhatIs.com. Available: https://whatis.techtarget.com/definition/descriptive-analytics [2 May 2019]. 32, 33

Tellis, G. J. & Ambler, T. (2007). *The SAGE handbook of advertising*. Sage. 45

Thomas, A. R., Lewison, D. M., & Hauser, W. J. (2007). *Direct marketing in action: cutting-edge strategies for finding and keeping the best customers*. Greenwood Publishing Group. 17, 18, 19, 21

Tsai, C.-F., Hu, Y.-H., & Lu, Y.-H. (2015). Customer segmentation issues and strategies for an automobile dealership with two clustering techniques. *Expert Systems*, *32*(1), 65–76. 30

Tsiptsis, K. K. & Chorianopoulos, A. (2011). *Data mining techniques in CRM: inside customer segmentation*. John Wiley & Sons. 12, 18, 23, 46

Upadhyay, T., Vidhani, A., & Dadhich, V. (2016). Customer profiling and segmentation using data mining techniques. *International Journal of Computer Science & Communication (IJCSC)*, *7*, 65–67. 28

USMA (2017). USMA working group. Unit for Systems Modelling and Analysis, Stellenbosch University. 44, 46

Wen, T.-H., Heidel, A., Lee, H.-y., Tsao, Y., & Lee, L.-S. (2013). Recurrent neural network based language model personalization by social network crowdsourcing. In *INTERSPEECH*, (pp. 2703–2707). 65

Winer, R. S. (2001). A framework for customer relationship management. *California management review*, *43*(4), 89–105. 12

Wu, S. & Chow, T. W. (2004). Clustering of the self-organizing map using a clustering validity index based on inter-cluster and intra-cluster density. *Pattern Recognition*, *37*(2), 175–188. 46

Yang, L., Liu, S., Tsoka, S., & Papageorgiou, L. G. (2017). A regression tree approach using mathematical programming. *Expert Systems with Applications*, *78*, 347 – 357. 44

Zaino, J. (2015). Baptist Health sees big payoff using predictive analytics. Available: https://www.healthcarefinancenews.com/news/baptist-health-sees-big-payoff-using-predictive-analytics [23 July 2019]. 52

Zhang, D. (2011). Wikipedia edit number prediction based on temporal dynamics only. *arXiv preprint arXiv:1110.5051*. 53

# Appendix A

# Hyperparameter tuning for XGBoost and Artificial Neural Network

In this Appendix the results for the random search cross validation parameter selection for the XGBoost model and the ANN models explained in Chapter 6, will be given.

## Hyperparameter tuning

### XGBoost

The hyperparameter search space is shown in Table A.1.

Table A.1: Hyperparameters for XGBoost

| Parameter | Default | Range specified |
|---|---|---|
| n_estimator | 100 | [10,100,500,800,1000,1200,1500] |
| max_depth | 3 | [2,3,4,5,8,10,15] |
| booster | gbtree | [gbtree,gblinear] |
| learning_rate | 0.1 | [0.01,0.05,0.1,0.15,0.2] |
| min_child_weight | 1 | [1,2,3,4] |
| base_score | 0.5 | [0.25,0.5,0.75,1] |

Fifty iterations of randomly selected parameters from the specified lists in Table A.1 were performed and the results of the tests that ranked 1-28 are shown in Table A.3. This table shows the parameters that were used in each test, and the metrics, negative mean absolute error on the training set and on the validation set. The rank is based on the best mean_val_score which is the negative mean absolute error on the validation set.

### Artificial Neural Network

The hyperparameter search space for the ANN can be seen in Table A.2. The results of the 20 configurations can be seen in Table A.4. Again the parameters are shown for each test, and the same metrics was used as for the XGBoost model.

Table A.2: Hyperparameter search space for the Neural Network model

| **Parameter** | **Specified search space** |
|---|---|
| layers | 1 layer : neurons[[10],[20]]<br>2 layers: neurons[[10,5],[20,10]]<br>3 layers: neurons[[45,30,15]] |
| activation functions | [relu, sigmoid] |
| batch size | [50,100] |

Table A.3: Results of hyperparameter random search cross validation tests for XGBoost

| n_estimators | min_child_weight | max_depth | learning_rate | booster | base_score | mean_val_score | std_val_score | rank_val_score | mean_train_score | std_train_score |
|---|---|---|---|---|---|---|---|---|---|---|
| 10 | 2 | 15 | 0.2 | gbtree | 0.5 | -5.025719 | 0.015918 | 1 | -3.899562 | 0.014266 |
| 100 | 3 | 10 | 0.05 | gbtree | 1 | -5.313197 | 0.020201 | 2 | -4.924274 | 0.009026 |
| 1000 | 2 | 3 | 0.1 | gbtree | 0.5 | -5.315115 | 0.019206 | 3 | -5.229103 | 0.006633 |
| 800 | 2 | 2 | 0.15 | gbtree | 0.5 | -5.315455 | 0.018365 | 4 | -5.280555 | 0.006747 |
| 800 | 1 | 2 | 0.2 | gbtree | 0.5 | -5.316316 | 0.019359 | 5 | -5.270856 | 0.006755 |
| 1200 | 3 | 3 | 0.15 | gbtree | 0.75 | -5.317467 | 0.019307 | 6 | -5.175183 | 0.006675 |
| 1000 | 4 | 4 | 0.1 | gbtree | 0.5 | -5.318834 | 0.020483 | 7 | -5.132436 | 0.007283 |
| 1000 | 3 | 4 | 0.1 | gbtree | 0.25 | -5.319705 | 0.019031 | 8 | -5.131524 | 0.006416 |
| 1200 | 4 | 4 | 0.1 | gbtree | 0.25 | -5.320889 | 0.020740 | 9 | -5.102741 | 0.007108 |
| 800 | 1 | 5 | 0.01 | gbtree | 0.5 | -5.323133 | 0.018274 | 10 | -5.286021 | 0.006128 |
| 1000 | 3 | 8 | 0.01 | gbtree | 0.25 | -5.324602 | 0.021085 | 11 | -5.075601 | 0.009034 |
| 100 | 2 | 10 | 0.1 | gbtree | 1 | -5.348569 | 0.019996 | 12 | -4.729604 | 0.008678 |
| 100 | 1 | 2 | 0.05 | gbtree | 0.75 | -5.349042 | 0.018397 | 13 | -5.346384 | 0.006622 |
| 1500 | 3 | 4 | 0.2 | gbtree | 0.25 | -5.353561 | 0.018636 | 14 | -4.855878 | 0.007486 |
| 500 | 1 | 8 | 0.1 | gbtree | 0.75 | -5.364682 | 0.020407 | 15 | -4.455917 | 0.008657 |
| 1200 | 2 | 8 | 0.1 | gbtree | 1 | -5.422223 | 0.020750 | 16 | -3.700489 | 0.010417 |
| 1500 | 3 | 3 | 0.2 | gblinear | 0.5 | -5.422543 | 0.019117 | 17 | -5.422392 | 0.006928 |
| 1200 | 3 | 15 | 0.2 | gblinear | 1 | -5.422543 | 0.019118 | 18 | -5.422392 | 0.006928 |
| 1500 | 1 | 5 | 0.15 | gblinear | 0.75 | -5.422547 | 0.019118 | 19 | -5.422396 | 0.006928 |
| 1200 | 1 | 8 | 0.15 | gblinear | 0.25 | -5.422548 | 0.019117 | 20 | -5.422397 | 0.006929 |
| 1500 | 4 | 15 | 0.15 | gblinear | 1 | -5.422549 | 0.019118 | 21 | -5.422398 | 0.006928 |
| 1000 | 2 | 8 | 0.15 | gblinear | 0.75 | -5.422612 | 0.019119 | 22 | -5.422461 | 0.006930 |
| 1200 | 3 | 2 | 0.1 | gblinear | 0.75 | -5.422855 | 0.019125 | 23 | -5.422706 | 0.006933 |
| 1000 | 4 | 3 | 0.1 | gblinear | 0.5 | -5.423024 | 0.019129 | 24 | -5.422876 | 0.006937 |
| 1000 | 3 | 2 | 0.1 | gblinear | 1 | -5.423237 | 0.019135 | 25 | -5.423089 | 0.006938 |
| 1000 | 3 | 8 | 0.05 | gblinear | 0.75 | -5.425897 | 0.019200 | 26 | -5.425756 | 0.006990 |
| 1000 | 2 | 5 | 0.05 | gblinear | 0.75 | -5.425897 | 0.019200 | 26 | -5.425756 | 0.006990 |
| 1000 | 3 | 5 | 0.01 | gblinear | 0.25 | -5.426177 | 0.018893 | 28 | -5.426059 | 0.007251 |

Table A.4: Results of hyperparameter random search cross validation tests for Neural Network

| layers | epochs | batch_size | activation | mean_val_score | std_val_score | rank | mean_train_score | std_train_score |
|---|---|---|---|---|---|---|---|---|
| [10, 5] | 50 | 100 | relu | -0.663780 | 0.004288 | 1 | -0.663276 | 0.001740 |
| [20, 10] | 50 | 50 | sigmoid | -0.664977 | 0.004998 | 2 | -0.664389 | 0.001484 |
| [20, 10] | 50 | 100 | sigmoid | -0.665307 | 0.003323 | 3 | -0.665271 | 0.000903 |
| [10] | 50 | 50 | relu | -0.666151 | 0.002631 | 4 | -0.666364 | 0.001557 |
| [45, 30, 15] | 50 | 50 | relu | -0.666273 | 0.003643 | 5 | -0.665849 | 0.002992 |
| [20] | 50 | 50 | relu | -0.666744 | 0.002324 | 6 | -0.666737 | 0.001322 |
| [10] | 50 | 50 | sigmoid | -0.667305 | 0.002496 | 7 | -0.667317 | 0.001725 |
| [20] | 50 | 50 | sigmoid | -0.667659 | 0.002733 | 8 | -0.667649 | 0.001146 |
| [20] | 50 | 100 | sigmoid | -0.667907 | 0.004132 | 9 | -0.668047 | 0.002172 |
| [10] | 50 | 100 | sigmoid | -0.668367 | 0.003554 | 10 | -0.668401 | 0.001913 |
| [20] | 50 | 50 | relu | -0.668663 | 0.003692 | 11 | -0.669286 | 0.001232 |
| [20] | 50 | 100 | relu | -0.669903 | 0.001002 | 12 | -0.669628 | 0.002606 |
| [45, 30, 15] | 50 | 100 | relu | -0.670222 | 0.002422 | 13 | -0.669780 | 0.001470 |
| [10, 5] | 50 | 50 | relu | -0.671980 | 0.018740 | 14 | -0.671506 | 0.019172 |
| [20, 10] | 50 | 50 | relu | -0.672976 | 0.021213 | 15 | -0.673235 | 0.017681 |
| [45, 30, 15] | 50 | 100 | sigmoid | -0.676032 | 0.016947 | 16 | -0.675576 | 0.017154 |
| [20] | 50 | 100 | relu | -0.677677 | 0.018773 | 17 | -0.677309 | 0.015696 |
| [20] | 50 | 50 | sigmoid | -0.677723 | 0.015809 | 18 | -0.677175 | 0.016449 |
| [10, 5] | 50 | 100 | sigmoid | -0.682041 | 0.020813 | 19 | -0.681871 | 0.023290 |
| [45, 30, 15] | 50 | 50 | sigmoid | -0.684067 | 0.023328 | 20 | -0.683522 | 0.020918 |

# Appendix B

# Results of the NPD Predictor for Cluster 3

In this Appendix the results of the NPD Predictions is presented. Table B.1 shows the predictions made in days for each user product pair. A sample of 100 predictions are shown. The table also includes the product name. These predictions are used in Chapter 8 to generate individualised advertisements.

Table B.1: NPD Predictions

| user_id | product_id | Product | Predict |
|---|---|---|---|
| 17 | 7350 | Natural Lime Flavor Sparkling Mineral Water | 8.39 |
| 17 | 17762 | Light Oaked Chardonnay | 10.41 |
| 17 | 18534 | Grade A Extra Large Eggs | 8.98 |
| 50 | 5612 | Reduced Fat Milk | 9.53 |
| 50 | 23165 | Organic Leek | 9.96 |
| 50 | 6182 | Natural Artesian Bottled Water | 8.18 |
| 140 | 19894 | Enlightened Organic Raw Kombucha | 8.73 |
| 140 | 21616 | Organic Baby Arugula | 13.75 |
| 140 | 34262 | Hint Of Sea Salt Almond Nut Thins | 13.86 |
| 140 | 36865 | Non Fat Raspberry Yogurt | 12.74 |
| 140 | 47610 | Pretzel Sticks, Gluten Free | 11.28 |
| 140 | 19057 | Organic Large Extra Fancy Fuji Apple | 13.10 |
| 140 | 432 | Vanilla Almond Breeze Almond Milk | 10.15 |
| 140 | 5973 | Spicy Avocado Hummus | 3.90 |
| 187 | 29436 | Original Party Mix Catt Treats | 4.76 |
| 210 | 4799 | Shredded Parmesan | 8.57 |
| 210 | 4920 | Seedless Red Grapes | 8.85 |
| 210 | 10292 | Family Size Shells & White Cheddar Macaroni & ... | 8.77 |
| 210 | 21903 | Organic Baby Spinach | 9.19 |
| 210 | 23375 | Marinara Sauce | 9.56 |
| 210 | 23909 | 2% Reduced Fat Milk | 4.33 |
| 210 | 24852 | Banana | 4.76 |
| 210 | 29388 | Mandarin Orange Crispy Chick'n | 10.66 |
| 210 | 32030 | Extra Sharp White Cheddar | 6.38 |
| 210 | 33043 | Crescent Rolls | 10.76 |
| 210 | 35547 | Organic Baby Kale | 6.26 |
| 210 | 41065 | Organic Yellow Squash | 7.04 |
| 210 | 42585 | Organic Extra Firm Tofu | 10.74 |
| 210 | 47209 | Organic Hass Avocado | 12.79 |
| 210 | 40198 | Blueberry Yoghurt | 7.02 |

Table B.1 continues on next page

| user_id | product_id | Product | Predict |
|---|---|---|---|
| 223 | 8309 | Nonfat Icelandic Style Strawberry Yogurt | 11.88 |
| 223 | 27086 | Half & Half | 12.32 |
| 223 | 40396 | Guacamole | 11.16 |
| 223 | 5250 | Blueberry Muffin Bar | 12.39 |
| 236 | 19660 | Spring Water | 8.68 |
| 243 | 27104 | Fresh Cauliflower | 18.18 |
| 290 | 47540 | Savory Delights Ham & Egg Flavor With Potato &... | 11.46 |
| 290 | 43511 | Sunrise Grilled Steak and Eggs Canine Cuisine ... | 10.69 |
| 290 | 43409 | Frosted Mini-Wheats Original Cereal | 11.77 |
| 290 | 42895 | Filets In Sauce New York Strip Flavor Dog Food | 11.63 |
| 290 | 34355 | Spearmint Sugar-Free Gum | 8.95 |
| 290 | 31981 | 1% Low Fat Milk | 13.00 |
| 290 | 30251 | All Natural Apple Juice Drink | 12.58 |
| 290 | 23405 | Pure Baking Soda | 10.11 |
| 307 | 6184 | Clementines | 14.96 |
| 307 | 8057 | Organic Chicken & Mozzarella Ravioli | 15.72 |
| 334 | 17830 | Premium Organic Strawberry Spread | 30.24 |
| 334 | 22035 | Organic Whole String Cheese | 10.12 |
| 334 | 5120 | Organic Vanilla Soymilk | 13.74 |
| 334 | 9076 | Blueberries | 9.37 |
| 334 | 10749 | Organic Red Bell Pepper | 8.77 |
| 334 | 21137 | Organic Strawberries | 10.23 |
| 334 | 37119 | Uncured Pepperoni | 7.76 |
| 334 | 49086 | Coconut Bliss Frozen Dessert, Dark Chocolate | 10.33 |
| 334 | 17807 | Gluten Free Blueberry Waffles | 30.95 |
| 398 | 26369 | Organic Roma Tomato | 9.99 |
| 398 | 21903 | Organic Baby Spinach | 11.84 |
| 454 | 28993 | Iceberg Lettuce | 8.66 |
| 454 | 8390 | Diet Ginger Ale | 6.69 |
| 454 | 11463 | Honey Ham | 12.02 |
| 516 | 9018 | All Natural Virgin Lemonade | 10.88 |
| 516 | 13629 | Organic Snipped Green Beans | 11.49 |
| 516 | 8277 | Apple Honeycrisp Organic | 4.65 |
| 516 | 45948 | Gluten Free 7 Grain Bread | 10.65 |
| 516 | 39589 | Grassmilk Organic Fat Free Milk | 10.67 |
| 516 | 21137 | Organic Strawberries | 10.87 |
| 516 | 25837 | Mini Babybel Light Semisoft Edam Cheeses | 6.96 |
| 516 | 13176 | Bag of Organic Bananas | 13.28 |
| 516 | 19741 | Organic Unsweetened Soy Milk H | 11.60 |
| 516 | 36082 | Organic Mango | 10.96 |
| 529 | 36929 | Milk, Vitamin D | 10.23 |
| 562 | 38928 | 0% Greek Strained Yogurt | 5.75 |

Table B.1 continues on next page

| user_id | product_id | Product | Predict |
|---|---|---|---|
| 583 | 13083 | Lowfat Small Curd Cottage Cheese | 15.89 |
| 599 | 15511 | Draft Sake | 13.05 |
| 646 | 19057 | Organic Large Extra Fancy Fuji Apple | 8.77 |
| 646 | 35921 | Organic Large Grade A Brown Eggs | 5.78 |
| 646 | 35535 | Organic Red Chard Greens | 9.49 |
| 646 | 13176 | Bag of Organic Bananas | 10.82 |
| 646 | 21137 | Organic Strawberries | 10.83 |
| 646 | 27521 | Organic Lacinato (Dinosaur) Kale | 8.95 |
| 699 | 19057 | Organic Large Extra Fancy Fuji Apple | 8.75 |
| 699 | 16617 | Organic Muenster Cheese Slices | 6.53 |
| 818 | 19019 | Uncured Slow Cooked Ham | 10.03 |
| 818 | 1577 | Unsweetened Whole Milk Peach Greek Yogurt | 8.57 |
| 818 | 8575 | Unprocessed American Singles Colby-Style Cheese | 12.57 |
| 818 | 44479 | Organic Stringles Mozzarella String Cheese | 12.49 |
| 818 | 24852 | Banana | 3.96 |
| 818 | 11520 | Large Alfresco Eggs | 13.32 |
| 818 | 41665 | Organic Mexican Blend Finely Shredded Cheese | 11.58 |
| 818 | 32747 | Low Fat 1% Milk | 13.99 |
| 818 | 18531 | Organic Heavy Whipping Cream | 5.00 |
| 849 | 36127 | Freshly Squeezed Orange Juice | 10.62 |
| 868 | 24852 | Banana | 5.41 |
| 868 | 27845 | Organic Whole Milk | 6.49 |
| 868 | 37687 | Organic Spring Mix | 9.48 |
| 911 | 27845 | Organic Whole Milk | 5.34 |
| 911 | 41220 | Organic Romaine Lettuce | 4.34 |
| 911 | 13176 | Bag of Organic Bananas | 18.54 |
| 117 | 1194 | Natural Artisan Water | 13.45 |
| 119 | 49621 | Challah Bread | 12.20 |
| 119 | 432 | Vanilla Almond Breeze Almond Milk | 12.67 |

End of Table B.1