REVA
UNIVERSITY
Bengaluru, India

A Project Report on
Product Affinity Analysis using Machine Learning

Submitted in Partial Fulfilment for Award of Degree of
Master of Business Administration
In Business Analytics

Submitted By
Sharon Joseph
R19DM053

Under the Guidance of
Mithun D J
Senior Manager -Data Science, RACE

REVA Academy for Corporate Excellence - RACE
REVA University
Rukmini Knowledge Park, Kattigenahalli, Yelahanka, Bengaluru - 560 064
race.reva.edu.in

August, 2022

REVA
UNIVERSITY
Bengaluru, India

Candidate's Declaration

I, **Sharon Joseph** hereby declare that I have completed the project work towards a Master of Business Administration in Business Analytics at, REVA University on the topic entitled **Product Affinity Analysis using Machine Learning** under the supervision of **Mithun D J, Senior Manager -Data Science, RACE.** This report embodies the original work done by me in partial fulfilment of the requirements for the award of the degree for the academic year **2022.**

Sharonjoseph

Place: Bengaluru
Date: 27th August 2022

Name of the Student: Sharon Joseph
Signature of Student

REVA
UNIVERSITY
Bengaluru, India

Certificate

This is to Certify that the Project work entitled **Product Affinity Analysis using Machine Learning** carried out by **Sharon Joseph** with **R19DM053,** a bonafide student at REVA University, is submitting the project report in fulfilment for the award of Master of Business Administration (MBA) in Business Analytics during the academic year 2022. The Project report has been tested for plagiarism and passed the plagiarism test with a less than 15% similarity score. The project report has been approved as it satisfies the academic requirements in respect of the project work prescribed for the said Degree.

Signature of the Guide	Signature of the Director
Name of the Guide: **Mr Mithun D J**	Name of the Director: **Dr Shinu Abhi**

External Viva

Names of the Examiners

1. Dr. Sai Hareesh, Research Expert, SAP Labs India
2. Pradeepta Mishra, Director – AI, L&T InfoTech

Place: Bengaluru

Date: 27th August 2022

REVA
UNIVERSITY
Bengaluru, India

Acknowledgement

I would like to acknowledge the support provided by the founder and Hon'ble Chancellor, **Dr P Shayma Raju**, Vice-Chancellor, **Dr M. Dhanamjaya**, and Registrar, **Dr N Ramesh**. for supporting the RACE program, and providing the necessary facilities and infrastructure required for the best learning experience. I am immensely proud of being part of this program and REVA University.

I would like to thank **Dr Shinu Abhi**, Director, RACE (REVA Academics for Corporate Excellence) for her guidance and constant supervision in providing necessary information regarding the course, subjects, and a plethora of other topics, thus ensuring that the learning experience was immensely powerful.

I sincerely thank all my mentors who have taught and guided me in this journey to achieve my degree. Their vast knowledge and practical experience paved way for a great learning experience.

A special thanks to my project guide **Mr Mithun D J**, Senior Manager -Data Science RACE for his valuable support, encouragement, and guidance provided in understanding the concept for executing the project.

I am also thankful to my classmates for their constructive criticism and friendly advice during the course period.

Last but not the least, I would like to thank the Almighty, my family, and my colleagues who have extended their invaluable cooperation and encouragement throughout this project lifecycle.

Sharon Joseph

Place: Bengaluru
Date: 27th August 2022

Name of the Student: Sharon Joseph
Signature of Student

REVA
UNIVERSITY
Bengaluru, India

Similarity Index Report

This is to certify that this project report titled "**Product Affinity Analysis using Machine Learning**" was scanned for similarity detection. The process and outcome are given below.

Software Used: Turnitin

Date of Report Generation: 26th August 2022

Similarity Index in %: 4%

Total word count: 6751

Name of the Guide: Mr Mithun D J

Sharonjoseph

Place: Bengaluru

Date: 27th August 2022

Name of the Student: Sharon Joseph

Signature of Student

Verified by: M N Dincy Dechamma

Signature

Dr Shinu Abhi,

Director, Corporate Training

List of Abbreviations

Sl. No	Abbreviation	Long Form
1	ERP	Enterprise resource planning
2	SCM	Supply chain management
3	CRM	Customer relationship management
4	OEE	Overall Equipment Effectiveness
5	MES	Manufacturing Execution Systems
6	EAM	Enterprise asset management
7	TMS	Transportation management system
8	WMS	Warehouse Management Systems
9	PLM	Product Lifecycle Management
10	TURF	Total Unduplicated Reach and Frequency
11	CART	Classification and Regression Trees
12	KNN	K-Nearest Neighbour
13	MBA	Market Basket Analysis
14	EDA	Exploratory data analysis
15	CHAID	Chi-squared Automatic Interaction Detector

List of Tables

List of Figures

Abstract

Decision-making and understanding consumer habits have become critical and difficult problems for companies wanting to maintain their position in competitive industries.

Aptean is a product-based company that provides mission-critical, industry-specific software owning an exclusive range of high-end products that serve in areas like, “Enterprise resource planning (ERP), Supply chain management (SCM), Customer relationship management (CRM), Overall Equipment Effectiveness (OEE), Manufacturing Execution Systems (MES), Enterprise asset management (EAM), Transportation management system (TMS), Warehouse Management Systems (WMS), Product Lifecycle Management (PLM), Compliance and Business Solutions”.

This study aims to leverage customer firmographic data and product sales transaction data to build a solution to project the likelihood of a purchase from our existing customer base. A better understanding of what was sold, why and to whom would help improve the ability to identify new cross-sell and up-sell opportunities. It discovers the associations among products using machine learning and predicts the products that could be projected for potential sales opportunities. Machine learning algorithms like Market Basket Analysis using Apriori, TURF Analysis for frequency study, CHAID algorithm- Decision tree, K-Nearest neighbour (KNN) and Multilayer Perceptron as part of Deep Learning are the techniques used to derive the desired outcome to the problem.

The specific outcome includes product affinity analysis and recommendation of products that can be cross-sold or upsold to existing customers or new customers. This analysis could help determine new sales opportunities thereby helping the organization.

Keywords: Machine Learning, Product Affinity Analysis, Cross-Selling, Market Basket Analysis, Apriori, CHAID, KNN, Deep Learning, Multilayer perceptron

Contents

Chapter 1: Introduction

Every business aims to improve its revenue and profits, and this is mainly achieved through increasing sales.

Some of how we can increase sales are to acknowledge the current customer behavior, requesting customer feedback, running promotions for current customers, providing excellent customer service, creating packages, deals and free trials to attract customers, conducting a content audit, to do something noteworthy or unique, to optimize your social media profiles, to advertise on social media platforms, to spread by word of mouth, put a call to action on your website, and to stay in touch with email marketing (David Gargaro, 2022).

These methods are traditional methods which most companies would follow to increase their sales. However, this study focuses on some machine learning techniques to promote cross-selling and up-selling opportunities.

"Cross-selling is a sales technique involving the selling of an additional product or service to an existing customer whereas, Up-selling is a sales technique where a seller invites the customer to purchase more expensive items, upgrades, or other add-ons to generate more revenue" (Wikipedia, 2022).

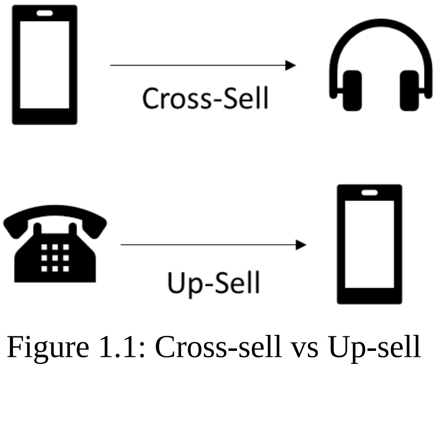


Figure 1.1: Cross-sell vs Up-sell

From figure 1.1, cross-sell indicates that headphones can be sold to a customer who has purchased a mobile phone and up-sell indicates that a mobile phone could be sold to a customer who purchased a telephone. Here, the headphone is a complementary product that may interest a customer, whereas a mobile phone is an upgraded version of a telephone that a customer may purchase.

Product co-occurrence analysis is one area of research that can help determine the relationship between various products. Affinity analysis and association rule learning are two analytics techniques that aim to discover the relationships and connections between specific objects. This market basket analysis is probably the most well-known example. Market Basket Analysis (MBA) identifies product associations by looking for product combinations that frequently co-occur in exchanges. Individuals who purchase flour and sugar, also buy eggs since a large proportion of them intend to bake a cake. (Gupta et al., 2016).

In this project, the aim is to identify and study the customers' purchase behavior from the product sales transaction data and customer firmographics data, to find out the products that are more likely to be cross-sold or up-sold to the different customer categories. This would help the sales team to improve sales and increase revenue, by promoting the existing customers to purchase better or related products.

Chapter 2: Literature Review

For this study, research papers have been extensively reviewed on various topics related to Affinity analysis and the different methodologies used to analyse product cross-selling and upselling. Some of these research papers have been discussed herein.

This paper focuses on "a binary classification framework for predicting the successful upsell of products and services, using data from one such telecommunications provider. Through this prediction model, the recommender system for voice products/services to corporate customers of the telecommunications company is demonstrated. Logistic regression classifier to automate the selection of customers that are most likely to upsell. Application of a predictive model to recommend a set of target customers to approach foupsellingll, illustrating the different accuracy results for different cost weightings and also show that the success rate of upselling products to the selected customers is dramatically improved when compared to the traditional approach" (Dookeram et al., 2022).

In this study, "an SVM ensemble learning method is proposed for classification using tensor data. The method is used in identifying cross-selling opportunities to recommend personalized products and services to customers. Two real-world databases are used to evaluate the performance of the method. In this study, a support vector machine (SVM) ensemble learning method, as a new data mining method, is proposed for classification using tensor data. Computational results show that the SVM ensemble learning method has good performance on these databases" (Chen et al., 2015).

This paper proposes that "affinity analysis and association rule mining encompasses a broad set of analytics techniques aimed at revealing the associations and correlation between specific objects. The purpose of this analysis is to generate a set of rules that relate two or more products together.

Each of these rules should have a lift greater than one. The interest is in the support and confidence of those rules such as higher confidence rules are ones where there is a higher probability of items on the RHS being part of the transaction given the presence of items on the LHS. R is a great statistical and graphical analysis tool, well suited to more advanced analysis which is used to perform the market basket analysis" (Gupta et al., 2016b)

The main objective of this paper is to "analyse the huge amount of data thereby exploiting consumer behaviour and making the correct decision leading to a competitive edge over rivals. Experimental analysis has been done employing association rules using Market Basket Analysis to prove its worth over the conventional methodologies. OLAP tools have been commercially used for in-depth analysis such as data classification, clustering and characterization of data that changes over time" (Raorane AA et al., 2012)

The goal of this project is to "use anonymized data from customers' transactional orders to focus on descriptive analysis of customer purchase patterns, items purchased together, and units purchased frequently from the store to facilitate reordering and maintaining adequate product stock. It is possible to accomplish this by analyzing the available data in such a way that a frequent item set can be identified and analyzed in order to the association rule. The Apriori algorithm is one of the algorithms that aid in the discovery of association rules for frequent item sets and the identification of correlations. The apriori algorithm model is being developed to investigate approaches for applying association rules to recommender systems" (Shruthi Gurudath, 2020)**.**

This paper reveals that, "the relationship between the purchased goods by using the Apriori algorithm to find out the data of shopping baskets from the massive data of consumers and then applying the association rules and CART decision tree algorithm to reveal the characteristics of the customer group and the target customers' classification. It is convenient for the goods to be better configured and sold to extract more detailed and valuable information, as well as to improve the market's operational efficiency" (Chen et al., 2015).

In this paper, the primary purpose is to build a data mining method in Excel using an XLMiner add-in to accelerate cross-selling. It does not necessitate a great deal of expertise in data mining but only some parameter setting. Furthermore, almost everyone is proficient in Excel. Therefore, the proposed Excel-based method is simple to learn. This paper also presents an example through mining association rules(Hewen et al., 2008).

Therefore, this project will emphasize on data mining and machine learning models to profile customers and products and propose product upsell or cross-sell opportunities.

Chapter 3: Problem Statement

At Aptean, traditional methods like publishing about the products on the company website, digital marketing, advertisements, and setting up campaigns are followed to sell the different products to customers. Aptean focuses on delivering purpose-built, industry-specific, mission-critical enterprise software, tailored to the industry's needs and so it is a time-consuming process to identify the potential customers and to sell products to them, which includes their requirements and customizations.

Product affinity analysis could be done to identify the different products that have a strong affinity and that can be considered for up-selling and cross-selling. The different products could be profiled based on their type, usage, customer firmographic data, revenue, and other valuable data. Customer profiling is another important study that could eventually help us to identify the potential customers to whom we could propose our findings, which in turn would help determine new sales opportunities, thereby increasing the organization’s revenue.

Leverage customer firmographic data and product sales transaction data to build a model to project the likelihood to purchase from our existing customer base and also solve the business challenge of achieving and meeting the sales targets of the organization. Increasing sales through product cross-selling and up-selling and identifying the customers to whom products can be sold.

Chapter 4: Objectives of the Study

The primary goal of this research is to help the company understand current customer behavior and predict future customer purchasing behavior. Leveraging customer transaction data can assist in understanding customer purchasing behavior, providing the correct bundles and promotions, assortment planning, and inventory management in order to retain customers, improve sales, and extend their customer relationship.

The important objectives of this study are stated below:

1. To understand the product purchase pattern from product sales transaction data.

2. To study and profile customers based on their purchase behavior.

3. To recommend and suggest products to customers, thereby increasing cross-sell and up-sell opportunities.

Chapter 5: Project Methodology

This Project uses the CRISP-DM framework to implement and carry out the different phases of the data mining and machine learning process involved in this study.

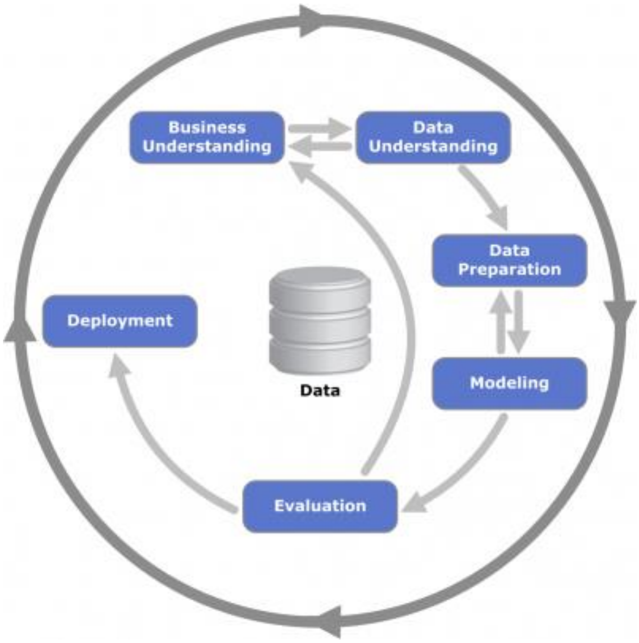
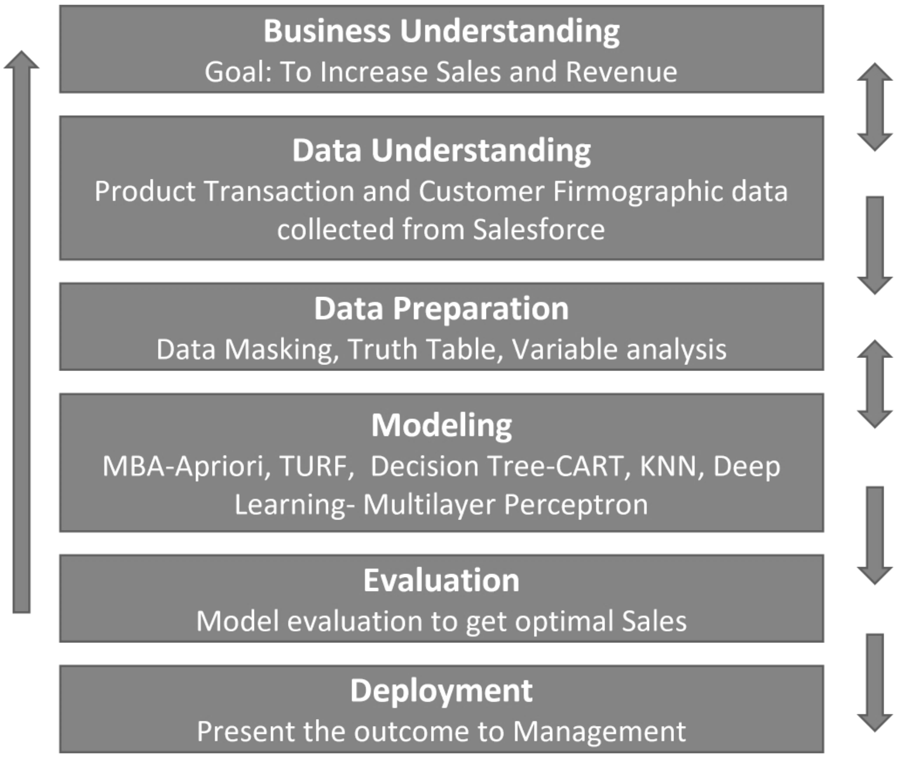


Figure 5.1: CRISP-DM Framework (Think Insights, 2022)

As in Figure 5.1, the CRISP-DM process begins with Business Understanding where the goal is to understand the project aim and fulfillment from the business view. Then comes the Data Understanding phase which starts with initial data collection, familiarizes the data, identifies data quality and discovers first insights into the data. The next phase is the most important phase called the data Preparation, which covers all the steps to build the final dataset from the initially collected data from salesforce systems. The next phase is called Modeling, here, the evaluation, selection and application of the appropriate modelling techniques is done. Later in the Evaluation phase, the model builds and right model selection that appear to have high-quality, is tested and validated such that the models sufficiently cover all the important business concerns. Finally, Deployment phase which deploys the model in to and it also includes mechanisms to evaluate or group new hidden data as it arises (Think Insights, 2022).



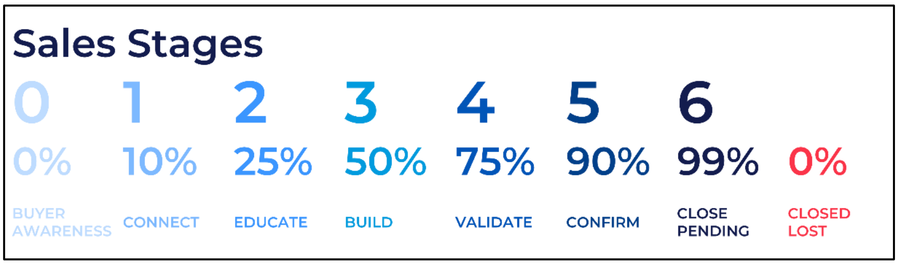
Business Understanding
Goal: To Increase Sales and Revenue
Data Understanding
Product Transaction and Customer Firmographic data collected from Salesforce
Data Preparation
Data Masking, Truth Table, Variable analysis
Modeling
MBA-Apriori, TURF, Decision Tree-CART, KNN, Deep Learning- Multilayer Perceptron
Evaluation
Model evaluation to get optimal Sales
Deployment
Present the outcome to Management

Figure 5.2: Project Pipeline

Figure 5.2 illustrates the detailed project pipeline. Initially, as part of the business understanding, discussions and meetings with the relevant stakeholders were conducted. The goal of maximizing the sales and the revenue of the organization was captured and the project plan was created. The data such as the product sales transaction data along with the customer firmographics data was collected from the internal Salesforce system. Later the data was organized, explored, visualized and verified for data quality. As part of the data preparation, relevant data was selected, cleaned, constructed, integrated and re-formatted to run the various techniques for analysis and modelling. Exploratory data analysis is carried out and different modelling techniques were performed. The models were then assessed, evaluated for the results and reviewed. Finally, the model is iterated to verify if the business objectives were met and deployed for end users. In this case, the deployment was to present the findings of this study to management and provide them with recommendations on what products can be sold and to which category of customers probable sales can be proposed.

Chapter 6: Business Understanding

Aptean launched “Mission-500”, a strategy to become a $500M business, driven by organic growth of the core business, expansion through acquisitions, streamlining processes and building an amazing team. The foundation of Mission 500 was to provide market-leading ERP, Supply Chain and Compliance solutions such that it created exceptional value for customers. Upon achieving this great mission, Aptean has launched the next growth horizon called, **“Operation 10^x ”** and the main goal of this initiative is to drive 10% organic revenue growth each year and become a $ 1 billion revenue company by the end of the next 5 years.



Sales Stages

0	1	2	3	4	5	6	
0%	10%	25%	50%	75%	90%	99%	0%
BUYER AWARENESS	CONNECT	EDUCATE	BUILD	VALIDATE	CONFIRM	CLOSE PENDING	CLOSED LOST

Figure 6.1: Sales Stages

The organization’s sales department currently follows 7 stages for achieving the sales targets and Figure 6.1 illustrates the same.

Stage 1- Buyer Awareness: The buyer has expressed interest and qualified meetings have been set.

Stage 2- Connect: Sales teams agree to accept the meeting and qualify for an opportunity.

Stage 3- Educate: Sales team further qualifies the opportunity fit on both sides.

Stage 4- Build: Sales and buyers are jointly building and confirming requirements for deal progression.

Stage 5- Validate: The buyer is finalising the solution fit and selecting the vendor of choice.

Stage 6- Confirm: Parties negotiate, and the buyer conducts the final vendor assessments to make the purchase decisions.

Stage 7- Close: The deal is marked as closed from a sales booking perspective.

a) Closed Pending: Awaiting final finance team approval.

b) Closed Won: Opportunity successfully closed after getting finance approval.

c) Closed Lost: Opportunity is longer viable.

7 Stages	**Sales Team Activities**
Buyer Awareness	The opportunity is reviewed, and the sales team will determine if they accept the meeting.
Connect	The sales team confirms the solution fit and plans the priority, consequences, budget range and timeline and the next meeting is scheduled within 2 weeks.
Educate	Conducts detailed qualifications. Identifies buyer roles, potential catalysts, competitors, and purchase process. Conducts Overview Demo.
Build	Business and fitness assessments. Validates buyer insights with stakeholders. Create an action plan. Prepare buyer scorecard/ Decision criteria. Create a preliminary proposal.
Validate	Proactive project implementation plan discussions. Address customer security requirements. Create a formal proposal/quote.
Confirm	Facilitate reference introductions. Review contract and confirm purchasing process. Negotiate/support cost justifications.
Close	Check with finance teams and prepare the closure plan.

Table 6.1: Sales Team Activities

Table 6.1 explains the activities carried out by the sales team members at the different sales stages in the organization. Data mining and machine learning techniques can be used to achieve the goal of increasing revenue and sales. Data mining is frequently used in business to unearth hidden knowledge in massive datasets. Cross-selling, which is the technique of recommending complementary goods or services to a customer who is thinking about making a purchase, is crucial to marketing and sales.

Therefore, this study's primary goals are to help the firm comprehend the behavior of its current customers and foretell the purchasing habits of its future ones. Utilizing consumer transaction data can assist businesses in extending their relationships with customers, retaining current customers, and increasing sales, revenue and profits.

Chapter 7: Data Understanding

The second stage of the CRISP-DM methodology includes processes to identify, collect, and analyse datasets that will assist in meeting the project objectives.

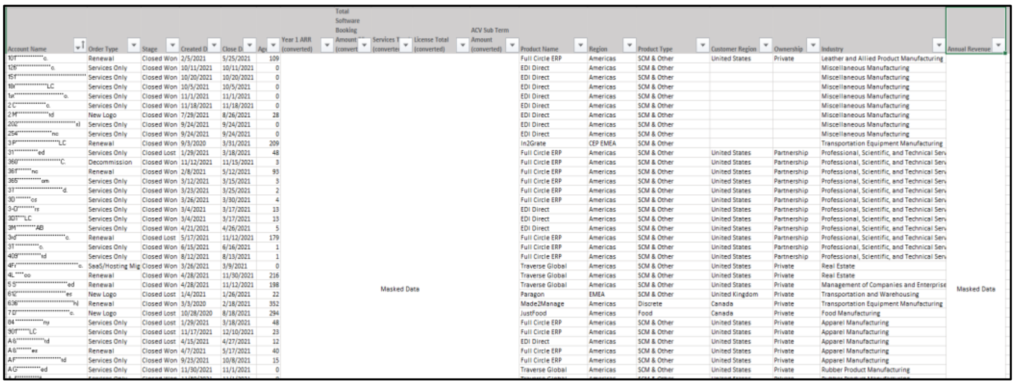
7.1 Collecting the Initial data

The data required for this study is stored in the Salesforce system, which is managed by the organization. During the data collection phase, meetings were held between stakeholders of the sales team, renewals team and the director of the Strategy Department. After discussions about the business problem and the goal to achieve, the customer firmographic data and the product purchase transaction data were merged and shared in the form of an excel file. Since the data involved Customer firmographics and certain financial data, the requirement was to use the data in a masked format, so that the organization's and customers' confidentiality can be maintained.

Features	Description
Account Name	Name of the Customer in masked format
Order Type	The type of order/transaction
Stage	The Sales Stage
Created Date	Date when the order was created
Close Date	Date when the order was closed
Age	Time is taken to close the order (in days)
Year 1 ARR	Financial Data which is masked
Total Software Booking Amount PE	
Services Total	
License Total	
ACV Sub Term Amount	
Product Name	Name of the Product
Region	Product selling region
Product Type	Category of the product
Customer Region	Product buying region
Ownership	Type of ownership of the business
Industry	The type of Industry buying the product
Total Revenue	Customers' Total annual revenue masked

Table 7.1: Data Dictionary

Table 7.1 shows the data dictionary with the explanation of features involved in this study and figure 7.1 shows the sample data in a masked format, which will be used in the further study. Data masking was done in excel itself by replacing the middle few characters of the customer's name with an asterisk symbol and by simply hiding the financial data.



Account Name	Order Type	Stage	Created D	Close D	Ag	Year 1 ARR (converted)	Total Software Booking Amount (convert	Services T (converte	License Total (converted)	ACV Sub Term Amount (converted)	Product Name	Region	Product Type	Customer Region	Ownership	Industry	Annual Revenue
10T************o.	Renewal	Closed Won	2/5/2021	5/25/2021	109	Masked Data					Full Circle ERP	Americas	SCM & Other	United States	Private	Leather and Allied Product Manufacturing	Masked Data
126**************o.	Services Only	Closed Won	10/11/2021	10/11/2021	0						EDI Direct	Americas	SCM & Other			Miscellaneous Manufacturing	
15T**************************	Services Only	Closed Won	10/20/2021	10/20/2021	0						EDI Direct	Americas	SCM & Other			Miscellaneous Manufacturing	
18i*************LC	Services Only	Closed Won	10/5/2021	10/5/2021	0						EDI Direct	Americas	SCM & Other			Miscellaneous Manufacturing	
1st*******************o.	Services Only	Closed Won	11/1/2021	11/1/2021	0						EDI Direct	Americas	SCM & Other			Miscellaneous Manufacturing	
2 C*************o.	Services Only	Closed Won	11/18/2021	11/18/2021	0						EDI Direct	Americas	SCM & Other			Miscellaneous Manufacturing	
2 M*************td	New Logo	Closed Won	7/29/2021	8/26/2021	28						EDI Direct	Americas	SCM & Other			Miscellaneous Manufacturing	
202***********************el	Services Only	Closed Won	9/24/2021	9/24/2021	0						EDI Direct	Americas	SCM & Other			Miscellaneous Manufacturing	
254**************nc	Services Only	Closed Won	9/24/2021	9/24/2021	0						EDI Direct	Americas	SCM & Other			Miscellaneous Manufacturing	
3 P*****************LC	Renewal	Closed Won	9/3/2020	3/31/2021	209						In2Grate	CEP EMEA	SCM & Other			Transportation Equipment Manufacturing	
31***********ed	Services Only	Closed Lost	1/29/2021	3/18/2021	48						Full Circle ERP	Americas	SCM & Other	United States	Partnership	Professional, Scientific, and Technical Serv	
360*****************C.	Decommission	Closed Won	11/12/2021	11/15/2021	3						Full Circle ERP	Americas	SCM & Other	United States	Partnership	Professional, Scientific, and Technical Serv	
361*******nc	Renewal	Closed Won	2/8/2021	5/12/2021	93						Full Circle ERP	Americas	SCM & Other	United States	Partnership	Professional, Scientific, and Technical Serv	
365***********om	Services Only	Closed Won	3/12/2021	3/15/2021	3						Full Circle ERP	Americas	SCM & Other	United States	Partnership	Professional, Scientific, and Technical Serv	
3T*******************d.	Services Only	Closed Won	3/23/2021	3/25/2021	2						Full Circle ERP	Americas	SCM & Other	United States	Partnership	Professional, Scientific, and Technical Serv	
30 ********os	Services Only	Closed Won	3/26/2021	3/30/2021	4						Full Circle ERP	Americas	SCM & Other	United States	Partnership	Professional, Scientific, and Technical Serv	
3-O********rs	Services Only	Closed Won	3/4/2021	3/17/2021	13						EDI Direct	Americas	SCM & Other	United States	Partnership	Professional, Scientific, and Technical Serv	
3DT****LC	Services Only	Closed Won	3/4/2021	3/17/2021	13						EDI Direct	Americas	SCM & Other	United States	Partnership	Professional, Scientific, and Technical Serv	
3M**********AB	Services Only	Closed Won	4/21/2021	4/26/2021	5						EDI Direct	Americas	SCM & Other	United States	Partnership	Professional, Scientific, and Technical Serv	
3xd*********************o.	Renewal	Closed Lost	5/17/2021	11/12/2021	179						Full Circle ERP	Americas	SCM & Other	United States	Partnership	Professional, Scientific, and Technical Serv	
3T ***********o.	Services Only	Closed Won	6/15/2021	6/16/2021	1						Full Circle ERP	Americas	SCM & Other	United States	Partnership	Professional, Scientific, and Technical Serv	
409***********td	Services Only	Closed Won	8/12/2021	8/13/2021	1						Full Circle ERP	Americas	SCM & Other	United States	Partnership	Professional, Scientific, and Technical Serv	
4Fi**************************c.	SaaS/Hosting Mig	Closed Won	3/26/2021	3/9/2021	0						Traverse Global	Americas	SCM & Other	United States	Private	Real Estate	
4L ****oo	Renewal	Closed Won	4/28/2021	11/30/2021	216						Traverse Global	Americas	SCM & Other	United States	Private	Real Estate	
5 S***********************ed	Renewal	Closed Won	4/28/2021	11/12/2021	198						Traverse Global	Americas	SCM & Other	United States	Private	Management of Companies and Enterprise	
612*********************es	New Logo	Closed Lost	1/4/2021	1/26/2021	22						Paragon	EMEA	SCM & Other	United Kingdom	Private	Transportation and Warehousing	
636***********************hl	Renewal	Closed Won	3/3/2020	2/18/2021	352						Made2Manage	Americas	Discrete	Canada	Private	Transportation Equipment Manufacturing	
7 D***********************o.	New Logo	Closed Lost	10/28/2020	8/18/2021	294						JustFood	Americas	Food	Canada	Private	Food Manufacturing	
84 *************ny	Services Only	Closed Lost	1/29/2021	3/18/2021	48						Full Circle ERP	Americas	SCM & Other	United States	Private	Apparel Manufacturing	
90T******LC	Services Only	Closed Lost	11/17/2021	12/10/2021	23						Full Circle ERP	Americas	SCM & Other	United States	Private	Apparel Manufacturing	
A 8*************td	Services Only	Closed Lost	4/15/2021	4/27/2021	12						EDI Direct	Americas	SCM & Other	United States	Private	Apparel Manufacturing	
A 8********ex	Renewal	Closed Won	4/7/2021	5/17/2021	40						Full Circle ERP	Americas	SCM & Other	United States	Private	Apparel Manufacturing	
AF*******************td	Services Only	Closed Won	9/23/2021	10/8/2021	15						Full Circle ERP	Americas	SCM & Other	United States	Private	Apparel Manufacturing	
AG***********ed	Services Only	Closed Won	11/30/2021	11/1/2021	0						Traverse Global	Americas	SCM & Other	United States	Private	Rubber Product Manufacturing	

Figure 7.1: Sample data in masked format

7.2 Describing the data

The initial data extracted from salesforce into excel contains the purchase transactions and customer firmographics data, in which the orders were closed between January 2021 and September 2021. The dataset contained 30596 rows and 19 columns, of which the transactions of each customer were repetitive in case of multiple orders or transactions.

Some of the columns that contained financial data of the organization and also the customer's name, were masked for data breach reasons during this study. Columns like the region, industry, ownership type and product name, and product type are all important feature variables that may help in this study to achieve new sales goals.

In the next phase of data preparation, the data will further be processed and cleaned for accurate analysis and outcome.

7.3 Exploring the data

Data has been explored and some findings could be made from the initial raw data. Exploratory data analysis (EDA), based on the customer and the product transactions have been done separately and several trends, patterns and analysis are done to investigate data sets and summarize their main characteristics.

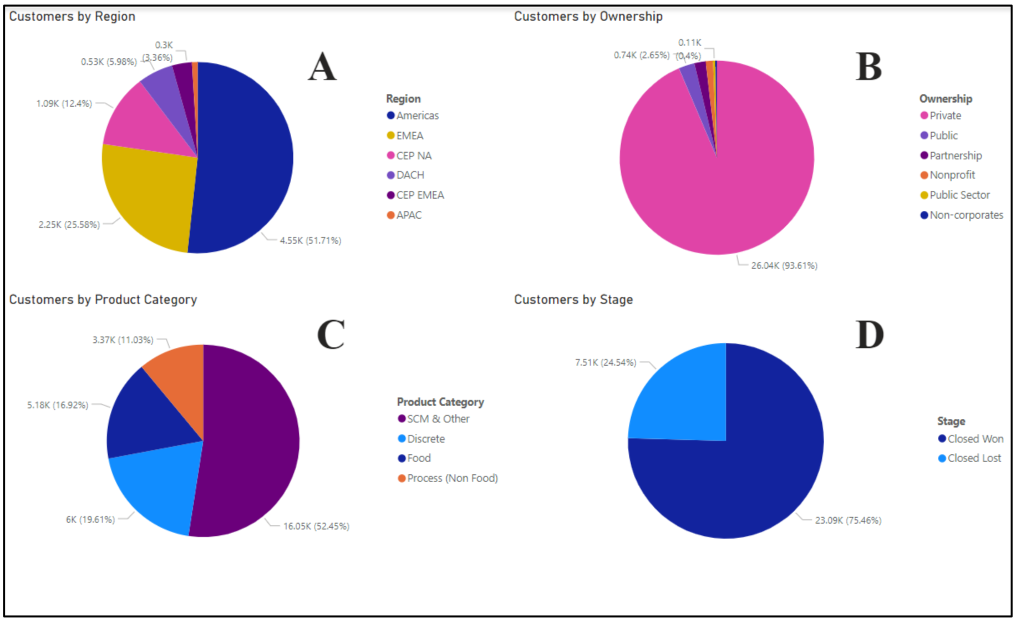


Figure 7.2: EDA based on customer data

Figure 7.2 has 4 pie charts, A, B, C, and D. Here, pie chart A shows the customers by region, and it can be concluded that close to 52% of customers are from the Americas region. Pie chart B shows the customers by ownership, where about 94% of customers are private owners of the business. Pie chart C depicts that 52% of customers have purchased products from the SCM and Others category. Finally , pie chart D shows that out of all the product transactions, only 75% of the opportunities or orders were successfully closed.

The next set of EDA is done on the Product data, where important findings can be made related to the products being sold.

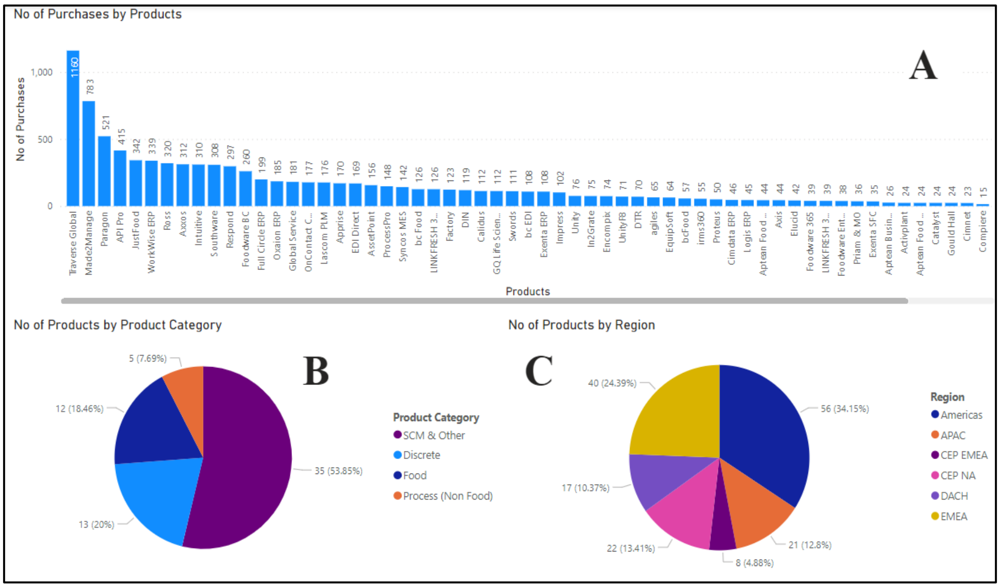


Figure 7.3: EDA based on Product data

Figure 7.3 has visuals A, B and C. Bar chart A shows the top selling products in the descending order. Traverse Global, Made2Manage and Paragon are the top 3 selling products. Pie chart B shows that 35 products fall under the product category SCM and other. This could be reason why the most transactions have happened for the same category. Pie chart C shows the number of products by region and there are 56 products sold in the Americas region, thereby making Americas the major selling site across the globe.

7.4 Verify data quality

There are some quality issues identified from the initial data that was collected. Data had to be masked to maintain the customer and organizations confidential data.

The financial details of all the customers were not available. Hence no cost analysis could be performed.

A product called “BC Food” was also available in the system as “bcFood”, therefore this had to be managed to avoid count mismatches.

There are several lines of transactions where the product and industry information are not rightly entered. This leads to blank values that have to be handled.

Chapter 8: Data Preparation

In the third phase of Data preparation, the final data which is needed for modelling is created. It is a step that requires the most time and effort.

8.1 Data Selection

This step determines which data sets will be used in the project scope. Salesforce is a cloud-based CRM software that is widely used by the Sales, Support and Marketing teams worldwide. Aptean sales team uses salesforce to track business information and hence the product transaction data and the customer firmographics data has been selected and exported into excel format for the analysis.

8.2 Data Integration

As part of the data integration, the product transaction data and the customer firmographics data has been merged to form one combined dataset. The integrated dataset now contains 26 columns and 30596 rows of transactional data. However, this dataset has more than one row of data for each customer causing redundant rows of data for the different dates that the transactions have occurred. We need to handle this by aggregating and grouping the transactions into one single row, such that there is only one row of transaction data per customer.

8.3 Data Cleaning

This is quite often the most time-consuming task. During this task, it is a standard procedure to adjust, impute or eliminate erroneous values. The data now needs to be cleaned and handled for missing values, duplicate and mis-spelled data.

Based on the Exploratory Data Analysis, the domain experts' suggestions and the recommendations from the management and relevant stakeholders, the following features as shown in Table 8.1 have been removed for further analysis in this study.

Features	Reasons for dropping the features
Account Number	Using the unique Account name as the ID
Account Type	Descriptive field that is not important for analysis
Order Type	Descriptive field that is not important for analysis
Stage	Descriptive field that is not important for analysis
Created Date	No analysis done with the transaction created date
Closed Date	No analysis done with the transaction closed date
Age	No analysis done with transaction age
Year 1 ARR	Organization and Customer financials that are masked and not used for any analysis in this project as per the recommendations from Organization and relevant stakeholders.
Total Software Booking Amount	
Services Total	
License Total	
Annual Revenue	
ACV Sub Term Amount	
CapDB Grade	Irrelevant data for study
DNB industry	Duplicate data, same as Industry
Difference	Irrelevant data for study

Table 8.1 Reasons for dropping features

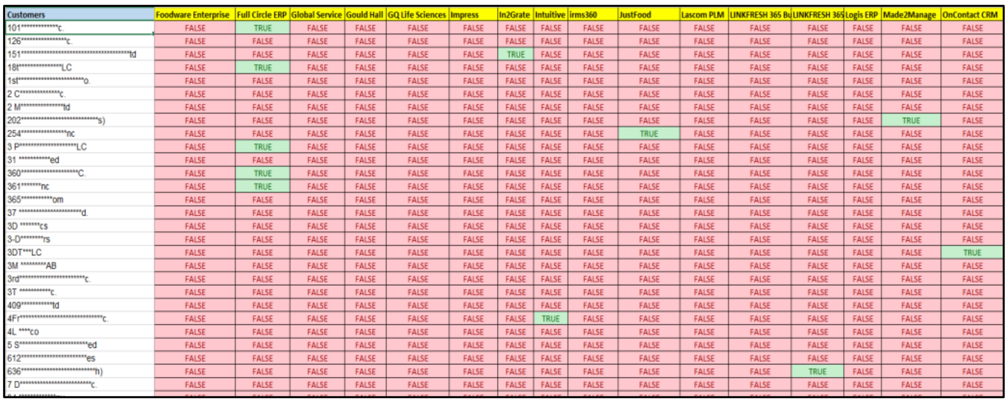
8.3.1 Missing value handling: There were a few rows of customer data that had missing values for customer region, ownership and industry. Since these were all categorical variables, the missing values were replaced with the most frequent value, after observing similar such rows of information from within the dataset. There were also few product names that were not entered in the salesforce system, which led to having blank values, this data was deleted to avoid null value entries.

8.3.2 Handling erroneous/misspelled data: There was a product called “BC Food” which was also available in the system as “bcFood”, these were both referring to the same product. Therefore, the value of “bcFood” was replaced with the correct product name “BC Food”.

8.3.3 Handling duplicate data: There were some customer data that had several rows of duplicated data, that occurred due to data merging. Such data was simply deleted to avoid data redundancy.

8.4 Data Formatting

As part of the analysis, the first step was to create the “Truth Table” to run the market basket analysis algorithm. Truth table contains the customer names against the product names with true and false values that denote their purchase pattern and to analyse purchases that commonly happen together. Figure 8.1 below, shows the truth table built to analyse and study the product affinity from the customers’ purchase pattern.



Customers	Foodware Enterprise	Full Circle ERP	Global Service	Gould Hall	GQ Life Sciences	Impress	In2Grate	Intuitive	irms360	JustFood	Lascom PLM	LINKFRESH 365 Bu	LINKFRESH 365	Logis ERP	Made2Manage	OnContact CRM
101*************c.	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
126***************c.	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
151************************************td	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
18t***************LC	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
1st***********************o.	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
2 C*************c.	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
2 M***************td	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
202****************************s)	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
254***************nc	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
3 P******************LC	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
31 ***********ed	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
360*******************C.	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
361*******nc	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
365***********om	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
37 ***********************d.	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
3D *******cs	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
3-D********rs	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
3DT***LC	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
3M *********AB	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
3rd**********************c.	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
3T ***********c.	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
409***********td	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
4Fr****************************c.	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
4L ****co	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
5 S**********************ed	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
612**********************es	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
636***************************h)	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
7 D************************c.	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

Figure 8.1: Truth Table

8.5 Feature Engineering

Feature engineering is a step that is done to derive new attributes that will be helpful for analysis and prediction in the modelling phase. To build the final dataset for modelling, the dependent variables need to be created. In this case, the dependent variable is the purchase pattern of a customer when two products are bought together. Table 8.2 below details out the features created in this study.

Dependent variables	Explanation
EDI Direct and FULL Circle ERP	When both the products EDI Direct and FULL Circle ERP are bought together
OnContact CRM and Workwise ERP	When both the products OnContact CRM and Workwise ERP are bought together
Foodware 365 and Foodware BC	When both the products Foodware 365 and Foodware BC are bought together
BC EDI and BC Food	When both the products BC EDI and BC Food are bought together

Table 8.2 Features newly created

8.6 Down-sampling

When some modelling techniques were tried on the initially prepared data, the accuracy seemed to be 100%, it wouldn't learn anything interesting, and would have a full rate of false negatives.

To solve this issue, the down sampling technique has been used. This is basically done to work with the imbalanced data in the final dataset. If there is an imbalanced data set, first training is done on the true distribution and if that does not solve the issue, we use the down-sampling technique.

Down-sampling is the process of randomly expelling observations from the majority class to prevent the majority class's signal from controlling the learning algorithm.

The following steps are taken:

- Divide the findings from each class into separate Data groups.
- Resample the class distribution without replacement, using the same number of samples as in the minority class of data.
- Merge the initial minority class data group with the down sampled majority class data group.

Upon down sampling the final dataset , the models seemed to provide more accurate results to the problem. This is further discussed in the modelling phase.

Chapter 9: Modeling

The Modeling phase is the next important phase of the CRISP-DM methodology where several modeling techniques are applied to create and access the models built.

9.1. TURF analysis

“TURF analysis is the Total Unduplicated Reach and Frequency Analysis, which is a statistical research methodology that that ranks combinations of products by how many people will like these combinations” (TURF analysis, 2019).

<table>
<tr><th colspan="6">“Best TURF Results”</th></tr>
<tr><th colspan="6">Statistics</th></tr>
<tr><th>Features</th><th>Size of group</th><th>“Reach”</th><th>“% of Cases”</th><th>“Frequency”</th><th>“% of Responses”</th></tr>
<tr><td>ADDED: TraverseGlobal</td><td>1</td><td>1160</td><td>13.2</td><td>1160</td><td>16.1</td></tr>
<tr><td>ADDED: Made2Manage
KEPT: TraverseGlobal</td><td>2</td><td>1941</td><td>22.1</td><td>1943</td><td>27.0</td></tr>
<tr><td>ADDED: Paragon
KEPT: Made2Manage, TraverseGlobal</td><td>3</td><td>2462</td><td>28.0</td><td>2464</td><td>34.3</td></tr>
<tr><td>ADDED: APIPro
KEPT: Made2Manage, Paragon, TraverseGlobal</td><td>4</td><td>2877</td><td>32.7</td><td>2879</td><td>40.0</td></tr>
<tr><td>ADDED: JustFood
KEPT: APIPro, Made2Manage, Paragon, TraverseGlobal</td><td>5</td><td>3216</td><td>36.5</td><td>3221</td><td>44.8</td></tr>
<tr><td>ADDED: WorkWiseERP
KEPT: APIPro, JustFood, Made2Manage, Paragon, TraverseGlobal</td><td>6</td><td>3549</td><td>40.3</td><td>3560</td><td>49.5</td></tr>
</table>

Table 9.1: Turf Analysis

Table 9.1 shows the best frequency and reach values among each group size. Unduplicated reach describes the proportion of customers who selected at least one of the products within a portfolio. Frequency is the number of items desired within the portfolio. Here, in the group size 1 "Traverse Global" product has the reach and frequency of 1160, this is also the total number of customers who purchased this product. The percent of cases or respondents saying "yes" to the product, which is 13.2% of total customers who have reach to this product and the Percent of response is the single response out of total responses from the given dataset which is 16.1% of total customers who respond to the product. Similar is the understanding for the other group sizes, where the percentage of cases and responses keeps increasing with increasing group size, where the number of products increases.

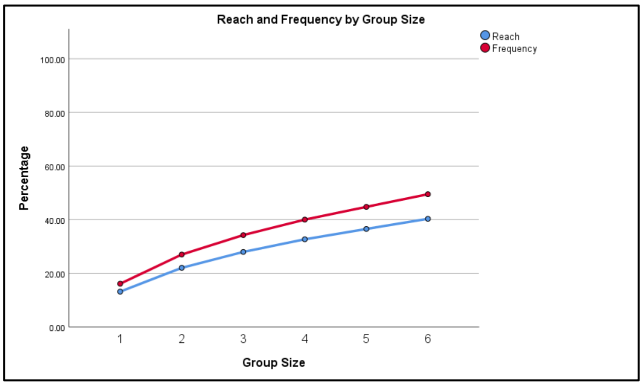
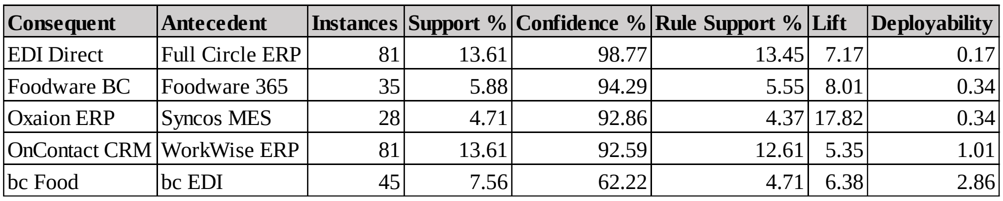


Figure:9.1: GGraph for Turf analysis

Figure 9.1 depicts the GGraph for reach and frequency by group size. The maximum group size was set to 6 and so the graph shows the reach and frequency percentages that increase per increase in the group size.

9.2 Market Basket Analysis using Apriori Algorithm.

Market Basket Analysis (MBA) discovers product associations by looking for product combinations that often co-occur in transaction data. Apriori is the most basic algorithm for mining frequent patterns from transaction information and pattern mining has indeed been widely used in market basket analysis to reveal hidden patterns in transactional data (Gupta et al., 2016).



Consequent	Antecedent	Instances	Support %	Confidence %	Rule Support %	Lift	Deployability
EDI Direct	Full Circle ERP	81	13.61	98.77	13.45	7.17	0.17
Foodware BC	Foodware 365	35	5.88	94.29	5.55	8.01	0.34
Oxaion ERP	Syncos MES	28	4.71	92.86	4.37	17.82	0.34
OnContact CRM	WorkWise ERP	81	13.61	92.59	12.61	5.35	1.01
bc Food	bc EDI	45	7.56	62.22	4.71	6.38	2.86

Figure 9.2: Output of MBA

Some of the main terms to understand here, are:

- **Items:** The products that we are identifying affinity between.
- **Antecedent**: The items on the LEFT, which the customer purchases.
- **Consequent:** The items on the RIGHT, which the customer follows to buy.
- **Support**: "Support of a product or set of products is the fraction of transactions in our data set that contain that product or set of products."
- **Confidence**: "Confidence is a conditional probability that customer purchasing product A will also buy product B".
- **Lift:** "Lift is if someone purchases product A, then what % of chance of buying product B would increase" (Gupta et al., 2016).

From the Figure 9.2, the understanding is that if a customer purchases a product from the antecedent column, he is more likely to purchase the consequent. In the dataset, we have 81 instances where "EDI Direct" and "Full Circle ERP" are bought together. Where support is 13.6%, Confidence is 98.7% and Lift is 7.17 which is above 1 indicating that the output response is more likely to occur than the average response. Therefore, the association rules improvise the likelihood of the results. Figure 9.3 shows the web chart with strong product associations.

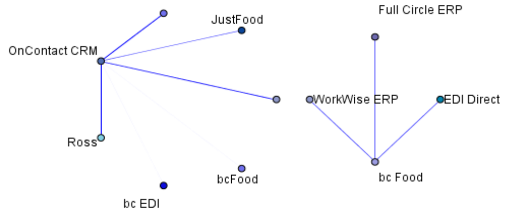


Figure 9.3: Web-Chart showing product associations

9.3 Decision Tree - CHAID

Decision trees are one of the effective methods for data mining. The tree structure provides an easy way to interpret the results. According to the correlation between products based on the association rules, there are different target groups of customers. To get the characteristics of the target customer, variables such as Customer Region, Product Selling Region, Industry, Ownership and Product Type are used as the independent variables. CHAID method (Chi-squared Automatic Interaction Detector) has been used as a growing method and the key drivers are – Product Region, Product Type and Ownership.

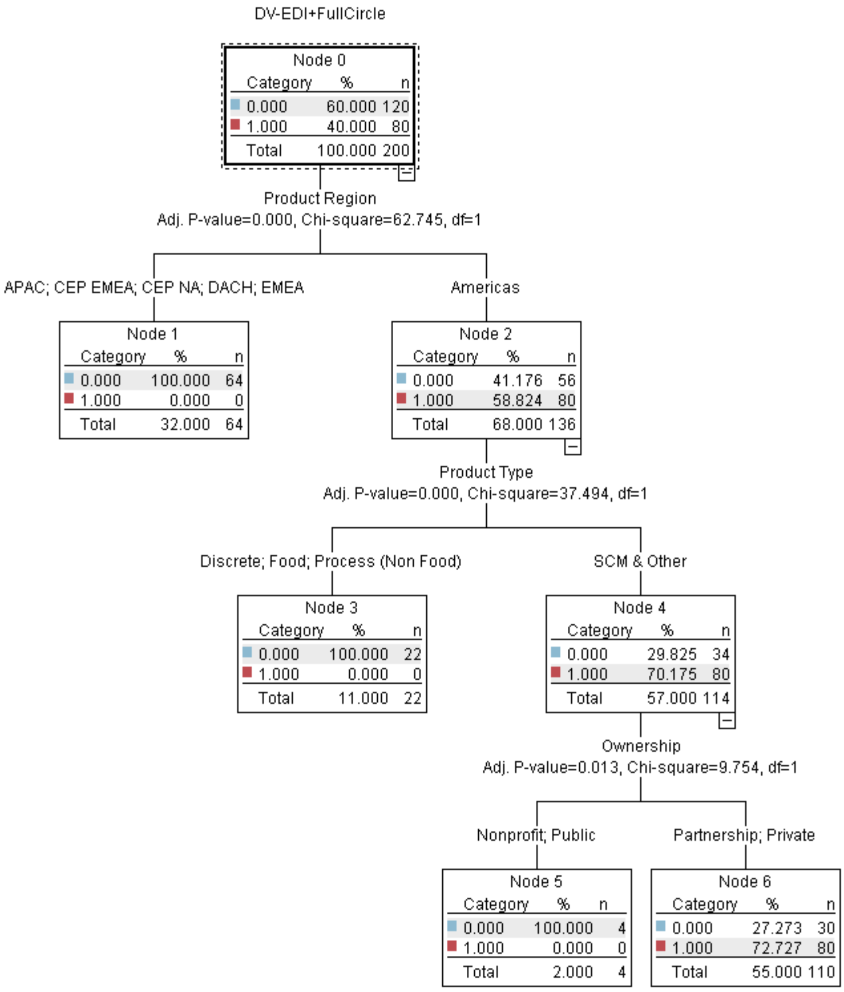


Figure 9.4: CHAID- Decision Tree Output

Figure 9.4 shows the CHAID decision tree output and the key factors responsible are depicted in the Figure 9.5 shown below, which is the predictor importance chart from which the Association Rule is made, stating that if Product Region is "Americas", Product Type is "SCM & Others" and Ownership is either "partnership or private", then there is a high probability that the customer will purchase both products "EDI Direct" and "Full Circle ERP" together.

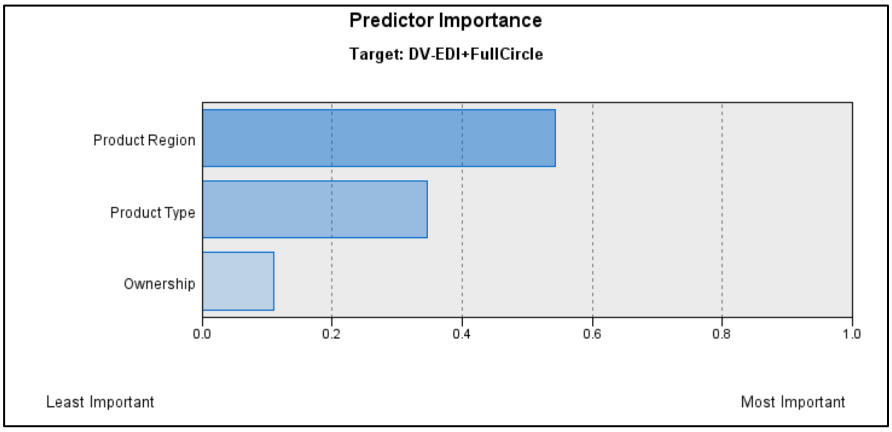


Figure 9.5 Predictor importance chart

9.4 Supervised Machine Learning: K- Nearest Neighbour

"KNN algorithm is a method for classifying cases based on their similarity to other cases. In machine learning, it was developed to recognize patterns of data without requiring an exact match to any stored patterns, or cases. Similar cases are near each other, and dissimilar cases are distant from each other. Thus, the distance between two cases is a measure of their dissimilarity."

From figure 9.6 we can see the distribution of the different products based on the 5 predictors, there is a relation between the product type "Process non-food" in the region of Sweden and Poland. Since this gives the distribution of the association between products in a 5- dimensional space, we cannot make accurate groupings of data categories

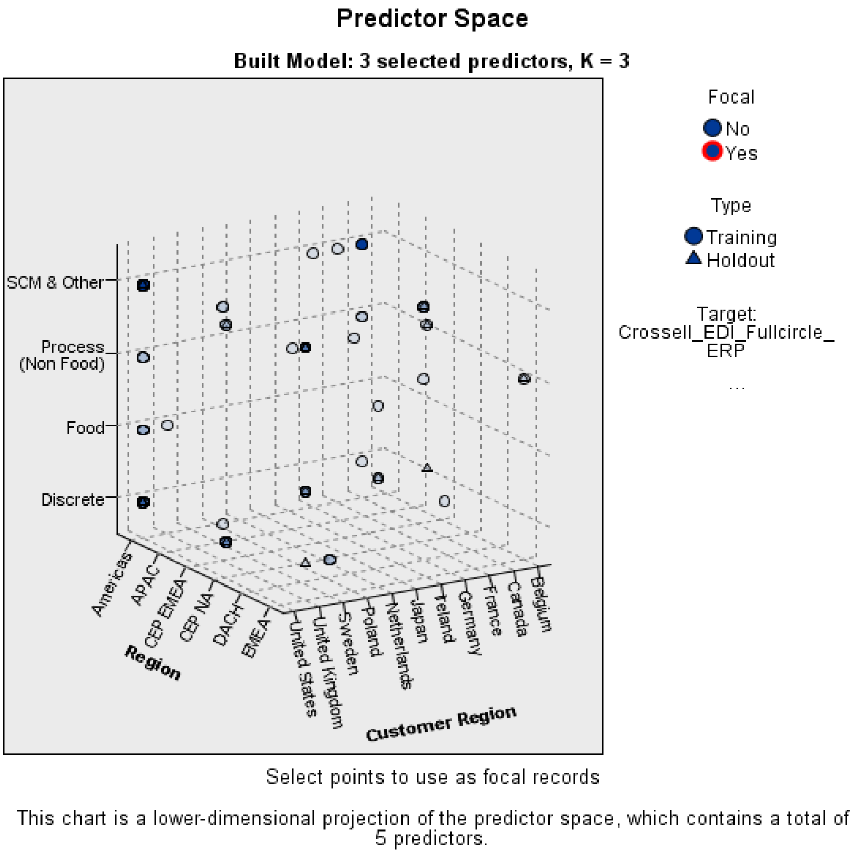


Figure 9.6: Output of KNN Algorithm.

9.5 Deep Learning - Multilayer Perceptron

Finally, the last model is built using the Deep learning method, in which the multilayer perceptron is used.

“The Multilayer Perceptron (MLP) procedure produces a predictive model for one or more dependent (target) variables based on the values of the predictor variables. It is a fully connected class of feedforward artificial neural network. The term MLP is used ambiguously, sometimes loosely to mean any feedforward, sometimes strictly to refer to networks composed of multiple layers of perceptron.”

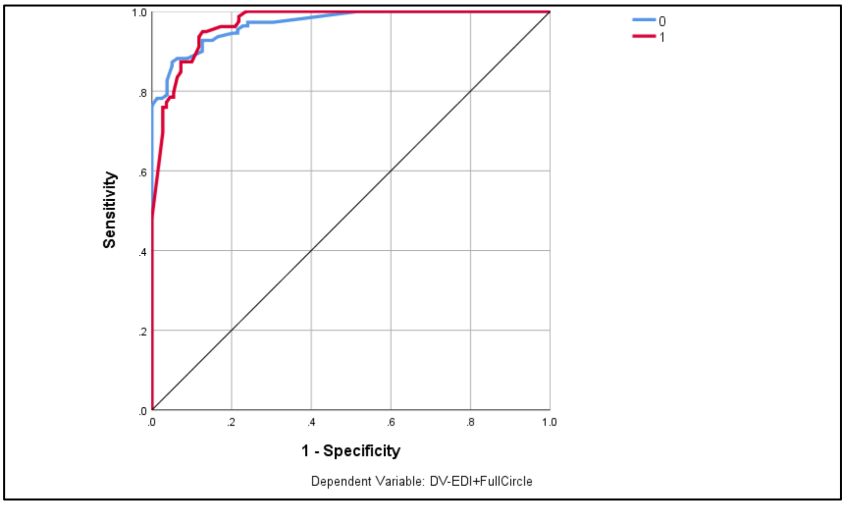


Figure 9.7 Sensitivity-Specificity chart

From figure 9.7 above, the Sensitivity-Specificity chart tells us the quality of the model. Since the AUC for this combination of products EDI and full circle ERP is 0.97, the model has a high accuracy and better than the baseline model.

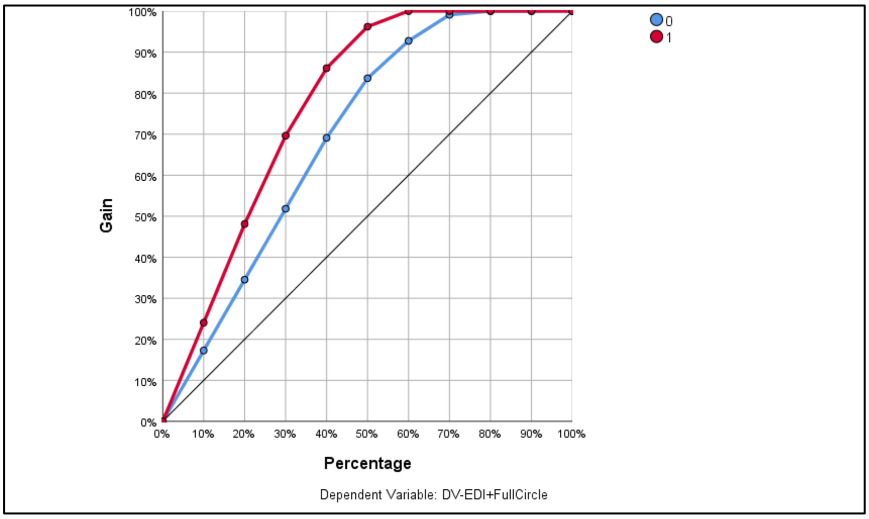


Figure 9.8 Gain chart

The above graph in Figure 9.8, shows the gain percentage for the different decile distribution. We observe that the gain percentage is the highest at the 60th percentage.

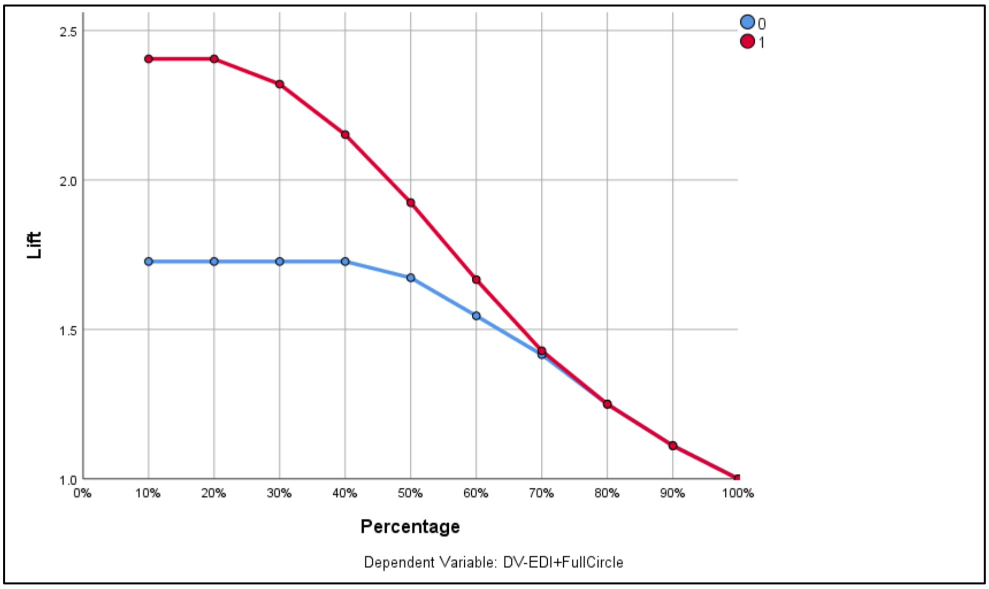


Figure 9.9 Lift chart

This lift chart as shown in Figure 9.9 depicts that lift is the highest at 10% and starts decreasing after the 60th percentile.

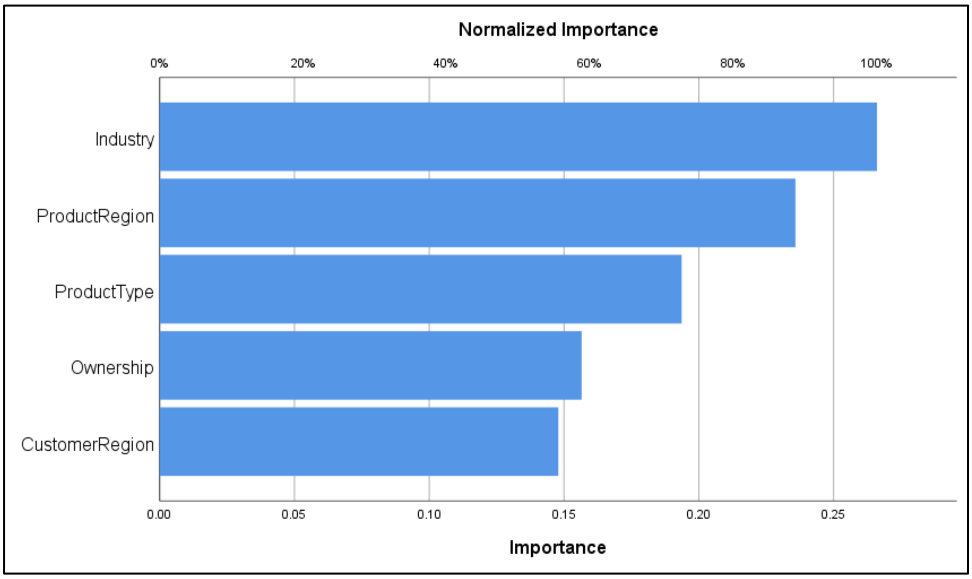


Figure 9.10 Normalised Importance chart

Figure 9.10 gives the importance chart which depicts the importance of the independent variables in the descending order. Where the normalised importance score of Industry is 100%, Product region is 88.6%, Product type is 72.8%, Customer region is 55.6% and Ownership is 58.8%

Chapter 10: Model Evaluation

In this section, an overview of the results obtained from different machine learning procedures is provided and explained.

10.1: Classification Report

“A classification report is a performance evaluation metric in machine learning. It is used to show the precision, recall, F1 Score, and support of the trained classification model” (Aman Kharwal, 2021).

Classification			
	Forecasted		
Actual	0	1	Correct %
0	116	4	96.70%
1	13	67	83.80%
Overall %	64.50%	35.50%	91.50%

Table 10.1: Classification Report for CHAID Decision Tree

From Table 10.1, it is observed that the overall accuracy of the model is close to 92% indicating that the model is good.

Classification			
Actual	Forecasted		
	0	1	Correct%
0	79	5	94.00%
1	17	42	71.20%
Overall Percent	67.10%	32.90%	84.60%

Table 10.2: Classification Report for KNN

From Table 10.2, it is observed the overall accuracy of the model is 85% indicating that the model is good, but not as good as the CHAID Decision tree algorithm.

We now model using the deep learning methodology of multilayer perceptron (MLP). In this model we divide the dataset into train and test where 70% of the data is utilized to train the model and remaining 30% is utilized for testing.

Classification				
		Forecasted		
Training actual	0	71	9	88.8%
	1	4	58	93.5%
	Overall Percent	52.8%	47.2%	90.8%
Testing actual	0	25	5	83.3%
	1	0	17	100.0%
	Overall Percent	53.2%	46.8%	89.4%

Table 10.3: Classification Report for Multilayer Perceptron

From Table 10.3, it is observed the overall accuracy of the model is 91% for the training data and 89% for the test data. indicating that the model is good, but not as good as the CHAID Decision tree algorithm. However, it is better than the KNN methodology used. From the Classification report tables 10.1, 10.2, 10.3, we could collect the confusion matrix values such as the "True Positive (TP) which is the correctly predicted event values, False Positive (FP) which is the incorrectly predicted event values, True Negative (TN) which is the correctly predicted no-event values, and False Negative (FN) which is the incorrectly predicted no event values".

10.2: Model Performance metrics

Almost all model-performance metrics are calculated on the model's predictions being compared to the value of the dependent variable in a dataset. Formulae to calculate the model performance metrics:

"Accuracy = (TP+TN) / (TP+TN+FP+FN)"

"Precision = TP/ (TP+FP)"

"Recall = TP / (TP+FN)"

"F1 score = (2* Precision* Recall) / (Precision +Recall)

Model Performance metrics				
Models	**Precision**	**Recall**	**F1-score**	**Accuracy**
Decision Tree	94%	84%	89%	92%
KNN	89%	71%	79%	85%
MLP- Test	77%	100%	87%	89%
MLP-Train	87%	94%	90%	91%

Table 10.4: Model performance metrics

From Table 10.4, it can be concluded that the Decision tree algorithm using the CHAID mechanism is the best model that could be built for the analysis of the right products that can be sold to customers. It has the highest precision and recall values of 94% and 92% respectively proving to be the best model.

10.3: Accuracy Score

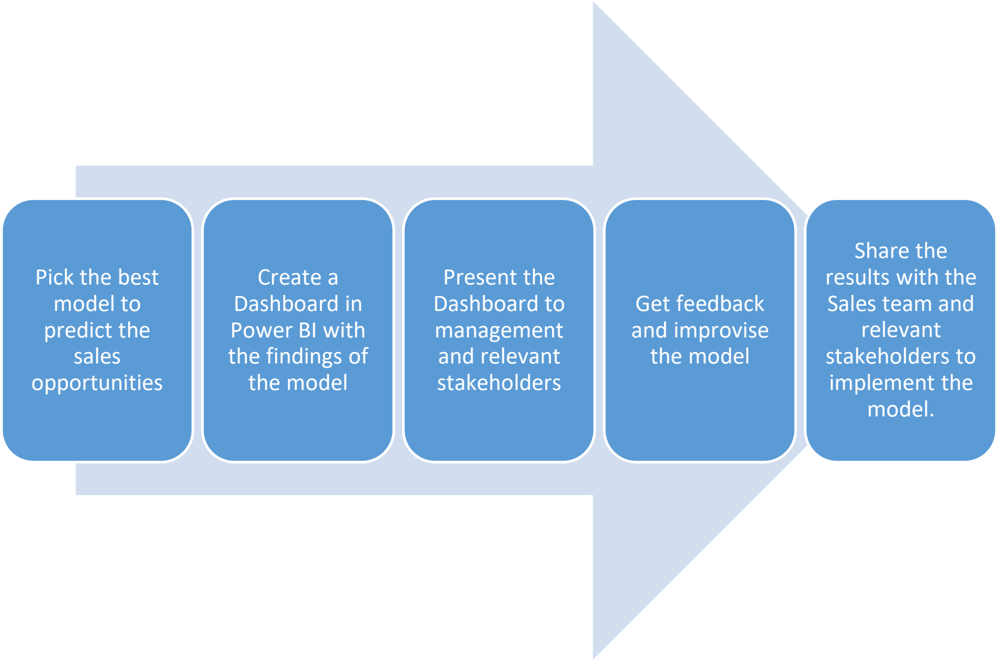
One metric for evaluating classification models is accuracy. In an informal way, “accuracy is the percentage of correct predictions made by our model”. Formally, “accuracy is defined as Number of correct predictions/ Total number of predictions” Table 10.5 below shows that the decision tree is the best model out of all the models that were built.

Models	Accuracy
Decision Tree	92%
KNN	85%
MLP- Test	89%
MLP-Train	91%

Table 10.5: Accuracy score

Chapter 11: Deployment

Model deployment is the process of putting machine learning models into production. This makes the model's predictions available to relevant stakeholders, so they can make business decisions based on the data. The best model is chosen and used to build similar analysis for new datasets so that we could propose increased product sales through product cross-selling or product up-selling.



Pick the best model to predict the sales opportunities

Create a Dashboard in Power BI with the findings of the model

Present the Dashboard to management and relevant stakeholders

Get feedback and improvise the model

Share the results with the Sales team and relevant stakeholders to implement the model.

Figure 11.1: Deployment Plan

Upon implementation as described in Figure 11.1. The final dashboards are built in Power BI which is the most commonly used tool at Aptean by the Executive leaders to view the progress of all teams within the organization.

Chapter 12: Analysis and Results

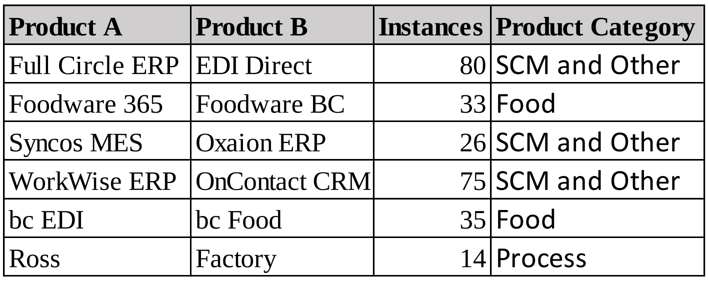
In the study presented in this paper, the dependent variable is the variable indicating the purchase pattern of the customers who have purchased 2 or more products together and the independent variables are continuous variables that are used to segment and profile the customers and the products to enable up-sell and cross-sell opportunities.

From the model evaluation though 3 different machine learning approaches were incorporated. The most efficient of the 3 is the Decision tree using the CHAID method, because of its high accuracy and precision value.

Key drivers influencing the product sales as shown in the decision tree are –
Product Region – Selling region of the product
Product Type – To which category of product it belongs
Ownership – The type of ownership of the customer company.

Other factors such as customer region and Industry are having huge value categories. Therefore, they have not played an important role in the decision tree algorithm for predicting the products that can be sold together.

The several combinations that can be sold together are as given below. This could be found from the market basket analysis backed by the Decision tree CHIAD algorithm.



Product A	Product B	Instances	Product Category
Full Circle ERP	EDI Direct	80	SCM and Other
Foodware 365	Foodware BC	33	Food
Syncos MES	Oxaion ERP	26	SCM and Other
WorkWise ERP	OnContact CRM	75	SCM and Other
bc EDI	bc Food	35	Food
Ross	Factory	14	Process

Figure 12.1: Cross-selling Products

From table 12.1 we can understand the scope of more products that could be sold together to a customer. In this case, if the Customer purchases product "A" he is more likely to purchase product "B". There are instances already existing in the dataset provided. When backed by the decision tree modelling technique, we can conclude on the association rules to optimize the results.

Chapter 13: Conclusions and Recommendations for future work.

The main objective of the project was to understand the purchasing pattern of products from the product sales transaction data, study and profile customers based on their purchase behavior and to recommend and suggest products to customers, thereby increasing cross-sell and up-sell opportunities.

One reason for market basket analysis' growing acceptance in the data research fields is that researchers can evaluate the existence of association rules by using an inductive approach to theorizing. Taking everything into account, a recommendation system can significantly make an effect on marketing and sales studies that can be used to derive at strategic business decisions.

In this project, different modelling techniques have been tried and evaluated to find out the key drivers responsible for cross-selling certain products. We could profile customers that belong to different categories based on these key drivers and propose that for new customers who belong to any of these categories, such products could be sold, thereby increasing sales opportunities in organization and enabling the organization to reach its new goal of achieving sales targets and increasing customer base and maintain niche enterprise products.

Recommendations for further work:

This project does not cover the cost and financial analysis, if the financial data could be used for analysis, we could probably recommend the best possible products for upselling or cross-selling thereby increasing sales.

A similar analysis can be used to model other combinations of data in which more than 2 products are sold together. As of today, the data for this kind of analysis is almost negligible, that is, out of the total 8799 customer base there are only 9 customers who have purchased 4 products together.

Bibliography

Aman Kharwal. (2021). *Classification Report*. https://thecleverprogrammer.com/2021/07/07/classification-report-in-machine-learning/#:~:text=A%20classification%20report%20is%20a,this%20article%20is%20for%20you.

Chen, Z. Y., Fan, Z. P., & Sun, M. (2015, July 28). A SVM ensemble learning method using tensor data: An application to cross selling recommendation. *2015 12th International Conference on Service Systems and Service Management, ICSSSM 2015*. https://doi.org/10.1109/ICSSSM.2015.7170282

David Gargaro. (2022). *12-ways-to-increase-sales*. https://www.business.com/articles/12-ways-to-increase-sales/

Dookeram, N., Hosein, Z., & Hosein, P. (2022). A Recommender System for the Upselling of Telecommunications Products. *International Conference on Advanced Communication Technology, ICACT, 2022-February*, 66–72. https://doi.org/10.23919/ICACT53585.2022.9728818

Gupta, T., Karthiyayini, R., Balasubramanian Professor, R., & And, P. (2016a). Affinity and Association Affinity Analysis and Association Rule Mining using Apriori Algorithm in Market Basket Analysis. In *International Journal of Advanced Research in Computer Science and Software Engineering* (Vol. 6, Issue 10). www.ijarcsse.com

Gupta, T., Karthiyayini, R., Balasubramanian Professor, R., & And, P. (2016b). Affinity and Association Affinity Analysis and Association Rule Mining using Apriori Algorithm in Market Basket Analysis. In *International Journal of Advanced Research in Computer Science and Software Engineering* (Vol. 6, Issue 10). www.ijarcsse.com

Hewen, T., Zengfang, Y., Pingzhen, Z., & Honglin, Y. (2008). Using data mining to accelerate cross-selling. *2008 International Seminar on Business and Information Management, ISBIM 2008, 1*, 283–286. https://doi.org/10.1109/ISBIM.2008.186

Raorane AA, Kulkarni RV, & Jitkar BD. (2012). *Association Rule-Extracting Knowledge Using Market Basket Analysis*. www.isca.in

Shruthi Gurudath. (2020). *Market Basket Analysis & Recommendation System Using Association Rules*. https://doi.org/10.13140/RG.2.2.16572.05767

Think Insights. (2022). *CRISP-DM – A framework for Data Mining & Analysis*. https://thinkinsights.net/data-literacy/crisp-dm/

Turf Analysis. (2019). https://conjointly.com/blog/turf-analysis/

Wikipedia. (2022). *Upselling*. https://en.wikipedia.org/wiki/Upselling

Appendix

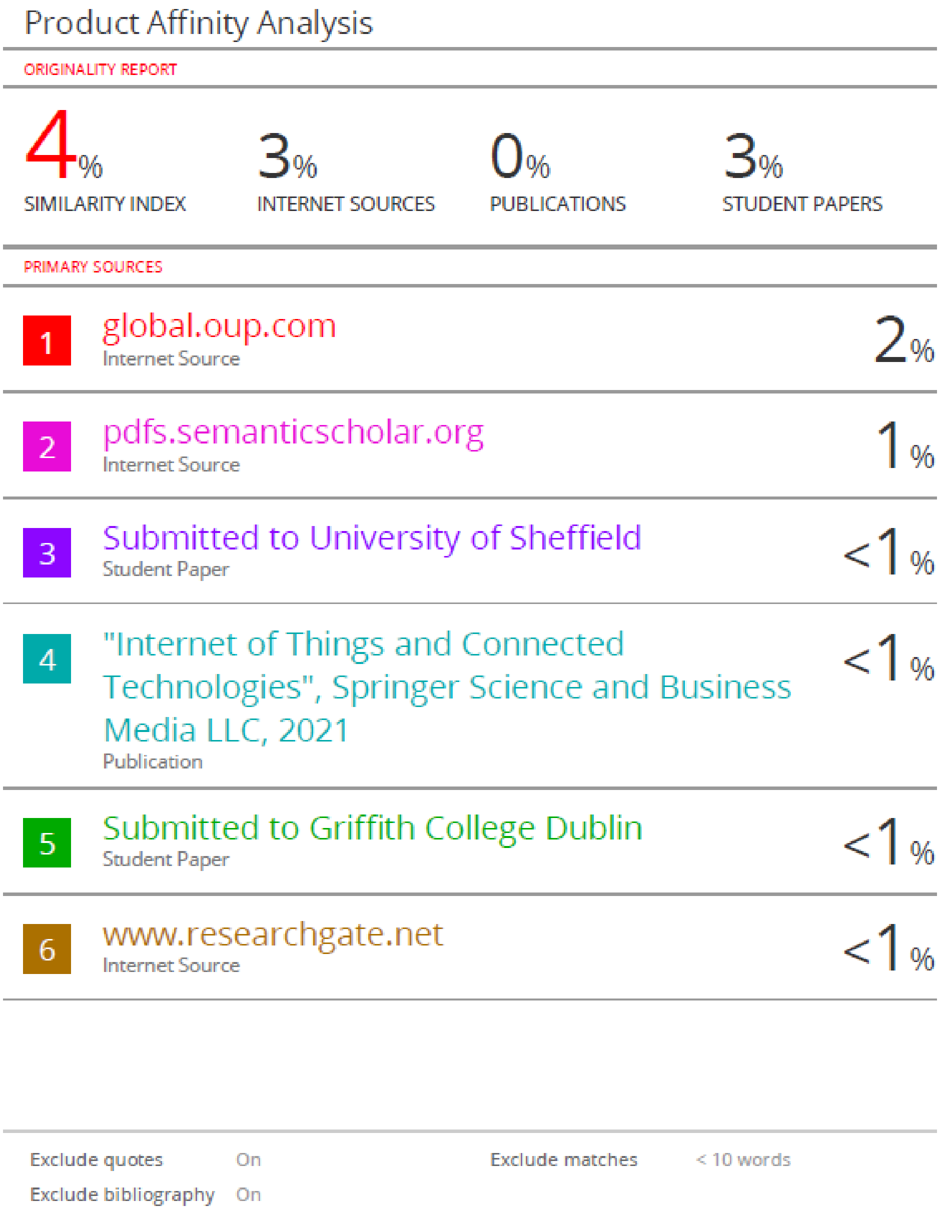
Plagiarism Report[1]

Product Affinity Analysis

by Sharon Joseph

Submission date: 26-Aug-2022 11:22PM (UTC+0530)
Submission ID: 1887572415
File name: Product_Affinity_Analysis.docx (1.11M)
Word count: 6751
Character count: 36685

[1] Turnitin report to be attached from the University.



Product Affinity Analysis

ORIGINALITY REPORT

4%	3%	0%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	global.oup.com Internet Source	2%
2	pdfs.semanticscholar.org Internet Source	1%
3	Submitted to University of Sheffield Student Paper	<1%
4	"Internet of Things and Connected Technologies", Springer Science and Business Media LLC, 2021 Publication	<1%
5	Submitted to Griffith College Dublin Student Paper	<1%
6	www.researchgate.net Internet Source	<1%

Exclude quotes	On	Exclude matches	< 10 words
Exclude bibliography	On		

Publication in a Conference

Paper Submitted:

Sharon Joseph, Mithun D J, Rashmi Agarwal "Product Affinity Analysis to Increase Sales using Machine Learning." Third International Conference on Smart Technologies in Computing,

Electrical and Electronics (ICSTCEE'22), Submission Date: 23rd September 2022

Product Affinity Analysis to Increase Sales using Machine Learning

Sharon Joseph
Reva Academy for Corporate Excellence – RACE REVA University
Bangalore, Karnataka 560064
sharon.ba06@reva.edu.in

Mithun D J
Reva Academy for Corporate Excellence – RACE REVA University
Bangalore, Karnataka 560064
mithun.dj@reva.edu.in

Rashmi Agarwal
Reva Academy for Corporate Excellence – RACE REVA University
Bangalore, Karnataka 560064
rashmi.agarwal@reva.edu.in

***Abstract*— Decision-making and understanding customer behavior has become critical and crucial for companies wanting to maintain their position in today's competitive markets. Every business aims to improve its revenue and profits by increasing sales. The objective of this study is to leverage customer firmographic data and product sales transaction data, which is drawn from the organization's internal Salesforce system, to build a solution to project the likelihood of a purchase from our existing customer base. The capacity of individuals to identify cross-sell and up-sell opportunities would be improved with a greater understanding of what was sold, why it was sold, and to whom. It also discovers the associations among products and predicts the products that could be projected for potential sales opportunities. Machine learning algorithms like Market Basket Analysis (MBA) using Apriori, Total Unduplicated Reach and Frequency Analysis (TURF) for frequency study, Chi-square Automatic Interaction Detector (CHAID) algorithm for the Decision tree, K-Nearest Neighbour (KNN) and Multilayer Perceptron (MLP) as part of Deep Learning are the techniques used to derive the desired outcome to the problem. The specific outcome includes product affinity analysis and recommendation of products that can be cross-sold or upsold to existing customers or new customers, where the decision tree algorithm achieves the best results among the other machine learning algorithms. Organizations can profile customers that belong to different categories based on these key drivers and propose the same for new customers who belong to any of these categories. Such products could be sold, thereby increasing the sales opportunities in the organization and enabling the organization to reach its goal of achieving sales targets, increasing the customer base and maintaining niche enterprise products.**

Keywords*— *Machine Learning, Product Affinity Analysis, Cross-Selling, Market Basket Analysis, Apriori, CHAID, KNN, Deep Learning, Multilayer Perceptron

I. Introduction

Every business aims to improve its revenue and profits, and this is mainly achieved through increasing sales. Factors which can increase sales are, acknowledging the current customer behavior, requesting customer feedback, running promotions, determining sales strategies, launching sales presentation techniques and methods and providing excellent customer service. Other traditional methods are, creating packages, deals and free trials to attract customers, conducting a content audit, doing something noteworthy or unique, optimizing social media profiles, advertising on social media platforms, spreading by word of mouth, putting a call to action on users' website, and to stay in touch with email marketing [1]. However, this study focuses on some machine learning techniques to promote cross-selling and up-selling opportunities.

"Cross-selling is a sales technique involving the selling of an additional product or service to an existing customer whereas, Up-selling is a sales technique where a seller invites the customer to purchase more expensive items, upgrades, or other add-ons to generate more revenue" [2].

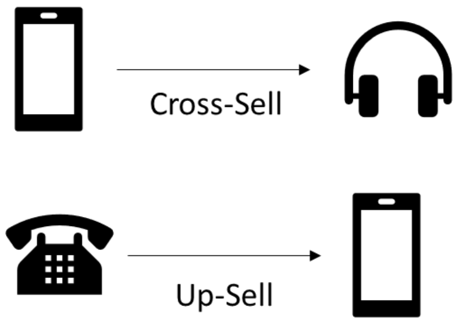


Fig. 1. Cross-sell vs Up-sell

From Fig. 1, cross-sell indicates that headphones can be sold to a customer who has purchased a mobile phone and up-sell indicates that a mobile phone could be sold to a customer who purchased a telephone. Here, the headphone is a complementary product that may interest a customer, whereas a mobile phone is an upgraded version of a telephone that a customer may purchase.

Product co-occurrence analysis is one area of research that can help determine the relationship between various products. Affinity analysis and association rule learning are two analytics techniques that aim to discover the relationships and connections between specific objects. Market Basket Analysis (MBA) is probably the most well-known example that identifies product associations by looking for product combinations that frequently co-occur in exchanges. Individuals who purchase flour and sugar, also buy eggs since a large proportion of them intend to bake a cake [3].

In this study, the aim is to identify and understand the customers' purchase behavior from the product sales transaction data and customer firmographics data, to find out the products that are more likely to be cross-sold or up-sold to the different customer categories. This would help the sales team to improve sales and increase revenue, by promoting the existing customers to purchase better or related products.

II. STATE OF ART

Extensive literature reviews have been done on various topics related to affinity analysis and the different methodologies used to analyse product cross-selling and upselling.

The primary purpose of the work of the author here is to build a data mining method in excel using an XLMiner add-in tool to accelerate cross-selling. It does not necessitate a great deal of expertise in data mining but only some parameter setting. Furthermore, almost everyone is proficient in Excel. Therefore, the proposed excel-based method is simple to learn and it also presents an example through mining association rules [4].

The main objective of the work done by the author is to analyse large datasets thereby exploiting consumer behavior and making the correct decision leading to a competitive edge over rivals. Experimental analysis has been done employing association rules using MBA to prove its worth over the conventional methodologies. Online Analytical Processing (OLAP) tools have been commercially used for in-depth analysis such as data classification, clustering and characterization of data that changes over time [5].

Support Vector Machine (SVM) is a supervised machine learning algorithm used for both classification and regression. The ensemble learning method is proposed for classification using tensor data. The method is used in identifying cross-selling opportunities to recommend personalized products and services to customers. Two real-world databases are used to evaluate the performance of the method and the SVM ensemble learning method, is proposed for classification using tensor data. Computational results show that the SVM ensemble learning method has good performance on these databases [6].

The next set of authors have mostly used the Apriori algorithm for MBA and revels, the relationship between the purchased goods by using the Apriori algorithm to find out the data of shopping baskets from the massive data of consumers and then applying the association rules and Classification And Regression Tree (CART) algorithm to reveal the characteristics of the customer group and the target customers' classification. It is convenient for the goods to be better configured and sold to extract more detailed and valuable information, as well as to improve the market's operational efficiency[7].

The author here, claims that affinity analysis and association rule mining encompasses a broad set of analytics techniques aimed at revealing the associations and correlation between specific objects. The purpose of this analysis is to generate a set of rules that relate two or more products together. Each of these rules should have a lift greater than one. The interest is in the support and confidence of those rules such as higher confidence rules are ones where there is a higher probability of items on the Right Hand Side (RHS) being part of the transaction given the presence of items on the Left Hand Side (LHS). "*R*" is a great statistical and graphical analysis tool, well suited to more advanced analysis which is used to perform the MBA [8].

Alternatively, the use of anonymized data from customers' transactional orders to focus on descriptive analysis of customer purchase patterns, items purchased together, and units purchased frequently from the store to facilitate reordering and maintaining adequate product stock, is possible to accomplish this by analyzing the available data in such a way that a frequent item set can be identified and analyzed to the association rule. The Apriori algorithm aids in the discovery of association rules for frequent item sets and the identification of correlations and the same is developed to investigate approaches for applying association rules to recommender systems [9]**.**

The latest study in 2022 by the authors states a binary classification framework for predicting the successful upsell of products and services, using data from a telecommunications service provider. Through this prediction model, the recommender system for voice products and services to corporate customers of the telecommunications company is demonstrated. Logistic regression classifier to automate the selection of customers that are most likely to upsell. Application of a predictive model to recommend a set of target customers to approach for upselling, illustrating the different accuracy results for different cost weightings and also showing that the success rate of upselling products to the selected customers is dramatically improved when compared to the traditional approach [10].

III. PROBLEM DEFINITION

Traditional methods like publishing about the products on the company website, digital marketing, advertisements, and setting up campaigns are followed to sell the different products to customers. The organization focuses on delivering purpose-built, industry-specific, mission-critical enterprise software, tailored to the industry's needs and so it is a time-consuming process to identify potential customers and to sell products to them, which includes their requirements and customizations.

Leveraging customer firmographic data and product sales transaction data to build a model to project the likelihood to purchase from our existing customer base and also solve the business challenge of achieving and meeting the sales targets of the organization is time-consuming. Increasing sales through product cross-selling and up-selling and identifying the customers to whom products can be sold is challenging.

IV. METHODOLOGY

This paper uses the Cross Industry Standard Process for Data Mining (CRISP-DM) framework to implement and carry out the different phases of the data mining and machine learning processes involved in this study.

The CRISP-DM methodology mainly consists of six phases and all these phases have an important role in the implementation of the model.

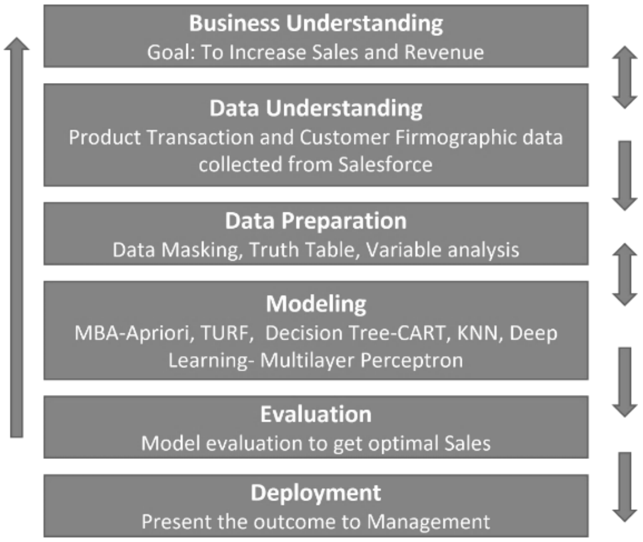

Business Understanding
Goal: To Increase Sales and Revenue
Data Understanding
Product Transaction and Customer Firmographic data collected from Salesforce
Data Preparation
Data Masking, Truth Table, Variable analysis
Modeling
MBA-Apriori, TURF, Decision Tree-CART, KNN, Deep Learning- Multilayer Perceptron
Evaluation
Model evaluation to get optimal Sales
Deployment
Present the outcome to Management

Fig. 2. Workflow of the Model

Fig. 2 illustrates the detailed workflow of the model developed. Initially, as part of the business understanding, discussions and meetings with the relevant stakeholders are conducted. The goal of maximizing the sales and the revenue of the organization is captured and the project plan is created. The data such as the product sales transaction data along with the customer firmographics data is collected from the internal Salesforce system. Later the data is organized, explored, visualized and verified for data quality. As part of the data preparation, relevant data is selected, cleaned, constructed, integrated and re-formatted to run the various techniques for analysis and modelling. Exploratory data analysis is carried out and different modelling techniques are performed. The models are then assessed, evaluated for the results and reviewed. Finally, the model is iterated to verify if the business objectives are met and deployed for end users. In this case, the deployment was to present the findings of this study to management and provide them with recommendations on what products can be sold and to which category of customers probable sales can be proposed.

V. DESCRIPTION OF DATASET

The data understanding phase of CRISP-DM methodology includes processes to identify, collect, and analyse datasets that will assist in meeting the project objectives. The data required for this study is stored in the Salesforce system, which is managed by the organization. Since the data involved customer firmographics and certain financial data, the requirement was to use the data in a masked format, so that the organization and customers' confidentiality can be maintained, and no data privacy is breached.

The initial data is extracted from salesforce into excel which contains the purchase transactions and customer firmographics data. In this file, the data about the orders which are completed between January 2021 and September 2021, is considered. The dataset contains 30596 rows and 19 columns, of which the transactions of each customer were repetitive in case of multiple orders or transactions.

Some of the columns that contained financial data of the organization and also the customer's name, were masked for data breach reasons during this study. Columns like the region, industry, ownership type, product name, and product type are all important feature variables that help this study to achieve new sales goals.

VI. MODELING

The modelling phase is the next important phase where several modelling techniques are applied to create and access the models built.

A. *Total Unduplicated Reach and Frequency (TURF) Analysis*

TURF analysis is a statistical research methodology that ranks combinations of products by how many people will like these combinations [11].

TABLE I. TURF ANALYSIS RESULTS

Features	Size of group	"Reach"	"% of Cases"	"Frequency"	"% of Responses"
ADDED: TraverseGlobal	1	1160	13.2	1160	16.1
ADDED: Made2Manage	2	1941	22.1	1943	27
KEPT: TraverseGlobal					
ADDED: Paragon	3	2462	28	2464	34.3
KEPT: Made2Manage, TraverseGlobal					
ADDED: APIPro	4	2877	32.7	2879	40
KEPT: Made2Manage, Paragon, TraverseGlobal					
ADDED: JustFood	5	3216	36.5	3221	44.8
KEPT: APIPro, Made2Manage, Paragon, TraverseGlobal					
ADDED: WorkWiseERP	6	3549	40.3	3560	49.5
KEPT: APIPro, JustFood, Made2Manage, Paragon, TraverseGlobal					

Table I shows the best frequency and reach values among each group size. Unduplicated reach describes the proportion of customers who selected at least one of the products within a portfolio. Frequency is the number of items desired within the portfolio. Here, in group size 1 "Traverse Global" product has a reach and frequency of 1160, this is also the total number of customers who purchased this product. The percentage of cases or respondents saying "yes" to the product, is 13.2% of total customers who have a reach to this product and the Percent of response is the single response out of total responses from the given dataset which is 16.1% of total customers who respond to the product. Similar is the understanding for the other group sizes, where the percentage of cases and responses keeps increasing with increasing group size, where the number of products increases.

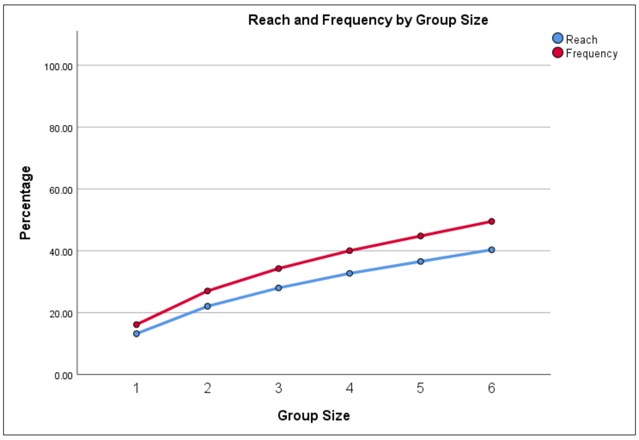


Fig. 3. *GGraph* for TURF analysis

Fig. 3 depicts the *GGraph* for reach and frequency by group size. The maximum group size was set to six and so the graph shows the reach and frequency percentages that increase per increase in the group size.

B. *MBA using Apriori Algorithm.*

MBA discovers product associations by looking for product combinations that often co-occur in transaction data. Apriori is the most basic algorithm for mining frequent patterns from transaction information and pattern mining has indeed been widely used in MBA to reveal hidden patterns in transactional data [3].

Confidence is 98.7% and Lift is 7.17 which is above 1, indicating that the output response is more likely to occur than the average response. Therefore, the association rules improvise the likelihood of the results. Fig. 4 shows the web chart with strong product associations.

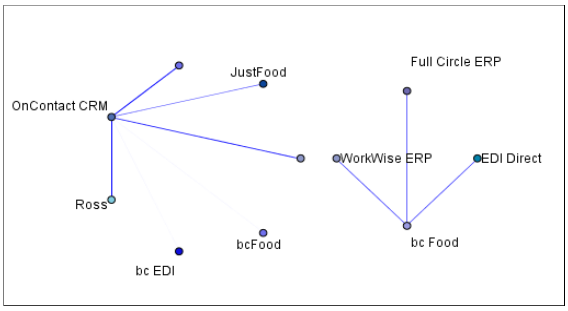


Fig. 4. Web Chart showing product associations

C. *Decision Tree - Chi-squared Automatic Interaction Detector (CHAID)*

Decision trees are one of the effective methods for data mining. The tree structure provides an easy way to interpret the results. CHAID method has been used as a growing method and is frequently used in direct marketing to select groups of consumers to predict how their responses to some variables affect other variables. CHAID, like other decision trees, has the advantage of producing highly visual and easy-to-interpret output. CHAID uses multiway splits by default, and therefore it requires rather large sample sizes to work effectively, as small sample sizes can quickly lead to respondent groups that are too small for reliable analysis.

According to the correlation between products based on the association rules, there are different target groups of customers. To get the characteristics of the target customer, variables such as customer region, product selling region, industry, ownership and product type are used as the independent variables. The key drivers identified from the importance chart are – product region, product type and ownership.

Fig. 5 shows the CHAID decision tree output and the key factors responsible are depicted in Fig. 6, which is the predictor importance chart from which the Association Rule is made, stating that if the product region is "Americas", product type is "SCM & Others" and ownership is either "partnership or private", then there is a high probability that the customer will purchase both products "EDI Direct" and "Full Circle ERP" together.

TABLE II. OUTPUT OF MBA

Consequent	Antecedent	Instances	Support %	Confidence %	Rule Support %	Lift	Deployability
EDI Direct	Full Circle ERP	81	13.61	98.77	13.45	7.17	0.17
Foodware BC	Foodware 365	35	5.88	94.29	5.55	8.01	0.34
Oxaion ERP	Syncos MES	28	4.71	92.86	4.37	17.82	0.34
OnContact CRM	WorkWise ERP	81	13.61	92.59	12.61	5.35	1.01
bc Food	bc EDI	45	7.56	62.22	4.71	6.38	2.86

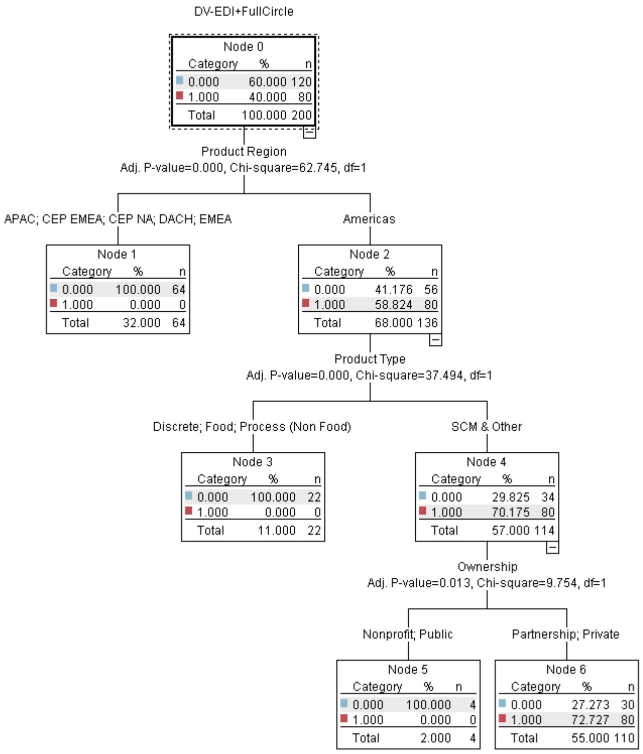


Fig. 5. CHAID- Decision Tree Output

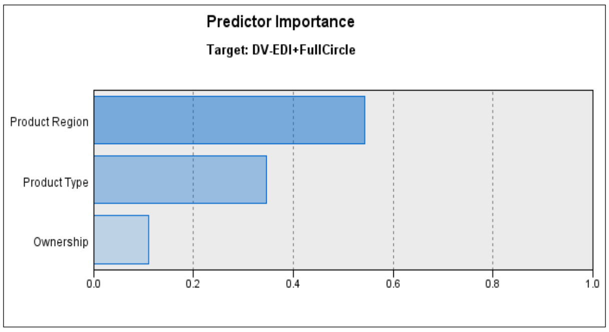


Fig. 6. Predictor importance chart

D. *Supervised Machine Learning: K- Nearest Neighbour (KNN)*

KNN algorithm is a method for classifying cases based on their similarity to other cases. Machine learning was developed to recognize patterns of data without requiring an exact match to any stored patterns, or cases. Similar cases are near each other, and dissimilar cases are distant from each other. Thus, the distance between two cases is a measure of their dissimilarity.

Fig. 7 shows the distribution of the different products based on the 5 predictors, there is a relation between the product type “Process non-food” in the region of Sweden and Poland. Since this gives the distribution of the association between products in a 5- dimensional space, accurate groupings of data categories cannot be made.

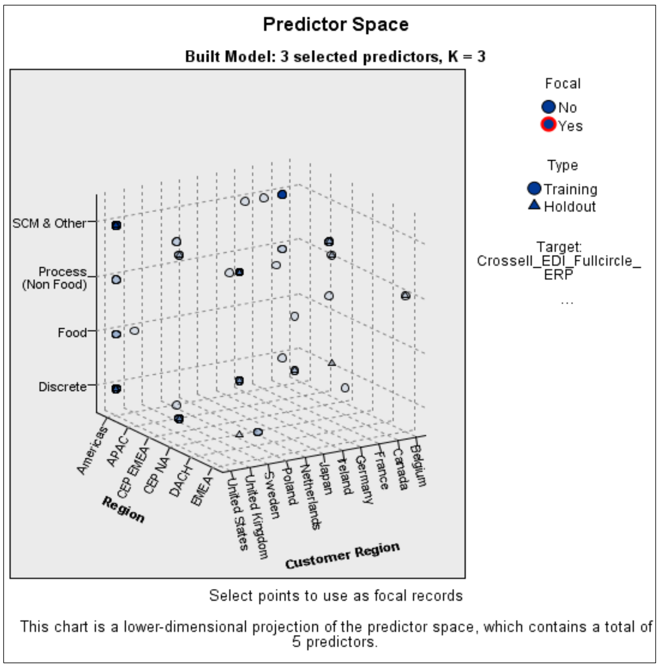


Fig. 7. Output of KNN Algorithm.

E. *Deep Learning - Multilayer Perceptron*

Finally, the last model is built using the Deep Learning method, in which the multilayer perceptron is used.

The Multilayer Perceptron (MLP) procedure produces a predictive model for one or more dependent (target) variables based on the values of the predictor variables. It is a fully connected class of feedforward artificial neural networks. The term MLP is used ambiguously, sometimes loosely to mean any feedforward, sometimes strictly to refer to networks composed of multiple layers of the perceptron.

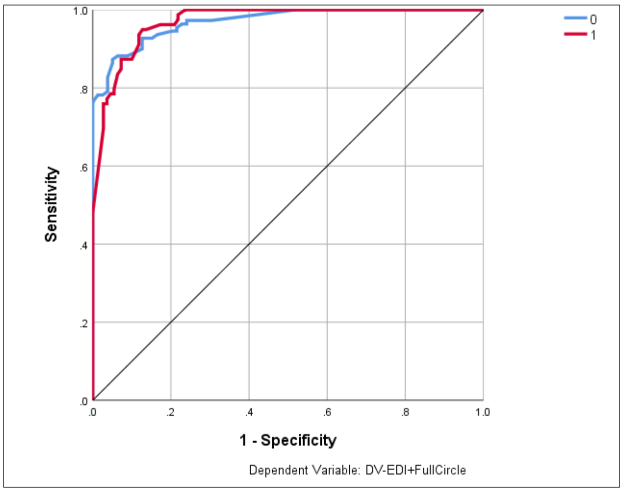


Fig. 8. Sensitivity-Specificity chart

Fig. 8 depicts the Sensitivity-Specificity chart and tells us the quality of the model. Since the AUC for this combination of products EDI and full circle ERP is 0.97, the model has high accuracy and is better than the baseline model.

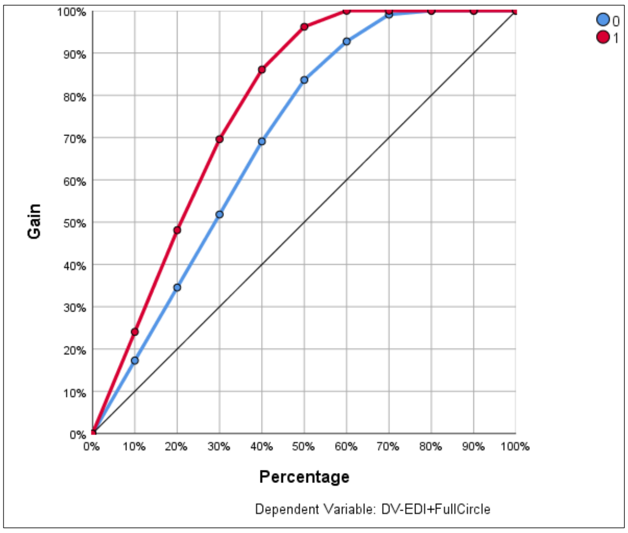


Fig. 9. Gain chart

The graph in Fig. 9, shows the gain percentage for the different decile distributions. It is observed that the gain percentage is the highest at the 60th percentage.

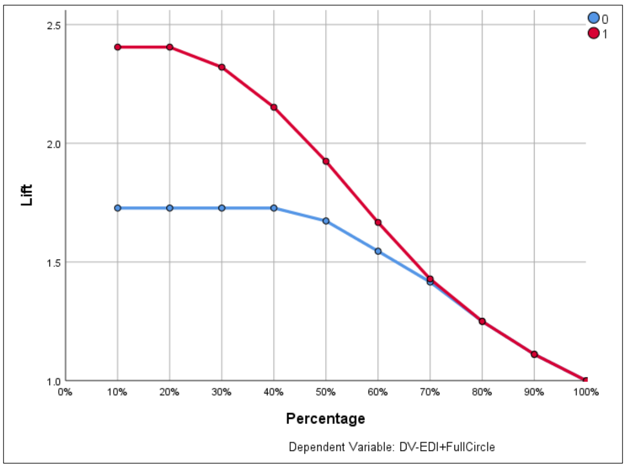


Fig. 10. Lift chart

This lift chart as shown in Fig. 10 depicts that lift is the highest at 10% and starts decreasing after the 60th percentile.

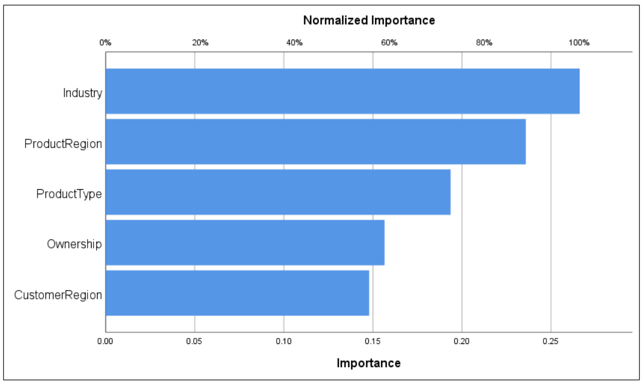


Fig. 11. Normalized Importance chart

Fig. 11 gives the importance chart which depicts the importance of the independent variables in descending order. Where the normalized importance score of Industry is 100%, Product region is 88.6%, Product type is 72.8%, Customer region is 55.6% and Ownership is 58.8%.

VII. MODEL EVALUATION

In this section, an overview of the results obtained from different machine learning procedures is provided and explained.

A. Classification Report

A classification report is a performance evaluation metric in machine learning. It is used to show the precision, recall, F1 Score, and support of the trained classification model [12].

TABLE III. CLASSIFICATION REPORT FOR CHAID DECISION TREE

	Forecasted		
Actual	***0***	***1***	***Correct %***
0	116	4	96.70%
1	13	67	83.80%
Overall %	64.50%	35.50%	91.50%

From Table III, it is observed that the overall accuracy of the model is close to 92% indicating that the model is good.

TABLE IV. CLASSIFICATION REPORT FOR KNN

	Forecasted		
Actual	***0***	***1***	***Correct %***
0	79	5	94.00%
1	17	42	71.20%
Overall %	67.10%	32.90%	84.60%

From Table IV, it is observed the overall accuracy of the model is 85% indicating that the model is good, but not as good as the CHAID Decision tree algorithm.

The model uses the deep learning methodology of MLP. In this model, the dataset is divided into train and test where 70% of the data is utilized to train the model and the remaining 30% is utilized for testing.

TABLE V. CLASSIFICATION REPORT FOR MULTILAYER PERCEPTRON

		Forecasted actual		
		0	***1***	
Training actual	***0***	71	9	88.80%
	1	4	58	93.50%
	Overall Percent	52.80%	47.20%	90.80%
		0	***1***	
Testing actual	***0***	25	5	83.30%
	1	0	17	100.00%
	Overall Percent	53.20%	46.80%	89.40%

From Table V, it is observed the overall accuracy of the model is 91% for the training data and 89% for the test data. indicating that the model is good, but not as good as the CHAID Decision tree algorithm. However, it is better than the KNN methodology used.

From the Classification report Tables III, IV, and V the confusion matrix values such as the True Positive (TP) which is the correctly predicted event value, and False Positive (FP) which is the incorrectly predicted event value, True Negative (TN) which is the correctly predicted no-event values, and False Negative (FN) which is the incorrectly predicted no event values.

B. Model Performance metrics

Almost all model-performance metrics are calculated on the model's predictions being compared to the value of the dependent variable in a dataset.

TABLE VI. MODEL PERFORMANCE METRICS

Model Performance metrics				
Models	***Precision***	***Recall***	***F1-score***	***Accuracy***
Decision Tree	94%	84%	89%	92%
KNN	89%	71%	79%	85%
MLP- Test	77%	100%	87%	89%
MLP-Train	87%	94%	90%	91%

From Table VI, it can be concluded that the Decision Tree algorithm using the CHAID mechanism is the best model that could be built for the analysis of the right products that can be sold to customers. It has the highest precision and recall values of 94% and 92% respectively proving to be the best model.

C. Accuracy Score

One metric for evaluating classification models is accuracy. Informally, “accuracy is the percentage of correct predictions made by our model”. Formally, “accuracy is defined as the number of correct predictions/ Total number of predictions” Table VII, shows that the decision tree is the best model out of all the models that were built.

TABLE VII. ACCURACY SCORE

Models	**Accuracy**
Decision Tree	92%
KNN	85%
MLP- Test	89%
MLP-Train	91%

VIII. RESULTS

In the study presented in this paper, the dependent variable is the variable indicating the purchase pattern of the customers who have purchased two or more products together and the independent variables are continuous variables that are used to segment and profile the customers and the products to enable up-sell and cross-sell opportunities.

From the model evaluation, the most efficient model is the Decision Tree using the CHAID method, because of its high accuracy of 92% and precision value of 94%.

Key drivers influencing the product sales as shown in the decision tree are:

a) Product Region – Selling region of the product.
b) Product Type – To which category of product it belongs.
c) Ownership – The type of ownership of the customer.

Other factors such as customer region and Industry are having huge value categories. Therefore, they have not played an important role in the decision tree algorithm for predicting the products that can be sold together.

From this paper, the scope of more products that could be sold together to a customer can be understood. That is; if the Customer purchases product “A”, then the customer is more likely to purchase product “B”. Thus, the decision tree modelling technique helps to conclude the association rules to optimize the results.

IX. CONCLUSION AND FUTURE WORKS

The main objective of the study is to understand the purchasing pattern of products from the product sales transaction data, study and profile customers based on their purchase behavior and recommend and suggest products to customers, thereby increasing cross-sell and up-sell opportunities.

One reason for MBA’s growing acceptance in the data research fields is that researchers can evaluate the existence of association rules by using an inductive approach to theorizing. Taking everything into account, a recommendation system can significantly make an effect on marketing and sales studies that can be used to derive strategic business decisions.

In this study, different modelling techniques are carried out and evaluated to find out the key drivers responsible for cross-selling certain products. It is possible to profile customers that belong to different categories based on these key drivers and propose that for new customers who belong to any of these categories, such products could be sold, thereby increasing sales opportunities in the organization and enabling the organization to reach its new goal of achieving sales targets and increasing customer base and maintain niche enterprise products.

The study does not cover the cost and financial analysis, if the financial data could be used for analysis, recommendation of the best possible products for upselling or cross-selling will be possible, thereby increasing sales.

A similar analysis can be used to model other combinations of data in which more than two products are sold together.

An understanding and study of the customers who add products to the cart but have incomplete transactions would be a good study, to retain or increase the customer base.

REFERENCES

[1] David Gargaro, “12-ways-to-increase-sales,” 2022. https://www.business.com/articles/12-ways-to-increase-sales/ (accessed Aug. 10, 2022).

[2] Wikipedia, “Upselling,” 2022. https://en.wikipedia.org/wiki/Upselling (accessed Aug. 10, 2022).

[3] T. Gupta, R. Karthiyayini, R. Balasubramanian Professor, “Affinity Analysis and Association Rule Mining using Apriori Algorithm in Market Basket Analysis,” 2016. [Online]. Available: www.ijarcsse.com

[4] T. Hewen, Y. Zengfang, Z. Pingzhen, and Y. Honglin, “Using data mining to accelerate cross-selling,” in *2008 International Seminar on Business and Information Management, ISBIM 2008*, 2008, vol. 1, pp. 283–286. doi: 10.1109/ISBIM.2008.186.

[5] Raorane AA, Kulkarni RV, and Jitkar BD, “Association Rule-Extracting Knowledge Using Market Basket Analysis,” 2012. [Online]. Available: www.isca.in

[6] Z. Y. Chen, Z. P. Fan, and M. Sun, “A SVM ensemble learning method using tensor data: An application to cross-selling recommendation,” Jul. 2015. doi: 10.1109/ICSSSM.2015.7170282.

[7] L. Wang and J. Sun, “Market Basket Analysis based on Apriori and CART,” 2019, doi: 10.25236/etmhs.2019.311.

[8] T. Gupta, R. Karthiyayini, R. Balasubramanian Professor, “Affinity Analysis and Association Rule Mining using Apriori Algorithm in Market Basket Analysis,” 2016. [Online]. Available: www.ijarcsse.com

[9] Shruthi Gurudath, “Market Basket Analysis & Recommendation System Using Association Rules,” Jun. 2020, doi: 10.13140/RG.2.2.16572.05767.

[10] N. Dookeram, Z. Hosein, and P. Hosein, “A Recommender System for the Upselling of Telecommunications Products,” in *International Conference on Advanced Communication Technology, ICACT*, 2022, vol. 2022-February, pp. 66–72. doi: 10.23919/ICACT53585.2022.9728818.

[11] “Turf Analysis,” 2019. https://conjointly.com/blog/turf-analysis/ (accessed Aug. 25, 2022).

[12] Aman Kharwal, “Classification Report,” 2021. https://thecleverprogrammer.com/2021/07/07/classification-report-in-machine-learning/#:~:text=A%20classification%20report%20is%20a,this%20article%20is%20for%20you. (Accessed Aug. 13, 2022).

GitHub Link

https://github.com/sharonjoseph/Capstone-2--Product-Affinity-Analysis-to-Increase-Sales-using-Machine-Learning