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.

978-1-6654-5664-7/22/$31.00 ©2022 IEEE

"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].

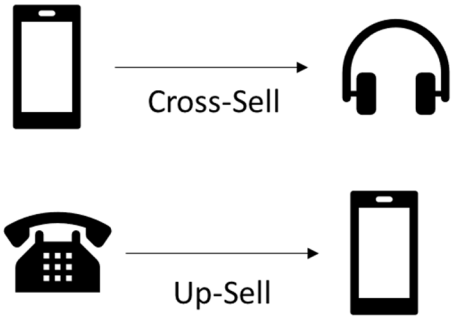

Fig. 1. Cross-sell vs Up-sell

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 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.

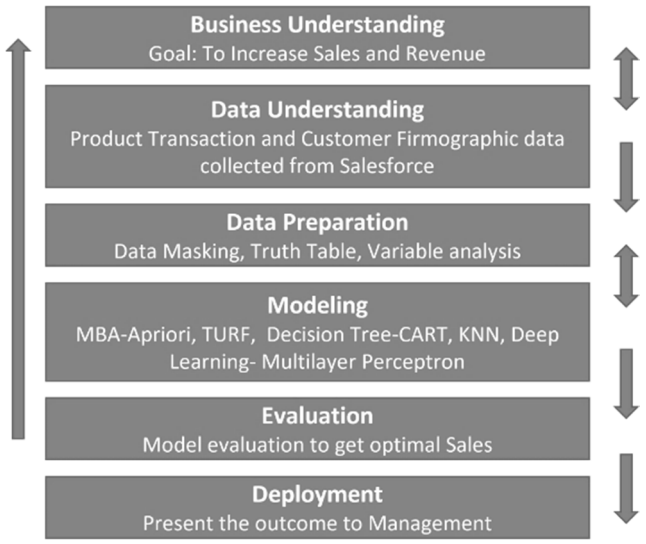

Fig. 2. Workflow of the Model

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.

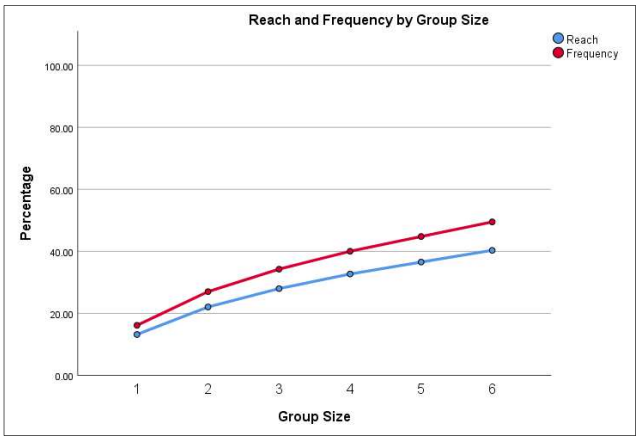

Fig. 3. *GGraph* for TURF analysis

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.

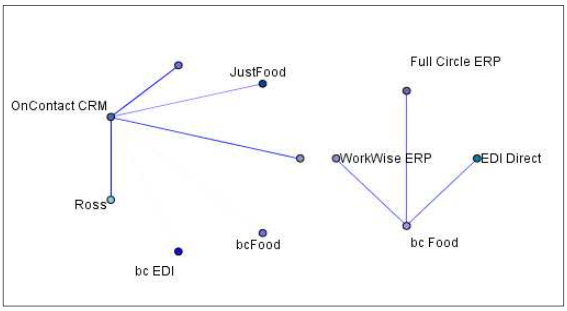

Fig. 4. Web Chart showing product associations

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

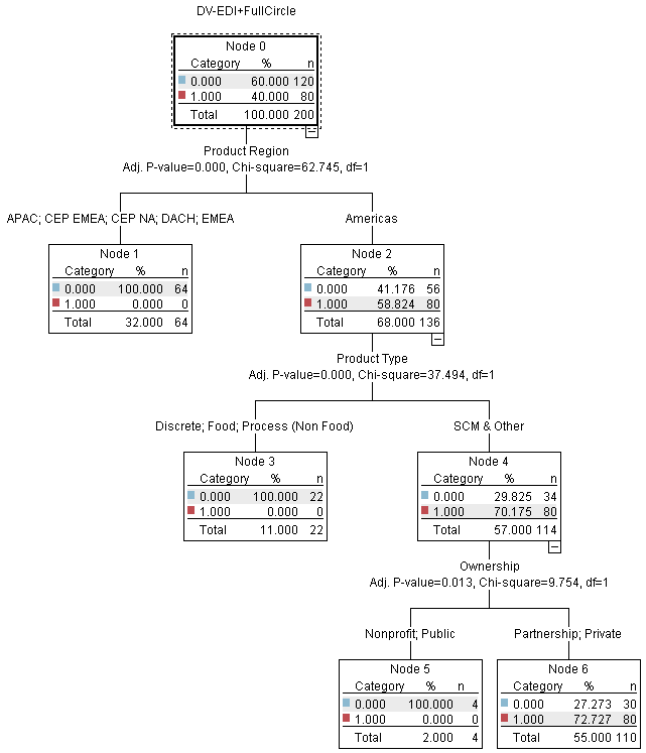

Fig. 5. CHAID- Decision Tree Output

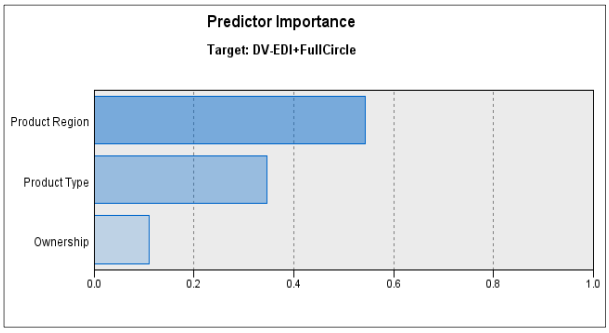

Fig. 6. Predictor importance chart

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

The 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 the 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.

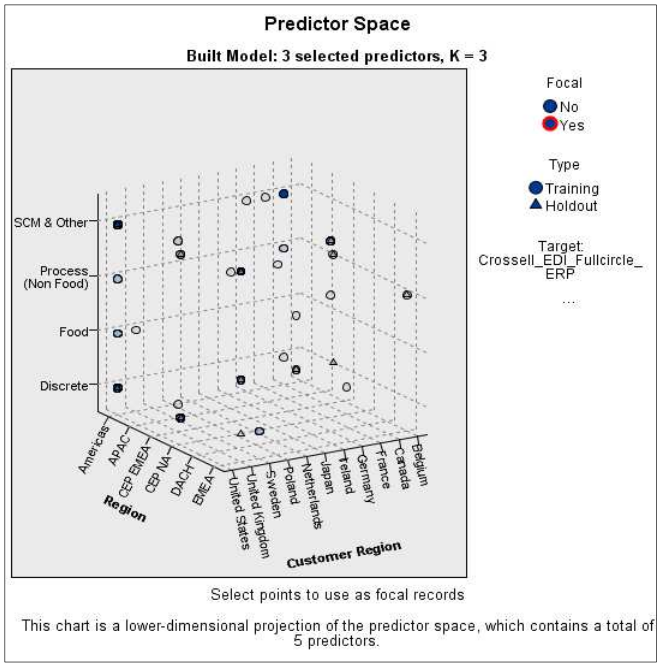

Fig. 7. Output of KNN Algorithm.

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.

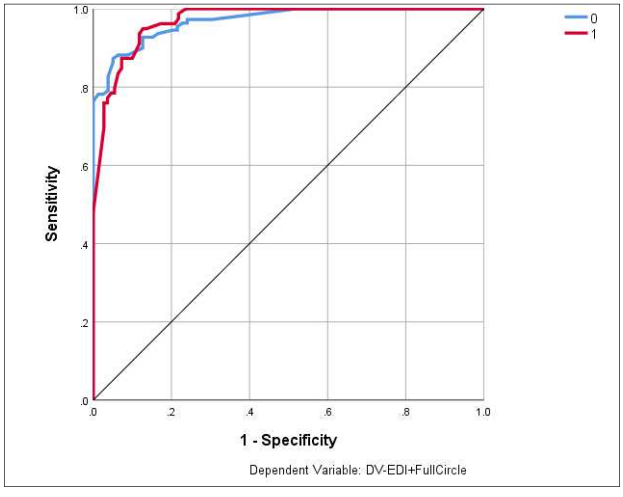

Fig. 8. Sensitivity-Specificity chart

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.

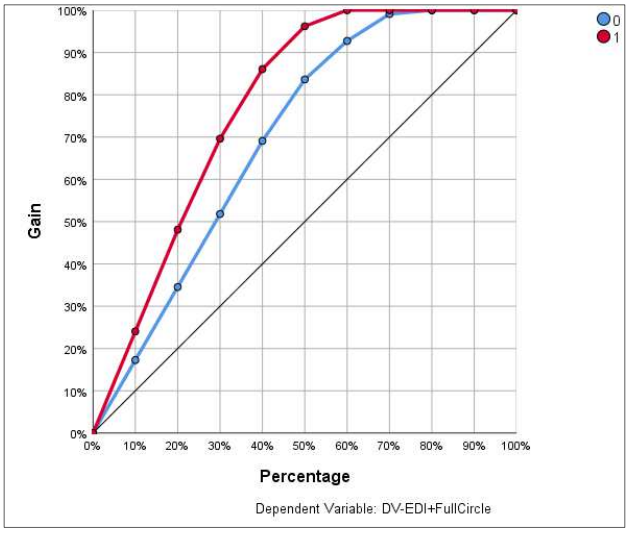

Fig. 9. Gain chart

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.

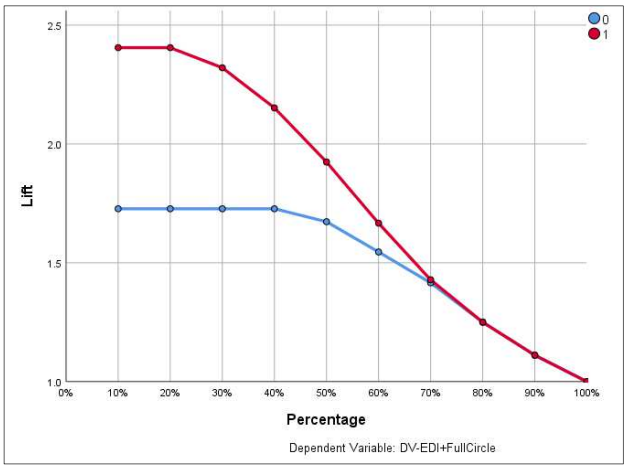

Fig. 10. Lift chart

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

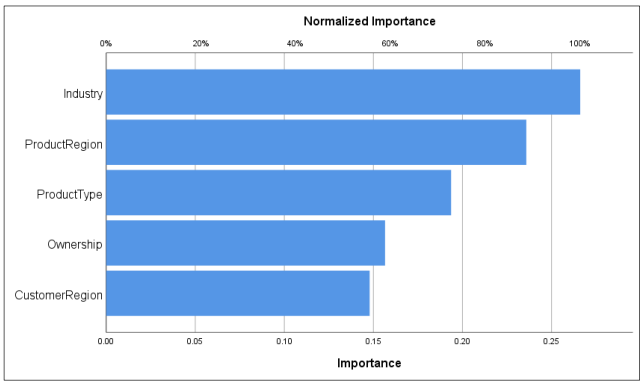

Fig. 11. Normalized Importance chart

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, 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 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).