National College of Ireland

Enhancing Retail Strategies: An Integrated Framework for Market Basket Analysis using Apriori and MLP in Consumer Behavior Modeling

MSc Research Project
Data Analytics

Rishika Poojari
Student ID:X2021421

School of Computing
National College of Ireland

Supervisor: Rejwanul Haque & John Kelly

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Rishika Poojari
Student ID:	X20214201
Programme:	Msc Data Analytics
Year:	2023
Module:	Msc Research Project
Supervisor:	Prof. Rejwanul Haque and Prof. John Kelly
Submission Due Date:	14/08/2023
Project Title:	Enhancing Retail Strategies: An Integrated Framework for Market Basket Analysis using Apriori and MLP in Consumer Behavior Modeling
Word Count:	5700
Page Count:	18

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Rishika Poojari
Date:	16th September 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Enhancing Retail Strategies: An Integrated Framework for Market Basket Analysis using Apriori and MLP in Consumer Behavior Modeling

Rishika Poojari
x20214201

Abstract

Market basket analysis has emerged as one of the benchmark techniques that offers insights into market structure and product relationships. These insights empower informed decisions about product promotion and positioning, ultimately optimizing revenue. Classic association rule mining identifies frequent itemsets and generates rules indicating item relationships. To address the problem of efficiency, prior research has explored combining association rules and deep learning for market basket analysis. However, deep learning's limitations like scalability and pattern recognition gaps persist. There is a research gap to fully investigate this combination.

This study proposes a hybrid framework that leverages machine learning algorithms for improved market basket analysis performance. Using a publicly available dataset, it examines extending association rules and integrating deep learning techniques. The results support the feasibility of this integrated approach, although not meeting initial objectives. This spurs the rejection of the prime hypothesis. Future research could refine association rules through advanced techniques such as customer segmentation or dimensionality reduction. The method used in this study is effective in extracting features from the Apriori Algorithm to be fed into a neural network, producing a satisfactory outcome. This approach also explains the best ways to include certain techniques in the topic to make the results even better.

1 Introduction

In today's highly competitive marketing environment, businesses must employ every advantage available to them in order to stay ahead of their competition. Market basket analysis has emerged as one of the fundamental product marketing techniques that provides valuable insights into the structure of the market and the relationships or associations between products. This information can be used to make smart choices about how to promote and position products, which in the end helps to maximize revenue and profit ultimately. One practical application of market basket analysis is the grouping of products based on characteristics such as orderprice, size, material or brand, which facilitates customers in finding their desired products more easily. Additionally, market analysis can be used to recommend products based on customers' prior purchases, thereby introducing them to new items of potential interest. Research has shown that tailored

recommendations from market basket analysis-based recommendation systems can significantly increase profits for stores by enhancing cross-selling and up-selling and reducing returns. Consequently, incorporating market basket analysis into recommendation systems represents a cost-effective approach to strategic promotional activities compared to traditional methods. This analysis also holds promise for identifying opportunities for new product development, as it can uncover associations between certain items.

Association Rule Learning and its Application in Market Basket Analysis: In the field of machine learning and deep learning, numerous studies have focused on developing algorithms and methodologies for market basket analysis, which involves analyzing transaction data to identify patterns or relationships between items. This process determines the associations between items, allowing businesses to understand how one item is associated with others. Association rule learning, a data mining technique, is often used to uncover correlations between elements or products in transactional data. Unsupervised learning techniques such as association rule learning enable the identification of meaningful patterns or rules within large datasets. The widely used A-priori algorithm is recognized as a classic approach for learning association rules. This algorithm first identifies frequent itemsets, which refer to groups of items that frequently appear together in the data, and subsequently generates association rules that demonstrate the relationships between these itemsets. Support and confidence are two metrics applied to assess the strength of an association rule. Support indicates the percentage of transactions that contain each item in an association rule, while confidence represents the likelihood of a transaction containing the antecedent of a rule also containing the consequent. An association rule is considered interesting if both support and confidence values meet the minimum thresholds. However, mining large datasets using association rule algorithms can be computationally expensive, particularly for online retailers with extensive product assortments like Walmart.com or Amazon.com, due to the exponential increase in possible item combinations and association rules as the dataset size grows.

Combining Association Rules and Deep Learning: To address the need for efficient analysis, previous research has explored the use of machine learning algorithms, including deep learning techniques, to analyze massive data from the online market basket and provide product recommendations. Recommending similar products not only increases sales but also optimizes inventory management, leading to improved customer satisfaction and reduced costs. Deep learning methodologies like Neural Networks (including Convolutional Neural Networks and Recurrent Neural Networks) have been employed; however, they have demonstrated limitations in terms of computational power, scalability, and the ability to identify specific patterns and relationships within a dataset. Currently, there is a gap in research investigating the combination of association rules and deep learning in market basket analysis. This study aims to propose a hybrid framework that leverages machine learning algorithms to enhance the performance of market basket analysis, surpassing the capabilities of standalone methods. The proposed research will be conducted using a publicly available dataset from an online grocery store. It will examine why the extension of association rules and their integration with deep learning techniques are necessary. Furthermore, it will explore the selection of the Multi-Layer Perception (MLP) neural network from various neural network architectures for the dataset. The study will assess the efficiency and effectiveness of data mining and deep learning approaches for product recommendations, specifically considering market basket analysis. The proposed

hybrid framework will be an A-priori MLP-based algorithm for market basket analysis.

Hence, the primary motive of this document revolves around the research question:

How efficiently can market basket analysis be performed using an integrated framework of association rules and neural networks for consumer behavior prediction?

This dissertation has been documented as follows: Section 2 introduces market basket analysis and the developments in the state of art through a comprehensive literature review, followed by a description of the research procedure followed in Section 3. Section 4 summarizes the framework of the novel algorithm that underlies the implementation of the project and Section 5 discusses the solution development or the implementation to the system design architecture in Section 4. The results are evaluated (Section 5) and discussed in Section 6.

Figure 1: Pictoral representation of groceries in a supermarket

2 Related Work

Market Basket Analysis (MBA) is a prominent technique widely employed in the retail industry to quantify the statistical correlation among various items or products. By utilizing this technique, retailers can make informed decisions regarding marketing strategies, shelf arrangements, cross-selling approaches, and more. MBA scrutinizes purchasing patterns by extracting information from transactions or shopping baskets, where a transaction represents a set of items purchased in a single transaction. Multiple studies, including those by Gouda et al. (2003) , Sorensen et al. (2017), and Valle et al. (2018), have highlighted the effectiveness of MBA in enabling retailers to optimize their business operations intelligently. Kutuzova et al.(2018) employed various data sources to improve the grocery store recommendation system. Their techniques encompassed collaborative filtering, clustering, and association rule mining, with the latter yielding satisfactory outcomes.

The use of association rules for data mining and market basket analysis is not limited to retail data. Standalone association rule mining like A-priori have been used by researchers to discover previously unknown associations between Chinese tradional herbs and a toxic herb called Fuzi (Tai et al. (2021)). Similarly, for a metal trading industry, Yudhishtyra et al. (2020) found that CARMA which is a hybrid algorithm that combines decision tree classification and association rule mining, outperformed Apriori in terms of precision and efficiency for recommendation in big data implementations. This forges the motivation further to implement a hybrid algorithm for market basket analysis.

2.1 Association Rules Mining

Current approaches for market basket analysis can be categorized into four factors: general methods, sequential methods, pattern-based methods and hybrid methods. Among these, general methods are the most fundamental. They involve identifying associations between items in a dataset without considering any specific order or sequence. However, pattern based techniques for finding frequent item-sets are much more efficient and faster. A convenience store named Toko Warga faced significant decrease in sales that required them to think of another way to increase their sales during during pandemic covid-19. Guidotti et al. (2019) proposed A-priori algorithm and an extension of it's deficiencies called the FP-Growth algorithm as the data mining solution for this problem. The difference between previous researches and the research conducted by the author is that their research compares two algorithm which is Fp-Growth and Apriori algorithm, and uses Toko Warga Dataset. It is seen that to address the deficiencies in the Apriori algorithm, a new algorithm was developed. However, this new algorithm aka FP-Growth, even though was efficient enough, the accuracy of the former was better given the small batch of dataset. Therefore, exploring the impact of using a larger dataset on the empirical results of the paper could provide better and vast insights. Addressing such challenges might be more effectively managed by the implementation of a hybrid algorithm. This type of algorithm has the potential to leverage the strengths of both methodologies while overshadowing their individual limitations.

The Apriori algorithm has two main downsides. Firstly, it creates a lot of potential groups of items to check. Secondly, it needs to scan through the database many times. So, with this consideration the necessity of employing the apriori algorithm on a large database creates a complexity on the computational power which was a limitation in the research of Guidotti et al. (2019). Hossain et al. (2019) suggests in their work that using top selling products, it is possible to get the same frequent itemsets and association rules within a short time compared with that outputs which are derived by computing all items together. This was tested on two datasets available on Kaggle, the first being the dataset that provides transaction information over the course of a week at a French Retail Store and the second consists of observations from a bakery which provides transaction of bakery items. Hence, this technique could be considered to downsize a large dataset similar to the one used in this research which is the InstaCart dataset.

In a study by Kavitha and Subbaiah (2020), they used the Apriori to look at data from a grocery store with a similar motive to find which groups of items customers often buy together. To do this, the authors developed an application that uses Apriori using a package in the tool R. Pictorial depictions of the patterns were also created using R in

the application. However, the research has some limits. They only used a set of data that was already pre-processed, and didn't use real-time information. Therefore, the research lacks feature extraction and feature selection techniques that could make the results more significant.

2.2 Neural Networks for Market Basket Analysis

The limitation identified in previous researches as mentioned talks about Apriori algorithms scanning the data multiple times. For this, Mathur et al. (2014) identified the use of the artificial neural network technique to overcome the problem of the Apriori algorithm in the market basket analysis by using the single-layer feedforward partially connected neural networks technique. However, it should be noted that this FFNN focused only on seasonal data.

Artificial neural networks, such as Multilayer Perceptron (MLP), can be used in different retail sectors. According to Polat, M. et. al (2019), combining MLP and LSTM Recurrent Neural Networks (RNNs) improves the accuracy and execution time of predicting what online buyers will buy in real-time. The data used for this prediction job will include attributes such as demographic information, browsing habits, and purchase history.

Sequential or time-series data can be easily processed by RNNs. However, they might struggle with long sequences of data, which can be a problem when trying to make fast predictions in real-time. This research shows that combining LSTM and MLP networks can help solve this issue. MLPs are feedforward neural networks that can process a lot of input data at once, making them great for large datasets. The MLP is used to make predictions by giving it new input data and gathering the results from the output layer after training. Since MLP can recognize patterns and dependencies in a shopper's behavior over time, it's a useful network for predicting online shopping behavior, especially with the constantly growing amount of data available. Convulational Neural Networks (CNNs) are also used to conduct market basket analysis, which generally uses images as the data input to extract features for the prediction of the next product. As the number of users and assets in the e-commerce system grows, the time required for searching for a specific user among the entire user population also increases substantially. Consequently, this negatively impacts the overall performance of the system if the data captured is image based (Ghadekar et al. 2019). Hence, considering this research which focuses on transactional data, MLPs are considered to be a good choice for combining with data mining and deep learning.

Furthermore, the integrated framework for market basket analysis is effectively exhibited in a Nigerian retail study by Eboka et al. (2018), showcasing its potency in predicting inventory needs. This is done by leveraging association analysis, artificial neural networks, and a memetic algorithm where this methodology successfully forecasts item demand by transforming sales data and enhancing artificial neural network predictions with these rules to capture and predict item relationships more effectively.

To conclude, this section has aimed to offer a comprehensive overview of the literature covering from conventional association mining algorithms, their limitations, and their progression towards enhanced market basket analysis techniques, including deep learning methodologies and the diverse applications of market basket analysis across industries. It becomes evident why a hybrid integration of algorithms is important and how the deficiencies of individual algorithms can be combined through the complementary strengths

of alternative algorithms.

3 Methodology

This study utilizes the widely adopted data mining approach known as the Cross Industry Standard Process for Data Mining (CRISP-DM). To ensure a systematic and methodical research process, the following steps are implemented.

3.1 Business Understanding

Market basket analysis, enabled by association rule mining and deep learning methodologies, holds significant importance in the success of retail businesses. It can be a valuable technique used to uncover hidden patterns within customer buying behavior. By studying these frequent patterns, businesses can provide personalized future recommendations to customers, enhancing their shopping experience. The proposed Apriori-MLP based framework can be employed by retail companies to analyze real-time data from social media, browsing history, and historical supermarket records.

The application of market basket analysis within recommendation systems offers a cost-effective strategy to tailor promotional activities compared to conventional methods. By utilizing an Apriori-MLP based framework, retailers can gain insights into customers' purchasing behavior and discover meaningful associations between products frequently bought together. For instance, if the analysis reveals that customers who purchase hiking boots are also likely to buy outdoor backpacks, the retailer can strategically position these items together in-store or showcase them as a bundled deal on their website, increasing the chances of cross-selling and higher average order value. Similarly, on analysis if it is found that customers who buy fitness trackers are also likely to purchase resistance bands, the retailer can explore the development of a fitness kit that includes both items, catering to health-conscious customers seeking comprehensive workout solutions.

Retailers can leverage customer satisfaction, increase sales, and improve inventory management using these data-driven insights. They can also efficiently cross-sell related items and introduce attractive product associations.

3.2 Data Understanding

This research utilizes the data available in Kaggle [1] which is for non-commercial use only and contains a set of related files that describe a consumer's orders for a said timeframe.

The aisles.csv contains different aisles with a total of 134 unique aisles. Departments.csv contains different departments and there are total 21 unique departments. order_products_prior.csv file gives information about which products were ordered and in which order they were added to the cart. It also tells us the product was reordered or not. Products.csv contains the list of a total of 49688 products and their aisle and department. The number of products in different aisles and different departments is different.

This dataset contains a total of 99,574 baskets that contain 1,048,575 products in total. This means that an average of 10.5 products are purchased within a basket. The dataset contains a range of 4 to 100 orders per customer, and sequence of the products purchased in each order. Additionally, data regarding the week and time of day when

[1] https://www.kaggle.com/competitions/instacart-market-basket-analysis/data

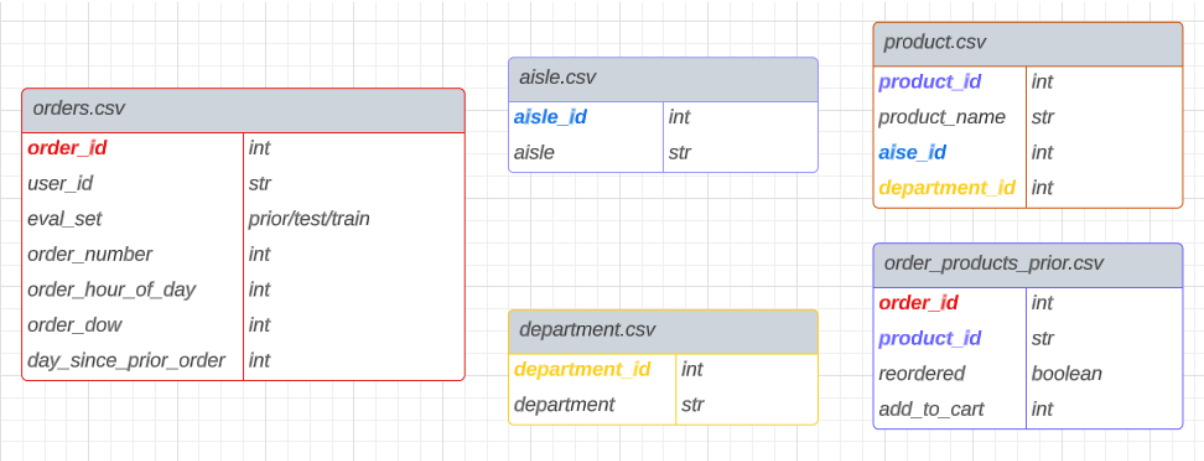

Figure 2: Correlation between the csv files.

each order was placed is included, along with a measurement of the time elapsed between consecutive orders.

3.3 Preparation of Data

Preliminary analysis of the data set is carried out using statistical descriptives and pandas in Python that helps to find duplicate entries, missing values, duplicate records, skewness, and other relevant factors, such as cardinality.

- The 5 files orders.csv, order_products_prior.csv, products.csv, aisles.csv and departments.csv are combined into one dataframe for data analysis based on order_id, product_id, aisle_id and department_id.
- The column that shows the 'days since prior orders' showed exceptionally high number of null values and hence were made to zero.
- The dataset contains 36,864 unique products that are purchased at least once
- The majority of products (90%) are bought less than 50 times and 8,392 different products are bought only once in the dataset.
- The product that is bought most often is a banana; this product is bought 14,136 times.

3.4 Exploratory Data Analysis

This phase of the project research helps to explore the data and understand the kind of information that will be worked on, for example, user preferences, item details, and maybe even when items were ordered and reordered and the time frame between consecutive orders. Hence it is important to spot any weird or missing info that can mess up the predictions.

When plotting the products with the highest frequency counts, it was observed that organic products such as organic bananas, organic strawberries, organic spinach etc. make up the top 10 products bought. In a similar fashion, the reorder ratio has the same organic products making up the top 10 highest reordered products. When checking the sales of the products department-wise, the produce contributed to the highest sales. Sunday and Monday appeared as the days of the week that witnessed great number of unique orders. Consequently, Sunday and Monday also appeared to be the busiest Days of The Week for ordering products.

Further analysis also depicted that the number of orders placed usually peaked from 10 AM in the morning to 4 PM in the evening. As depicted in Fig.3 which depicts the time distribution between consecutive orders, it is evident that in a timeframe of 2 seconds, about 10-30 orders are added to the cart.

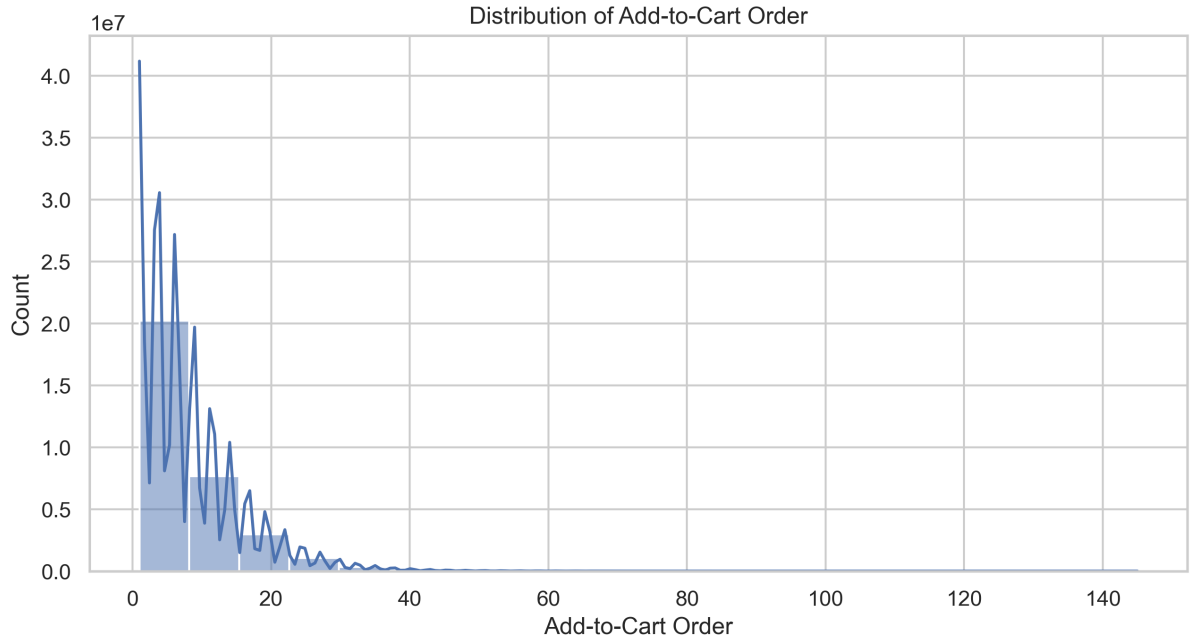

Figure 3: Time intervals between consecutive orders

3.5 Modeling

3.5.1 Association rules

Association rules are a popular method for discovering interesting relations between variables in large databases. They are used to identify and analyze patterns of items in large data sets. The apriori algorithm is a classical algorithm used for mining frequent itemsets and generating association rules. It is designed to operate on databases containing transactions, such as purchases by customers of a store.

Evaluation Metrics

- **Support**: Support is the probability or percentage of occurrences of a specific item or item set in a data set that indicates its popularity or frequency of appearance.

It helps in identifying the most common items or itemsets in a dataset (Blattberg et al. 2008).

- **Confidence**: Confidence is the measure of how often a particular association rule is found to be true in the dataset. It represents the likelihood that an item or itemset B is purchased, given that item or itemset A has already been purchased. Confidence is expressed as a percentage and helps to understand the strength of association between different items or sets of items (Blattberg et al. 2008).

In this research study, the model of association rules will use the 'apriori' function from the 'efficient_a priori Python package to find frequent itemsets and rules. Further, a dictionary is created to store the confidence of each itemset. A confidence of 0.5 in rule A =>B means that in 50% of the cases where A appears, B also appears.

3.5.2 Multilayer perceptron (MLP)

The perceptron is the most basic neural network architecture containing just one neuron, multiple inputs, and a single output. A multilayer perceptron (MLP) is created when multiple perceptons are joined together, making it a fundamental feedforward neural network with hidden layers (Dunham, 2003). Using weighted sums and transfer functions like the hyperbolic tangent, the inputs in an MLP are combined in hidden nodes before aggregating in the output node. Training methods are used to determine the weights of the node connections. However, determining ideal weights on a restricted time scale is difficult. For this research, the model employed is an MLP with several layers, including dropout layers for regularization. The MLP is trained to predict whether a product will be reordered based on the characteristics of the data, including the new features generated from the association rules.

The architecture of the MLP is deployed using TensorFlow and Keras in Python using Jupyter for a binary classification task on the empirical dataset. The step-by-step breakdown of the architecture of MLP is as follows:

- Input layer: The input layer has a shape based on the number of features extracted from the dataset.
- Dropout Layer: A regularization technique that randomly sets a fraction of input units to zero during training (here, 30%).
- BatchNormalization Layer: This ormalizes the activations of the previous layer.
- Hidden Layers: Three hidden layers with decreasing numbers of neurons (256, 128, 64) and ReLU activation functions. Each is followed by a dropout layer and a batch normalization layer.
- Output Layer: A single neuron with a sigmoid activation function which is suitable for binary classification tasks.

The model defined is configured and compiled using the Adam optimizer with a learning rate of 0.0001, binary cross-entropy as the loss function which is suitable for binary

classification problems, and a callback function from Keras called the 'EarlyStopping' instance that shall monitor the validation loss and will stop training if the validation loss does not improve for a certain number of epochs (controlled by the patience parameter, set to 10 in this case). These aspects of the architecture were tweaked to get the optimal accuracy and hence these parameters were finalized. The fit() method is used to train the model with X_train and y_train being the training data and labels respectively. A batch size of 64 is taken and the model is run for 50 epochs. The early_stopping callback is used during training.

4 Design Specification

This research project's methodology involves a novel approach to blend two well-known machine learning algorithms: Association Rules and Multi-Layer Perceptron (MLP). This fusion aims to construct a hybrid framework model for analyzing InstaCart data baskets. The Association Rule algorithm will extract frequent itemsets and association rules from transaction data. Simultaneously, the MLP algorithm will construct a neural network, that shall learn the association or patterns among related items and forecasts future purchases. Fig. 4 is a depiction of the system design to followed in order to implement this methodology.

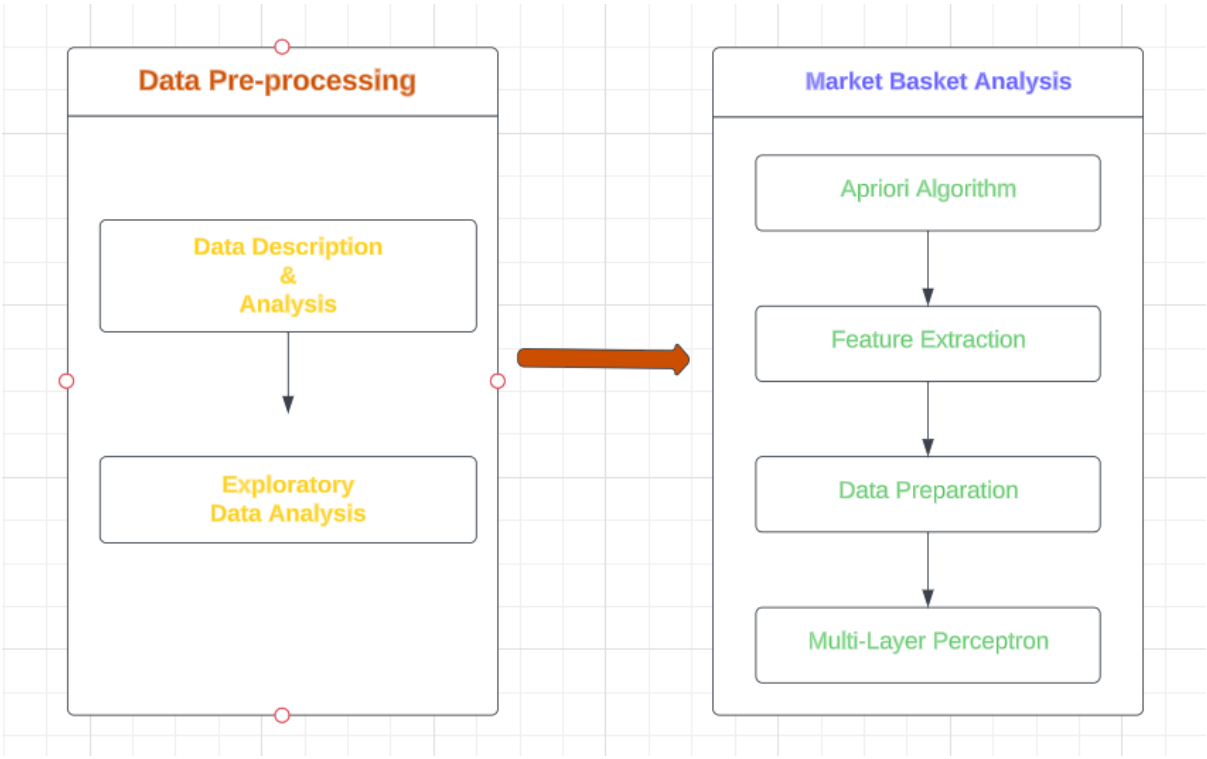

Figure 4: Architecture of the System Design

5 Implementation

This section will delve into the steps undertaken for the implementation of this research project that includes the data mining methodology undertaken using the association rules to extract features for the MLP model. The implementation performed in Jupyter using Python involves two major phases in this study:

5.1 Applying Apriori Algorithm For Feature extraction

Agarwal et al. introduced the Apriori algorithm through their groundbreaking works in 1994. This algorithm was designed to solve the challenge of finding patterns of frequently co-occurring items in extensive databases of transactions, with the aim of creating rules that describe these associations.

All the data pre-processed as mentioned in section 3.3 is loaded and grouped together by order and product. Unwanted columns are dropped. The frequencies of orders based on the products are found and 100 most frequent order items are considered. The dataset contains huge amount of data, thus a subset of the data is taken to extract the association rules from it. Here the basket is segmented by considering 100000 records.

The pre-processed data is then encoded into a binary format suitable for the algorithm. The encoding function is used to convert the values in the basket pivot table into a binary format. If the value is greater than or equal to 1, it's encoded as 1 (indicating the product was reordered), and if it's less than or equal to 0, it's encoded as 0 (indicating the product was not reordered). This kind of binary format is necessary in association rule mining to find patterns like "if a customer orders product A, they are likely to also order product B." Once all the necessary transformations have been carried out on the data, it is subsequently feeded to the the 'apriori' function from the 'efficient_apriori' Python package responsible for generating the association rules. These generated rules act as features for the neural network used in the next phase.

5.2 Multi-Layer Perceptron Model for Classification Prediction

MLPs have the capability to generalize training data to make predictions on unseen data. Given that the task here is to conduct market basket analysis and that the feature extractions from the Apriori algorithm are to be feeded as input to the MLP, it can be expected that the MLP model could give better predictions for the MBA. For this, the MLP model built in the modelling phase is saved for further experiments. Fig.5 shows the architecture of the saved MLP model.

6 Evaluation

In order to provide a comprehensive analysis for the novel proposed methodology, the MLP-Apriori Algorithm results shall be compared with the stand-alone MLP Algorithm. For this, the research is executed into two experiments:

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 256)	3072
dropout (Dropout)	(None, 256)	0
batch_normalization (BatchNormalization)	(None, 256)	1024
dense_1 (Dense)	(None, 128)	32896
dropout_1 (Dropout)	(None, 128)	0
batch_normalization_1 (BatchNormalization)	(None, 128)	512
dense_2 (Dense)	(None, 64)	8256
dropout_2 (Dropout)	(None, 64)	0
batch_normalization_2 (BatchNormalization)	(None, 64)	256
dense_3 (Dense)	(None, 32)	2080
dropout_3 (Dropout)	(None, 32)	0
dense_4 (Dense)	(None, 1)	33

Total params: 48,129
Trainable params: 47,233
Non-trainable params: 896

Figure 5: Architecture of MultiLayer Perceptron

6.1 Prediction using Local MLP

The first experiment is to simply execute the MLP model using the combined empirical dataset without extracting the features from the apriori algorithm. Hence, this experiment focused only on the loading of the preprocessed data, and the categorical features in the dataset are label encoded as the input to the MLP model.

A 10% fraction of the loaded pre-processed data is randomly sampled using the sample() method so as to reduce the complexity of the data and speed up processing. It is then split into two dataFrames based on the values in the 'reordered' column where 'data_majority' contains rows where 'reordered' is 1 and 'data_minority' contains rows where 'reordered' is 0. 'data_majority' is downsampled to avoid biasness and to match the size of the minority class 'data_minority' by randomly sampling. This makes sure that the state of randomness is reproducable.

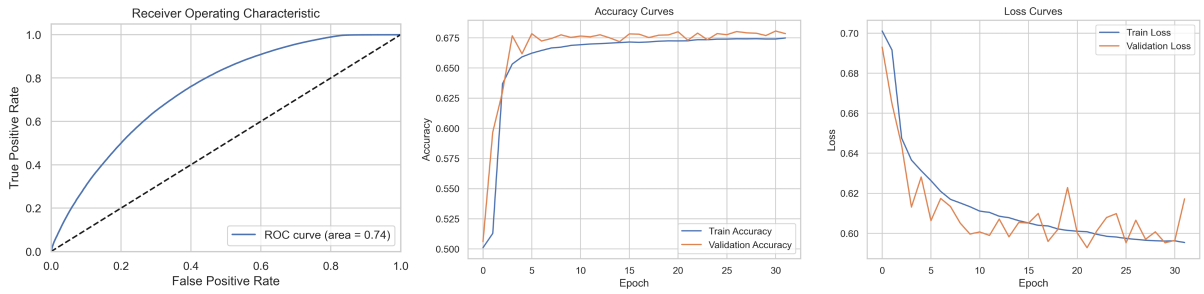

Figure 6: ROC, AUC and Loss Curves Plot respectively for local MLP.

Furthermore, the data is then split into features (X) and the target variable (Y), where the target variable is 'reordered' and the model compiled.

6.2 Evaluating A-priori Algorithm

The apriori algorithm is applied to a basket containing 100,000 randomized records to enhance computational efficiency. It is configured such that the minimum support threshold is set to 0.01, meaning that item sets that appeared in at least 1% of the transactions are considered. Using the frequent item sets association rules are then generated. Using the "lift" metric to generate the strength of these associations, the top five rules are listed in descending order showing the rules with the strongest relationships as in Fig. 5.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
29	(Organic Raspberries)	(Organic Strawberries)	0.04394	0.08810	0.01032	0.234866	2.665899	0.006449	1.191817
28	(Organic Strawberries)	(Organic Raspberries)	0.08810	0.04394	0.01032	0.117140	2.665899	0.006449	1.082912
14	(Banana)	(Organic Fuji Apple)	0.17149	0.02710	0.01078	0.062861	2.319587	0.006133	1.038160
15	(Organic Fuji Apple)	(Banana)	0.02710	0.17149	0.01078	0.397786	2.319587	0.006133	1.375773
4	(Bag of Organic Bananas)	(Organic Raspberries)	0.13489	0.04394	0.01322	0.098006	2.230446	0.007293	1.059940

Figure 7: Association rules generated based on the strength of 'lift'

- Antecedents: The item or set of items that is typically found in the basket. Here 'Organic Raspberries' are the antecedents.
- Consequents: The item or set of items which is likely to be found in the same basket as the antecedents. Here, 'Organic Strawberries' is the consequent of 'Organic Raspberries'.
- antecedent support: This is the proportion of transactions in the dataset that contain the antecedents. Here in the first row, 'Organic Raspberries' are present in approximately 4.394% of all transactions.
- consequent support: This is the proportion of transactions in the dataset that contain the consequents. In the first row, 'Organic Strawberries' are present in approximately 8.81% of all transactions.
- lift: This is the ratio of the observed support to that expected if the antecedent and the consequent were independent. A lift value greater than 1 indicates that the antecedent and consequent are more likely to be bought together than would be expected if they were purchased independently. In the first row, the lift of 2.67 suggests that 'Organic Raspberries' and 'Organic Strawberries' are bought together more than 2.67 times as frequently as would be expected if they were purchased independently.
- leverage: Leverage computes the difference between the observed frequency of the antecedent and the consequent appearing together and the frequency that would be expected if they were independent. A leverage value of 0 indicates independence.
- conviction: A high conviction value means that the consequent is highly dependent on the antecedent. For example, in the first row, the conviction value of 1.19 suggests that the rule Organic Raspberries ->Organic Strawberries is 1.19 times more confident than the independence assumption of the items.

Therefore, in the first row, it can be interpreted that 'Organic Raspberries and 'Organic Strawberries are purchased together more frequently than would be expected if they were purchased independently. If a customer buys 'Organic Raspberries', they are likely to also buy 'Organic Strawberries'. This is based on the measures of support, confidence, lift, leverage, and conviction.

6.3 Prediction using MLP with Association Rule

The second experiment is replicated using the first experiment with the same architecture of the MLP model and same pre-processed data is loaded. However, before label encoding the categorical features, the association rule mining is applied. For this, top 1000 products (N) in the sampled dataset is determined (Hossain et al. (2019) (Ref Section 2.1)). N can be adjusted based on how many products are to be focused on. The data is organized in such a way that each row represents an order and each column represents a product to form a 'basket. The values in the table tell us if a product was reordered in that order (1) or not (0). Moving on, patterns like "if product A is reordered, then product B is also reordered." are looked for. This is done using the Apriori algorithm that helps find groups of products that tend to be bought together often. From the frequent itemsets the "association rules" are generated. Here, "lift" is the metric that is used to evaluate the rules with the minimum threshold set to 1. For each association rule, new features in our data are created. This is done by iterating a loop through the generated association rules. These new features represent the application of the association rules to the dataset.

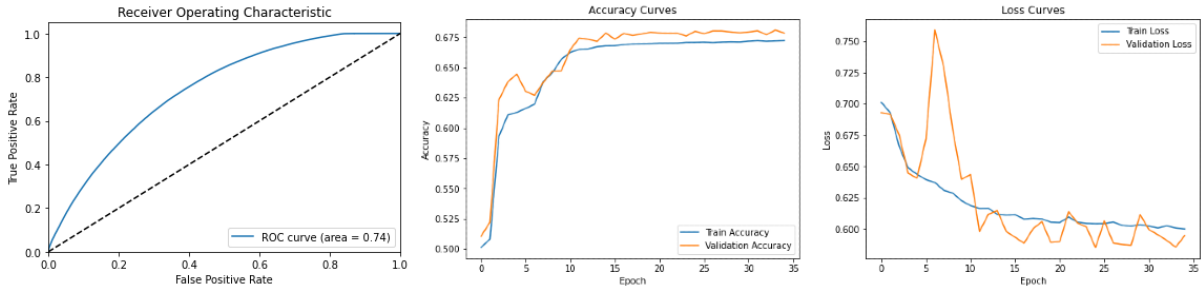

Figure 8: ROC, AUC and Loss Curves Plot respectively for Hybrid MLP.

6.4 Discussion

The predictive model proposed was compared with the stand-alone model and both were evaluated based on the metrics criteria in the methodology (refer section 3.6). Figure 9 shows a summary of the evaluation of the classification report and Fig. 10 provides the Confusion Matrix for Local MLP and Hybrid MLP.

It is worth observing that despite integrating association rules to detect potential patterns of products frequently bought together with a neural network, both trials exhibited the comparatively same performance measurements. This indicates that the enhancements derived from the introduction of features based on association rules are not significant in this research. The accuracy resulted the same for both models, possibly because both use the same architecture. However, the classification report from the MLP-AR model was expected to give better predictions than standalone MLP, which it failed.

Measuring Metrics	MLP Classifier	MLP-AR Hybrid Classifier
Accuracy	0.68	0.68
Precision 0	0.69	0.72
Precision 1	0.67	0.65
Recall 0	0.65	0.58
Recall 1	0.7	0.77
F1 Score 0	0.67	0.64
F1 Score 1	0.68	0.7

Figure 9: Evaluation Metric Summary

Thus, here it is evident that the proposed hypothesis fails and the alternate hypothesis is favoured.

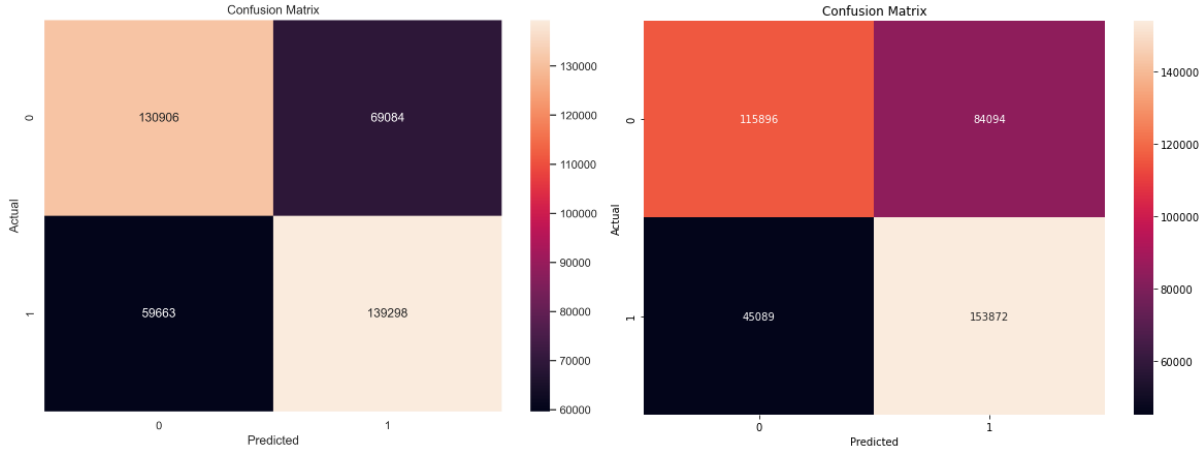

Figure 10: Confusion Matrix for Local MLP(Left) and Hybrid MLP (Right)

Although the ideology behind the objective of finding frequent pattern itemsets through Apriori was to enhance the prediction of the MLP model. However, as visible in the confusion matrices in Fig. 10, the true positive rate for the Hybird MLP was just 0.03 greater than the local MLP model, which is clearly insiginifcant compared to the local MLP.

7 Conclusion and Future Work

This dissertation research presents a novel model that integrates data mining association rules with a deep learning neural network for the task of market basket analysis on an empirical transactional data from an online store.

The results of the proposed research **do support the feasibility of using an integrated framework** for Market Basket Analysis using Apriori and MLP in Consumer

Behavior Modeling. **However, this approach does not produce expected results for the objectives and research question** stated in Section 1.

The research unlike other studies did use an empirically large dataset to facilitate more improved performance in predictions and also used subset baskets for a better reproudcibilty and efficient execution time as mentioned in previous studies. It also experimented with different hyperparameters in the MLP model to optimize its architecture at its best, including validations like binary cross-entropy to ensure the robustness of results. Yet, the fact that there was no improvement in the metrics using association rules brings in scope for improvement. **Thus, this leads to rejecting the hypothesis and accepting the alternate hypothesis.**

In future research endeavors, an advanved or enhanced technique could be used to refine association rule, for example, customer segmentation or using dimensionality reductions.

Acknowledgment

I'd like to take this moment to express my heartfelt gratitude and warm regards to Prof. Rejwanul Haque and Prof. John Kelly. Throughout this project, they have been a guiding mentor. Their unwavering support and direction have helped me overcome the challenges that I encountered during my research.

References

Agrawal Rakesh, Srikant Ramakrishnan, (1994). Fast algorithms for mining association rules. In Proceedings of the 20th International Conference on Very Large Data Bases. 487–499.

Bhargav, A., Mathur, R. P. and Bhargav, M. (2014) 'Market basket analysis using artificial neural network', International Conference for Convergence for Technology-2014. doi: 10.1109/i2ct.2014.7092091.

Blattberg, R.C., Kim, BD., Neslin, S.A. (2008). Market Basket Analysis. In: Database Marketing. International Series in Quantitative Marketing, vol 18. Springer, New York, NY.

Dunham M.H. (2002). Data Mining Introductory and Advanced Topics. Prentice Hall,

Gouda K, Zaki M (2001) Efficiently mining maximal frequent item sets. In: Proceedings of ICDM. IEEE Computer Society, pp. 163–170

Guidotti, R., & Tatti, M. (2019). Comparison of market basket analysis to determine consumer purchasing patterns using fp-growth and A-priori algorithm. In 2019 24th International Conference on Information Systems Analysis and Innovation (ISAI) (pp. 1-6). IEEE.

H. Sorensen, S. Bogomolova, K. Anderson, G. Trinh, A. Sharp, R. Kennedy, B. Page, M. Wright (2017) Fundamental patterns of in-store shopper behavior J. Retail. Consum. Serv., 37 , pp. 182-194

Hossain, M., Sattar, A. H. and Paul, M. K. (2019) 'Market basket analysis using Apriori and FP growth algorithm', 2019 22nd International Conference on Computer and Information Technology (ICCIT). doi: 10.1109/iccit48885.2019.9038197.

Kavitha, M. & Subbaiah, D. S. (2020). Association Rule Mining using Apriori Algorithm for Extracting Product Sales Patterns in Groceries. International Journal of Engineering Research and Technology (IJERT), 8(3), 5–8.

Kutuzova T, Melnik M (2018) Market basket analysis of heterogeneous data sources for recommendation system improvement. Procedia Computer Science 136: 246–254.

Ojugo, A.A. and Eboka, A.O., 2019. Inventory prediction and management in Nigeria using market basket analysis associative rule mining: memetic algorithm based approach. Int J Inf Commun Technol ISSN, 2252(8776), p.8776.

Polat, M., Katircioglu, M., & Kastro, Y. (2019). Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks. Neural Computing & Applications.

Tai CC, El-Shazly M, Chen YH. (2021). Uncovering Modern Clinical Applications of Fuzi and Fuzi-Based Formulas: A Nationwide Descriptive Study With Market Basket Analysis. Front Pharmacol. 2021;12:641530.

Valle, Mauricio A.; Ruz, Gonzalo A.; Morrás, Rodrigo (2018): Market basket analysis: Complementing association rules with minimum spanning trees". In: Expert Systems with Applications. 97, S.&;146–162, DOI: 10.1016/j.eswa.2017.12.028.

Yudhistyra, Wecka Imam, Evri Marta Risal, I-soon Raungratanaamporn, and Vatanavongs Ratanavaraha. "Using Big Data Analytics for Decision Making: Analyzing Customer Behavior using Association Rule Mining in a Gold, Silver, and Precious Metal Trading Company in Indonesia." International Journal of Data Science, vol. 1, no. 2, pp. 57-71, Jun. 2020.