**EXECUTIVE SUMMARY**

The high degree of competition in the retail industry has forced many retail stores such as us to use growing technology to remain successful in the business. One key aspect to do that is to understand the needs of the customers and provide offers that would be beneficial to them. The approach we have adopted here is to analyse the customer transaction data and form patterns of products frequently purchased together. By doing so, we were able to identify a pattern of items frequently bought together. This data can be used to improve the store layout and increase cross-selling opportunities. Furthermore, we have developed a recommendation system for personalizing product suggestions to individual customers based on their buying history and identified patterns.

In simple words, many customers who have bought milk have also bought bread along with it. We can leverage this information to

   i.     Offering bread and milk at a combo price, which the customers would find a good deal,

ii.     Placing bread and milk nearby, so that when customers buy either of them and see the other one nearby, it would make them buy it subconsciously.

Implementing this into our system has several benefits to the company. Some of them are

   i.     **Increased Sales**: By optimizing the layout of the store based on items that are frequently bought together, it would be easier for them to buy their day-to-day groceries without having to go all around the store which many customers don’t prefer. We can improve the discoverability of related products, leading to better opportunities across cross-selling items.

ii.     **Enhanced Customer Experience**: Personalizing product recommendation and offers to customers will increase the brand loyalty to our store and make them stay longer as our customers. We would be their go-to choice for day-to-day grocery needs. This would be reflected in repeated purchases and positive word-of-mouth advertising.

   iii.     **Competitive Advantage**: Using advanced analysis techniques will set us ahead of our competitors. It shows our commitment to understanding customer requirements and delivering personalized experiences.

We have used a representative dataset to build the proposed model. However, it is built with consideration of the need to scale the system to process larger volumes of transactions. It is worth noting that the system built is capable of handling datasets of up to one million customer transactions, making it scalable without compromising performance.

In conclusion, implementing the proposed system has the potential to significantly increase the sales of our store and enhance customer experience. Although the solution is scalable to process large volumes of data, we recommend proceeding with the implementation stage, making sure it is constantly monitored and customers are encouraged to give feedback so that we can enhance the system to work in a live environment.

**INTRODUCTION**

The report focuses on presenting the findings and recommendations of our project on analysing customer behaviour by using a dataset containing customer transaction details and providing personalised recommendations. The aim of undertaking this project is to find out frequently bought items and using this data to increase the sales of the store by modifying the store layout in such a way that frequently bought together items are kept closer, providing combination offers on a set of products as promotions and increasing the sales of items sold less by pairing it with items that are sold more often to improve stock rotation.

With the increase in competition in the retail industry and increase in costs of input involved, it is critical for companies to constantly improve performance to be successful in the business. Among others, some ways to achieve this are updating the sales techniques and customer experience. Our project is aimed at identifying patterns in the co-occurrence of items using customer transaction obtained through the loyalty program. The dataset contains features such as date, customer ID, and items bought with one column per item bought by the customer. By analysing the dataset, we were able to find relationships between different items and create association rules, which would give us a better idea on optimizing the store layout for increasing cross-selling opportunities. Furthermore, we have also developed a recommendation system to personalize product suggestions to customers based on their buying history and frequently purchased items.

The primary objectives of the project were

   i.     **Discover frequent item sets**: Using advanced data mining techniques, specifically the Apriori algorithm, we aimed to identify the item sets that are bought together frequently by customers. This would give us some idea on what products can be placed closer to each other, so that it would be easier for the customers to navigate in search of them. It would enable us to optimize the layout of the store accordingly.

ii.     **Personalized product recommendations**: Using the identified frequently purchased items from the customers’ transaction history, we were able to understand what they were mostly buying. Having this along with the frequent item sets, we were able to generate better offers to customers. For example, if customer A was buying milk frequently and from the frequent dataset, we understood that milk and bread were usually bought together, we were able to offer milk and bread to customer A at a discounted price together.

   iii.     **Optimise the store layout**: Optimizing the store layout is the by-product of this project. It is the way we plan to achieve the ultimate outcome of increasing the sales of the store. However, it is worth noting that optimizing the store layout would have other benefits such as efficient use of resources, being able to rotate stock effectively etc.

**EXPLORATORY ANALYSIS**

This section deals with the methodology of this project from the collection of data to the presentation of the results. Initially, the collection of data was carried out by encouraging the customers to join the loyalty program which would enable us to get the transaction details of the customers. It should be noted here that, even without joining the loyalty program we have the transaction history, but for the purpose of personal recommendation, it is required for the customers to be mapped to the relevant transaction. The transactions of non-member customers can be used in pattern analysis to get a broader picture of what the new or non-regular customers also buy together frequently. This would get a more sophisticated pattern derivation from the dataset. This analysis was aimed to identify patterns in the frequently bought items, and generate personal recommendations based on that to member customers.

**1.** **Data Analysis:**

a. **Overview of the dataset:** The characteristics of the dataset was examined which included the date of transaction, the unique customer identifier and the items purchased. The date of the transaction was useful in identifying seasonal trends in the patterns generated. It helped us to understand the scale of the data and assess the feasibility of our model to handle bigger datasets in the future.

b. **Item Distribution:** Analyzing the dataset gave us an idea of the distribution of individual items and identified the ones which are purchased most frequently. It also helped us understand the popularity of items and how can they be exploited to monetize cross-selling opportunities.

c. **Co-occurrence of items:** We were also able to understand the co-occurrence of items in each item set. Co-occurrence means the presence of one item when another one is present. For example, the presence of milk or bread when the other one is present in a transaction will help us understand how many people strongly want to buy milk and bread together and plan offers accordingly.

**2.** **Proposed Solution:**

a. **Frequent Mining Algorithm:** Apriori algorithm was used to identify the frequent itemset from the transaction data. This algorithm efficiently identified the items that occur together and formed item sets accordingly, which in turn helped us understand customer buying habits better.

b. **Association Rule Generation:** From the obtained frequent patterns, we then generated association rules. We used certain metrics such as support, confidence and lift to measure the degree of associativity between items. These rules helped us to generate recommendations based on the relationships between each individual item and identify cross-selling opportunities.

6. **Collaborative filtering**

Collaborative Filtering (CF) is a technique for recommendation in recommender systems. (Terveen and Hill, 2001) They have two types, one is narrow, and the other is general. (Melville and Sindhwani, 2017) The Collaborative filtering method that can make automatic predictions for a user by collecting the preference information from many users(collaborating) is called narrow. The hidden hypothesis of the collaborative filtering approach is that if person A and person B share the same opinion on a topic, A is more likely to share B's opinion on a different theme than a randomly selected individual. A collaborative filtering recommendation system for television preferences may forecast which television show a user will enjoy based on a partial list of that user's preferences (likes or dislikes). (Meyer,2012)

6.1 **Methodology**

Memory-based CF and Model-Based CF are the two algorithms used in the recommender system. (Gong, Ye and Tan,2009) Memory-based algorithms determine the similarity between two users by analyzing their ratings/purchase history on a group of items. (Gong, Ye and Tan,2009) However, these methods have encountered two significant challenges: sparsity and scalability. (Gong, Ye and Tan,2009) The use of model-based approaches can resolve these problems. (Gong, Ye and Tan,2009) The methodology initially uses memory-based CF to fill the vacant ratings of the user-item matrix, followed by user-based CF as a model-based to from the nearest neighbors of every user. (Gong, Ye and Tan,2009) The predictions for the target user can be produced in real-time using this methodology. (Gong, Ye and Tan,2009) The combined Memory-based CF and Model-based CF can provide better recommendations than traditional collaborative filtering. (Gong, Ye and Tan,2009)

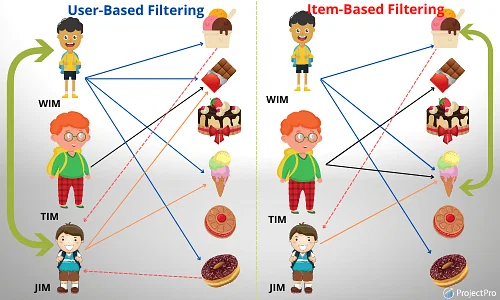


Figure 1. The difference between User-based and Item-based

Collaborative filtering (Model-based)

6.1.1 **Calculating Similarity Matrix**:

The similarity between the user A and active user B must be calculated in-order to find the nearest neighbor where r is the purchase history. Cosine-Similarity is used to compute the similarity between two users and stored in a 2-D array for easy calculation.

Where is the dot product of the vectors A and B.

is the length of vectors A and B.

is the cross product of vectors A and B.

6.1.2 **Finding Nearest Neighbors:**

After calculating the similarity matrix between every user with the other users, the computed similarity values are used in selecting K users, having highest similarity with the active user ‘B’ and the neighborhood is formed for active user.

|  |  |
| --- | --- |
| Member\_number | neighborhood |
| 1000 | [2272, 403, 2267, 2631, 2722] |
| 1001 | [1293, 353, 2804, 92, 3100] |
| 1002 | [2758, 907, 1868, 2910, 3607] |
| 1003 | [1884, 56, 192, 449, 816] |
| 1004 | [1959, 1309, 1674, 480, 2505] |
| 1005 | [3144, 1015, 1670, 96, 374] |
| 1006 | [1733, 3268, 3789, 2558, 2743] |
| 1008 | [333, 820, 2058, 3149, 2942] |
| 1009 | [1546, 770, 1037, 2835, 504] |
| 1010 | [3031, 982, 2091, 3304, 1314] |

Table 1. shows the neighborhood of corresponding customers.

6.1.3 **Recommendation:**

Produce recommendations for the target user using the similarity and the neighborhood.

Figure 2. flow of the model

6.2. **Metrics used for evaluation on test dataset.**

‌ The F1 score, based on precision and recall, has been used as an evaluation criterion for the test dataset. Precision is the ratio of the number of appropriate items to the number of those chosen. (Samundeeswary and Krishnamurthy,2017) It represents the probability that a selected item is relevant such as,

where is the true positive and is the false positive.

Recall is defined as the number of appropriate items that have been chosen together with the total number of available items. (Samundeeswary and Krishnamurthy,2017) It represents the probability that a relevant item will be selected.

where , is a true positive and false negative.

The F1 Score is the harmonic mean of precision and recall.

The harmonics mean weights low values more heavily, which makes the F1 score sensitive to an imbalanced dataset. It considers both precision and recall providing a balanced evaluation of the model's performance. A strong F1 score indicates the model does well in both precision and recall, while a lower F1 score is suggestive of trade between these two metrics. (Samundeeswary and Krishnamurthy,2017

**7. Recommendation method from frequent patterns.**

Apriori is an algorithm for frequent item set mining and association rule learning over relational databases. (Agarwal and Ramakrihnan,1998) It further identifies the frequent individual items in a database and extend them to larger, wider sets so long as these item sets appear more frequently in that database. Association Rules which show general trends in the database may be established using frequent item sets set up for Apriori: they can be applied to different domains such as market clusters analysis. The Apriori algorithm produces results with columns like antecedents, consequents, support, confidence, and lift with frequent items in antecedent and consequent columns.

The most frequent pair or pattern from the dataset are items in the preceding and subsequent columns. As the antecedent is highly likely to have the consequent, pattern values can also be viewed as key value pairs. The method incorporates the recommendations produced from user-based collaborative filtering using similar users and the patterns produced using Apriori algorithm.

The method starts by producing patterns from the dataset with an optimal support value. The Pattern and Recommendation combine to make a new recommendation for the selected user. The Patterns are processed to get the key-value pair of antecedent and consequent. The items in the Recommendation set are taken as the input and compared with the Pattern created by the Apriori algorithm. If there is a resultant for any of the items in the Recommendation set, then they are added to the Recommendation set. The resultant is added to the list when the item is not in the Recommendation set. The output of this method will be a List of Recommendations in addition to the items obtained by Pattern.

8a How patterns and recommendations were ranked.

The recommendations were ranked by the Personalized ranking method. This method considers the user’s personal preference and behavior to generate a personalized ranking. It acknowledges the historical interactions of target user, like past purchases, and incorporates them into the ranking algorithm and sorts them to the frequency of the items in descending order.

c. 10 examples of recommendations from these patterns, two examples from each of the above patterns.

The Table 3 shows the recommendations from the patterns produced by the Apriori algorithm. The recommendation result should have at most five items, but when Patterns have an impact, they tend to have more than five as few items that are bought together are added to the recommendation. For example, for Member 1003 Whole milk is added to the suggestion as whole milk is a pair with sausage.

|  |  |
| --- | --- |
| **Member\_number** | **Recommendation** |
| **1003** | ['other vegetables', 'rolls/buns', 'root vegetables', 'bottled water', 'sausage', 'whole milk'] |
| **1005** | ['rolls/buns', 'shopping bags', 'whipped/sour cream', 'brown bread', 'margarine', 'soda'] |
| **1014** | ['whole milk', 'yogurt', 'tropical fruit', 'sausage', 'pip fruit', 'rolls/buns'] |
| **1016** | ['other vegetables', 'rolls/buns', 'pip fruit', 'shopping bags', 'bottled beer', 'soda', 'whole milk'] |
| **1018** | ['rolls/buns', 'yogurt', 'root vegetables', 'tropical fruit', 'sausage', 'whole milk'] |
| **1019** | ['other vegetables', 'shopping bags', 'beef', 'hamburger meat', 'UHT-milk', 'soda'] |
| **1024** | ['other vegetables', 'root vegetables', 'tropical fruit', 'chicken', 'grapes', 'rolls/buns', 'soda'] |
| **1025** | ['yogurt', 'sausage', 'shopping bags', 'canned beer', 'coffee', 'whole milk', 'soda'] |
| **1029** | ['canned beer', 'frankfurter', 'beef', 'fruit/vegetable juice', 'dessert', 'other vegetables'] |
| **1034** | ['other vegetables', 'soda', 'tropical fruit', 'bottled water', 'bottled beer', 'whole milk'] |

Table 3. Examples showing recommendation from patterns.

e. Table or chart of metrics with brief discussion, showing results of recommendations on training and test sets, **with and without frequent patterns used.**

The metrics used for evaluating the model are Precision, Recall, F1 score. The precision is the measure of positive predictions to the total predicted positives, and recall is the measure of true positives to the total number of true positive and false negative. F1 score is the harmonic mean of precision and recall which gives a balanced measure of the model's performance.

|  |  |  |  |
| --- | --- | --- | --- |
| **Recommendation** | **Precision** | **Recall** | **F1 score** |
| **With patterns** | **0.39** | **0.85** | **0.52** |
| **Without patterns** | **0.15** | **0.5** | **0.19** |

Table 4. Comparison of metrics for recommendation with and without frequent patterns.

The evaluation of the model with the selected metric proves that the recommendation with frequent patterns have high recall and a balanced f1 score. High recall value means that the model can predict positives and negatives better, and it indicates that the model is better to predicting positive examples.

|  |  |
| --- | --- |
| **Member number** | **Recommendation With Pattern** |
| **1003** | ['other vegetables', 'rolls/buns', 'root vegetables', 'bottled water', 'sausage', 'whole milk'] |
| **1005** | ['rolls/buns', 'shopping bags', 'whipped/sour cream', 'brown bread', 'margarine', 'soda'] |
| **1014** | ['whole milk', 'yogurt', 'tropical fruit', 'sausage', 'pip fruit', 'rolls/buns'] |
| **1016** | ['other vegetables', 'rolls/buns', 'pip fruit', 'shopping bags', 'bottled beer', 'soda', 'whole milk'] |
| **1018** | ['rolls/buns', 'yogurt', 'root vegetables', 'tropical fruit', 'sausage', 'whole milk'] |

|  |  |
| --- | --- |
| **Member\_number** | **Recommendation without pattern** |
| **1003** | ['other vegetables', 'rolls/buns', 'root vegetables', 'bottled water', 'sausage'] |
| **1005** | ['rolls/buns', 'shopping bags', 'whipped/sour cream', 'brown bread', 'margarine'] |
| **1014** | ['whole milk', 'yogurt', 'tropical fruit', 'sausage', 'pip fruit'] |
| **1016** | ['other vegetables', 'rolls/buns', 'pip fruit', 'shopping bags', 'bottled beer'] |
| **1018** | ['rolls/buns', 'yogurt', 'root vegetables', 'tropical fruit', 'sausage'] |

**REFLECTION**

This project emphasizes the importance of analyzing data to improve in the business. This can be done in many ways, from cutting unnecessary costs to increasing profitable spendings. The success of the analysis and recommendation is mainly dependent on customer transaction data.

Collecting user feedback, keeping track of certain key metrics and conducting regular performance analysis would help us, as a business, on a periodic basis to identify areas of weaknesses and strengthen them rather than to wait until it has turned into something catastrophic.

**References:**

Terveen, L. and Hill, W. (02 2001) ‘Beyond Recommender Systems: Helping People Help Each Other’.

Melville, P. and Sindhwani, V. (2017) ‘Recommender Systems’, Encyclopedia of Machine Learning and Data Mining, pp. 1056–1066. doi:10.1007/978-1-4899-7687-1\_964.

Meyer, F. (2012) Recommender systems in industrial contexts. Cornell university.

Gong, S., Ye, H. and Tan, H. (2009). Combining Memory-Based and Model-Based Collaborative Filtering in Recommender System. 2009 Pacific-Asia Conference on Circuits, Communications and Systems. doi: https://doi.org/10.1109/paccs.2009.66.

Samundeeswary, K & Krishnamurthy, V 2017, Comparative study of recommender systems built using various methods of collaborative filtering algorithm, IEEE Xplore, pp. 1–6, viewed 20 January 2023, <https://ieeexplore.ieee.org/document/8272663>.

‌ Kumar, A. (2022). Accuracy, Precision, Recall & F1-Score - Python Examples. [online] Data Analytics. Available at: <https://vitalflux.com/accuracy-precision-recall-f1-score-python-example/#:~:text=The%20higher%20the%20recall%20score>.

Agrawal, R. and Ramakrishnan Srikant (1998). Fast algorithms for mining association rules. pp.580–592.

‌