MEF UNIVERSITY

GROCERY SHOPPER MARKETING APPLICATIONS: MARKET BASKET ANALYSIS AND RECOMMENDER SYSTEM

Capstone Project

Dost Karaahmetli

ISTANBUL, 2020

EXECUTIVE SUMMARY

GROCERY SHOPPER MARKETING APPLICATIONS:
MARKET BASKET ANALYSIS AND RECOMMENDER SYSTEM

Dost Karaahmetli

Advisor: Asst. Prof. Hande Küçükaydın

SEPTEMBER 2020, 33 pages

The rapid growth of consumer adoption of online grocery shopping catalyzes the generation of ubiquitous data on shopping behavior, along with a strong competition among the players in this field. The data all present and accounted for all of these players becomes a vital source for their profitability, quality of service, and customer satisfaction. Like all e-commerce platforms, online grocery shopping services rely on transactional data. They can benefit from it to improve their services, create targeted campaigns for cross-selling their products, and establish sustainable relationships with their customers through personalized communications. There is no doubt that understanding customer behavior and predict their future actions would lead to an increase in sales. This project introduces market basket analysis using the apriori algorithm and a predictive machine learning model using the XGBoost classification algorithm to forecast the next item to be repurchased. It proposes a couple of collaborative filtering-based recommender systems to capture Instacart's customers' shopping patterns.

Key Words: Market Basket Analysis, Association Rule Mining, Apriori Algorithm, Predictive Model, XGBoost Classifier, Random Search Cross-Validation, Recommender System

ÖZET

MARKET ALIŞVERİŞÇİ PAZARLAMASI UYGULAMALARI:
SEPET ANALİZİ VE ÖNERİCİ SİSTEM

Dost Karaahmetli

Tez Danışmanı: Dr. Öğr. Üyesi Hande Küçükaydın

EYLÜL 2020, 33 sayfa

Çevrimiçi alışverişin tüketici tarafından benimsenmesindeki hızlı büyüme, bu alandaki oyuncular arasında güçlü bir rekabetin yanı sıra, müşteri davranışı hakkında yaygın ve ulaşılabilir verilerin üretilmesine zemin oluşturmuştur. Tüm bu oyuncular için mevcut veriler, karlılık, hizmet kalitesi ve müşteri memnuniyeti için yadsınmaz bir kaynak haline gelmiştir. Tüm e-ticaret platformları gibi, çevrimiçi market alışveriş platformları da hizmetlerini iyileştirmek, ürünlerini çapraz satışa sunmak için hedefli kampanyalar oluşturmak ve kişiselleştirilmiş iletişim yoluyla müşterileriyle sürdürülebilir ilişkiler kurmak için bundan yararlanabilir. Hiç şüphe yok ki müşteri davranışını anlamak ve gelecekteki eylemlerini tahmin etmek satışlarda artışa yol açacaktır. Bu proje, "apriori" algoritması kullanan bir sepet analizi ve sepete girecek bir sonraki ürünü tahmin etmek için XGBoost sınıflandırma algoritması kullanan bir kestirimsel makine öğrenimi modeli ortaya koyarak, Instacart müşterilerinin satın alma davranışlarını yakalamak için işbirliğine dayalı filtreleme (collaborative filtering) yöntemiyle öneri sistemleri ileri sürmektedir.

Anahtar Kelimeler: Pazar Sepet Analizi, Birliktelik Kural Çıkarımı, Apriori Algoritması, Kestirimsel Model, XGBoost Sınıflandırma Algoritması, Rastgele Arama, Öneri Sistemleri

TABLE OF CONTENTS

.

1. INTRODUCTION

Along with the overall development of e-commerce, a boost for the growth of online grocery shopping is expected. However, the expected boost is delayed up until the Coronavirus pandemic. Despite the fact that the value of the U.S. online grocery market has grown from $12 billion in 2016 to $26 billion in 2018 [1], it is still a tiny portion of the overall market size, which was $632 billion in 2018, according to IBISWorld [2]. Although online retail develops in many sectors, food and grocery are one of the less developed ones compared to apparel and homewares spending. However, the Food Marketing Institute and Nielsen (FMI) have updated their predictions on online grocery shopping from $100 billion to $143 billion by 2025 in their latest report [3].

We definitely know that the main reason behind the delayed growth of online grocery shopping is habitual. We will see whether the physical restrictions, such as staying home, introduced with the pandemic shall change the habits of grocery shoppers or not. While this is a question to be answered in the near future, this project tries to employ a market basket analysis to investigate shopper behavior by analyzing previous purchases and building a recommender system in order to enrich the experience for Instacart's customers.

Instacart, founded in 2012 in San Francisco, is an online grocery delivery company – like Getir in Turkey – offering services via a website and a mobile app over 5,500 cities in all U.S. states and Canada in partnership with over 350 retailers that have more than 25,000 grocery stores [4]. Instacart does not own any grocery store but provides a platform where they enable retailers to sell their products. The revenue model of Instacart depends on the income from delivery fees and advertising earnings from local grocery manufacturers or retailers.

This project provides a framework for Instacart to make use of their customer transaction data focusing on descriptive analysis of purchase patterns to find out which items are bought together and when they are bought. As well as exploring customer shopping behavior by implementing market basket analysis, this study builds a recommender system that suggests the next likely item to purchase to the customer to improve the shopping experience and drive higher engagement.

1.1. Literature Review on Market Basket Analysis and Association Rule

Market Basket Analysis (MBA), also known as affinity analysis, is a widely used technique for revealing the relationship between the items in a transaction. It is instrumental in determining future strategy and decision making. Thus, a massive number of studies have been conducted in this field.

The procedure of MBA offers an insight into shopper behavior. In earlier research, data mining with market basket analysis method is implemented using Minimarket X data, which is provided by a supermarket chain in India. In order to understand the shopping habits, the Apriori algorithm is performed, in which popular items in the data and the pair of items in a transaction are explored [5]. The knowledge of products that frequently go together within the same deal is generated as per the Hybrid-dimension Association Rules criteria. Pursuant to the Association Rules process, the executives of Minimarket X could benefit from the considerable correlation between the data in terms of support and confidence for their future decisions.

Retailers, especially grocery stores and e-commerce mega-sites, deal with thousands (if not millions) of items supplied to their customers. As most techniques are elusive and computationally expensive, high dimensionality can be an issue, making it awfully hard for the businesses to draw conclusions from their analysis. van Maasakkers [6] introduces a three-step autoencoder-based model aiming to tackle this issue. The first step is to summarize the basket by bringing it into a low dimensional code via autoencoder. Then, the prediction of the next basket proceeds by using this code. The set of items a customer might buy next, a.k.a. “the next basket,” is revealed in the third step by decoding.

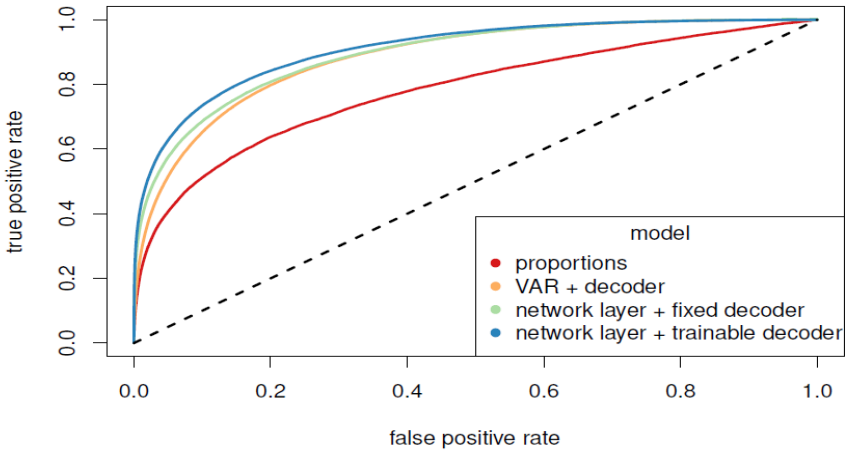

Figure 1: ROC Curve comparing the three models to the benchmark.

As seen in Figure 1, all three models tested in the study outperform the benchmark, while the neural network with the trainable decoder comes out as best performing according to the evaluation by means of the ROC curve (van Maasakkers, 2019).

Another way to tackle the high dimensionality issue causing a huge number of Association Rules might be employing a graph-based methodology based on minimum spanning trees (MSTs). A complementary methodology to association rules mining has been proposed and applied to a grocery store transactional data by Valle et al. [7]. The graphics-oriented approach helps to display the strong interdependencies between the same category products. In addition, minimum spanning trees give the opportunity to expose the most important products with links to different categories. The MST enables the determination of association rules in the simplest possible manner. Another advantage of the MST disclosed by the study is that it facilitates the detection of taxonomic relations and clusters of subcategories.

1.2. Literature Review on Recommender Systems

Recommender Systems, which are sometimes referred to as Recommendation Engines, are supporting systems that help users/buyers to find information, products, and/or services by analyzing suggestions from other users and/or historical transactional data. Most of the time, the output is a match of offer and demand. In the marketing field, they are explicitly used for cross-selling and up-selling purposes and the enhancement of customer experience.

Since the beginning of the century, many researchers have studied new approaches and real-life application of recommender systems. In an in-depth reflection, diverse characteristics, process phases, and means of various prediction models of recommender systems, including collaborative filtering, content-based filtering, and hybrid filtering techniques have been examined. While discussing learning algorithms used in generating recommendation models and evaluation metrics of these algorithms, recommender systems ease the information overload complication and allow customers to attain products and services that are not readily available [8].

Based on the input data and the nature of the problem, the filtering method to be used varies. When something like a web page, a news article, or any publication such as books and movies is to be recommended, conducting a content-based filtering technique would be

more successful. It is a domain-dependent technique often relying on explicit input, which is declared by users regarding their interest in products. The emphasis in content-based filtering is on the attributes of items, rather than the user profiles. On the other hand, collaborative filtering depends mostly on implicit data, which is historical, previous transactions of users in this case. Collaborative filtering (CF) aims to determine user-item associations by analyzing relationships between users and interdependencies among products [9].

Media consumption consisting of purchases of content-based products (e.g., books, movies, music) stayed in the heart of recommender systems studies for long before the applications in other commercial domains. Ming Li et al. [10] illustrate a recommender system concerning grocery shopping where consumers make repetitive purchases and buy more than one of a particular product. The authors introduce a personalization algorithm modeling shopper behavior to identify new items that are likely of interest to a particular customer. The study compares their unique approach, a basket-sensitive random walk model for a personalized recommendation, to the conventional collaborative filtering model by experimenting on three real-world data sets. The performance of traditional CF models is limited due to their inability to explore transitive associations between the products that have never been purchased together. Basket sensitive random walk model, where the similarity is defined as the transition probability between products instead of users, has overcome the limitations of the rating-oriented model.

The remainder of the report is organized as follows. Section 2 defines the project, where the vision, the scope, and the stepwise work effort of the study are presented. Section 3 introduces the source of the data, the datasets, the data preprocessing steps, and the insights provided by data as a result of exploratory data analysis. Section 4 contains the approaches, techniques, and algorithms to be conducted. Finally, the results of the research are given with remarks leading to further studies in Section 5.

2. PROJECT DEFINITION

As the online grocery shopping industry gets more competitive every day, Instacart has to extract more insights from data to acquire more customers, to increase customer retention, and to bring the shopping experience to perfection through its channels. In order to fulfill these objectives, Instacart, like other retailers, can benefit substantially from drawing valuable insights from customers' purchase histories.

1.1. Problem Statement

We would not be surprised when we see that a bottle of ketchup and a bottle of mayonnaise are pulled out of the same shopping bag. That is why we often see some items are bundled to be sold together, or many fast food joints create menus, including meals, beverages, and desserts. These actions are obviously done to increase sales. Similarly, the items which tend to be bought together are displayed close to one another to improve the customer shopping experience.

However, in most cases, the fact that particular objects bought together may not be as apparent as the "ketchup & mayonnaise" or the "burger & fries" cases. The famous "beer & diapers" story of Wal-Mart is a particularly good example of such a situation [11]. After combining the data from their loyalty card system with those from their POS systems, Wal-Mart marketers found out that on Friday afternoons, young males who have bought diapers also had a predisposition to buy beer. After Wal-Mart moved the beer next to the diapers, beer sales went up. The origins of this story are still in question, and some indicate that it is a myth, but in any case, it is an excellent example of data-driven marketing. The relationship between two different items out of the same shopping bag reveals valuable information and can be used for promotions, cross-selling, up-selling, and other recommendations.

By analyzing historical purchasing data from the Instacart Public Datasets, this study intends to explore and seek solutions for the following problems: identifying products bought together, predicting which products will be in a customer's next order, and recommending new products to customers.

2.2. Project Objectives

Enabling the customers to have an improved shopping experience, therefore increasing engagement, and boosting the effectiveness of promotion and sales strategies, are the two main marketing objectives of online grocery retail businesses.

This project aims to address the above-mentioned marketing objectives of Instacart while performing basket analysis, shopper behavior analysis, and building a recommender system using the historical data on customer orders. While analyzing the data of 3 million grocery orders from Instacart customers, this project aspires to:

- Perform Explanatory Analysis of the data to reveal a fact about transactions,
- Find out hidden patterns between products for better cross-selling and up-selling:
    - To identify the frequent items from a transaction based on support and confidence,
    - To generate Association Rule from the frequent item sets,
- Develop a model to predict the next likely product the customer would purchase and whether it will be reordered or not,
- Build a system to predict and suggest products to individual customers.

2.3. Project Scope

This project tries to find patterns in the data, such as how often people reorder the same product they have purchased before and which items are purchased at different times or intervals. It aims to predict which previously purchased products are more likely to be reordered by the customer and what other products could be offered to him/her which he/she would be more likely to be interested. The recommender system does not only engage the customers by personalizing suggestions and enhance their user experience but also the predictions by the system can be used for targeted marketing activities.

The scope of the work consists of data preprocessing, explanatory data analysis, market basket analysis, predictive model to identify which items would appear in the next order, and a recommender system. The deliverables of the project are project report, Python code partitions in ".ipynb" format, output visualizations (figures, tables, graphs), and a reference list.

3. ABOUT THE DATA

Instacart released an anonymized dataset in 2017 for public use [12], which is also made use of in this project. “The Instacart Online Grocery Shopping Dataset 2017” is available on the Instacart website as well as on Kaggle [13].

Jeremy Stanley, the former V.P. of Data Science at Instacart, has provided the data dictionary on his GitHub repository, to be referred to for further information about file descriptions [14].

3.1. Datasets and Features

The dataset contains a sample of over 3 million grocery orders from more than 200,000 users. It comes with six .csv files, each of which containing different tables:

- orders.csv: specifies which set (prior, train, test) an order belongs as well as detailed order information
    - order_id: order identifier
    - user_id: customer identifier
    - eval_set: which evaluation set this order belongs in
    - order_number: the order sequence number for this user
    - order_dow: the day of the week the order was placed on
    - order_hour_of_day: the hour of the day the order was placed on
    - days_since_prior: days since the last order capped at 30
- order_products_prior.csv & order_products_train.csv: specify which products were purchased in each order, their add-to-cart order and whether they are reordered
    - order_id: foreign key
    - product_id: foreign key
    - add_to_cart_order: order in which each product was added to cart
    - reordered: 1 if this product has been ordered by this user in the past, 0 otherwise
- aisles.csv: specifies the ids and names of aisles
    - aisle_id: aisle identifier
    - aisle: the name of the aisle

- departments.csv: specifies the ids and names of departments
    - department_id: department identifier
    - department: the name of the department
- products.csv: specifies the products information
    - product_id: product identifier
    - product_name: name of the product
    - aisle_id: foreign key
    - department_id: foreign key

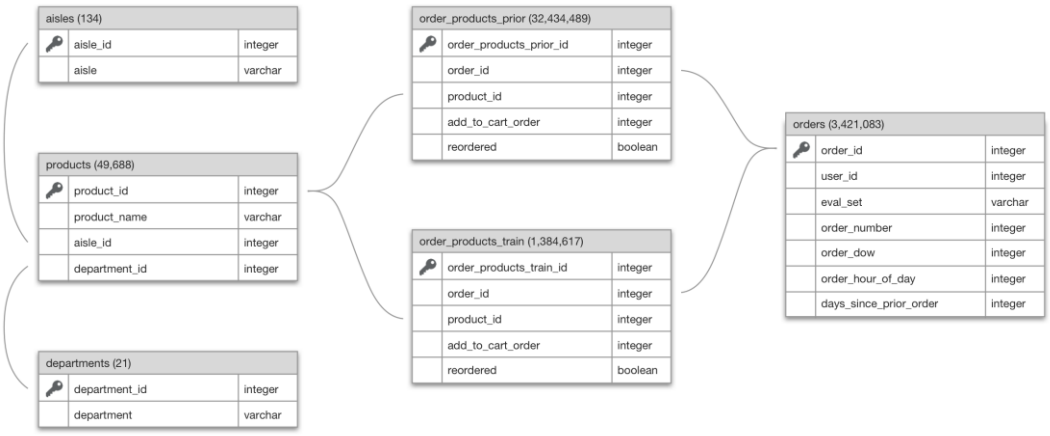

Figure 2: Data Structure

Figure 2 displays the relation between the datasets used in this project [15]. This diagram helps to determine the index columns when merging the datasets.

3.2. Data Preparation

The information in the data is delivered through six datasets. Examination prior to the data wrangling process shows that the data is mostly clean and only some minor cleaning needed. Merely, the "Orders" dataset had NaN values in "days_since_prior_order" variable, which is then converted to zero. For the purposes of the project, the datasets need to be merged together for further analysis. The merger of datasets is done in the following order:

1. Combine aisles, departments, and products
2. Combine Prior Orders and Products
3. Combine Train Orders and Products
4. Combine Prior orders and Order Details
5. Combine Train Orders and Order Details
6. Merge the Train and Prior Dataset

3.2.1. Feature Construction

In order to train a more robust and reliable model, new features are extracted from the existing data and added to the existing ones. Thanks to this process, new information will be accessible for the model construction and, therefore, hopefully, result in a more accurate model. Reig Grau [16] states that the task of feature construction is about retrieving features that add value to the model by creating and transforming of variables that represent the behavior and tendencies hidden in the data.

At the end of this process, a final dataset that captures the behavior of the customers will be generated. Extracted new features are:

1. Order Size: The number of items added per order.
2. Total Order Count: The number of transactions carried out per customer.
3. Loyal Customers: The ones who ordered more than average order counts.
4. User Product Count: The number of times a customer purchased a product.
5. User Department Count: The number of times a customer purchased a product from the particular department.
6. Weekend Customers: Whether a customer prefers to buy on weekends or on a weekday.

3.3. Explanatory Data Analysis

After the data preprocessing phase, an explanatory data analysis is conducted to gain more insight from data and to have more understanding about the problem at hand. Explanatory data analysis demonstrates a general idea about customer shopping behavior, which will help the inferences further in this study.

The first question to be asked to the data is that whether people usually reorder previously ordered products. Figure 3 shows that 60% of the products are reordered.

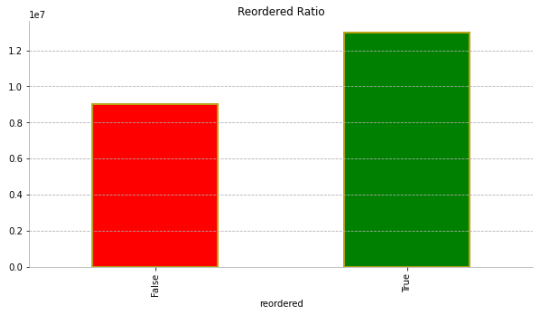

Figure 3: Reorder Frequency

In Figure 4, a word cloud of the top 400 most popular products is displayed. The larger the name of the product, the more popular it is in Instacart shopping carts. It looks like organic products are trending.

Figure 4: Word Cloud of Popular Products sold at Instacart

Moving on with the popular products, the Top 25 most bought products in Instacart are plotted in Figure 5. This plot reinforces the finding that organic products are among the most popular purchases.

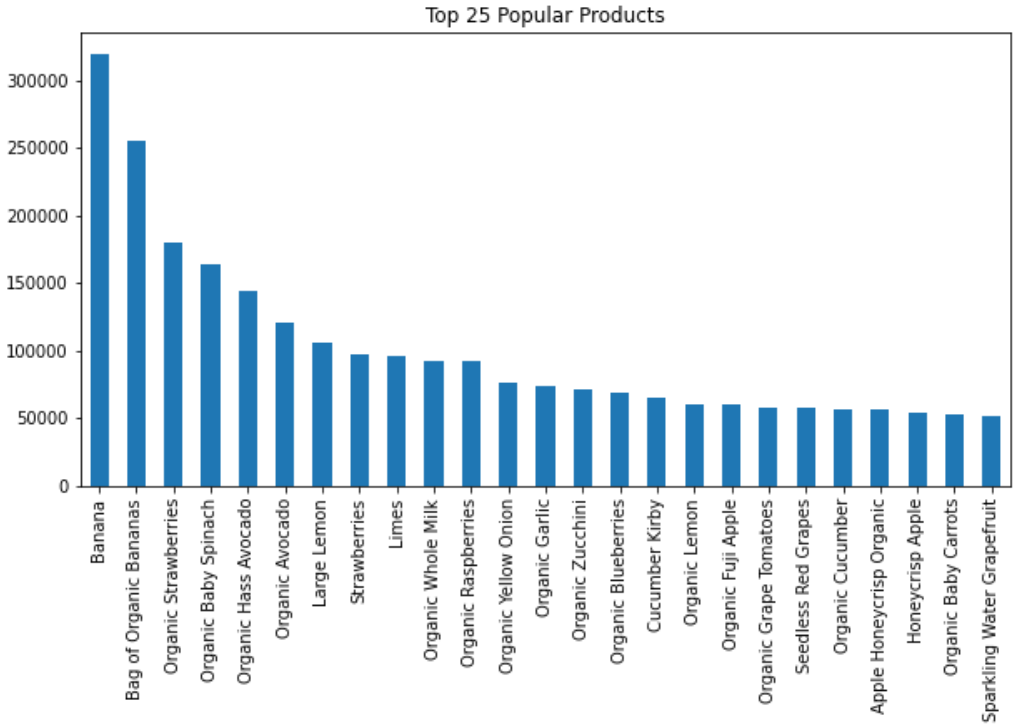

Figure 5: Top 25 Popular Products

There is a strong positive correlation between the most popular products and the most reordered products, as seen in Figure 6.

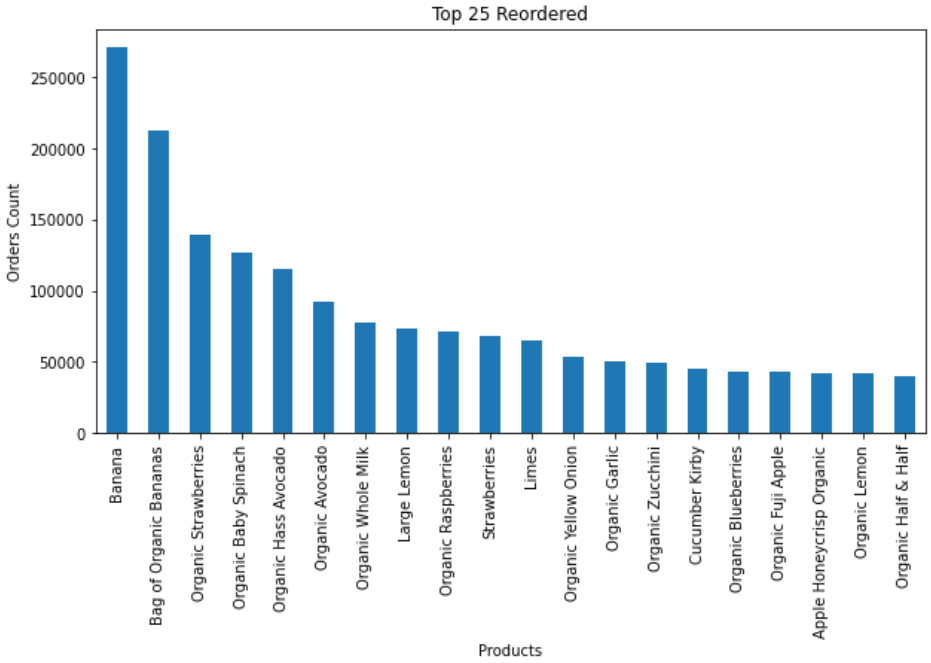

Figure 6: Top 25 Reordered Products

The distribution of purchases according to departments in Figure 7 reveals that more than half of the products sold at Instacart comes from only 3 departments.

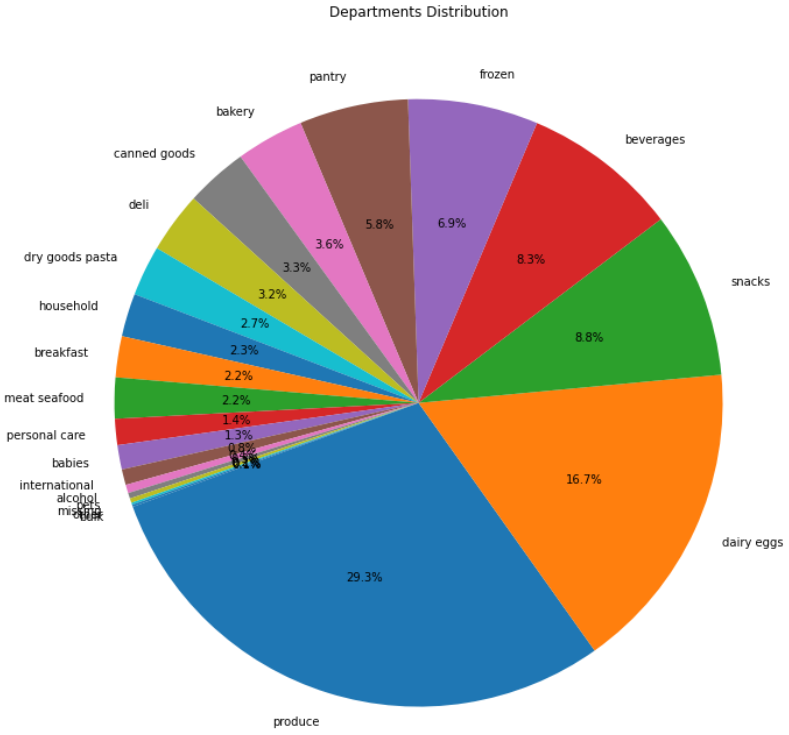

Figure 7: Order Distribution by Department

A similar correlation between most frequently purchased and repurchased items is observed when plotting the order distribution by departments as displayed in Figure 8.

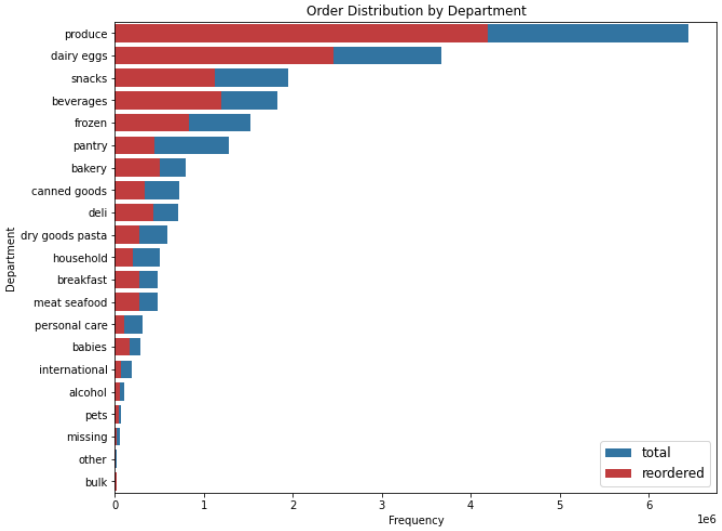

Figure 8: Order/Reorder Distribution by Department

Figures 9 and 10 display the frequency of transactions in terms of the day of the week and hour of the day, respectively. It is no surprise that weekends and afternoons are the busiest periods of the business.

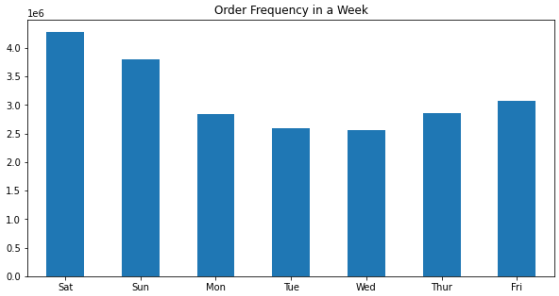

Figure 9: Weekly Order Frequency

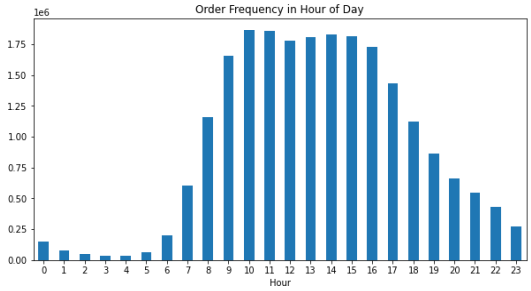

Figure 10: Hourly Order Frequency

Transactional data suggests that customers usually order 10 to 20 items at a time as Figure 11 indicates.

Figure 11: Order Size Trends

Customers tend to put in reorders within a week as shown in Figure 12. The reason why the frequency in reorders is high on the 30th day is because the plot displays the first 30 days after the prior order. The orders put in after a month are counted as the 30th day. Thus, customers can place orders right after their prior order and even a month apart.

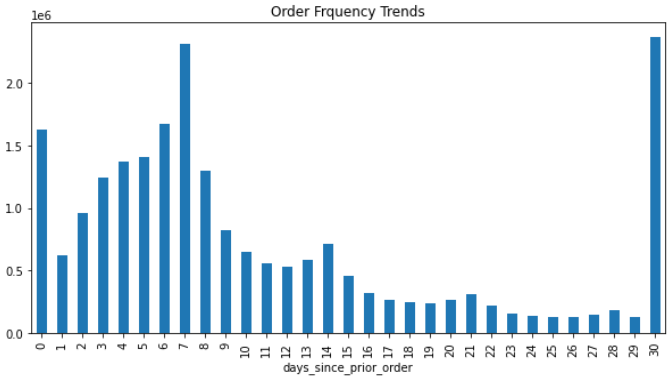

Figure 12: Order Frequency: Order Interval of Customers

However, the average frequency of customers placing a new order is 14 as introduced by the histogram in Figure 13.

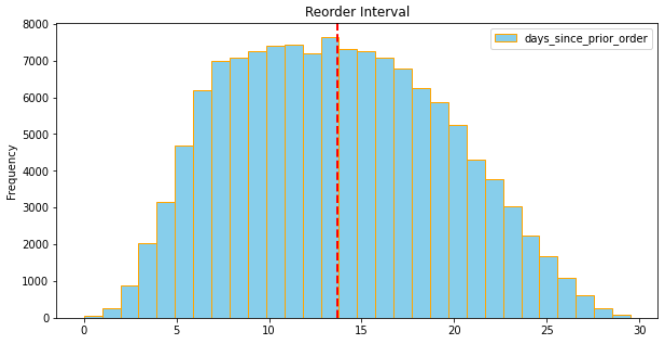

Figure 13: Average Reorder Interval

The scatter plot in Figure 14 presents the relation between customers and the number of transactions. There is a strong positive correlation between the number of unique customers and the number of purchases.

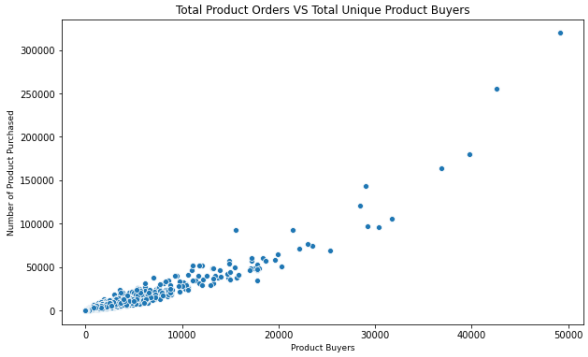

Figure 14: Total Product Orders vs. Total Unique Customers

Figures 15 and 16 demonstrate the plot of the reorder ratio against total orders and total customers, respectively. These plots display a ceiling effect. While certain products are reordered regularly, many people try some products only once and they do not reorder them.

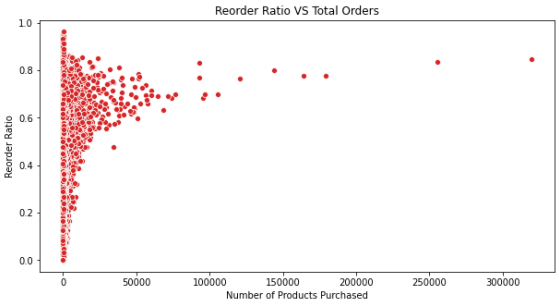

Figure 15: Reorder Ratio vs. Total Orders

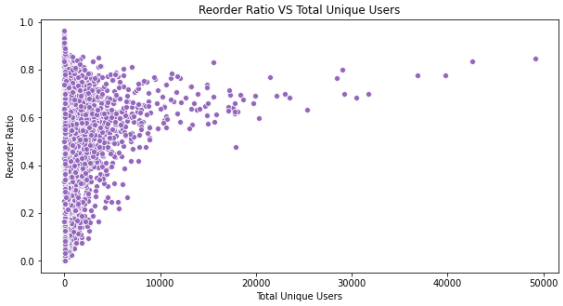

Figure 16: Reorder Ratio vs. Total Customers

As seen in Figure 17, less than 16% of customers are returning users meaning that they have placed reorders.

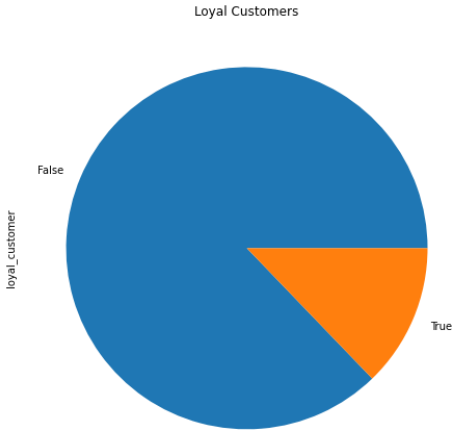

Figure 17: Loyal Customers

4. METHODOLOGY

By analyzing the customer's shopping trends, which are in common with the customer's shopping patterns, this project aims to enable Instacart the ability to offer bundles of products to their customers for cross-selling or up-selling purposes. The recommendations depending on the analysis in this report, are intended to be attractive so that the customers will be more likely to end up buying the bundled products instead of just the original item.

In this section, market basket analysis is performed with the help of the apriori algorithm, a powerful technique to extract association rules. Then XGBoost Machine Learning model is built and evaluated to predict what grocery items each Instacart customer reorders based on the user's purchase history. The machine learning task is a binary classification problem taking the “reordered” feature as target variable. Finally, two recommender systems are built with collaborative filtering method using cosine similarity and bigrams to recommend product bundles. Figure 18 displays the flowchart of the proposed methodology of the project.

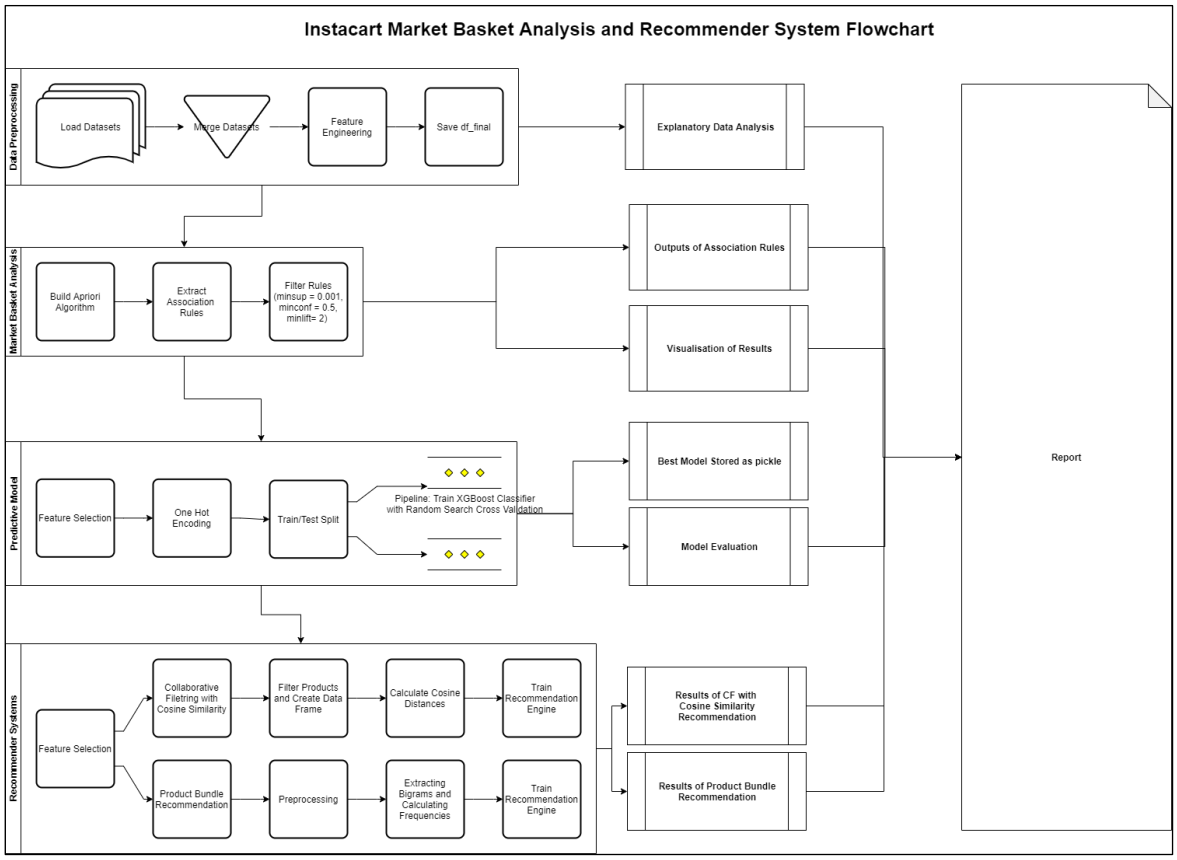

Figure 18: Overview of the proposed methodology

4.1 Market Basket Analysis (MBA)

Simply put, market basket analysis explains the combination of merchandise that regularly co-occur in transactions. The method uses data to decide which products should be cross-sold or promoted together based on the products customers tend to buy together. Association Rule Mining is used to find associations, a.k.a. frequent patterns between different objects in a transaction database. Association Rule Mining using the apriori algorithm is employed in this study.

4.1.1. Affinity Analysis (Association Rule Mining)

Association rule mining or affinity analysis is one of the common technologies proved to be very helpful when dealing with customers, especially while implementing efficient business strategies for attaining more sales [17]. The main purpose of this technique is to create frequent patterns and to create association rules. The analysis is performed on the collection of data items, market basket in this case, by selecting one or more items and support by measuring the dependency of each item on others.

Let $I = \{I_1, I_2, \ldots, I_n\}$ be a set of items and let $D = \{T_1, T_2, \ldots, T_n\}$ be a set of transactions. Every transaction in *D* has a unique transaction I.D. and contains a subset of the items in *I*. A rule in the database is defined as an implication of the form: $I_1 => I_2$. The sets of items I_1 and I_2 are called antecedent (left-hand side) and consequent (right-hand side) of the rule, respectively [18].

The significance of an association rule is measured based on three parameters, namely support, confidence, and lift.

Support is the default popularity of an item. In mathematical terms, the support of item *A* is the ratio of transactions involving *A* to the total number of transactions.

$$Support_A = (\text{Number of Transactions that Contains } A) / (\text{Total Number of Transactions}) \quad (4.1)$$

Confidence is the ratio of the number of transactions involving both *A* and *B* and the number of transactions involving *B*. In other words, it is the likelihood that customers who bought item *A* would also buy item *B*.

$$Confidence_{A,B} = Support_{A,B} / Support_A \quad (4.2)$$

Lift is the ratio that indicates how efficient the rule is in finding consequences compared to the random selection of a transaction, and calculated as follows:

$$Lift_{A,B} = Confidence_{A,B} / Support_B \quad (4.3)$$

If $Lift_{A,B}$ turns out to be equal to 1, then there is no correlation within the item set. If it is greater than 1, then there is a positive correlation. Else, there is a negative correlation within the item set.

4.1.2. Apriori Algorithm

Although computationally expensive, apriori algorithm is widely used in affinity analysis due to its ease of implementation and effectiveness. It is the algorithm behind MBA. Apriori algorithm assumes that any subset of a frequent item set must be frequent. Therefore, the support of an item set never exceeds the support of its subsets. This is known as the anti-monotone property of support.

Apriori principle allows pruning all the supersets of an item set, which does not satisfy the minimum threshold condition. The minimum threshold condition is a parameter set by users for association rule generation. These parameters can be set for support and/or confidence and are used to exclude rules in the result that have support or confidence lower than the minimum support and minimum confidence, respectively. For example, if $\{a, b\}$ does not satisfy the threshold of minimum support (minsup), any item added to this item set will also not cross the threshold [19].

Apriori algorithm uses a bottom-up approach, first generating all frequent item sets, and then extracting all confident association rules from frequent item sets. The algorithm stops when there are no more items to add that meet the minimum threshold requirement.

Association rules mining (ARM) have some limitations other than being computationally expensive, especially when generalizing results. ARM algorithms normally discover a huge quantity of rules and do not guarantee that all the rules found are relevant. So, while working on a fully sophisticated system for recommendations, other techniques such as collaborative filtering, which depends on the similarity of users instead of depending only on the frequency of item sets, are also commonly used.

4.2. Prediction of Next Item by XGBoost Algorithm

For the purposes of this project, a machine learning model that predicts which products a customer will buy again using transactional data is built. This is a supervised classification task. The target variable to be predicted is a binary feature under the "reordered" column.

Before deploying the algorithm, feature selection is applied to the data that has already been obtained at a preprocessing step. The training and test sets are split according to the "eval_set" column so that the "prior" observations set aside as the training set and the "train" observations set aside as the test set. Then, one hot encoding is applied to the categorical variables. After these preparatory steps, a training set of 20,641,991 observations and 13 features is obtained.

XGBoost Algorithm is used on this large dataset due to its reliability, efficiency, and scalability. XGBoost is short for Extreme Gradient Boosting. It is well focused on both computational speed and model performance. It is a tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. Ensemble learning offers a systematic, aggregated solution by combining the predictive powers of multiple learners.

On top of this advantage, the boosting in XGBoost enables the trees to be built sequentially so that each tree learns from its predecessors and updates the residual errors [20]. Randomized Search, where random combinations of hyperparameters are polled to find the best solution for the model, is used in the pipeline for hyperparameter tuning with cross-validation.

4.3. Recommender System

Recommender systems are algorithms aimed to suggest relevant items, such as movies to watch, text to read, products to buy to users. They are very critical in some industries as they can generate a huge amount of income as well as catering a competitive edge to a company in order to stand out from their competitors. In grocery shopping, people tend to make repeated purchases, and product preference is conveyed implicitly in the transaction data instead of being expressed explicitly as ratings like Amazon or Netflix [10].

In this project, two different approaches are employed for building a recommender system for Instacart, which is better suited to the characteristics of grocery shopping data. Both approaches utilize collaborative filtering techniques, while the first one uses cosine similarity measure, the second one is a product bundle recommendation based on the bigram frequency.

4.3.1. Collaborative Filtering Using Cosine Similarity

Collaborative filtering (CF) methods are based solely on past transactions in order to produce new recommendations. These transactions are stored in matrices called "user-item interactions matrix". They are very popular due to their simplicity and relatively good performance.

CF is a similarity-based algorithm that assumes the customers are likely to accept product recommendations that are similar to what they have bought before [21]. Thus, if person *A* has a similar shopping behavior as person *B*, person *A* is more likely to have person *B*'s opinion on a different product than that of a randomly chosen person. The information is gathered by many users, but the predictions are specific to one person.

The effectiveness of CF depends on the similarity measure. The most commonly used similarity measures are cosine-based and conditional probability-based similarities. The conditional probability-based similarity is asymmetric. It is basically the ratio between the number of customers who bought both items and the number of users who bought only one of them. On the other hand, the cosine similarity measure is symmetric and is calculated as the normalized inner product of two feature vectors [22]. In other words, it is the similarity between two vectors by calculating the cosine of the angle between each other. The cosine of a 0-degree angle is 1. Therefore, the more the value is closer to 1, the more similar the items are.

The cosine similarity recommender system built in this study aims to extract recommended products based on similar profiles with similar purchase histories as the target customer.

4.3.2. Product Bundle Recommendation Based on Bigram Frequency

The main idea of this approach is that if the right product bundles can be predicted, relevant products can be offered. Thus, in order to build a product bundle recommendation engine, a two-step process is required. First, finding out which products are frequently bought together, and second, given these product pairs, generating a recommendation list by predicting the next item to be purchased. However, first and foremost, the whitespace between the words under 'product_name' is replaced with underscore "_" for both prior and test datasets, so that each product name is one word with no space in between.

For the first step, bigrams are used, and bigram frequency is calculated. A bigram is a sequence of two adjacent elements from a string. They are typically letters, syllables, or words [23]. The frequency distribution of every bigram in a string is almost exclusively used for simple statistical analysis of text in many applications. In this case, the bigrams are the names of two products that are bought together. For example, if person *A* adds milk, soda, and water to his/her cart one by one, the bigrams will be "milk soda", "soda water." And if person *B* adds water, soda, milk in another order, then the bigrams are "water soda" and "soda milk."

After extracting bigrams, the bigram frequency is calculated and stored in a nested dictionary, where:

- the first layer key is the first word,
- the second layer key is the second word,
- the second layer value is the frequency.

Extracting and training bigrams are deployed by the PySpark package. In the second step, a recommendation list is generated based on the bigram frequency.

5. RESULTS

In this section, the results of the MBA, XGBoost prediction, and recommender systems are demonstrated and interpreted.

5.1. Results of MBA

Association rules mining is employed by Python's 'apriori' library. Figure19 displays the association rules extracted. Antecedents are the original items, combined with the consequences that are used to compare possible combinations of items in the basket. Once these pairs are identified as having a positive relationship, recommendations can be made.

	antecedents	consequents	support	confidence	lift
9160	(Sparkling Water Grapefruit, Sparkling Lemon W...	(Lime Sparkling Water)	0.001816	0.500627	28.202348
9165	(Lime Sparkling Water)	(Sparkling Water Grapefruit, Sparkling Lemon W...	0.001816	0.102305	28.202348
9161	(Sparkling Water Grapefruit, Lime Sparkling Wa...	(Sparkling Lemon Water)	0.001816	0.364563	28.167727
9164	(Sparkling Lemon Water)	(Sparkling Water Grapefruit, Lime Sparkling Wa...	0.001816	0.140315	28.167727
9162	(Sparkling Lemon Water, Lime Sparkling Water)	(Sparkling Water Grapefruit)	0.001816	0.536962	17.722511
...	...	...	...	...	...
3361	(Organic Half & Half)	(Unsweetened Almondmilk)	0.000587	0.019223	1.001412
4076	(Sparkling Natural Mineral Water)	(Sparkling Water Grapefruit)	0.000504	0.030315	1.000566
4077	(Sparkling Water Grapefruit)	(Sparkling Natural Mineral Water)	0.000504	0.016629	1.000566
1970	(Orange Bell Pepper)	(Organic Whole Milk)	0.000719	0.058540	1.000273
1971	(Organic Whole Milk)	(Orange Bell Pepper)	0.000719	0.012285	1.000273

11878 rows × 5 columns

Figure 19: Association Rules with 'min_threshold=1'

After the first experiment, where the minimum threshold is 1, 11,878 rules are displayed. Several experiments are run with various filtering based on support, confidence, and lift. Since Instacart has an enormous number of SKUs (stock-keeping units), in order to have a reasonable number of rules, the minimum support is then set to 0.001. In Figure 20 rules generated where the minimum lift is 2.0, and the minimum confidence is 0.20. 299 rules are returned with these parameters.

As not surprisingly seen in Figure 20, different flavors of similar class products such as lemon and grapefruit flavors of sparkling water have high lift and confidence scores. Similarly, organic foods, fruits, and vegetables, to be more precise, have strong relationships with each other.

	antecedents	consequents	support	confidence	lift
9160	(Sparkling Lemon Water, Sparkling Water Grapef...	(Lime Sparkling Water)	0.001816	0.500627	28.202348
9162	(Sparkling Water Grapefruit, Lime Sparkling Wa...	(Sparkling Lemon Water)	0.001816	0.364563	28.167727
9161	(Sparkling Lemon Water, Lime Sparkling Water)	(Sparkling Water Grapefruit)	0.001816	0.536962	17.722511
1746	(Sparkling Lemon Water)	(Lime Sparkling Water)	0.003382	0.261312	14.720800
1749	(Lime Sparkling Water)	(Sparkling Water Grapefruit)	0.004981	0.280623	9.262002
4074	(Sparkling Lemon Water)	(Sparkling Water Grapefruit)	0.003628	0.280279	9.250632
9394	(Organic Yellow Onion, Limes)	(Organic Garlic)	0.001014	0.258351	6.761840
10690	(Organic Garlic, Organic Hass Avocado)	(Organic Yellow Onion)	0.001483	0.256353	6.317567
5746	(Organic Garlic, Bag of Organic Bananas)	(Organic Yellow Onion)	0.001789	0.241041	5.940223
9926	(Organic Yellow Onion, Organic Baby Spinach)	(Organic Garlic)	0.001413	0.213436	5.586264
10688	(Organic Yellow Onion, Organic Hass Avocado)	(Organic Garlic)	0.001483	0.213111	5.577774
5744	(Organic Yellow Onion, Bag of Organic Bananas)	(Organic Garlic)	0.001789	0.202574	5.301984
10750	(Organic Garlic, Organic Strawberries)	(Organic Yellow Onion)	0.001043	0.214854	5.294865
9928	(Organic Garlic, Organic Baby Spinach)	(Organic Yellow Onion)	0.001413	0.213143	5.252693
9396	(Limes, Organic Garlic)	(Organic Yellow Onion)	0.001014	0.210510	5.187802
8832	(Organic Avocado, Limes)	(Large Lemon)	0.002049	0.288181	5.167818
11671	(Organic Strawberries, Organic Hass Avocado, B...	(Organic Raspberries)	0.001539	0.258101	4.778384
8858	(Organic Garlic, Large Lemon)	(Limes)	0.001155	0.235294	4.776845
8833	(Organic Avocado, Large Lemon)	(Limes)	0.002049	0.235008	4.771033
1526	(Jalapeno Peppers)	(Limes)	0.002959	0.232971	4.729692
8958	(Organic Avocado, Organic Garlic)	(Large Lemon)	0.001106	0.261134	4.682800
721	(Bunched Cilantro)	(Limes)	0.003040	0.227636	4.621379
1773	(Organic Cilantro)	(Limes)	0.004995	0.219789	4.462059
10905	(Organic Raspberries, Organic Lemon)	(Organic Hass Avocado)	0.001361	0.376888	4.336422
8856	(Limes, Organic Garlic)	(Large Lemon)	0.001155	0.239773	4.299751

Figure 20: Association Rules with min_sup = 0.001, min_conf = 0.20, min_lift = 2

Mexican cuisine has some unique food products that go well together. This is clearly reflected when “Organic Avocado” is chosen as antecedent. The products with high confidence and lift scores as consequents to organic avocado shows how Mexican cuisine is sui generis, in Figure 21.

```
[76] rules[rules['antecedents'] == {'Organic Avocado'}].sort_values(['lift','confidence'],ascending=False).head(10)
```

	antecedents	consequents	support	confidence	lift
11860	(Organic Avocado)	(Limes, Organic Baby Spinach, Large Lemon)	0.000519	0.007430	4.776809
11818	(Organic Avocado)	(Banana, Organic Baby Spinach, Large Lemon)	0.000758	0.010857	4.478374
11790	(Organic Avocado)	(Limes, Banana, Large Lemon)	0.000761	0.010890	4.398331
8517	(Organic Avocado)	(Broccoli Crown, Large Lemon)	0.000651	0.009317	4.356308
11832	(Organic Avocado)	(Limes, Banana, Organic Baby Spinach)	0.000587	0.008406	4.179011
8529	(Organic Avocado)	(Broccoli Crown, Organic Baby Spinach)	0.000683	0.009784	4.020382
11776	(Organic Avocado)	(Cucumber Kirby, Banana, Organic Baby Spinach)	0.000536	0.007679	3.936337
8967	(Organic Avocado)	(Organic Garnet Sweet Potato (Yam), Large Lemon)	0.000523	0.007484	3.824379
6687	(Organic Avocado)	(Banana, Broccoli Crown)	0.001262	0.018070	3.722687
8973	(Organic Avocado)	(Organic Grape Tomatoes, Large Lemon)	0.000712	0.010196	3.701181

Figure 21: Items to be recommended to organic avocado buyers

Similar relations can be observed with lemon, organic lemon, and limes.

The scatter plot in Figure 22 represents the relationship between support, confidence, and lift in the dataset. The plots give an overview of the distribution of support, confidence, and lift in the rule set. The support values of association analysis, in general, are lower, and confidence values are well-distributed. There is a positive correlation between lift and confidence values.

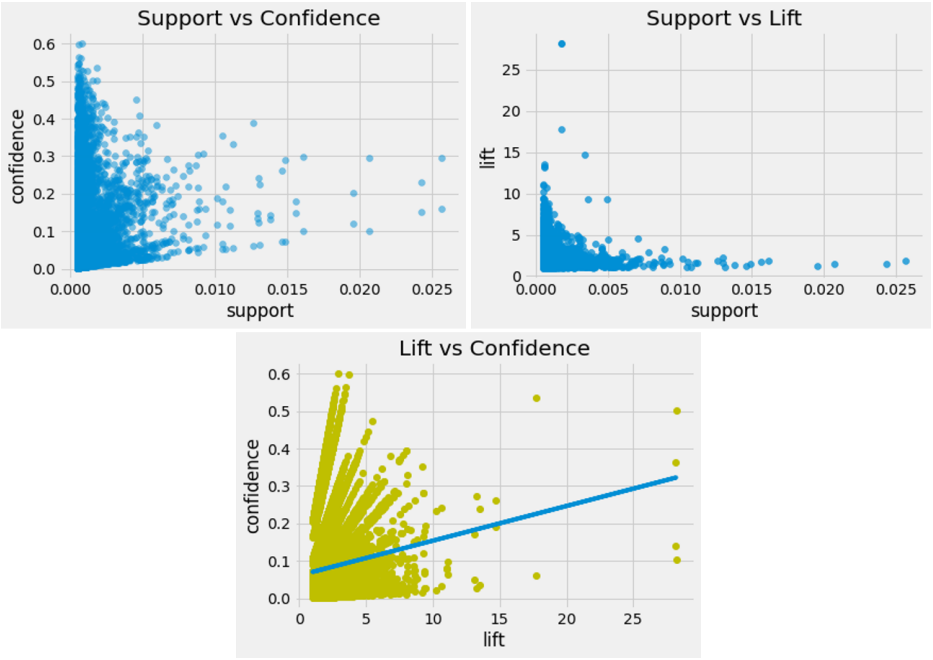

Figure 22: Scatter Plot of the relationship between support, confidence, and lift

Banana is the most popular product at Instacart. Since it has a high score of support, it has as many as 353 rules with 0.001 minimum support. Figures 23 and 24 display banana's relations with its top 10 and top 20 consequent item sets, respectively.

	antecedents	consequents	support	confidence	lift
6585	(Banana)	(Bartlett Pears, Organic Fuji Apple)	0.000816	0.003984	2.937438
6580	(Banana)	(Organic Avocado, Bartlett Pears)	0.000759	0.003706	2.739002
6955	(Banana)	(Organic Fuji Apple, Granny Smith Apples)	0.000553	0.002700	2.663623
6598	(Banana)	(Organic Whole Milk, Bartlett Pears)	0.000530	0.002589	2.616742
8074	(Banana)	(Strawberries, Organic Fuji Apple)	0.001205	0.005881	2.579770
6573	(Banana)	(Bartlett Pears, Large Lemon)	0.000582	0.002841	2.550644
6699	(Banana)	(Organic Fuji Apple, Broccoli Crown)	0.000508	0.002482	2.504499
11777	(Banana)	(Organic Avocado, Cucumber Kirby, Organic Baby...	0.000536	0.002619	2.497487
7015	(Banana)	(Half & Half, Organic Fuji Apple)	0.000655	0.003196	2.492984
6718	(Banana)	(Strawberries, Broccoli Crown)	0.000556	0.002715	2.481618

F

Figure 23: Top 10 Banana Rules

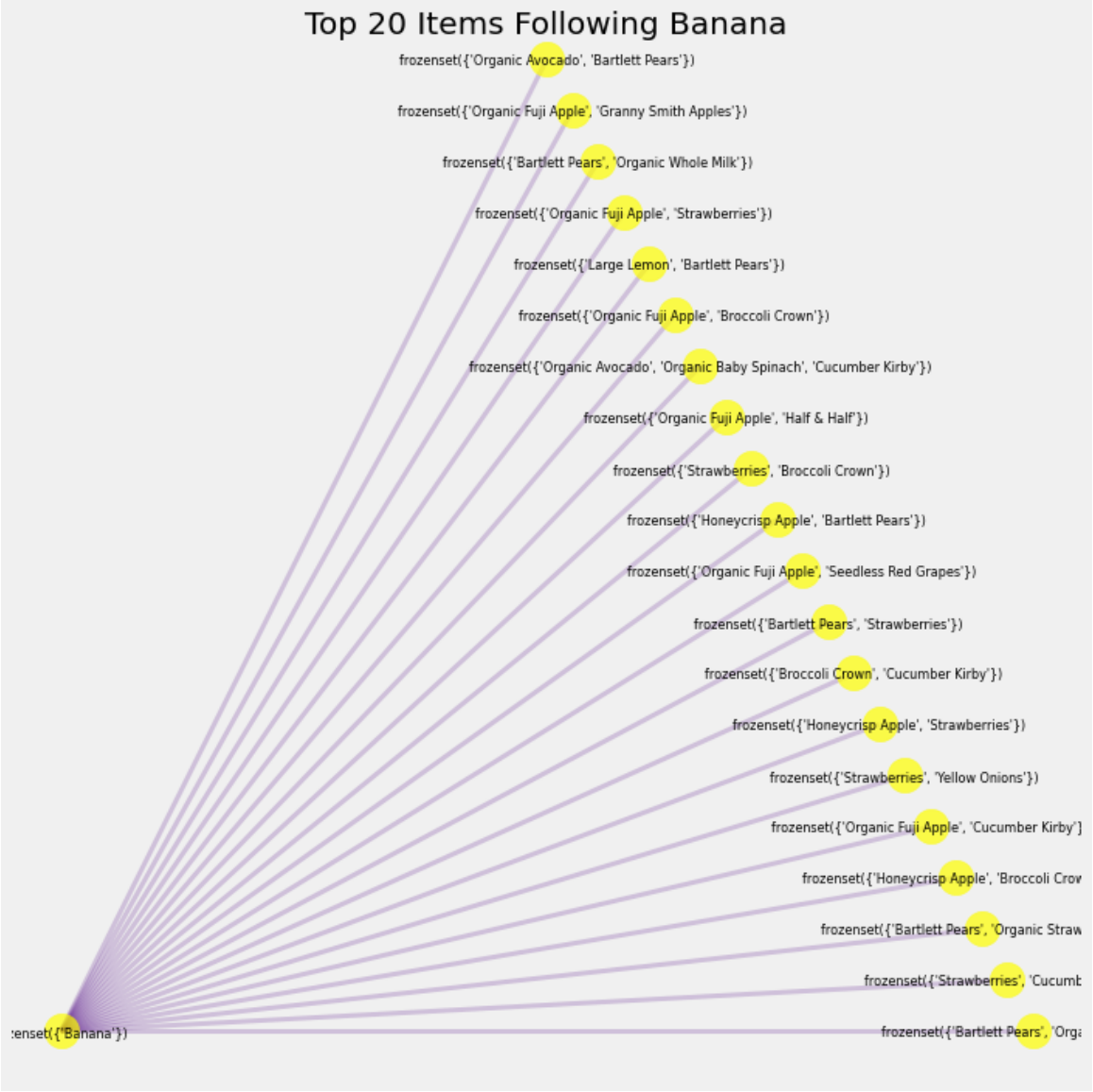

Figure 24: Top 20 item sets following 'Banana'

5.2. Results of XGBoost Algorithm

The pipeline employed for the prediction of the next item to be reordered has normalization, random search, and XGBoost classifier steps. 3-fold cross-validation is applied. Table 1 displays the best values for all parameters after hyperparameter tuning.

XGB Classifier Best Parameters	
Parameter	**Value**
Booster	gbtree
Objective	binary: logistic
Verbosity	1
Gama	0
Learning Rate	0.1
N Estimators	1000
Max Depth	10
Regularization: alpha	0
Regularization: lambda	0.5

Table 1: Best Parameters

The classification report demonstrating model's performance is shown in Table 1. The model returns a mean cross-validation ROC AUC accuracy of 93% and an F1-Score of 0.86. Given that the F1-Score captures the trade-off between precision and recall, the score evaluates the model's ability to be both accurate and generalizable.

XGB Classifier:				
	Precision	Recall	F1-Score	Support
0	0.91	0.75	0.82	2,795,780
1	0.84	0.95	0.89	4,016,078
Accuracy			0.86	6,811,858
Macro Avg.	0.87	0.85	0.85	6,811,858
Weighted Avg.	0.87	0.86	0.86	6,811,858

Table 2: Classification Report

Figure 25 plots the confusion matrix for the model. The model obtains a high precision score of 0.91 by returning only 217,033 false negatives. The model stands out in accurately predicting 95% of reordered products, which is very helpful for the marketing department of Instacart when recommending products to their customers.

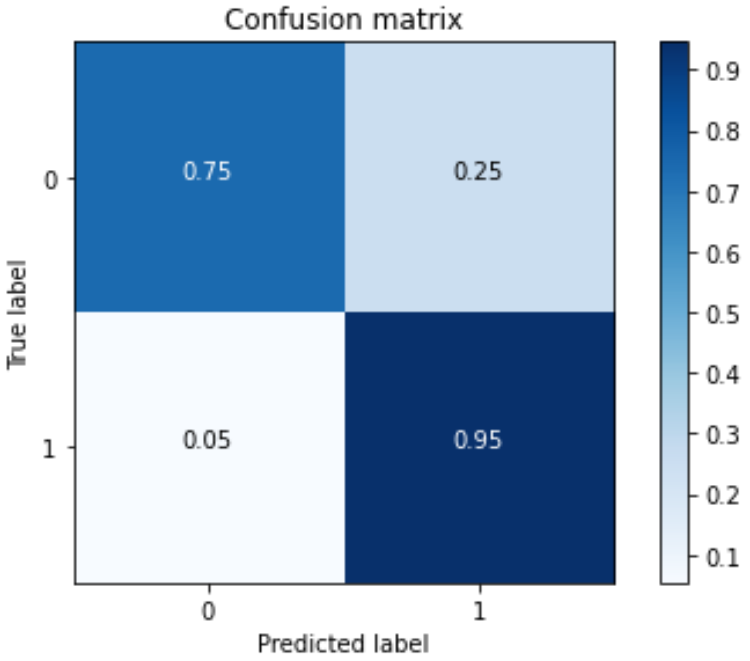

Figure 25: Confusion Matrix of XGBoost Classifier

5.3. Results of Recommender Systems

Before implementing the two recommender systems, some basic recommendations are extracted from the data with regard to customer purchases and to the departments from which customers buy. As clearly seen in Table 3, regardless of departments, most popular products and top reordered products turn out to be congruent.

	Department	List of Products
Top 5 Most Popular Products	All	'Banana', 'Bag of Organic Bananas', 'Organic Strawberries', 'Organic Baby Spinach', 'Organic Hass Avocado'
	Beverages	Sparkling Water Grapefruit', 'Spring Water', 'Lime Sparkling Water', 'Sparkling Natural Mineral Water', '100% Raw Coconut Water'
Top 5 Most Reordered Products	All	Banana', 'Bag of Organic Bananas', 'Organic Strawberries', 'Organic Baby Spinach', 'Organic Hass Avocado'
	Beverages	Sparkling Water Grapefruit', 'Spring Water', 'Lime Sparkling Water', 'Sparkling Natural Mineral Water', '100% Raw Coconut Water'

Table 3: Basic Recommendations of Top 5 Products

5.3.1. Results of Recommender System with Cosine Similarity

CF model with cosine similarity is employed using the “sklearn” library. Recommended products based on similar profiles with similar purchase histories as the target customer are extracted. Ten products to be recommended to the customer with user ID 1 are shown in Figure 26.

```
Recommended Products for User :  1

                        product_name      value
2926                            Soda  17.515693
202            Bag of Organic Bananas  17.087347
2423              Original Beef Jerky   8.432863
2306            Organic String Cheese   6.263991
2599                        Pistachios   6.000000
3470               Zero Calorie Cola    5.573143
618                       Clementines   5.053264
2053              Organic Half & Half   4.633295
902   Extra Fancy Unsalted Mixed Nuts   3.911176
2055              Organic Hass Avocado  3.646706
```

Figure 26: Top 10 Products Recommended to User_1

Products that the customer with user ID 1 has actually bought are: 'Soda', 'Organic String Cheese', '0% Greek Strained Yogurt', 'XL Pick-A-Size Paper Towel Rolls', 'Milk Chocolate Almonds', 'Pistachios', 'Cinnamon Toast Crunch', 'Aged White Cheddar Popcorn', 'Organic Whole Milk', 'Organic Half & Half', 'Zero Calorie Cola'. 5 of the 10 recommended products are actually bought by the customer.

As seen in the example above, by comparing the recommended product list with the actual items in a customer's basket, the model is evaluated. The mean score of the cosine similarity model is calculated as 12%, which means that from every 100 products recommended to a customer, 12 of them would actually be bought. This is a really good ratio concerning that Instacart has a huge number of products in its portfolio.

5.3.2. Results of Bundle Recommender with Bigrams

The model extracts and trains bigrams in order to recommend products to customers. For this, product name and number of recommendations are taken as inputs. For example, five recommendations to a banana buyer are listed as Organic Avocado, Organic Fuji Apple, Honeycrisp Apple, Organic Baby Spinach, and Organic Strawberries.

The model is evaluated by running on the test set. Evaluation starts by recommending what can be bought together with the first product, and ten recommendations are listed. The next nine actually bought products are compared with these recommendations. If there is a match, 1 point is added to the total score. Then the model moves to the second actually bought the product and repeats the process. After all ten products are processed, the total scores are computed and divided by the order size to get the final score. The mean of all final scores in the list gives the model score.

The mean score of the model is 0.205, which points out that the bundle recommender shows superior performance over the CF model with cosine similarity.

6. CONCLUSION AND FUTURE WORK

For the purposes of creating business value through analyzing and predicting shopper behavior, this study proposes a three-step framework model to Instacart, a market basket analysis, a machine learning model to predict if an item will be reordered or not, and a comparison of two product recommender systems. The first two steps introduce basket recommendations, which are particularly useful where a customer makes regular, and recurring purchases, whereas the third step introduces product recommendations, which can have further scope such as helping customers to discover new products.

Based on the association rules generated by means of the apriori algorithm, and the model to predict the reorder of products, frequent patterns from the transaction database are revealed. These patterns and association rules are classic MBA tools for retailers to comprehend the purchasing behavior of their customers and provide valuable business insights. However, they have the drawback of the tendency to produce a high number of rules. Therefore, it would be helpful to compare the returns of promotional campaigns run by trying out different thresholds of minimum support, confidence, and lift values. Since companies offering a huge number of stock-keeping units like Instacart could be overwhelmed by excessive association rules, a product clustering model can be implemented prior to the MBA.

Another future research to address this problem could be employing the minimum spanning tree (MST) approach, which is a simplified representation of an association network limiting the association rule search space and enabling greater control over spurious relations and noise [7].

As for the predictive model, the achieved accuracy score of 86% can be improved by systematically tuning the models. With higher computational power, grid search can be employed and compared with the results of the best model parameters obtained by the random search, which is implemented in this project.

Moreover, prediction performance and can be further improved by building a deep learning model using recurrent neural networks (RNN), which is a type of neural network architecture that allows previous outputs to be used as inputs while having hidden states. In other words, each layer learns from its predecessor, make decisions accordingly, and carry information to the next layer [24].

Since the data consists of transactions reflecting past user behavior, the focus of this project is to identify user-item relationships in order to generate recommendations. One of the proposed methods of product recommender systems is a commonly used method, a collaborative filtering model using cosine similarity, and the alternative method is a product bundle recommender using bigram frequencies. The experimental results based on Instacart data demonstrated a performance advantage of the alternative method over the conventional CF model.

Recommender systems work best in the presence of both implicit and explicit data. The methods in this project are chosen due to the lack of explicit data. The downside of these methods is their inability to match products and users new to the system. A content-based approach that takes explicit data to create profiles for each customer and product should be joined to the existing methods to form a hybrid approach to address this problem. Thus, for further research, collecting explicit data through questionnaires and/or social media profiles of users that would help to profile customers and items would be necessary to build a hybrid filtering method. The obtained profiles allow the recommender system to associate customers with matching products, as well as give the marketers the opportunity to facilitate a sound customer segmentation.

APPENDIX

The codes are stored in .ipynb notebooks and accessible through the links below.

1. Data Preparation Codes: https://drive.google.com/file/d/1qKMXMFZLn-uIUGeOSao86klU1XUoT24f/view?usp=sharing
2. Explanatory Data Analysis Codes: https://drive.google.com/file/d/1tGLeq1q1QeqRHIPbcDQtYP4F-zCpI-jr/view?usp=sharing
3. Market Basket Analysis Codes: https://drive.google.com/file/d/1VlAQmk2MlaPcCSXvSkUk95KYccwrY0ox/view?usp=sharing
4. Prediction of Next Item (XGBoost Algorithm) Codes: https://colab.research.google.com/drive/1rCEFfZH6eEvk2q0lZWKmSIn9Ps4lF6e0?usp=sharing
5. Recommender Systems Codes: https://colab.research.google.com/drive/1j4kxSGWkCUo1eVS-0WdvVWpJFVHCnyKX?usp=sharing

REFERENCES

[1] Keyes, D. (2020). The Online Grocery Report: The coronavirus pandemic is thrusting online grocery into the spotlight in the U.S. — here are the players that will emerge at the top of the market. *Business Insider Intelligence.* Retrieved from https://www.businessinsider.com/online-grocery-report

[2] Supermarkets & Grocery Stores Industry in the U.S. - Market Research Report. (2020). *IBISWorld*. Retrieved from https://www.ibisworld.com/united-states/market-research-reports/supermarkets-grocery-stores-industry/

[3] Nielsen, & FMI. (2017). The digitally engaged food shopper. Retrieved from https://www.fmi.org/forms/store/ProductFormPublic/the-digitally-engaged-foodshopper

[4] Instacart. (n.d.). *Wikipedia*. Retrieved from https://en.wikipedia.org/wiki/Instacart

[5] Setiabudi, D. H., Budhi, G. S., Purnama, I. W. J., & Noertjahyana, A. (2011, August). Data mining market basket analysis' using hybrid-dimension association rules, case study in Minimarket X. *In 2011 International Conference on Uncertainty Reasoning and Knowledge Engineering* (Vol. 1, pp. 196-199). IEEE.

[6] van Maasakkers, F.J.L. (2019, April 30). *Predicting Purchase Decisions Using Autoencoders in a High-Dimensional Setting*. (Master's thesis, Erasmus University Rotterdam).

[7] Valle, M. A., Ruz, G. A., & Morrás, R. (2018). Market basket analysis: Complementing association rules with minimum spanning trees. *Expert Systems with Applications*, (pp. 97, 146-162).

[8] Isinkaye, F. O., Folajimi, Y. O., & Ojokoh, B. A. (2015). Recommendation systems: Principles, methods, and evaluation. *Egyptian Informatics Journal*, (pp. 16(3), 261-273).

[9] Hu, Y., Koren, Y., & Volinsky, C. (2008, December). Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE International Conference on Data Mining* (pp. 263-272). IEEE.

[10] Li, M., Dias, B. M., Jarman, I., El-Deredy, W., & Lisboa, P. J. (2009, June). Grocery shopping recommendations based on basket-sensitive random walk. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1215-1224).

[11] Power, D. J. (2017). *What is the" true story" about data mining, beer, and diapers?* DSS News.

[12] Stanley, J. (2017, May). *Instacart Press Release*. Retrieved from https://tech.instacart.com/3-million-instacart-orders-open-sourced-d40d29ead6f2

[13] The Instacart Online Grocery Shopping Dataset. (2017). *Instacart* [Data file]. Retrieved from https://www.instacart.com/datasets/grocery-shopping-2017

[14] Stanley, J. (n.d.). The Instacart Online Grocery Shopping Dataset 2017 Data Descriptions. *GitHub*. Retrieved from https://gist.github.com/jeremystan/c3b39d947d9b88b3ccff3147dbcf6c6b

[15] Charissa, R. (2018). Optimal Coupon Targeting for Grocery Items: an Instacart Case Study. *Towards Data Science*. Retrieved from https://towardsdatascience.com/optimal-coupon-targeting-for-grocery-items-an-instacart-case-study-128e8d169c7c

[16] Reig Grau, G. (2017). *Market basket analysis in retail*. (Master's thesis, Universitat Politècnica de Catalunya).

[17] Prasad P. & Malik L. (2011). Using association rule mining for extracting product sales patterns in retail store transactions. *International Journal on Computer Science and Engineering* (IJCSE).

[18] Sharma, N., & Verma, C. K. (2014). Association rule mining: An overview. *International Journal of Computer Science and Communication*. (pp. 5, 10-15).

[19] Tan, P. N., Steinbach, M., & Kumar, V. (2013). Data mining cluster analysis: basic concepts and algorithms. *Introduction to data mining*, (pp. 362-367).

[20] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

[21] Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001, April). Item-based collaborative filtering recommendation algorithms. *In Proceedings of the 10th international conference on World Wide Web* (pp. 285-295).

[22] Li, M., Dias, B. M., Jarman, I., El-Deredy, W., & Lisboa, P. J. (2009, June). Grocery shopping recommendations based on basket-sensitive random walk. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1215-1224).

[23] Collins, M. (1996). A new statistical parser based on bigram lexical dependencies. *arXiv preprint cmp-lg/9605012*.

[24] Ko, Y. J., Maystre, L., & Grossglauser, M. (2016, November). Collaborative recurrent neural networks for dynamic recommender systems. In *Asian Conference on Machine Learning* (pp. 366-381).