ECE/CIS 568 DATA MINING, Fall 2022

Final Project Report

Analysis of Customer Purchase Behavior using Market Basket Analysis

Team Members: Ankita Namdeo (UMID:14356614)
Vandana Dattatray Pathare (UMID: 13618379)

Department Name: Computer and Information Science

Responsibilities:

Milestones	Responsibility
Data Collection	Ankita
Data Analysis	Vandana
Technology Survey	Ankita
Data Modelling	Ankita
Evaluation Parameter/Visualization	Ankita, Vandana
Presentation	Ankita
Final Report	Ankita, Vandana

1. Introduction

Today online retailers, physical retailers, and a supermarket need to analyze their customer behavior. By analyzing the customer's behavior, the customer purchasing pattern gets to know which product is in high demand. This helps the retailer to increase their sales by either discounting on products occurring together or by arranging the shelves according to the frequently ordered products. One of the important techniques used to achieve this is the market basket analysis which helps identify the relationship between products based on their occurrences in every transaction and also predicts the next item.

In our project, the aim is to find the frequent items and generate association rules using the apriori algorithm and to predict which products that were previously purchased will be present in the customer's next order. In this project, we must use data mining techniques like market basket analysis, and association rules generation using the apriori algorithm. For prediction, we have used logistic regression where model evaluation parameters are accuracy, precision, recall, f-1 score, and roc-AUC.

1.1 Dataset: This anonymized dataset has a sample of 3 million grocery orders from more than 200k Instacart users. These are the 5 tables order, product priority, product, department Ailse, and order product train data. For this project, we are using an online resource dataset. The size of the Dataset is consisting of 1,384,617

Dataset Link: https://www.kaggle.com/competitions/instacart-market-basket-analysis/data

Field	Description
Reordered = 1	Product reordered
Reordered = 0	Non-re-ordered product

In this report, we will be looking at the steps that are followed to get the predictions, and frequent items using the market basket analysis technique. We will look at the following stages namely, data pre-processing, exploratory data analysis, implementation of apriori algorithm, mining of association rules, and prediction of the user's next purchase.

2. METHODS USED IN THE PROJECT

2.1 Market Basket Analysis – Look for groups of items that commonly appear together in transactions.

2.2 **Apriori Algorithm-** The Apriori Algorithm finds the things in a set that appear the most frequently and then expands those items to a bigger set. The join and prune phases of the Apriori Algorithm are repeated until the most frequent itemset is discovered.

2.3 **Association Rules-** Finding rules and figuring out which things are linked to a group of items are the objectives of association rule mining.

2.4 **Logistic Regression-** It is a classification technique that aids in the prediction of binary results from a group of independent factors.

3. IMPLEMENTATION

3.1 **Setup and Installation: Software:** Python 3.7, Tableau Desktop
Hardware: Microsoft Windows 8/8.1, Windows 10/11,1.5 GB minimum free disk space for Tableau desktop, CPUs must support SSE4.2 and POPCNT instruction sets, CPU: 2 x 64-bit 2.8 GHz 8.00 GT/s CPUs, Memory: 34 GB, Storage: 300 GB.

3.2 **Data Preparation:** In our project, we have done all the data preprocessing in python and stored the final data in a specific file, and then used that specific file in tableau for visualization.

1. **Check the Null Values:** Check the null values and remove the null values from the tables.

2. **Merge the tables:** To make a relationship between the tables apply the join function

3. **Conversion:** Converted the object data type into category type

4. **Applied one-hot encoding:** To convert the categorical variables into binary form

5. **Split the Data:** For training our model we split our dataset into 80% training set and 20% test set.

3.3 **Apriori Algorithm Implementation:** We applied the apriori algorithm to two important features which are product name and order id. For doing this, we selected a basket (15000) sample and got association rules based on lift and support of 0.01. The underlying assumption for the apriori algorithm is that all the subsets of a frequent itemset must be frequent. The first step is to calculate the support of each itemset. Then we decide on a support threshold (which is 0.01 in our case) and eliminate all the itemsets that are below the support threshold. We repeat this procedure of k-item sets. We then create association rules based on the lift.

3.4 **Prediction:** For prediction, we took 100000 data rows. The columns taken are 'add to cart order', ''order number', order day of the week', order hours of the day, and 'reordered'. We replaced the null values in the row with zero. The target variable is 'reorder' which is binary type. Used the train and test dataset in logistic regression. We also used Decision Tree Classifier, random forest, and KNN for prediction to get the comparatively best prediction.

3.5 **Model Evaluation:** To evaluate the model we have taken 5 evaluation parameters. Below are the:

3.5.1 **Accuracy:** Out of all the data points, it provides the number of data points that were accurately anticipated.

3.5.2 **Precision:** It facilitates our understanding of the model's dependability in judging the model to be positive.

3.5.3 **Recall:** It gauges how well the model can identify positive samples.

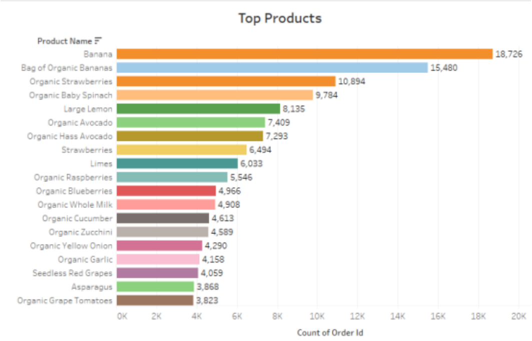
3.5.4 **F1-Score:** Combines precision and recall, two measurements that would normally compete, to provide a summary of a model's prediction performance.

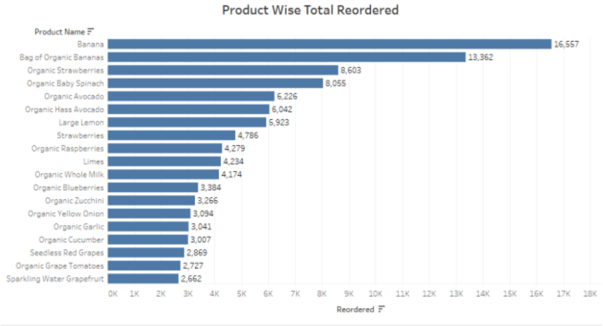
3.5.5 AUC-ROC: The AUC-ROC statistic aids in determining and informing us about a model's capacity for classifying data.

3.4 Result:

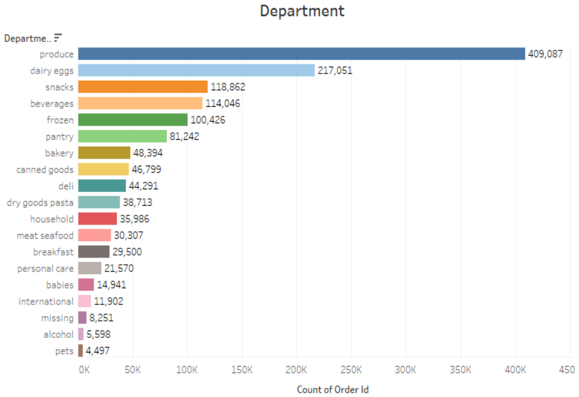
3.4.1 Tableau Output: After preprocessing the data, we stored it in an output file and used that output file in tableau for visualization. Below are a few insights from the data:

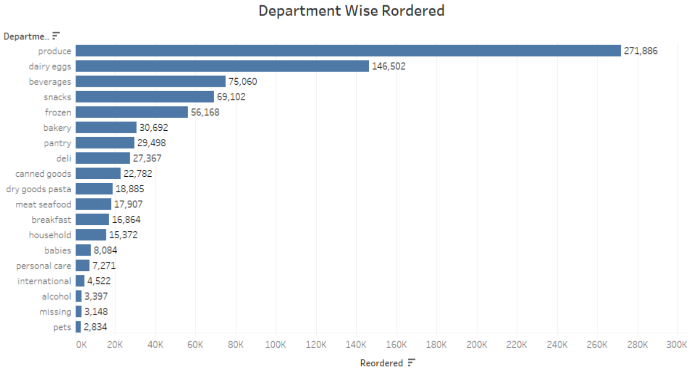
(a) Product: From the below 2 bar graph we can see which product was ordered most based on users and reordered by customers.



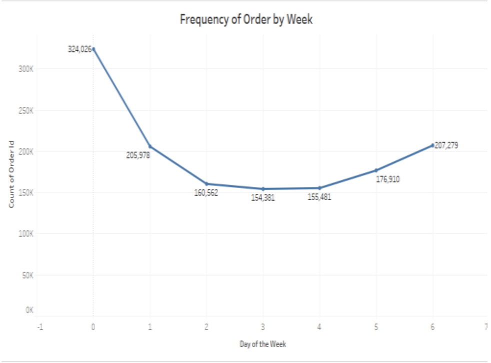


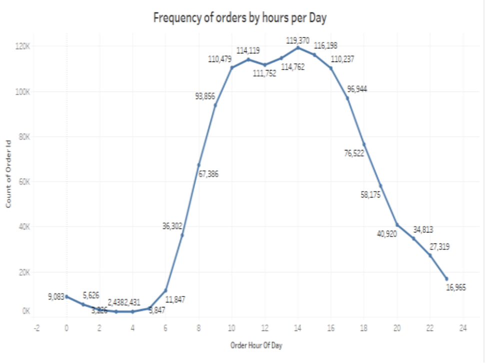
(b) Department: Below 2 bar graph gives us detailed information about the department which departments have the highest order and what is the count of reordering by customers from that department.



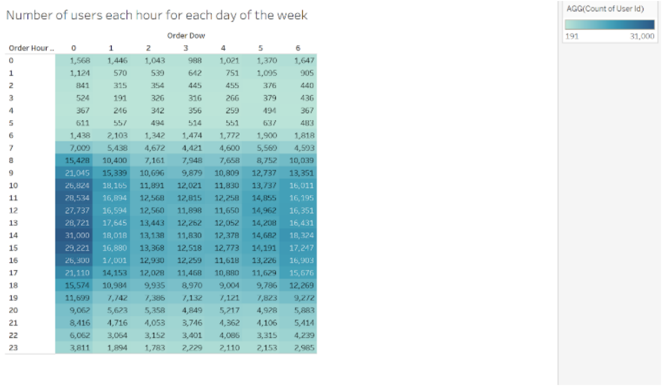


(c) Order Frequency: From the below two trend charts we can see on Sunday and Sat order count was higher as compared to other days whereas from the second chart we can see from 10 AM to 6 PM order count is high as compared to other hours.





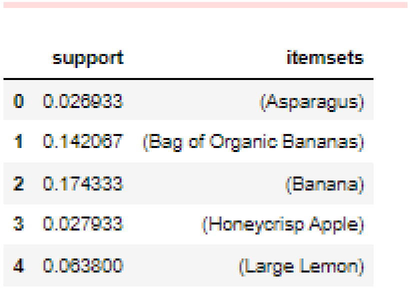
(d) Heat Map: Below is the heat map showing the number of user purchases broken down by each day of the week with each hour of the day.



Number of users each hour for each day of the week

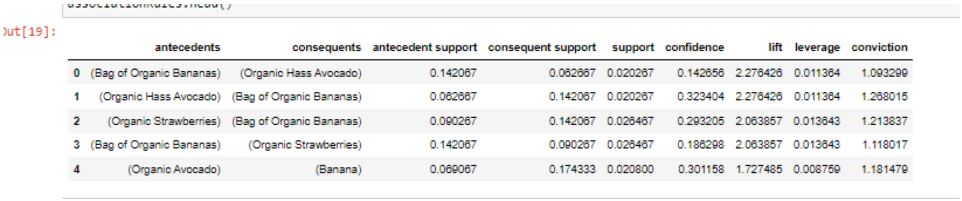
	Order Dow						
Order Hour ..	0	1	2	3	4	5	6
0	1,568	1,446	1,043	988	1,021	1,370	1,647
1	1,124	570	539	642	751	1,095	905
2	841	315	354	445	455	376	440
3	524	191	326	316	266	379	436
4	367	246	342	356	259	494	367
5	611	557	494	514	551	637	483
6	1,438	2,103	1,342	1,474	1,772	1,900	1,818
7	7,009	5,438	4,672	4,421	4,600	5,569	4,593
8	15,428	10,400	7,161	7,948	7,658	8,752	10,039
9	21,045	15,339	10,696	9,879	10,809	12,737	13,351
10	26,824	18,165	11,891	12,021	11,830	13,737	16,011
11	28,534	16,894	12,568	12,815	12,258	14,855	16,195
12	27,737	16,594	12,560	11,898	11,650	14,962	16,351
13	28,721	17,645	13,443	12,262	12,052	14,208	16,431
14	31,000	18,018	13,138	11,830	12,378	14,682	18,324
15	29,221	16,880	13,368	12,518	12,773	14,191	17,247
16	26,300	17,001	12,930	12,259	11,618	13,226	16,903
17	21,110	14,153	12,028	11,468	10,880	11,629	15,676
18	15,574	10,984	9,935	8,970	9,004	9,786	12,269
19	11,699	7,742	7,386	7,132	7,121	7,823	9,272
20	9,062	5,623	5,358	4,849	5,217	4,928	5,883
21	8,416	4,716	4,053	3,746	4,362	4,106	5,414
22	6,062	3,064	3,152	3,401	4,086	3,315	4,239
23	3,811	1,894	1,783	2,229	2,110	2,153	2,985

3.4.2 Market Basket Analysis Result: From the below output we can see frequent items and their support value.



	support	itemsets
0	0.026933	(Asparagus)
1	0.142067	(Bag of Organic Bananas)
2	0.174333	(Banana)
3	0.027933	(Honeycrisp Apple)
4	0.063800	(Large Lemon)

3.4.3 Association Rules Output

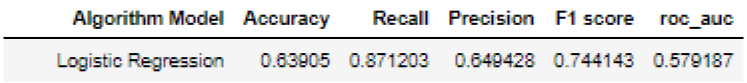


```
associationRules.head()
```

)ut[19]:

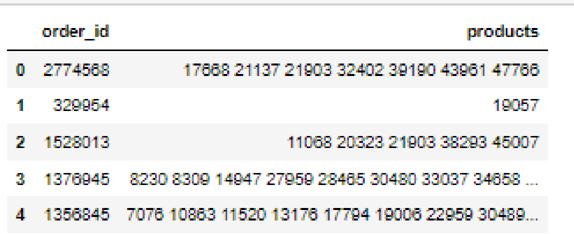
	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(Bag of Organic Bananas)	(Organic Hass Avocado)	0.142067	0.062667	0.020267	0.142656	2.276426	0.011364	1.093299
1	(Organic Hass Avocado)	(Bag of Organic Bananas)	0.062667	0.142067	0.020267	0.323404	2.276426	0.011364	1.268015
2	(Organic Strawberries)	(Bag of Organic Bananas)	0.090267	0.142067	0.026467	0.293205	2.063857	0.013643	1.213837
3	(Bag of Organic Bananas)	(Organic Strawberries)	0.142067	0.090267	0.026467	0.186298	2.063857	0.013643	1.118017
4	(Organic Avocado)	(Banana)	0.069067	0.174333	0.020800	0.301158	1.727485	0.008759	1.181479

3.4.4 Prediction Result: From logistic regression, we get an accuracy of about 0.63 %.



	Algorithm Model	Accuracy	Recall	Precision	F1 score	roc_auc
	Logistic Regression	0.63905	0.871203	0.649428	0.744143	0.579187

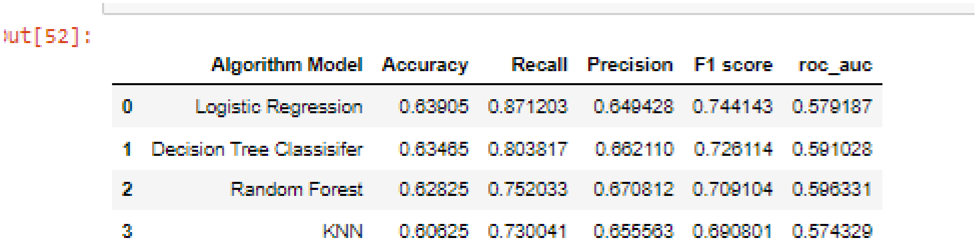
3.4.5 Next Items In Customer's Bucket: In the below output we can see each order ID have a product ID against them so might be from those products that the user will buy 1-2 product.



	order_id	products
0	2774568	17668 21137 21903 32402 39190 43961 47766
1	329954	19057
2	1528013	11068 20323 21903 38293 45007
3	1376945	8230 8309 14947 27959 28465 30480 33037 34658 ...
4	1356845	7076 10863 11520 13176 17794 19006 22959 30489...

4. EXPERIMENTS

We try to improve the accuracy, so we implemented other algorithms like decision tree random forest KNN and observed that from random forest algorithms we are getting higher as compared to others so we can say that for this dataset random forest is the best algorithm. Below is the output Table of all the algorithm values:



ut[52]:

	Algorithm Model	Accuracy	Recall	Precision	F1 score	roc_auc
0	Logistic Regression	0.63905	0.871203	0.649428	0.744143	0.579187
1	Decision Tree Classisifer	0.63465	0.803817	0.662110	0.726114	0.591028
2	Random Forest	0.62825	0.752033	0.670812	0.709104	0.596331
3	KNN	0.60625	0.730041	0.655563	0.690801	0.574329

5. CONCLUSION

We used the user purchase data and manipulated it to apply the market basket analysis on it. This helped us predict the user's next purchase. The problems we faced were in deciding which algorithms suit best for our problem statement. We also faced the challenge of which evaluation metrics to use on our way to get the best prediction. The logistic regression is best suited for the data and problem statement that we had.

We learned all the steps that a data science team typically follows to optimize supermarket sales based on consumer behavior. We also learned what all challenges are faced while carrying out all the correct steps.

Reference

https://select-statistics.co.uk/blog/market-basket-analysis-understanding-customer-behaviour/

https://www.researchgate.net/publication/322161863_Market_Basket_Analysis_to_Identify_Customer_Behaviours_by_Way_of_Transaction_Data/link/5a49f56a458515f6b0590f13/download

https://www.researchgate.net/publication/351168385_Research_and_Case_Analysis_of_Apriori_Algorithm_Based_on_Mining_Frequent_Item-Sets