INDIANA UNIVERSITY BLOOMINGTON

CSCI B 565

DATA MINING

MARKET BASKET ANALYSIS, CUSTOMER SEGMENTATION AND LIFETIME VALUE PREDICTION

Authors:
Akshat Arvind
Aniket Kale

Professor:
Yuzhen Ye

December 13, 2021

Abstract

Market Basket Analysis is a key technique used by companies to uncover associations between items. It works by looking at items which occur frequently in customer's order basket. Our project explores Association Rule Mining on the Instacart Market Basket Analysis Dataset from Kaggle. We generate suggestions based on the products in the customer's order basket using Association Rule Mining.

Further, we also used the data to perform customer segmentation and customer lifetime value estimation using RFM (Recency, Frequency and Monetary) analysis. Use clustering to analyze customer patterns.

Lastly, we discuss ways on how to use suggestions generated from Association Rule Mining to provide targeted offers to different customer segments which can help in increasing lifetime value for customers.

Dataset Information

Tables present in the dataset -

- **Products** –

  Columns - [product_id, product name, aisle_id, department_id]
  Size – 813.67 KB

- **Orders** –

  Columns -[order_id,user_id,eval_set,order_number,order_dow,order_hour_of_day
  days_since_prior_order]
  Size – 32.81 MB

- **Order_products__prior** –

  Columns - [order_id, product_id, add_to_cart_order, reordered]
  Size – 164.69 MB

- **aisles** –

  Columns - [aisle_id, aisle_name]
  Size – 1.91 KB

- **Department** –

  Columns - [department_id, department_name]
  Size – 804B

Introduction

Using the market basket data from Kaggle Instacart Market Basket Analysis competition, we have targeted 4 major objectives in our project –

- Exploratory Data Analysis on data to extract actionable insights.
- To come up with a model to predict customer segmentation and the customer lifetime value using RFM Analysis.
- Use Association Rule Mining to give suggestions of possible products which could be added in a customer's order basket.
- Also, we have tried to complement the prediction results by adding rewards, promos or clubbed offers to specific customer segments using Association Rule Mining to convert this model into a possible profitable business model.

Methods

Objective -1 EDA

Exploratory Data Analysis has been performed on the dataset to extract insights and trends from the data.

Tools Used: -

- Matplotlib
- Seaborn
- Pandas
- Plotly

Objective -2 Association Rule Mining

Association Rule Mining has been done to suggest additional products based on the products present in the customer's order basket.

Tools Used: -

- mlxtend
- fpgrowth
- association rules

Objective – 3 Customer Segmentation using RFM Analysis

RFM Analysis for Customer Segmentation, predicting segments, clustering on segments.

Tools Used: -

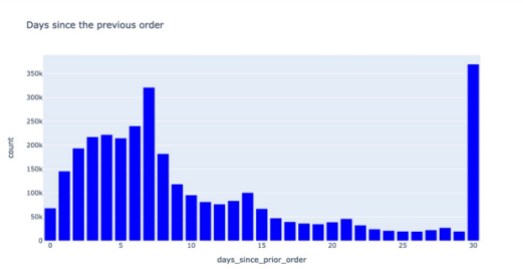
- Logistic Regression
- Random Forest
- XG Boost
- K-means clsutering

‘

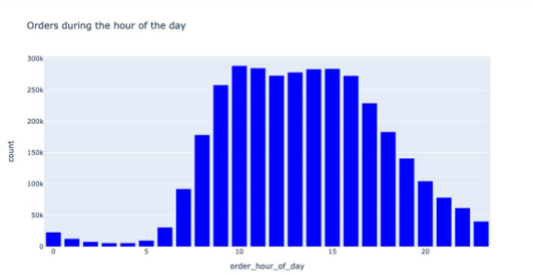
Results

Objective – 1 EDA

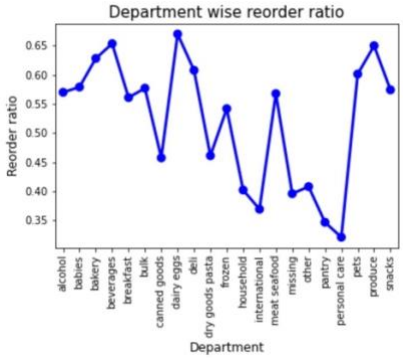
These are some of the interesting insights extracted from the data –



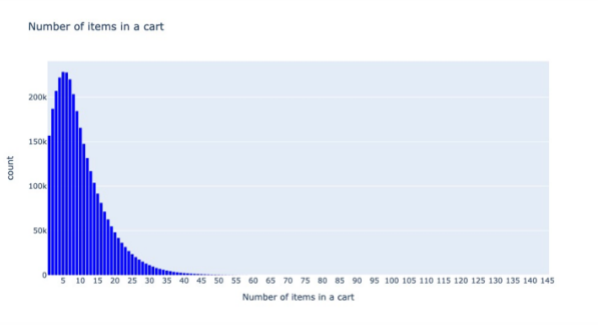
The plot shows the count of orders vs days since prior order for customers. We can observe the spikes on 7th, 14th, 21st and 30th day confirming the weekly and monthly trend of reordering supplies.



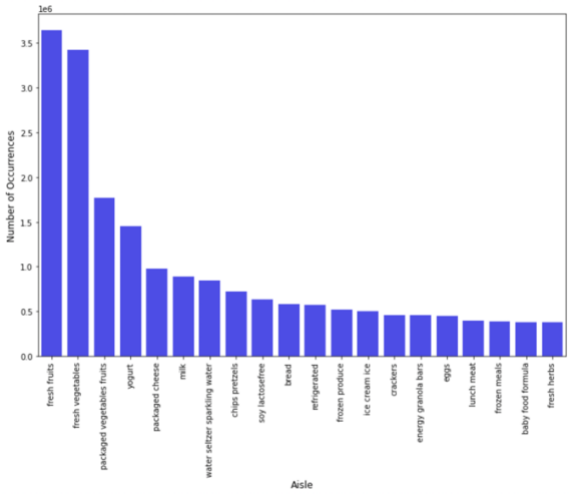
The plot shows the count of orders vs order hour of the day. The maximum orders are placed during the day hours from 9 AM – 5PM.



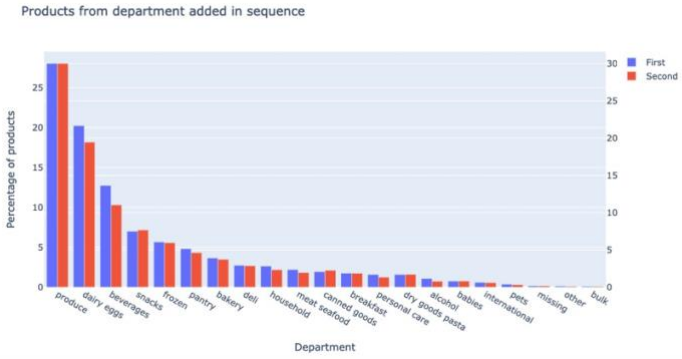
The plot shows the reorder ratio of items from different departments. We can see that item from bakery, supplies and produce have high reorder ratio compared to items like personal care and household which are products that are frequently ordered.



The plot shows the number of items that customers usually keep in their order basket. According to the data, we can see that an average of 5-10 items are added in the basket by customers before placing their orders.



The plot shows the number of occurrences of a product type across all the orders in the dataset. We can see that fresh fruits and vegetables are highest ordered products and baby food and herbs are least ordered.



The plot shows the first and second product that the customers add in their order basket while placing an order. Produce and dairy eggs are added first by majority of the customers.

Objective – 2 Association Rule Mining

Association rule mining technique when used in Market Basket Analysis enables one to find sets of items that are often found together in a customer's basket. This can be used to improve or create bundles of products to improve sales. Also, the created bundles can be used to provide special offers for customers with discounted pricing on bundled products. We have used fpgrowth algorithm from mlxtend module to extract frequent itemsets and association rules module to generate the rules.

Using the data from the Order_products__prior table, we create a truth table of order_id and products.

```
truth_table
```

	1	2	3	4	5	7	8	9	10	11	...	49677	49678	49679	49680	49681	49682	49683	49686	49687	49688
1	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False	True	False	False	False
36	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False	False	False	False	False
38	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False	False	False	False	False
96	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False	False	False	False	False
98	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

After the truth table is generated, it is used to generate frequent itemsets, that are used to generate association rules.

```
frequent_itemsets = fpgrowth(truth_table, min_support=5/len(truth_table), use_colnames=True)
```

```
rules = association_rules(frequent_itemsets, metric="confidence", min_threshold=0.8)
```

```
print("μ number of consequents:", rules['consequents'].apply(len).mean())
rules
```

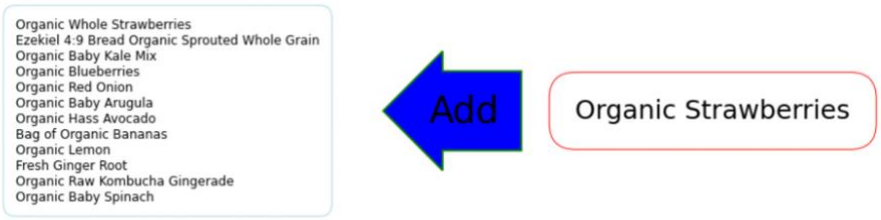
```
μ number of consequents: 1.0391897394136809
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(47626, 49683, 4605, 21903)	(24852)	0.000076	0.142719	0.000069	0.900000	6.306104	0.000058	8.572811
1	(26209, 49683, 28204, 16797)	(24852)	0.000046	0.142719	0.000038	0.833333	5.838985	0.000032	5.143687
2	(49683, 39275, 48679)	(24852)	0.000046	0.142719	0.000038	0.833333	5.838985	0.000032	5.143687
3	(27104, 49683, 24964, 47766)	(24852)	0.000038	0.142719	0.000038	1.000000	7.006782	0.000033	inf
4	(42265, 40706, 49683, 24852)	(21903)	0.000046	0.074568	0.000038	0.833333	11.175474	0.000035	5.552592
...	...	...	...	...	...	...	...	...	...

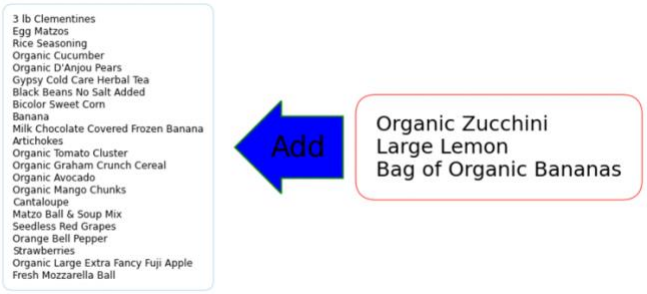
Antecedents – The original order basket for a customer.
Consequents – The suggestions added by the association rules.

Some of the suggestions generated from association rule mining –

Example -1:



Example-2:



Objective – 3 Customer Segmentation using RFM Analysis

1. **Analysis of the consumer-base for a brand with RFM analysis:**

- Knowing who your most valuable customers are, what their potential future spending amounts to, and how they engage with your brand is critical.
- Marketers care about retention, loyalty, and improving customer experience.
- Instead of focusing on the consumer-base as whole, it is helpful to segment customers in separate categories, based on their engagement and revenue that they generate.
- This helps the marketing team to prioritize certain strategies over others, based on the customer behavior patterns of a certain customer segment.
- A structured and easy way to segment customers is to perform RFM analysis. RFM stands for ***Recency, Frequency and Monetary*** analysis.
- RFM analysis quantitively ranks and groups customers based on the Recency, Frequency and Monetary factors:
    - **Recency**: How many days, on an average, does a customer takes before placing the next order? In other words, the average of the days between a customer's sequential orders.
    - **Frequency**: How many orders does a customer place in a given period of time? In other words, the number of transactions placed by a customer.
    - Monetary: How much money has the customer spent during the life span they were in business with the brand. This factor allows us to rank customers based on their spending patterns.
- These factors are ranked and scaled (usually form 1-5) according to the data and merged to get the customer segments (i.e., from 1-125). This score can be used as the Customer Lifetime Value (CLV) Score. But we'll discuss more about CLV in the next section.

- We can again divide this distribution into equal parts to define our customer segments.
- Further analysis is performed on classes 1-5, (5 being the highest) to determine the strategies as mentioned above. But now we have an idea about what kind of a strategy is applicable to each segment. Consider segment #5, here we have the highest paying user, with high frequency of orders. So, we might want to use better loyalty programs to improve their experience. But, say for Segment #3, there could be lacking frequency, or recency or lacking revenue generated from them. It would be interesting to further analyze the area that is dragging the score of customers in this segment and apply strategies like added discounts, more rewards on purchases and more.

2. **Use case: *Using Machine Learning Algorithms to predict the Customer Segments generated from RFM analysis***
    - To automate this process of RFM analysis, we can use Machine learning algorithms to predict the customer segments based on the raw features.
    - This would help collating all the sales data together and performing RFM analysis every time we want to check which segment the customer or a set of customer belongs to.

Modelling Pipeline:

a. **Data Pre-processing and imputations:**
Initially, the ***orders*** data, ***department***, and ***aisle*** information, as well as the orders and ***products*** were in separate files. These files were merged to include Customer ID, Order No, Product, Product Category in the same table.
Then this merged file was used to compute the average revenue generated for every customer. This was used as the **Monetary** factor in the RFM analysis.
The total number of orders for each customer were computed. This column was used as the **Frequency** factor.
In the ***order*** data, there was also information about days it took to place a new order from previous order, associated with the Order No. The values for "previous order days" were summed for each customer. This column was used as the **Recency** factor.

b. **Computing RFM scores:**
After computing all the 3 factors: **Recency, Frequency and Monetary,** these factors were scaled from 1-5 (using quantile-cuts of 5). Multiplied and scaled again from 1-5. The following diagram depicts this procedure:

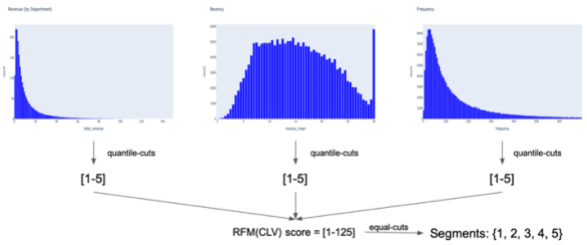


Figure 2.1 Computing RFM score.

As we can see the Revenue and Frequency are right-skewed, so we are quantile cutting all the values to ensure equal distribution of rows. After multiplying the factors, the cuts are made equally so that the Customers are segregated according to their scores, and this is how we have Customer Segments from 1-5.

c. Balancing the imbalanced classes:

Hence, this was posed as a classification problem. However, there was an imbalance in the classes. The classes 4 and 5 had the least amount users in them (almost 1% and 2.5% each), as opposed class 1 being the majority class (60%), and classes 2 and 3 being somewhere in the middle (30% and 7% respectively).

This was obvious, as most of the customer lie in the lower segment, and hence the lower are the most interesting to analyze and improve upon. But instead of changing the distributions we decided to use SMOTE to oversample the minority classes. We tried to strike a balance with the other classes so that the model won't overfit if it sees the same samples.

d. **Modelling**:

The predictors were: ***Frequency, Recency*** and ***Total revenue*** (not scaled according to the quantile-cuts but raw data was used to predict the target variable)
The target variable was: ***Customer Segment***

The following models were experimented with:

Model	**Accuracy**	**F1-score**
Logistic Regressor	33.62%	0.3361
Random Forrest	94.39%	0.9431
XG Boost	96.18%	0.9617

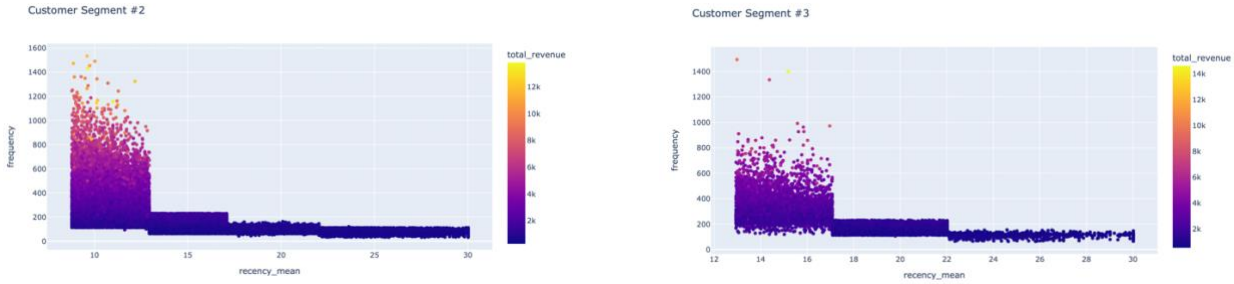
Note: Train-test split ratio was 70:30
The best model here was XG Boost in terms of accuracy, but the precision and recall scores were also high for this model, as compared to the other models [See appendix].

Thus, we have predicted the RFM score/segmentation.
Note: This RFM value can also be called the Customer Lifetime Value (CLV) score. But the important information needed for actual calculation of CLV is the date of the orders which is missing in our dataset. But the predicted score is good enough for our analysis.

Objective – 4 Clustering on Segments, predicting/improving the RFM/CLV score.

After obtaining the Customer Segments, we want to find out what the consumer patterns are within the segments themselves.

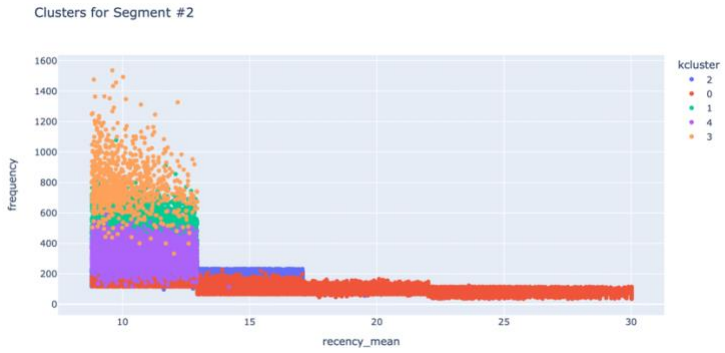
The following are the plots of how the users in Segments #2 and #3 are distributed.



In these plots we can see that there are some interesting clusters forming. These clusters would help us gain information about the customer patterns in the respective segments. These clusters would have varying Recency, Frequency and Monetary values. Using the cluster information, we can make an informed decision about the marketing strategy revolving that particular customer. For example, in Segment #2 we can see users with very high "total revenue", low "recency" with high/medium/low levels of "frequency. There could be different kinds of startegies that can be applied for varying frequency levels like an expiring coupon to increase frequency. And the end goal here is to improve the RFM/CLV score of the customer.

K-Means clustering:
We have performed k-means clustering on the data for Segment #2 but we have also estimated the optimal number of clusters for each of the Segments 1-5. But we are only mentioning the results for Segment #2 since the findings are the most interesting to us.

Visualizing the results of clustering:



Average values of RFM for the clusters:

kcluster	recency_mean mean	frequency mean	total_revenue mean
0	18.660012	93.495571	674.617065
1	10.429604	509.333435	3838.621582
2	12.642400	183.695115	1354.412842
3	10.158270	789.256563	6315.746094
4	10.694851	323.211556	2392.406494

Observations:

- In the above table, we can see that there are different attributes of the three values (frequency, recency, revenue)
- For example, cluster #3 has low average recency, but high frequency and revenue
- another example is cluster #0 which has high recency but low frequency and total revenue.
- We can also see these clusters in the clustering plot above.
- We can further analyze the purchasing behaviors of the customers belonging to these clusters from the same segment. Then come up with targeted strategies for these segments of customers.

Hence, this can be done for all the clusters belonging to each of the segments. And such finding could be used to build a better overall marketing strategy.

Discussion

How are we linking RFM/CLV analysis with Market Basket Analysis?

In other words, predicting Customer Lifetime Value using the finding from Market Basket Analysis:

Earlier we performed Market Basket Analysis to predict which products are associated to which other products. In other words, how a customer buying a particular product increases the likelihood of another customer buying the associated products. This can be used for product recommendation.
But we would like to propose a new way to use the results of Market Basket Analysis. When we trained out model, we used the previous purchase data to find the RFM/CLV score. But if we impute the potential revenue that can be generated from a customer by adding the associated products, it can help up even more with regards to which Customers to focus upon.

For example,
Say Customer #100 (belonging to Segment #2) has bought products P3, P10, P17. And our MBA tells us that P3 has a string association with P4 and P10 with P11, P12 and so on. Now with certain confidence we can say that the same user can buy the associated products {P4, P11, P12, ...}. Now we add the projected revenue (with its confidence) to the RFM model to see if there is an improvement (in comparison other users).

i.e.,
1. Update revenue: New Revenue = (Original Revenue) + (confidence level) * (Projected revenue)
2. Predict using RFM model
3. Compare the results with other users to prioritize and focus our decisions.

This can be used as an added metric to target customers who have more projected revenue. And can be used as a good strategy to predict customer behaviors.

Appendix

- **Random Forest Classifier classification report:**

Classification report:

	precision	recall	f1-score	support
1	1.00	0.99	0.99	37752
2	0.94	1.00	0.97	27377
3	0.84	0.93	0.88	19365
4	1.00	0.51	0.68	7537
5	0.98	1.00	0.99	7551
accuracy			0.94	99582
macro avg	0.95	0.89	0.90	99582
weighted avg	0.95	0.94	0.94	99582

- **XG Boost Classifier classification report:**

Classification report:

	precision	recall	f1-score	support
1	1.00	1.00	1.00	37752
2	1.00	0.99	0.99	27377
3	0.84	1.00	0.91	19365
4	1.00	0.53	0.70	7537
5	1.00	1.00	1.00	7551
accuracy			0.96	99582
macro avg	0.97	0.90	0.92	99582
weighted avg	0.97	0.96	0.96	99582

References

- http://rasbt.github.io/mlxtend/user_guide/frequent_patterns/fpgrowth/
- https://www.qualtrics.com/experience-management/customer/customer-lifetime-value/
- https://knowledge.dataiku.com/latest/kb/industry-solutions/rfm-customer-lifetime-value/rfm-customer-lifetime-value.html
- https://clevertap.com/blog/rfm-analysis/
- http://www.brucehardie.com/papers/rfm_clv_2005-02-16.pdf
- https://www.kaggle.com/jboros/market-basket-analysis-with-association-rules/notebook