Cover Sheet

Changes relative to Draft Report

Professor

- Changes in formatting of the report
- Equations written in latex
- Basic exploratory data analysis before the “Methods” and “Results” section
- Described hyperparameters in order for the reader to reproduce the results
- Solved computational limitation problem in apriori algorithm
- Informative plots on LDA results
- Perplexity used to find the ideal number of topics for LDA

Peer Reviews

- Applied Market-Basket Analysis on multiple clusters in LDA
- Number of users in each cluster mentioned in both K-means and LDA
- Labeling of section and divisions of sections
- Assigned numbers to each figure
- Complex long sentences were simplified in order for the reader to understand easily

Analysis on Instacart Online Grocery Shopping Dataset

Unsupervised Machine Learning
DS 5230

Crains Sudhirkumar Patel
Devarsh Harshad Bhupatkar
Madhunil Anil Pachghare

Khoury College of Computer Sciences
Northeastern University

1. Introduction

When it comes to online shopping, customers need to be provided with a personalized service in order to lure them towards buying more products. Companies often fail to predict their customer's requirements which this project will try to bridge by finding frequent itemsets for different cluster of users by considering the similarity between an individual's buying patterns as well as the similarity between the buying history of other similar users. The objective of our project is to apply unsupervised machine learning algorithms to segment users and mine association rules for that targeted user cluster, which shall increase revenue/sales for Instacart.

The dataset used in this project was published on Kaggle by Instacart under the name of "Instacart Online Grocery Shopping Dataset 2017". The dataset is a relational set of files describing customers' orders over time. It is anonymized and contains a sample of over 3 million grocery orders from more than 200,000 Instacart users.

The problem interests us as we believe this is a real world problem that we see every day where an online shopping site gives different recommendations and offers to different users on the basis of their usage patterns and the approach to the problem is to try a variety of algorithms for clustering and then applying market basket analysis on them in order to get more targeted results for individual users, which from an unsupervised machine learning point of view would be a good way of giving offers to users on the basis of their usage patterns. The exploratory data analysis depicts that 65% users have ordered less than 15 times from Instacart and thus doing a more targeted marketing on those users will be a better marketing strategy. The proposed model first segments the users who have ordered less than 15 times and on this segment of users, Topic Modeling (Latent Dirichlet Allocation (LDA)) is done for clustering of users based on their buying patterns. Finally, Association Rules Mining using Market Basket Analysis is done on each cluster obtained after LDA which gives a set of association rules. These rules can be used by the Instacart company for targeted advertisement and discount offers to increase their revenue/sales.

2. Exploratory Data Analysis

The dataset consists of multiple files which give the data of each anonymized user order. The dataset has between 4 and 100 orders of each user. Each order has multiple products and additionally each product is assigned to one of 21 departments and one of 134 aisles. For each order, the dataset gives information about the day of the week when the order was made, the time of the day and additionally the number of days it has been since the user made a previous purchase.

The entire dataset gives information about approximately 200,000 users who have collectively placed more than 3 million orders from a range of 50,000 products with total 32 million products being delivered by Instacart in these 3 million orders.

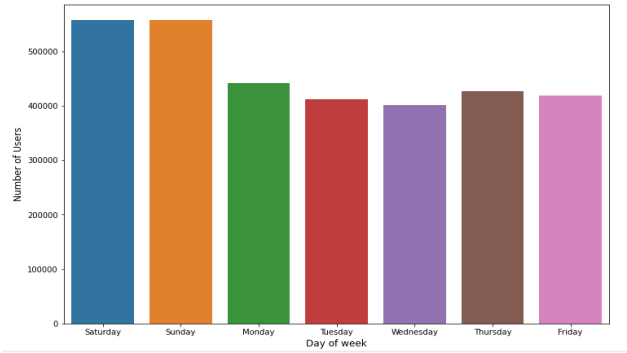

Fig 2.1 Frequency of order by Week Day

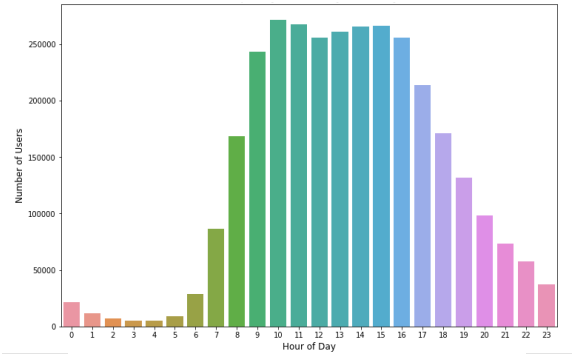

Fig 2.2 Frequency of order by Hour of Day

It can be seen that users usually order from Instacart on weekends and there is no clear preference amongst users when it comes to weekdays. Users usually use Instacart from 9 a.m. to 5 p.m. and the usage drops sharply post 6 p.m.

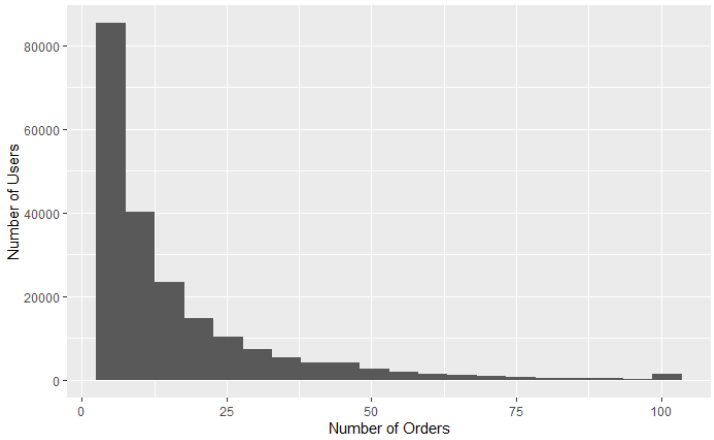

Fig 2.3 User count for Number of orders

The distribution of number of users shows how most users have not used Instacart a lot of times and on further analysis it can be found that 65% users have ordered less than 15 times from Instacart and just converting a few of these users to regular customers could lead to big financial gains for the company.

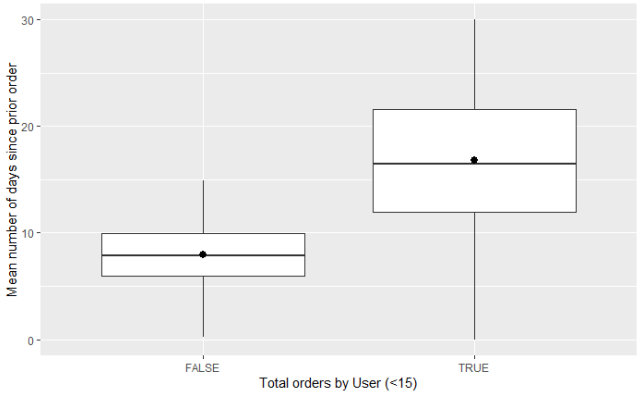

Fig 2.4 Average number of days since prior order for user segments

In order to specifically target users who do not order frequently on Instacart, users who had ordered less than 15 times from Instacart were segmented. These users showed a pattern of buying items at a larger gap than other users in the dataset.

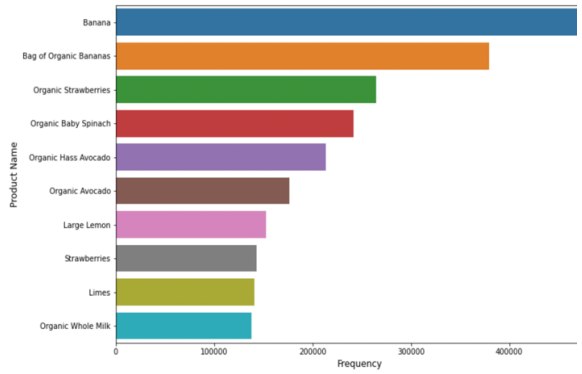

Fig 2.5 Top 10 Frequent products for total users

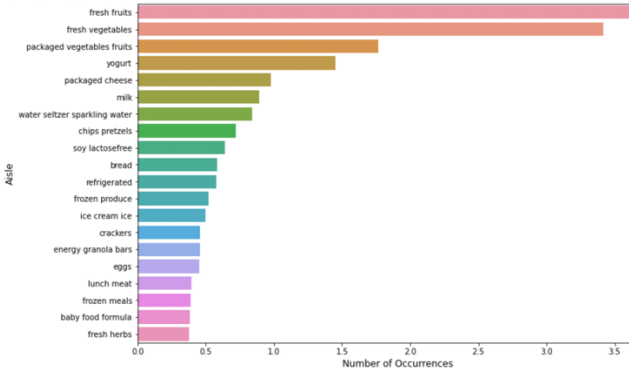

Fig 2.6 Top 20 frequent aisles for total users

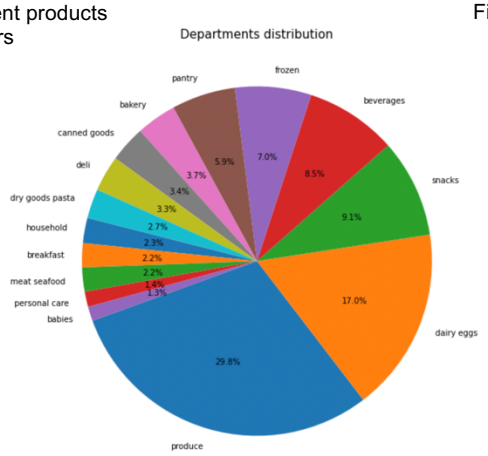

Fig 2.7 Pie chart showing buying percentage by departments

The users seem to prefer buying fresh fruits and vegetables from Instacart as their frequencies are almost twice that of the third most frequent aisle. Department sales reiterate that users buy produce from Instacart and the pie chart gives a representation of the overall buying patterns of the users showing the kind of products that users seem to prefer buying from Instacart.

3. Methods

3.1 K-Means Clustering

K-means clustering is a well-known unsupervised machine learning algorithm. It is used to divide the entire dataset into clusters based on the patterns in the data. The goal of the algorithm is to minimize the intra-cluster difference while maximize the inter-cluster similarity as much as possible. The algorithm groups all the data points in K clusters based on its distance from the centroid of the cluster. The algorithm initiates with a random value of mean and assigns each

point to a cluster. The mean of the cluster is again calculated with the new data points and points are clustered with certainty in one of the clusters. This process is iteratively repeated until the loss function reaches an optimal value.

The loss function which is optimized in this algorithm is defined as below:

$$L(\mu, z) = \sum_{k=1}^{K} \sum_{n=1}^{N} [\![I[z = k](x_n]\!] - \mu_k)$$

Where μ is the mean, *z* is the assignment of the data point, x_n is the data point, K is the number of clusters, N is the number of points.

3.2 Topic Modeling

Topic Modeling is an unsupervised machine learning based technique to represent a document using a few topics which are composed of words from the document. These topics are formed such that they give the combination of words which best explain the document. Here words are grouped into topics such that the group represents a topic. The topic modeling technique used in this paper is Latent Dirichlet Allocation.

The topic modeling does a repetitive task for each document in the corpora. It first assumes that there are k topics across the documents. Each of the document is assigned a topic k with some probability α. The assignment of probability α is done on the basis of the words present in the document. Each word is probabilistically assigned to a topic based on what topics are present in the document and how many times that word has been assigned to a particular topic for all the documents.

The form of topic modelling where Dirichlet priors are used is known as LDA (Latent Dirichlet allocation). These priors are input to the model. The below model is then used to obtain optimized values for topic proportions and topics.

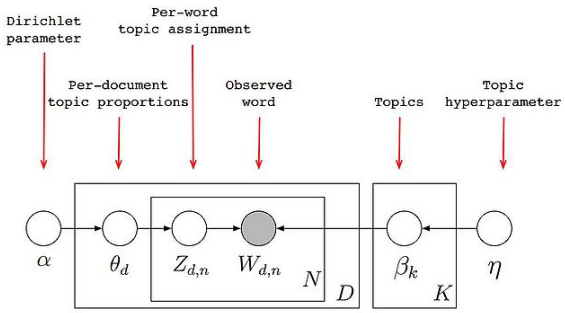

Fig 3.2.1 Latent Dirichlet Allocation Model

Where,

D denotes the number of documents
N is number of words in a given document (document d has Nd words)
α is the parameter of the Dirichlet prior on the per-document topic distributions
η is the parameter of the Dirichlet prior on the per-topic word distribution
θ_d is the topic distribution for document i
β_k is the word distribution for topic k

$z_{d,n}$ is the topic for the j -th word in document i
$W_{d,n}$ is the specific word.

So here the documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over all the words. So, to implement LDA we follow the below generative process.

1. Choose both the priors,

   θ_d ~ Dirichlet(α)
   where d is in {1,…,M}

   β_k ~ Dirichlet(β)
   where k is in {1,….K}

2. For each of the word positions d, n where d is in {1,…,M}, and n is in {1,…,N}

   Choose a topic $z_{d,n}$ Multinomial (θ_d)

   Choose a word $W_{d,n}$ Multinomial ($\beta_{k,n}$)

The above generative process implemented using techniques such as MAP estimation using EM algorithm and Variational Expectation Maximization. Both techniques make use of expectation maximization step to obtain optimized values for topic proportions and topics.

3.3 Market Basket Analysis

The process of analyzing the shopping trends of customers based on the itemsets present in the baskets is known as 'Market Basket Analysis'. Companies are interested in analyzing the data to learn about the purchasing behaviour of their customers. This information is used for various purposes like marketing promotions, inventory management and customer relationship management. The idea here is to analyze and find patterns, which would reveal that if someone buys an item, they are bound to buy other related product. Several measures which are useful in calculation of Market-Basket Analysis are as shown below.

Support: Less frequently occurring items can be filtered out using support, as it tells us about the combination of items bought together frequently.

$$Support(I) = \frac{Number\ of\ transactions\ containing\ I}{Total\ number\ of\ transactions}$$

Confidence: It reveals how frequently two items A and B are bought together, for the number of times A is bought.

$$Confidence(A \rightarrow B) = \frac{Number\ of\ transactions\ containing\ A\ and\ B}{Number\ of\ transactions\ containing\ A}$$

Lift: Strength of an association rule is determined by Lift, more the lift more is the strength of the rule.

$$Life = \frac{Support}{Support(A) \times Support(B)}$$

One of the ways to perform Market Basket Analysis is using Apriori Algorithm which uses frequently bought itemsets to generate association rules. The idea here is that the subset of a frequent itemset is also frequent and the frequent item sets are decided based on minimum support threshold.

Below lines sum up the Apriori Algorithm.

- Use k-1 itemsets to generate k itemsets
- Getting C[k] by joining L[k-1] and L[k-1], where C[k] is the candidate set and L[k-1] are the frequent itemset of length 'k-1'.
- Prune C[k] with subset testing
- Generate L[k] by extracting the itemsets in C[k] that satisfy minimum Support

4. Results

In order to get a baseline on the buying patterns of users who do not order frequently, the Apriori algorithm was applied on the set of users who had ordered less than 15 times from Instacart. As the number of products were as high as 50,000 and the number of orders being more than 1.2 million, due to computational limitations the Apriori algorithm took a lot of time to generate rules.

For a workaround of the computational limitations, items with a support count less than the threshold were removed manually before applying the Apriori algorithm which resulted in faster and better results as it helped in decreasing the number of items in the dataset. Due to less number of items, many transactions were completely eliminated from the dataset as they did not contain any of the remaining items and additionally it also had an effect on the average transaction width which led to less number of candidate itemsets that had to be examined and additionally less itemsets were present in each transaction which led to less hash tree traversals. The support count for the transactions left had to be adjusted in order for the initial threshold cutoff selected to hold true for the next traversal of data.

After using the workaround described above, the apriori algorithm was able to generate multiple rules at support of 0.01. This gave a baseline of the usual buying behavior of the users who did not order frequently from Instacart and what are the products that they prefer buying from Instacart and which items are bought together frequently.

antecedents	consequents	antecedent support	consequent support	support	confidence	lift
'Organic Lemon'	'Organic Hass Avocado'	0.084551	0.146123	0.035274	0.417189	2.855056
'Bag of Organic Bananas', 'Organic Strawberries'	'Organic Hass Avocado'	0.079441	0.146123	0.030665	0.386006	2.641650
'Banana', 'Limes'	'Large Lemon'	0.066704	0.169756	0.030355	0.455077	2.680766
'Banana', 'Organic Baby Spinach'	'Organic Avocado'	0.075436	0.164249	0.033116	0.439001	2.672770

Fig 4.1 Market-Basket Analysis: High Confidence and Lift

These rules are still dominated by the items that are brought the most frequently in the entire dataset as the set of users who have ordered less than 15 times from Instacart still represent 65% of the dataset and approximately 40% of the transactions which shows why these users need to be further clustered in order for more targeted analysis to be done on the users on the basis of their buying pattern.

To further break down the user segment, two approaches were tried, one is clustering users on the basis of K-means and another one is LDA. After obtaining the clusters, Market basket analysis is applied to the clusters using the apriori algorithm and the rules obtained are compared to evaluate the better clustering technique for our dataset.

4.1 K-means

In order to cluster users using K-means from the items they have bought; each feature was the number of times a particular user has bought a particular product from Instacart. For deciding the ideal number of clusters, the Silhouette Score was plotted against the number of clusters and k=10 was selected as the adequate number of clusters.

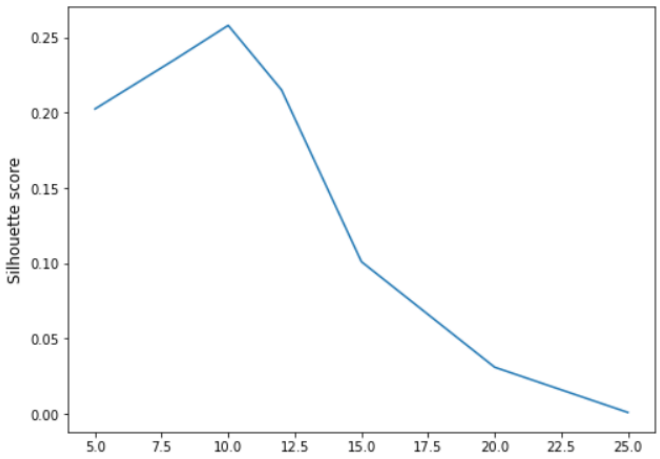

Fig 4.1.1 Silhouette Coefficient Plot for K-means

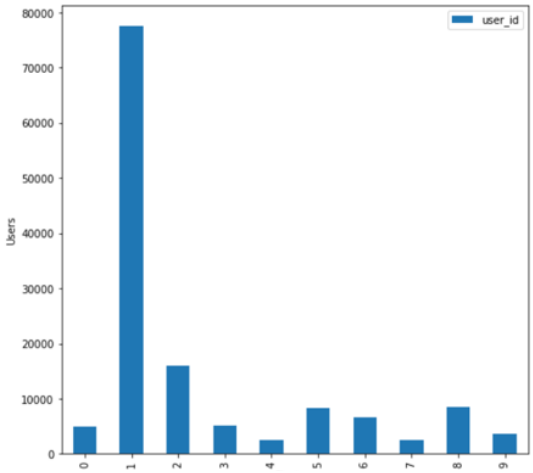

Fig 4.1.2 User Count for each K-means cluster

Even after multiple restarts, k-means still generated one big cluster which represented almost 55% of the users. On applying the apriori algorithm (support = 0.01) on the biggest cluster, the association rules generated were still similar and highly influenced by the frequent items in the entire dataset and as these rules were still very generic and affecting 55% of the users, it would not have helped in furthering our analysis.

antecedents	consequents	antecedent support	consequent support	support	confidence	lift
'Organic Italian Parsley Bunch'	'Organic Garlic'	0.068966	0.115322	0.028855	0.418397	3.628059
'Jalapeno Peppers'	'Limes'	0.060450	0.154693	0.028608	0.473254	3.059314
'Organic Garnet Sweet Potato (Yam)'	'Organic Baby Spinach'	0.050910	0.203752	0.026362	0.517818	2.541415
'Bunched Cilantro'	'Limes'	0.066930	0.154693	0.032360	0.483496	3.125521

Fig 4.1.3 Market-Basket Analysis on biggest cluster after k-means: High Confidence and Lift

antecedents	consequents	antecedent support	consequent support	support	confidence	lift
'Organic Baby Spinach', 'Banana'	'Organic Avocado'	0.075137	0.164591	0.033348	0.443824	2.696523
'Large Lemon', 'Banana'	'Limes'	0.077667	0.154693	0.030349	0.390752	2.525981
'Limes', 'Banana'	'Large Lemon'	0.067263	0.168627	0.030349	0.451193	2.675685
'Bag of Organic Bananas', 'Organic Strawberries'	'Organic Hass Avocado'	0.078247	0.145264	0.030188	0.385804	2.655888

Fig 4.1.4 Market-Basket Analysis on biggest cluster after k-means: Similar rules

On analyzing the second biggest cluster, it can be seen that even though each cluster produces a few distinct rules that can be applied on the individual clusters, there are still many similar rules reproduced.

antecedents	consequents	antecedent support	consequent support	support	confidence	lift
'Banana', 'Organic Baby Spinach'	'Organic Avocado'	0.075939	0.162977	0.035994	0.473988	2.908308
'Banana', 'Large Lemon'	'Limes'	0.080329	0.155013	0.030727	0.382514	2.467615
'Banana', 'Limes'	'Large Lemon'	0.066596	0.170502	0.030727	0.461394	2.706084

Fig 4.1.5 Market Market-Basket Analysis on second cluster (20k users) after k-means: Similar rules

4.2 LDA

A more appropriate method to find groups of users with similar buying behaviour was to use LDA on the users where the set of documents would be replaced with a set of users and the number of words in each document would be replaced by the number of products bought by the user in their transaction records.

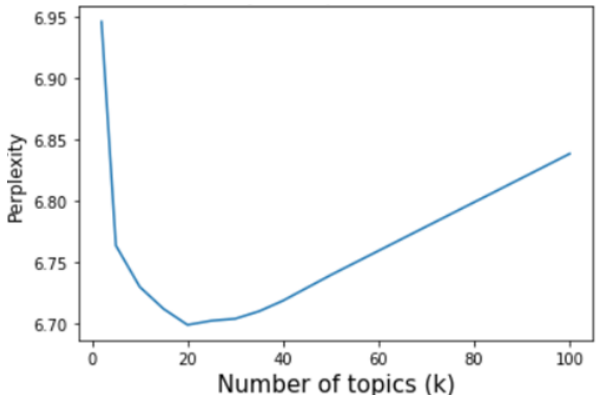

Fig 4.2.1 Perplexity vs Topic Count Relation

According to the perplexity on different values of k, 20 topics were selected to cluster the users on the basis of their buying pattern. On the basis of the results of LDA, products which seem to share properties are in the same topic and hence these topics can be categorized on the basis of the intuitions of the users and many discernable patterns can be seen in the topics that have been modelled on the basis of the products in them.

Fig 4.2.2 Topics of LDA

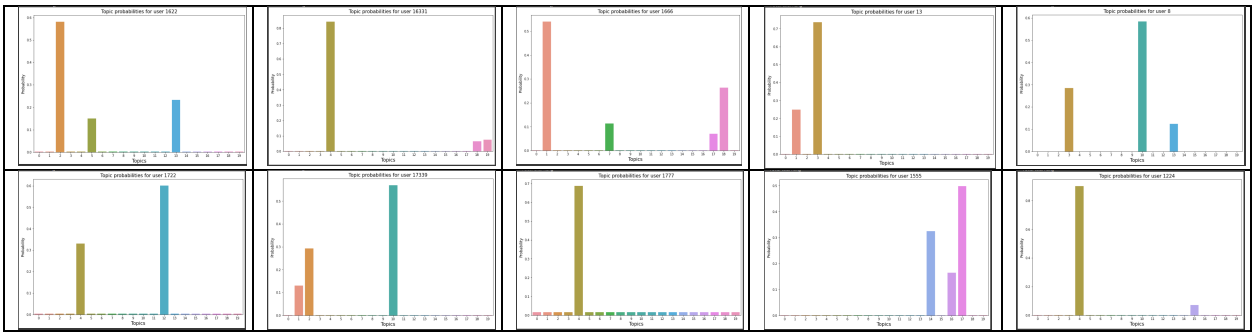

Fig 4.2.3 Topic Distribution

In order to further cluster the users on the basis of their buying pattern, users are clustered into the topics in which they have the highest probability. Due to the sparse nature of the topic distributions, most users just have one very dominant topic and the rest seem to have negligible influence towards the clustering of the user. Hence, it can be assumed that results obtained by simply clustering the user on the basis of the highest projection in topics is viable as LDA is still producing a lower dimensional representation of the documents in a corpus.

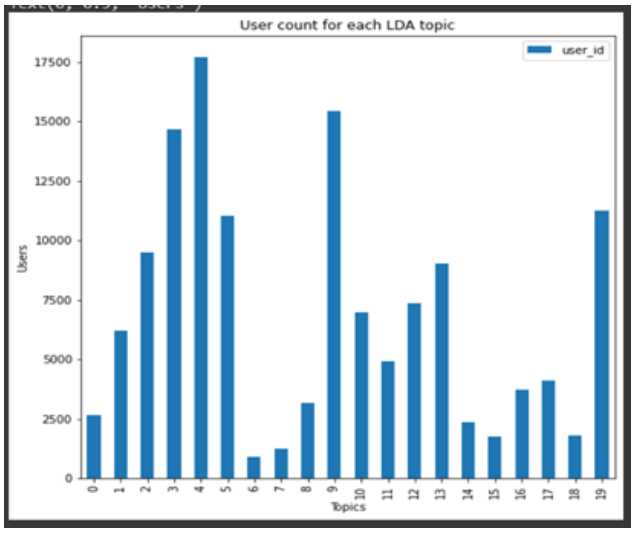

Fig 4.2.4 Number of users in each topic

The users are clustered more evenly using LDA as compared to K-means where the biggest cluster represented 55% of the users.

On applying market basket analysis (support = 0.01) on a few of the most frequent topic clusters, it can be seen how frequent products and association rules have products that are related to the topic associated with it and the rules with high confidence and lift are from the top probabilistic products in the topic.

Topic – “Organic Products” (17717 users)

antecedents	consequents	antecedent support	consequent support	support	confidence	lift
Large Lemon	Limes	0.106250	0.109835	0.024427	0.229904	2.093176
Limes	Large Lemon	0.109835	0.106250	0.024427	0.222399	2.093176
Organic Ginger Root	Organic Garlic	0.048592	0.105725	0.012942	0.266347	2.519242
Organic Italian Parsley Bunch	Organic Garlic	0.050458	0.105725	0.011048	0.218949	2.070926
Organic Yellow Onion	Organic Garlic	0.085145	0.105725	0.020871	0.245122	2.318483
Organic Lemon	Organic Hass Avocado	0.060223	0.095085	0.012709	0.211036	2.219434

Fig 4.2.5 Association Rules for Organic Products Topic: High Confidence and Lift

Topic – “Berries” (15462 users)

antecedents	consequents	antecedent support	consequent support	support	confidence	lift
'Blueberries'	'Strawberries'	0.075786	0.149261	0.026340	0.347561	2.328551
'Organic Blueberries'	'Organic Strawberries'	0.118762	0.198706	0.049908	0.420233	2.114849
'Organic Strawberries'	'Organic Raspberries'	0.198706	0.108133	0.054529	0.274419	2.537786
'Raspberries'	'Strawberries'	0.060074	0.149261	0.022181	0.369231	2.473732

Fig 4.2.6 Association Rules for Berries Topic: High Confidence and Lift

Topic – “Veggies” (14656 users)

antecedents	consequents	antecedent support	consequent support	support	confidence	lift
'Jalapeno Peppers'	'Large Lemon'	0.062650	0.168211	0.026777	0.427397	2.540833
'Large Lemon', 'Organic Baby Spinach'	'Organic Avocado'	0.056042	0.160831	0.026262	0.468606	2.913662
'Red Onion'	'Large Lemon'	0.067628	0.168211	0.027463	0.406091	2.414172
'Large Lemon', 'Organic Avocado'	'Organic Baby Spinach'	0.054411	0.206488	0.026262	0.482650	2.337421
'Bunched Cilantro'	'Large Lemon'	0.066598	0.168211	0.028665	0.430412	2.558758

Fig 4.2.7 Association Rules for Veggies Topic: High Confidence and Lift

Topic – “Apples” (11031 users)

antecedents	consequents	antecedent support	consequent support	support	confidence	lift
'Apple Honeycrisp Organic'	'Organic Large Extra Fancy Fuji Apple'	0.071168	0.149166	0.026847	0.377232	2.528942
'Honeycrisp Apple'	'Organic Fuji Apple'	0.073392	0.157109	0.034154	0.465368	2.962074
'Bag of Organic Fuji Apple'	'Apple Juice'	0.065131	0.203177	0.026688	0.409756	2.016743
'Golden Delicious Apple'	'Apple Juice'	0.122478	0.203177	0.051311	0.418936	2.061927

Fig 4.2.8 Association Rules for Apple Topic: High Confidence and Lift

5. Discussion

The main goal of the project was to increase the revenue/sales for Instacart. To achieve this first a user segment was selected using EDA, the users having less than 15 total orders were targeted because they account for about 65% of the total users but only 40% of total orders and targeting those users will lure them towards ordering more frequently. Then market basket analysis was applied on this user cluster and the association rules were formed.

After applying market basket on this user segment, the most frequent itemsets found were still the products which were most frequently bought by the entire user cluster, which showed that this segment of customers may still have many properties of the entire customer dataset. So, it was learnt that further clustering of users is required using some clustering algorithm based on their buying patterns to generate more useful association rules.

The first algorithm implemented to cluster the users was K-means clustering method. The segment of users with less than 15 orders is clustered based on the products they order. The number of clusters K is selected based on the analysis of Silhouette Coefficient score. Thus 10 clusters were formed after applying K-means with multiple restarts and the number of users in each cluster were calculated and it was discovered that almost 55% of the users with less than 15 orders were in one cluster. Further when market-basket analysis was applied on the user clusters, it was found that the rules formed contained a few distinct rules but most of them were almost similar in every cluster and it was a repetition of the total user cluster. Thus K-means algorithm was rejected based on two points that it was unable to break the large cluster which consisted of around 55% users of less than 15 orders and that it was unable to mine a lot of distinguishable rules among the clusters.

Further LDA was used, where users were treated as documents and products as words. LDA was implemented taking all the products as dimensions, it proved to be a more appropriate method to find cluster of users based on their product affiliations. The number of topics was selected to be 20 based on the perplexity score. Now the users are assigned to the topic which is the most dominant. Thus, the users are divided into 20 clusters based on the group of products which are the most dominant for that user. After this market basket analysis was applied on these clusters and the rules were evaluated.

The results showed how topic modelling affects the market basket analysis. It can be seen that the rules obtained on the LDA user clusters are not repeating as much and are less redundancy than the rules obtained on the targeted user segment using K-means. The reason for this is LDA can come up with topics which have similar kind of products in them and when user clusters are made on the basis of these topics, they all tend to have an affiliation towards certain types of products. Hence when market basket is applied to this cluster, it tends to create rules only for those certain types of products. Hence the rules obtained are less redundant and are distinguishable enough within the clusters.

Limitations of results

Absence of real-time data has produced some limitations on the evaluation for the final results of the experiment. As no real-time data is available to work with, the actual potential of these final rules is still not precisely known and testing them on a real user base is required. One more problem faced during the analysis is that irrespective of the clustering methods used, some products which have an exceptionally high frequency count when compared to other products are dominating the association rules obtained.

Future work

The future work includes trying different clustering algorithms other than LDA and testing the results on the real-world data. Later choosing a better clustering algorithm which gives more meaningful clusters, and which are helpful in effective marketing. Customers can be lured to buy more products by recommending them products based on the products they put in their cart and the buying patterns of other similar customers. So, a recommendation system can be implemented on the dataset to recommend products when the user adds a product to their cart.

References

[1] “The Instacart Online Grocery Shopping Dataset 2017”, Accessed from https://www.instacart.com/datasets/grocery-shopping-2017
[2] “David M Blei; Probabilistic Topic Models”, Accessed from https://www.ccs.neu.edu/home/jwvdm/teaching/ds5230/spring2021/assets/pdf/blei-topic-models.pdf
[3] “Association Analysis: Basic Concepts and Algorithms ”, Accessed from https://www.ccs.neu.edu/home/jwvdm/teaching/ds5230/spring2021/assets/pdf/tan-steinbach-kumar-ch6.pdf
[4] “John P. Cunningham, Zoubin Ghahramani; Linear Dimensionality Reduction: Survey, Insights, and Generalizations”, Accessed from https://www.ccs.neu.edu/home/jwvdm/teaching/ds5230/spring2021/assets/pdf/cunningham-ghahramani-linear-dimensionality-reduction.pdf
[5] “Namita Dave, Karen Potts, Vu Dinh, and Hazeline U. Asuncion; Combining Association Mining with Topic Modeling to Discover More File Relationships” Accessed from https://www.thinkmind.org/articles/soft_v7_n34_2014_9.pdf
[6] “Shi Na, Liu Xumin, Guan Yong; Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm”, Accessed from https://ieeexplore.ieee.org/document/5453745
[7] “Fei Wang, Hector-Hugo Franco-Penya, John D. Kelleher, John Pugh, Robert Ross; An Analysis of the Application of Simplified Silhouette to the Evaluation of k-means Clustering Validity” Accessed from https://www.researchgate.net/publication/318109824_An_Analysis_of_the_Application_of_Simplified_Silhouette_to_the_Evaluation_of_k-means_Clustering_Validity
[8] “Tushar Kansal; Suraj Bahuguna; Vishal Singh; Tanupriya Choudhury; Customer Segmentation using K-means Clustering” Accessed from https://ieeexplore.ieee.org/document/8769171