Smart Shopping: Transaction Predictor

Aditi Kumari[1] and Aditya Doshatti[2] and Caroline Chandraguptrajah[3] and Devashish Nyati[4]

Abstract— **Transaction prediction is a task that predicts the next items a customer is likely to buy. This will help businesses gain insights into customer behavior and preferences and make recommendations accordingly. In this project we look at the task of transaction prediction as a binary classification problem, of whether a customer is likely to buy the item in the next order or not. This prediction is based off of a user's past transactions and general preferences. For a given input pair (u, v) of user u and product v, we predict whether the user will buy the product in the next order (1) or not (0). We use logistic regression for this purpose and since logistic regression can also give us probabilities between the interaction of these two variables, we use the predicted probabilities to sort the products and generate a recommendation list for each user. Major part of the work was done on generating meaningful features from the large dataset containing around 3 million Instacart orders. The resulting accuracy was comparable to other state of the art approaches. We also explore customer segmentation to better adjust our recommendations to different segments of customers.**

Index Terms— **Transaction Prediction, Customer Segmentation, Logistic Regression, Principal Component Analysis, K-Means Clustering**

I. INTRODUCTION

With the growth of internet and online shopping solutions, we now can provide products to customer with a click of a button. Through time these online e-commerce websites have gathered significant amount of data on their customers. These websites also have tools to make their marketing more user centric and personalized. However, one important aspect of this is, recommending the right products to the customers. This has a potential for making huge profits, as it has been found that users spend a great amount of time on searching for products and reading reviews on them before deciding to buy a particular product. A recommendation list will be a valuable resource for such customers as they will quickly be able to look at products that match their preferences and preferences of other users belonging to the same customer segment.

In hindsight it will help businesses better understand their customers, segment these customers, and address them accordingly. They will be able to calculate customer lifetime value for every customer which will help them bias decisions based on long term revenue rather than short term profits. It will help them adjust their marketing strategies and personalize them to each

Recommendation systems have been widely researched since the end of 1990's. It began with a novel method of recommending products called collaborative filtering. A typical recommendation system on an online e-commerce website is a list of products that you will see under the title "Other products you may like" , "frequently bought together" and "Users who bought this also liked" etc. A lot of ongoing research is being done into perfecting this list of recommended products through different learning algorithms and association rules.

A. Problem: Predict the next items an Instacart user will purchase and perform customer segmentation

Instacart is leading online grocery shopping website. They released their anonymized data set for researchers and machine learning practitioners to predict the next items a user will purchase. The data set contains over 200,000 users who placed more than 3 million orders in total. The scope of this data set is very huge and can be used to find new and interesting patterns, associations, and predictions. Our project is one such attempt to find out the future transaction prediction of the customers and to segment customers based on their purchase history.

B. Data : The Instacart dataset

The data set is a relational set of 4 main csv files containing details about all the orders, the products, the aisles the product belongs to and the department the product belongs to. For each Instacart user, we have between 4 and 100 orders, and we also have information of the sequence in which the products were placed in the cart. Information about the day of the week the order was placed, time of the day the order was placed and relative time between orders has also been given.

C. Outline

The structure of the report follows the typical path of a machine learning project. In section 2 we explain the exploratory data analysis conducted on the data set and our observations. Section 3 briefs about the current state of the art approaches to similar market basket predictions. We then explain the feature engineering step, detailing the type of feature and method used to extract them. In section 5 we explain our model and different variations of the model used. Part 2 of the report explains our customer segmentation methodology and its results. We then conclude with evaluation results.

A link to our project on GitHub can be found here : *GitHub Link*

II. DATASET AND EXPLORATORY DATA ANALYSIS

A. *Dataset Details*

Instacart releases it data open source. We selected the data set which contains almost 3 million instacart transaction. The data set has details about the orders, products, aisles, departments and data of products ordered prior and order data for training. The link for the data set is *Instacart data set.*

B. *Exploratory Data Analysis*

Going according to our aim to predict future transaction we analyzed the data for any unnecessary information and reducing the data size to save the computation time.From the analysis of orders data we got that we have 131209 transaction record to train our model with minimum of 4 unique items per transaction to maximum of 100 items. Below graph clearly depicts these details in the dataset.

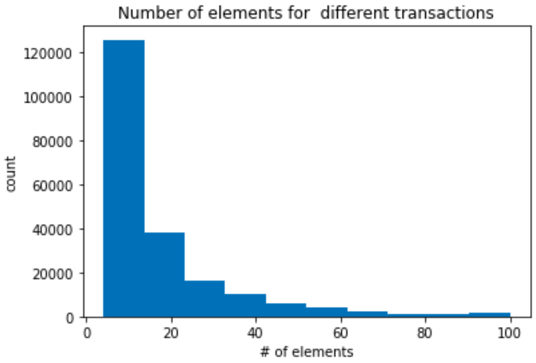


Fig. 1. Number of different elements

Analyzing more about the time and dates of order we get that most of the shopping is done on Saturday and Sunday but we see transaction every day. Also maximum number of transaction of first week of every month but there are transaction on every week of every month. The frequency of transaction by hour is always high between 9AM to 4PM. In the prior order data set checking the probability of reorder data we got its 0.59. Hence, any product bought has good chances of reordering.

The graph in Fig. 2 clearly states the trend of the data, that maximum shopping is done on weekends during day time. But it also states that the data contains transactions for everyday and various hours of day. So the data set has sufficient transactions with variety.

Merging the order data, aisle data and product data we found the item with maximum frequency of ordering and getting their aisle we got that maximum products are from fruits and dairies departments and from nearby aisles. When analyzed through out the data with department wise orders the maximum number of orders where from produce and dairies. The maximum frequency of reorder was also from the same department. While getting the reorder ratio based on aisles we got that beverages, breakfast, fresh fruits and fresh vegetables are most reordered products. People who order weekly go for this and generally re order same things.

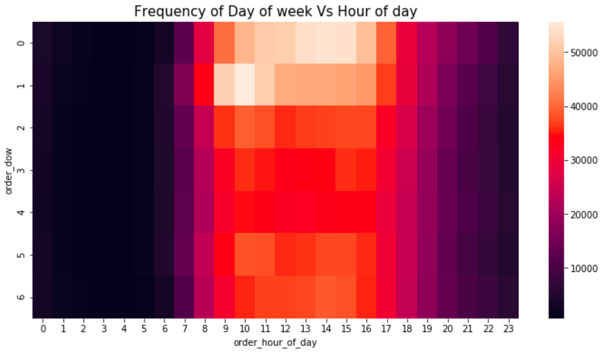


Fig. 2. Frequency of Day of week vs hour of day

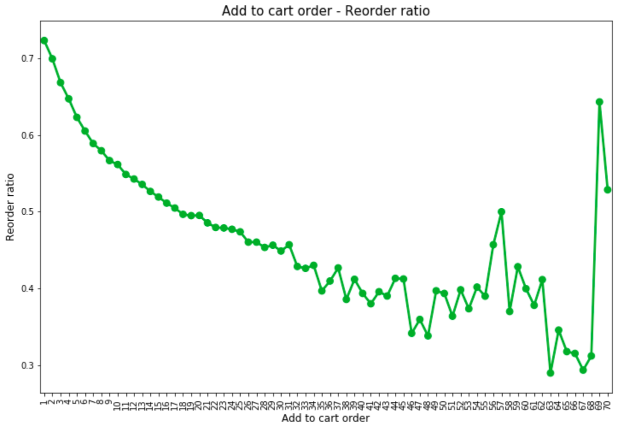


Fig. 3. Add to cart reorder ration

The data set also has another feature of add to cart data, Generally people tend to add to cart the items which they order regularly and are their mostly ordered products. As we already discussed the maximum orders has 4 items, Hence in the data set the add to cart has most number of 1 items and them the frequency for more number of items goes on decreasing. There are few more high ratio data where the add to cart number of items is around 50 that can be because we might less data with orders with that much items in order hence there are high frequency values in that range in the reorder ratio.

III. RELATED WORKS

Market basket predictions are aimed at understanding user behavior and building an effective recommendation system. There have been many approaches to building this recommendations system, which can be categorized as collaborative filtering, content based filtering, sequential, customer pattern based, hybrid and sequence matching method.

A. Collaborative filtering method

Collaborative filtering is the most popular approach to recommendation systems. This article [5] discusses the various collaborative filtering techniques that have been widely used. These techniques produce recommendations based on general preferences of the user. The Collaborative filtering algorithm believes that when users have similar ratings on products that might have similar preferences as well. The product recommendations are thus extracted from neighbors of the target user.

B. Content based filtering method

In content-based filtering [6] the product features and functions are taken into account. These properties are analyzed and matched with the product the user is interested in purchasing. The user is then recommended products with similar properties.

C. Sequential recommendation method

These systems are built on markov chains and leverage sequential information in a user's purchase history [7]. It recommends products based on both a user's long-term preferences and small sequential patterns called short term dynamics in the user's search history or most recent market baskets.

D. Pattern Based

These methods identify patterns in items that are frequently brought together [8]. They discard any sequential information about a user's historical purchases. They extract item sets from the order history of all customers and predict the next market basket for a user. It overcame the cold-start problem that existed in other methods.

E. Hybrid method

This method combines the ideas in the collaborative filtering method and sequential method. There are many versions of this method, one of which combines collaborative filtering and content based filtering [9]. It's the most widely used method in e-commerce website and provides recommendations with good accuracy.

F. Sequence Matching Method

This method [10] proposes a novel method for market basket prediction by means of similarities in sequential purchase histories across customers. It attempts to predict market baskets of customers by identifying similar purchase habits among different customers. They use a method called sub- sequential dynamic time warping for identifying cross-customer patterns. The similarities are measured by Wasserstein distance between different purchase histories. Once this distance is measured, they use the k-Nearest Neighbor Matching to cluster customers with similar purchase histories. The prediction of the next market basket is then calculated by taking into account the most common items purchased within that customer segment.

IV. FEATURE ENGINEERING

We extracted a lot of features to be given to our models. All these features scaled in the range 0 to 1. We created a feature matrix and the target labels. Then the problem was a simple classification problem. Here is the definition of all the features extracted:

A. User Features

- User Product Total Order: The total number of times a user has ordered any particular product.
- Latest Cart: Set of all the products a user has ordered overall in the train data set.
- In Cart: Current product id in the latest cart is present or not. This is our Y_label. 0: Not present 1: Present
- User Total Orders: Total number of orders that user_id has.
- User Avg Cart Size: Average number of products in a user's cart.
- User Total Products: The number of products a user has purchased.
- User Avg Days Since Prior Order: Mean of any user_id day since prior order.
- User Avg Order DOW: The day of the week when the user orders averagely.
- User Avg Order Hours Of Day: The hour of the day when the user orders averagely.
- User Avg Days Since Prior Order: The average number of days since the user ordered priorly.

B. Product Features

- Product Total Orders: Total number of times any product has been ordered.
- Product Avg Add To Cart Order: Average number of times any product has been added to cart.
- Product Avg Order DOW: The day of the week when that product is ordered averagely.
- Product Avg Order Hours Of Day: The hour of the day when that product is ordered averagely.
- Product Avg Days Since Prior Order: The average number of days since that product was ordered before.
- Product Total Orders Delta Per User: The difference between total number of product orders and the total number of times a user has ordered that product.
- Product Avg Add To Cart Order Delta Per User: The difference between the average number of times a product has been added to cart and the average number of times the user has added that product in the cart.
- Product Avg Order DOW Per User: The day of the week when the particular product is ordered averagely per user.
- Product Avg Order Hours Of Day Per User: The hour of the day when the particular product is ordered averagely per user.
- Product Avg Days Since Prior Order Per User: The average number of days since the product is ordered priorly per user.

- All The Departments: All the departments given in the departments.csv file.

C. *User-Product Features*

- User Product Avg Add To Cart Order: A user adds a particular product how many times averagely.
- User Product Avg Frequency: The total number of user-product orders divided by the total number of user total orders.
- User Product Avg Days Since Prior Order: The average number of days since the user ordered that product priorly.
- User Product Avg Order DOW: The day of the week when the user orders that particular product averagely.
- User Product Avg Order Hours Of Day: The hour of the day when the user orders that particular product averagely.

V. MODELS

After Extracting the features, we used 2 models to test the features:

- Naive Bayes: Since the data is too huge, Naive Bayes runs very quickly. Here, A can be replaced as the Y_Label that is the classes and B can be replaced as the features. So it gives the probability of a class over the features. This gave an accuracy on the training data as 0.70 and on the test data as 0.24.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Fig. 4. Naive Bayes

- Logistic Regression: Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). Here we used logistic regression as it is good for binary classification. It uses a sigmoid function that classifies the classes as either 0 or 1.
  Also, logistic regression is interpretable.

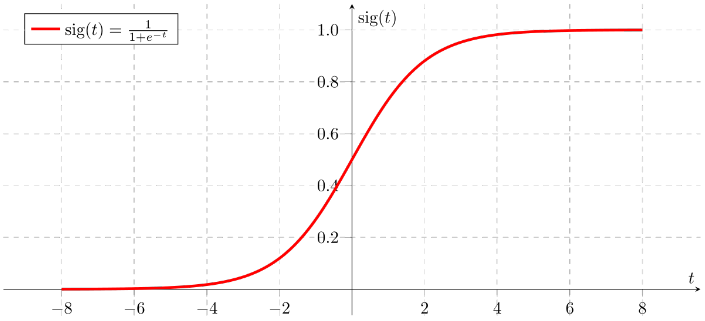


Fig. 5. Logistic Regression

  In logistic regression, we used 3 models:
  - Simple Logistic Regression
  - Class Weight Balanced
  - Class Weight Manual

  After training the models, we calculated the accuracy on the test data. For simple logistic regression, the accuracy was 0.20, for class weight balanced it was 0.37 and for manual class weight, it was 0.39. Since the classes were imbalanced, this makes sense. We decided to go with manual class weights. The training accuracy came out to be 0.83 and the test accuracy came out to be 0.39. Figure 6 displays the confusion matrix:

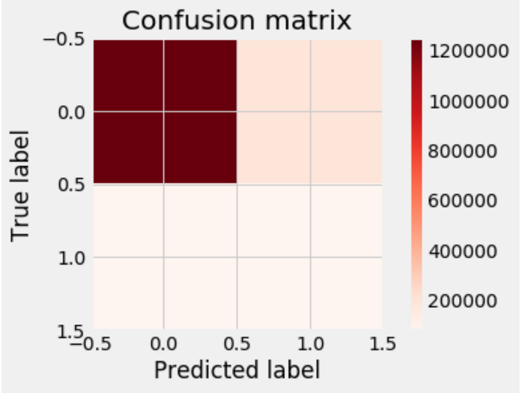


Fig. 6. Confusion Matrix

VI. CUSTOMER SEGMENTATION

Customer segmentation refers to classifying customers by inferring information from their past purchase patterns. The customers were classified into various groups based on the purchases made from various aisles.

A. *Models*

We used two models for segmenting the customer based on their previous transactions.

- Principal Component Analysis(PCA): The main idea behind this algorithm is to reduce the dimensionality of the data set and perform data analysis. It was used to reduce the dimensionality of the variables that were correlated heavily while retaining the variation present in the dataset. This was achieved by transforming the variables into a new set of variables known as principal components. We used 6 principal components for this purpose.
  The data frame representing all the purchases made by the customer was created using the data exploration techniques. Principal Component Analysis was then executed to reduce the number of features from the number of aisles to 6.

- K-means clustering: K-means is an unsupervised learning algorithm that aims to partition the dataset into k distinct non-overlapping subgroups. Every data point of the dataset belongs to only one cluster group. The algorithm identifies k number of centroids and assigns each data point to the nearest cluster. Centroids are kept as small as possible. Data points are said to be more homogeneous if there is less variation within the clusters.

We used this algorithm on the dataset after performing the dimensionality reduction. It divided the dataset into four customer groups based on their purchase history. The output values correspond to aisles present in the grocery stores.

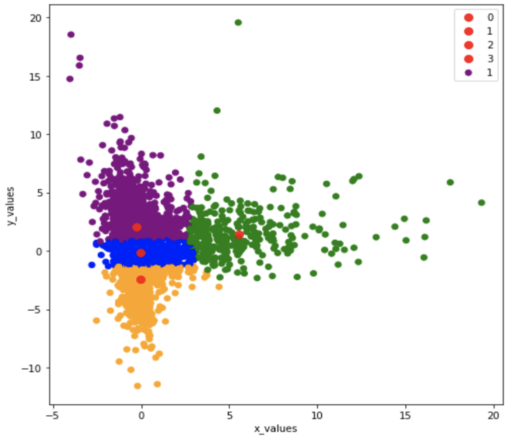
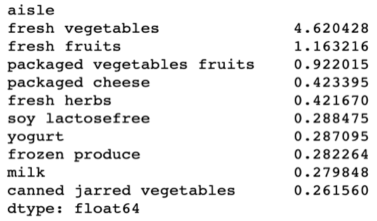


Fig. 7. Customers segmented into four groups based on past purchases



```
aisle
fresh vegetables                 4.620428
fresh fruits                     1.163216
packaged vegetables fruits       0.922015
packaged cheese                  0.423395
fresh herbs                      0.421670
soy lactosefree                  0.288475
yogurt                           0.287095
frozen produce                   0.282264
milk                             0.279848
canned jarred vegetables         0.261560
dtype: float64
```

Fig. 8. Top 10 items purchased by people of the segment 1

The table in fig 8 represents the top ten goods that were purchased by people in one of the segmented clusters. We relied on the absolute data followed by percentage among the top 8 products for each cluster.

VII. EVALUATION

For evaluation, we used the accuracy score and confusion matrix as a parameter. Since this was a classification problem, the confusion matrix suited the best. The ROC curve shows the relationship between false positives and true positives.

VIII. CONCLUSIONS

Our simple logistic regression model with tailored features and manual class balancing has given us good results. Other state of the art approaches to this problem on the same data set has given a maximum accuracy of 0.40, while our approach is comparable with an accuracy of 0.39. By fine tuning our coefficients of the model, we observed that the frequency of a customer's order was the highest contributing factor to whether the product will be reordered or not. The total orders for a particular product also played a major role in its chances of being reordered again.

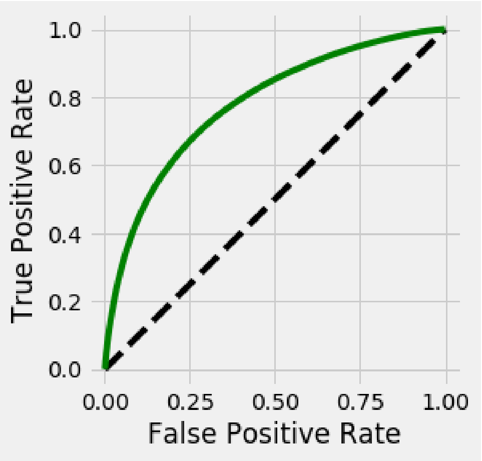


Fig. 9. False Positive Rate

We explored customer segmentation based on aisles, and conjecture that recommending products based on which segment the user belongs to will fetch us better results. This segmentation information could be an added feature to our initial feature set used for product recommendation. The result of such segmentation features looks promising and is yet to be studied.

REFERENCES

[1] M. Kraus, S. Feuerriegel. "Personalized Purchase Prediction of Market Baskets with Wasserstein-Based Sequence Matching"

[2] A.Krishna Kumar, D.Amrita, N.Swathi Priya. "Mining Association Rules between Sets of Items in Large Database" , In International Journal of Science and Modern Engineering (IJISME).

[3] Riccardo Guidotti, Giulio Rossetti, Luca Pappalardo, Fosca Giannotti, and Dino Pedreschi. 2017. "Market basket prediction using user-centric temporal annotated recurring sequences." In IEEE International Conference on Data Engineering

[4] Guidotti, R., Gabrielli, L., Monreale, A., Pedreschi, D., & Giannotti, F. (2018). "Discovering temporal regularities in retail customers' shopping behavior". EPJ Data Science, 7, 1-26

[5] J. Xia, "E-Commerce Product Recommendation Method Based on Collaborative Filtering Technology," 2016 International Conference on Smart Grid and Electrical Automation (ICSGEA), Zhangjiajie, 2016, pp. 90-93.doi: 10.1109/ICSGEA.2016.81

[6] A. Pal, P. Parhi and M. Aggarwal, "An improved content based collaborative filtering algorithm for movie recommendations," 2017 Tenth International Conference on Contemporary Computing (IC3), Noida, 2017, pp. 1-3.

[7] R. He and J. McAuley, "Fusing Similarity Models with Markov Chains for Sparse Sequential Recommendation," 2016 IEEE 16th International Conference on Data Mining (ICDM), Barcelona, 2016,

pp. 191-200.doi: 10.1109/ICDM.2016.0030

[8] E. Lazcorreta, F. Botella and A. Fernández-Caballero, “Towards personalized recommendation by two-step modified Apriori data mining algorithm”,Expert Systems with Applications, vol. 35, no. 3, pp. 1422-1429, 2008. Available: 10.1016/j.eswa.2007.08.048.

[9] A. Romadhony, S. Al Faraby and B. Pudjoatmodjo, “Online shopping recommender system using hybrid method,” 2013 International Conference of Information and Communication Technology (ICoICT), Bandung, 2013, pp. 166-169. doi: 10.1109/ICoICT.2013.6574567

[10] Kraus, M., Feuerriegel, S. (2019). Personalized Purchase Prediction of Market Baskets with Wasserstein-Based Sequence Matching.arXiv preprint arXiv:1905.13131.