A close up of a logo

Description automatically generated

FINAL PROJECT

RONAK SOMANI

NORTHEASTERN UNIVERSITY

MPS ANALYTICS – WINTER 2019

SUBJECT: ALY 6040 21414 DATA MINING APPLICATIONS

INSTRUCTOR NAME: PROF. JUSTIN GROSZ

SUBMISSION DATE: March 25, 2019

**INTRODUCTION**

In the recent years, analyzing shopping baskets turned out to be very appealing to retailers. Sophisticated technology made it possible for them to collect information of their customers and what they purchase. The introduction of electronic point-of-sale expanded the utilization and application of transactional data in Market Basket Analysis (MBA). In retail business, analyzing such information is exceedingly valuable for understanding purchasing behavior. Mining purchasing patterns allows retailers to adjust promotions, store settings and serve consumers better. In statistics, exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task. Regression analysis is a powerful statistical method that allows us to examine the relationship between two or more variables of interest. While there are many types of regression analysis, at their core they all examine the influence of one or more independent variables on a dependent variable. It provides detailed insight that can be applied to further improve products and services. Predictive analysis is an advanced branch of data engineering which generally predicts some occurrence or probability based on data. It uses data-mining techniques in order to make predictions about future events and make recommendations based on these predictions. The process involves an analysis of historic data and based on that analysis to predict the future occurrences or events. Predictive Analytics is composed of various statistical & analytical techniques used to develop models that will predict future occurrence, events or probabilities. It is able to not only deal with continuous changes, but discontinuous changes as well. Classification, prediction, and to some extent, affinity analysis constitutes the analytical methods employed in predictive analytics.

**WHAT IS THE DATA SET?**

In 2014, Market Basket which is a grocery chain in US combined with Instacart. Instacart operates in six largest cities nationwide, began working with Market Basket to deliver its food online. Instacart, a grocery ordering and delivery app, for on-demand grocery shopping with same-day delivery service, aims to make it easy to fill refrigerator and pantry with personal favorites and staples when needed. It uses a crowdsourced marketplace model, akin to that of Uber or Lyft. After selecting products through the Instacart app, personal shoppers review the order and do the in-store shopping and delivery for customers. The Instacart shopping process is as follows. First, an app user places their grocery order through the app. Then, a locally crowdsourced “shopper” is notified of the order, goes to a nearby store, buys the groceries, and delivers them to the user.

There are three ways that Instacart generates revenue: delivery fees, membership fees, and mark-ups on in-store prices. Instacart uses transactional data to develop models that predict which products a user will buy again, try for the first time, or add to their cart next during a session. The objective for this project is to use anonymized data on customer orders over time to predict which previously purchased products will be in a user’s next order and also to demonstrates a detailed analysis of Instacart Market Basket. With ever increasing cost of acquiring new customers and increasing competition, leveraging existing customer is the best option at the disposal of businesses. The best solution for this turns out to be analyzing using big data techniques. This proposal includes methods of data collection, storing, cleaning, validating, optimizing, predicting and analyzing various trends and patterns of Customer behavior.

**DATA SOURCE AND DATASET INFORMATION**

I have collected the dataset from “The Instacart Online Grocery Shopping Website”. And it can be downloaded from the link mentioned below:

<https://www.instacart.com/datasets/grocery-shopping-2017>

This dataset is a relational set of files describing customers’ orders over time. The goal of this project is to predict which products will be in a user's next order. The dataset is anonymized and contains a sample of over 3 million grocery orders from more than 200,000 Instacart users. For each user, we provide between 4 and 100 of their orders, with the sequence of products purchased in each order. We also provide the week and hour of day the order was placed, and a relative measure of time between orders. Also, each entity (customer, products, order, and aisle) has an association unique id. This dataset includes orders from many different retailers and is a heavily biased subset of Instacart’s production data. The values and all the ID’s in the dataset are entirely randomized and they cannot be linked back to any other ID. Also, the information provided about the users is sequence of orders and product in that order. Retail ID is not provided, only products that are brought by multiple people at various retailers are included. Many interesting patters can be found using these datasets.

**DATA OBSERVATIONS**

Total six datasets were imported. The description of each file is mentioned below:

orders (3.4m rows, 206k users):

* order\_id: order identifier
* user\_id: customer identifier
* eval\_set: which evaluation set this order belongs in (see SET described below)
* order\_number: the order sequence number for this user (1 = first, n = nth)
* 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 (with NAs for order\_number = 1)

products (50k rows):

* product\_id: product identifier
* product\_name: name of the product
* aisle\_id: foreign key
* department\_id: foreign key

aisles (134 rows):

* aisle\_id: aisle identifier
* aisle: the name of the aisle

deptartments (21 rows):

* department\_id: department identifier
* department: the name of the department

order\_products\_\_SET (30m+ rows):

* 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

where SET is one of the four following evaluation sets (eval\_set in orders):

* "prior": orders prior to that users most recent order (~3.2m orders)
* "train": training data supplied to participants (~131k orders)
* “test”: test data reserved for machine learning competitions (~75k orders)

**DATA CLEANING AND ADDITIONAL REWORK**

* The following dataset requires additional rework, in “orders, products, aisles, departments”, structure of some features are wrong type, we need to do some recoding and convert variables to factors.
* It can be noticed in “orders” csv file that only column with “days\_since\_prior\_order” has NA values summing up to 206209, which is approximately equal to 6% of the total. Also, it is equal to total of “user id”, so it can be assumed that every user has at least one missing record entry.
* By analyzing the data, it can be noticed that almost NA values of “days\_since\_prior\_order” have “eval\_set” at prior, and “order\_number” at 1. The percentage 6% of missing value is big because Prior factor which relate to Na’s values of “days\_since\_prior\_order” is the biggest part in “eval\_set”. Hence, the missing values could not be dropped. We will explore their interaction.

**DATA PROFILING**

After performing EDA on this dataset many interesting facts were uncovered like

1. What time of the day do people order the highest selling products?

A close up of a map

Description automatically generated

From the above plot it can be stated that, there is a clear effect of hour of day on order volume. Most orders are between 8.00-18.00 hours. Also, the hour at which the maximum number of orders placed is 10AM which is approximately equals to 2874905 orders which is when people usually state their meal and day. Instacart can accordingly plan to hire persons for delivery during days shifts.

1. On which day do people order the highest selling products?

A screenshot of a cell phone

Description automatically generated

This analysis shows on what day of the week are the products being sold the highest, clearly Sunday (which is represented by “0”) and Monday (which is represented by “1”) are when there is highest sale in the products which means that this analysis shall help the retailers restock their stock by Saturday (which is represented by “6”) night for better sales and growth.

1. When do customer reorder since prior order day?

A picture containing screenshot

Description automatically generated

This analysis shows us the usual reordering frequency of consumers who order through Instacart, looking at the Bar plot above, we can deduce that people seem to order more often after exactly 1 week and once a month (peaks at 7th and 30th day of a month) which shows weekly intervals.

1. What are the top 10 maximum sold product from market basket?

A screenshot of a cell phone

Description automatically generated

From the above horizontal bar plot, it can be clearly stated that “Bananas” and “Bag of Organic Bananas” stands out as the top products in terms of highest sales. So, we can recommend Instacart to always maintain the stock of these products as they are more in demand.

1. What are the top 20 aisle and department by variety of products?

From the bar plots of top 20 aisles by variety of products, it can be deduced that Missing, candy chocolate, Ice cream ice and vitamin supplements aisle are the aisles with maximum variety of products. Moreover, it can be inferred from the below bar chart of department by variety of products offering that Instacart has maximum number of product offerings across personal care and edible item departments.

A screenshot of a cell phone

Description automatically generated

Based on above done analysis of the data set we can recommend the best product fitting the consumers need which will save their time as well as efforts also, not only it will help the consumers for easy checkout but also by predicting this user based purchase, retailers could run an inventory check stocking the most used department to retain their consumers which would act as a source for their commercial and organizational benefits. Effective recommendations are a valuable service to the customers and a profitable service to the retailer, so to help the customer have a hassle-free efficient shopping experience and for retailers to increase their sales and manage their existing customers. We can recommend items to a user which are most popular among users of same type, also we will divide users to multiple segments based on their preferences and recommend items based on where they belong to.

**MODEL SELECTION**

In order to achieve this, I applied logistic regression model to examine the relationship between one or more independent variables on a dependent variable. Regression analysis provides detailed insight that can be applied to further improve products and services. I made two models using logistic regression taking account of two predictors: like the length of time since the previous order and the fraction of previous orders on which the item was bought for 1st model and which user reorder or what product has the best chance for reorder, or which particular order is repeated for 2nd model. Also, I calculated F1 score which is a measure of a test's accuracy in statistical analysis of binary classification. It considers both the precision p and the recall r of the test to compute the score. Furthermore, to check or visualize the performance of the multi-class classification problem, I used AUC (Area Under the Curve) ROC (Receiver Operating Characteristics) curve. It is one of the most important evaluation metrics for checking any classification model’s performance.

**DATA PRE-PROCESSING**

The raw data is spread across 7 tables and isn't amenable to standard classification algorithms. To create a tractable data set, for 1st model there was a need to get the data into a particular form to do this; we need to make a data frame with the following columns: order\_id, previous items ordered by the user, fraction of previous orders the user bought this item, how many days was it since the previous order? and did the user buy it this time yes/no?

For 2nd model I created reproducible random sampling of the orders and order\_product datasets. 40000 users are randomly chosen (~20% sample). Before sampling: orders dataset has 3.4 million rows and order\_products have 32.4 million rows. After sampling: orders dataset has 660 thousand rows and order\_products have 6.23 million rows.

**INSIGHTS FROM THE 1st MODEL**

For this model, we have to make a data frame which, for every order in the training set, lists all previous orders by the user and their probability. Next, we have to train a logistic regression model with days\_since\_prior\_order and frac as the regressors. Since, the model object is huge, and we only need the coefficients, so we’ll just save those and do the predictions.

A close up of a logo

Description automatically generated

Now, this model outputs an updated “probability” of a certain item being in an order which is an increasing function of the fraction of prior orders and a decreasing function of the time since the previous order was made — the longer a user waits between orders, the less likely it is that they will buy something!

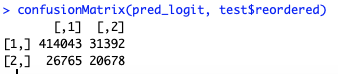
Using these estimates of the regression coefficients, I’ve defined a function to give the response for a given value of frac and days\_since\_prior\_order. I made a plot of the response as a function of frac.



It shows how it varies with the number of days—the impact of days\_since\_prior\_order is small but not negligible. Now, with this model trained, we can make predictions on the test dataset. We need to get a data frame which has the test order numbers with frac and days\_since\_prior\_order as its columns.

**INSIGHTS FROM THE 2nd MODEL**

For 2nd model, I used 20% of the total data and after loading these files I cleaned the data by changing the variable types. I build the model only on the prior dataset and to cross-validate it. I used train dataset results. We are going to want data specific to each user, each product, and each user-product interaction. For this I created my own dataset. For example, adding user specific features and finally adding the `test` and `train` orders for each user to the users table to be used later for predictions and in separating of cross-validation and test prediction sets. I joined all of the feature tables and currently only considering possible products that a user can order from the products they have already ordered. Finally created training and testing data for running logistic regression model on the above-mentioned data frame. Below is the confusion matrix which helps in visualization of the performance of logistic regression model.



Plotting Logistic Regression Model



**FEATURE EXTRACTION ANALYSIS**

I performed user clustering to categorize users based on their shopping behavior. i.e., Can we cluster users into different groups based on the products they ordered, the time of their orders and if they reordered or not? For this I will apply K-means, agglomerative clustering and affinity propagation were applied to various combinations of features. I will use K-means clustering to cluster the users. I choose K-means clustering because it provides a clear and straightforward visualization of the groups. Also, I can adjust how many groups I want. The variables used to cluster users include what products they ordered, during what time of the day they placed the order, the order of products they added the cart, if the user has reordered and how many days the user reordered. I plan to choose to categorize the users into five clusters, which we directly used as a feature for our model. But before plotting it we need to do additional rework on the dataset like the dataset is too large. Therefore, I only included the first 1000 rows in the sample data. Also, there is a need of normalizing the variables before clustering.

I plotted product\_id on the x-axis and order\_hour\_of\_day on the y-axis.



Then I plotted add\_to\_cart\_order and days\_since\_prior\_order.



Next, I tried hierarchical clustering (bottom\_down) for the users. I chose this method because it provides a good visualization of the different levels and how the clusters are made. The variables we are going to use to cluster is reordered users.

Clustered based on the unnormalized variables.

A screenshot of a cell phone

Description automatically generated

Clustered based on the normalized variables.

A picture containing tool

Description automatically generated

Even though the hierarchical clusters are very dense, I can still compare the differences between the unnormalized and normalized one. For the unnormalized graph, it is clear that there are a few dominant variables that divide the groups. On the other hand, for the normalized graph, it's a combination of many little things. Overall, the hierarchical cluster are hard to read because one user has many characteristics and it is hard to plot all of them in one clear tree. The hierarchical cluster might be more useful for products other than users. The K-means gives a good and clear clustering of the users. Furthermore, the purpose of using K-means and Hierarchical clustering techniques is to find out which is more suitable for clustering the users and products into categories.

**WHY MODELLING METHOD IS IMPORTANT?**

Predictive modeling offers the potential for firms to be proactive instead of receptive. It uses transactional data to create particular challenges which need to be carefully addressed to develop valuable models. Classification, prediction, and to some extent, affinity analysis constitutes the analytical methods employed in predictive analytics. By using the analyses made in this assignment, we would be able to help and recommend our customers that which product they should buy with the product already present in their cart. For example, users buying product like Baby diapers from the baby department would also like to buy baby shampoo or other products for a baby. So, a recommendation system predicts what a user is likely to buy based on what they have brought earlier. Also, leading retailers can drive more profitable advertising and promotions, attract more customers, increases the value of market basket and much more. Consumers, planners, merchandisers and store administrators have started to recognize how this new era of easy-to-use market basket analysis tools helps to work more intelligent and compete more successfully.

**REFERENCES**

Instacart Market Basket Analysis. (2017). Retrieved from <https://www.kaggle.com/c/instacart-market-basket-analysis>

Tech @ Instacart. (2017). 3 Million Instacart Orders, Open Sourced. Retrieved from <https://www.instacart.com/datasets/grocery-shopping-2017>