Exploratory Analysis on E-Commerce Sales

Finding product associations and optimizing channel sales

Aditya Wakade[1]
wkaditya@uw.edu
iSchool, University of Washington

Harsh Dev[2]
harshdev@uw.edu
iSchool, University of Washington

Manasi Kulkarni[3]
manasik@uw.edu
iSchool, University of Washington

Manika Nangia[4]
manika2@uw.edu
iSchool, University of Washington

Prem Shah[5]
prems2@uw.edu
iSchool, University of Washington

***Abstract*— Flavors of My City is an Indian online retail business focusing primarily on delivering region-specific delicacies to other parts of India. Their business model is based on creating tie ups with third party vendors of a specific region, so that they can deliver delicacies to other states/regions. As they are scaling up in terms of products, customers and services, their data is increasing massively. In such a scenario, we strongly feel that statistical analysis of their data can help find insights which will enhance the older business model. We decided to take the data that came primarily from four segments: Products, Customers, Regions and Web Traffic.**

Typically e-commerce datasets are proprietary and consequently hard to find among publicly available data. This dataset contains real world data. We will not only perform our analyses but also advise the 'Flavors of my city', with our findings and provide recommendations for scaling up their business. Our research intends to understand trends and patterns between different characteristics of their data and develop inferences from them . These can help the company improve their supply chain and better manage their inventory. This analysis would also give us an idea of the shopping patterns, on the basis of which, the business can derive smarter promotions. We performed exploratory and predictive analytics on transactional data provided to us from the business owners of Flavours of My City and on the basis of our analysis we came up with the following recommendations:

1)Diverting metropolitan customers to online channel.
2)Improving sales and marketing strategies for juices.
3)Predicting inventory & maintaining a smart catalogue for festivals.
4)Driving sales with personalized recommendations.

***Keywords*— market basket analysis, flavours of my city, sales analysis**

I. Introduction

To boost e-commerce sales an online retailer must understand its customer buying patterns. This not only helps in recommending right products to the customers at the right time but also helps in maintaining a smart inventory that adapts to seasonal changes in buying patterns. The transactional data collected by an organization can be mined and explored to build a strong understanding of the business in terms of predicting sales, releasing promotional offers for specific items and for specific regions , and building a recommendation engine to drive sales and profitability.

As different aspects of the above mentioned objectives require specific type of data science analysis we have chosen to do the following :

1) Perform an exploratory analysis to understand trends & patterns in buying behaviour.
2) Build association rules for products that sell together.
3) Understand and evaluate the factors that contribute towards higher bill value.

One of the key techniques that we have used is called Market Basket Analysis, which is a is a data mining method focusing on discovering purchase patterns of the customers by extracting association or co-occurrences from a store's transactional data. For example, when the person checkout items in a supermarket all the details about their purchase goes into the transaction database. Later, this huge data of many customers are analyzed to determine the purchasing pattern of customers. Also decisions like which item to stock more, cross selling are determined.

Another technique that we have used falls under predictive analysis and required us to build a logistic model to find out relations of categorical variables such as 'Price' 'Category', location etc with respect to 'Bill Amount' to draw observations which can be used to implement business strategies.

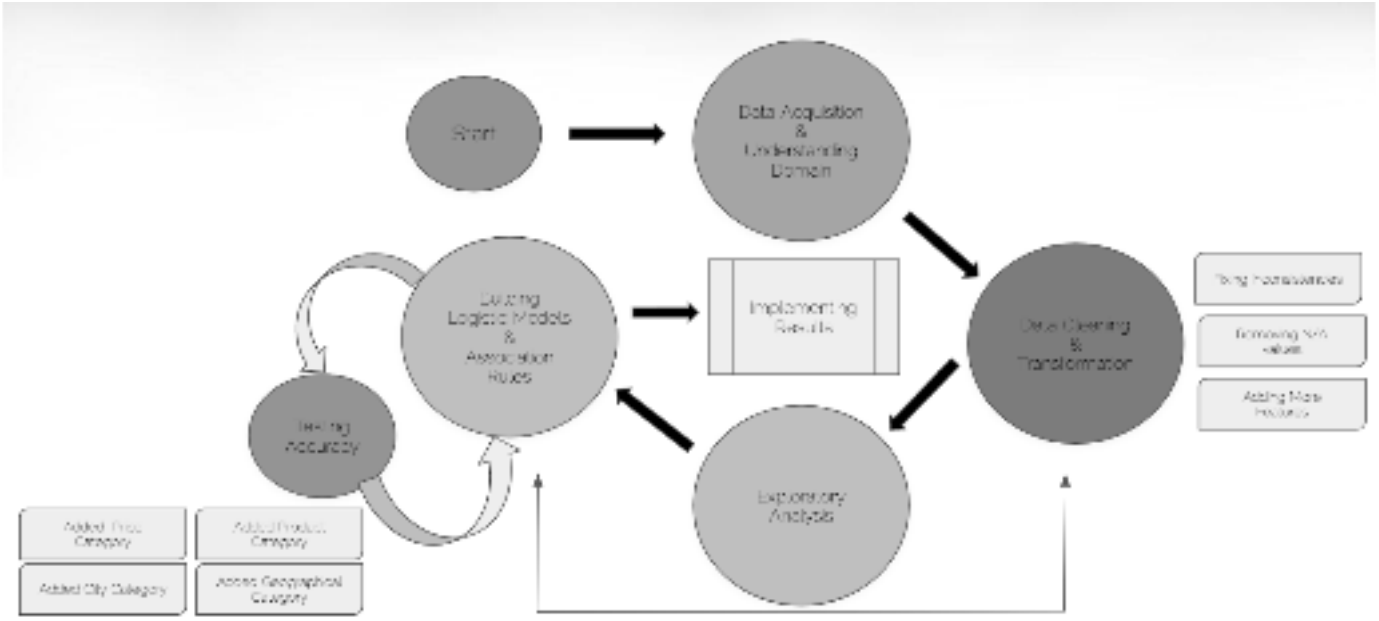
The methods and the results of all our analyses will be discussed in the subsequent sections.

II. Project Questions

We wanted to chart an exploratory and predictive route for this project, and with the limited time and resources that we had, we reached a consensus in the team on the following questions that we wanted to answer through our analysis. The answer to these questions would help us recommend actionable changes to the business owners to drive the sales up.

2014,2015,2016, 2017 and 2018 (till February) . We augmented existing data by adding the following columns - Geographic Location (North, South, East, West) , 'Price' 'Category' (1 <- Rs 0 - 99 , 2 <- Rs 100 - 199 & 3 <- for greater than Rs 200).

These features were added to analyze the effect of 'Price' categories and geographic location on bill value.



1. What is the association in terms of buying patterns of products? Which product is bought most frequently & what is the order?

2. Is the sale of organic products more in Metropolitan cities compared to non-metropolitan cities?

3. Do the sales increase seasonally? Specifically, by any occurence of an Indian festival?

4. How does the product ('Category', 'Price', 'Quantity') affect the 'Bill Amount'?

III. Scope & Limitations

This project is limited to the 236 unique products, the data of which was provided to us by the business owners. Also, we did not perform any time series analysis as the data that was provided to us was only split by month and year and not by day/hour/min.

IV. methods

The results that we have presented in our research is based on analysis done on transactional data provided to the project team by the business owners. The core analysis of the research was done on transactional data of 236 products of years

We followed the following project cycle from the inception of the project. In the first phase the data was acquired and business domain was understood through multiple sessions with business owners. In the next phase the data was cleaned, transformed and features were added post which we moved to the exploratory phase.

Our results, which will be discussed in the consequent sections are dependent on the following methods of data analysis.

Association rule mining - Association rule mining is primarily focused on finding frequent co-occurring associations among a collection of items. It is sometimes referred to as “Market Basket Analysis”. (Bone, 2014) .

Predictive Modeling : Predictive modeling is a process that uses data mining and probability to forecast outcomes. Each model is made up of a number of predictors, which are variables that are likely to influence future results. (Rouse, 2016).

V. Findings & Analysis

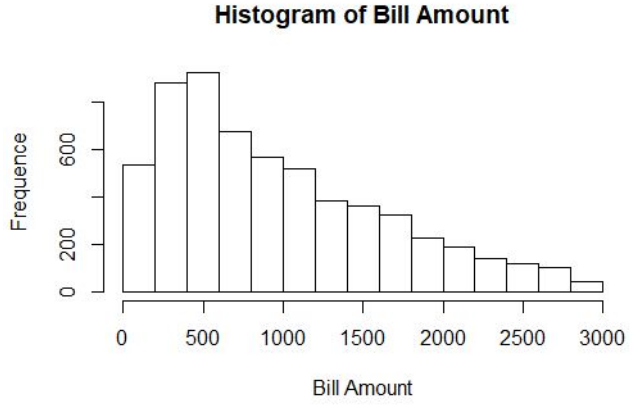
Our findings and analysis can be divided into two parts, exploratory and predictive analysis.

A. Exploratory Analysis

Our first step for exploring the data was to find out the nature of data offered in every column. We identified the data types of every column along with their range and central tendencies using the summary function in R.

```
summary(fomc)

##       X.1             X              Nos.          order_id
##  Min.   :   1   Min.   :   1   Min.   :   1   Min.   :  1
##  1st Qu.:1501   1st Qu.:1501   1st Qu.:1501   1st Qu.:126
##  Median :3000   Median :3000   Median :3000   Median :254
##  Mean   :3000   Mean   :3000   Mean   :3000   Mean   :252
##  3rd Qu.:4500   3rd Qu.:4500   3rd Qu.:4500   3rd Qu.:378
##  Max.   :6000   Max.   :6000   Max.   :6000   Max.   :500
##
##           Item       Item.Name                                  :2889
##  Product 485:  68                                               :  68
##  Product 236:  59   Laxminarayan's Poha Chiwda (Pack of 2):  68
##  Product 173:  58   Aloo Mixture Namkeen                    :  59
##  Product 36 :  58   KC Das's Rossogullas                    :  58
##  Product 93 :  56   Vajani's Methi Khakra                   :  58
##  Product 341:  41   Kala Khatta                             :  56
##  (Other)    :5660   (Other)                                 :2812

##     Category      Category.1       Price             City
##         :2579             :2579   Min.   : 50.0   Mumbai   : 427
##  Snacks :1180   Snacks :1180   1st Qu.:111.0   pune     : 220
##  Sweets : 866   Sweets : 866   Median :179.0   Delhi    : 203
##  Pickle : 313   Pickle : 313   Mean   :174.9   Kolkata  :  97
##  #N/A   : 312   #N/A   : 312   3rd Qu.:241.0   Bengaluru:  76
##  Spices : 216   Spices : 216   Max.   :299.0   Chennai  :  75
##  (Other): 534   (Other): 534                   (Other)  :4902

##    City.1          Quantity        Bill.Amount       State.Of.Delivery
##  Mumbai   : 427   Min.   : 1.000   Min.   :  50.0   Maharashtra   :1113
##  Pune     : 220   1st Qu.: 3.000   1st Qu.: 420.0   Tamil Nadu    : 616
##  Delhi    : 203   Median : 5.000   Median : 796.0   Andhra Pradesh: 588
##  Kolkata  :  97   Mean   : 5.518   Mean   : 966.3   Kerala        : 546
##  Bengaluru:  76   3rd Qu.: 8.000   3rd Qu.:1404.0   Gujarat       : 455
##  Chennai  :  75   Max.   :10.000   Max.   :2990.0   Karnataka     : 454
##  (Other)  :4902                                     (Other)       :2228

##   State.of.Delivery        Month           Year         pricecateg
##  Maharashtra   :1113   February :1006   Min.   :2014   Min.   :1.000
##  Tamil Nadu    : 616   October  : 862   1st Qu.:2015   1st Qu.:2.000
##  Andhra Pradesh: 588   March    : 828   Median :2016   Median :2.000
##  Kerala        : 546   July     : 609   Mean   :2016   Mean   :2.191
##  Gujarat       : 455   January  : 586   3rd Qu.:2017   3rd Qu.:3.000
##  Karnataka     : 454   September: 359   Max.   :2018   Max.   :3.000
##  (Other)       :2228   (Other)  :1750
##

 geogloc          item_id
##  Center:  58   Min.   :  1.0
##  East  : 578   1st Qu.:131.0
##  North :1395   Median :254.0
##  South :2302   Mean   :255.1
##  West  :1667   3rd Qu.:380.0
##                Max.   :500.0
##
```

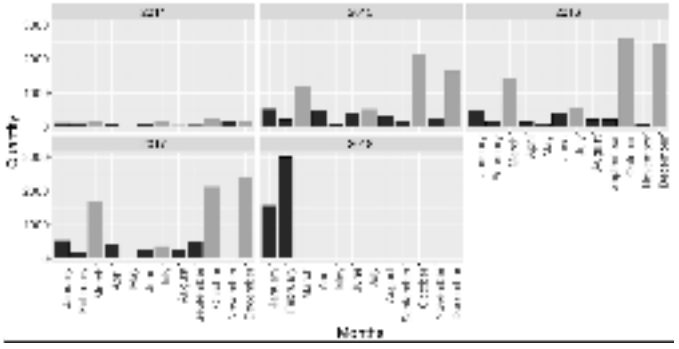
Our dataset has only one numeric feature, which was the 'Bill Amount' which changes for every order. 'Quantity' and 'Price' had fixed numeric values. All other features are categorical or qualitative. Hence, we decided to plot a histogram for 'Bill Amount'.



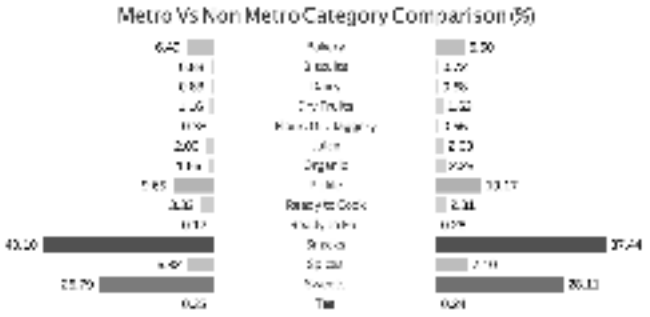
The histogram shows a skewed distribution to the right, as shown below. A distribution skewed to the right is said to be positively skewed. This kind of distribution has many occurrences in the lower value cells (left side) and few in the upper value cells (right side). It causes the mean and median values to shift to the left (PQ Systems). We went ahead to try find relations between attributes which we felt can be related based on our logical understanding of the business. We then tried to find buying patterns over geographical regions which shows us the average 'Bill Amount' of every state and number of orders delivered to every state. We came to know that the average 'Bill Amount' was almost the same for every state but South region accounted for a lot of orders.

mean.fomc.Bill.Amount..fomc.geogloc....elem1....na.rm...T.	nrow.fomc..fomc.geogloc....elem1.....	elem1
956.598080383923	1667	West
970.334926151173	2302	South
969.886021505376	1395	North
969.074394463668	578	East
972.810344827586	58	Center

We also found seasonal trends in the dataset. We can see that the months March, July, October December account for increased sales. We speculate that this could be because major Indian festivals fall in these months (Holi, Janmashtami, Diwali and Christmas).



Another interesting finding in our exploratory analysis was found after comparing sales of different product categories in metropolitan and non-metropolitan cities.



We found that despite most categories accounting equally for sales in metropolitan and non-metropolitan cities, organic foods accounted for higher percentage of sales in nonmetropolitan places. We felt that the reason behind this could be that People in metro's might have organic food

readily available (which we don't know), so they tend to buy region specific sweets and snacks which are not readily available in metro.

B. Association Rule Mining

We performed association rule mining to understand different product dependencies in order to facilitate better recommendations. Instead of looking for a particular rule specifically, we were looking for the strongest rules in the data in order to help the company provide their customers with better recommendations as well as boost sales.

We also looked at the best associations for each region (North, West, East, South) in order to understand each region's buying patterns.

1) Process: To perform the association mining, we used Python and specifically the *mlxtend* library in it. First we took the product IDs for each order and the order IDs. We first constructed a one-hot matrix which encodes categorical integer features into a matrix. The matrix has the order IDs as rows and product IDs as columns. As you can see in the figure below, each cell is a value 1 if the product is bought in the order, otherwise it is a 0. This matrix is popularly known as a 'basket' in market basket analysis.

This matrix is taken as an input to an apriori rule making function which takes into account minimum support and confidence. We define support and confidence as follows:

Support: Support is defined as the percentage of transactions which contain both the products i.e. the antecedents and the consequents in the association

Support: $P(A \cup B)$

Confidence: Confidence is the percentage of transactions containing the consequent, given that the antecedent is already in the basket (order).

Confidence: $P(A|B)$

For each region we define the minimum support and confidence to get strong rules and also for the overall dataset. We have kept our minimum confidence at 50% and each association mining set (Whole, North, West, East & South) has different minimum support since we want the best rules for each.

Additionally, we look at the lift of each rule. Lift is defined as the ratio of observed support to the expected support. A higher value of lift signifies higher interdependency between the products in the association rule. We discuss the results from the same in the subsequent section.

2) Results: We got the following results by performing association mining on our data. Each section describes the results and what we infer from it.

Whole Data Set

	antecedants	consequents	support	confidence	lift
0	(Peanut Chikki)	(Kala Khatta)	0.029070	0.500000	4.000000
1	(Desai's Talni Mirchi)	(Varhadi Mutton Rasa Masala(Pack of 4))	0.014535	0.800000	8.877419
2	(Kolhapuri Mutton Rasa Masala(Pack of 4))	(Varhadi Mutton Rasa Masala(Pack of 4))	0.023256	0.500000	5.548387
3	(Bikalananda's Khirmohan)	(Aloo Mixture Namkeen)	0.023256	0.500000	3.909091
4	(Kolhapuri Thecha (Set Of 2))	(Laxminarayan's Poha Chiwda (Pack of 2))	0.023256	0.500000	3.127273
5	(Athavale's Amrakhand Wadi (2 packs))	(Vajani's Methi Khakra)	0.034884	0.583333	4.666667

Here, we see that the most important rule is #5 which is between Wadi & Khakhra (both Indian Snacks). They have a support of 0.034884 meaning 3.48% of the transactions include both of these products. And a lift of 4.67 signifies that interdependency of these products is 4.67 times more than expected which shows that it is a strong, rather very strong association.

Now we divide the data according to the regional parts and analyze their association rules.

North

	antecedants	consequents	support	confidence	lift
0	(Athavale's Amrakhand Wadi (2 packs))	(Vajani's Methi Khakra)	0.044944	0.75	6.068182
1	(Sadanand's Dink Ladu)	(Kala Khatta)	0.056180	0.60	6.675000
2	(Tandoori Paste Masala Pk2)	(KC Das's Rossogullas)	0.033708	1.00	6.357143

We observe that the most important rule in the northern region is the same as the rule in the whole dataset.

West

We observe that the most important rules contain items (products) which are local to the western region. This might be due to the fact that some specialties are available in some cities only, and people from other cities in the western region might want to order them.

	antecedants	consequents	support	confidence	lift
0	(Peanut Chikki)	(Kala Khatta)	0.054945	0.60	4.550000
1	(Kolhapuri Thecha (Set Of 2))	(Laxminarayan's Poha Chiwda (Pack of 2))	0.032967	1.00	5.055556
2	(Athavale's Bhajani Chakali)	(Kala Khatta)	0.065934	0.50	3.791667
3	(Xacuti)	(Vajani's Methi Khakra)	0.043956	0.75	4.875000
4	(Kaka Halwai's Kaju Gajak)	(Vajani's Methi Khakra)	0.054945	0.60	3.900000
5	(Takatali Mirchi (Pack of 2))	(Aloo Mixture Namkeen)	0.043956	0.75	4.875000

East

	antecedants	consequents	support	confidence	lift
0	(Mango Pickle (Pack Of 2))	(Caramel Stick Jaw Butter Toffee)	0.129032	0.500000	3.875000
1	(Caramel Stick Jaw Butter Toffee)	(Mango Pickle (Pack Of 2))	0.129032	0.500000	3.875000
2	(Vajani's Methi Khakra)	(Desai's Mango Cubes)	0.096774	0.666667	6.888889
3	(Desai's Mango Cubes)	(Vajani's Methi Khakra)	0.096774	0.666667	6.888889
4	(Athavale's Assorted Wadi (4 packs))	(Athavale's Puran Poli (2 packs))	0.064516	1.000000	10.333333
5	(Athavale's Puran Poli (2 packs))	(Athavale's Assorted Wadi (4 packs))	0.096774	0.666667	10.333333

Here, with a lift of 10.3333, Puran Poli & Wadi serve as the most important items. Predominantly, these are famous delicacies in the western region but turn up more frequently in the eastern region. A plausible explanation might be the non-availability of such delicacies in the eastern region.

South

	antecedants	consequents	support	confidence	lift
0	(Madras Fish Masala (Set of 4))	(KC Das's Rossogullas)	0.039370	0.60	5.861538
1	(Kolhapuri Thecha Red + Green (Pack of 4))	(KC Das's Rossogullas)	0.039370	0.60	5.861538
2	(Ginger Chips (Pack Of 2))	(Kala Khatta)	0.047244	0.50	3.342105
3	(Kolhapuri Tam&PhanRasa Masla(Pack of 4))	(Varhadi Mutton Rasa Masala(Pack of 4))	0.047244	0.50	5.291667
4	(Venkateshwara's Mysore Pak)	(Laxminarayan's Poha Chiwda (Pack of 2))	0.023622	1.00	8.466667
5	(Desai's Hapus Amba Wadi)	(Laxminarayan's Poha Chiwda (Pack of 2))	0.039370	0.60	5.080000

From the above analyses, we draw two interesting inferences:
1. People from the same region order products local to that region sometimes, probably because they are specialties from a different city.
2. Some orders contain products from completely different regions

C. Predictive Analysis

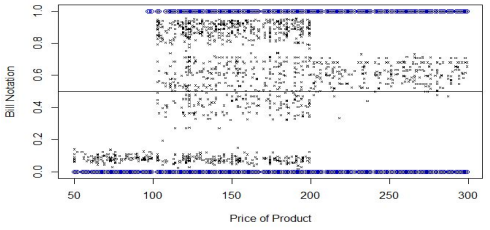
The idea of using predictive analytics started with the need of finding how different parameters in the data affected the 'Bill Amount' of the user's order. The table below shows us all the parameters in the dataset that were available to us.

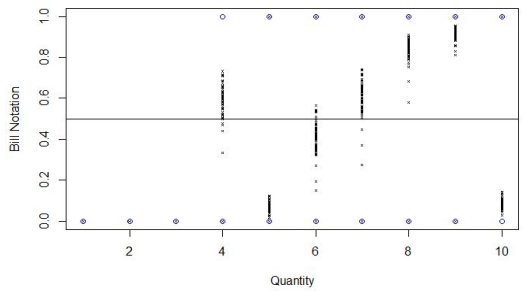
As you can observe, all the data columns hold three types of variables, which are: Ordinal, Nominal, Boolean or Interval. Boolean variables indicate the presence or absence of a property. Ordinal variables hold values which increase or decrease in a fixed order. Nominal variables hold categories or multiple fixed values which are neutral and have no relations between them. Interval variables are stratified ranges of a continuous numerical data (Institute of Digital Research and Education, 2017). In our dataset, ordinal variables would include 'Months' (which increase from January to December), 'Quantity' of Products in an order (which increase from 1 to 10) and Year (2014 to 2018). Nominal variables include 'Category' of products and Geographical Location. Interval variable is the 'Price' range which divides product 'Price's into three categories. The only Boolean variable in our dataset tells if the city is a metropolitan city or not. One can observe that almost all our data is qualitative. Hence, we decided to use a logistic regression to find out how significantly and how reliably different values of these variables affected the 'Bill Amount'.

Price Category				
0-100 (1)	100-200 (2)	200-above (3)		
Quantity				
1 to 10				
Geographical Location				
North	South	Center	West	East
Months				
January to December				
Is the City of the order a Metropolitan City?				
1 if Yes. 0 if No				
Category of Products				
Sweets, Snacks, Organic, Tea & Coffee, Spices, Dairy, Biscuits, Pickle, Ready to Eat, Flour/Jaggery				

To optimize the ''Bill Amount'', we had to decide a threshold to test variables which contribute towards the 'Bill Amount' being higher or lower than this threshold. We selected the threshold to be Rs. 1000 because, firstly, it was very close to the central tendency of ''Bill Amount'' (Rs. 966 mean and Rs. 970 mean), secondly, Rs. 1000 would make a good number for the business to craft offers around. Hence, we added a column in the data which held the value 0 if the 'Bill Amount' was lower or equal to 1000 and 1 if it was higher than 1000.

We decided to find how the product characteristics affect the 'Bill Amount' ('Price', 'Quantity' and 'Category'). We decided to add product 'Category' into the model along with 'Price' and 'Quantity', as it would give us some perspective about nature of the product, which is important to reason why certain values of 'Price' and 'Quantity' contribute towards higher 'Bill Amount'. The model provided us numbers associated with every value of these variables, which would increase or decrease the log odds of having a higher 'Bill Amount'. The results were satisfactory, indicating 'Price' range Rs 100 to Rs 200 increased the odds better than 'Price' range >Rs.200. However, this inference had a large standard error. The 'Quantity' values 9 and 8 significantly contributed for the 'Bill Amount' to be higher than Rs. 1000.





Next, we decided to run a logistic regression which included Month, Geographical location and Metropolitan City as parameters along with the product metadata. The regression gave us similar results with slightly deviated standard errors and odds multiple.

1) Results: We found how every variable affected the log odds of the bill amount being more than Rs.1000. We have tabulated them into two parts, one shows all factors that improve the odds, and other shows all factors that reduce the odds.

Column	Value	Odds multiplied by
Category of product	**Organic**	**0.05**
	Dry Fruits	**0.068**
	Sweets	0.4
	Snacks	0.68
Month	August	0.66
	Septem ber	0.84

The table above shows the significant values which reduce the odds of the 'Bill Amount' being higher than 1000 ie, multiply the odds by numbers less than 1. As we can see, Dry Fruits and Organic foods multiply the odds by 0.05 and 0.068 which significantly influence the Bill to be lower. Other categories of food like sweets and Snacks also play a role in having the 'Bill Amount' lower. Surprisingly, Snacks and Sweets are the most common type of product bought over different regions. This tells us that people tend to buy snacks and sweets in lower quantities (mean 'Price' of both Sweets and Snacks is around Rs. 173). However, the approximate range of 'Prices for Snacks and Sweets is Rs. 55 to Rs. 299, which makes it unreliable to make such an inference. However, we can predict that Dry Fruits and Organic Food reduce the odds of the 'Bill Amount' being higher very significantly. There were other variables like 'Category': Dairy, Pickle and Month: June but were not considered because of low multiplication factor or High Standard Error.

The table above shows the significant values which increase the odds of the 'Bill Amount' being higher than 1000 ie, multiply the odds by numbers more than 1. As we can see, Juice and Spices are foods that multiply the odds by 1.53 and 1.28 which significantly influence the Bill to be higher. We can predict that Juice and Spices increase the odds of the 'Bill Amount' being higher very significantly. There were other variables like Region: East, North and 'Category': Ready to Eat and Month: March but were not considered because of low multiplication factor or High Standard Error.

Column	Value	Odds multiplied by
Categor y of product	**Juice**	**1.53**
	Spices	1.28
Month	**October**	**1.4**
	Novembe r	1.23
Region	**South**	**4.3**
	West	3.7

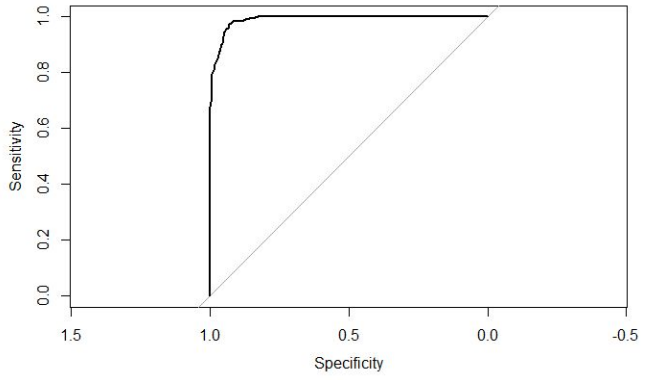
One interesting finding was that, if we compare the mean 'Bill Amount' of orders with 'Category' Juice and Spices, Juice accounts for mean of Rs. 985 and Spices account for mean of Rs. 1095. This was interesting because Juice indicated better odds than spices but accounted for lower mean 'Bill Amount'. On further analysis we discovered that the total orders with 'Bill Amount' greater than 1000 were more than total orders with 'Bill Amount' lower than or equal to 1000 for Juice. However, total orders with 'Bill Amount' more than 1000 were less than total orders with 'Bill Amount' lower or equal to 1000 for Spices. This was the probably the reason for better odds of Juice. Spices see orders with much higher 'Bill Amount' than Juice. If the threshold of the model was higher, spices would have performed better. Hence, we feel offers for spices should not be made with this threshold. The business

can go ahead to design offers for spices which target 'Bill Amount's much more than 1000.

Note: We did not consider any values which provided a multiplier of odds between 0.9 and 1.1, as it would hardly cause any significant change in the odds ratio. Moreover, because of a finite standard error associated with that value, the inference would be unreliable.

The graphs which plot the optimal 'Price' and 'Quantity' predict range of 100-200 and 'Quantity' 8 and 9 for optimal Bills but involve significant standard error which makes them unreliable.

2) Evaluation of the model: We evaluated the model by finding the percentage error and finding the ROC and AUC values. We calculated the percentage error in the model, which is (false positives + true negative)/all outcomes. It calculated to around 6% (5.889%). We also found the ROC and AUC values for the model. An ROC curve is a commonly used way to visualize the performance of a binary classifier. It is a plot of the True Positive Rate (on the y-axis) versus the False Positive Rate (on the x-axis) for every possible classification threshold. Basically, assesses the model's performance in predicting positive value as positive and negative values as negative. To quantify the performance of the classifier, we use AUC which stands for Area Under the Curve. AUC is the percentage of box that is under this curve (Data School, 2014).



We divided the whole dataset into a 80:20 proportion. We used the 80 part as training data and 20 part as testing data. We ensure that the training data is not skewed in terms of having significantly larger value of Bills higher than 1000 or lower than 1000. We iterated this process over the dataset picking different parts as 80 and 20. The range in which we got the area under the curve was 0.97 and 0.98. We tried to tackle chances of overfitting my performed the evaluation on more test and training datasets created in different proportion (60:40 and 70:30). However, they led to similar kind of results. We speculate the overfitting problem to be caused due to lack of user data. Right now, the model is encountering similar transactions due to lack of user id associated with the transaction. Once the user data is added, the model will be able to further classify based on which user bought which set of products while order was placed. We were not given user data by the company because of their privacy policy.

Moreover, the logistic model aligned well with our exploratory analysis by not predicting any conflicting results. The model predicted 'South' as a factor which improves the odds and months 'November' and 'October' as factors that improve the odds. Also, 'August' and 'September' are predicted as factors that reduce the odds.

VI. Recommendations

As we were dealing with a real dataset which had a real business associated with it, our analysis was more of a business project. We had to present recommendations to the business on the basis of the analysis that we had done. We carried out tests to target three specific areas of the business: Inventory, Product, and Customers. On the similar lines, we targeted a specific business function, i.e Marketing, and Service. Based on the logistic regression and the association mining, we recommend the following strategies to start/change in their current process.

A. Diverting Metropolitan customers to the Online channel for Organic products

We compared the percentage sales of Metropolitan and Non-Metropolitan cities and observed that the sale of organic food was more in the non-metropolitan cities. Also, from the look of the analysis on a granular level, we found that sweets and snacks have a similar sales pattern. We based our conclusion on the assumption that, metropolitan cities, being more developed than non-metropolitan cities will have the organic foods readily available. To boost its sales in the metropolitan cities, Flavors of my City can possibly create third-party tie-ups with the local shops dealing with organic foods. In this case, there is a window of opportunity for the business to boost the sales for the local shops whilst also increasing their sales on the online channel. In fact, Flavors of my City could also try to effectively create customer segments (Bell, 2016) in terms of understanding the customers in two city segments, their choices of products and the time they are buying. Precisely, they could channel the segments into two categories based on the city categories and particularly try to

improvise their sales during the non-festival seasons for the organic products.

B. Improve Sales and Marketing Strategy for Juices

We wanted to find out the relationship between different categorical variables and their effect on the Bill Amount. For this, we set the threshold amount of the Bill as 1000 and ran a logistic regression to determine the log odds. Two specific observations were recorded in this analysis. The odds for the bill amount to be greater than 1000 for Juices increases by 50 percent while the odds for the bill amount to be more than 1000 was reduced by 95 percent for Organic foods. Also, the region that came primarily in the picture was South specifically during the month of October. Particularly, for October, in the South region, the business can experiment with selling different types of juices to confirm the analysis and increase the breadth of the product sales. We also saw that the odds of bill amount being greater than 1000 decreased with Organic foods. We recommend the business to not give offers on bulk quantities of organic and dry fruits.

C.Predict Inventory and maintain a smart catalog for festivals

When we found out the seasonal patterns of sales, we figured out that the sales increased marginally during the festival months, specifically during Diwali, Christmas, New Year and Holi. This could be a valuable insight to the business in terms of managing inventory. Also, we found out that perishable items account for the maximum inventory. In this case, we recommend the business to stock up their inventory, particularly with non-perishable items during the non-festive seasons. It opens up the window of opportunity in terms of increasing sales during the festive seasons, considering that the sweets and snacks are non-perishable items.

D. Drive sales with personalized recommendations

Based on our logistic regression models and association mining models, we recommend Flavors of my City to build a personalized recommendation system that increases their sales. This can be done by observing the patterns of sales and trends in regions after implementing the Market Basket Analysis technique as demonstrated in the performed analysis and association mining.

VII. Future Work

We have spent more than two months to complete the analysis until this point and provide the recommendation to the business. Our data was real data and it came from real customers and involved real orders. The inconsistencies that came with it were massive. Given the short time and the extensive scope of the problem statement, we spent a lot of time cleaning and organizing the dataset. We further plan to perform customer segmentation using PCA and determine the target strategies for an individual segment by performing logistic regression using different combinations of segments. We were currently limited by data, but we expect to get the reorder data for every product. This, in conjunction with the customer data, can help us perform analysis and derive strategies for loyal customers, star customers, the probability of reordering etc.

VIII. Conclusion

Our aspect of analysis that stood out the most was the specific time of October that brought in the maximum sales. In our analyses, we have shown the different categories of products that were bought together and the pair that stood out the most, in terms of category, was sweet and spicy. We also concluded from our analyses that South region accounts for the maximum of sales and people in non-metropolitan cities buy organic food more, but that does not mean that they are more health conscious. To summarize, Flavors of my City should make a clear marketing strategy that segments customers based on regions than simply performing customer segmentation on buying patterns and orders. This is essential because the pillar of the business model is based on region-specific delicacy delivery.

Acknowledgements

Working on this project, finding analyses and developing a paper would not have been possible if we wouldn't have had the dataset from Flavors of my City. Ms Shanti Nair supported us with providing their real-world data and trusting us with their sensitive customer information. Many people have helped us to improvise on our research. Mr Sangeet Agarwal the Machine Learning Head at Flavors of My City for providing a preliminary analysis of the research they were already conducting. Based on their feedback, we improvised and focussed our research on another domain so that it could be beneficial to the organization.

REFERENCES

[1] MargaretRouse(2016),http://searchdatamanagement.techtarget.com/definition/p redictive-modeling

[2] Dr. Kirk Borne (2014), https://mapr.com/blog/association-rule-mining-not-your-typical-data-science-algorithm/

[3] PQ Systems. (n.d.). Histogram: Study the shape. Retrieved from PQ Systems: http://www.pqsystems.com/qualityadvisor/DataAnalysisTools/interpretation/histogram_shape.php

[4] Data School. (2014). *ROC curves and Area Under the Curve explained* . Retrieved from Data School: http://www.dataschool.io/roc-curves-and-auc-explained/

[5] Institute of Digital Research and Education. (2017). *WHAT IS THE DIFFERENCE BETWEEN CATEGORICAL, ORDINAL AND INTERVAL VARIABLES?* Retrieved from UCLA Institute of Digital Research and Education: https://stats.idre.ucla.edu/other/mult-pkg/whatstat/what-is-the-difference-between-categorical-ordinal-and-interval-variables/

[6] Bell, E. (2016, June 29). *Customer Segmentation: 5 ways to divide the consumer base*. Retrieved from The Bridge Corp: http://www.thebridgecorp.com/customer-segmentation/