Instacart Case Study

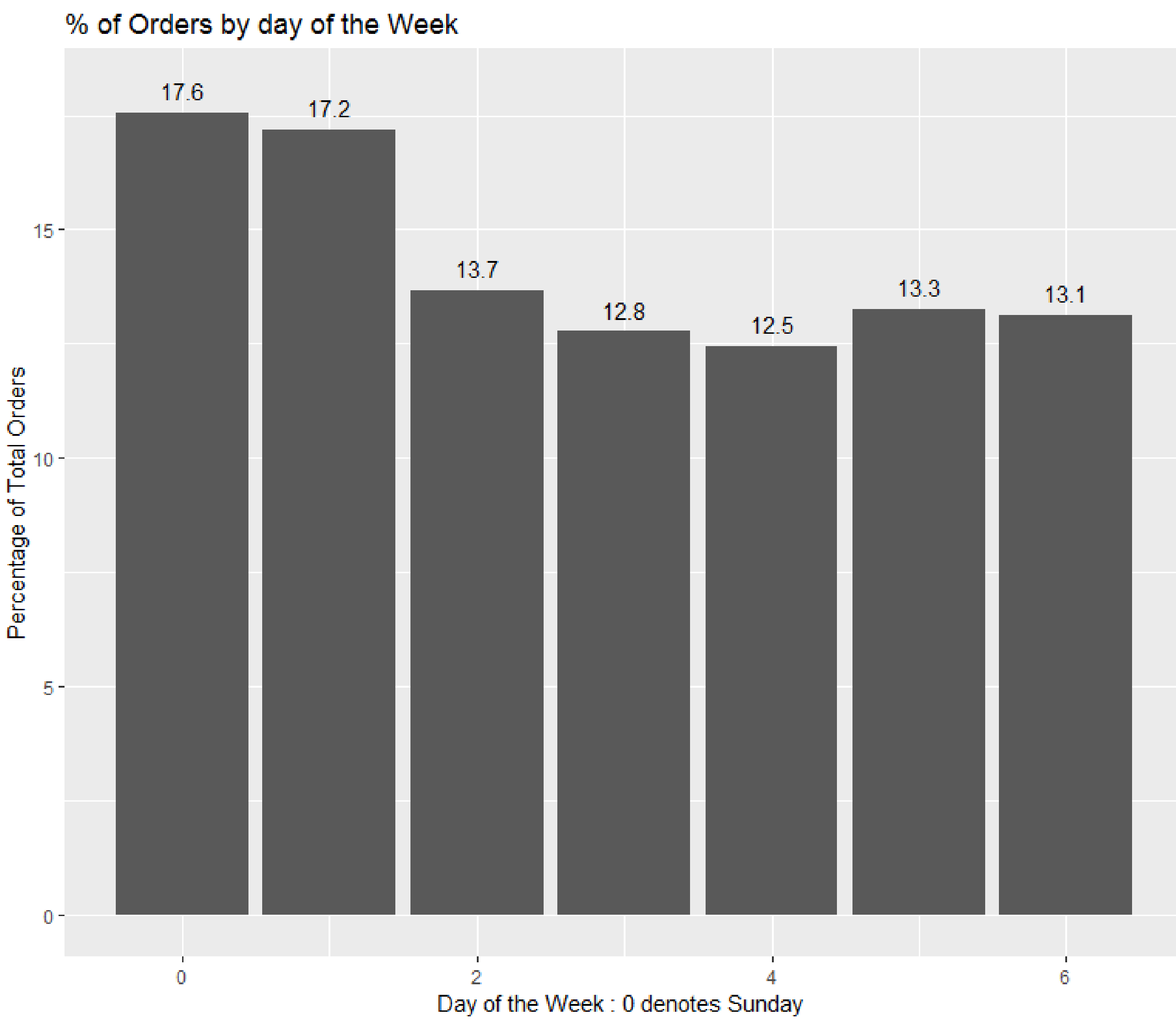
Chirag Alawat (chirag.alawat@gmail.com)

Objective - The aim of this case study is to predict products which are ordered frequently by the customers based on their purchase history. Every month thousands of customers purchase from the various store of instacart varying from frequent customers to the ones who came in once, through the power of analytics we can predict most frequent purchased products and have impactful results to furthermore improve the service and analyze the week sections.

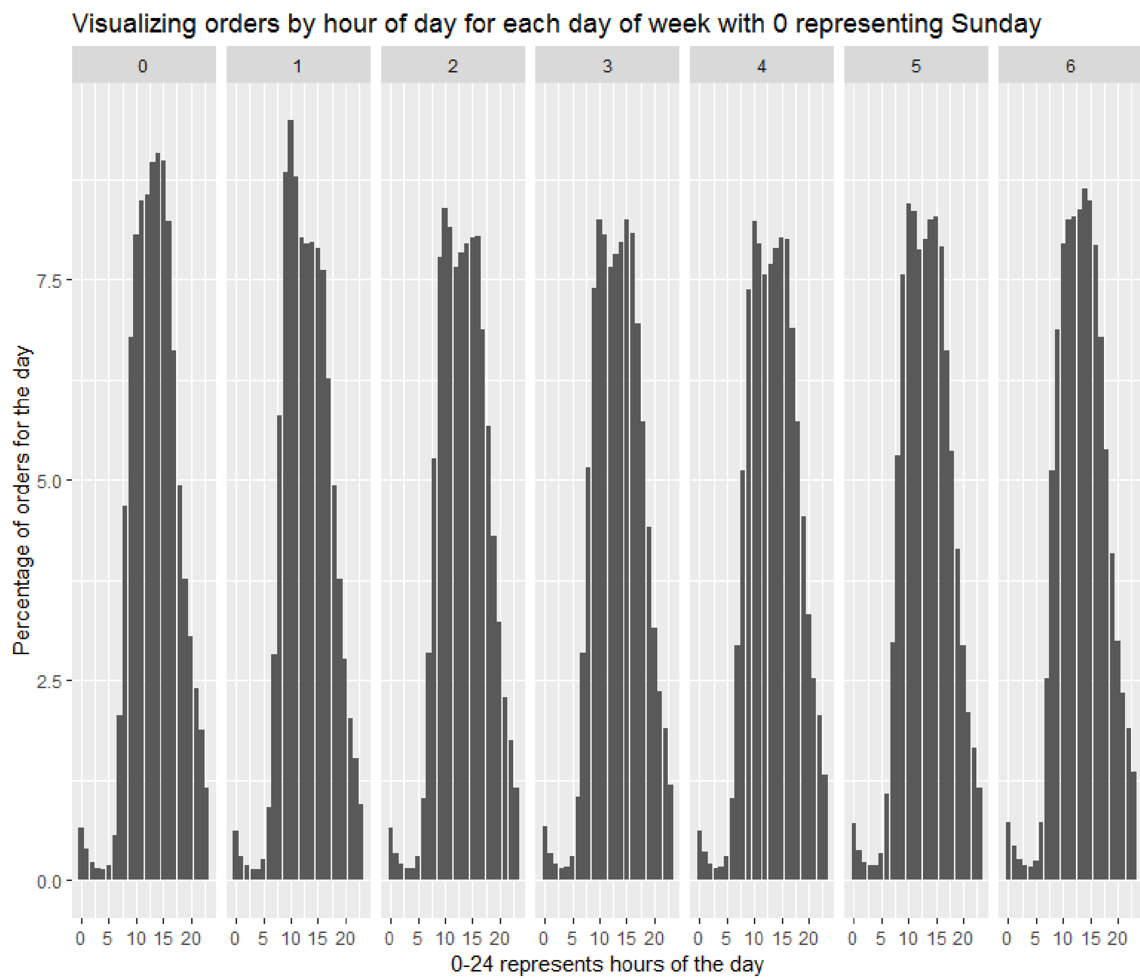
Limitations - having said that there are few limitations of this case study, like predicting re-orders won't give us any insight on what is the current trending product in the market and absence of age of the customer in the data have the major effect in the context of the type of purchased products.

Exploratory Data Analysis:

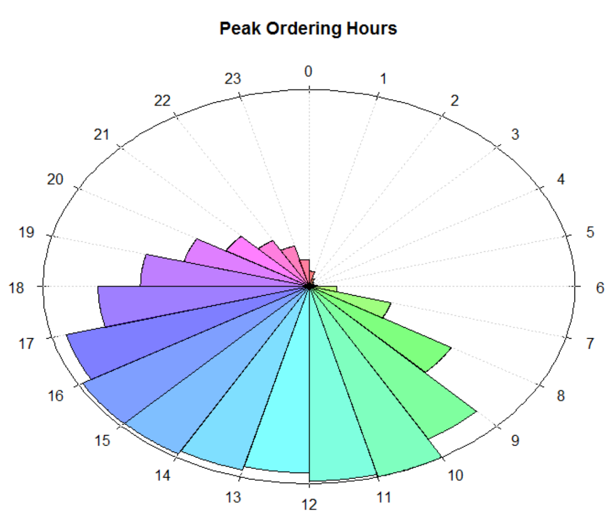
1) Order_dow - Which day of the week purchase was done denoted by 0-6 i.e. Sunday to Saturday, given below is plot stating Sunday and Monday are the busiest days in the week.



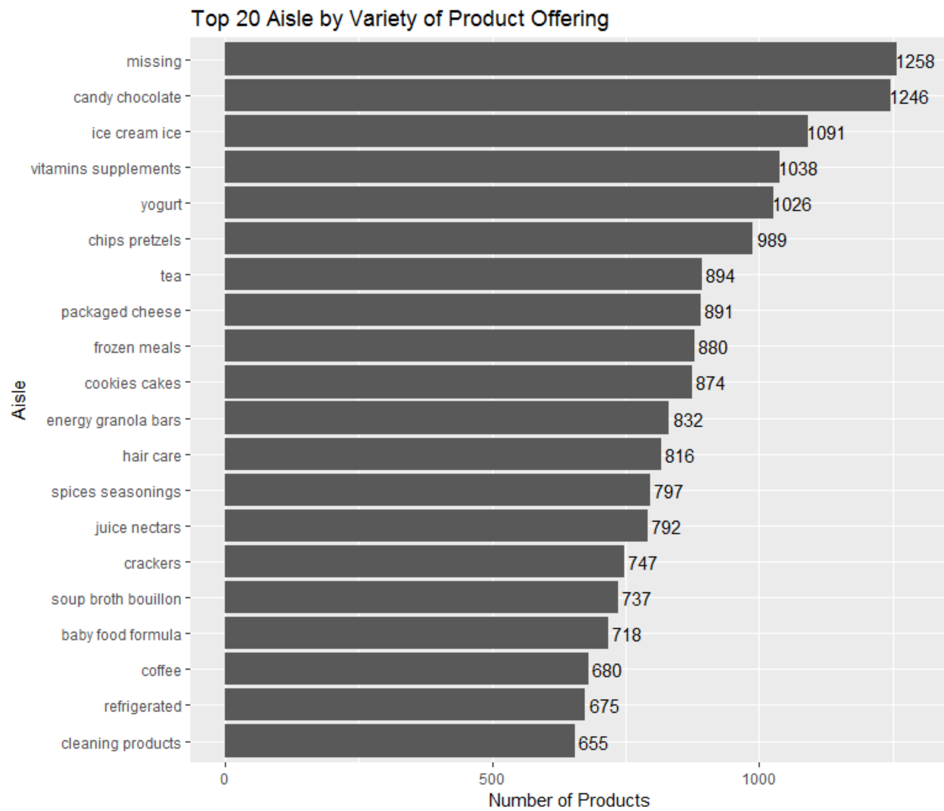
2) Order_hod_dow - this shows orders every hour of the day and every day in the week, given below is the visualization which states that on Monday around 10 AM has the highest sales.

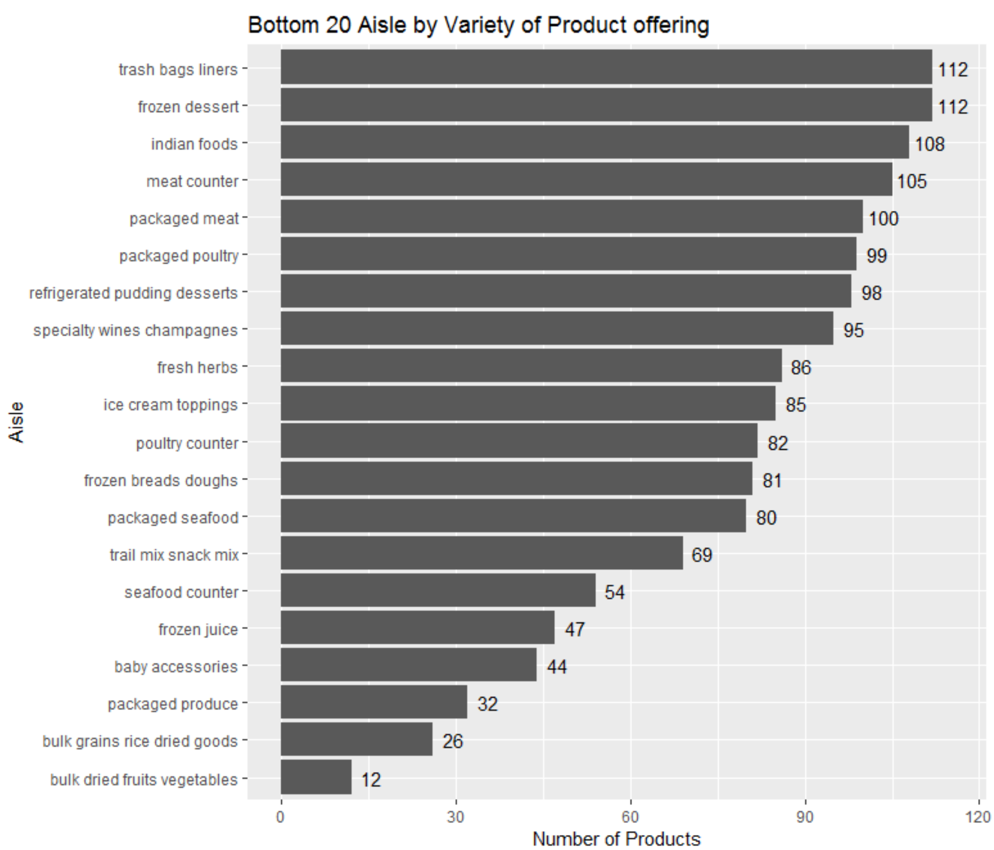


3) Order_hod - This shows orders every hour of the day, given below is the clock visualization which states that from 9 AM to 6 PM there is the highest footfall.



4) Aisle - Top and bottom 20 Aisle representation:





5) Bulk orders – Usually people buy bulk products either on the 7th day and on 30th day of their prior purchase.

Data Pre-processing:

1) Feature Engineering: I have worked on important variable and introduced some variables :

• Reorderchance – Chances in terms of probability that the customer will re-order

• Usertimesreordered – no of times the customer re-ordered products

• Prodreorderchance- Chances of getting a reorder of a particular product (in terms of probability)

• Prodtimesreordered- No of times a particular product got reordered.

2) Missing value Treatment:

Our Data contains which is in the variable days_since_prior_order. It was due to the fact that the first order of every customer got a NA. I have replaced all the missing values with 0.

Machine learning Model Building:

I have used Decision Tree for the classification of predicted variables, tested and compared with Logistics Regression and naiveBayes algorithms.

I was able to achieve 69% accuracy on trainmodel with removing all the insignificant variables and computed the result in one CSV file.

Conclusion:

This Market Basket analysis is very **subjective analysis for this data** and depends heavily **on customer to customer buying pattern and service provided by the store**. It also depends on time of the day and day of the week plus the age group and gender of the customer to have more accurate results in the applied algorithm. There can be more things which could have been impactful to the model but this the best **structured solution** in my opinion and the **feature techniques** of developing variable importance based on **probabilities and re-ordering chance**.

Thanks!