Amazon
ML engineer Hiring Problem

Name: Manish M. Dalvi

Problem statement

- Predict the customer category.
- Based on Customer category recommend correct products.

Problem Type: Binary Classification Problem

Performance Metric: Macro Precision score

Business Objective

- No low latency requirement.
- Error not very critical but should be in respectable limit.
- Interpretability is important on why a person was classified into a category based on which features.
- Recommend Products based on the predicted category of the customer.

Given Dataset

Dataset with 10 features and 1 output

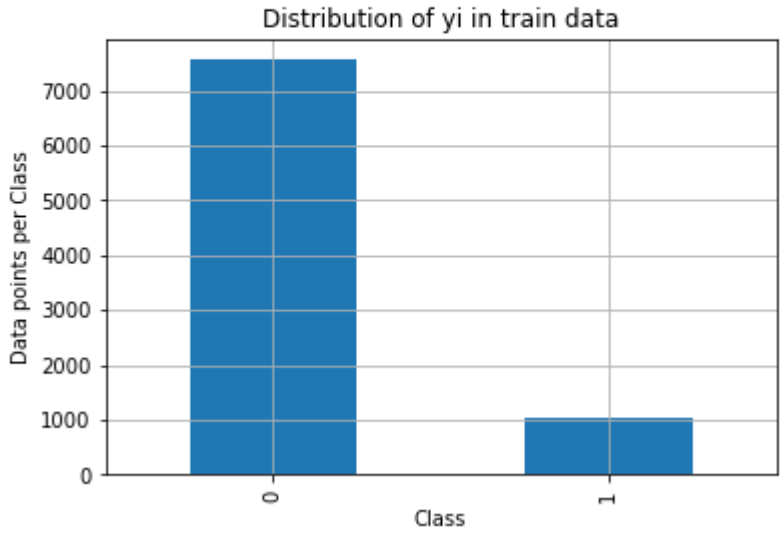
customer_visit_score	A score based on how regularly the customer visits the website
customer_product_search_score	Quality or price range of product that the customer searches for. For ex: a customer Searching for a laptop will have more weightage than someone looking for a book
customer_ctr_score	How many of the searched links does the customer click
customer_stay_score	A score based on the time spent on an avg. by the customer
customer_frequency_score	A score based on how many times in a day the customer visit the website
customer_product_variation_score	A score based on how many varieties of products does a customer search for, for ex. electronics, apparels, etc.
customer_order_score	Score based on the no. of orders that has been successfully delivered and not returned
customer_affinity_score	An internal overall score calculated which signifies the affinity of the customer towards the website
customer_category	The cluster/group to which the customer should belong to
customer_active_segment	The categorization of the customers based on their activity
X_1	Anonymized feature based on loyalty of the customer

Dataset Distribution

- Since the test dataset did not have the output category, the train data is split into train and cross validation data in the 80:20 ratio.

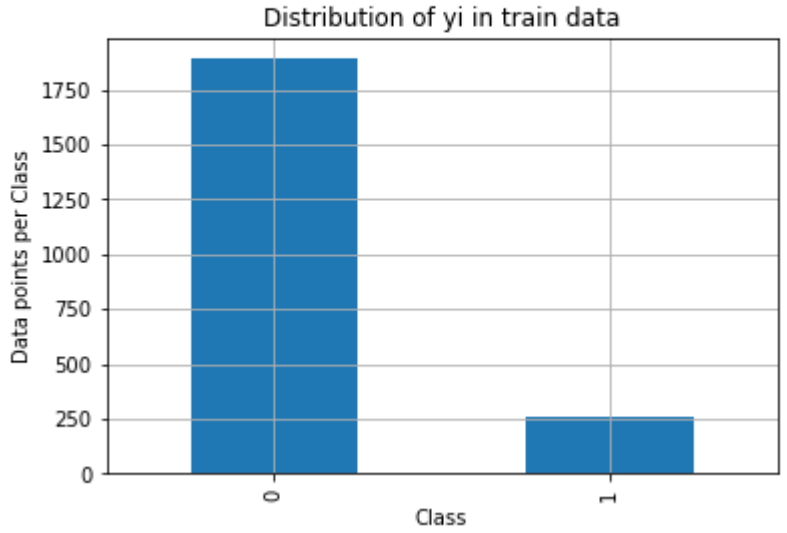
Training Dataset

```
0    7554
1    1036
Name: customer_category, dtype: int64
```



Cross Validation Dataset

```
0    1889
1     259
Name: customer_category, dtype: int64
```



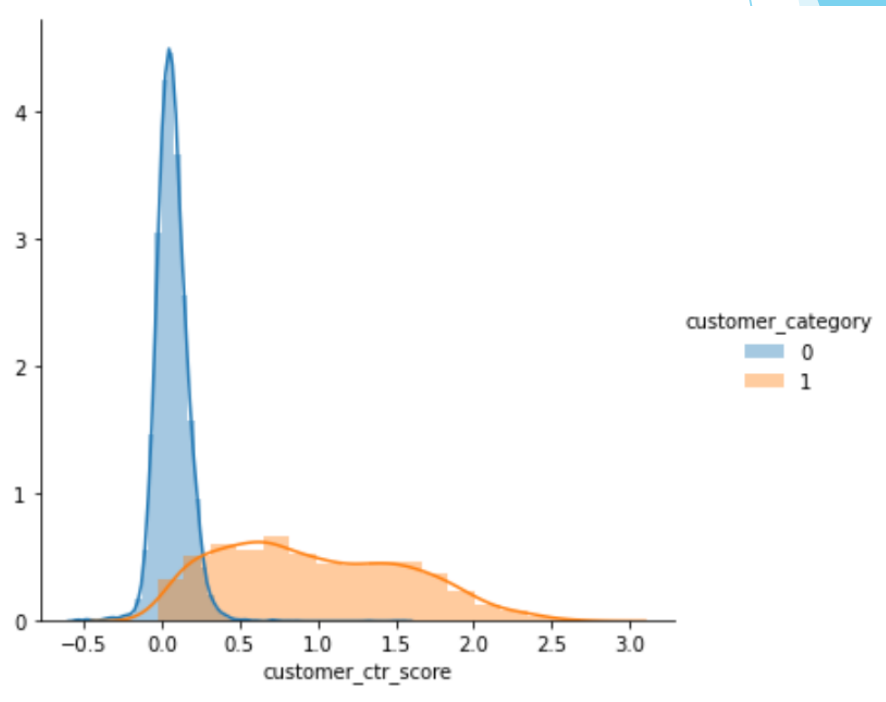
Clearly, the given dataset is an imbalanced dataset.

Feature Engineering

From the given dataset which features are important and why?

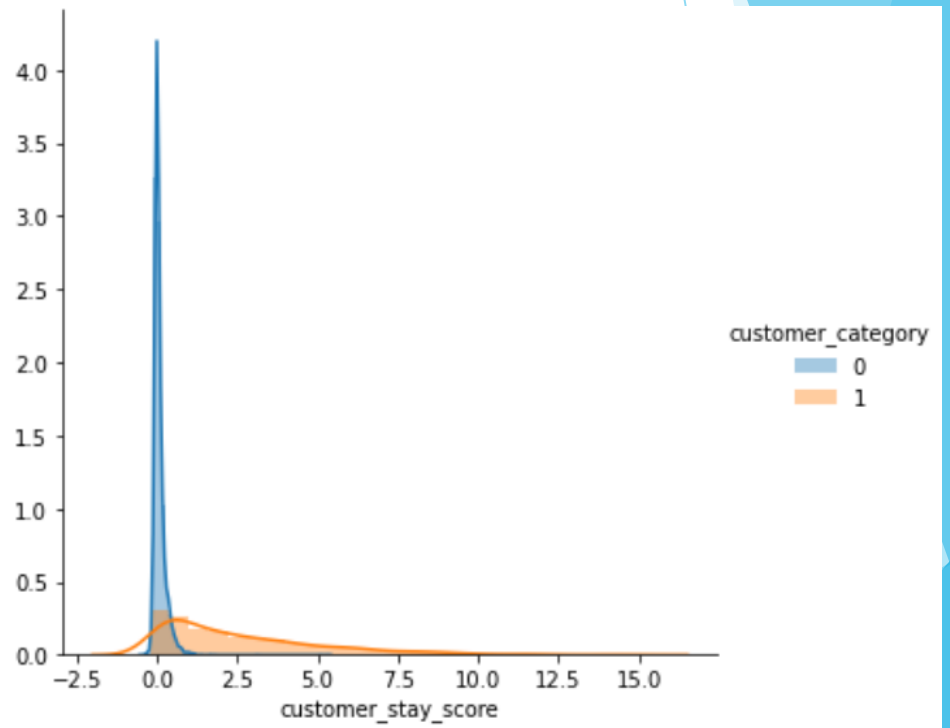
Customer CTR score – Most important Feature

- Higher the ratio of click to search, the customer is categorized into category “1”.
- The graph distribution of the two categories has very less overlap due to which this feature is very important.



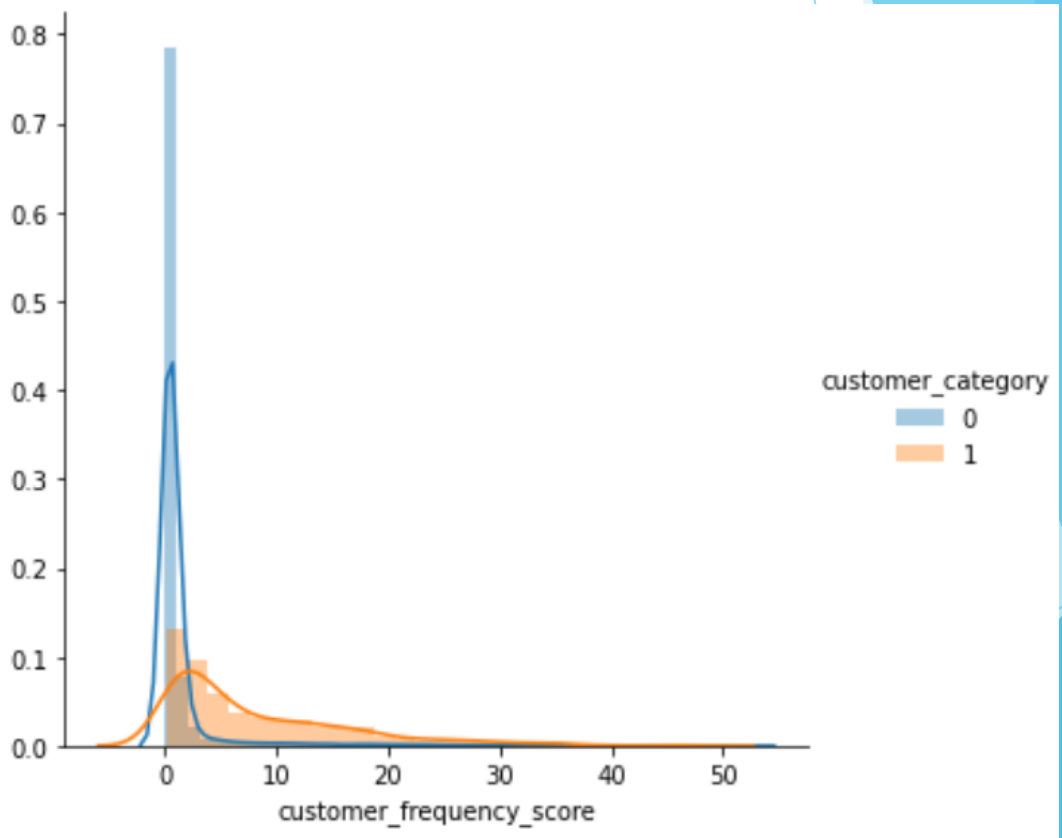
Customer Stay Score – 2nd Most important feature

- The more the time spent by the customer on the website, the more likely the customer is categorized into category “1”.
- This graph too has less overlap between the two categories which helps in classification.



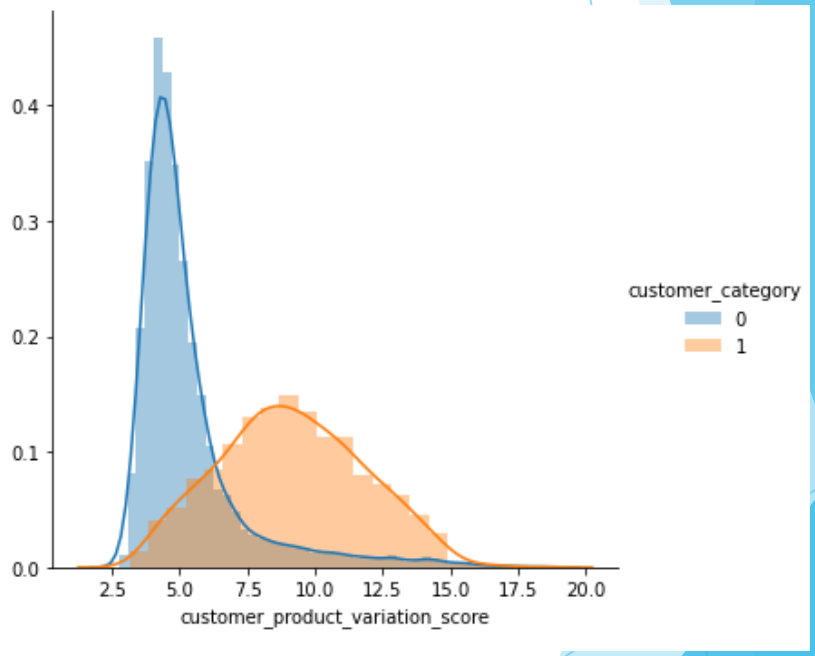
Customer Frequency Score – 3rd most important feature

- Customer who visited the website more in a day will be categorized into category “1”.
- This graph too has a fair amount of non overlapping areas (after 5) which helps in categorizing the customer.



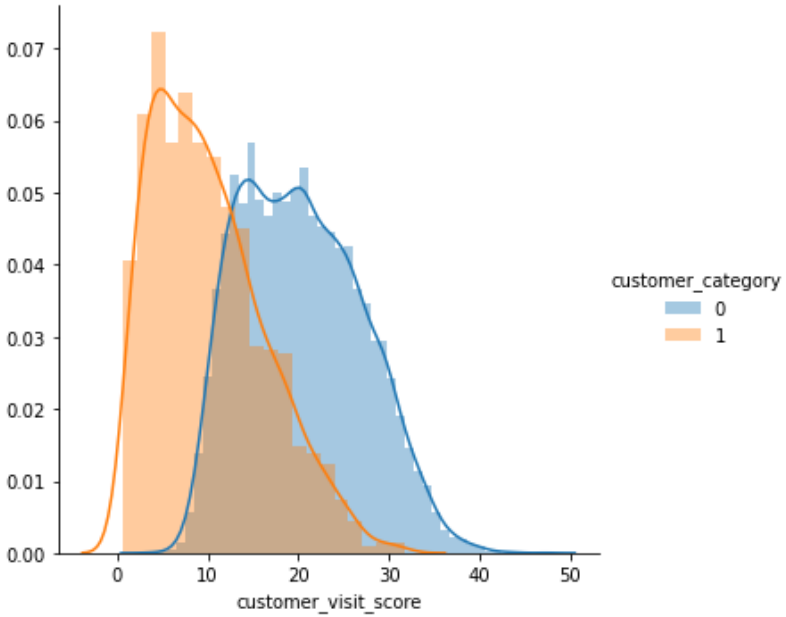
Customer Visit Score – 4th Most important feature

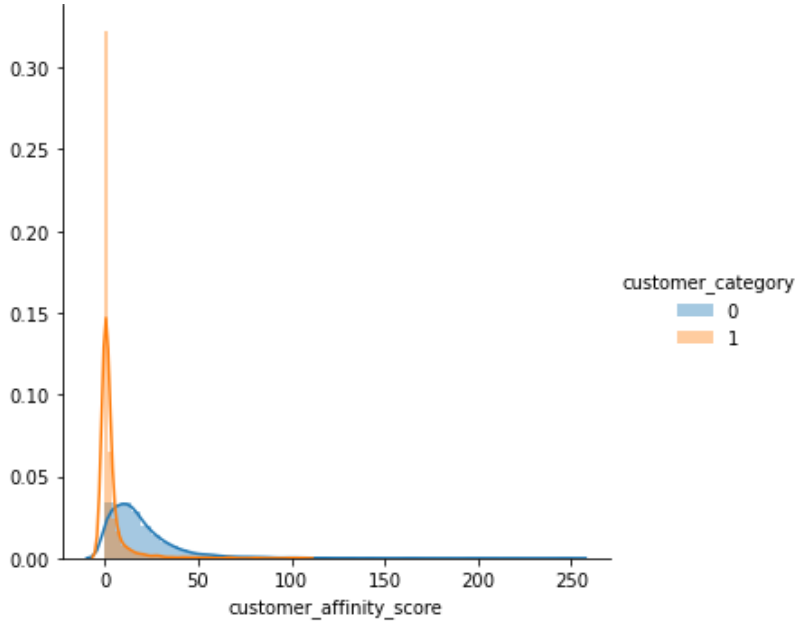
- The customer who visited more variety of products is more likely in category “1”.
- This graph has considerable amount of overlap and due to which, there is difficulty in using this feature alone for deciding the category of the customer.



Other Features

- The Remaining numerical features tend to have overlapped distributions due to which it is difficult to decide the category.
- Among the other features: Customer visit score and customer affinity score tend to have more importance since their distributions has lesser overlaps.



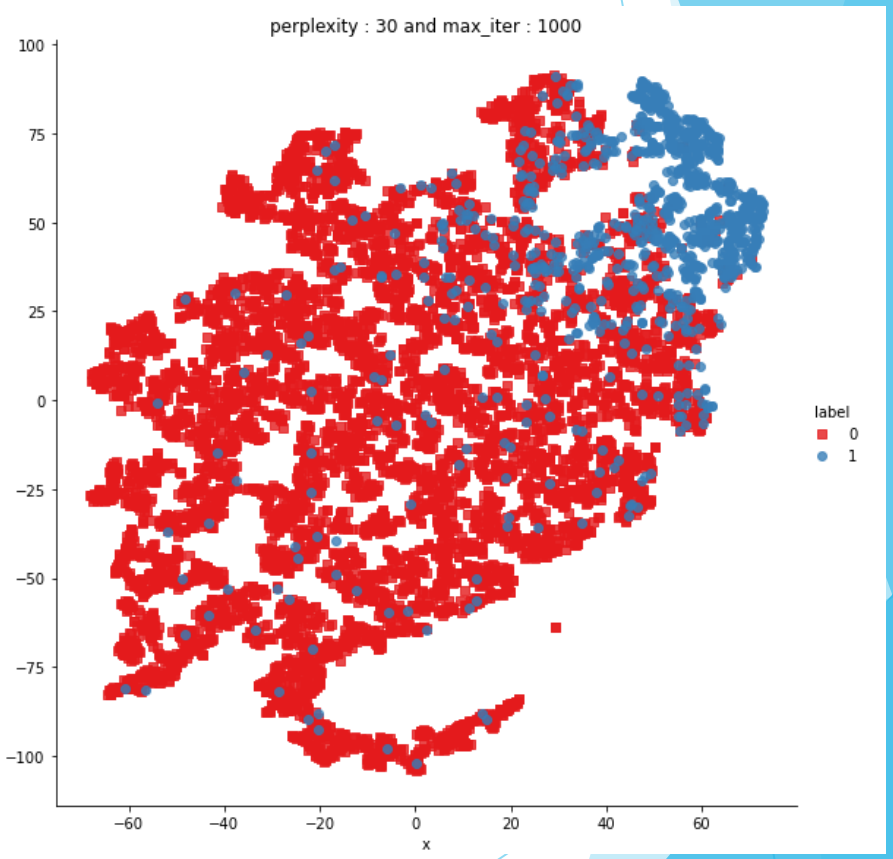


Categorical Features

- Customer Active Segment – Just by using this feature we obtained a 71% precision score.
- Which means that this feature is an important contributor.
- X1 Feature - Anonymized feature based on loyalty of the customer.
- The precision by this feature was not great hence this feature is not relevant in deciding the customer category.

T-SNE

- TSNE can be used to visualize N dimensional data in 2 dimensional .
- It helps in knowing how separable is the data.
- The blue data of category “1” seems quite separable from the red points of category “0”.
- However, there are several blue points overlapping in the red regions where some algorithms like KNN will not work well.



ML algorithms and their Precision Scores

- Decision Tree based algorithms, SVM were used since they handle Imbalanced dataset well.
- Train dataset split into Train and Cross validation dataset in ration 80:20.

ML Algorithm	Precision
Decision Tree	93.54
XG boosting DT	94.35
SVC (balanced class weight, rbf)	87.68

- Other ML algorithms which can be tuned to handle such datasets.

ML algorithms	
KNN (N=10)	95.69
Logistic Regression	94.71
Stacking Classifier (KNN, DT, XGB, meta=Logistic Regression)	95.69

Deciding on the Best Algorithm

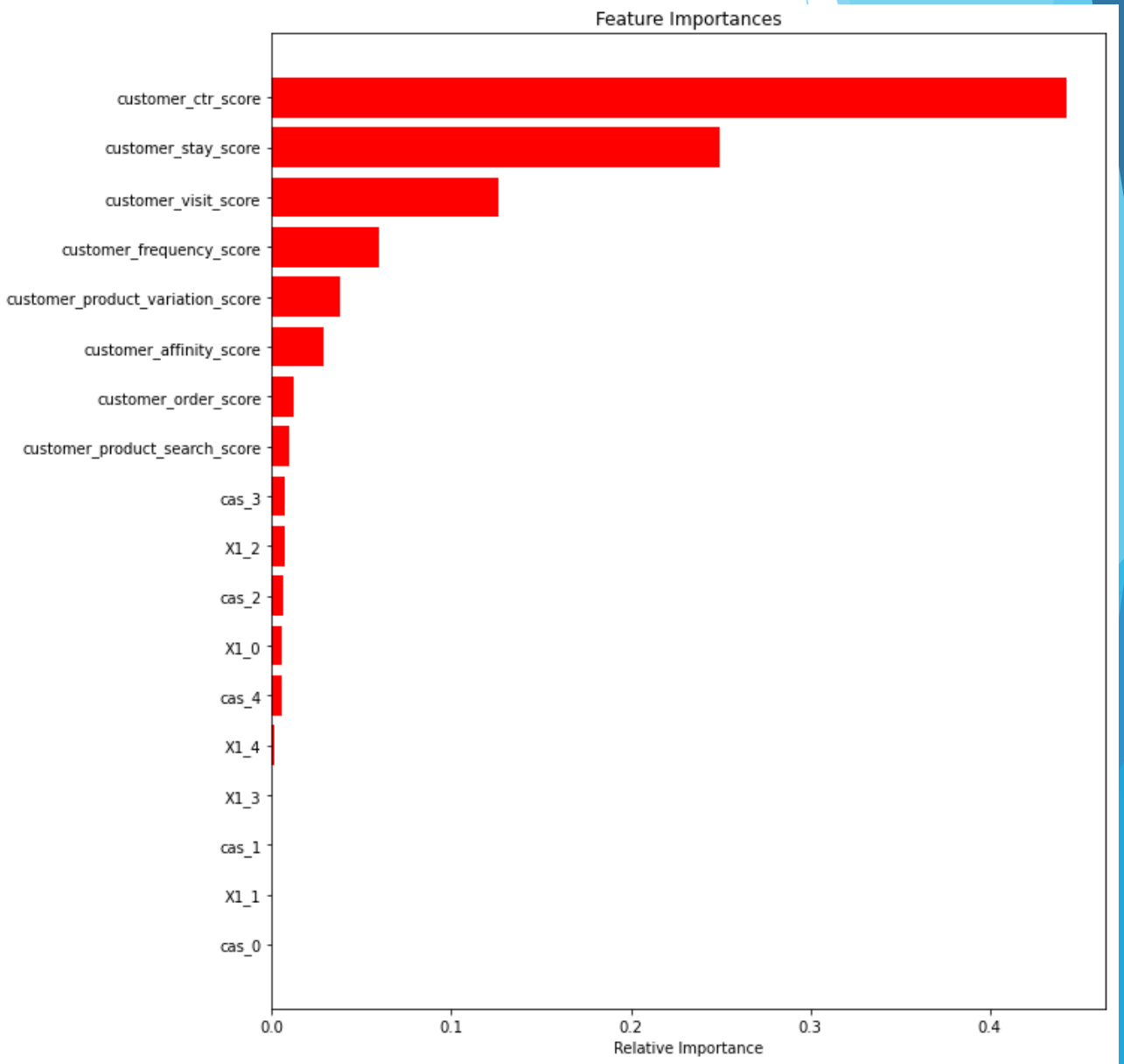
- XGBoost and KNN provided good results during Cross validation stage.
- They were chosen to be trained over the entire dataset.
- XGBoost provided the best results at 92.16268.

Algorithm	Macro Precision Score
XGBoost DT	92.16268
KNN (N=10)	91.81054

Feature Importance with XGBoost

From the graph we see that,

- Customer Ctr score, Customer Stay Score and Customer Visit score are top 3 important features
- We had also seen the same during analysis of the distributions of the graphs during the exploratory data analysis.
- The important features tend to have visually separable distributions for each category.
- Among the categorical feature – Customer Activity score when added up is higher



Key points

Customer belongs to category “1” if,

- The customer had a higher ratio of click to search for searched links.
- The customer stayed on the website for more time.
- The customer visits the website frequently in a day.
- The more categories the customer explores.
- The customer has lesser returns to delivered ratio.

Analysis of the Customer and category

- Category 1 of customers are more likely an experienced buyer and checks the reviews, searches alternate products, more in-depth analysis before purchase.
- Since the research about the product is in depth before purchase, Category 1 are less likely to return the products on delivery
- Category 0 customers are the ones who buy the product with less time spent on it, lesser analysis and quicker buying decision.
- Reason the Category 0 customers have higher “return to delivered” ratio is due to fact that lesser time is spent on the website which means improper analysis of reviews before making a purchase decision.

Pros and Cons of the Recommendation

Pros:

- 2 Simple Categories.
- Recommend Similar Products as being searched by customer.
- Quicker recommendations and less processing for categorizing customer.
- Advertise on other website so that customer visits frequently and chance of purchase increases

Cons:

- Not very in-depth recommendation system since very few features in analysis.
- May recommend products to category “0” customers that are not having good ratings due to which higher returns done by the category “0” person.
- Small number of categories, can be increased based on customer preferences.
- Recommend different kind of products rather than similar products.

Proposed Architecture

- Recommend best products of the current searched item using Content Based Filtering.
- Recommend other products, which similar users like the one under observation using Collaborative Filtering.
- Recommend quality products with better ratings and reviews to both categories to avoid higher return to delivered ratio.
- Perform Text analysis over reviews of the product for recommendations.
- Instead of using 2 class classification recommendation system. Use User-User Matrix to recommend Products which provides more vast recommendations.
- Use Features like User Geography, Previous purchases, Quality of products purchased, Income, Preferred brand in the category.
- Based on User Product Search, recommend what similar users would have preferred.