Report on Machine Learning Technical Test

Exploratory Data Analysis

1. The data is comprised of 10000 observations under 20 different column variables

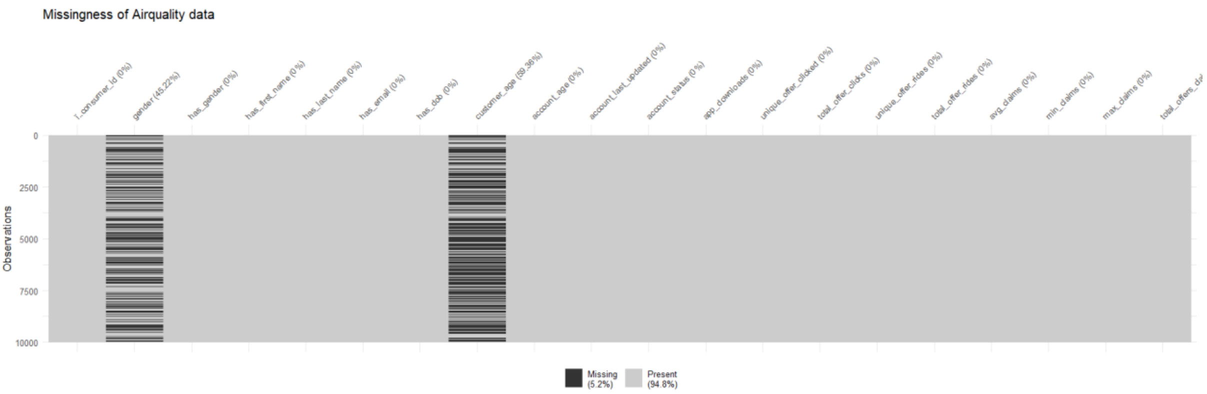


Figure 1: Missing Plot

2. Missing Values
   - I. 5.2% of total observations are missing data where gender is missing by 45% and customer age by 59%.
   - II. As it is known imputation restores the missing observations but it is most likely to increase model variance and bias as gender variable is nearly missing 50% of the observation
   - III. Since customer age is heavily missing (> 50% missing) it is a good candidate to be dropped as a predictor

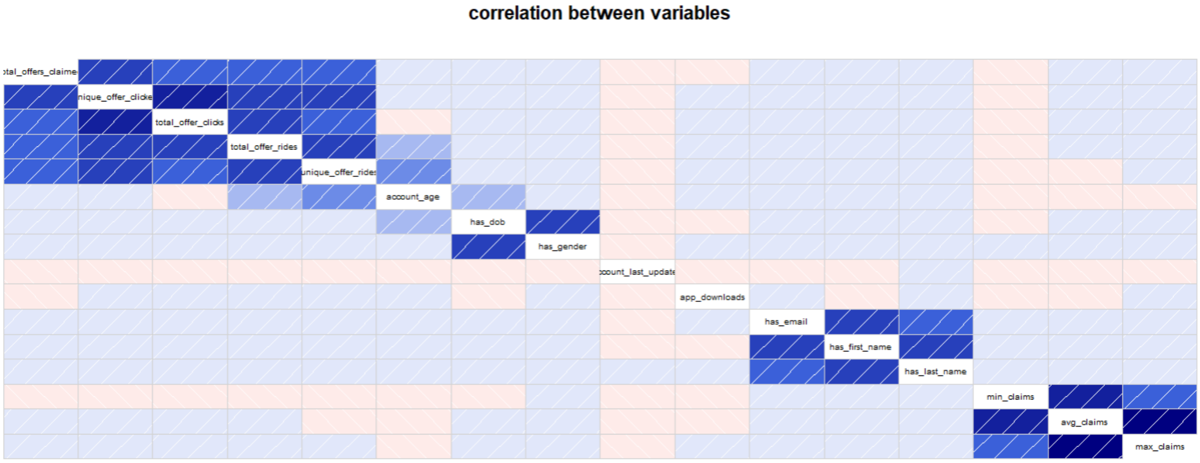
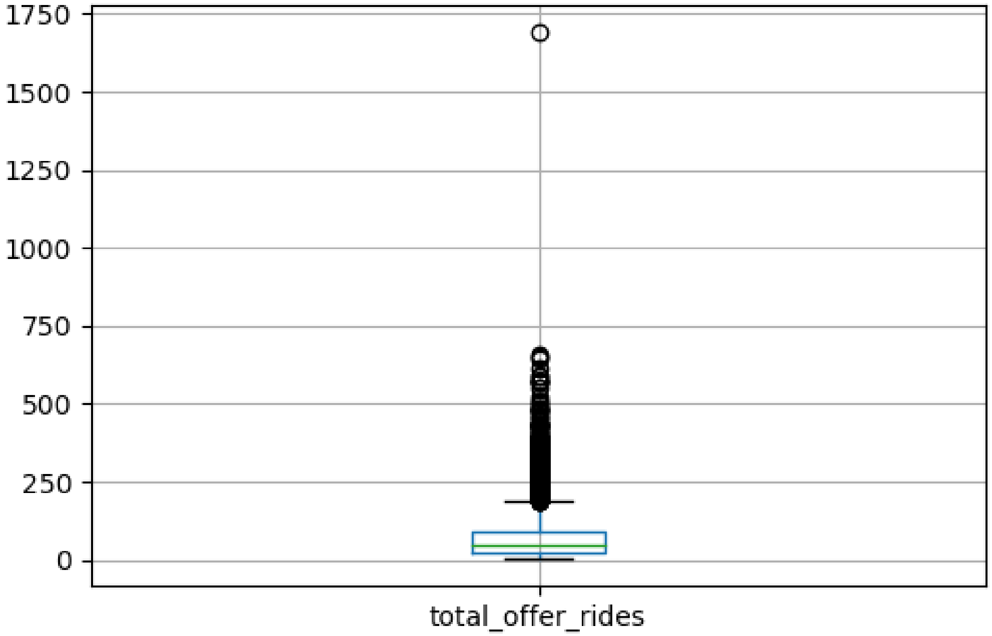


Figure 2: Correlation Plot

3. There is a strong positive correlation between feature columns,

- Total offer claims, unique offer clicks, total offer clicks, total offer rides and unique offer rides.
- First name, last name and email
- Min claims, avg claims and max claims
- DOB and gender

4. There are 112 duplicate “consumer id” present in the dataset. All the duplicate id’s share same attributes except app downloads. I assume that it could be the count i.e, number of times app is downloaded. It might be because of multiples times app downloaded by same consumer
5. Consumer Id “1a4117f1-045c-482d-81d8-359bd14072d4” has unusual “total offer rides” value to be 1693 which could be potential outlier



6. Account status is identical for all the observations, so it is disregarded from using it as feature column. Similarly, “consumer id” column cannot be used as part of features as it plays unique identifier role
7. Out of 20 columns only 16 of them are considered for statistical analysis. Those are, 'has_gender', 'has_first_name', 'has_last_name', 'has_email', 'has_dob', 'account_age', 'account_last_updated', 'app_downloads', 'unique_offer_clicked', 'total_offer_clicks', 'unique_offer_rides', 'total_offer_rides', 'avg_claims', 'min_claims', 'max_claims', 'total_offers_claimed'

Statistical modelling

Clustering

To perform clustering K-means clustering algorithm was used.

To choose the ideal cluster size elbow Criterion and Silhouette Coefficient methods were tried and applied.

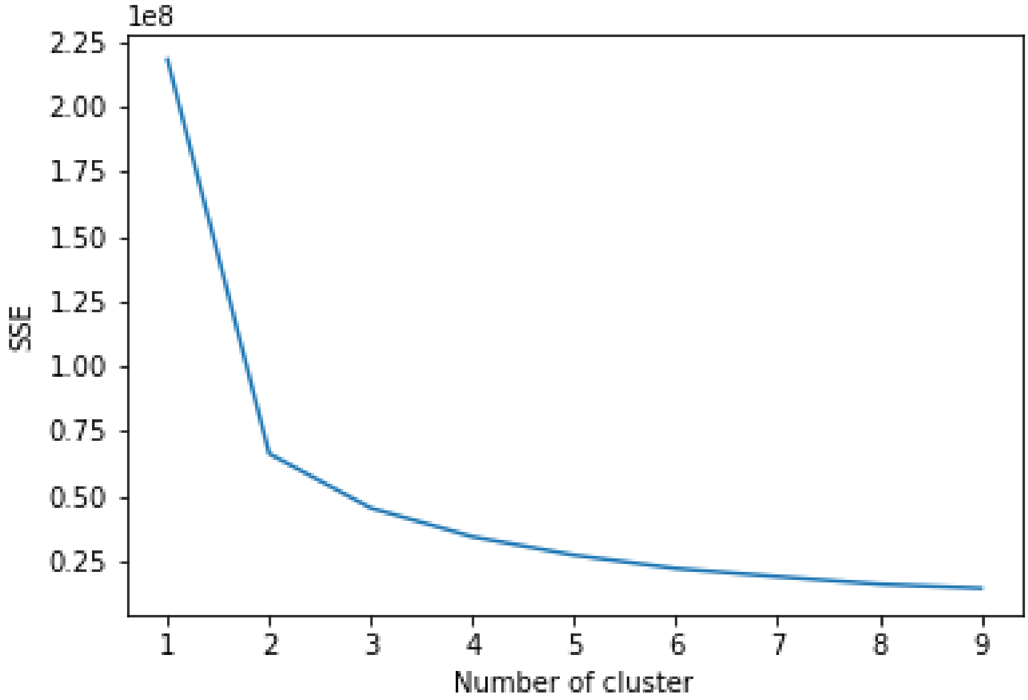


Figure 3: Elbow Criterion Plot (SSE vs Cluster size)

Bent arm like curve in the plot points to ideal k size to be 2 or 3 which still satisfies low Sum of Squared Error (SSE)

Silhouette Coefficient - higher the Silhouette Coefficient score relates to a model with better-defined clusters.

For K=2, The Silhouette Coefficient is 0.756214036347971
For K=3, The Silhouette Coefficient is 0.626216246989207
For K=4, The Silhouette Coefficient is 0.6235650451360049
For K=5, The Silhouette Coefficient is 0.5712903592708712
For K=6, The Silhouette Coefficient is 0.5670980737710015

Based on the above results, K = 2 was chosen

K-means clustering was performed on the selected feature columns which resulted in forming 2 clusters with cluster 0 containing 1592 observations and cluster 1 containing 8408 observations.

To evaluate quality of the clusters as well as its content Silhouette score was used, which measures the separability between clusters based on the distances between and within clusters.

K-means model with 2 clusters has got 0.7564460864016483 Silhouette score with metric set to ‘euclidian’.

Visualizing Clusters - Given 16 feature columns, it is not sensible to plot 16-dimension at the same time. So PCA dimensionality reduction technique is used to visualize the clustering results. The first two principle components PCA1 and PCA2 were chosen which explained 79.1%, 20.6% of total variance respectively.

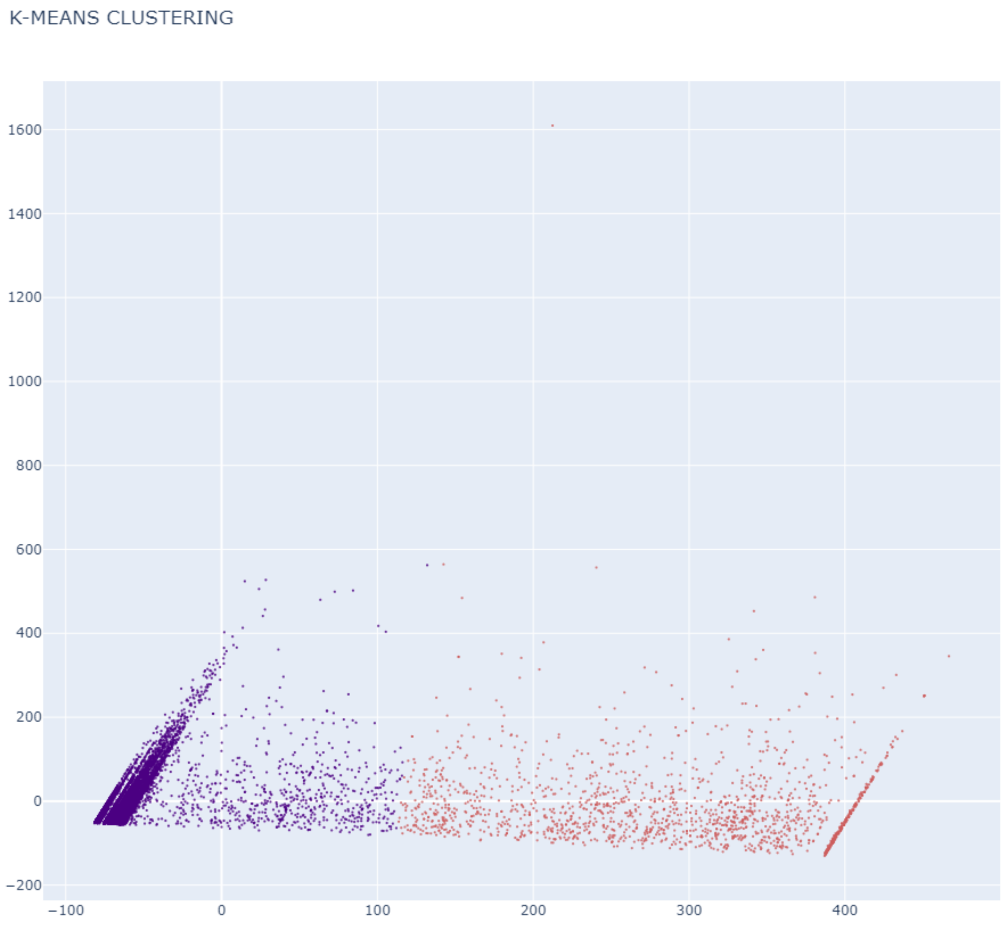


Figure 4: PCA 2d scatter plot for K-means cluster visualization

Classification

Now obtained clusters are set as target variable for carry on modelling with supervised machine learning techniques. Random Forest was chosen as the classifier since it gives better model interpretability, proven to be best in fraud detection/consumer classification problem and also to understand which features are important in the generation of the clusters.

Random forest algorithm achieved 99.8% accuracy in predicting test data with 20% split proportion. The other metrics calculated that were calculated are,

Accuracy Score - *0.9984*
f1_score - *0.9970652144565189*
Recall - *0.9902200488997555*
Precision - *1.0*

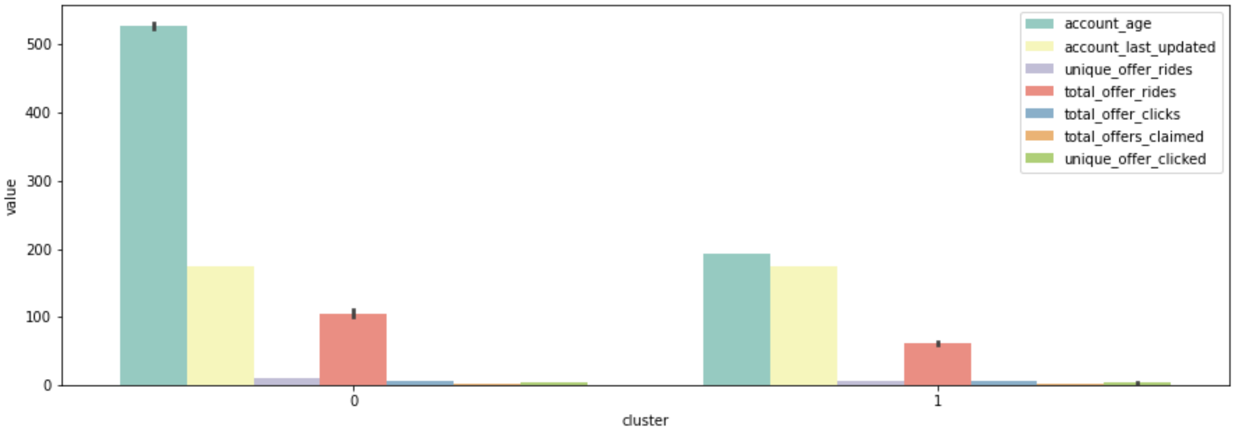


Figure 5: Feature Importance in clusters

Based on Random forest feature importance function it is clearly seen that clusters are basically distinguished by 3 feature columns “account age”, “account last updated” and “total offer rides”.

Assuming that the values represent number of days, when looking at the random data points, it seems that there is significant gap between account age and account last updated number in cluster 0, for instance account age is 572 and account last updated value is 173, account age is nearly 3 times greater than account last updated in most of the cases. On other hand, account age is close to account last updated value in cluster 1, for instance account age is 188 and account last updated is 174. So the consumers are grouped based on how long they have been using the platform.

On further interpretation to meanings of the cluster, analysis was extended with another highly interpretable machine learning algorithm which is decision tree.

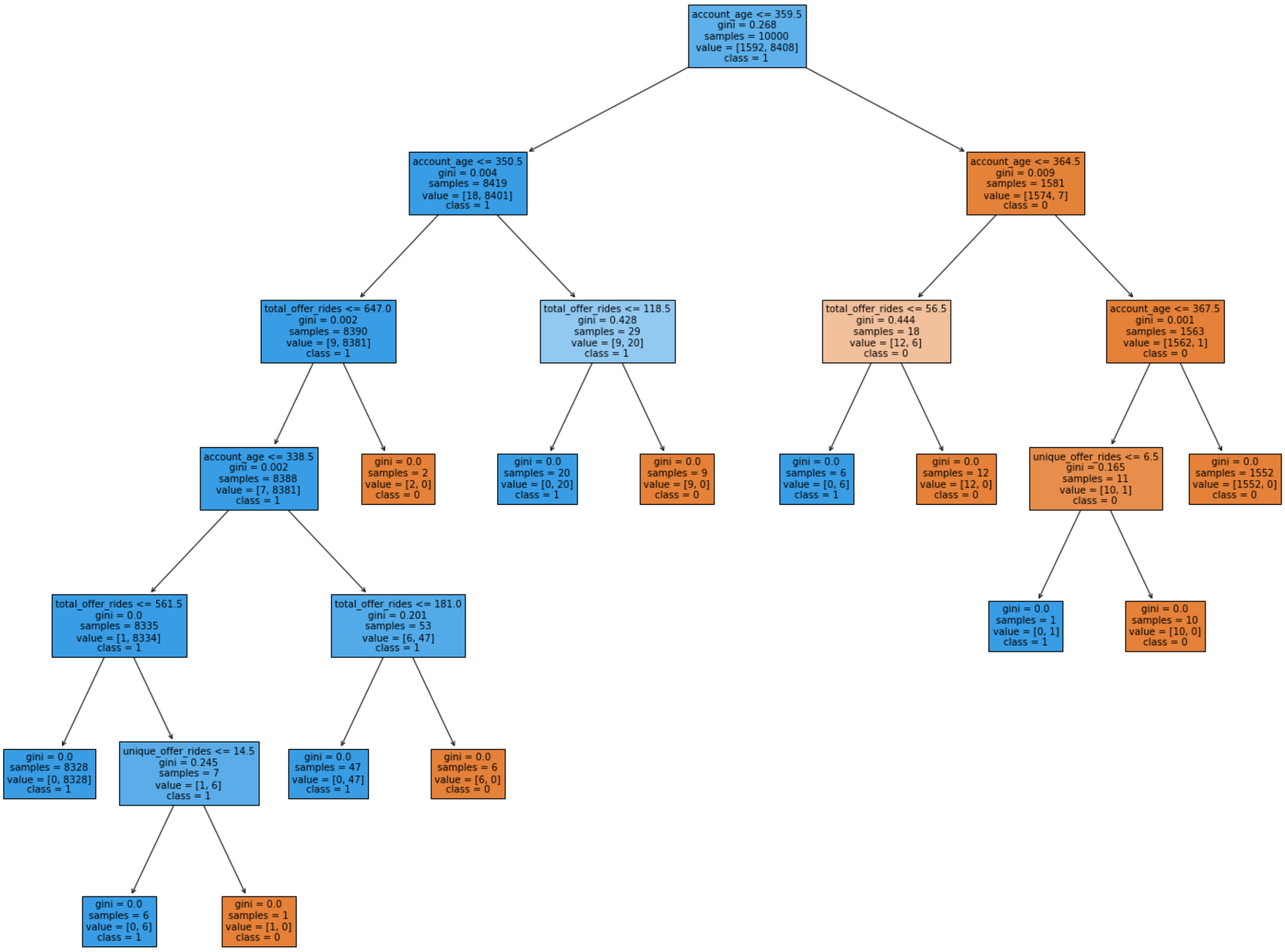


Figure 6: Decision tree graph plot

The ‘value’ row in each node tells us how many of the observations that were sorted into that node fall into each of our three categories. We can see that “account age” feature was able to completely distinguish one cluster class from other.