**Assessment submission form: MIS41270**

|  |  |
| --- | --- |
| **Student name & number** | Name: Abhishek Chauhan (ID - 20200248)  Name: Muzammil Memon (ID – 20200387)  Name: Robert Tuke (ID - 17501176)  Name: Zechen Hu (ID - 20203292) |
| **Assessment title** | Insure ABC - New product purchase prediction using Data Analytics |
| **Module code** | MIS41270 |
| **Module title** | Data Management and Mining |
| **Module coordinator** | Aoife D’Arcy |
| **Date submitted** | 16/04/21 |
| **Date received** |  |
| **Grade mark** |  |

**A SIGNED COPY OF THIS FORM MUST ACCOMPANY ALL SUBMISSIONS FOR ASSESSMENT.**

**STUDENTS SHOULD KEEP A COPY OF ALL WORK SUBMITTED.**

**Procedures for Submission and Late Submission**

**Ensure that you have checked the School’s procedures for the submission of assessments.**

**Note: There are penalties for the late submission of assessments. For further information please see the University’s Policy on Late Submission of Coursework, (**[**http://www.ucd.ie/registrar/**](http://www.ucd.ie/registrar/)**) Plagiarism: the unacknowledged inclusion of another person’s writings or ideas or works, in any formally presented work (including essays, examinations, projects, laboratory reports or presentations). The penalties associated with plagiarism designed to impose sanctions that reflect the seriousness of University’s commitment to academic integrity. Ensure that you have read the University’s Briefing for Students on Academic Integrity and Plagiarism and the UCD Plagiarism Statement, Plagiarism Policy and Procedures , (**[**http://www.ucd.ie/registrar/**](http://www.ucd.ie/registrar/)**) Declaration of Authorship**

**I declare that all material in this assessment is my own work except where there is clear acknowledgement and appropriate reference to the work of others.**

**Signed: Robert, Jack, Abhishek & Muzammil Date: 16/04/2021**

**Word count not including chapter titles, appendix or references.**

# **MIS41050 – Data Mining and Management**

# **Situation Assessment**

**Background**

Our client Insure ABC is launching a new home insurance product, Insure ABC currently provide health insurance, motor insurance and travel insurance, using the dataset provided on current customers for other insurance products we need to create a data driven marketing campaign to target previous customers appropriately.

**Business objectives**

1. Identify the existing customers most likely to purchase the new home insurance product.
2. Identify the most successful strategy for marketing the new product to these customers including the proper channel to be used.

**Success criteria**

1. Identify several key target segments of previous customers who will be most receptive to buying Home Insurance from Insure ABC.
2. Have a misclassification rate on our channel preference prediction model less than 35%.
3. Create a specific marketing strategy for each of the target segments in our data.

**Segmentation Methods**

K-Means Clustering - When the cluster count is known, K Means uses Euclidean distances to find intra-class similarity for cluster point assignments followed by virtual median assignment for understanding inter-class links. Due to the linear time complexity, it is known as one of the best choices for larger data sets (MacKay, 2003).

Hierarchical Clustering – The cluster count can be easily decided with an easy understandable dendrogram, allowing us to make a more informed choice of the number of clusters we desire from our data. Hierarchical method repetitively finds the closest cluster points and links them to form a tree like structure which can be best used for understanding the embedded data structures. It produces iteration results unlike K-means method (Maimon and Rokach, 2005).

**Analytical Methods**

**Linear & KNN Regression**

Linear regression is easy to fit, to interpret and implement in our data. However Linear regression is a parametric model meaning it can only be used effectively when there is a linear relationship between the data and the response variable (Freedman, 2009). KNN regression, in contrast is non-parametric model and assumes no such relationship, however this method is less interpretable than Linear regression and performs worse under situations where the relationship between the data is linear. In situations where we know the relationship between variables is non-linear or if we don’t know the true relationship, we should use KNN regression (Altman, 1992).

Both Linear regression and KNN regression assume that the response variable is quantitative but, in our case, we are trying to segment different customers making our response variable qualitative or categorical, neither of these methods is therefore ideal and what we need for this task are methods that allow us to classify different datapoints.

Logistic regression – It is an upgrade to linear regression. It is more suited to dealing with classification. It is used to model the probability of binary events and is useful for prediction based on binary variables. However just like linear regression, it performs poorly for non-linear problems because of its necessity to have a linear decision surface (Tolles and Meurer, 2016).

Linear discriminant analysis – LDA is ideal when working with continuous variables to predict a qualitative response variable, it also assumes a normal distribution. In situations however where the x variables are not continuous, LDA can’t be used (McLachlan, 2004).

**Decision Tree & Random Forest**

Decision Tree is a simple structured model which can handle multiple datatypes. The main problem is regarding the choice of splits criterion. Random Forest resembles decision tree, except that it implements a snowflake like multiple tree structure. This structure taxes the processing time but helps overcome the problem of automatically choosing the splits when the categories are multifarious. In our case we are aiming to achieve a more understandable and interpretable output structure. On experimenting with both the techniques, a decision tree returned favourable results for our analysis (Ho, 1995; Kamiński, Jakubczyk and Szufel, 2018).

# **Data Quality Report**

Data quality is extremely important since low-quality data leads to multiple versions of the truth, conflicting metrics and increased time in which IT spends cleansing data for each individual request from the business.

The outliers found in the training dataset were normalized to their normal format by looking up the data dictionary and their respective list of values. The final files used for analysis were cleaned and noise-free, data was treated as the final source-of-truth. Explanations revolving around missing values, cardinality, outliers, and visualizations to approach clustering practices are explained below.

1. **Missing Values**

Observing the data quality report, we can see continuous and categorical features have significant numbers of missing: MotorValue (17.25%), HealthDependentAdults (37.82%), HealthDependentKids (37.82%), Occupation (38.04%), and HealthType (37.82%) The absent values in the Occupation feature look representative of this type of data. A little over a third of the customers in the dataset appear to have merely not supplied this piece of evidence. Imputation is perhaps not a good approach given the percentage of misplaced values and the high cardinality of this feature. Instead, this feature may be a good candidate for a derived flag which again, does not add monumental value to our clustering analysis. We can easily see the reason for the missing values in the HealthDependentAdults, HealthDependentKids, and HealthType by inspecting sectional details. When the HealthInsurance function is set to no, these features still have values unaccounted for.

It is a fair perception from a business standpoint—if a patient does not have health insurance, the specifics of a health insurance program would not be populated. This also explains why the percentages of missing values for each of these features are the same. The same reason applies to the missing values for the MotorValue element. When we look at the particulars, we can see that whenever the MotorInsurance feature is set to no, the MotorValue feature is empty. This is entirely plausible, if a consumer does not have a motor insurance policy, none of the policy's particulars will be available.

1. **Cardinality**

The age function has low cardinality (given that the dataset has over 4000 instances). This isn’t surprising, considering that ages are given in full years, and there is only a limited range of possible ages in this case, 18–80.

For HealthDependentAdults and HealthDependentKids, Missing values are considered when calculating cardinality. For example, the only values in the data for HealthDependentAdults are 0, 1, and 2, so cardinality is 4 due to the existence of missing values. Given the limited number of definite values, we convert these features to categorical features after careful consideration of their importance.

The occupation feature is inconsistent, and fairly so, since the data seems to be manually entered. The fact that a categorical function has 1,830 levels renders it useless for model building. The fact that the mode percentage for this feature is only 0.34 percent emphasizes this point even more. We decided to remove this function from the ABT due to its high cardinality. A data classification technique based on occupation umbrellas (tester and developer -> IT) can be employed, but we chose to skip it for out activity since it does not add enough relevance and developing the algorithm would be a challenging task that would require mapping classification guides.

1. **Outliers**

Amongst the dataset, only the MotorValue feature really has an issue with outliers. We can see this in a couple of ways. First, the difference between the median and the 3rd quartile and the difference between the 3rd quartile and the maximum values is wayward. This suggests the presence of outliers. Second, the histogram of the MotorValue feature shows a huge skew to the right-hand side (Please refer to [Fig 1.7](#_Appendix)) Extremely large values were spotted in the dataset as well. These outliers should be investigated with the business to determine whether they are valid or invalid, and based on this, a strategy should be developed to handle them. If valid, winsoring or a clamp transformation is the best approach. To avoid complexities in the modelling process and, we rendered these values invalid.

1. **Visualisations**

The visualizations are the MVPs of the data quality report for understanding the distributions of different features. The distribution of the age function is peculiar. In a large population, we would expect age to follow a structured distribution, but this histogram certainly indicates a dual distribution. There are three distinct clusters of customers: one in their early twenties, another with a mean age of around 48, and a small group of older customers with a mean age of around 65 (Please refer to [Fig 1.5](#_Statistical_Visualisations)) From a modelling viewpoint, we could hope that these three groups might be good predictors of the target feature, PrefChannel. On analysing the HealthDepAdults and the HealthDepKids columns, the firm could use data that majority of consumers with dependent children have multiple children.

The most interesting thing to learn from the representations for categorical features is that the target variable is slightly imbalanced. Email connections are preferred by many customers over phone or SMS contacts. The Age descriptive feature and the target feature, PrefChannel, have quite a close relationship throughout our visualizations.

On further digging, we can see that the customers whose preferred channel is SMS are largely younger. This is palpable from the average age region of the bar and the fact that there are very few instances in the age range above 60 (Please refer to [Fig 1.6](#_Statistical_Visualisations)). We can see the contrasting pattern for those customers whose preferred channel is phone. There are limited customers below 40 in this group. On further analysis, there seems to be no relation between age and gender. For those customers whose favoured channel is phone, the overall ratio between rural and urban locations is switched—more rural customers prefer this channel. In the other two channel preference groups, SMS and e-mail, there are quite a few more urban inhabitants. The findings suggest that location is relatively analytical of the prefChannel feature and should be incorporated for clustering setups.

# **Clustering results**

Our clustering results demonstrated 5 segments - Segmentation is the process of putting customers into groups based on their similarities. We may often want to analyse each segment separately, since when we segment and study, we know who to target basis preferences and behaviours.

Our statistical model’s analysis establishes and validates 5 different cluster groups. All of these clusters have different profiles and can be converted to insurance user-portraits. The clusters are segmented based on the input variables and centroids are formed in the data. The largest cluster is cluster 3 – with a frequency of 674 instances, followed closely by cluster 2 and 4, respectively.

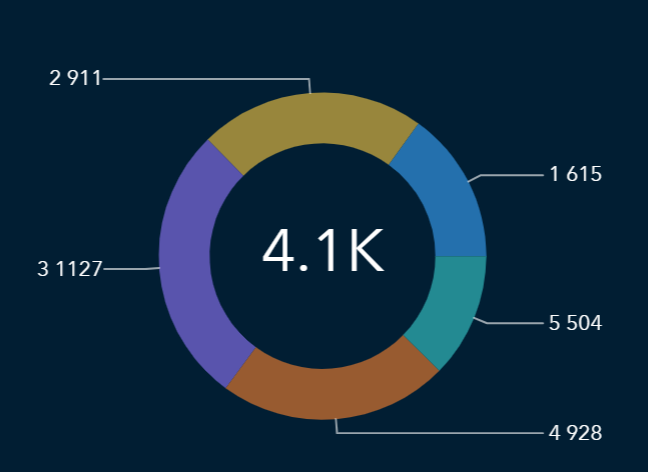


Fig 1.8 – Clustering results provided by the statistical model.

**Final solution inputs**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Age | Gender | HealthDependentAdult | HealthDependentKids | HealthType |
| Location | MotorType | MotorVlaue | TravelType |  |

We decided not to include the inputs for HealthInsurance, MotorInsurance and TravelInsurance as it minimised the inputs and there was no extra information gain from using them as other variables already specify whether an individual had health insurance or not. For example, there was no need to include TravelInsurance because people who had no travel insurance were included in our model as a null value in TravelType. We decided not to use the occupation variable as there are too many occupations to possibly classify and could not realistically be included in any data analysis due to its cardinality.

**Silhouette index**

The silhouette index is one of the most effective ways of measuring how well a dataset has been clustered, it does this by adding up all the distances between the different points on a cluster and its centroid and taking this away from the distance between the different distances for the clusters.

For our final cluster we had a final silhouette index of 17611946, this is relatively poor for a set of clustering however there were key reasons why we chose this over alternative. Additionally, other clusters we had created had poor results, we had created a demographics cluster, a cluster in which we used all the variables and a cluster where we only included binary variables such as TravelInsurance, MotorInsurance and HealthInsurance. The cluster that received the best results was the demographics cluster, however this cluster left out all the key information that was more useful than demographic information such as what form of insurance an individual bought.

**Outline of cluster results**

**Cluster 1**

* **Size –** 16% of total user base

**Unique properties**

* **Health Type** – All of cluster 1 have some form of health insurance from ABC, over 50% have level 3 health insurance.
* **Health dependent kids –** Over 50% have 3 dependent children on their health insurance plan.
* **Gender –** Over 80% are female.

**Cluster 2**

* **Size –** 21% of total user base

**Unique properties**

* **Age –** 90% of this cluster is young people in the 18–25 age range.
* **Motor type –** 87% have bundled motor insurance.
* **Health type –** Most people in this group do not have elite health insurance.

**Cluster 3**

* **Size –** 28% of our customers

**Unique properties**

* **Location -** 100% located in an urban area.
* **Motor type –** Over 70% have unbundled motor insurance.
* **Age** *-* Over 80% are in the age range 38-50

**Cluster 4**

* **Size –** 23% of the dataset

**Unique properties**

* **Location –** 100%of this target group is Rural.
* **Motor value –** Have very reasonably priced vehicles, mostly less than 25,000 euro
* **Age –** This group is typically young rural living people, typically in the 18-25 or 38-50 range.

**Cluster 5**

* **Size –** 10% of the dataset

**Unique properties**

* **Location –** 100% of the people live in a Rural location
* **Age –** This age group is typically made up of older people with over 50% being over the age of 50
* **Motor value –** This group also has much more high valued vehicles, indicating that they are more affluent.

# **Prediction Model Implementation**

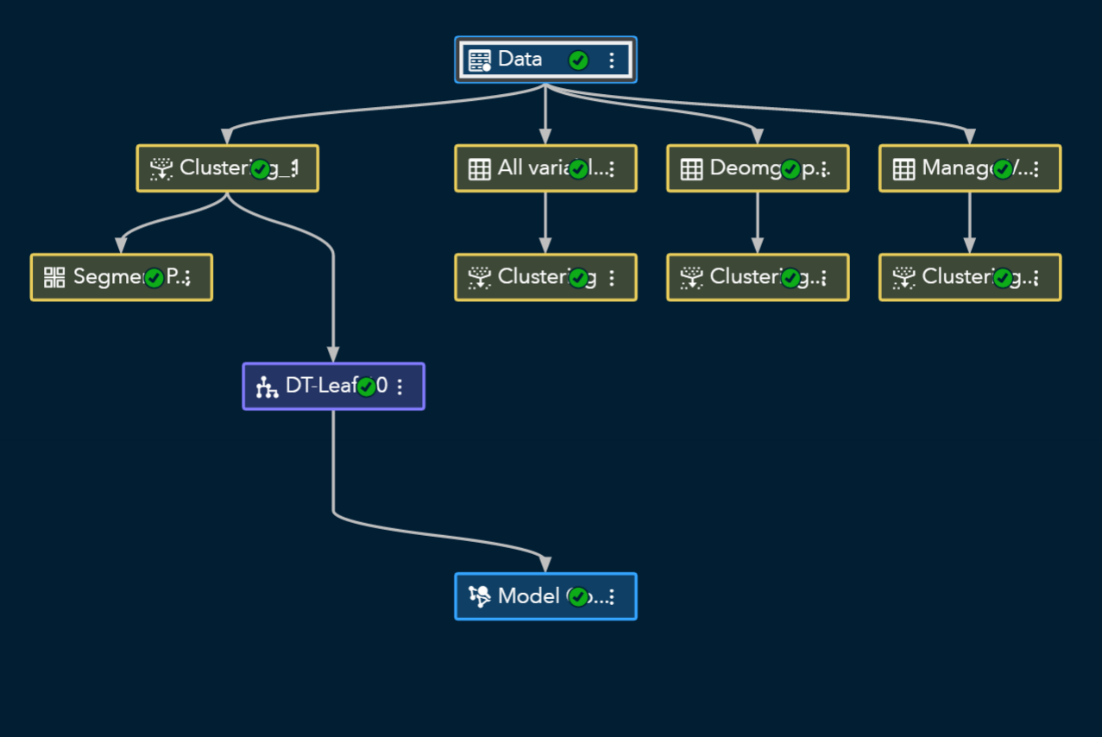


Fig 2.1: SAS Implementation

As we now have the segments, the next step is to predict the demand and connect each individual (each data row) with one of our clusters. Here we implemented the proposed supervised analytical approaches on our current data for validation.

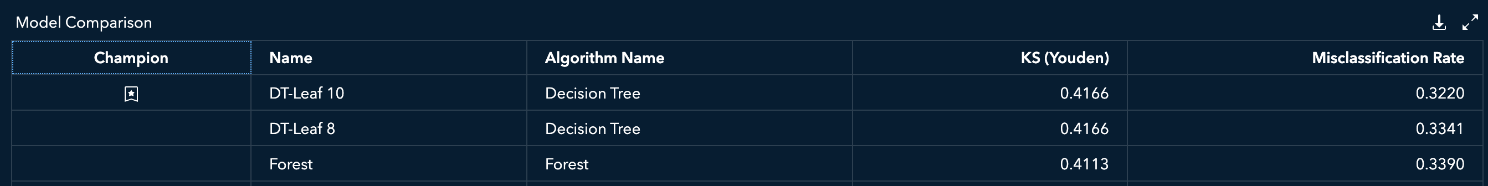


Fig 2.2: Model Comparison

|  |  |  |
| --- | --- | --- |
| **Algorithm** | **Tuning** | **Misclassification Rate** |
| **Decision Tree** | Leaf-10, Branch-2, leafsize-14 | **0.3220** |
| **Decision Tree** | Leaf-8, Branch-8, leafsize-13 | 0.3341 |
| **Forest** | Trees-40, Branch-4, leafsize-5 | 0.3390 |

Table 2.1: Tuning Features

**Decision tree**

For this model we chose to use a decision tree which is an extremely simple form of prediction within machine learning, a decision tree segments the prediction space into several simple regions according to previously set form of splitting rules.

The best results were achieved using the 10-leaf **decision tree.** Here’s why:

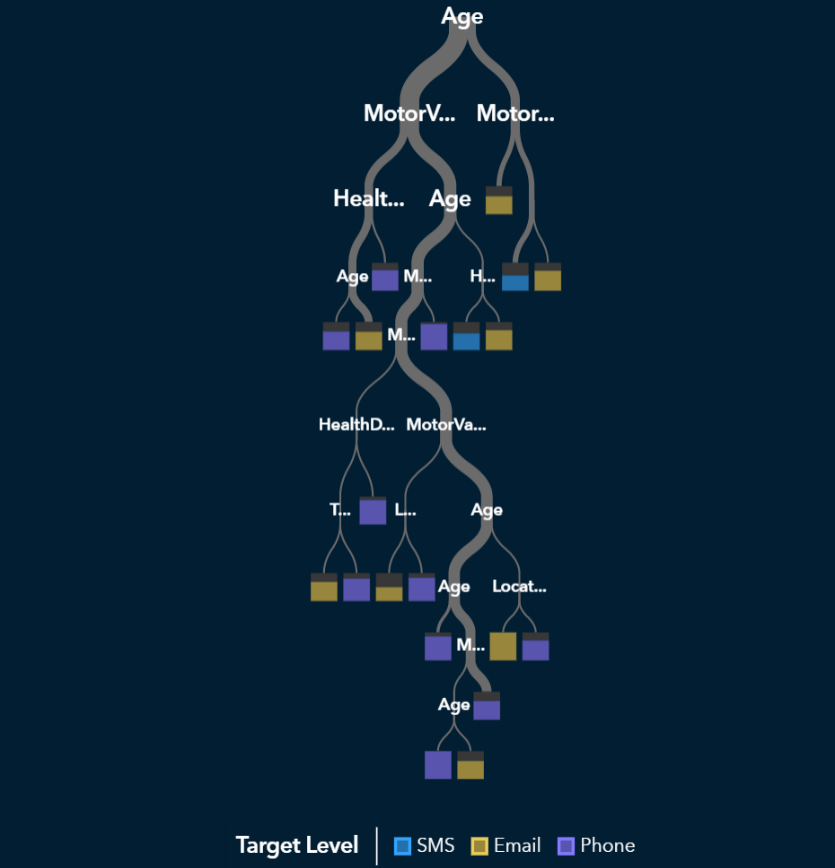
Choice of Split Criterion – We tried and tested our model with Entropy, Gini Index, and Information gain ratio. Information gain produced the best results, explaining that the data required small partitions. This is due to the filtering of available features to feed only the most relevant ones based on segment profiles.

Tree Structure – To achieve an easier understandable structure we tried keeping the branches to ‘2’ splits and hence increasing the depth to incorporate the data volume. The leaf size of 14 was used as suggested by cross-validation folds to avoid overfitting.

Cross validation – In order to prevent bias and overfitting and selection bias, this is done by splitting the dataset into 4 parts training data and 1 part validation data (in the case of 5-fold cross validation) and repeating this process until all 5 parts of the dataset have been validated.

Bias/variance trade-off – The importance of cross validation is most evident when we are trying to get an appropriate balance between bias and variance, variance occurs when we overfit our model on to the training data while bias occurs when we do not underfit our data to our model, the ideal therefore is to balance bias and variance to maximise the future predictive accuracy.

Results/Validations – The cumulative lift, which is a measure of prediction effectiveness, for training and validation performed equally in all the quantile segments. It backed the overall misclassification results for all these data splits being in +-0.01% range. This proves our model to be rightly tuned (no over-fitting or under-tuning) for unseen data. The measure of predictive probability, cumulative captured response also gives the accuracy of over around 50% in the first 20% quartile which is a decent number for incorporating the model to business use. Finally, the ROC curve, draws the cut-offline at 0.15, i.e., the false positive rate 0.805, far greater than the false negative rate 0.322.



Future Enhancements – The use of machine learning models can be incorporated for better accuracy. Neural networks are more details neural structures that artificial biological neurons to calculate the node interconnectivity. This complicates it, making them tough to interpret. These models do not require their outputs to be redesigned. Their two-way node structure allows them to learn from a generated output and change the nets internally, i.e., incremental learning.

# **Model Deployment**

After determining the optimal clustering and prediction analytics settings on the Training Dataset, we applied the techniques with same settings to the Scoring Dataset. Scored Dataset was exported from SAS to perform data visualisation in Tableau.

In short, customer segmentation (clustering) for the Scored Dataset demonstrates similar patterns to the Training Dataset as described above, in which cluster 3 showed the highest customer count of 318, followed closely by cluster 4 (count = 253) and cluster 2 (count = 231), whereas cluster 1 (count = 173) and cluster 5 (count = 115) had lower customer counts, respectively (Fig 3.1).

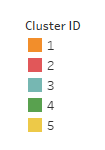
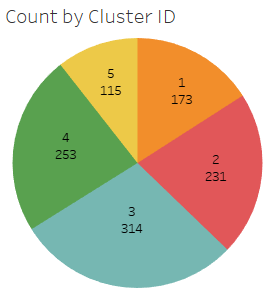


Fig 3.1– Customer count by Cluster ID for the Scored Dataset.

Decision tree algorithm was applied to the scored dataset to obtain the predicted channel of communication for the customer base. In summary, email was predicted to be the most preferred channel of communication with 513 instances, phone ranked second with 407 instances, and SMS ranked the lowest with 166 instances (Fig 3.2).

Customer segmentation was integrated with predicted channels of communication in Fig 3.3, in which further business insight can be developed to create specific marketing strategies, which will be further discussed in the section below.

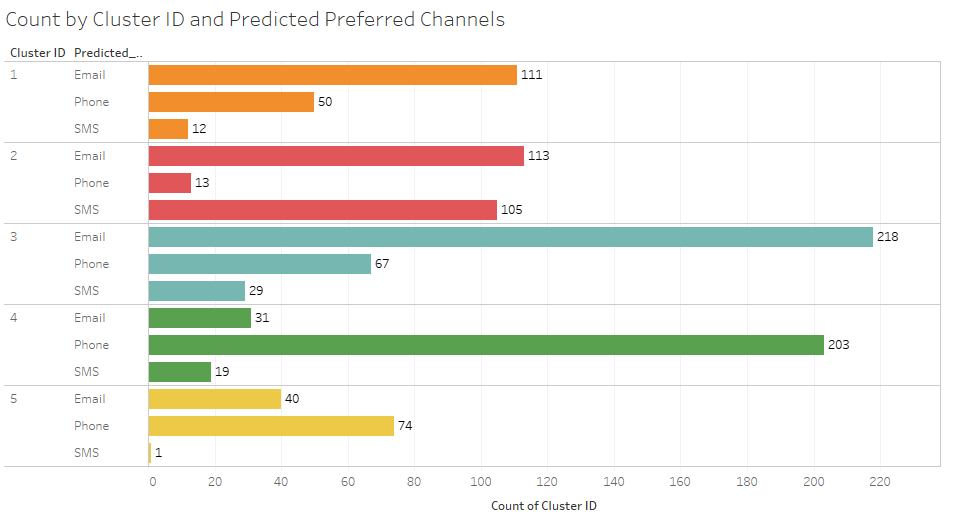


Fig 3.3– Customer count by Cluster ID and Predicted Preferred Channel for the Scored Dataset.

Notably, comprehensive, and detailed visualisations are illustrated here:

<https://public.tableau.com/views/ScoreData_16186052169610/Story1?:language=en-GB&:display_count=y&publish=yes&:origin=viz_share_link>

# **Final report findings for senior management**

After careful data analysis we were able to uncover 5 target segments which we would be able to focus our marketing on and identify areas where we can focus any marketing campaigns on.

This is useful because this dataset specified what the customers’ preferred communication channel, we were able to train an AI model on this dataset and can now apply this AI model to their datasets where customers had not specified their preferred marketing communication channel. It does this by identifying which customer fits into which target segment or cluster and choosing the dominant marketing communication channel based on the cluster which the customer belongs.

Additionally, by splitting our data into several different target segments we can tailor our marketing campaigns based on their unique characteristics of the different segments. A list of each target segment is below and recommendations based on how we could market to each different group.

**Cluster 1**

Cluster 1 represents individuals who all have health insurance with ABC with over 50% having level 3 the highest level of insurance, what also separates these people is that most of them have several dependent adults and children on their health insurance indicating that cluster 1 is made of parents, primarily mothers as over 80% are female.

This group is very risk averse, are generally affluent and typically have purchased a home as opposed to renting, because of these reasons we believe that cluster 1 is our most promising target group and should be a primary focus of any marketing campaign.

These customers can be targeted for the home insurance product with health and safety pitches, they are most likely to buy the product since they care about their children and are financially stable. What we found was this group is most receptive to email and phone as their preferred channel.

**Cluster 2**

Cluster 2 is overwhelmingly made up of young people and they have a bundled motor insurance scheme. This tells us that Insure ABCs bundled motor insurance must be popular with people 18 – 25, who are buying their first car insurance. We believe that Insure ABC should not focus on this group as young people overwhelmingly are either renting or living with their parents, therefore marketing targeting this group will likely be wasted as most will not own a home to insure.

Expensive marketing campaigns should wait until these customers become older and are in a position where they are likely to have purchased a home, when this happens this target segment generally prefers communication via SMS or through email.

**Cluster 3**

Cluster 3 is all people who are living in an urban area, it is mostly made up of middle-aged men, of the ones who have motor insurance their vehicle value is relatively high. We think this is a key target segment for Insure ABC to pursue, as customers in this segment are generally affluent and live in urban centres where home insurance is important.

A marketing campaign should focus on issues that plague the individuals in this segment such as home burglaries and vandalism. Our model predicts this segment will respond best to emails overwhelmingly, with the second best being by phone.

**Cluster 4**

The members of these clusters are middle-aged people living in rural areas, they are typically less well-heeled than some of our other target segments, but we still think that it has some potential as a target segment.

This customer as predicted by our model prefers to communicate overwhelmingly by phone, this indicating that they may respond well to telemarketing. This could be because people living in rural areas typically have poor Wi-Fi and many prefer to use landlines over mobile phones making communication via SMS and email not useful.

The needs for rural people to have home insurance are much different than that of those living in urban areas, this distinction can be key in communicating the value proposition. Marketing campaigns towards these customers should focus on home insurance in the event of a natural disaster such as floods.

**Cluster 5**

Our final cluster are typically older people living in rural areas, what separates this target segment is that they are typically more affluent than cluster 4, we believe this to be one of our key target segments along with cluster 3 and cluster 1.

This individuals within this cluster prefer communicating by phone or by email, which is broadly similar to cluster 5, similar marketing could be focusing on things like natural disasters would be a key focus for this segment.

## **References**

* Altman, N.S., 1992. An introduction to kernel and nearest-neighbour nonparametric regression. *The American Statistician*, *46*(3), pp.175-185.
* Freedman, D.A., 2009. Statistical models: theory and practice. Cambridge university press.
* Ho, T.K., 1995, August. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition* (Vol. 1, pp. 278-282). IEEE.
* John D. Kelleher, Brian Mac Namee, Aoife D'Arcy - Fundamentals of Machine Learning for Predictive Data Analytics\_ Algorithms, Worked Examples, and Case Studies-The MIT Press (2020)
* Kamiński, B., Jakubczyk, M. and Szufel, P., 2018. A framework for sensitivity analysis of decision trees. *Central European journal of operations research*, *26*(1), pp.135-159.
* MacKay, D., 2003. An example inference task: Clustering. In Information theory, inference and learning algorithms (Vol. 20, pp. 284-292). Cambridge, UK: Cambridge Univ. Press.
* Maimon, O. and Rokach, L. eds., 2005. Data mining and knowledge discovery handbook.
* McLachlan, G.J., 2004. *Discriminant analysis and statistical pattern recognition* (Vol. 544). John Wiley & Sons.
* Tolles, J. and Meurer, W.J., 2016. Logistic regression: relating patient characteristics to outcomes. *Jama*, *316*(5), pp.533-534.

# **Appendix**

**Work Log**

19th - 20th March – Abhishek and Robert assess situation and consider analytical approaches.

19th - 20th March – Muzammil and Jack Create data quality report.

22nd March – Muzammil and Jack conduct data pre-processing, upload clean data and create an ABT in SAS cloud.

12th - 13th April – Robert and Abhishek perform the clustering in SAS and Python.

13th -14th April – Muzammil and Jack visualise data in Tableau, Excel, and SAS.

13th – 14th April – Abhishek creates Random Forest prediction model.

14th - 15th April – Abhishek and Jack use random forest to make predictions.

15th – 16th April – Report is written by Muzammil, Robert, Abhishek, and Jack.

**Graphs and figures**

|  |  |  |
| --- | --- | --- |
| **Feature** | **Insurance Training File** | **Insurance Scoring File** |
| CreditCardType | Missing values (17.65%) | Missing values (18.8%) |
| Occupation | Missing values (38.04%) | Missing values (36.6%) |
| Gender | Inconsistent data (low) | Inconsistent data (Very low) |
| Age | Outliers (low) | Outliers (Very Low) |
| MotorValue | Outliers (low)  Missing values (17.82%) | Outliers (low)  Missing values (17.20%) |
| MotorType | Missing values (17.82%) | Missing values (17.20%) |
| HealthType | Missing values (37.82%) | Missing values (36.62%) |
| HealthDependentsAdults | Missing values (37.82%) | Missing values ( 36.62%) |
| HealthDependentsKids | Missing values (37.82%) | Missing values ( 36.62%) |
| TravelType | Missing values (48.46%) | Missing values (46.92%) |
| PrefChannel | Inconsistent data (low) | NA |

Fig 1.a) Overall data quality statistics

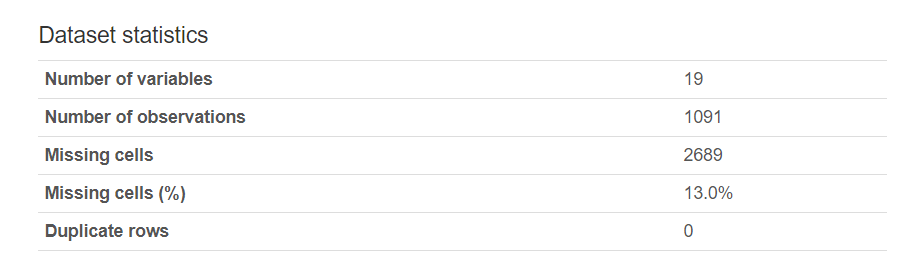


Fig 1.b) Dataset statistics – QUB\_Insurance\_Scoring

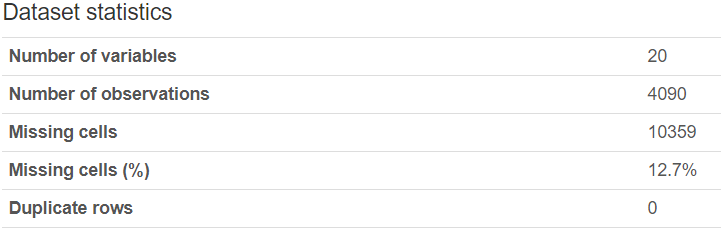


Fig 1.c) Dataset statistics – QUB\_Insurance\_Training



Figure 1.d) Categorical Features – Statistical Numbers – Training Dataset



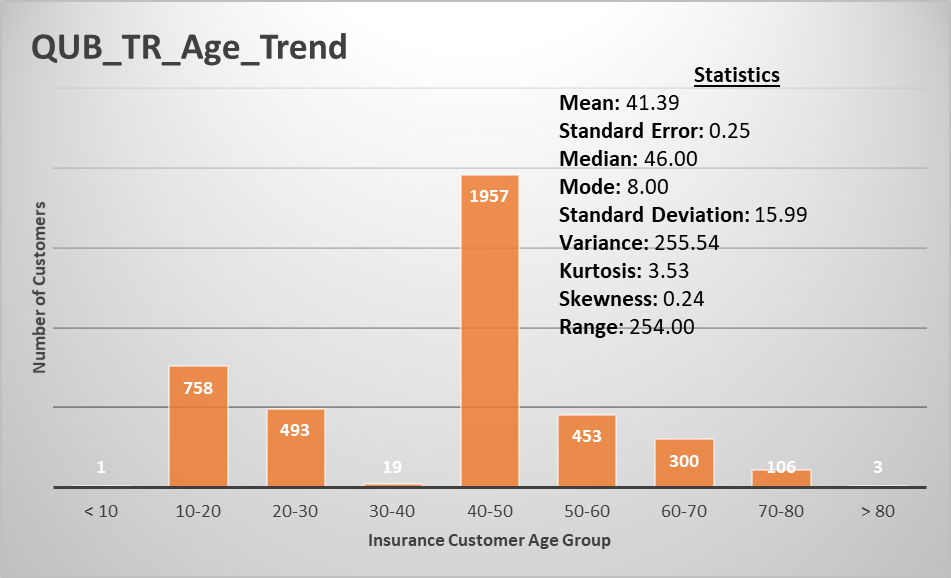
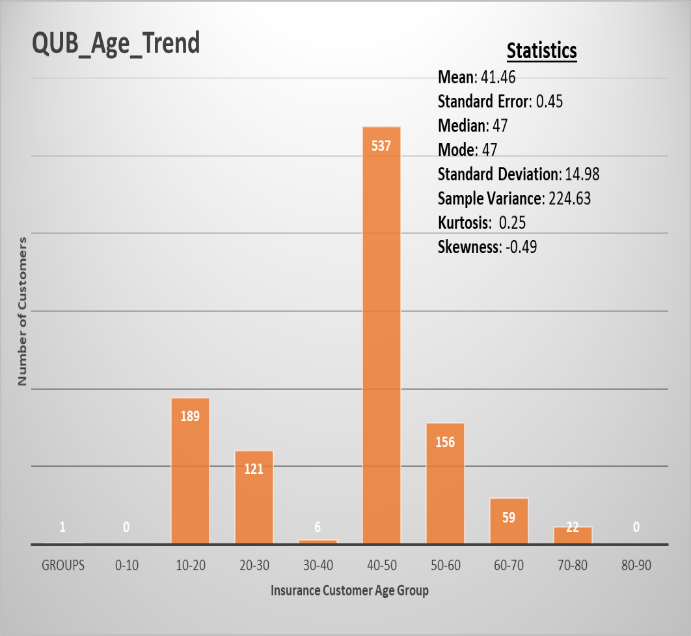
Figure 1.e) Categorical Features – Statistical Numbers – Scoring Dataset

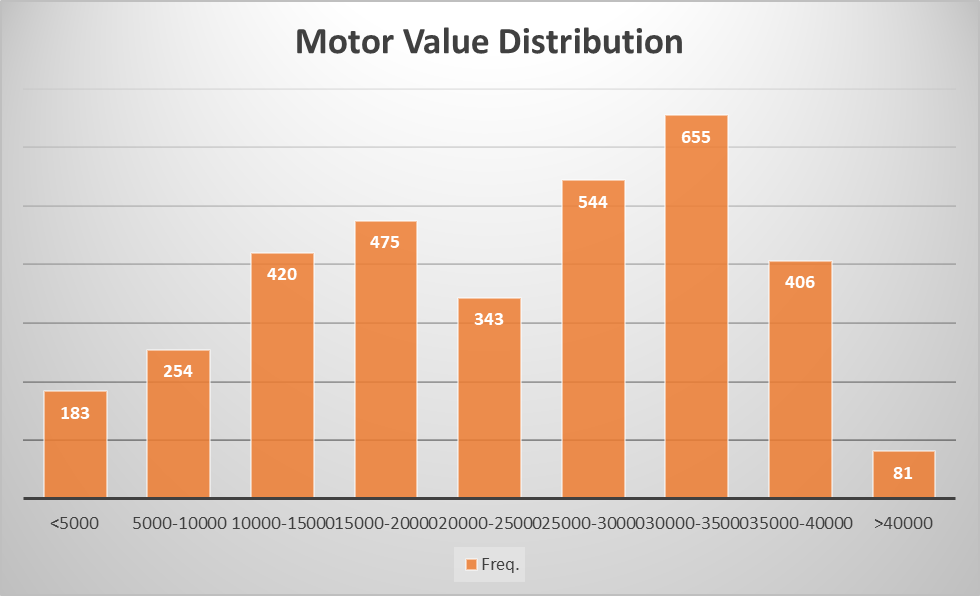
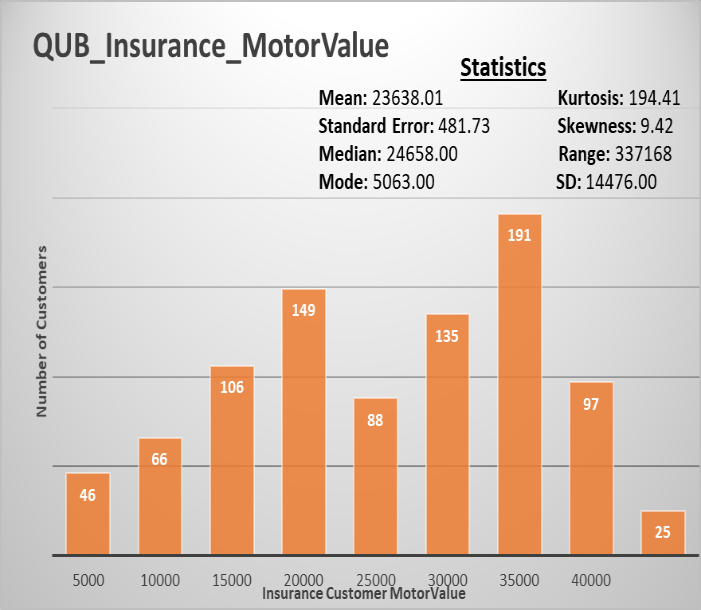


Figure 1.f) Continuous Features – Statistical Numbers – Traning Dataset



Figure 1.g) Continuous Features – Statistical Numbers – Scoring Dataset

Figure 1.1 – Comparisons and analysis based on qualitative data – Age

 Figure 1.2 – Comparisons and analysis based on qualitative data – MotorValue

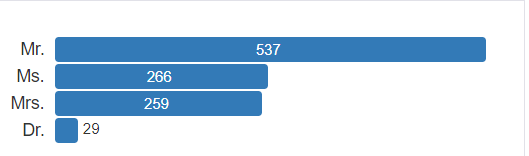
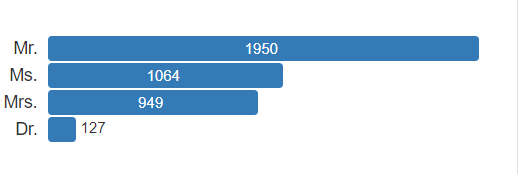


Figure 1.3 – Comparisons and analysis based on clean categorical data – Title

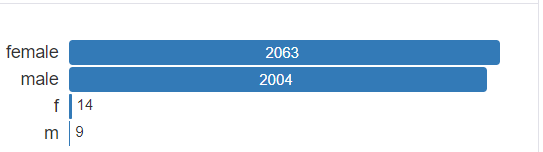
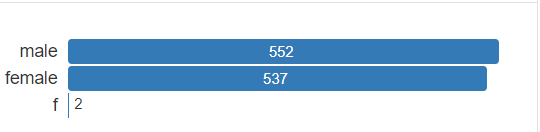


Figure 1.4 – Comparisons and analysis based on clean categorical data – Title

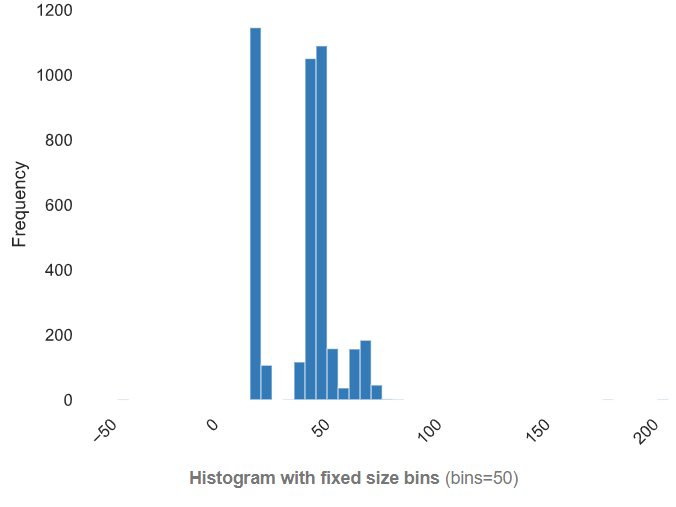
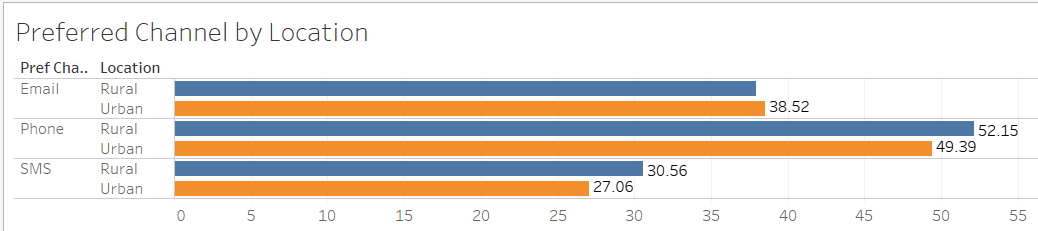


Fig 1.5: Frequency distribution for customers by age

 Fig 1.6: Average age frequency distribution for customers by location

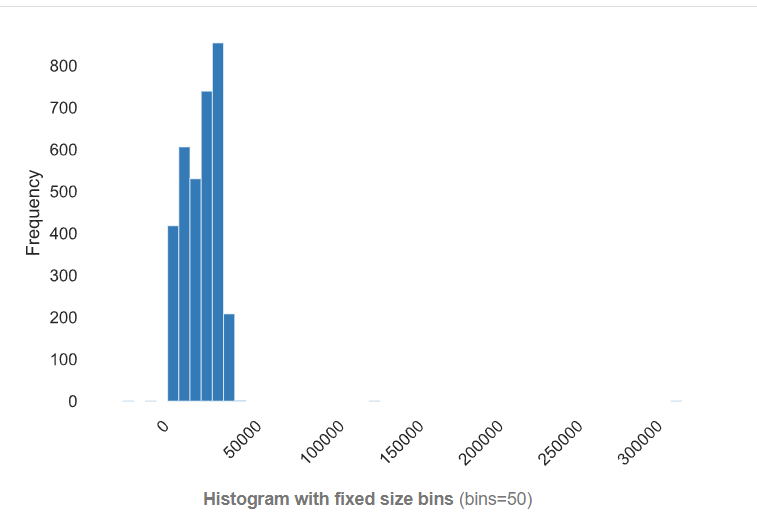
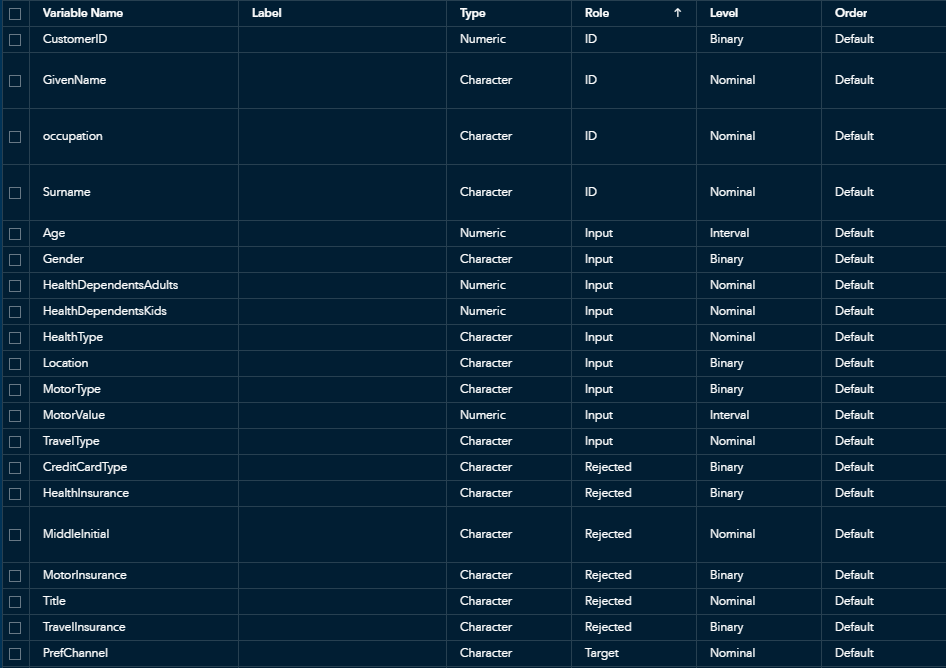


Fig 1.7: MotorValue histogram, right skew.

 Fig 1.8: Clustering inputs provided to the model.

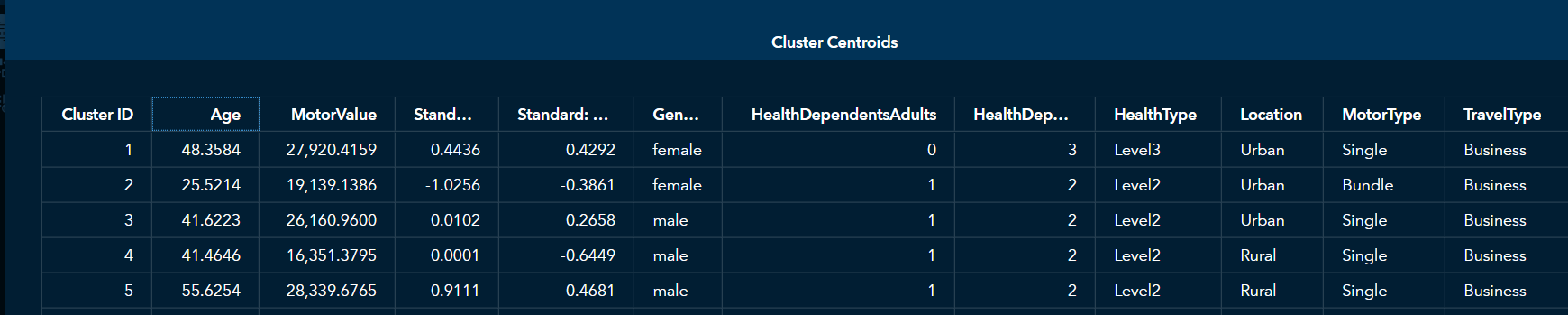
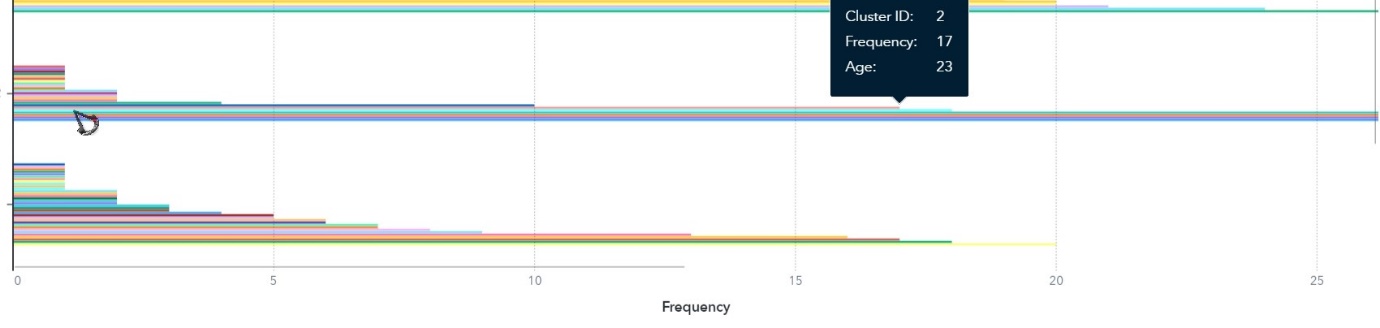


Fig 1.9 – Cluster details and centroids provided by the statistical model.

Fig 1.10: Cluster representation – Cluster 2 grouped by age.

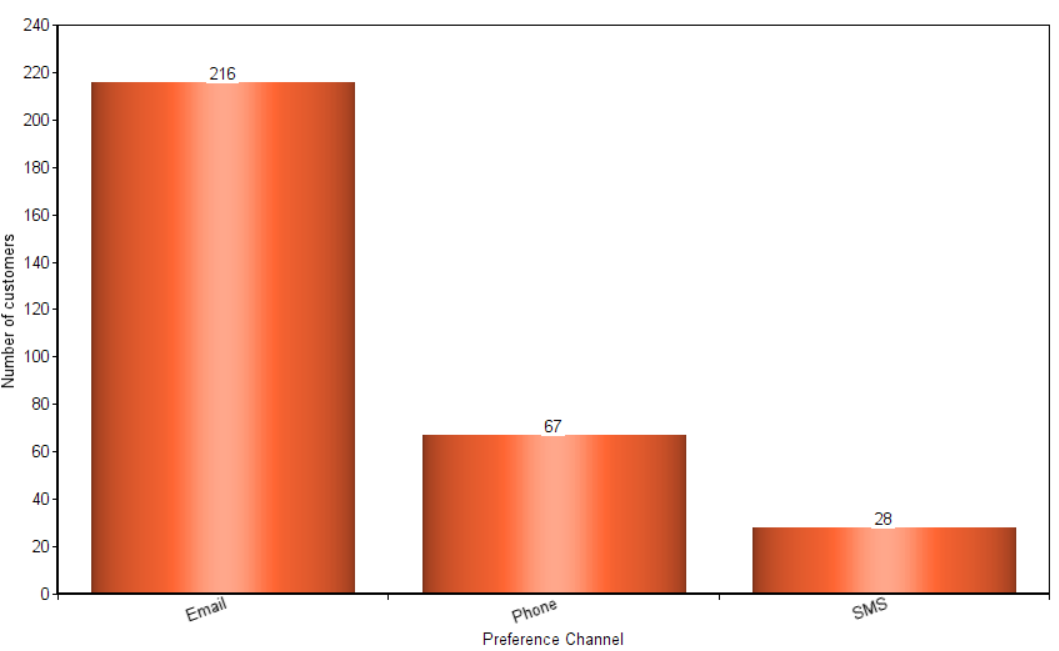


Fig 1.11: Customer preference channel (Scored Dataset) – Cluster 3.

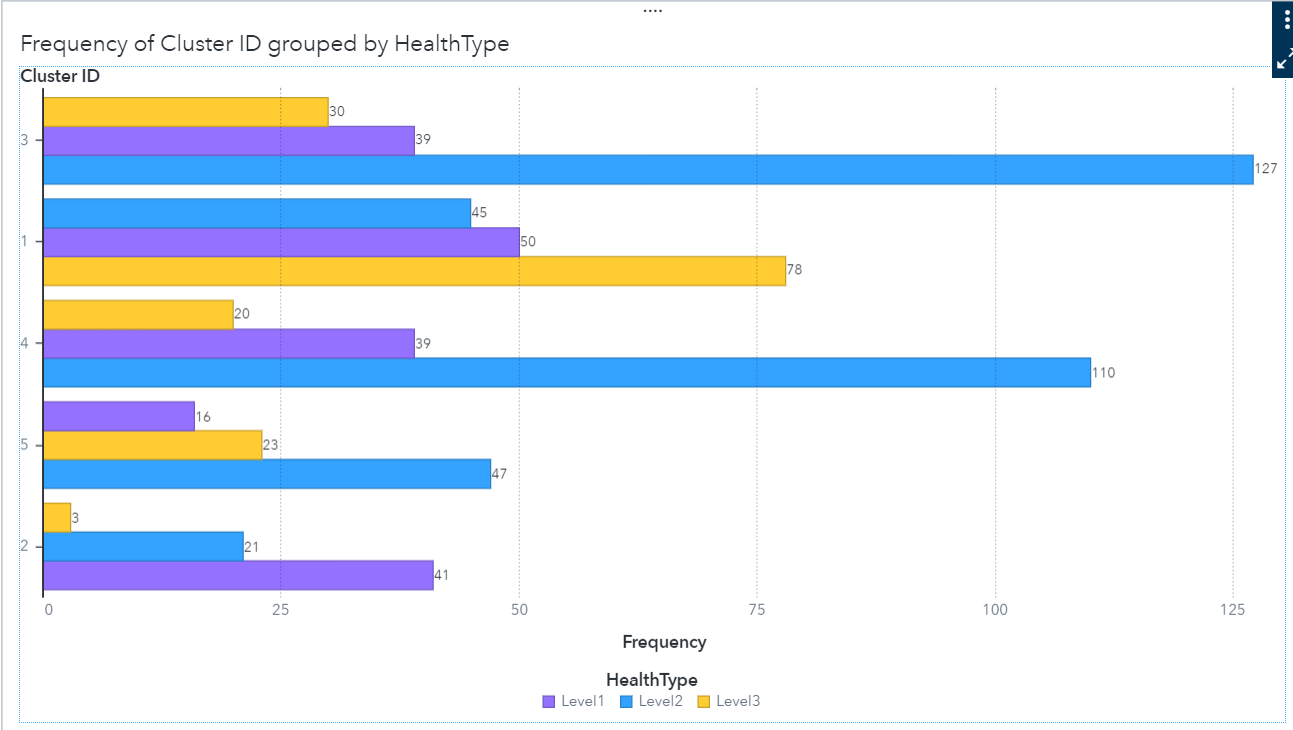


Fig 1.12: Customer count by Health Insurance Type

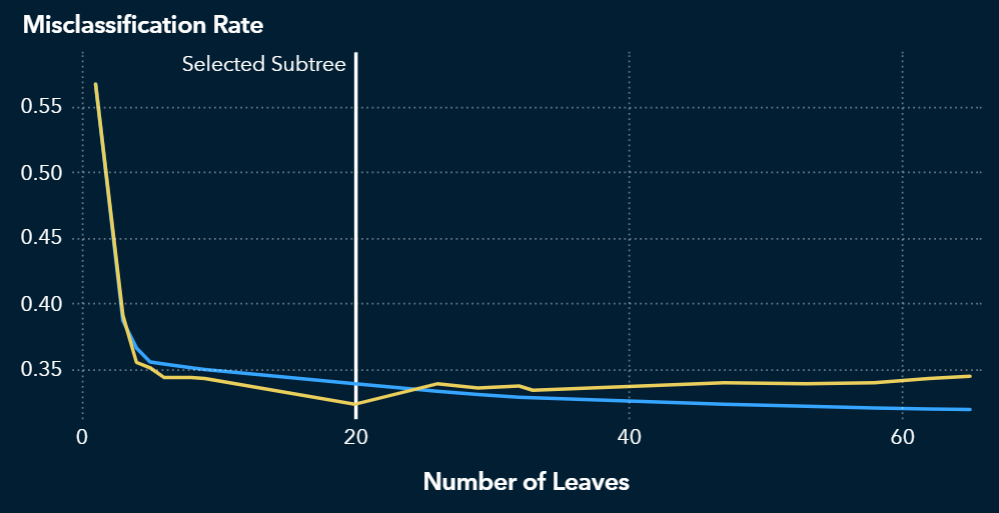


Figure 1.13 Misclassification rate of decision tree

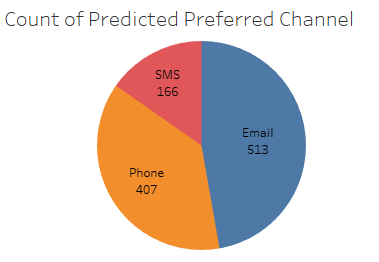


Fig 1.15: Customer count by Predicted Preferred Channel.