ABC INSURANCE COMPANY

Marketing Strategy for Home Insurance

This document is the submitted in form of a report & is submitted to UCD Smurfit Business School inpart fulfilment of the requirements for the degree of Master's in Business Analytics.

Group 21

Nivedha Venkatraman	**20200034**
Akshay Hande	**20200052**
Shweta Soni	**20200233**
Siva Thirumavalavan	**20200235**

Contents

Introduction

In a report published by Deloitte in October 2020 on The future of Home & Motor Insurance, the main findings stated that "customers favour simplicity; they want products that are easy to understand, purchase and use, they also want to be confident that they are paying a fair price, and to trust that claims will be paid fairly; with this insight, insurers can use a human-centred approach to redesign products and exceed customers' expectations". "Internet-connected offerings, such as motor telematics and home insurance linked to home sensors, cause apprehension among many customers; they do not feel comfortable sharing data from car and home sensors with insurers". In a business article by Facebook- "When the time comes to purchase or renew insurance, people have a variety of resources at their fingertips to help them—from discovery to purchase. Today, online channels provide an alternative to traditional conversations with insurance agents in person and on the phone." It also provided facts such as 22% of shoppers discover property insurance online of which 83% customers compared 1-3 brands offerings, and 80% customers were highly influenced by recommendations from family and friends. Whereas, for millennials, 32% of them discovered their options online, 76% of them drawing comparison between 1-3 brands and 95% of them being influenced by family & friends, 38% purchasing the insurance online.

With the decision of purchasing insurance highly influenced by recommendations, it is important to provide quality service to existing customers. ABC Insurance company currently has 4091 customers who have purchased Health, Travel and/ or Motor Insurance products. ABC is now planning to launch Home Insurance product with it's initial outreach for the existing customers through a preferred mode of channel for communication. This report is a detailed explanation of customer segments, their distinct behaviour and marketing & product strategy which will offer in hand recommendations for ABC's home insurance launch. This report will also highlight the process flow, it's scalability and scope of improvement.

The data used for this analysis had variables describing the demographic and Insurance policy details for each customer.57% customer belong to Urban, 82% customers hold credit card, 82% customers hold motor insurance, 42% customers who hold motor insurance also hold travel insurance. 78% customers who are aged between 45 to 60 years have a health insurance policy. 43% customers prefer email as their channel, 38% prefer phone and the rest prefer SMS.

Customer Segments

Based on only available quantitative data (Age & Motor value) we performed initial clustering. This approach has helped us to identify three distinct clusters. Underlying profile of all customers in cluster are later identified using addition available categorical variables. To look at the profiles of customer we looked at existing available variables as well as the features we have created for this analysis. List of following variables were used to understand the underlying profiles of customer: Salary Indicator, Family Indicator, Location, Age buckets, Motor Type, Health Insurance Type, Number of Dependents (Adults & Child), Travel Insurance Type & Married Flag. Using clustering & segmentation approach, we distinctly identified three profiles. Keeping in mind our Business Objective we have classified customer profiles on priority. Here are the details of these customer profiles.

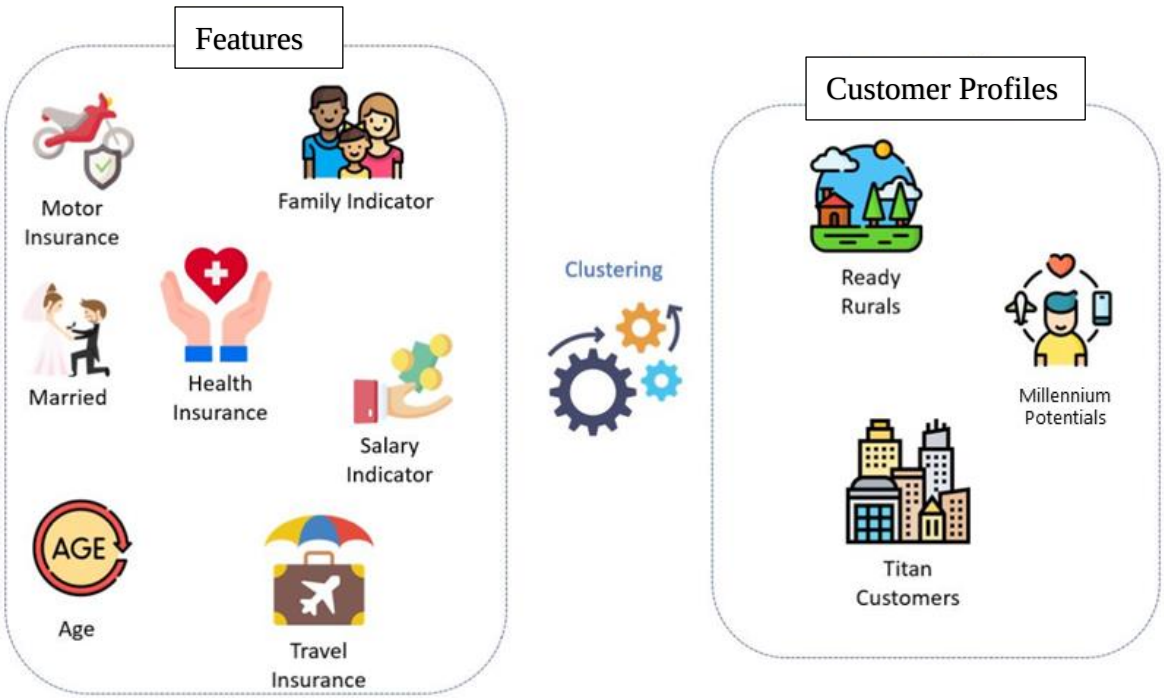


Fig 1: Customer Profiling

Titan Customers

Customers belonging to this profile are our High Value customers and are to be given the top priority to sell our new Home Insurance product.

Demographic Behaviour:

- Middle aged and older adults: They fall under the age bucket >36. Out of 1483 customers, 1054(71%) customers are middle aged (36-50) and 429(29%) customers are older adults (>50).
- Townsfolk: Out of 1483 customers, 1063(72%) customers live in urban area.
- Large Family: Out of 1483, 1084(80%) customers have opted for a Health Insurance and majority (77%) among them have covered their dependents indicative of having a large family.
- Financially Well-Off: Majority (89%) of customers belonging to this cluster fall under the High - Medium Salary Buckets as most of them have purchased motor insurance of either medium or high value, opted for 2 or 3 insurance products and higher levels of Health Insurance indicative of earning high.

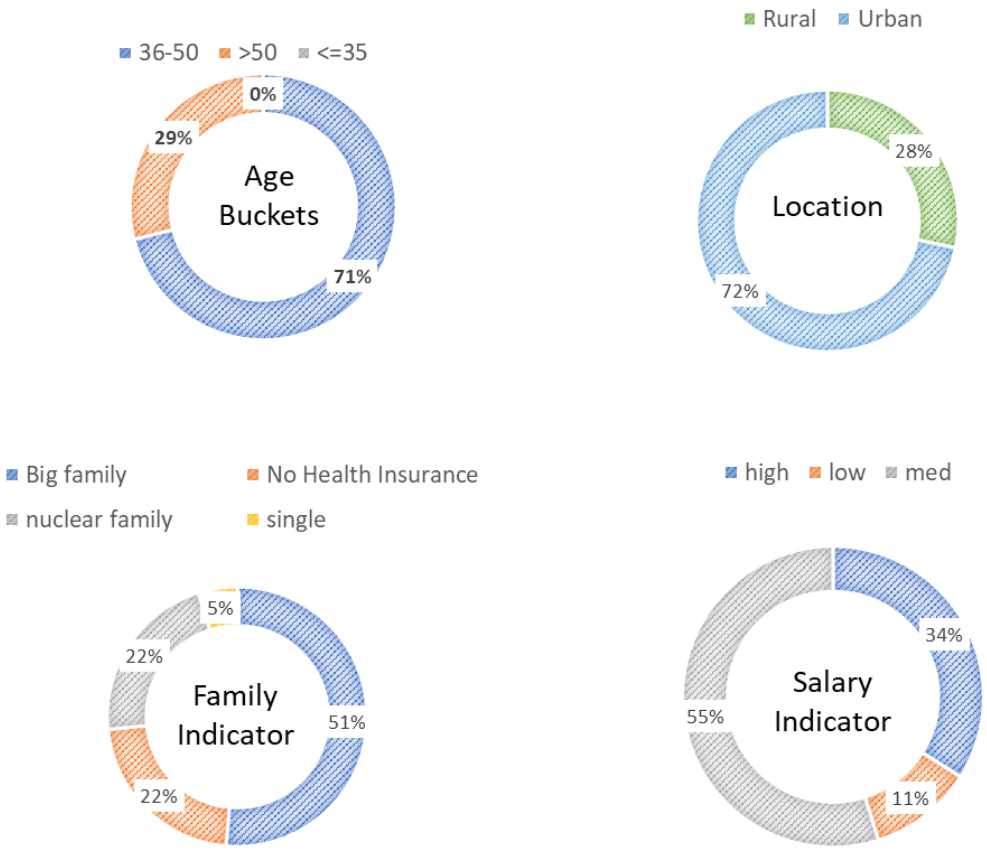


Fig 2: Titan Demographic Behaviour

Insurance Purchase Behaviour:

- Instances of insurance purchased: Majority belonging to his cluster know the importance of having an insurance. Out of 1483 customers, 645(44%) customers have opted for all 3 insurance products and 707(48%) customers have opted for 2 insurance products.
- Health Insurance: Out of 1483 customers, around 1156(78%) have opted for Level 2 or Level 3 Health Insurance indicative of their high spending potential.
- Business Travellers: Most of the customers belonging to this cluster have opted for Business and Senior Travel Type Insurance product.
- Preferred Channel for majority in the cluster: 56% of the customers in this cluster prefer Email and 36% of the customers in this cluster prefer Phone as the channel to purchase the insurance.

Strategy:

Titan Customers are to be given the highest priority to sell our Home Insurance product, since they are trustworthy given that among the 3 insurance products that we currently sell, majority of them have purchased 2 and more than 2 of our products. They have mostly opted for level2 and level 3 Health Insurance and Business Travel Insurance as well. Also, they are most likely to own a house as they fall in the middle aged and older adults' group with potentially high purchasing power. There are high chances that these customers purchase our 4th Insurance product (Home Insurance) even if the value of the product is high as they have already purchased the other 3 categories of products from us. These customers understand the need of Insurance & are price insensitive, using this to our advantage, majority of them preferring Email as their preferred channel, should be targeted through Email Marketing campaign to sell Home Insurance Product. Personalized email campaign should be designed catering the need of individual customers. As these customers are not price sensitive, bundling of offers should not be considered. Special Relation Managers should be used to cater these high value customers.

Ready Rurals

Customers belonging to this profile are to be given the second priority to sell our new Home Insurance product.

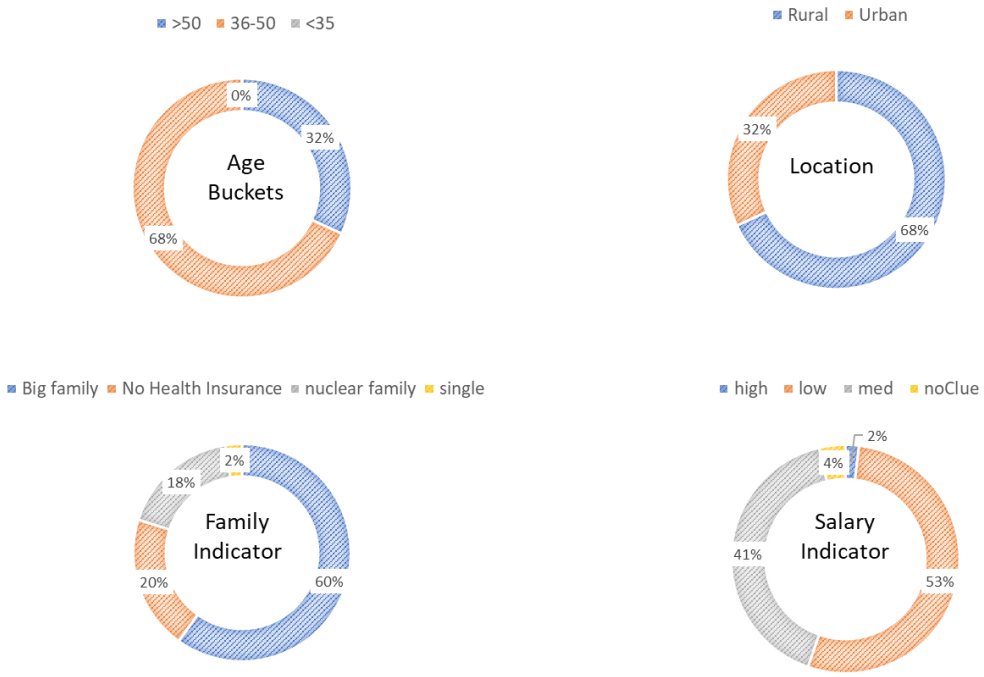


Fig 3: Ready Rurals Demographic Behaviour

Demographic Behaviour:

- Middle aged people: They fall under the age bucket 36-50
- Country-dwellers: Out of 1354 customers, 922(68%) customers live in rural area.
- Large family: Out of 1354, 811(60%) customers have a big family with more than 2 dependants covered in their health insurance.
- Hard Working class: Out of 1354, 1278(94%) customers fall under the Medium - Low Salary Buckets indicative of low to medium income which is evident from their value of motor insurance, count of insurances purchased and Health insurance level.

Insurance Purchase Behaviour:

- Instances of insurance purchased: Out of 1354 customers belonging to this group, 1152 (85%) customers have opted for 1 or 2 of our insurance products.
- Health Insurance: As majority in this cluster have a large family, they know the importance of Health Insurance and out of 1354 customers, around 1083(80%) have opted for Level 1 or Level 2 Health Insurance.
- Sedentary lifestyle: People belonging to this group are not much into travelling, though 885 have opted for Single type Motor Insurance, only 391 have purchased the travel insurance policy.
- Preferred Channel for majority in the cluster: 70% of the customers in this cluster prefer Phone as the channel to purchase the insurance.

Strategy:

Ready Rurals are to be given the second priority to sell our Home Insurance product, since majority of them have a large family with an average earning. They are price sensitive and exhibit the characteristics of a traditional customer. They are most likely to own a house as they fall in the middle age group and have already purchased our Level 1 and Level 2 Health Insurance covering their dependents', so they might also opt for purchasing the new Home Insurance product from us. Most of them belonging to this cluster do not have high earnings and live in rural, we can target them by giving discounts on the Home Insurance so that they can afford getting one from us. As majority of them prefer Phone as the preferred channel, specialized Tele-sales executive should be hired for targeting the customer base of Ready Rurals while targeting Home Insurance product. Bundled offers of Health Insurance & Home Insurance should be offered for this customer base.

Millennium Potentials

Customers belonging to this profile are to be given the least priority to sell our new Home Insurance product.

Demographic Behaviour:

- Young age people: They fall under the age bucket less than 35.
- Townsfolk: Out of 1253 customers, 828(66%) customers live in urban area.
- Teenagers: Out of 1253, 949(76%) customers have not opted for a Health Insurance indicative of being single or not having a family and only 226 customers have purchased Health Insurance but have not covered dependents indicative of having a family.
- Hard Working class: Out of 1253, 662(53%) customers have fall under the Low Salary bucket and 475(38%) fall under Med Salary bucket indicative of low to medium income which is evident from the low and medium value of motor insurance, count of insurances and very less Health Insurance policy purchased.

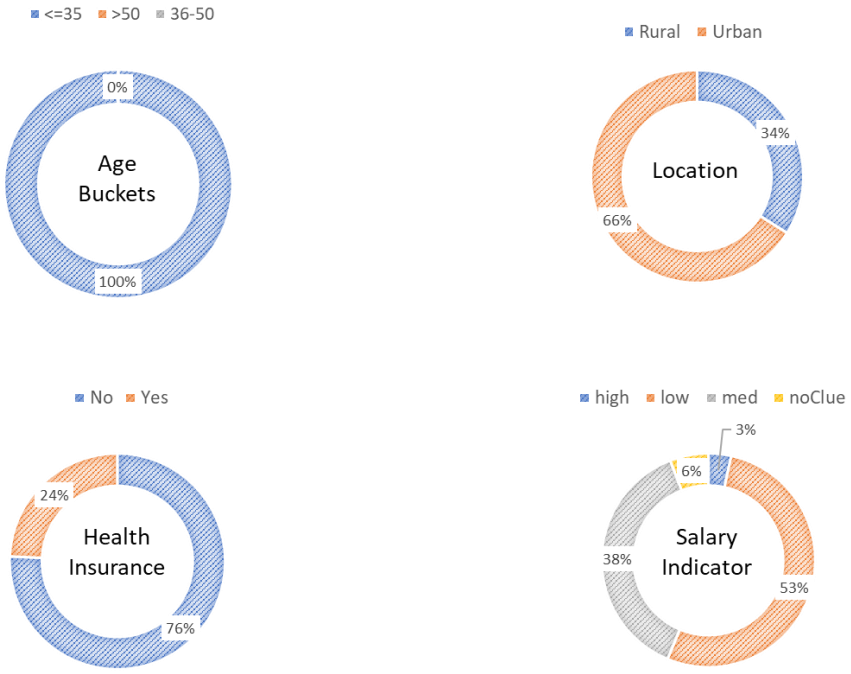


Fig 4: Millennium Potentials Demographic Behaviour

Insurance Purchase Behaviour:

- Instances of insurance purchased: Out of 1253 customers belonging to this group, 1000(80%) customers have opted for 1 or 2 of our insurance products.
- Health Insurance: As majority (76%) in this cluster are noticed to not have a Health Insurance, indicative of being single, there are chances that they are covered as dependents under their parents' insurance policy.
- Explorer: Out of 1253 customers, 876(70%) customers have purchased our Travel Insurance policy which is also symbolic of their young age and 854(69%) of them have opted for a Bundle Motor Type Insurance indicative of attracted to offers and discounts while purchasing an Insurance product.
- Preferred Channel for majority in the cluster: 50% of the customers in this cluster prefer Email and 43% of the customers in this cluster prefer SMS as the channel to purchase the insurance.

Strategy:

Millennium Potentials are to be given the least priority to sell our Home Insurance product as they are price sensitive urban population who do not understand the importance of having an insurance. Majority of them have not opted for a family insurance. But around 25% of customers in this cluster do have a family and there is a rough chance that they will own a house in the near future. We can target these customers by offering worthy discounts on our Home Insurance product. As customers belonging to this profile prefer Email and SMS as preferred channel, Personalized Email & SMS campaign marketing strategy should be used to attract customer. Bundle offers of insurance + noninsurance product (E.g., Gift Vouchers/Concert Passes) should be created to attract customer from this profile. Email & SMS being the most cost-effective communication channel, we should use this medium to push through marketing content.

Process Flow

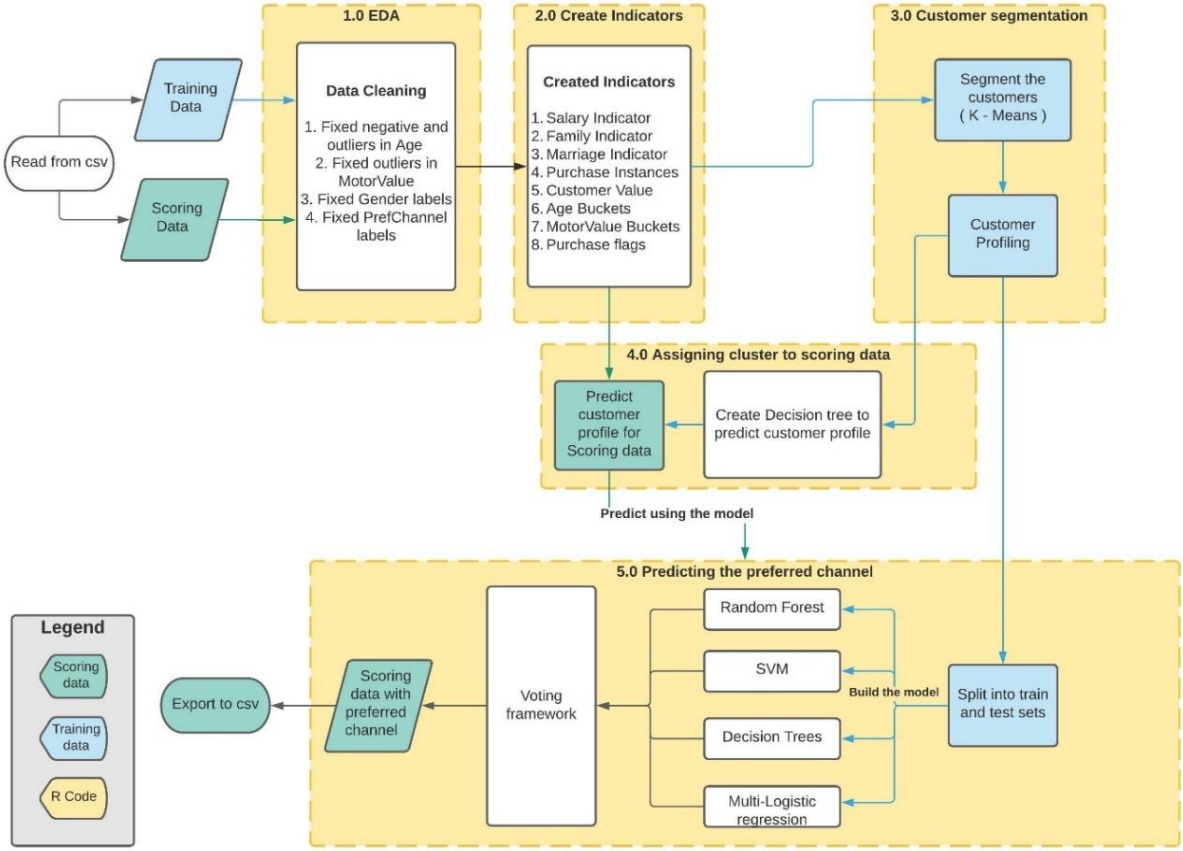


Fig 5: End to End process flow.

The process flow diagram above depicts how we went about solving the problem statement from scratch. We used the open-source application "R" for data cleaning, clustering and modelling. In addition to the report, the required R code files are submitted for reference. Below text here describe the process in detail.

- Data Explorations & Feature Engineering:

  Before proceeding to modelling phase, we started with Data Exploration. Imputation techniques were used to treat unusual values present in the raw dataset (refer *Data quality report* in Appendix). Since the number of records were less, we used technique of imputation over deletion. New features were created to add value to the dataset and aid in the consumer profiling process.

Indicators	Type	Levels	Columns used	logic
CardHolder	Char	"Yes" "No"	CreditCardType	No - NA , Yes - otherwise
Married	Char	"Yes" "No"	Title, dependent kids	Yes - Title = Mrs or Dep.Kids >0 No - Otherwise
AgeBuckets	Char	">50" "36-50" "<=35"	Age	
MotorValueBuckets	Char	"medium" "low" "0" "high"	MotorValue	low - <25 percentile med - 25 -75 percentile High - >75 percentile 0 - missing
MotorFlag	Num	1 , 0	MotorInsurance	1- yes , 0 -No
MotorLevel	Num	0, 1, 2, 3	MotorValueBuckets	0- 0 , 3-high , 2-med, 1-low
HealthFlag	Num	1, 0	HealthInsurance	1- yes , 0 -No
HealthLevel	Num	0, 1, 2, 3	HealthType	0- no insurance , 1- level 1 2- level 2 , 3- level 3
TravelFlag	Num	1, 0	TravelInsurance	1- yes , 0 -No
TravelLevel	Num	0, 1, 2, 3	TravelType	0- no insurance ; 1- Backpacker 2- Standard, Senior ; 3-Premium , Business
Instances	Num	0, 1, 2, 3	MotorFlag, HealthFlag, TravelFlag	MotorFlag + HealthFlag + TravelFlag
customerValue	Num	min - 0 , max- 9	MotorLevel, HealthLevel, TravelLevel	MotorLevel + HealthLevel +TravelLevel
SalaryIndicator	Char	"low" "med" "high" "noClue"	customerValue	1:3 -low ; 4:6 -med ; 7:9 - high ; 0-noClue
FamilyIndicator	Char	"No Health Insurance" "Big family" "nuclear family" "Single"	HealthDependentsAdults + HealthDependentsKids	No HI - No HI; 0 - single 1,2 - nuclear family >2 - big family

Fig 6: Derived Variables

- Customer Segmentation & Profiling:

  After completing initial data exploration steps, Machine learning technique was used to identify underlying clusters in our dataset. To do so, we have used available quantitative variables Age & Motor Values. K-Means clustering technique was applied to process raw data and finalized 3 clusters exhibiting distinct characteristics.

Fig 3: Customer Segments

Customer profiling was performed after studying the characteristics exhibited by these segments. (*Please refer to the customer segmentation part of the report for detailed explanation*).

These profiles reflect consumer behaviour and can assist us in developing appropriate marketing campaigns. We built a decision tree (with 95 percent accuracy) using the profiles generated on the Training Dataset and used it to classify the customers of the scoring dataset.

- Identifying Preferred Channel for individual Customer:

  It was vital to understand the contributing reasons to for a customer to choose a specific preferred channel, even before building a predictive model for it. Hence, we performed a function to identify the variable importance leading to it. We employed these critical variables and the customer segments in predicting the preferred channel. We used advance machine learning techniques like SVM, Random Forest, Decision Tree & Logistic Regression to build the classifiers. These techniques were capturing unique essence of the customers which pushed us to use **Ensemble approach** to predict the preferred channel with the weighted- voting mechanism based on accuracy of individual performing models. The overall accuracy of 70% on training set was observed.

- Validation of Clusters & Model built:

  We classified the customers in the scoring dataset as consumer segments with 95% accuracy.

  The ensemble model to predict preferred channel was built on *Training Dataset*. With the limited data availability, to capture significant customer behaviour with the train set and the purpose of validation met with test set, we opted not to further create a validation set. The ensemble model was run on the test set to see if the model developed was performing well and an accuracy of 65 percent was achieved.

  To further validate the results, the Train and Test sets were merged and exported to csv. When viewed for major preferred channel, in a cluster-wise pivot,

  - 53.1% customers out of 56% preferring Email as their channel in Titan customers, were predicted correctly offering 95% accuracy.
  - 69.8% customers of the 69.9% customers preferring Phone as their channel in Ready Rural, were predicted correctly offering 99% accuracy. And,
  - Similarly, 67% for Email and 69% for SMS as their accuracy of preferred channel for millennium potentials were observed.

- Scoring Data:
  - The scoring dataset was imported in csv, data quality check, variable imputations and new features were created as per the logic used above.
  - The profile segments were assigned using the decision tree.
  - Ensemble model along with weighted ranks was used to predict preferred channels for these customers.
  - This process remains intact for any future data prediction for preference channel and customer segmentation. *(The attached codes can be fully automated, and the predictions will be available in a matter of minutes. For possible scoring, please refer to the user manual in the appendix. This is to ensure scalability)*

Deployment Strategy & Recommendations

- We will incorporate test-control method to further validate the performance of the models. **5%** random allocation from each customer segment will go in **control** and 95% will remain in test.

- The control population will be communicated via random channel (Email/ Phone/ SMS) whereas the test population will strictly be communicated via channels the model predicted.

- After one month of reaching out, we can track whether the customer has responded through the predicted channel and will look at the impact. ***(Please refer to the tracking sheet in the zip folder).***

Segment	Base Customers	Responsive Customers	Collection %
Titan Customers	416	300	72%
Ready rurals	365	250	68%
Millenium potentials	310	200	65%
Grand Total	1091	750	69%

Populate the count of customers segment wise from **C3 to C5** who responded to the campaign after one month. We will then see collection % automatically. **For demonstration purpose, we have included dummy numbers.**

Segment	Control			Test			Impact in NOC	Lift in NOC
	Base Customers	Responsive Customers	Collection %	Base Customers	Responsive Customers	Collection %		
Titan Customers	21	10	48%	395	290	73%	102	1.54
Ready rurals	18	6	33%	347	244	70%	128	2.11
Millenium potentials	16	5	31%	294	195	66%	103	2.12
Grand Total	55	21	38%	1036	729	70%	333	

**NOC is number of customers

Populate the **CONTROL** count of customers segment wise from **C11 to C13** who responded to the campaign after one month. We will then see control collection % automatically. **For demonstration purpose, we have included dummy numbers.**

Populate the **TEST** count of customers segment wise from **F11 to F13** who responded to the campaign after one month. We will then see TEST collection %, Impact in number of customers & Lift of model automatically. **For demonstration purpose, we have included dummy**

- Content of email & SMS campaign would differ across segments as the age group differs across each segment. Titan customers are aged >=36years whereas millennium potentials are <36year old. Location also factors in when majority of Titan live in Urban and majority of ready rurals belong to Rural locality. Hence designing the campaigns would vary and good amount of time should be spent with marketing and advertising professionals as it implies as the first impression of Home Insurance product.

- Personalized emails, SMS with consistent follow-ups should be the go-to strategy. For the segment consisting middle-aged and older people with large family with high salary indicator, marketing strategy should be conveyed in less text rather than information in brochure. The text in the first paragraph should highlight the important takes.

- When reaching out to the customers, **Good-Time & Day** to reach out should identified with previously available campaign data. This approach will ensure maximum response rate and the customer will feel respectful of his time. Once the customer confirms appointment, appropriate performing agent should be allocated the customer.

- Customers with higher vintage should be reached out by same agents as those agents have built in trust & communication which will make the entire process of selling home insurance hassle free.Customers with less vintage and belonging to Titan segment should be reached out by high performing agents having quality communication whereas the customers from ready rurals should be reached out by mid-performing agents and millennium potentials can be reached out by a mix of high-mid and low performing agents.

Scope of improvement

Customer segmentation helps in informing, sensing, and persuading the different segments where the potential users are available. The segmentation will help the ABC company in dividing the market into small segments where the customer needs are identical. With the limited variables shared with us, we would in future like to see if the following features are shared by the organization.

Demographic Segmentation

- Household sector: Salaried class, Self-employed, Retired employees
- Industrial sector: Public sector and Private sector
- Trade sector: Small-scale and Large-scale business
- Institutional sector: Universities, Colleges, Schools, and Institutes
- Level of Education: Graduates, Masters, PHD
- Smoker/ Non-smoker
- House information such as – house alarm existence, smoke detector installed, security locks, how long the person has been living in that house.

Geographic Segmentation

- Region wise: Central zone, Eastern zone, Western zone, Northern zone and Southern zone
- City
- Postal Code
- Region wise – Burglary rate, Natural Calamity prone identification

Behavioral Segmentation

- Spending & purchasing pattern.
- Mode of association with the company (Online/ Via Agent/ Other Sourcing Channel)
- Customer Loyalty
- Customer feedback from surveys
- Agents' association
- Campaign data
- Product information
- Policy details – Premium, Sum Assured, Period of policy, Policy Type – Traditional / ULIP

The behaviors captured above would help us create better customer profiles, make our predictive models stronger with higher accuracy and confidence. The segmentation then created would assist ABC organization in developing marketing promotional campaign. It would be useful in creating awareness of the prospects. Advertisement professionals would tailor advertisement appeals, messages, and campaigns to the target audience's receiving capacity. Given the foregoing, it is safe to say that segmentation is critical for insurance professionals. It turns prospects into customers.

Appendix

1. You must first assess the situation and consider all of the various analytics approaches that could be useful for the business problem described above.

Insure ABC is introducing Home Insurance product in the coming months. The organization as of now don't have any comparative item, however they do offer Travel, Health and Motor insurance at present. The company is presently figuring the promoting methodology for this new item and might want to guarantee that they suitably focus on the current clients destined to purchase the new item.

We approached this situation with initial Data Exploration step and formulated the following set of Hypothesis. These hypotheses were created not just by limiting our thought process to available data features but also understanding the insurance domain.

- Gender will play significant role in sale of new insurance.
- Married customer will help us target segment.
- Salary is missing in the dataset; Indicator should be created to identify spending patterns.
- Location of customer is crucial, Explore this variable in analysis.
- Explore variable "Title" feature – This can be used in customer profiling.
- Occupation variable can be used for profiling, Requires lot of manual efforts.
- Age of customer will be good indicator if customer is interested in purchasing Home Insurance.
- Preferred Channel may have interlinked location & age, Explore this further.

Above hypothesis formed the base for our approach to the analysis. While formulating the hypothesis, we realized importance of missing information in our dataset. Using business logic, we created few indicators which helps us to answer the hypothesis.

Next, to develop marketing strategies, it was important to identify varied segments of the customers by demographics & insurance policy behaviour, for which we went ahead with k-means clustering. After studying the 3 segments of clusters, we identified distinct characteristics for customer profiling. We used advance ML techniques like SVM, Random Forest, Decision Tree & Logistic Regression to build the classifiers. These techniques pushed us to use Ensemble approach to predict the preferred channel with the weighted- voting mechanism based on accuracy of individual performing models. We incorporated test- control method to further validate the performance of the models and as a deployment strategy.

2. Provide a data quality report based on descriptive statistics for each of the variables in the dataset (use both statistical and graphical output). Comment on anything unusual or noteworthy that you see in the data.

When we performed basic health check, we found the following:

Column	Class	Levels	Missing Values (%)	Issues	Treatment
CustomerID	Num		0%		
Title	Char	"Mrs." "Ms." "Mr." "Dr."	0%		used in creating 'Marriage' indicator
GivenName	Char		0%		Removed - No use in Model
MiddleInitial	Char		0%		Removed - No use in Model
Surname	Char		0%		Removed - No use in Model
CreditCardType	Char	"AMEX" "Visa" NA	18%	had NA (Train – 18% , Score – 19%)	renamed NAs as 'No Card'
Occupation	Char	1592 levels	38%	had 1592 levels and 1556 missing values - 38% (train) Had 623 levels and 399 missing values - 37% (test)	used in creating 'isEmployed' indicator
Gender	Char	"female" "male" "f" "m"	0%	Had duplicate levels	Renamed f as female and m as male
Age	Num		0%	had negative values and outliers	made the column absolute, removed the 0s induced by mistake
Location	Char	"Urban" "Rural"	0%		
MotorInsurance	Char	"No" "Yes"	0%		
MotorValue	Num		18%	had outliers	squished the outliers to 5th and 95th percentile, Used for Salary Indicator
MotorType	Char	"Single" "Bundle" NA	18%	NA when the customer doesn't have motor insurance	renamed NAs as 'No Insurance'
HealthInsurance	Char	"No" "Yes"	0%		
HealthType	Char	NA "Level1" "Level2" "Level3"	38%	NA when the customer doesn't have health insurance	renamed NAs as 'No Insurance'
HealthDependentsAdults	Num		38%	NA when the customer doesn't have health insurance	renamed NAs as 'No Insurance', used for Family Indicator
HealthDependentsKids	Num		38%	NA when the customer doesn't have health insurance	renamed NAs as 'No Insurance', used for Family Indicator
TravelInsurance	Char	"No" "Yes"	0%		
TravelType	Char	NA "Business" "Backpacker" "Standard" "Premium" "Senior"	48%	NA when the customer doesn't have travel insurance	renamed NAs as 'No Insurance'
PrefChannel	Char	"SMS" "Phone" "Email" "E" "P" "S"	0%	Had duplicate levels	renamed E, P and S as Email, Phone and SMS respectively

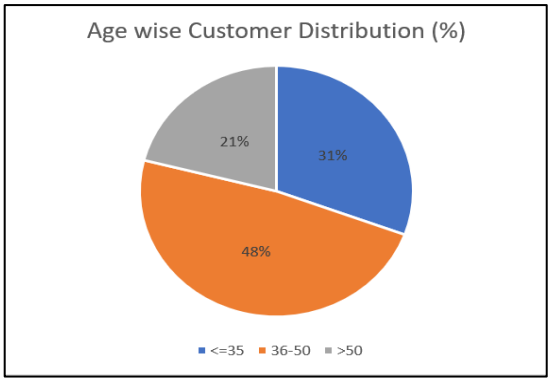
Fig 1: Data Health Check

Mentioned above is the way we went ahead with treating the abnormal and missing values. Below table gives the descriptive statistics for Age and Motor Value.

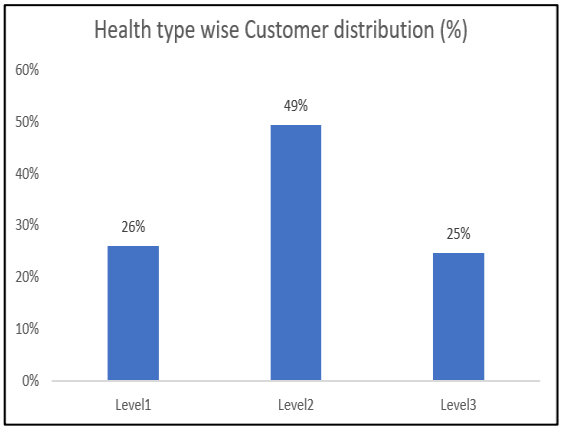
	Age	MotorValue
count	4090	3361
mean	41.4	23450.9
std	16.0	11985.6
min	-44.0	-25686.0
25%	22.0	14837.0
50%	46.0	25045.0
75%	50.0	32289.0
max	210.0	325940.0
Skewness	0.2	4.8
Kurtosis	3.5	121.2
Mode	48	5069
Median	46	25045
1st Percentile	19	5238
99th percentile	73	41583.4

Fig 2: Descriptive Statistics

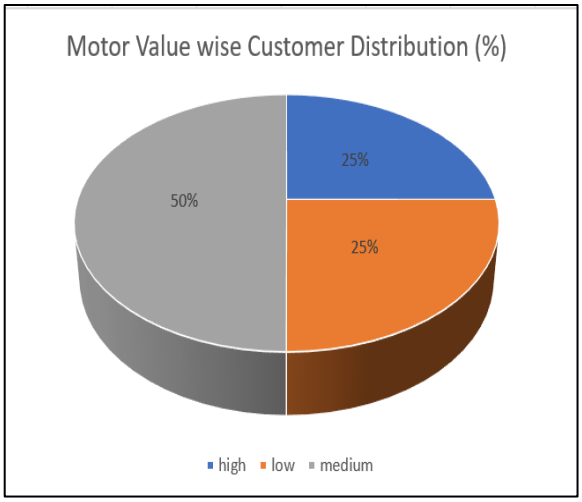
We performed univariate and bivariate analysis. To mention a few interesting insights:



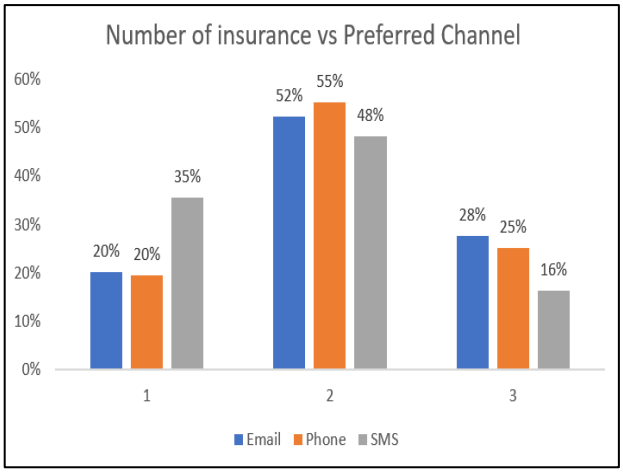
Majority of the customers belong to the age group 36 to 50 years.



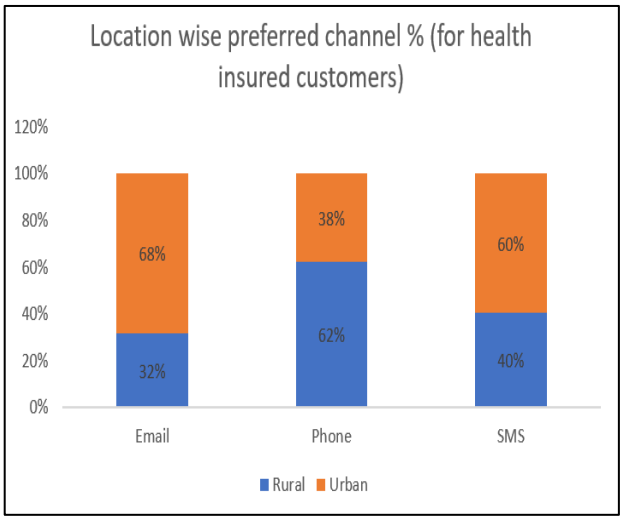
Majority of the health insured customers have Level 2 as their product whereas the purchase of level 1 & level 3 is approximately the same i.e., 25%



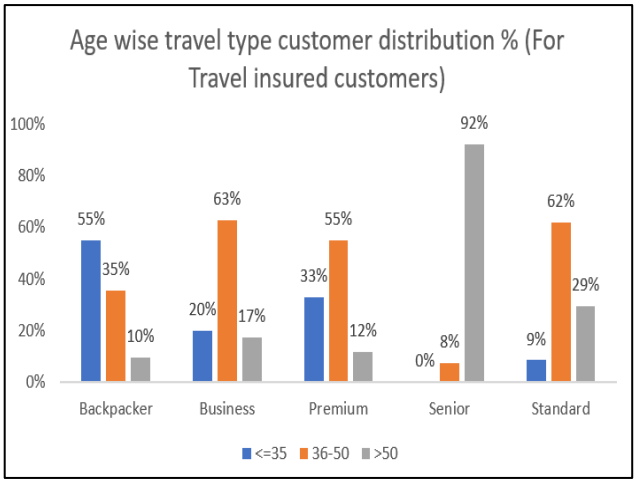
25% of the customers who have motor insurance come under the high value bucket which can be considered as premium customers.



Majority customers who have only one insurance prefer SMS as their communication channel whereas for customers with 3 insurance prefer email.



For customers with health insurance, 62% of the customers with preferred channel as phone are from Rural.



For customers with travel insurance, backpackers include 55% of customers with age <=35 whereas Senior travel type include 92% of >50 years of age.

3. Use the training dataset to create a customer profile (segmentation) by selecting relevant data. Justify your choice of inputs & final cluster solution. Describe the final solution using variables not used in the cluster definition.

After completing initial data exploration steps, Machine learning technique was used identify underlying clusters in our dataset. To do so we have used available quantitative variables Age & Motor Values. K-Means clustering technique was applied to process raw data. Looking at the below scree plot, initially we decided to cluster our raw dataset into 4 cluster.

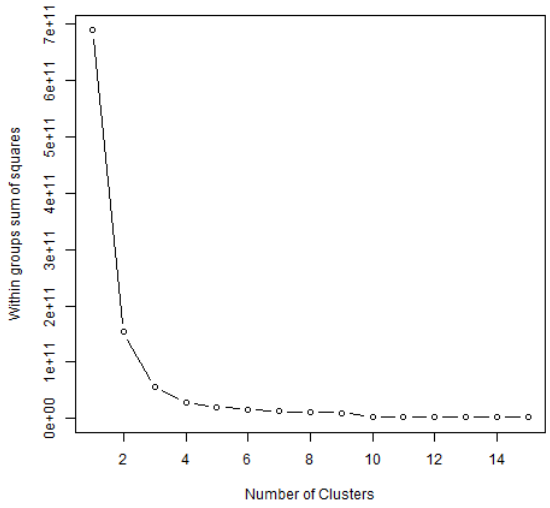


Fig3: Scree Plot showing optimum number of clusters.

After the initial clustering & profiling of customer using features, we created, we realized our dataset exhibits the characteristics of three distinct clusters. Using the Business logic along with statistical knowledge, we finalized that our dataset has distinct THREE clusters. **Please refer to the management report to understand the behaviour of customers which is explained using variables not used in the final cluster solutions.**
Below plot show the different cluster we have identified.

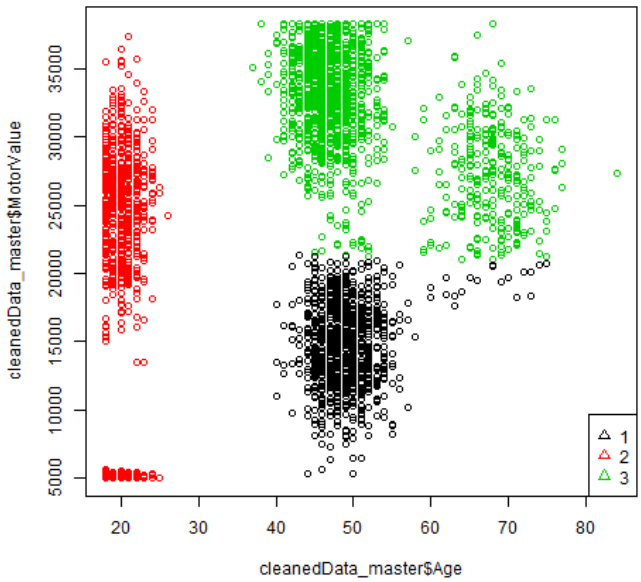
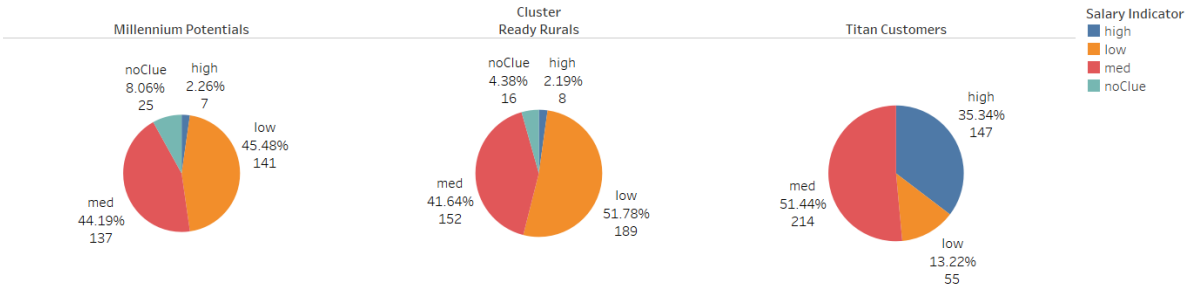
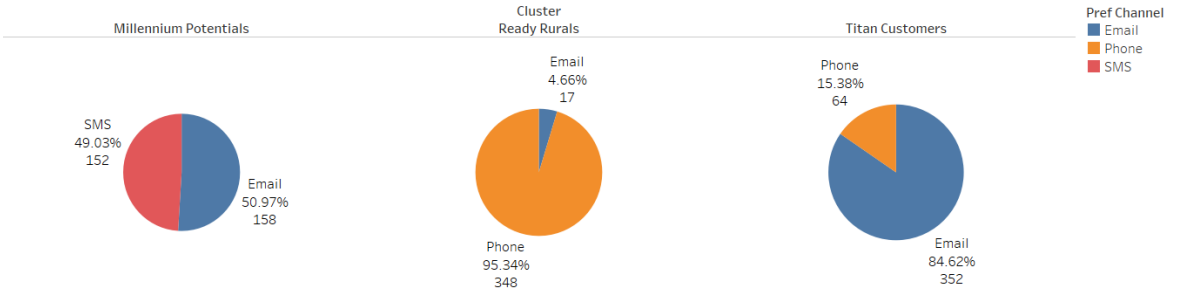


Fig 4: Finalized Clusters

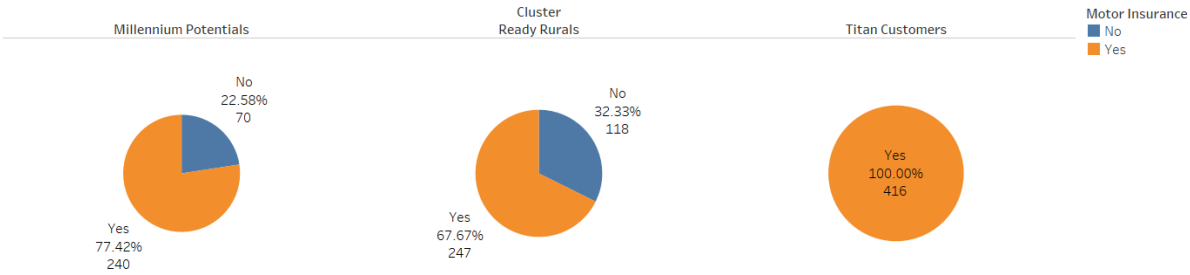
Salary Indicator



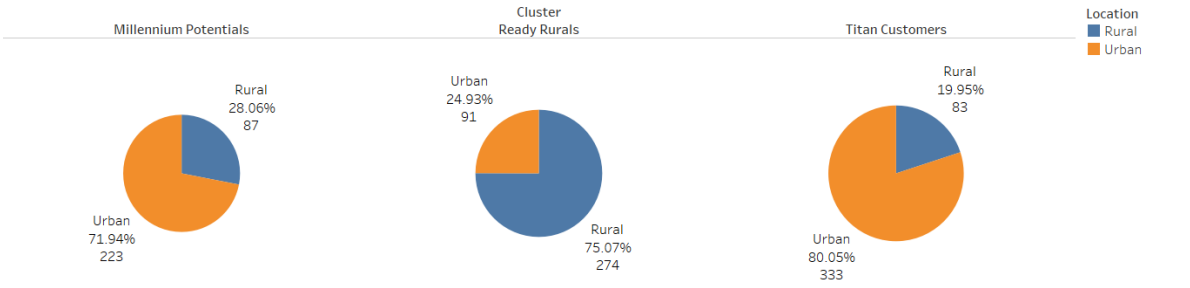
Preferred Channel



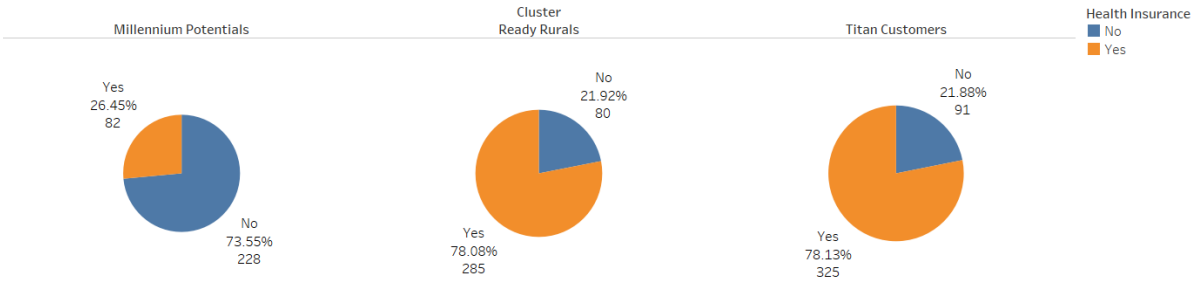
Motor Insurance

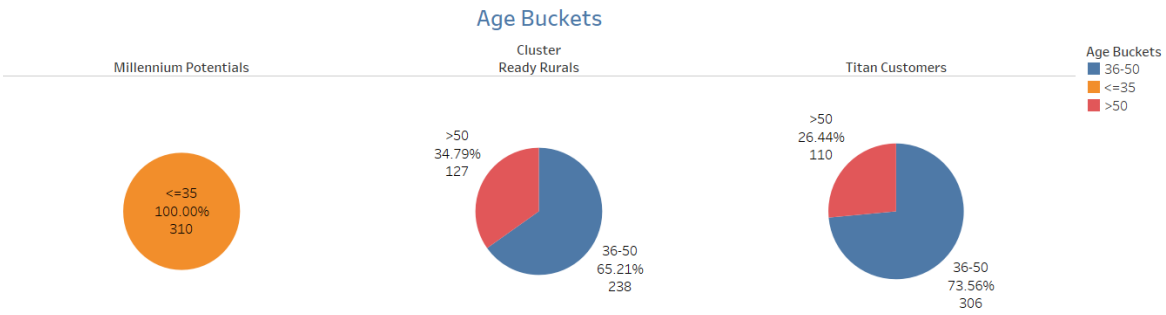


Location



Health Insurance





These identified cluster exhibited underlying characteristic of customer which we were used customer profiling. Features created during the data exploration steps were used to understand different characteristic of the customer. Important variable from the raw dataset were identified before proceeding to model building. For reference PDF file of variable importance is attached. Using Decision Tree Algorithm, we clustered our scoring dataset with accuracy of **95%**. Following chart depicts the decision-tree based clusters we obtained for segmentation process.

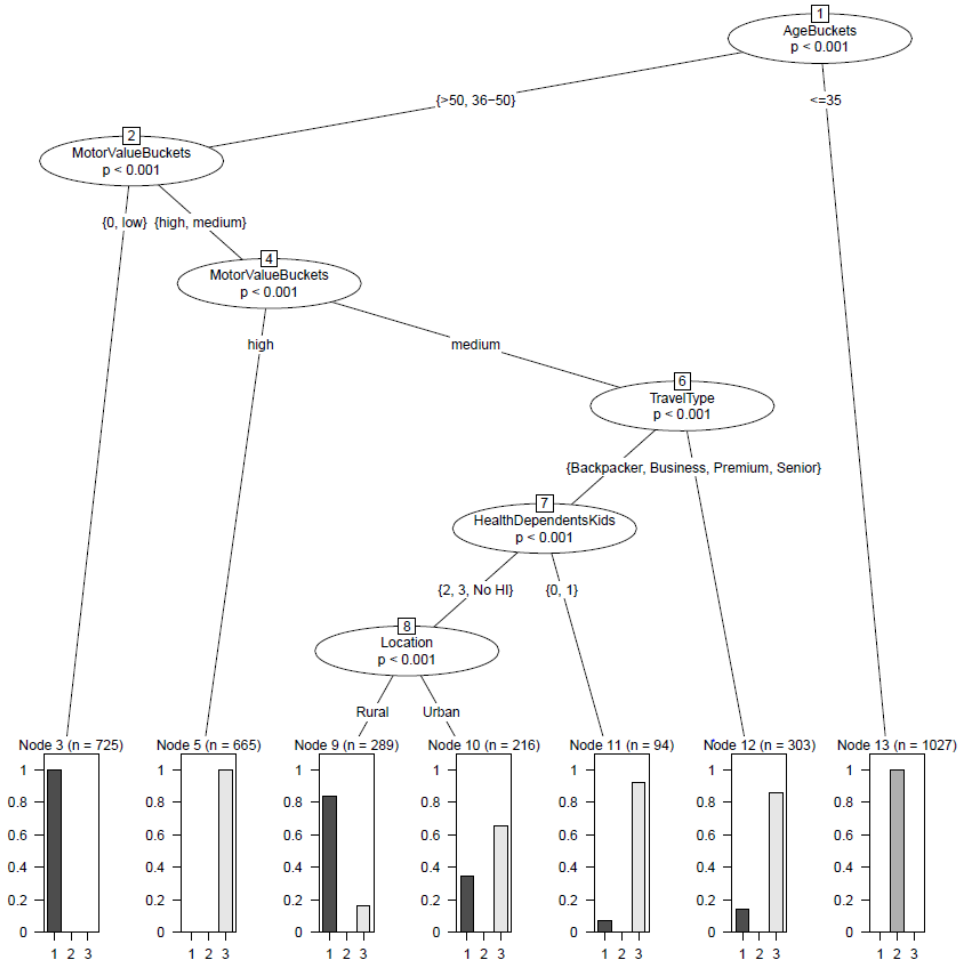


Fig 5: Decision Tree for segmentation

4. Use the training dataset and an appropriate prediction analytics technique to predict which communication channel would be most effective for each customer in the target segment. Describe the final solution using making reference to appropriate model validation to test the accuracy of the predictions.

For selecting relevant variable for ML models, we used advance feature selection algorithm "Boruta". This is how Boruta algorithm works in nutshell.

- Firstly, it adds randomness to the given data set by creating shuffled copies of all features (which are called shadow features).
- Then, it trains a random forest classifier on the extended data set and applies a feature importance measure (the default is Mean Decrease Accuracy) to evaluate the importance of each feature where higher means more important.
- At every iteration, it checks whether a real feature has a higher importance than the best of its shadow features (i.e. whether the feature has a higher Z score than the maximum Z score of its shadow features) and constantly removes features which are deemed highly unimportant.
- Finally, the algorithm stops either when all features get confirmed or rejected or it reaches a specified limit of random forest runs.

Below is screenshot of depicting important variables present in our raw dataset.

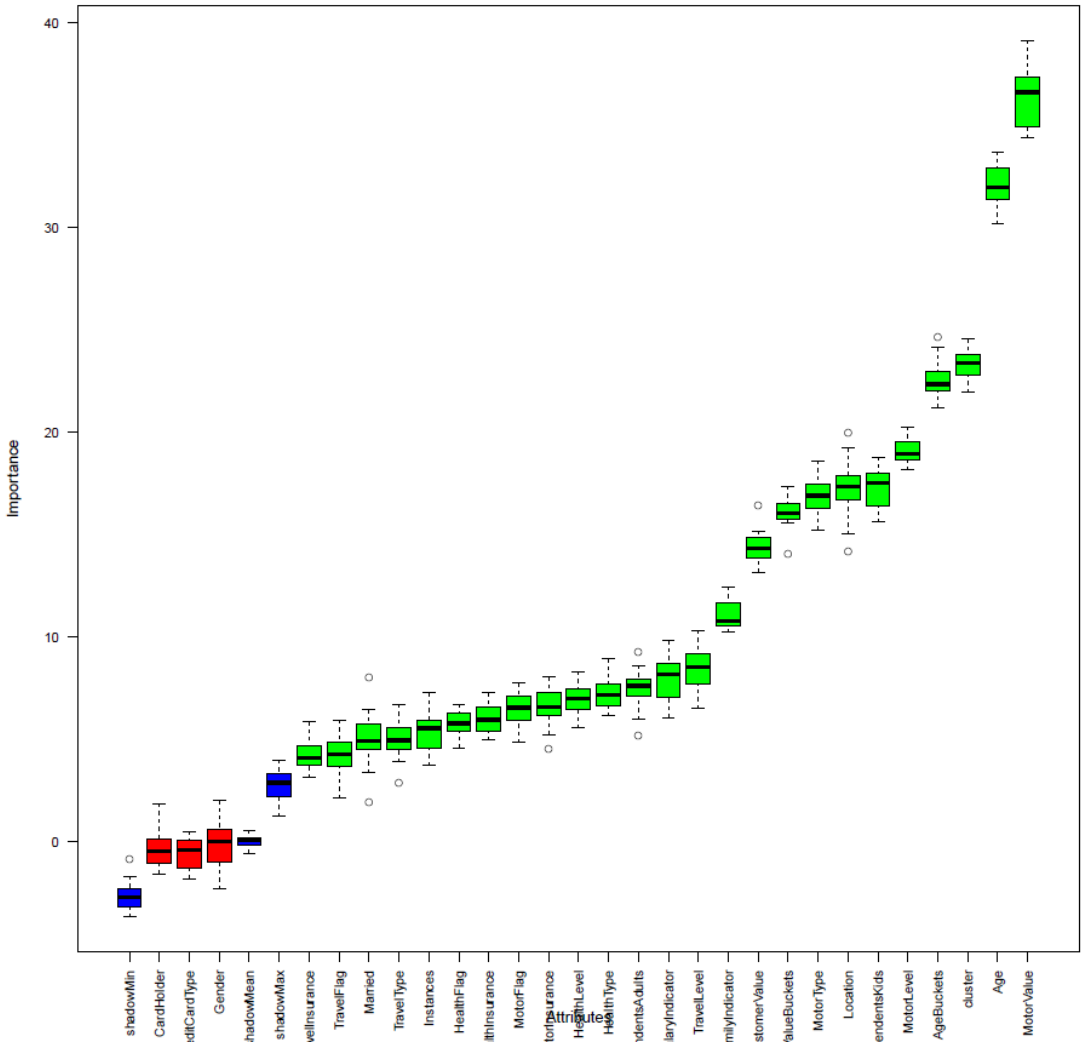


Fig 6: Variable Importance

Using Ensemble approach, we predicted the preferred channel for each customer. Clustered data was divided into training & test with spilt of 80-20. This is approach of splitting data into training & test was consistent across various models we have used. Various advance machine learning technique like SVM, Random Forest, Decision Tree & Logistic Regression were used to build the classifiers.

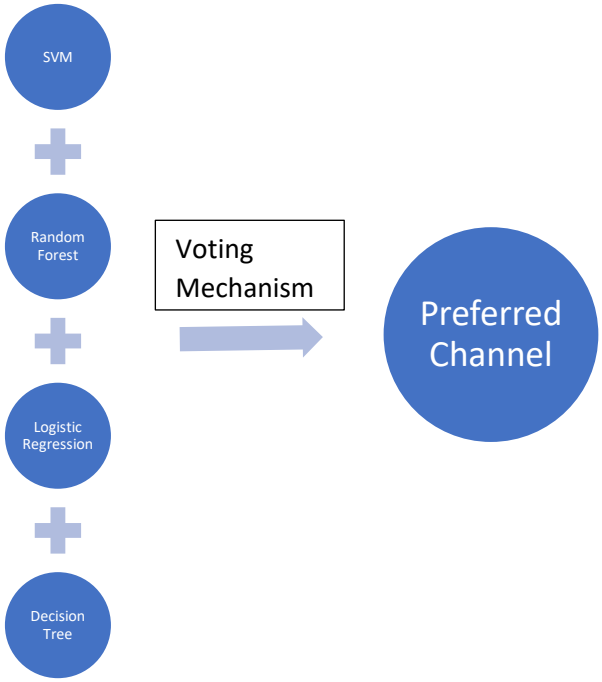


Fig 7: Multiple Model approach in predicting Preferred channel of customer

Voting mechanism was used to predict the preferred channel. Model with highest accuracy rate was given higher weights. These weights helped during voting process & which was used to predict the preferred mode of communication for the customer. With overall accuracy of 70% on training set & 65% of test set we predicted the preferred channel for individual customer.

	Models Created			
	Random Forest	SVM	Decision Trees	Logistic Regression
Train Accuracy(%)	81.6	67	66.3	67
Test Accuracy(%)	64.3	65	64.9	63.8
Weights	0.29	0.24	0.23	0.24

Fig 8: Key metrics of individual machine learning models are listed in table above.

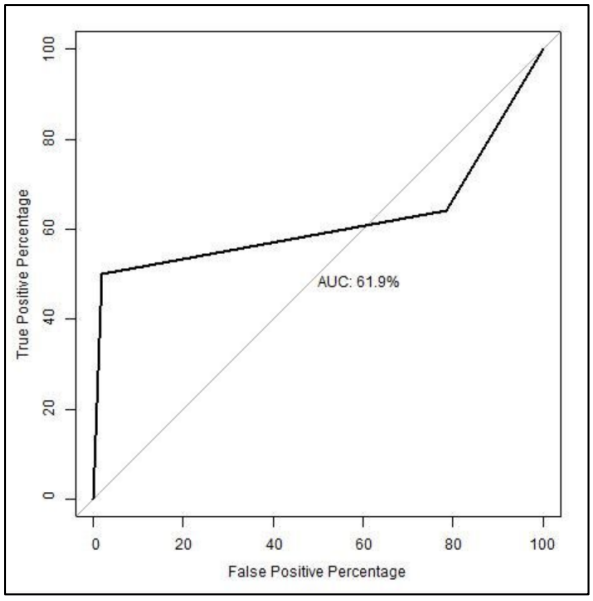


Fig 9: Decision Tree ROC Curve

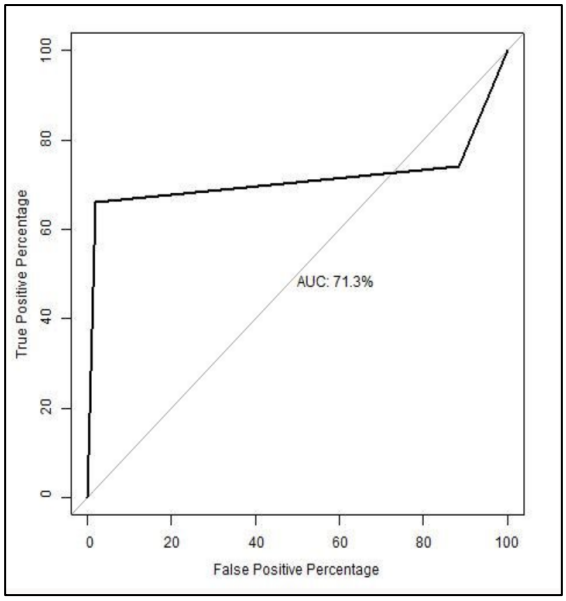
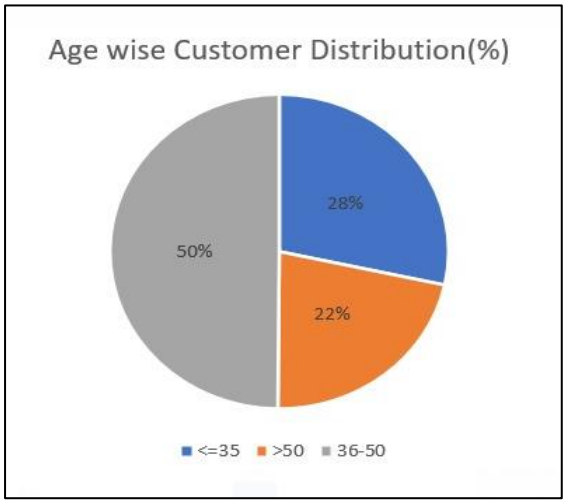


Fig 10: Random Forest ROC Curve

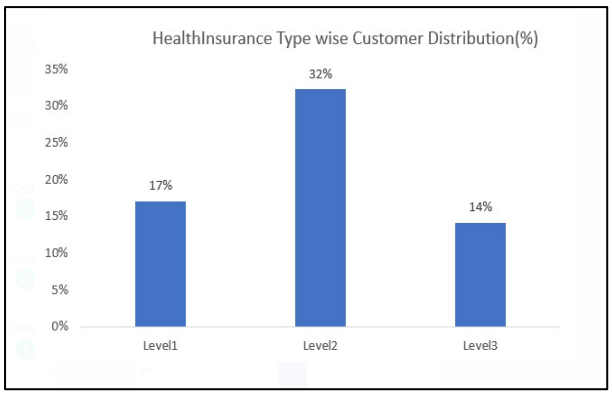
The AUC captured by Decision Tree is 61.9% and for Random Forest is 71.3% on training dataset which indicates good model performance.

5. Use the scoring dataset to apply the models created in step 3 and 4 to the current customer base. Describe the scored dataset using appropriate descriptive statistics and specify how it may be deployed with specific reference to marketing strategy.

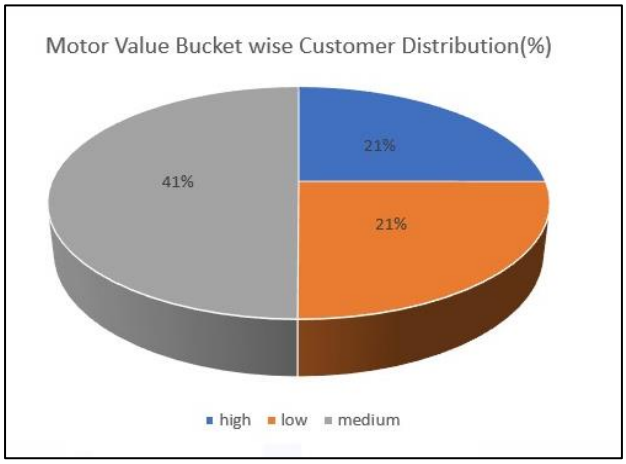
The scoring dataset was segmented and scored as mentioned with steps under Process Flow> Scoring Data in the management report. The proposed method of deployment is included in Deployment Strategy & Recommendations part of the report. Descriptive Stats for these include:



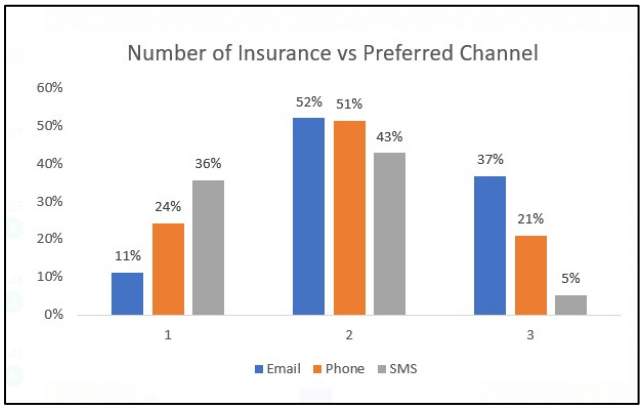
Majority of the customers belong to the age group 36 to 50 years which holds like the Training data set.



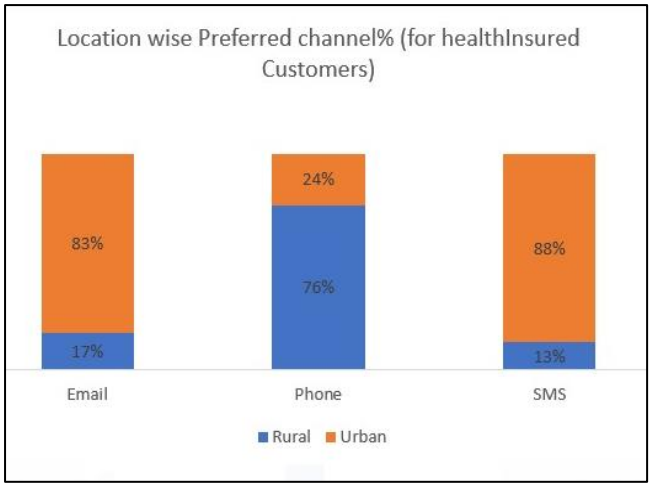
Majority of the health insured customers have Level 2 as their product whereas the purchase of level 1 & level 3 is approximately the same around 15% which is 10% less than the training dataset.



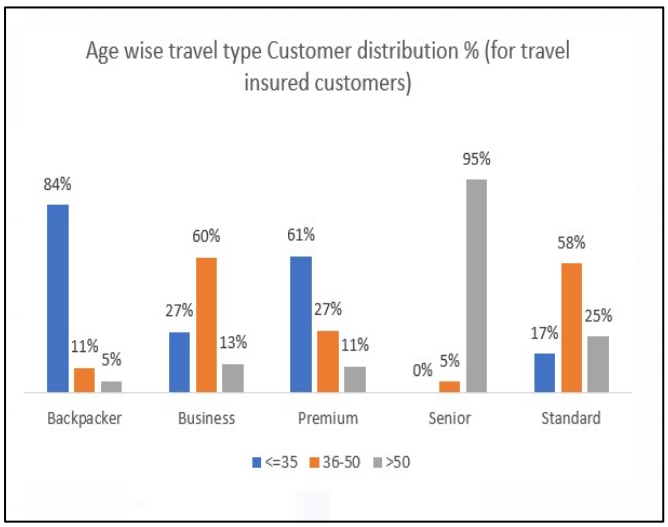
21% of the customers who have motor insurance come under the high value bucket which can be considered as premium customers.



Majority customers who have only one insurance prefer SMS as their communication channel whereas for customers with 3 insurance prefer email which holds in line with the training dataset.



For customers with health insurance, 76% of the customers with preferred channel as phone are from Rural.



For customers with travel insurance, backpackers include 84% of customers with age <=35 whereas Senior travel type include 95% of >50 years of age.

Note: Please refer to the attached zip-file for the model outcomes, analysis reports and R-codes

Work Log

Date	Tasks	M.O.M	Member's responsible for Task
01-04-2021	Introduction Meeting	Understand the concept of Classification and Clustering	All
		Understand the data	`
		Formulate a set of Hypothesis	All
03-04-2021	Exploratory Data analysis	Explore the data and come up with strategies to treat the outliers, missing data	All
04-04-2021	New Indicators	Discuss on logics for new indicators that can add more value to our dataset	All
06-04-2021	Customer Segmentation	K-means Clustering to segment the customers	Siva
08-04-2121	Customer Profiling	Analyse the characteristics of each cluster and Profile them	All
10-04-2021	Build Predictive models	Predictive modelling in R	Siva, Akshay
		Predictive modelling in SAS	Shweta, Nivedha
11-04-2021	Ensemble Approach in modelling	Ensemble approach to predict the preferred channel with the weighted-voting mechanism based on accuracy of individual performing models in R	Siva
13-04-2021	Developing Marketing Strategy	Understand the Home Insurance Market and come up with strategies to market the product	All
13-04-2021	Report Writing		Shweta, Nivedha, Akshay

Member	Student ID	Contribution
Siva Thirumavalavan	20200235	25%
Akshay Hande	20200052	25%
Shweta Soni	20200233	25%
Nivedha Venkatraman	20200034	25%

References

- CCPC Consumers. 2021. Home insurance - CCPC. [online] Available at: <https://www.ccpc.ie/consumers/money/insurance/home-insurance/> [Accessed 16 April 2021].

- Demandjump.com. 2021. Types of Consumers: Who Buys and When. [online] Available at: <https://www.demandjump.com/blog/types-of-consumers-in-marketing> [Accessed 16 April 2021].

- Facebook IQ. 2021. Understanding the journey of the connected insurance consumer. [online] Available at: <https://www.facebook.com/business/news/insights/understanding-the-journey-of-the-connected-insurance-consumer> [Accessed 16 April 2021].

- Deloitte Insights. 2021. The future of home and motor insurance. [online] Available at: <https://www2.deloitte.com/xe/en/insights/industry/financial-services/future-of-insurance-survey.html> [Accessed 16 April 2021].