

Safe Insurance

Graduate Certificate in Customer Analytics - Practice Module Project Report

Submission Date: 22 May 2020

|  |  |
| --- | --- |
| Names | Student Number |
| Norvin Chandra | E0384966 |
| Tan Zhi Yang | E0384308 |
| Lee Suan Ching | E0384993 |
| CI Rui Xiang | E0384938 |

Table of Contents

[1 Background 4](#_Toc41080771)

[2 CRISP-DM Model 4](#_Toc41080772)

[3 Business Problem 4](#_Toc41080773)

[4 Data Exploration 5](#_Toc41080774)

[4.1 Overview 5](#_Toc41080775)

[4.2 Response By state 5](#_Toc41080776)

[4.3 Response by Vehicle size 6](#_Toc41080777)

[4.4 Response By Gender 6](#_Toc41080778)

[4.5 Response By Coverage 7](#_Toc41080779)

[4.6 Response By Education 7](#_Toc41080780)

[4.7 Response by location code 8](#_Toc41080781)

[4.8 Response by Marital Status 8](#_Toc41080782)

[4.9 Response by Policy type 9](#_Toc41080783)

[4.10 Response by Policy 9](#_Toc41080784)

[4.11 Response by Renew offer type 10](#_Toc41080785)

[4.12 Response by Sales Channel 10](#_Toc41080786)

[4.13 CLTV Distribution 11](#_Toc41080787)

[4.14 CLTV Distribution by Coverage 11](#_Toc41080788)

[4.15 CLTV Distribution by Employment Status 12](#_Toc41080789)

[4.16 CLTV Distribution by Sales Channel 12](#_Toc41080790)

[4.17 CLTV Distribution by Gender 13](#_Toc41080791)

[5 Business Problem 1 – Evaluating the effectiveness of Sales Channel 14](#_Toc41080792)

[5.1 Data Exploration Result 14](#_Toc41080793)

[5.2 Action Plan for Business Problem 1 15](#_Toc41080794)

[6 Business Problem 2 - Determining and building relationship with valuable customers 16](#_Toc41080795)

[6.1 Clustering 16](#_Toc41080796)

[6.1.1 Net CLTV 16](#_Toc41080797)

[6.1.2 Calculate Gowers distance 16](#_Toc41080798)

[6.1.3 Silhouette Width 16](#_Toc41080799)

[6.1.4 K=3 Clusters scatter plot 18](#_Toc41080800)

[6.2 Description of each cluster 19](#_Toc41080801)

[6.2.1 Analysis of cluster value 20](#_Toc41080802)

[6.3 Classification of new customers 20](#_Toc41080803)

[6.4 Action Plan for Business Problem 2 22](#_Toc41080804)

[7 Dealing with Poor Renewal Rate - Churn Prediction 24](#_Toc41080805)

[7.1 Business Recommendation 26](#_Toc41080806)

[8 Conclusion 28](#_Toc41080807)

[8.1 Limitations 28](#_Toc41080808)

[References 29](#_Toc41080809)

[Appendix A 29](#_Toc41080810)

[Appendix B 30](#_Toc41080811)

# Background

Safe Insurance PLC is an insurance company in the United States offering general insurance to both personal and corporate customers through its own direct sales channel namely call center and branch. The company also offer insurance via licensed intermediaries namely agent and broker.

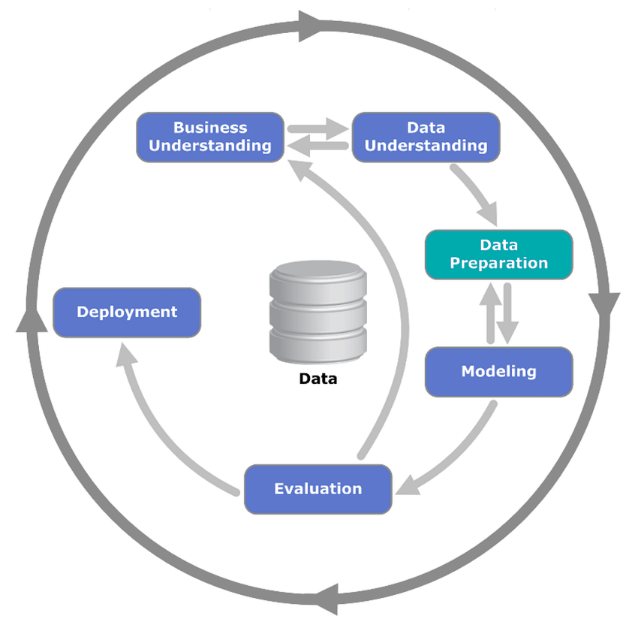
One of the main products that Safe Insurance is offering is motor insurance. However, this product has been facing stiff competition from other insurers resulting in increasing customer attrition rate, i.e. meaning customers are not renewing the policy. At the same time the underwriting profit is also under pressure due to increasing accident frequency and claim amount as well competitive premium. The management has tasked the underwriting team to improve the overall portfolio loss ratio.

The underwriting team has revamped its product and created 4 new products which will be offered to existing policyholders upon expiry of current policy which usually run for 12 months. Due to system restriction, Safe Insurance can only generate renewal invitation based on 1 of the 4 new products.

Safe Insurance Data Analytics team has been enlisted to assist in understanding the portfolio segment, helping in targeting the right customer mix and improving renewal success rate on valuable policyholders.

# CRISP-DM Model

Safe Insurance Data Analytics team have adopted CRISP-DM Model to solve the above business problem.



# Business Problem

The Business problems can be summarized as follows:

1. To identify policyholders that bring value to Safe Insurance and safeguard those from competition
2. Safe Insurance uses multiple channels to acquire business, and there is a need to evaluate the effectiveness of the current sales channels to justify maintaining the diverse channel
3. Improve Safe Insurance portfolio in terms of premium production and loss ratio through improvement on renewal product take up rate on valuable policyholders.

# Data Exploration

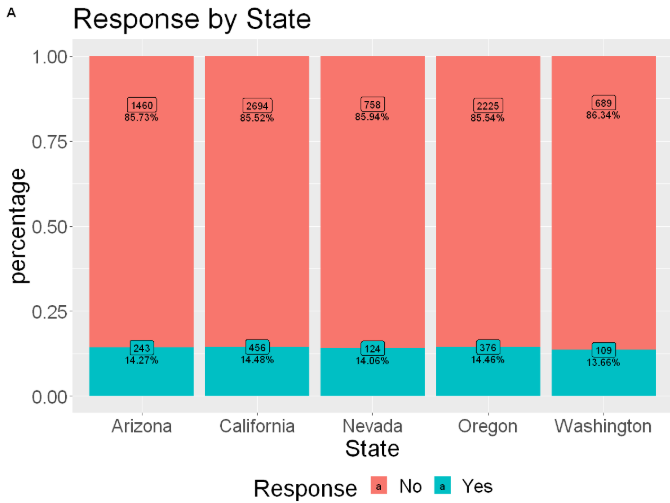
## Overview

The Data Analytics team have conducted Exploratory Data Analysis on the data provided. The data is collected from 9134 customers from 5 states in the United States spanning across January and February of 2011. The data also includes 24 variables. There is no missing or duplicate data.

There are two major dependent variables – the customer lifetime valuable (CLTV) which is pre-calculated by Safe Insurance, and the response to the renewal campaign. Predictor variables include both customer demographic information and transactional data.

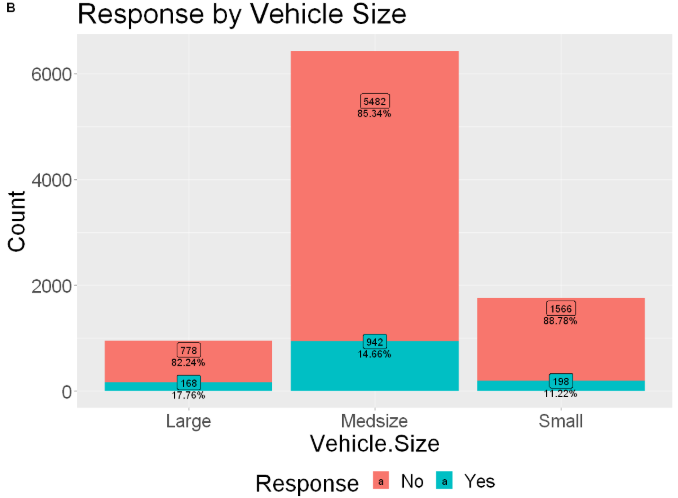
The part below illustrates detailed data exploration on the two dependent variables – CLTV and response to renewal.

## Response by state



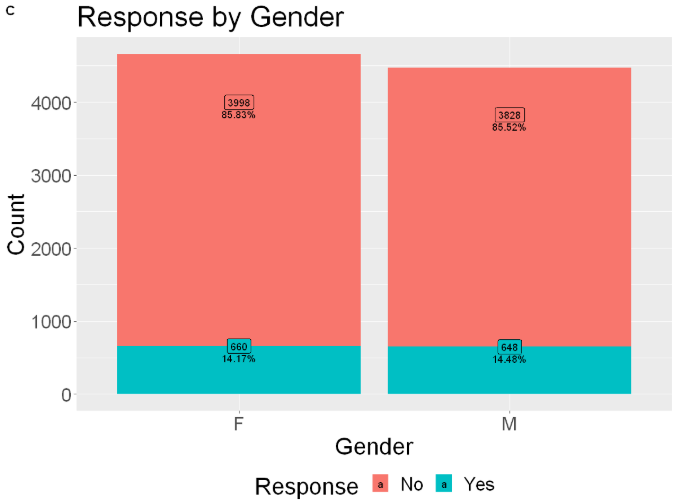
By plotting the number of customers from each state and their responses to the renewal campaign, it can be seen that the largest number of customers for Safe Insurance comes from California. The highest number of positive respondents also comes from California with 14.48% of respondents agreeing to renew their insurance.

## Response by Vehicle size



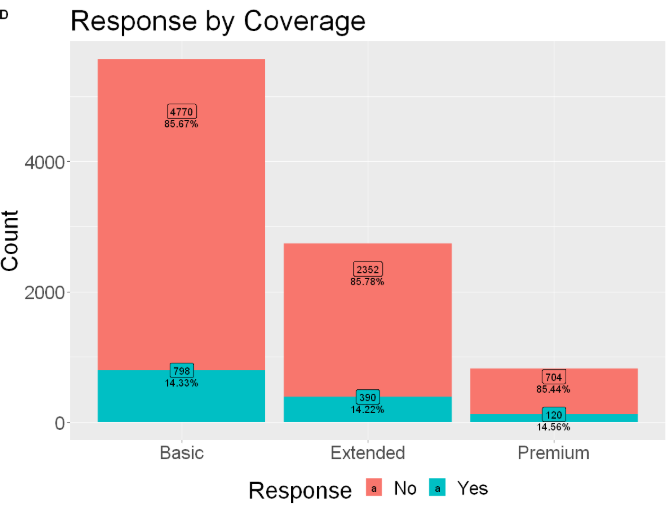
When analyzing the vehicle size by the response, it can be seen that most customers drive medium cars. When the response is overlaid, it can be seen that even-though most customers have medium sized cars, the car size with the best response are large vehicle size, with a 17.76% positive response rate. Hence, Safe insurance should target customers with large vehicles as they are more likely to respond positively to a sales campaign.

## Response by Gender



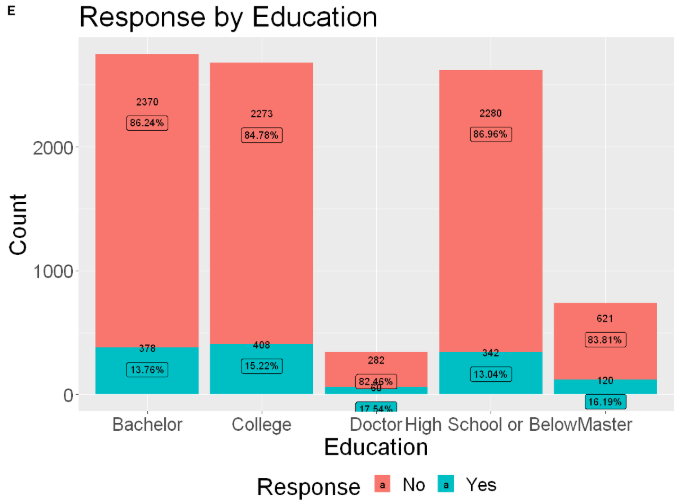
When comparing the response rate among the different genders, it can be seen that there is not much difference between the genders when it comes to a response to a sales campaign with both genders having similar proportion of respondents.

## Response by Coverage



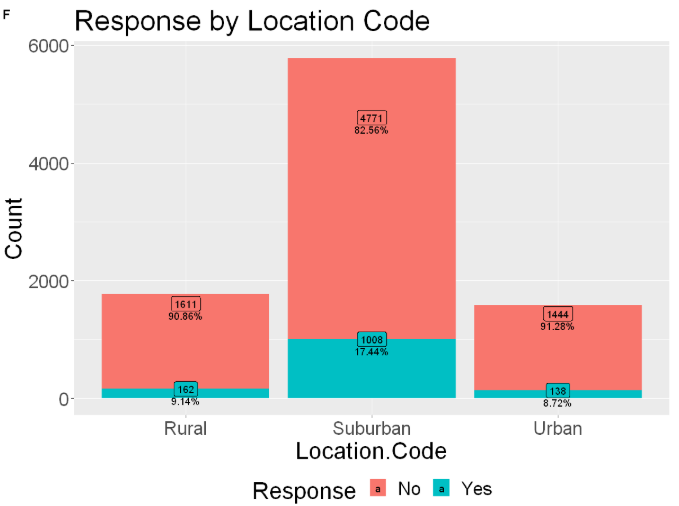
It can be seen that most customers purchased basic auto coverage with Premium coverage being the least purchased. When compared to response, Premium coverage has a slightly better proportion of respondent as compared to the other two coverages.

## Response by Education



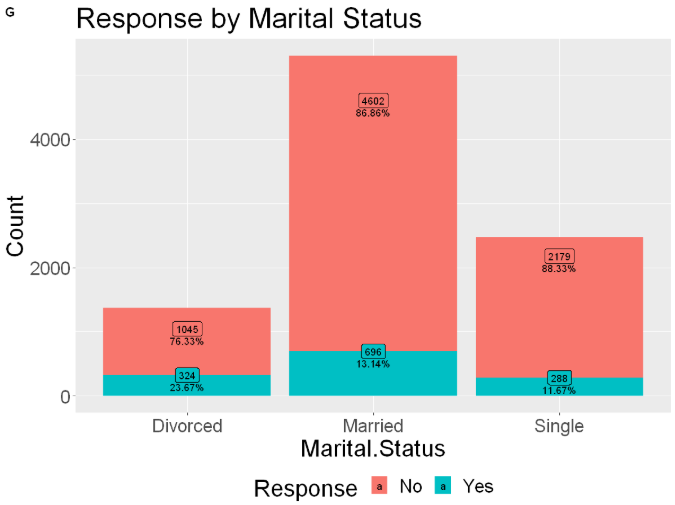
From the education with response graph, it can be seen that there is a higher number of respondents from higher education background. Customers with Doctorates and Masters degrees have the highest number of positive respondents with 17.54% and 16.19% respectively. Customers with lower education qualifications High School and below have the lowest number of respondents.

## Response by location code



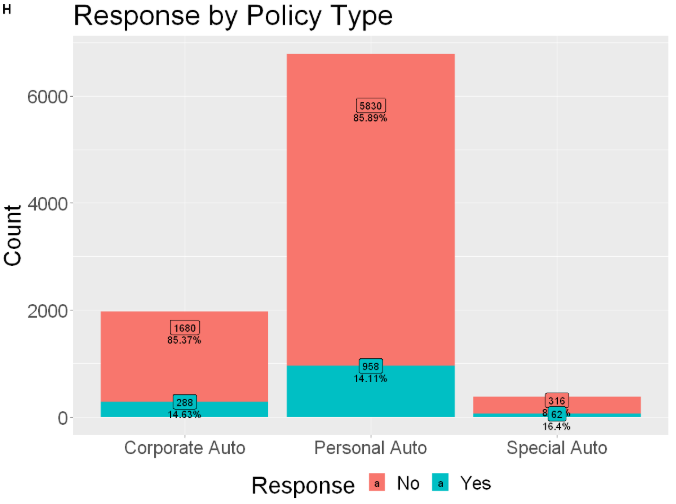
By comparing the location code, it can be seen that the largest population and proportion of respondents comes from the suburban class. With a 17.44% positive response as compared to urban and rural customers. Safe Insurance should target suburban customers as they are most likely to respond favorably.

## Response by Marital Status



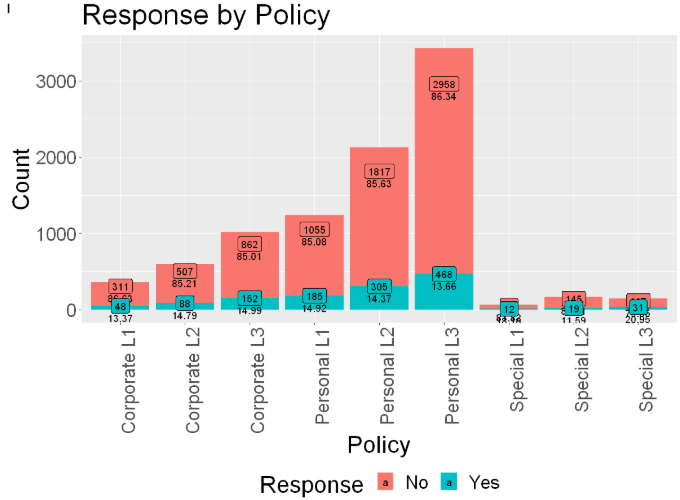
From the marital status of customers, it can be seen that most customers are married. However, the highest proportion of positive respondents comes from Divorced customers with 23.87% offering a positive renewal response.

## Response by Policy type



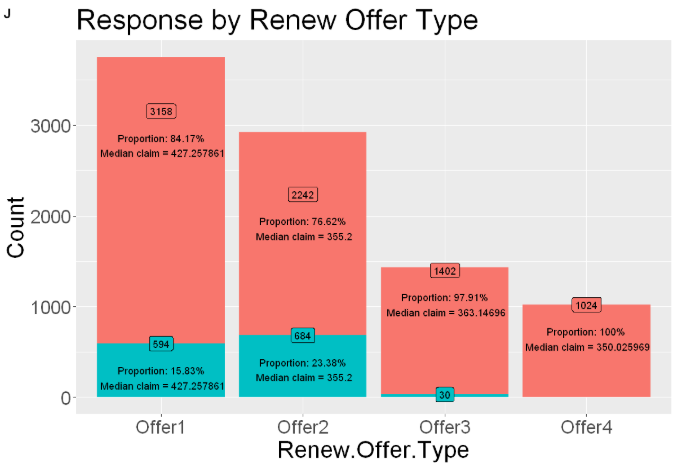
Most customers have personal auto insurance followed by corporate and then special auto policy type being in the minority. However, customers in the special auto policy type have a higher likelihood of responding favorably to a renewal campaign as compared to the other policy type holders.

## Response by Policy



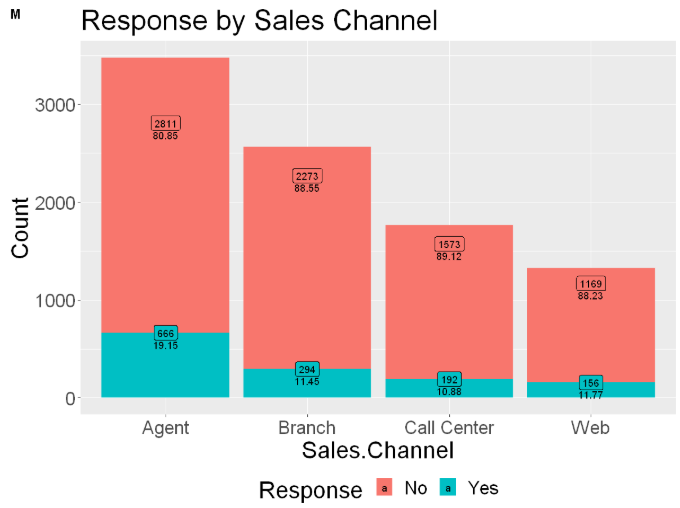
From the policy break down it is confirmed that most customers use Personal L3 policy with the three Personal policies having a higher take up rate as compared to the special or corporate policies.

## Response by Renew offer type



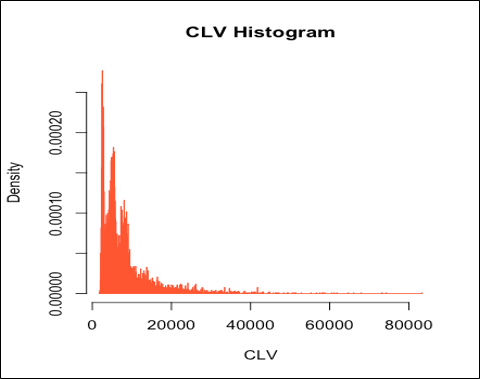
By comparing take up rates for the offer types, it can be seen that Offer 1 and Offer 2 are the most popular. With 15.83% and 23.38% of those offered responding favorably. However, their claim amount is also the highest indicating that it would cost Safe insurance more in the event that the customers do put up claims.

## Response by Sales Channel



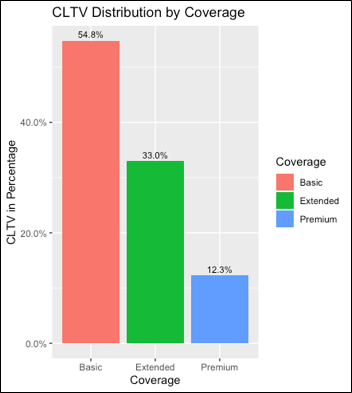
By plotting the different sales channels, it can be seen that most customers obtained their policies through sales agents and it represents the best sales model. Second being the branch and then call center and web sales. For the retention of customers, it can be also seen that having an insurance agent it improves customer retention as 19.15% of the customers renewed their offers after the contract ended.

## CLTV Distribution



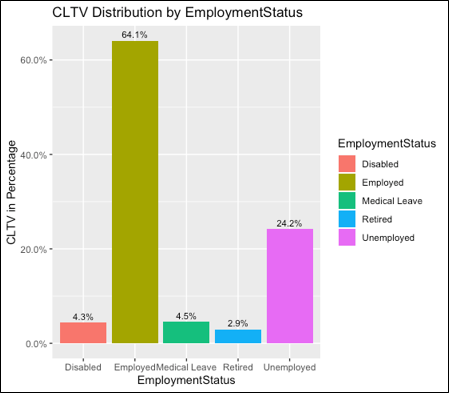
By plotting the histogram of CLTV distribution, it can be seen that CLTV is highly skewed to the left, with skewness of 3.03 and kurtosis of 16.82. As CLTV is not normally distributed, this means that it requires further processing in the modelling process where log transformation is required.

## CLTV Distribution by Coverage



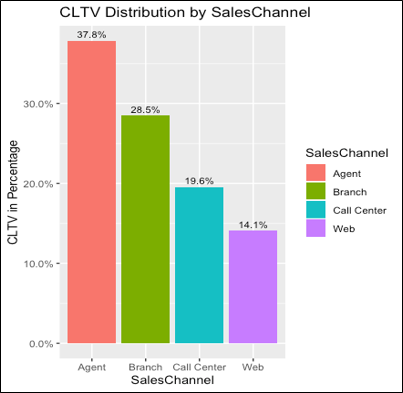
From the boxplots of the CLTV distribution by coverage types, it is found that basic coverage makes up the majority of CLTV, with 54.8% of the total CLTV value. By contrast, premium coverage has the least portion of CLTV out of the three types of coverages.

## CLTV Distribution by Employment Status



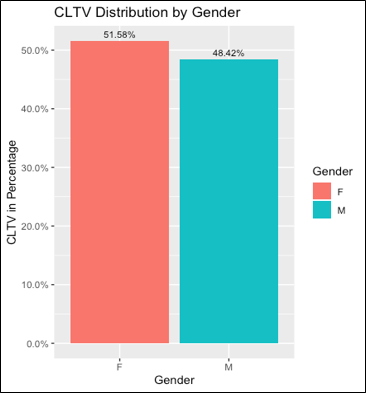
From the CLTV distribution by employment status, it can be seen that those policy holders who are employed make up the majority (64.1%) of the total CLTV value. The second largest group are unemployed policy holders. This makes sense as people who are employed are generally are more well-off and can purchase more insurance policies.

## CLTV Distribution by Sales Channel



By plotting bar charts of CLTV by various sales channels, it can be seen that agents have brought in customers with the highest amount of CLTV, followed by branch and call centre. Also, 14.1% of CLTV was brought in from the website. This will help us on the later portion for Business Problem 1, where the effectiveness of the different sales channel is being evaluated.

## CLTV Distribution by Gender



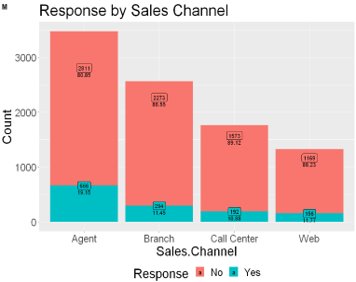
In terms of CLTV distribution by gender, it is quite similar between both gender, and the difference between males and females is very marginal. This behavior coincides as there is higher proportion of female policy holders compared to male ones (3% more).

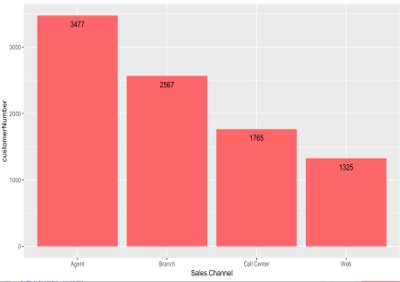
# Business Problem 1 – Evaluating the effectiveness of Sales Channel

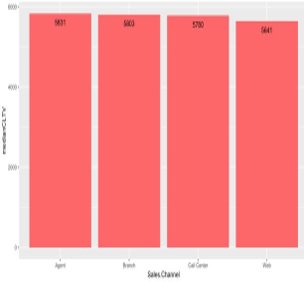
The first business problem is to evaluate the effectiveness of the four sales channels. This can be deduced from our data exploration. Through plotting bar charts of CLTV value and response rate with respect to the respective sales channels, we are able to evaluate the effectiveness accordingly.

## Data Exploration Result

The results are as below:







In terms of response to renewal, agents have achieved the highest renewal rate, and those policy holders from website have the lowest renewal rate. The renewal rate from call center is also noticeably lower than agents.

Agents have brought in the largest number of customers, and the website has brought in the lowest number. In terms of CLTV per customer, those brought in by agents are also higher than the other sales channels. Hence, it can be concluded that agent is the most effective sales channel while web is the least effective.

## Action Plan for Business Problem 1

Hence, the Data Analytics team advise Safe Insurance to consider recruiting more agents to increase sales revenue. On the other hand, since web is the least effective, Safe Insurance should consider to optimise their website design to attract more customers. It is also noticed that call centre did not perform as well as compared to agents for customer renewal. As a result, Safe Insurance should re-access the skill sets of the staff in the call centre, and adopt training to enhance their skills.

# Business Problem 2 - Determining and building relationship with valuable customers

The second business problem is to find out who are Safe Insurance’s valuable customers and to implement strategies to build a relationship with the different customer groups. The reason behind this is to reduce the likelihood that a customer would churn.

## Clustering

In order to define the valuable customers for Safe Insurance, the first step is cluster the existing customer base in order to determine the various customer groups under Safe Insurance’s auto insurance umbrella and to determine which group of customers is the most valuable to safe insurance. The metrics to determine value of the customer is below.

### Net CLTV

Customer Lifetime Value estimates the profit attributed to a customer and hence, it determines the value to the company. Insurance claims determines the cost of the customer to the company. Hence, in order to measure net value of a customer group, a derived variable called Net CLTV is produced. The equation for the variable is below.

*Net CLTV=CLTV-Claims*

Net CLTV provides a better determinant of value as compared to CLTV as it takes into account the claims which the insurance company has to pay to the customers.

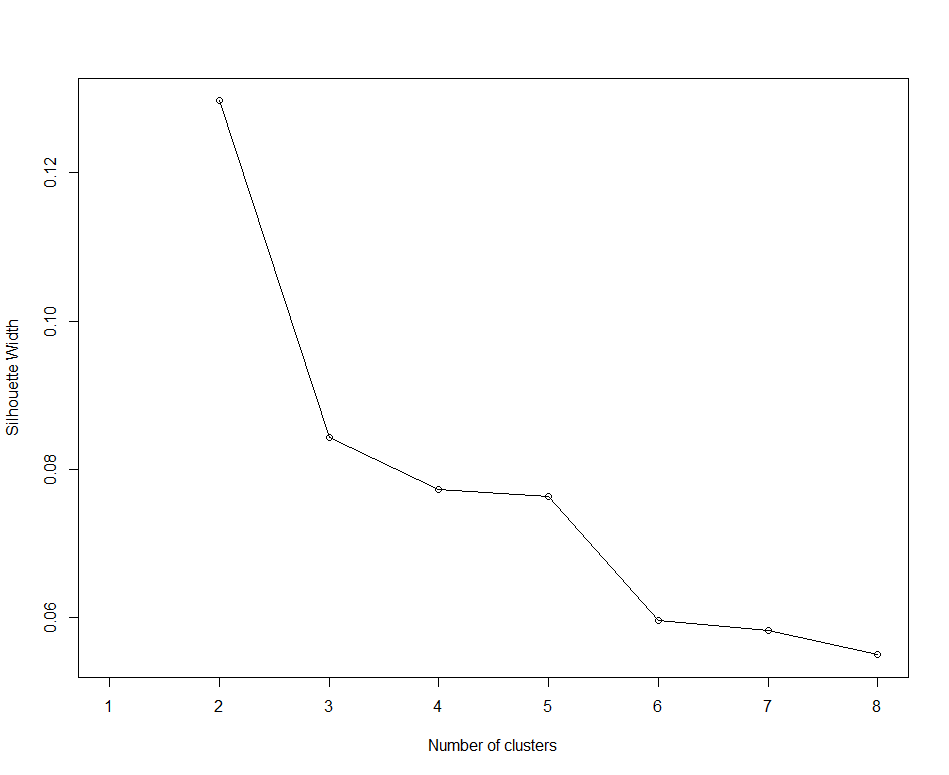
### Calculate Gowers distance

As the dataset contains a mix between continuous and categorical variables, it was decided to utilize Gowers distance to calculate the differences between the different customers. Gower Distance is a distance measure that can be used to calculate distance between two entity whose attribute has a mixed of categorical and numerical values. By calculating the Gowers distance, we can see which customers are the most similar and the most dissimilar. The customers who are most similar according to Gowers distance are grouped together into the same cluster.

### Silhouette Width

The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters.

With the Gowers distance, the Silhouette Width is calculated to determine the optimum number of clusters in the dataset to ensure the greatest distance between the different clusters.



From the above chart, it can be seen that 2 clusters produce the largest Silhouette Width this would indicate the largest difference between the different cluster. However, 3 clusters are picked instead as it offers more room for different strategies to treat each customer groups.

### K=3 Clusters scatter plot

Plotting out the scatter plot of the three clusters would showcase how similar/dissimilar are the two clusters.



From the scatter plot, cluster 2 is more distinct as compared to cluster 1 and 2. From the scatter plot, cluster 1 and 3 has some overlaps.

## Description of each cluster

After plotting the demographics for each of the clusters, the details can be

|  |  |  |  |
| --- | --- | --- | --- |
|  | Cluster 1 | Cluster 2 | Cluster 3 |
| Description | •Well educated  • Employed  •Mostly Female  • Married  •High median income  • Mostly from California  • Mostly lives in suburban areas  •Most have Policy personal L3  •Most purchased from Branches or agents | * Not highly educated * Mostly Unemployed * Mostly Male * Single * Lowest median income * Mostly from California * Mostly lives in suburban areas * Most have Policy personal L3 * Most purchased from Branches or agents * Highest median claim amount | * Well Educated * Employed * Equal Male/Female * Married * Highest median income * Mostly from Oregon * Living in Rural/Urban areas * Most have Policy personal L2 * Most purchased from Branches or agents * Lowest median claim amount. |

As shown earlier, the distance between cluster 1 and cluster 3 are close mainly differentiated by their state of origin as well as the type of policy which they hold. To improve the cluster distance, the variables which are similar in all the cluster or between cluster could be removed to make the cluster more unique. Some of the variables which are rather similar is the vehicle type, vehicle size and number of policies held. This separation increasing separation of clusters can be obtained in subsequent explorations.

### Analysis of cluster value

When plotting out the median net CLTV for the 3 cluster, the below table is obtained.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Cluster | Median CLTV | Median Total Claims | Median Net CLTV | Customer value |
| 1 | 5703 | 336 | 5362 |  |
| 2 | 5842 | 532 | 5217 | Least Valuable |
| 3 | 5847 | 350 | 5629 | Most Valuable |

From the Net CLTV (CLTV – claims) for each cluster, the most valuable cluster to Safe insurance is cluster 3 followed by cluster 1 and lastly cluster 3.

From business understanding of the auto insurance business, it can also be seen that cluster 1 and 3 are valuable customers. Firstly, customers in cluster 1 and 3 are employed and have high income. This would mean that these customers are able to afford the premiums for the auto insurance. Added to that the insurance claims for cluster 1 and 3 are lower as compared to cluster 2 this would indicate that cluster 1 and 3 are less risky insurance customers as compared to cluster 2

Conversely for cluster 2, most of the customers in that cluster are unemployed which would mean that they may have problems paying the insurance premiums. Added to that, cluster 2 has the highest median claim amount as compared to cluster 1 and cluster 3. This would indicate that this would cost Safe insurance more as more payouts to the customer are required.

## Classification of new customers

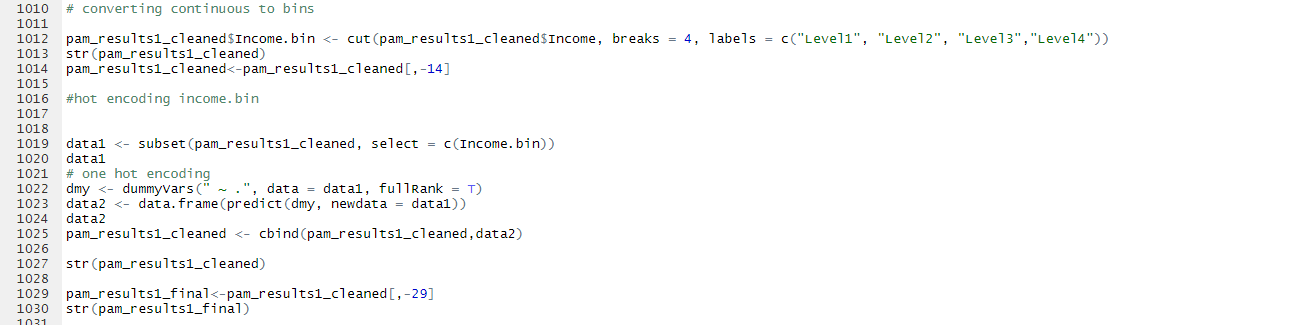
For new customers who have just purchased Safe insurance products, there will be an issue of how to classify them into the various clusters so that Safe insurance would know what strategy to extract value from them. In order to do this, we have created a decision tree classifier to classify the new joiners to Safe insurances customer pool. The variables used to train the classifier are below.

|  |  |  |
| --- | --- | --- |
| No. | Variable | Variable type |
| 1 | State | independent |
| 2 | Marriage status | independent |
| 3 | Gender | independent |
| 4 | Income | independent |
| 5 | Vehicle size | independent |
| 6 | Vehicle type | independent |
| 7 | Location code | independent |
| 9 | Employment | independent |
| 10 | Sales Channel | independent |
| 11 | Cluster | Dependent (to be predicted) |

The variables are limited to those which can only be provided by a new customer. Hence, variables such as join date and claim amount which only can be provided by older customers, are not considered in the model.

For the model, a decision tree classifier was being used.

1. **Data Preprocessing**
   * + Check and confirmed that there is no duplicates or missing data
     + Removed relationship data fields (I.e. claim amount, premium amount) which cannot be obtained from new joiners.
     + Split the dataset to 70% training and 30% testing dataset.
2. **Model buildings**
   * + Income variable binned into 4 bins as decision tree requires categorical variables.



* + - One hot encoding done to all categorical variables.

1. **Model Evaluation**
   * + From the decision tree model below, the new customers can be split into the three clusters.

A close up of a logo

Description generated with high confidence

The confusion matrix is as shown below:

|  |  |  |  |
| --- | --- | --- | --- |
| **Cluster/Predicted**  **Actual** | **1** | **2** | **3** |
| 1 | 863 | 56 | 238 |
| 2 | 80 | 615 | 31 |
| 3 | 207 | 30 | 620 |

Accuracy of model

Accuracy =

1. **Implementation**
   * + At the point of sales, the customer data would be fed into the decision tree model. This would help safe insurance to determine the customers value for the organization based on the cluster the customer is classified into.
     + This would allow Safe insurance to tune their strategies to each customer.

## Action Plan for Business Problem 2

To reduce churn rate and build a relationship with the customer, Safe Insurance’s status as a general insurance company should be leveraged upon to cross sell its other insurance products to the valuable customers. The offers for cross selling are in the table below.

|  |  |  |
| --- | --- | --- |
| **Cluster 1 programme** | **Cluster 2 programme** | **Cluster 3 programme** |
| * Offer Family Travel insurance at a special rate * Offer House insurance at special rate | * No special program as not considered a focus customer segment. | * Offer complementary insurance for technology gadgets. With added cover * Offer Travel insurance at a special rate * Offer House insurance at special rate |

For Cluster 1

* Mostly Females and married. Hence, Family travel insurance would be attractive for this cluster
* Higher income, hence willing to spend on Home Insurance.

For Cluster 3

* Cluster 3 has the highest median income among the 3 clusters. Hence, offering complementary insurance for technology gadgets and travel insurance will appeal more to this cluster.

By cross-selling other insurance products to these clusters at special rates, it will attract more customers to take up more insurance products. This is essential in building a long-term relationship with the customer by building Brand Loyalty among the customers for Safe Insurance. This also makes it more difficult for the customer to churn as they would lose their special discount on the other insurance policies. This could help in the reduction of customer churn rate.

# Business Problem 3 - Dealing with Poor Renewal Rate - Churn Prediction

Data exploration shows that 84% of the renewal offered for policies expiring in Jan 2011 received “No” response effectively leading to non-renewal of USD31 million Net.CLTV. This is a major problem for Safe Insurance PLC as this indicates very low renewal retention. A model to identify who are likely going to churn will allow in house sales channels including agent to focus their effort and pay attention to those policyholders to get them to renew with Safe Insurance PLC.

Given the limited amount of data, we will use Jan 2011 data for training and validation while keeping Feb 2011 data as Out of Time test data. While churn production built from limited data may not work well when put into actual use, the purpose of this model building exercise mainly to show management of Safe Insurance PLC the benefit of having a churn prediction model to aid its effort to improve renewal rate and secure management buy in to fund and commission better data collection to enable better model building.

The target variable in this scenario is the “Yes” or “No” response to renewal offer. We will use classification model to predict whether a policyholder will respond positively or negatively to the renewal offer. Modelling techniques that we have used include Logistic Regression, Decision Tree, Support Vector Machine and Random Forest.

For this project, the current renewal ratio of 16% will be served as the baseline to measure improvement gained after implementing churn prediction and customer retention measures.

Our churn prediction model should ideally be able to identify potential churner (non-renewing policyholder) as accurately as possible. In other words, model with higher the True Negative Rate is better.

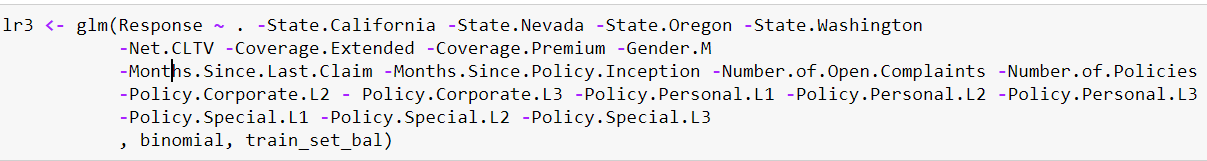
1. **Data Preprocessing**

* Check and confirmed no duplicates or missing data
* Convert Effective.To.Date to date format and use it to group data by its Effective.To.Date / its expiry date
* Jan 2011 data is used for training (80%) and validation (20%). Feb 2011 data is used for OOT testing
* One hot encoding done on categorical independent variables in data set used in Logistic Regression and SVM only. Tree and Random Forest model do not require one hot encoding.
* Random oversampling is used to balance the dependent variables in training set

1. **Model buildings**

* Logistic Regression on balanced training set, one hot encoded

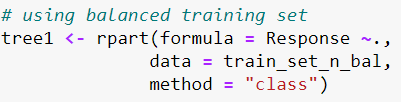
After removing statistically not significant data



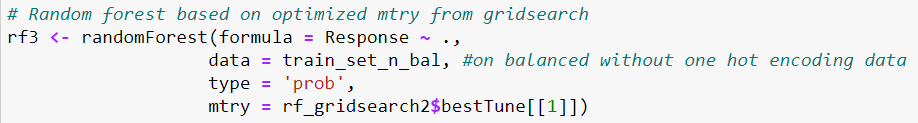
* Support Vector Machine on balanced training set, scaled numeric variables, one hot encoded categorical variables



* Decision Tree on balanced training set without one hot encoding

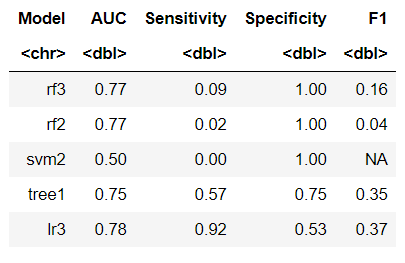


* Random Forest on balanced training set

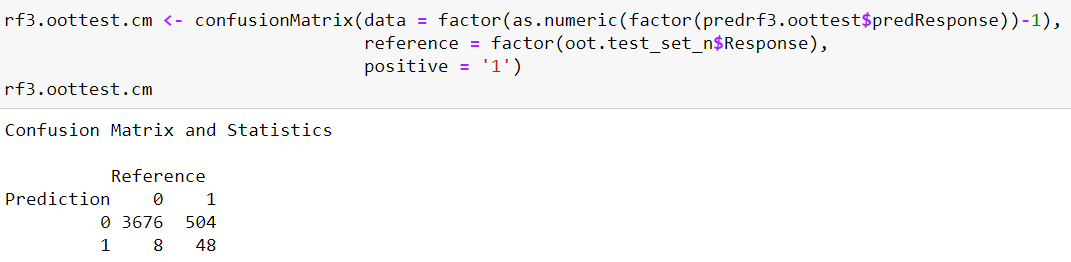


1. **Model Evaluation**

All models are tested against the Out of Time test set (Feb 2011 data) and the metrics are as follows:

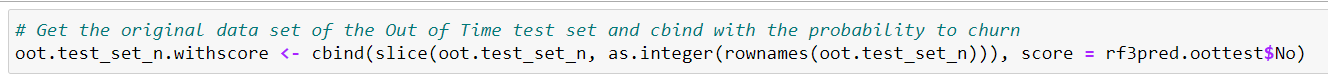


Based on the small dataset available, rf3 (Random Forest) model has highest specificity of 1.00 and able to identify churn rate best.

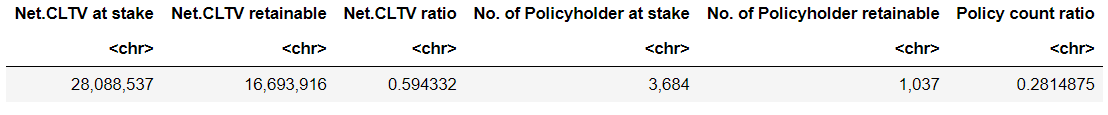


1. **Implementation**

* Using Jan 2011 as Observation Period, use the data to build model and score Feb 2011 renewal on 31 Jan 2011.
* Since Auto Insurance is compulsory insurance, the performance period need not be long as the definitive status of the renewal should be known within 30 days from Effective.To.Date
* We combine the predicted score with the Feb 2011 data



* For effectiveness of renewal targeting, we use the following arbitrary thresholds to demonstrate effectiveness of a churn model:
  + Churn score > 0.6
  + Net.CLTV > = 75th percentile Net.CLTV of Feb 2011 policyholder data (USD 8,494)
* Result:



By **targeting 28%** of potentially not renewing policyholders we can **reach out to almost 60% of the Net.CLTV** that are predicted to be not renewing and entice them to renew with Safe Insurance PLC. This allows limited resources to be channeled to higher value policyholders. The thresholds can be altered to meet operational constraints and maximize efficiency.

* Each distribution channel can further filter the list of policyholders and run renewal retention campaign for its respective channel to improve renewal retention.

# Business Recommendation

To effectively retain policyholder, we would recommend:

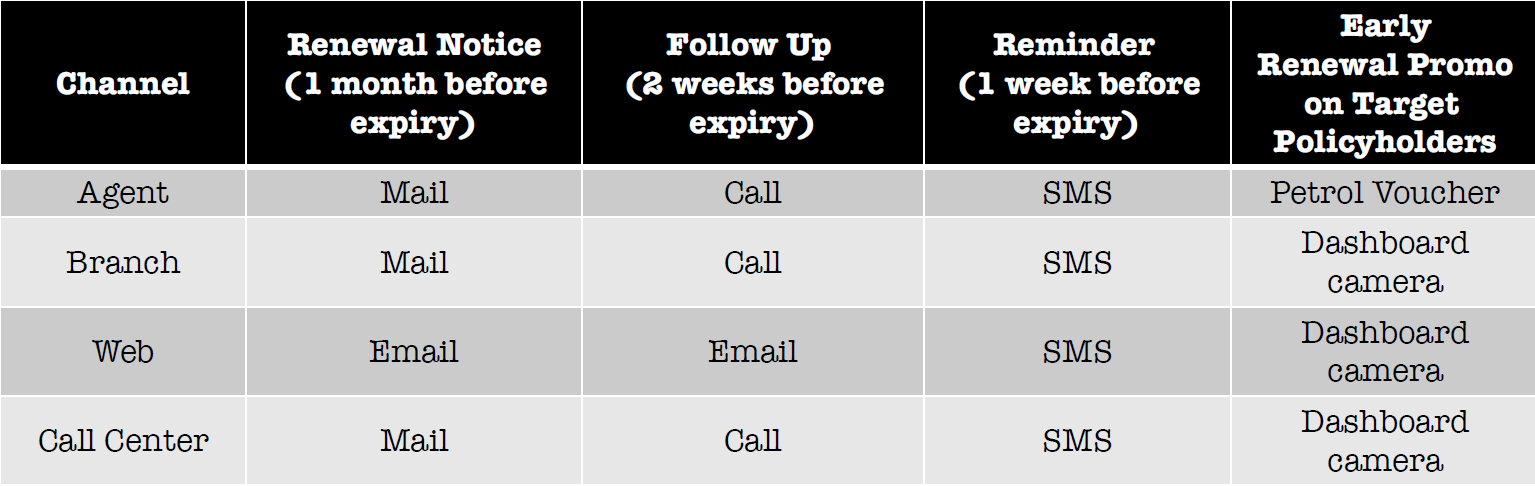
* 1. Early engagement and cross selling

It is important to engage the policyholder early and deepen their relationship with Safe Insurance. One way to achieve this is by cross selling other insurance products at preferred terms or rate for existing auto insurance policyholder. This way the policyholder may not so easily switch auto insurer given the preferential terms they enjoy.

* 1. Renewal Retention Campaign

Utilize churn score and Net.CLTV to scope valuable policyholders to be targeted at renewal retention campaign.

Renewal Retention Campaign Idea



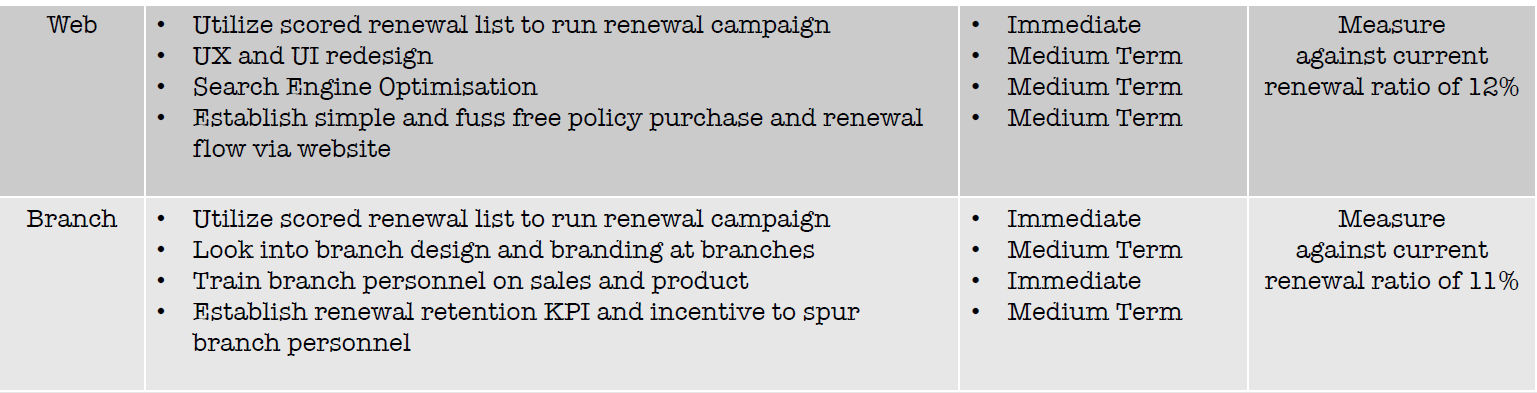
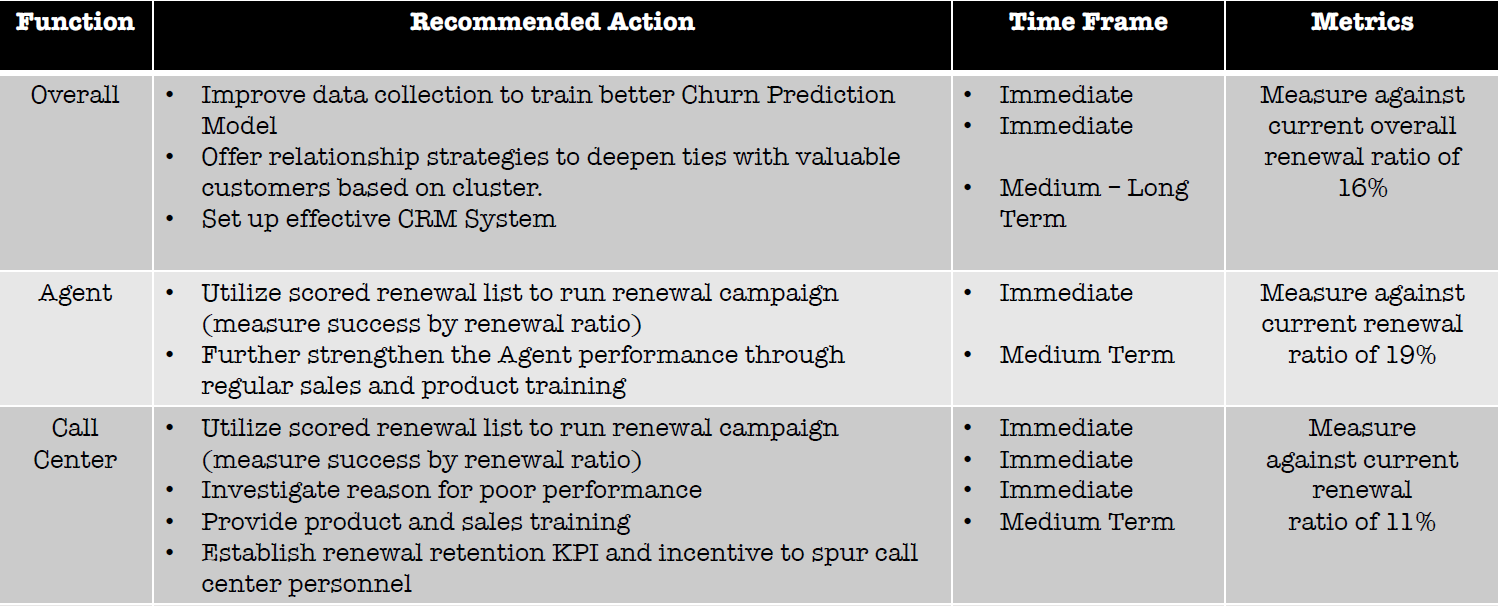
For Agents, Branch and Call Center customers, a renewal notice email would be sent 1 month before the expiry of their policy this mail would also contain a sign-up form outlining early renewal gift and customer service contact. For customers who purchased insurance through website, we assume they prefer to use their digital means of communications hence instead of a mail, an email should be sent instead. An early bird gift would incentivise customers to commit to the policy renewal earlier hence securing them as customers. Concluding the sales deal earlier would also free up resources for other customers.

2 weeks before the expiry, a sales representative will call the customer to ask if they need any explanation on the coverage and renewal terms as well as politely encourage them to renew with Safe Insurance.

Lastly, if policy is still not renewed 1 week before policy expiry, Safe Insurance will send an SMS to the customers to remind them not to forget to renew their auto insurance as it is an offence to drive without insurance.

By implementing multiple layers of renewal reminder, Safe Insurance is engaging policyholder at crucial renewal period allowing policyholder to seek clarification if necessary and hopefully reduce policy lapse (churn).

The following table summarizes our recommendation



# Conclusion

1. Given the limited data, we have showcased the benefit of analytics project to improve Safe Insurance’s auto insurance portfolio particularly in improving the renewal ratio.
2. To achieve a good renewal retention, we believe valuable policyholders have to be engaged early and encouraged to get other products from Safe Insurance at preferential terms to deepen their relationship with Safe Insurance.
3. A deeper relationship will also aid the renewal retention campaign nearer the expiry date of the auto insurance policy.
4. The cluster analysis and prototype models we have built using limited data can be further improved with larger dataset
5. We find the absence of product and renewal offer information limits the depth of our analysis thus this information will be necessary in next phase of this analytics project.
6. Lastly if Safe Insurance is going to engage in effective campaign, a CRM system will benefit them in the long run allowing a comprehensive customer interaction management and running sophisticated marketing campaign.

## Limitations

This mini-prototype has showcased success in addressing and solving the listed business problems with the limited dataset. We would like to seek management’s approval to expand into a full-scale project. Before working on the full-scale project, there are some limitations in this prototype due to the limited data provided, and we believe by providing more data, we would be able to overcome these limitations.

The limitations are as below:

1. The dataset only spans across two months (January and February of 2011), which could affect the accuracy of the models we built. The business nature of insurance is that insurance policies are usually annual contracts, and there could be underlying seasonality in the dataset. Hence, we believe that it is adequate to provide at least 2-3 years of data to build more accurate models.
2. Details pertaining to product information and renewal offers are not given in the dataset. As a result, it is difficult to perform further analysis on this information and study the underlying reasons why customers chose not to renew their policies. We believe that by providing these details, we will be able to produce better and more tailored strategies for customer renewal.
3. No information is provided on the cost of marketing campaigns and sales channels, which makes it difficult to conduct cost and ROI analysis for the different plans and sales channels. We believe that this information is important for proper campaign analysis.

In conclusion, we advise Safe Insurance to provide the abovementioned information to overcome these limitations, so that more accurate models and better sales and marketing strategies can be provided.

# References

**Source:** <https://www.kaggle.com/pankajjsh06/ibm-watson-marketing-customer-value-data>

# Appendix A

**R Code Files**

1. Data Exploration 1 (Response)
2. Data Exploration 2 (CLTV Distribution)
3. Business Problem 2
4. Business Problem 3

# Appendix B

**Variables of Source Data**

|  |  |  |
| --- | --- | --- |
| Variables | Description | Data type |
| Customer | Customer ID (vehicle registration number) | Character |
| State | State that the customer’s vehicle was registered in | Categorical |
| Customer Lifetime Value | Customer Lifetime Value of the customer | Numeric |
| Response | Response to previous campaign | Categorical |
| Coverage | Insurance type | Categorical |
| Education | Highest education level | Categorical |
| Effective To Date | Date that the insurance is effective to | Date |
| Employment Status | Employment status | Categorical |
| Gender | Gender | Categorical |
| Income | Income | Numeric |
| Location Code | Residential location | Categorical |
| Marital Status | Marital Status | Categorical |
| Monthly Premium Auto | Month premium for the insurance | Numeric |
| Months Since Last Claim | Months Since Last Claim | Numeric |
| Months Since Policy Inception | Months Since Policy Inception | Numeric |
| Number of Open Complaints | Number of Open Complaints | Numeric |
| Number of Policies | Number of Policies | Numeric |
| Policy Type | Policy Type | Categorical |
| Policy | Policy sub-type | Categorical |
| Renew Offer Type | Renew Offer Type | Categorical |
| Sales Channel | Sales Channel | Categorical |
| Total Claim Amount | Total Claim Amount | Numeric |
| Vehicle Class | Vehicle Class | Categorical |
| Vehicle Size | Vehicle Size | Categorical |