2022

BGL Insurance

Contents

1. Introduction

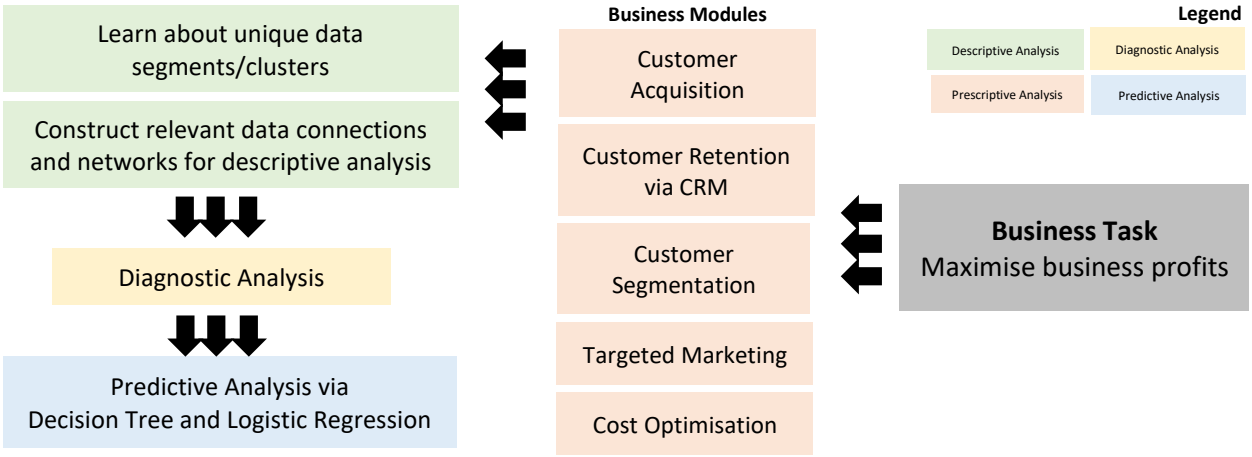
Facing the growing consumer market and increasing risk in the insurance industry, companies must understand customer behaviours and needs to effectively customize its products and services. With digitalization, increasingly tech savvy individuals, and Internet of Things (IoT) devices becoming prevalent, this gives rise to new insurance models leveraging on data analytics to truly add value to customers (Thompson, 2016). Hence, the competitive landscape in the insurance industry will be increasingly more focused on data and ensuring a closed loop data ecosystem to increase their market share, lock in customer loyalty and higher customer lifetime value thereby translating to higher profits.

In this project, we utilize the dataset from BGL Insurance company (BGL), a leading digital distributor of insurance products ranging from motor, home, and life insurance to approximately three million customers. BGL also owns comparethemarket.com, which is one of UK's leading price comparison websites providing customers with a convenient way to review a wide range of insurance products.

We've identified BGL's business use case as maximising its business profits. Further elaboration is outlined in the next section below. A combination of clustering and propensity models including decision tree and logistic regression was utilized in our data modelling process to derive data driven insights. An elaborate interpretation of our results for actionable intelligence as well as the importance of a closed loop data ecosystem and data governance will also be discussed in the later sections of the report.

2. Business Use Case/Problem Scope

Before any insightful data driven insights can be derived, it is important to first define the business use case/problem scope. Our identified business use case is to utilize data-driven insights to help BGL ***optimize business profits*** amidst the competitive insurance landscape. A factor analysis framework as outlined below provides an overview on the key business modules that we will be focusing on:



The above flow chart clearly defines BGL's business task, and shows how the various descriptive, prescriptive, diagnostic, and predictive analysis cascade accordingly.

It is important to construct relevant data connections during exploratory data analysis. Additionally, it is crucial to segment customers based on their unique and discriminative profile demographics obtained via BGL's internal data. This allows for a better understanding behind the decision-making process of each unique segments, thereby allowing BGL to optimise on limited resources for effective targeted marketing efforts.

With a proper understanding of its customers, BGL can also subsequently perform appropriate diagnostic and predictive analysis. Further elaboration on the predictive models that were built to predict if BGL's customers will renew their policies and the associated factors impacting retention rates are discussed within section 6.

3. Dataset Overview

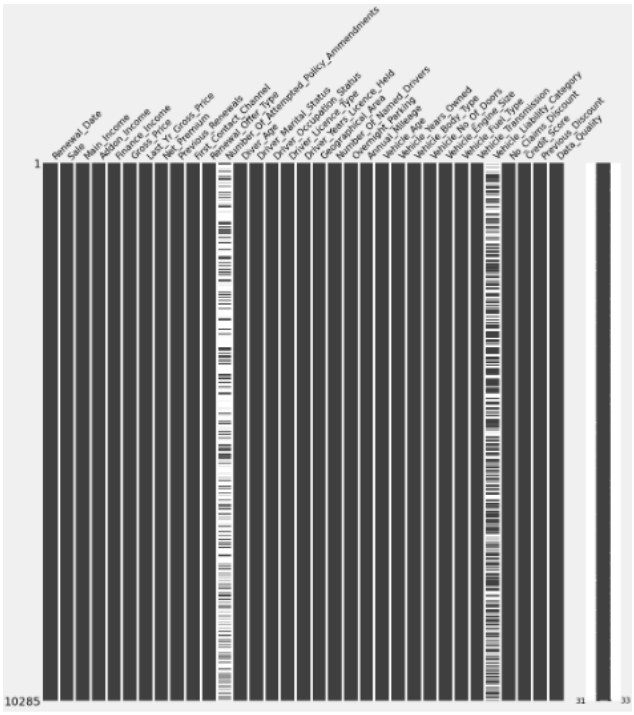
The current BGL dataset on hand focuses on motor insurance. Figure 1 in Appendix details the full data dictionary and description of BGL's current internal dataset that we have on hand. The dataset consists of a total of 10,286 records with 33 unique variables including a "Sale" variable which is identified as the target variable. Figure 2 in Appendix describes the existing data relationship and its associated connection within the data modelling stage. It is important for BGL to ensure that the dataset is collected, integrated, stored with clearly defined data schemas and cleansed to ensure accurate data analysis for relevant intelligent actionable insights.

4. Exploratory Data Analysis (EDA)

Exploratory data analysis is crucial to ensure a good understanding of the dataset as it helps with identifying the important and significant variables contributing to the prediction analysis. Business intuition coupled with data science thinking should be applied during the process of EDA.

4.1 Problematic records

We first begin by checking if there are any significant missing records that can potentially affect the model prediction results and its associated analysis. The plot below reveals a significant number of missing records within "Number of Attempted Policy Amendments" (7285 missing records) and 'Vehicle Liability Category' (4140 missing records). As such, we will proceed to exclude these variables in the subsequent analysis.

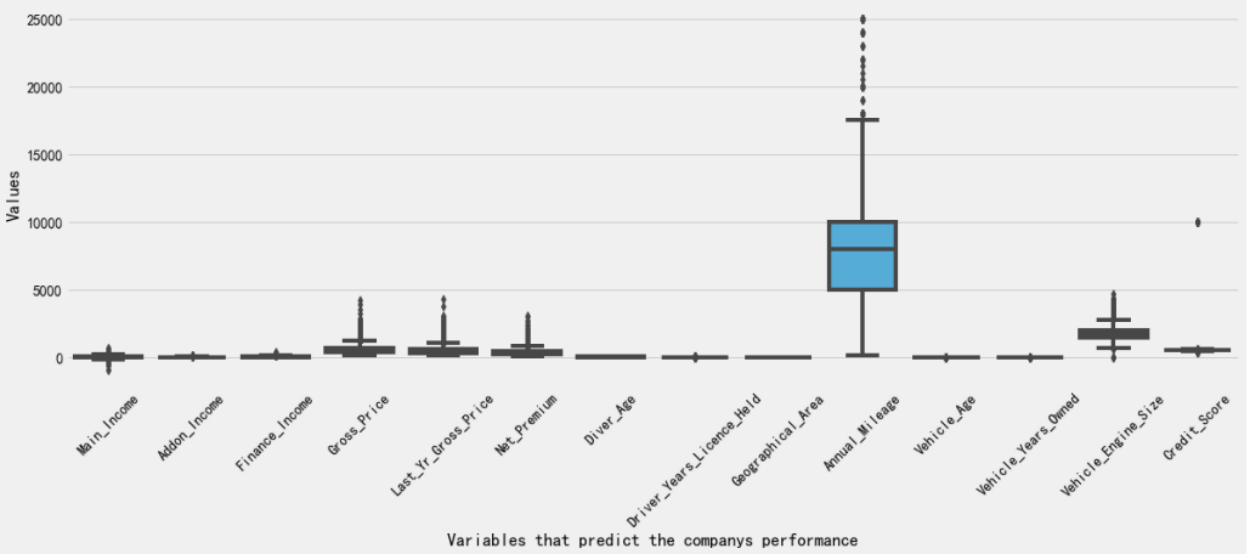


Additionally, as per the image below, within “Data Quality” column, there are 10167 “OK” records, 106 “Unacceptable” records, 9 “Quote Total Premium Error” records and 3 “Policy Total Premium Error” records. We delete the problematic records and retain the remaining “OK” records.

```
In [8]: df['Data_Quality'].value_counts()

Out[8]: OK                              10167
        Unacceptable                      106
        Quote Total Premium Error           9
        Policy Total Premium Error          3
        Name: Data_Quality, dtype: int64
```

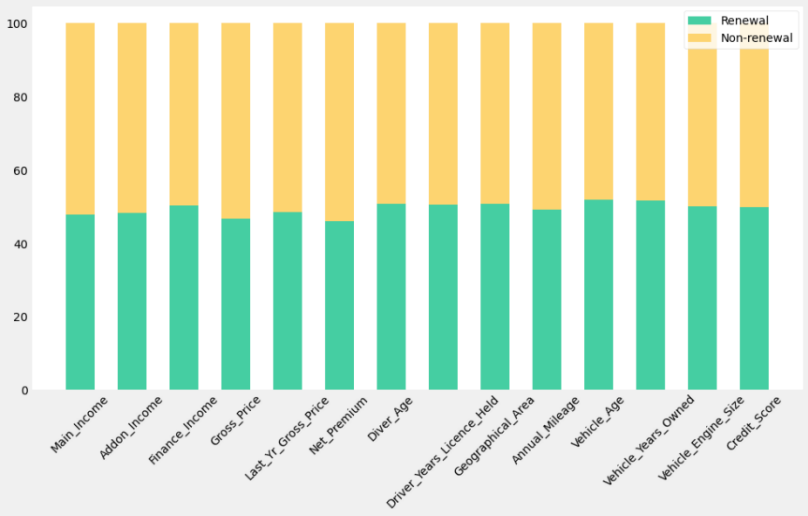
4.2 Outliers

The boxplot below reveals the presence of outlier within the variable “Annual Mileage”. These records will be retained as it can potentially indicate a unique group of customer profile. Additionally, “Annual Mileage” is an important variable as it provides information on the amount of distance that motorists clock on the roads, thereby making it a commonly used metric for insurance companies to predict a drivers’ risk based on the number of miles driven each year, alongside other variables such as age and experience (Hardesty, 2021).



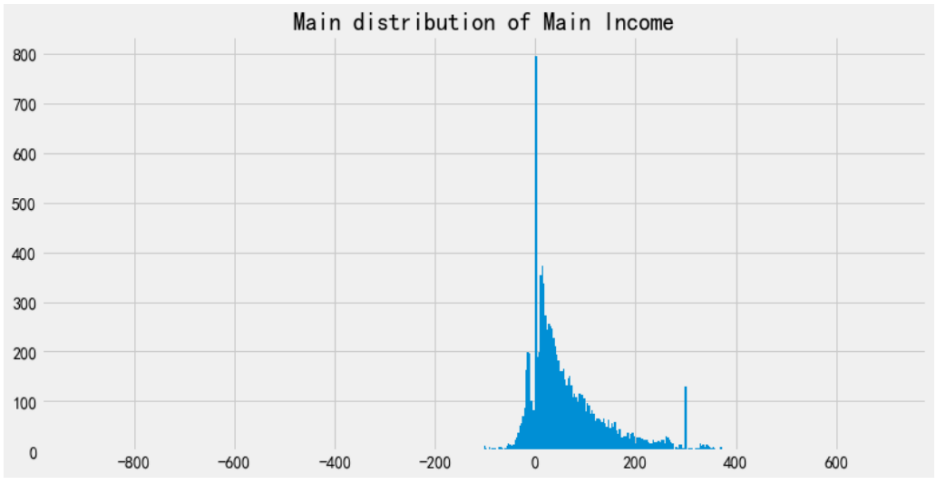
4.3 Descriptive Analysis

A stacked bar plot was also plotted to provide for a better understanding on the proportion of the customers who will renew across the various variables. “Renewal” refers to customers who renew their policy while “Non-Renewal” refers to customers who do not renew. We note that none of the variables seem to display any specific trend with regards to whether a customer will renew his or her policy with BGL.



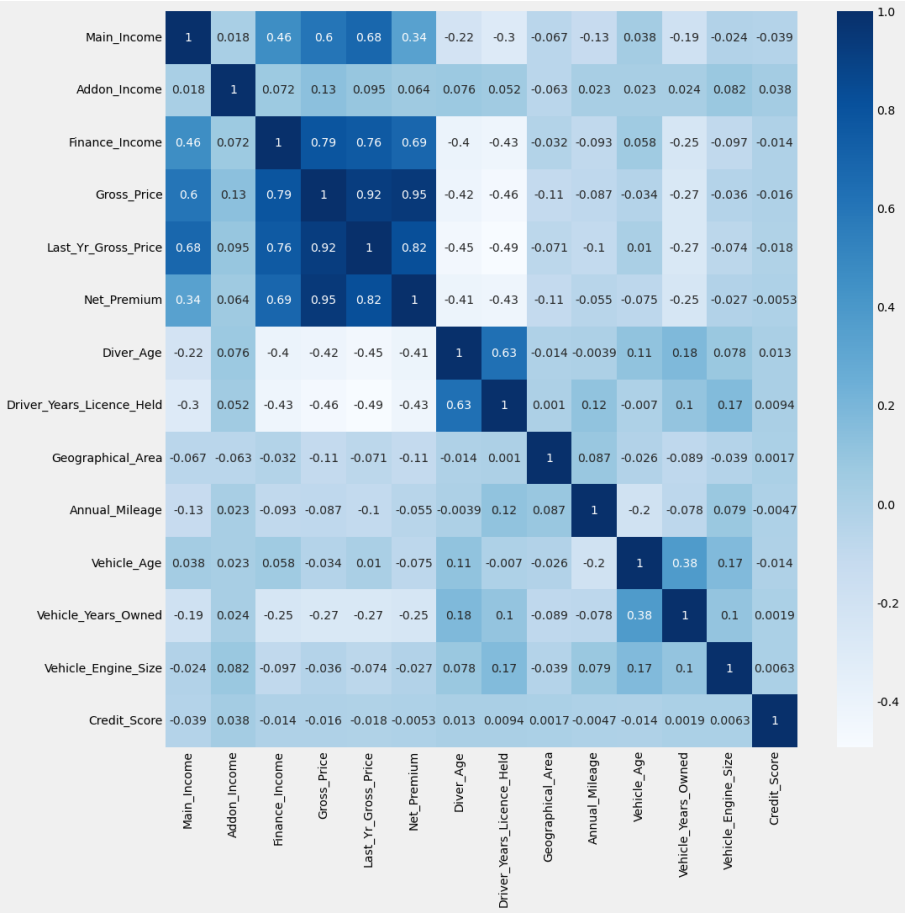
4.3.1 Univariate Analysis

As BGL’s business use case is to maximise profits, we took a closer look into the variable “Main Income”. From the figure below, majority of the “Main Income” amount is distributed at approximately $100, thereby indicating that BGL’s financials is relatively healthy.



4.3.2 Bivariate Analysis

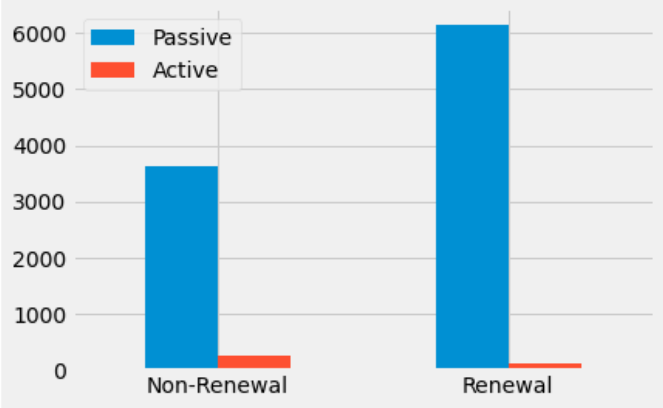
Given that numerical variables are utilized for clustering and logistic regression in propensity modelling, it is crucial to ensure that our variables are not highly correlated. The heat map of all numerical variables shows that high correlation of >0.8 exists amongst “Net Premium”, “Gross Price” and “Last Year Gross Price”. Hence, we retained “Net Premium” and dropped “Gross Price” as well as “Last Year Gross Price” variables.



4.4 Target Variable

The “Sale” variable within the dataset has been identified as our target variable, which takes on a binary value of 1 or 0, where 1 indicates a policy renewal and 0 indicates no policy renewal.

Taking a closer look at the data distribution between the two possible binary values, we can see signs of class imbalance. We can also deduce that customers tagged as “Active” within “Renewal Offer Type” variable have a lower renewal rate as compared to customers who are tagged as “Passive”. Customers who are tagged as “Active” have to actively call in to BGL to renew their policies while customers tagged as “Passive” will have their policies automatically renewed unless instructed otherwise. This seems to imply that a higher effort is required from customers who are tagged as “Active” to maintain their policies with BGL. Hence, BGL can step in and take actions accordingly to ensure that these customers do not seek out BGL’s insurance competitors but continue to renew their policies with BGL in the longer term.



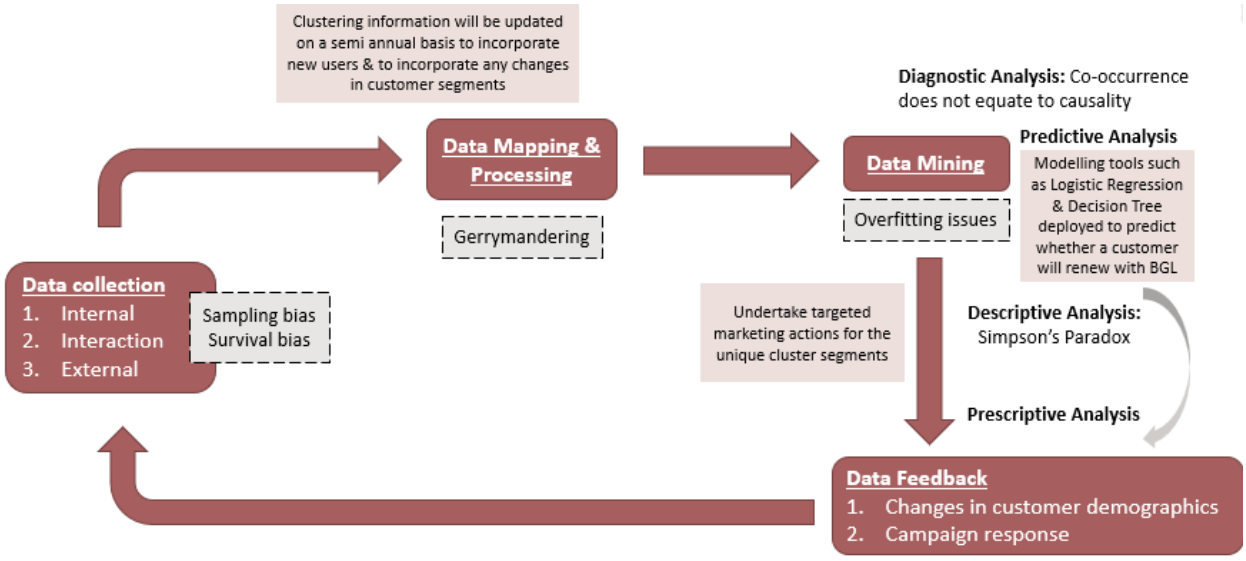
5. Data Preparation/Wrangling

5.1 Synthetic Minority Oversampling Technique

As identified in section 4.4, there are signs of imbalance present within the target variable. To avoid the situation where our subsequent machine learning models ignore the minority class, we leveraged on Synthetic Minority Oversampling Technique (SMOTE) to help with oversampling in our imbalanced classification dataset. SMOTE involves generating synthetic samples for the minority class by focusing on the feature space to create new instances via interpolation of the nearby positive instances (Brownlee, 2020).

6. Data Modelling

Before delving into data modelling, we mapped out BGL’s general existing data lifecycle per the image below.



During the data collection stage, BGL needs to be mindful of sampling and survival bias. Thereafter, the collected data will undergo data processing to arrive at our defined cluster segments. It is crucial for BGL to continually update the information on a semi-annual basis to incorporate new users and any changes in the existing customer segments. During data segmentation, BGL needs to be mindful of not gerrymandering where data segmentation is manipulated with the intent of creating desired analytical results.

Thereafter, data modelling can take place where logistic regression and decision tree models can be deployed to predict whether a customer will renew with BGL. At this stage, BGL's data scientists need to be mindful of overfitting issues. With the modelling and clustering results, targeted marketing actions can then be undertaken accordingly.

Data feedback will be concurrently collected to update any changes in dataset as well as to evaluate the effectiveness of the marketing efforts. This information will subsequently be incorporated in the next round of data collection, thereby establishing a closed loop data ecosystem.

6.1 Customer Segmentation via Clustering

Effective customer segmentation is crucial to ensure optimal usage of valuable resources while tailoring customized marketing efforts and products for BGL's various unique customer segments.

Discriminative pattern mining is a useful pattern mining technique to uncover a set of unique and significant patterns (He et al., 2017). This can help BGL to cater its service customization or marketing campaigns accordingly to different cluster segments.

An unsupervised clustering method, K-means was utilized to help sed light on the unique and discriminative characteristics of BGL's various cluster segments, some of which include:

- Better understanding of discriminative customer profiles across various clusters
- Better understanding on key clusters that offers the highest monetary value to BGL

In determining the optimal number of clusters, the elbow method was utilized. We arrived at an optimal cluster of 7, which has the steepest drop in the Sum of Squared Error (SSE) value.

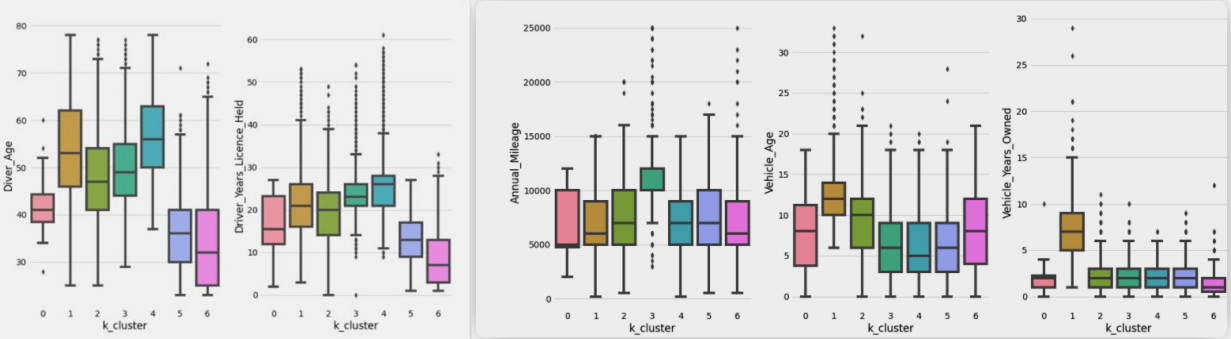
The customers are allocated to 7 unique segments based on the below features:

- Main_Income
- Addon_Income
- Finance_Income
- Net_Premium
- Driver_Age
- Driver_Years_License_Held
- Annual_Mileage
- Vehicle_Age
- Vehicle_Years_Owned
- Vehicle_Engine_Size
- Credit_Score

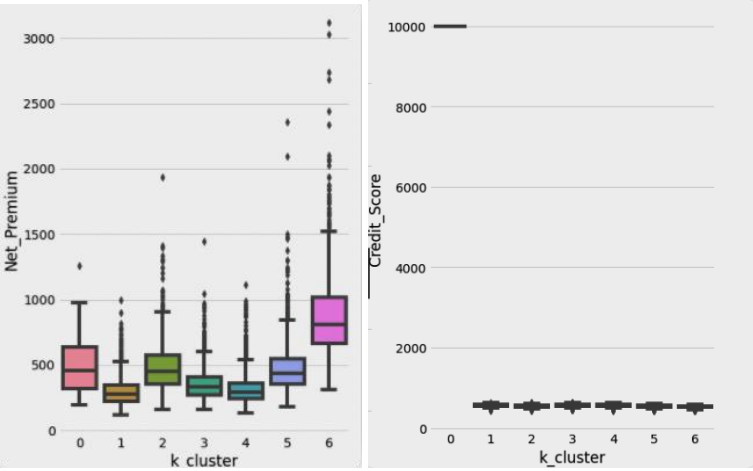
6.1.1 Clustering Result Interpretation

Boxplots were utilized to gain a better visual understanding of the dispersion of features across the various customer segments. Cluster 1, 2, 3, 4 have high average values for both "Driver Age" and "Driver Years License Held" variables, indicating an older age profile and more experienced drivers within that segment as compared to Cluster 0 and 5. Cluster 6 on the other hand indicates younger and more inexperienced drivers as compared to other clusters.

From the perspective of vehicle usage, Cluster 3 has the highest "Annual Mileage", whereas Cluster 1 has the highest "Vehicle Age" and "Vehicle Years Owned" which indicates the number of years that the vehicle has been owned for.



With regards to "Net Premium" variable, Cluster 6 has the highest amount followed by Cluster 0; Cluster 0 also have a significantly higher credit score of 9999 compared to other clusters as well.



The below table provides a summary on the characteristics of customers within each cluster.

Clusters	Clusters Description
Cluster 0	New car owners who are relatively young and have very high credit score
Cluster 1	Older and experienced drivers who own the car for a long period of time; customers' cars also tend to be old
Cluster 2	No discriminative characteristics
Cluster 3	Older and experienced drivers with a long annual mileage
Cluster 4	Older and experienced drivers
Cluster 5	New car owners who are relatively young
Cluster 6	New car owners who are relatively young who have owned their cars for a shorter time as compared to other clusters

To help BGL determine how profitable the various customer segments are, we calculated the average expected income per offer (IPO) for each customer segment via the below formula:

- IPO of each cluster = $\frac{\Sigma\,[Sale\ X\ (Main\_Income + Addon\_Income + Finance\_Income\ )]}{n}$

Additionally, the retention rate of each cluster was also computed via: $\frac{\sum_n count = 1}{total}$

Full details on the various clusters with their respective expected IPO and retention rates can be found in the table below. A few key interesting observations are highlighted below:

1. Customers within Cluster 6 generate the highest revenue (average IPO = $205.23) for BGL, while Cluster 4 generates the lowest revenue (average IPO = $47.62).
2. Cluster 6 has the lowest retention rate at 56.1%. Cluster 1 on the other hand, has the highest retention rate of 66.4% but has a low average revenue as compared to other clusters.

The above observations seem to suggest that there is room for BGL to explore when it comes to tailoring its marketing efforts, which will be further elaborated within section 7.

Clusters	**0**	**1**	**2**	**3**	**4**	**5**	**6**
Avg.IPO	$67.76	$50.43	$127.62	$52.64	$47.62	$70.13	**$205.23**
Retn. Rate	0.625	**0.664**	0.595	0.613	0.636	0.597	0.561

6.2 Propensity Models

For propensity modelling, we explored both decision tree and logistic regression due to their easy implementation, interpretation, and competitive performance.

Decision tree is a supervised information-based classification method, where parameters such as maximum depth, split ratio and number of features can be tuned accordingly. Logistic regression is a supervised algorithm that is useful for classification problems, with tuneable parameters ranging from regularization to help address overfitting issues. For both models, we are interested in determining whether BGL's customers renew their automotive insurance policies.

With a clearly identified objective and the associated model results, BGL can take immediate actions either to lock in customers during policy renewals or step in to close the gaps when customers are unlikely to renew their policies, thereby contributing to higher profits in the long run.

Before embarking on data modelling as detailed in section 6.2.1 and 6.2.2, correlated variables were removed and SMOTE was applied to correct for imbalanced class within the dataset as highlighted in section 5. Random_state = 42 was utilized to ensure consistency of the model output results. For both models, the dataset was split into 70% training and 30% test data set.

6.2.1 Decision Tree

The categorical variables were pre-processed to generate the appropriate features for modelling purposes. For example, "Renewal Offer Type" was converted into 0 and 1 which represents "Passive" and "Active" respectively. The full list of converted features can be found in the figure below.

```
First_Contact_Channel
{'I': 0, 'T': 1}
Renewal_Offer_Type
{'Passive': 0, 'Active': 1}
Driver_Marital_Status
{'M': 0, 'D': 1, 'S': 2, 'P': 3, 'A': 4, 'B': 5, 'W': 6}
Driver_Occupation_Status
{'E': 0, 'U': 1, 'R': 2, 'S': 3, 'H': 4, 'N': 5, 'F': 6}
Driver_Licence_Type
{'Z': 0, '5': 1, 'Y': 2, '4': 3, '6': 4, '3': 5}
Vehicle_Body_Type
{'H': 0, 'E': 1, 'S': 2, 'B': 3, 'C': 4}
Vehicle_Fuel_Type
{'P': 0, 'D': 1, 'E': 2}
Vehicle_Transmission
{'M': 0, 'A': 1}
```

Entropy was utilized as the measure for information impurity in our decision tree model as it provides more accurate results. Further parameter tuning was explored via maximum tree depth, minimum split ratio and maximum number of features by looping through different parameter settings to arrive at the best decision tree score.

6.2.1.1 Maximum Tree Depth

We first explored optimal tree depth selection for the decision tree model. While the highest score (0.6359) is obtained at tree depth of 2, we decided to go ahead with a tree depth of 13 to incorporate more features as it still has a relatively high decision tree score of 0.5729.

```
With Max-Depth 1, the decision tree score : 0.5084
With Max-Depth 2, the decision tree score : 0.6359
With Max-Depth 3, the decision tree score : 0.5542
With Max-Depth 4, the decision tree score : 0.5939
With Max-Depth 5, the decision tree score : 0.6106
With Max-Depth 6, the decision tree score : 0.5923
With Max-Depth 7, the decision tree score : 0.5801
With Max-Depth 8, the decision tree score : 0.5683
With Max-Depth 9, the decision tree score : 0.5667
With Max-Depth 10, the decision tree score : 0.5608
With Max-Depth 11, the decision tree score : 0.5651
With Max-Depth 12, the decision tree score : 0.5585
With Max-Depth 13, the decision tree score : 0.5729
With Max-Depth 14, the decision tree score : 0.5575
With Max-Depth 15, the decision tree score : 0.5615
With Max-Depth 16, the decision tree score : 0.5533
With Max-Depth 17, the decision tree score : 0.5654
With Max-Depth 18, the decision tree score : 0.5719
With Max-Depth 19, the decision tree score : 0.566
With Max-Depth 20, the decision tree score : 0.5706
With Max-Depth 21, the decision tree score : 0.5693
With Max-Depth 22, the decision tree score : 0.5556
With Max-Depth 23, the decision tree score : 0.5674
With Max-Depth 24, the decision tree score : 0.5657
With Max-Depth 25, the decision tree score : 0.5628
With Max-Depth 26, the decision tree score : 0.5634
With Max-Depth 27, the decision tree score : 0.5634
With Max-Depth 28, the decision tree score : 0.57
With Max-Depth 29, the decision tree score : 0.5677
With Max-Depth 30, the decision tree score : 0.5664
With Max-Depth 31, the decision tree score : 0.5624
With Max-Depth 32, the decision tree score : 0.5624
With Max-Depth 33, the decision tree score : 0.5624
With Max-Depth 34, the decision tree score : 0.5624
With Max-Depth 35, the decision tree score : 0.5624
With Max-Depth 36, the decision tree score : 0.5624
With Max-Depth 37, the decision tree score : 0.5624
With Max-Depth 38, the decision tree score : 0.5624
With Max-Depth 39, the decision tree score : 0.5624
With Max-Depth 40, the decision tree score : 0.5624
With Max-Depth 41, the decision tree score : 0.5624
With Max-Depth 42, the decision tree score : 0.5624
With Max-Depth 43, the decision tree score : 0.5624
With Max-Depth 44, the decision tree score : 0.5624
With Max-Depth 45, the decision tree score : 0.5624
With Max-Depth 46, the decision tree score : 0.5624
With Max-Depth 47, the decision tree score : 0.5624
With Max-Depth 48, the decision tree score : 0.5624
With Max-Depth 49, the decision tree score : 0.5624

The maximum score 0.6358570960340871 can first be obtained at tree depth 2
```

6.2.1.2 Split ratio

Next, we explored optimal split ratio selection for the decision tree model. We can see that a maximum tree depth of 13 and split ratio of 6% yields the best decision tree score of 0.6205.

```
Max-Depth 13, Split Ratio 1%, DT Score : 0.6168
Max-Depth 13, Split Ratio 2%, DT Score : 0.6064
Max-Depth 13, Split Ratio 3%, DT Score : 0.6123
Max-Depth 13, Split Ratio 4%, DT Score : 0.606
Max-Depth 13, Split Ratio 5%, DT Score : 0.5847
Max-Depth 13, Split Ratio 6%, DT Score : 0.6205
Max-Depth 13, Split Ratio 7%, DT Score : 0.6205
Max-Depth 13, Split Ratio 8%, DT Score : 0.6201
Max-Depth 13, Split Ratio 9%, DT Score : 0.6198
Max-Depth 13, Split Ratio 10%, DT Score : 0.5939
Max-Depth 13, Split Ratio 11%, DT Score : 0.5939
Max-Depth 13, Split Ratio 12%, DT Score : 0.609
Max-Depth 13, Split Ratio 13%, DT Score : 0.6123
Max-Depth 13, Split Ratio 14%, DT Score : 0.6123
Max-Depth 13, Split Ratio 15%, DT Score : 0.571
Max-Depth 13, Split Ratio 16%, DT Score : 0.571
Max-Depth 13, Split Ratio 17%, DT Score : 0.571
Max-Depth 13, Split Ratio 18%, DT Score : 0.571
Max-Depth 13, Split Ratio 19%, DT Score : 0.571

The maximum score 0.6204523107177975 can be obtained at tree depth 13 and split ratio 6%.
```

6.2.1.3 Max Feature Number

We then explored efforts on selecting the optimal maximum feature number for the decision tree model. It is crucial to be aware of the below points when deciding on the feature number:

1. Curse of dimensionality
2. Not all data are of good quality and can be acquired at the same cost

Factoring in the above considerations, we eventually concluded on selecting max feature number of 20, coupled with tree depth 13 and split ratio 6% as it yields the highest score of 0.6205.

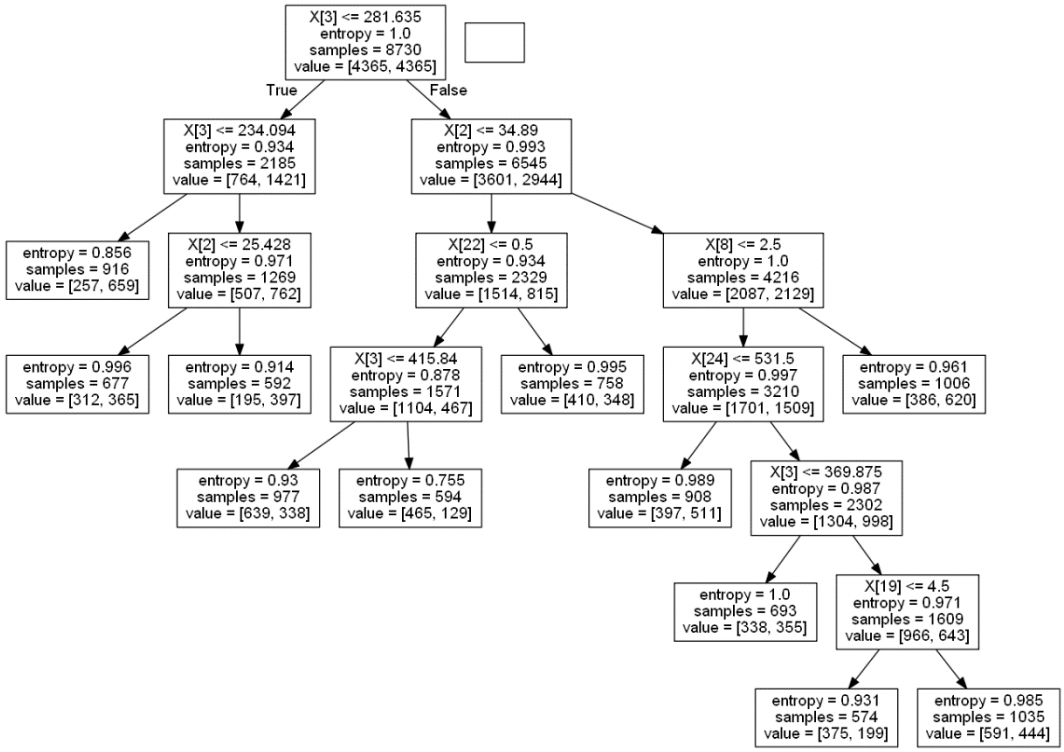
```
Max-Depth 13, Split Ratio 6%, Feature Number 2, DT Score/Error : 0.5461, 0.4539
Max-Depth 13, Split Ratio 6%, Feature Number 3, DT Score/Error : 0.5585, 0.4415
Max-Depth 13, Split Ratio 6%, Feature Number 4, DT Score/Error : 0.5382, 0.4618
Max-Depth 13, Split Ratio 6%, Feature Number 5, DT Score/Error : 0.566, 0.434
Max-Depth 13, Split Ratio 6%, Feature Number 6, DT Score/Error : 0.587, 0.413
Max-Depth 13, Split Ratio 6%, Feature Number 7, DT Score/Error : 0.589, 0.411
Max-Depth 13, Split Ratio 6%, Feature Number 8, DT Score/Error : 0.5788, 0.4212
Max-Depth 13, Split Ratio 6%, Feature Number 9, DT Score/Error : 0.5713, 0.4287
Max-Depth 13, Split Ratio 6%, Feature Number 10, DT Score/Error : 0.5611, 0.4389
Max-Depth 13, Split Ratio 6%, Feature Number 11, DT Score/Error : 0.569, 0.431
Max-Depth 13, Split Ratio 6%, Feature Number 12, DT Score/Error : 0.5942, 0.4058
Max-Depth 13, Split Ratio 6%, Feature Number 13, DT Score/Error : 0.5831, 0.4169
Max-Depth 13, Split Ratio 6%, Feature Number 14, DT Score/Error : 0.5792, 0.4208
Max-Depth 13, Split Ratio 6%, Feature Number 15, DT Score/Error : 0.5857, 0.4143
Max-Depth 13, Split Ratio 6%, Feature Number 16, DT Score/Error : 0.59, 0.41
Max-Depth 13, Split Ratio 6%, Feature Number 17, DT Score/Error : 0.5906, 0.4094
Max-Depth 13, Split Ratio 6%, Feature Number 18, DT Score/Error : 0.589, 0.411
Max-Depth 13, Split Ratio 6%, Feature Number 19, DT Score/Error : 0.5792, 0.4208
Max-Depth 13, Split Ratio 6%, Feature Number 20, DT Score/Error : 0.6205, 0.3795
Max-Depth 13, Split Ratio 6%, Feature Number 21, DT Score/Error : 0.5939, 0.4061
Max-Depth 13, Split Ratio 6%, Feature Number 22, DT Score/Error : 0.6168, 0.3832
Max-Depth 13, Split Ratio 6%, Feature Number 23, DT Score/Error : 0.6047, 0.3953
Max-Depth 13, Split Ratio 6%, Feature Number 24, DT Score/Error : 0.6205, 0.3795
Max-Depth 13, Split Ratio 6%, Feature Number 25, DT Score/Error : 0.6205, 0.3795

The maximum score 0.6204523107177975 can be first obtained at tree depth 13, split ratio 6% and feature number 20.
```

6.2.1.4 Decision Tree Plot

The below table provides a summary of the tuned parameters for our decision tree model and its respective score and error rate.

Parameters	Parameter's value	Decision Tree Result
Max tree depth	13	Score: 0.6205 Error Rate: 0.3795
Split ratio	6%	
Feature number	20	

A visual plot of the decision tree based off the above tuned parameters is illustrated below:



6.2.1.5 Decision Tree Feature Importance

The below table shows the features that are used in the decision tree model as well as their respective feature importance score sorted in descending order:

Features used	Feature importance score
Net_Premium X[3]	0.0294
Finance_Income X[2]	0.0141
Driver_Marital_Status X[8]	0.0055
Vehicle_Transmission X[22]	0.0048
Credit_Score X[24]	0.0036
Vehicle_No_Of_Doors X[19]	0.0009

We can see from the decision tree that “Net Premium” variable is used several times as a splitting feature and has the highest feature importance score at 0.0294. An initial split using “Net Premium” is used, followed by “Finance Income”.

6.2.2 Logistic Regression

Dummy variables, taking on binary values of either 1 or 0 were created for categorical features. Recursive Feature Elimination (RFE) was leveraged on to help with selecting key features that are most relevant in predicting the target variable (Brownlee, 2020). This is done by recursively selecting important features to be trained by the model and eventually pruning away the less

important features. To ensure that RFE converges, standardization using min max scaler was applied.

After standardization and RFE is done, we first perform an explanatory logistic regression. The below image highlights the 20 variables selected via RFE. We then analysed the p-values to determine which independent variables are significant, with a significance value threshold at 5%. “Driver Occupation Status U”, “Driver Occupation Status N”, “Vehicle Body Type C” have p-values lesser than 0.05. As such, we proceed to drop these insignificant features.

Results: Logit

Model:	Logit	Pseudo R-squared:	0.108
Dependent Variable:	Sale	AIC:	10838.2838
Date:	2022-07-02 06:48	BIC:	10979.7743
No. Observations:	8730	Log-Likelihood:	-5399.1
Df Model:	19	LL-Null:	-6051.2
Df Residuals:	8710	LLR p-value:	4.7430e-265
Converged:	1.0000	Scale:	1.0000
No. Iterations:	6.0000		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Finance_Income	4.1220	0.3200	12.8826	0.0000	3.4949	4.7492
Net_Premium	-10.6127	0.4917	-21.5854	0.0000	-11.5763	-9.6491
Driver_Marital_Status_A	2.0256	0.2071	9.7819	0.0000	1.6198	2.4315
Driver_Marital_Status_D	1.8207	0.1338	13.6039	0.0000	1.5584	2.0830
Driver_Marital_Status_M	1.7479	0.1069	16.3486	0.0000	1.5384	1.9575
Driver_Marital_Status_P	2.0293	0.1161	17.4755	0.0000	1.8017	2.2569
Driver_Marital_Status_S	2.0062	0.1174	17.0853	0.0000	1.7760	2.2363
Driver_Marital_Status_W	2.0218	0.1994	10.1402	0.0000	1.6310	2.4126
Driver_Occupation_Status_H	0.3963	0.1257	3.1533	0.0016	0.1500	0.6426
Driver_Occupation_Status_N	0.4583	0.2735	1.6758	0.0938	-0.0777	0.9943
Driver_Occupation_Status_U	0.1992	0.1470	1.3548	0.1755	-0.0890	0.4874
Vehicle_Body_Type_B	0.4865	0.1964	2.4776	0.0132	0.1016	0.8714
Vehicle_Body_Type_C	0.3913	0.2153	1.8173	0.0692	-0.0307	0.8134
Vehicle_Body_Type_E	0.7769	0.1158	6.7103	0.0000	0.5500	1.0038
Vehicle_Body_Type_H	0.5742	0.1098	5.2300	0.0000	0.3590	0.7893
Vehicle_Body_Type_S	0.7396	0.1343	5.5079	0.0000	0.4764	1.0028
Vehicle_Fuel_Type_D	-0.8142	0.1178	-6.9119	0.0000	-1.0451	-0.5833
Vehicle_Fuel_Type_P	-0.7452	0.1191	-6.2576	0.0000	-0.9786	-0.5118
Vehicle_Transmission_A	-0.6093	0.1242	-4.9062	0.0000	-0.8528	-0.3659
Vehicle_Transmission_M	-0.8858	0.1156	-7.6659	0.0000	-1.1123	-0.6593

We then arrive at new coefficients after the insignificant variables are dropped, where all variables are significant. The following features in the image below are then used to train our logistic regression model.

Results: Logit

Model:	Logit	Pseudo R-squared:	0.107
Dependent Variable:	Sale	AIC:	10840.1838
Date:	2022-07-02 06:48	BIC:	10960.4506
No. Observations:	8730	Log-Likelihood:	-5403.1
Df Model:	16	LL-Null:	-6051.2
Df Residuals:	8713	LLR p-value:	3.3483e-266
Converged:	1.0000	Scale:	1.0000
No. Iterations:	6.0000		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Finance_Income	4.1324	0.3196	12.9288	0.0000	3.5060	4.7589
Net_Premium	-10.6104	0.4911	-21.6039	0.0000	-11.5730	-9.6478
Driver_Marital_Status_A	2.0613	0.2064	9.9851	0.0000	1.6567	2.4659
Driver_Marital_Status_D	1.8537	0.1331	13.9222	0.0000	1.5928	2.1147
Driver_Marital_Status_M	1.7700	0.1063	16.6523	0.0000	1.5617	1.9784
Driver_Marital_Status_P	2.0531	0.1155	17.7728	0.0000	1.8267	2.2795
Driver_Marital_Status_S	2.0366	0.1166	17.4706	0.0000	1.8081	2.2650
Driver_Marital_Status_W	2.0472	0.1990	10.2849	0.0000	1.6571	2.4373
Driver_Occupation_Status_H	0.3845	0.1256	3.0618	0.0022	0.1384	0.6307
Vehicle_Body_Type_B	0.3900	0.1883	2.0713	0.0383	0.0210	0.7590
Vehicle_Body_Type_E	0.6861	0.1025	6.6961	0.0000	0.4853	0.8870
Vehicle_Body_Type_H	0.4827	0.0962	5.0182	0.0000	0.2942	0.6713
Vehicle_Body_Type_S	0.6474	0.1225	5.2863	0.0000	0.4074	0.8874
Vehicle_Fuel_Type_D	-0.7816	0.1159	-6.7415	0.0000	-1.0089	-0.5544
Vehicle_Fuel_Type_P	-0.7055	0.1171	-6.0256	0.0000	-0.9350	-0.4760
Vehicle_Transmission_A	-0.5632	0.1209	-4.6604	0.0000	-0.8001	-0.3264
Vehicle_Transmission_M	-0.8448	0.1129	-7.4801	0.0000	-1.0661	-0.6234

Subsequent evaluation metrics comparison between the decision tree and logistic regression model will be discussed in section 7.

Before moving on to the section on analysis and evaluation metrics, we round off this section with the recommendation that further parameter tuning can be explored in both the decision tree and logistic regression models. Additionally, other ensemble models such as Random Forest and XGBoost can also be utilized.

7. Analysis & Evaluation Metrics

The clustering results identified in section 6.1 can help BGL easily identify which customers segments are likely to renew their policies, thereby allowing BGL to design its targeted marketing actions accordingly.

Across the board, the various clusters have an average retention rate of 61.3%, which indicates that there is room for BGL to improve and retain its customers. For customers who have been identified as unlikely to renew their policies, BGL should work towards finding out the possible rationale and place extra effort to convert the customers accordingly.

While Cluster 6 generates the highest amount of revenue for BGL, its retention rate is the lowest at 56.1%. BGL should pay attention to this customer segment which generates the highest revenue and focus on increasing their retention rates. More efforts can be placed to conduct focus studies to understand the needs of this particular segment to allow BGL to effectively retain this relationship in the long run.

Cluster 1 on the other hand has the highest retention rate but yields the lowest amount of revenue for BGL. As such, when BGL is planning on the execution of its marketing strategy for this target group, it is especially crucial for BGL to weigh in cost constraints such as marketing budget.

Given that the cost of sending out marketing emailers is low, BGL can do this across all cluster segments to ensure that it remains on the top of its customers' mind amidst the competitive insurance landscape. Thereafter, it can then undertake more specific and targeted marketing efforts such as focus group to better understand or provide incentive coupons to lock in its highest revenue generating customers.

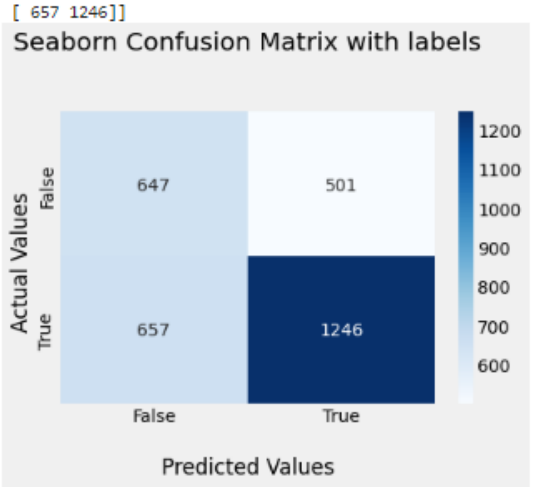
Next, we delve into the detailed classification report results which include precision, recall, f1-score, and a confusion matrix visual. The results for the decision tree and logistic regression model are on the left and right of the image below respectively.

Evalutation metrics

	precision	recall	f1-score	support
0	0.50	0.56	0.53	1148
1	0.71	0.65	0.68	1903
accuracy			0.62	3051
macro avg	0.60	0.61	0.61	3051
weighted avg	0.63	0.62	0.62	3051

Confusion matrix

```
[[ 647  501]
 [ 657 1246]]
```

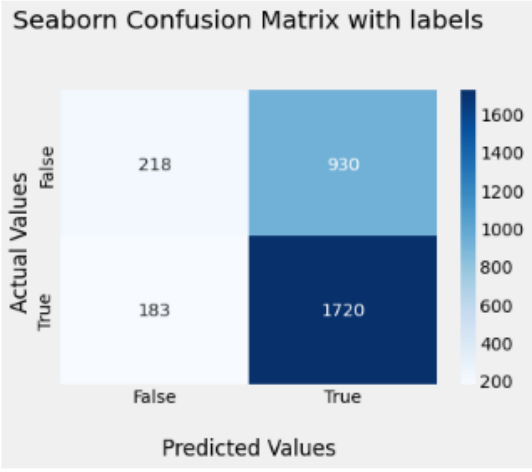


Evalutation metrics

	precision	recall	f1-score	support
0	0.54	0.19	0.28	1148
1	0.65	0.90	0.76	1903
accuracy			0.64	3051
macro avg	0.60	0.55	0.52	3051
weighted avg	0.61	0.64	0.58	3051

Confusion matrix

```
[[ 218  930]
 [ 183 1720]]
```

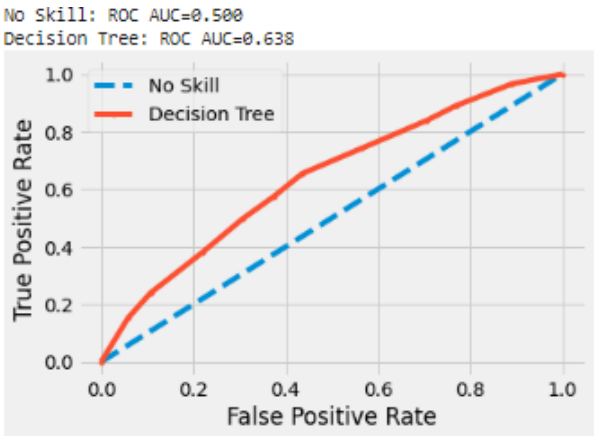


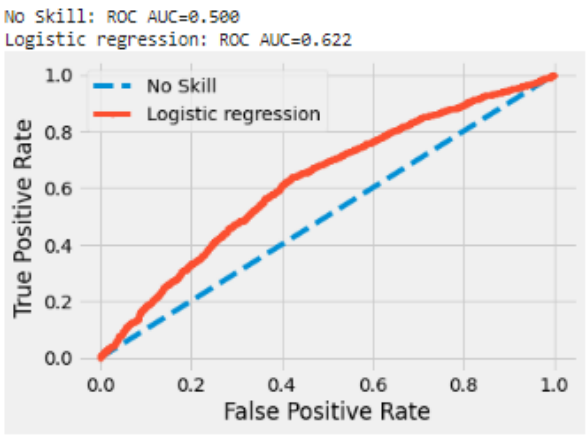
As our business use case is profit maximization, one of the ways BGL can achieve it is through high customer retention rates. With regards to the model evaluation metrics, we note that precision is indeed important as it allows BGL to have precise positive predictions on whether a customer will renew his/her policy. However, we also acknowledge that the cost of marketing via sending of emailers is low, and hence a higher weight should be placed on recall. Given these factors, we tap on F2 score to review our model evaluation metrics. F2 score is calculated from the harmonic mean between the precision and recall, where we assigned a higher weight to recall

using beta of 2. As shown in the table below, we can see that logistic regression has an approximate F2-score of 0.84 while decision tree has an F2-score of 0.67.

	Precision	Recall	F2 Score
Decision Tree	0.713	0.655	0.666
Logistic Regression	0.649	0.904	**0.838**

Looking at the below Receiver Operating Characteristic Curves, the logistic regression model yields an Area Under Curve (AUC) of 0.622 while the decision tree model yields an AUC of 0.638. Given that the AUC score difference is very minimal between both models, coupled with the fact that Logistic Regression yields a higher F2 Score, we recommend going ahead with operationalizing the Logistic Regression model.



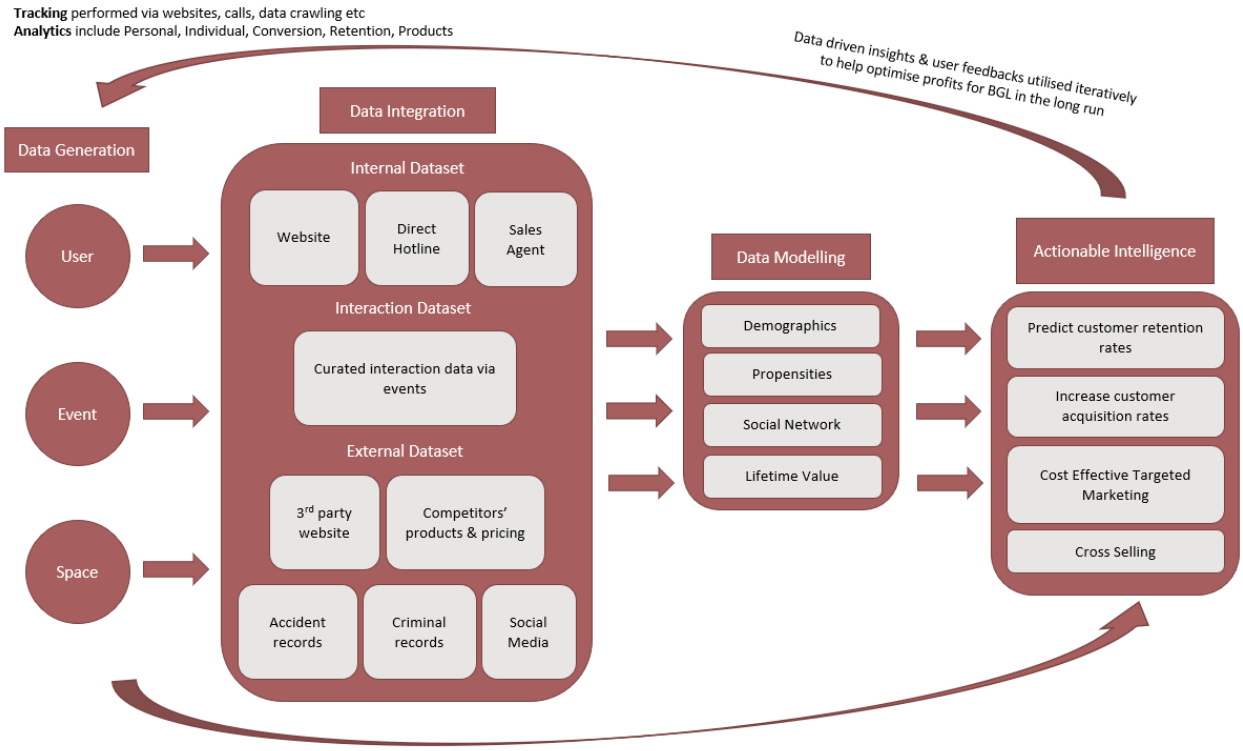


8. Closed Loop Data Ecosystem with User Centric Data Intelligence

Technological advancement will continue to allow BGL to accumulate huge amounts of data, otherwise known as Big Data. Hence, it is crucial to adopt a Data Science thinking approach with regards to the identified business use case/problem scope. Teams often skip to data analysis before properly defining the business task to be addressed, and this was identified as a key reason contributing to companies' low success rates with data science initiatives (Hoerl et al., 2022). Defining concise and targeted business use case/problem statements is a crucial step to ensure the successful implementation of data science initiatives.

Section 8.1 to 8.3 provides further elaboration on the end-to-end process of a closed-loop data eco-system for BGL. A closed-loop data ecosystem is akin to a rinse and repeat cycle, where companies can gain insights for relevant engagement or actions with customers that they engage, before subsequently taking back the engagement results for further analysis (Goetz, 2013). Companies have to continuously take inputs and feedbacks to learn the new and updated data in every closed loop iteration (Goetz, 2013). Enforcing a closed loop data ecosystem (image below) will allow BGL to derive real time analysis and feedback to perform cost optimised targeted

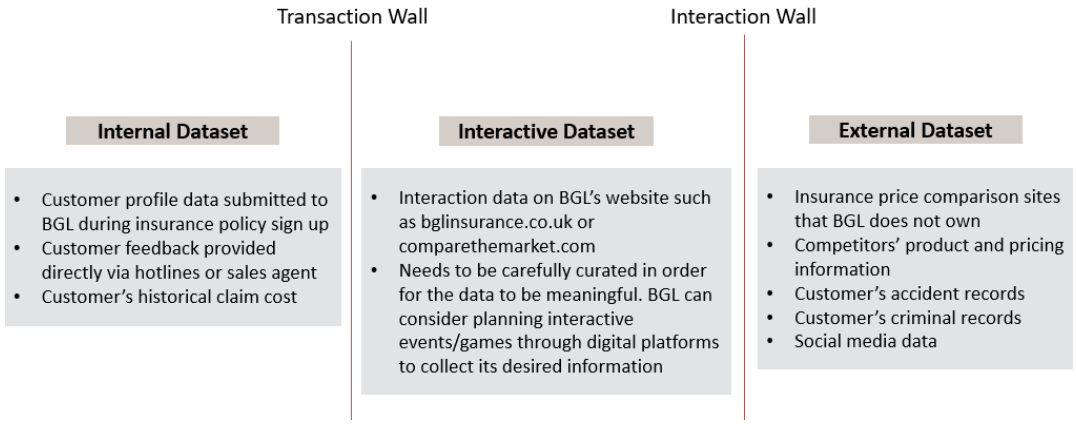
marketing, provide outstanding services and policies in order to effectively acquire and retain customers in the long run. This is especially important in today's saturated insurance landscape to ensure high and sustainable profits in the future.



8.1 Data Generation

A user centric data intelligence framework begins by focusing on the data that is generated from the users and their interaction activities (event) with BGL (space) via BGL's various touchpoints. Omnichannel engagement refers to how policyholders interact with insurers via various channels and is fast becoming a critical distribution strategy (Boosam, 2020). Consumers are increasingly beginning their insurance journey using one channel and eventually concluding their transaction through another. Hence, it is important for BGL to identify and appropriately frame its data generation process which will subsequently affect the data integration, modelling, and analysis stages.

The below figure provides a brief overview of the dataset relevant to BGL, but might not be limited to the below:



8.1.1 Internal Dataset

BGL has a rich source of customer profile data and feedback obtained via various channels. It is crucial for BGL to continually collect updated data from its existing policyholders to ensure that they have the latest record on hand for updated and effective modelling for the task on hand.

Our current dataset does not contain the historical cost of claims for each customer. However, BGL should weave in the cost perspective within its future data modelling approaches. This will enable it to have a more holistic view of each customers' profitability after accounting for the cost of claims.

Weaving in our analysis from the earlier modelling results, there are 7 key features (Finance Income, Net Premium, Driver Marital Status, Driver Occupation Status, Vehicle Body Type, Vehicle Fuel Type, Vehicle Transmission) that are important to collect with regards to customer retention prediction. BGL should continue to work closely with the respective individuals to ensure that customer data are updated and accurate during the customers' insurance profiling and application stage to ensure accurate predictive elements for data driven actionable intelligence as outlined in the image below.

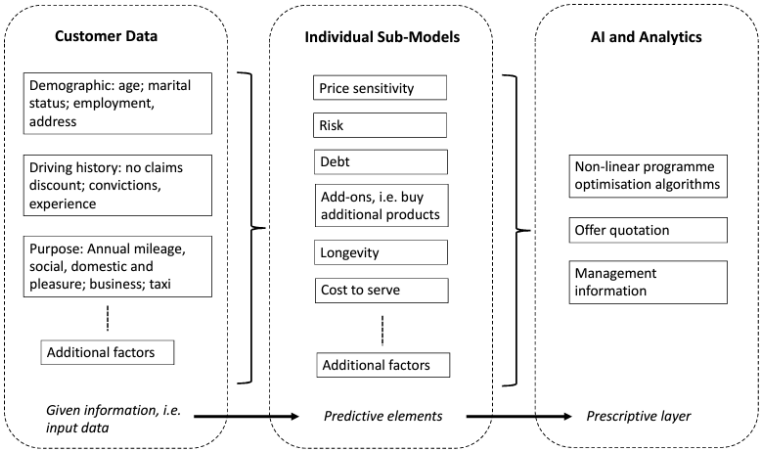


Figure on pricing model based off discussions with BGL's Director of Pricing (Holland, 2021).

8.1.2 Interactive Dataset

When registered BGL users visit comparethemarket.com, their interaction data is tracked and collected. If the user has a pre-existing record with BGL, the newly collected data should be linked to his/her pre-existing profile and policies. Continual data collection will enable BGL to effectively recommend cross-selling of other insurance products to existing car insurance customers.

BGL should also consider collecting customer risk profile information via an application that looks at determining how safe a driver is. This will benefit both BGL as well as its existing and potential customers in the long run, as safer drivers can get a higher score which will then translate to better pricing of auto insurance for the customers and lower insurance claims for BGL. Such an approach has been implemented by Aviva where it assimilates GPS based data on customers' cornering and braking skills to provide a discount to customers who are deemed as safer drivers (Gupta, 2021).

Volume of new business is key within the insurance industry as this directly affects BGL's bottom-line. To effectively acquire new customers, BGL needs to be visible so that it can remain on customers' top of mind when consumers seek for insurance protection. BGL can explore partnering with car dealers during roadshow events to promote their various automotive insurance products through interactive games or events conducted via digital platforms. The interactive dataset collected via the digital platform needs to be carefully planned in line with the business use case to ensure that actionable intelligence insights can be derived.

8.1.3 External Dataset

BGL can utilize external dataset to ensure that its product variety and price are competitive to other competitors. Customer's accident and criminal records can also be useful to BGL in terms of policy pricing for each individual customer. Such information records can also be useful for potential fraudulent claims detection. While BGL presently has an AI-powered counter fraud technology solution which has demonstrably reduced their loss ratios (FinTech Global, 2020), BGL should still continue to review the present dataset that contributes to the solution, constantly refining it and incorporating relevant data along the way.

As seen from the image below, social media can provide BGL with tremendous information to value add in a variety of activities, ranging from customer service to fraud investigation, customer profiling and risk analysis. The positive or negative sentiments around its brand can be continually collected via active social listening to allow for BGL to actively rectify any gaps on a timely basis.

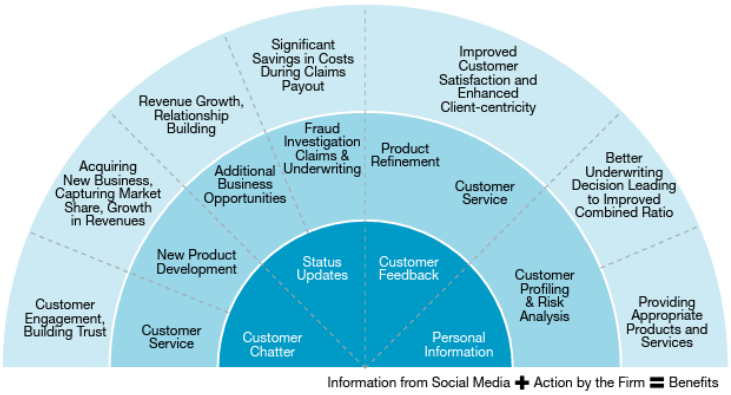


Figure on realizing social media's benefits through timely action (Capgemini, 2014).

8.2 Data Integration

BGL should ensure that it effectively weaves in the data collected from all touchpoints. Proper data pre-processing should also be performed and integrated through relevant employment of algorithm tools before deploying subsequent data modelling and intelligence.

8.3 Data Modelling and Intelligence

With the above foundations built, BGL can then begin data modelling to discover actionable outputs and intelligence based on the identified business modules within section 2 such as "Targeted Marketing" for specific cluster segments. Further details on the modelling results can has been discussed earlier within section 6 and 7.

To ensure a strong closed loop data ecosystem, BGL should continue to track feedbacks, collect updated data to iteratively improve its model for better data driven actionable insights.

9. Data Governance & Considerations

Given that an entire insurance company's reputation is built on trust, it is crucial that insurers walk the talk on ethics and ensure that customers' data are protected (Hohne, n.d.). BGL needs to ensure that there are proper data asset governance mechanism in place to guide and reinforce the integrity and robustness of the closed loop data ecosystem in the long run.

The usage of customer personal data comes with certain ethical considerations and data sensitivity. BGL needs to be mindful of country specific data privacy rules and regulations to ensure compliance. BGL should adhere to the "Minimalism" and "Accessibility" principle where data asset operation right is granted only to individuals on a "needs" basis.

Federated learning is a good tool for BGL to consider leveraging to circumvent privacy concerns. By utilizing for example horizontal federated learning, BGL can gain access to other insurance companies' data with potentially different customer base but similar business features. As seen

from the image below, BGL and other insurance companies can both train their model locally, before sending back model parameters to the global risk model, thereby avoiding any actual data transfer.

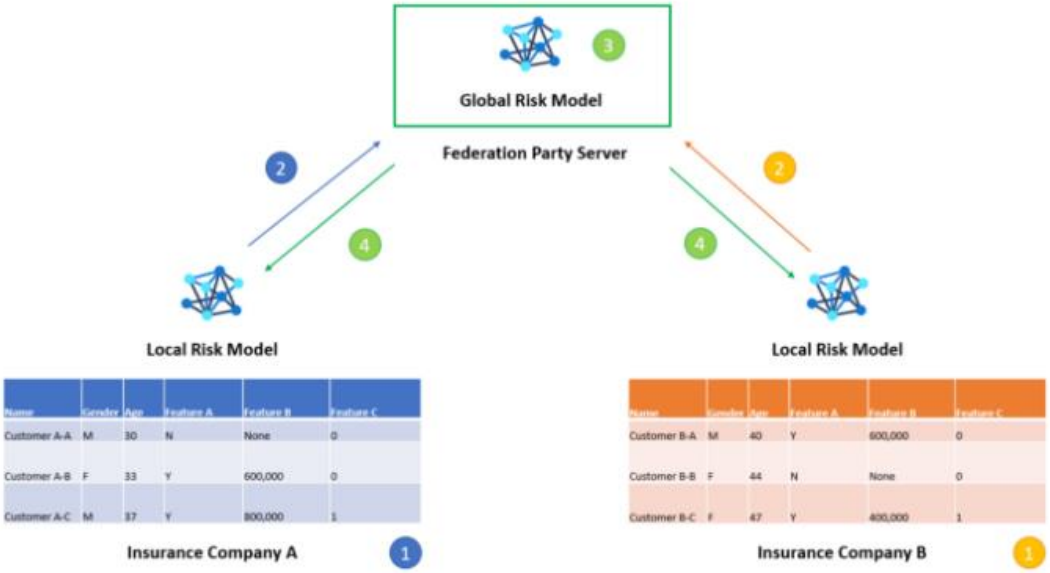


Figure showcasing an example of horizontal federated learning (Eason, 2021).

10. Recommendations & Conclusion

The current dataset on hand focuses on building models to predict customers' retention rate. Given that our business use case is to maximise BGL's profits, we recommend for BGL to continue collecting relevant data and effectively integrate it into their data warehouse and modelling for additional data driven actionable insights to help drive long run profits. We propose for continual collection of relevant information such as customer acquisition cost, claims cost and user demographic information from potential new customers that can be obtained from various touch points such as comparethemarket.com. This information can then actively contribute to the identified business modules of customer acquisition and cost optimisation on a going forward basis.

As consumers are increasingly tech-savvy and highly connected to their mobile phones, BGL should adopt a seamless avenue for individuals to conveniently make an insurance purchase or review their existing policies via an application. This is also another means of allowing BGL to effectively capture both user and interaction data for further analysis and actions.

Finally, while the current BGL dataset we have on hand does not contain a comprehensive suite of insurance policies owned by the individual customers, we note that Recommendation System for Car Insurance can be applied in insurance policies recommendations via cross selling of other

insurance products or upselling of existing car insurance policies (Lesage et al., 2020). BGL can also leverage on Recommender System by looking at the unique attributes and profiles of customers who purchased other insurance products (for example, home insurance) apart from automotive insurance. BGL can then perform effective targeted marketing and recommendation of similar products to users who are likely to be converted.

In conclusion, a well-designed and robust closed loop data ecosystem will set BGL for success in the long run within the competitive insurance industry.

11. Appendix

Figure 1

Sales and Pricing Data

Renewal_Date	The date the policy was due to be renewed
Sale	Did the policy renew?
Main_Income	The main income generated on the core policy
Addon_Income	The income made through additional products on the policy
Finance_Income	The income made off financing a policy if it is paid by installments
Gross_Price	The gross price the customer receives
Last_Yr_Gross_Price	The gross price the customer received last year
Net_Premium	The net rate provided by our underwriter
Renewal_Offer_Type	The type of offer the customer received (Active - the customer must actively call in to renew, Passive- the policy renews unless the customer instructs us otherwise)

Driver's Individuals

Diver_Age	The driver's age
Driver_Marital_Status	The driver's marital status
Driver_Occupation_Status	The driver's occupation status
Driver_Licence_Type	The driver's license type
Driver_Years_Licence_Held	The years the driver has held their license
Geographical_Area	The policy holder's geographic location
Number_Of_ Named_Drivers	The number of named drivers on the policy
Credit_Score	The customer's public credit score

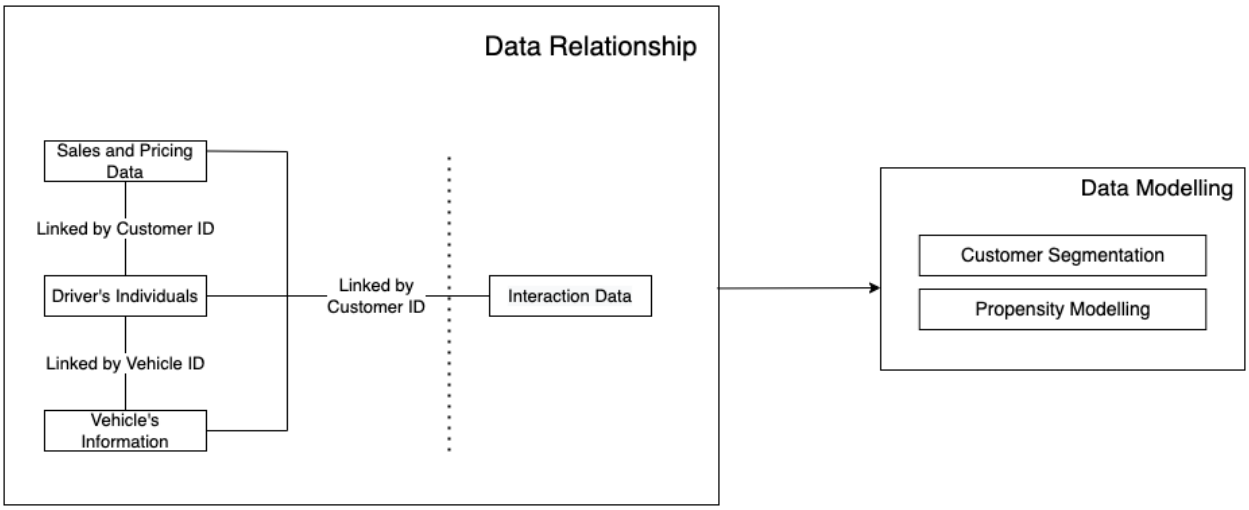
Vehicle's Information

Overnight_Parking	How the car is parked overnight
Annual_Mileage	The annual mileage stated on the policy
Vehicle_Age	The vehicle's age
Vehicle_Years_Owned	The number of years the current owner has owned the vehicle
Vehicle_Body_Type	The vehicles body shape
Vehicle_No_Of_Doors	The number of doors the vehicle possesses including doors
Vehicle_Engine_Size	The vehicle engine size
Vehicle_Fuel_Type	The vehicles fuel type, (P-Petrol, D-Diesel, E-Electric)
Vehicle_Transmission	vehicle transmission type (M-Manual, A-Automatic)
Vehicle_Liability_Category	The vehicles liability category 1= very low, 20=very high

Interaction Information – These data are collected based on past customers activities

Previous_Renewals	The number of times the customer has previously renewed.
First_Contact_Channel	The initial contact made by the customer (I-Internet, T-Telephone)
Number_Of_Attempted_ Policy_Ammendments	The number of attempted amendments to the policy in the previous year
No_Claims_Discount	How many years No Claims discount the customer has
Previous_Discount	Whether the customer has asked for a discount at previous renewals

Figure 2



12. References

Boosam, K. (2020, September 1). *Tech-savvy policyholders require a blend of digital and emotional capabilities from insurers.* https://www.capgemini.com/insights/expert-perspectives/tech-savvy-policyholders-require-a-blend-of-digital-and-emotional-capabilities-from-insurers/

Brownlee, J. (2020, January 17). *SMOTE for imbalanced classification with python.* https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/

Brownlee, J. (2020, May 25). *Recursive feature elimination (RFE) for feature selection in python.* https://machinelearningmastery.com/rfe-feature-selection-in-python/

Dandamudi, S., Deel, L.V., & Nadimpalli, R. (2014). *Leveraging social media across the insurance lifecycle.* Capgemini. https://www.the-digital-insurer.com/wp-content/uploads/2015/04/498-leveraging_social_media_across_the_insurance_lifecycle.pdf

Eason. (2021, November 23). *Exploring federated machine learning in the insurance sector.* https://medium.com/analytics-vidhya/exploring-federated-machine-learning-in-the-insurance-sector-cf9708f1d53d

FinTech Global. (2020, October 13). *BGL Group launches AI tool to prevent false auto claims.* https://member.fintech.global/2020/10/13/bgl-group-launches-ai-tool-to-prevent-false-auto-claims/

Goetz, M. (2013, July 26). *Data science and "closed-loop" analytics changes master data strategy.* https://www.forrester.com/blogs/13-07-26-data_science_and_closed_loop_analytics_changes_master_data_strategy/

Gupta, P. (2021, March 23). *Data analytics making auto insurance cheaper.* https://digital.hbs.edu/platform-digit/submission/data-analytics-making-auto-insurance-cheaper/

Hardesty, C. (2021, September 22). *Average miles driven per year: why it is important.* https://www.kbb.com/car-advice/average-miles-driven-per-year/#:~:text=What%20Are%20Average%20Miles%20Driven,about%2039%20miles%20per%20day.

He, Z., Gu, F., Zhao, C., Liu, X., Wu, J., & Wang, J. (2017). Conditional discriminative pattern mining: concept and algorithms. *Elsevier, 375,* 1-15. https://www.sciencedirect.com/science/article/abs/pii/S0020025516309860

Hoerl, R., Kuonen, D., & Redman, T.C. (2022, April 14). *Framing data science problems the right way from the start.* MIT Sloan Management Review. https://sloanreview.mit.edu/article/framing-data-science-problems-the-right-way-from-the-start/

Hohne, M. (n.d.). *The smart insurer: embedding big data in corporate strategy.* https://www.bearingpoint.com/en/insights-events/insights/the-smart-insurer-embedding-big-data-in-corporate-strategy/

Holland, C.P. (2021, May 7). *BGL group: Artificial intelligence (AI) strategy.* https://ssrn.com/abstract=3841656 or http://dx.doi.org/10.2139/ssrn.3841656

Kumbhar, A. (2022, April 8). *Digitalisation – Future of Insurance Industry?* https://www.myinsuranceclub.com/articles/digitalisation-future-of-insurance-industry

Laurent, L.., Deaconu, M., Lejay, A., Meira, J., Nichil, G., & State, R. (2020). A recommendation system for car insurance. *European Actuarial Journal, Springer, 10,* 377-398.

Thompson, C. (2016, July 27). *A tech-savvy future for insurance.* https://www.odgersberndtson.com/fr-fr/insights/a-tech-savvy-future-for-insurance