Carlytical - Data Analysis Report

Group 5

Hasitha Jayaweera - 190277L

Kajanan Selvanesan - 190287R

Zafra Rifky - 190525X

Mentor - Dr. Nisansa de Silva

Problem Overview

Telemarketing is widely used by businesses to promote themselves and reach target customers. This project revolves around SpeedStar, a vehicle insurance service provider, and its telemarketing campaign. Since it is important to balance between avoiding being a nuisance to the general public and reaching potential customers through the telemarketing campaign, this project focuses on identifying potential customers.

Given certain information about a customer, a predictive model would correctly predict whether that customer will purchase vehicle insurance or not. Additionally, valuable insights will be derived from the given dataset and presented as an interactive dashboard, making it easier for stakeholders to make important business decisions.

Description of The Dataset

The given dataset contains data collected from the company's most recent telemarketing campaign and their results. This includes data from 4000 call attempts out of which 1604 were successful in getting customers to subscribe to an insurance plan. This dataset will be preprocessed and used as input to a predictive model to obtain the prediction with the highest accuracy. The demographic information included in the dataset can be analysed and used to derive useful insights.

- **Feature Overview**

  The dataset consists of 18 predictor variables and one target variable.

Variable Name	Data Type	Attribute Type	Description
Id	String	Categorical Nominal	Unique ID number
Age	Integer	Metric Discrete	Age of the client
Job	String	Categorical Nominal	Job of the client

Marital	String	Categorical Nominal	Marital status of the client
Education	String	Categorical Ordinal	Education level of the client
Default	Integer	Categorical Nominal	Has credit in default?
Balance	Integer	Metric Discrete	Average yearly balance, in USD
HHInsurance	Integer	Categorical Nominal	Is household insured
CarLoan	Integer	Categorical Nominal	Has the client a car loan
Communication	String	Categorical Nominal	Contact communication type
LastContactMonth	String	Categorical Ordinal	Month of the last contact
LastContactDay	Integer	Metric Discrete	Day of the last contact
CallStart	Time	Metric Continuous	Start time of the last call (HH:MM:SS)
CallEnd	Time	Metric Continuous	End time of the last call (HH:MM:SS)
NoOfContacts	Integer	Metric Discrete	Number of contacts performed during this campaign for this client
DaysPassed	Integer	Metric Discrete	Number of days that passed by after the client was last contacted from a previous campaign (numeric; -1 means client was not previously contacted)
PrevAttempts	Integer	Metric Discrete	Number of contacts performed before this campaign and for this client
Outcome	String	Categorical Nominal	Outcome of the previous marketing campaign
CarInsurance	Integer	Categorical Nominal	Has the client subscribed a car insurance?

Exploratory Tool and Features used

Exploratory data analysis tools provide a better understanding of the data variables and their relationships. Through these tools it can be determined the best features. And it helps to create new features using the business knowledge as well.

The main tool which is used for exploratory data analysis is python. With the use of python libraries it is very easy to visualise and manipulate data efficiently. Libraries which are used for exploratory data analysis are numpy, seaborn, matplotlib, sklearn and pandas. Those libraries provide functions to get correlations among attributes as well. With using these libraries further, it can be identified patterns, outliers and trends. To impute missing values, KNN imputation will be used and there are functions for data cleaning.

Moreover for exploratory data analysis Microsoft Power BI is being used. This tool helps to analyse insights well and create interactive visualisations. It provides functionality to load excel data as a file and get visualisation plots such as stacked bar chart, stacked column chart, clustered bar chart, clustered column chart, 100% stacked bar chart , 100% stacked column chart, pie chart , donut chart and scatter chart etc.

Justification of technologies used for the tool

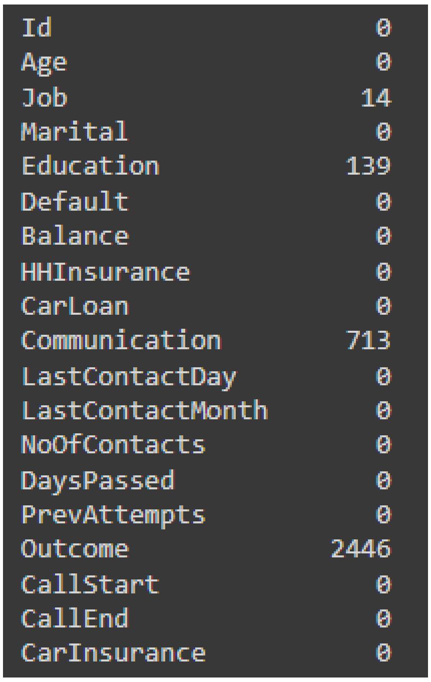
The Python pandas, numpy, seaborn, matplotlib libraries can be used for data preprocessing and exploratory data analysis. Pandas dataframes are capable of handling large datasets. Numpy is a powerful library that is extremely useful to analyse large datasets. Seaborn and matplotlib are great visualisation tools.

Power BI is a cloud-based solution backed by Microsoft. It is relatively cheaper and contains a free version for single users. It is user friendly, and no prior knowledge is required to use it. Chart types such as line graphs, bar graphs, pie charts, area charts, and scatter plots provide effective descriptive analysis. Multiple graphs can be combined to derive comparative insights between variables. It is also extremely user-friendly, with easy-to-use drag and drop features. It also contains clear documentation to refer to. Power BI graphs also contain tooltips which are customizable and will display more details upon hovering over the graph. Therefore, PowerBI is a good choice to create the dashboard in this web app.

Descriptive Data Analysis

The dataset contains 18 predictor variables and 1 target variable. It consisted of 3200 call records and was checked for duplicates, but no duplicates were found.

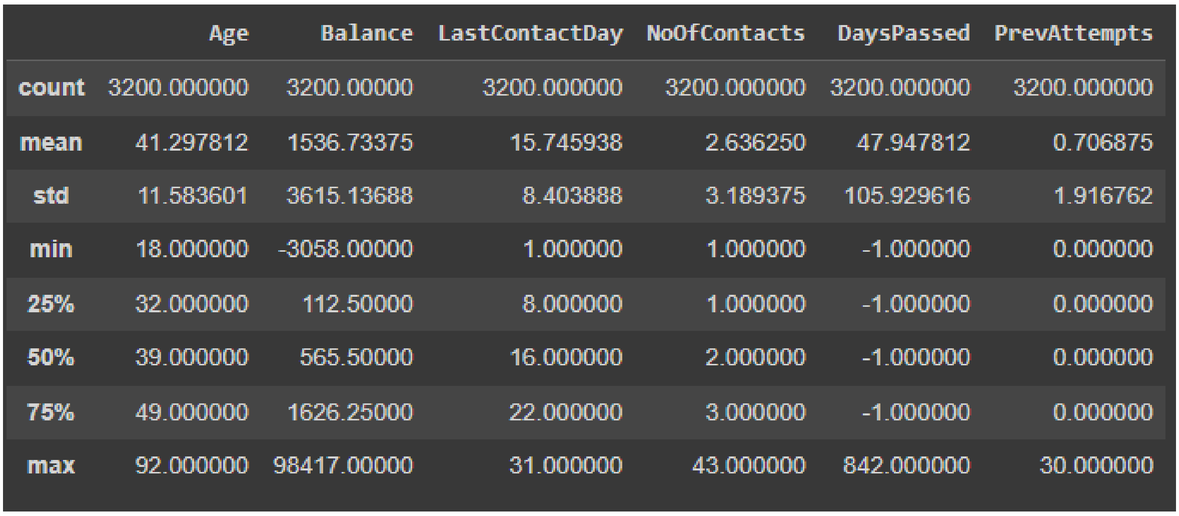
Then it was checked for missing values. The column name and the number of missing values it contained are depicted below.



```
Id                  0
Age                 0
Job                14
Marital             0
Education         139
Default             0
Balance             0
HHInsurance         0
CarLoan             0
Communication     713
LastContactDay      0
LastContactMonth    0
NoOfContacts        0
DaysPassed          0
PrevAttempts        0
Outcome          2446
CallStart           0
CallEnd             0
CarInsurance        0
```

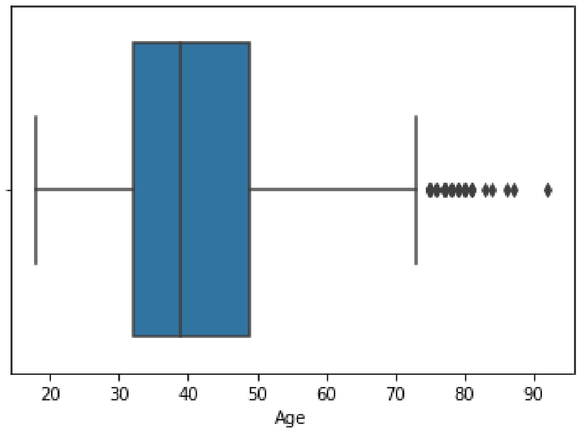
Missing values for features ‘Job’ and ‘Communication’ were replaced with a new variable named ‘Other’. Since the missing values for the ‘Outcome’ feature corresponded to customers not contacted through a previous campaign, they were replaced with ‘Not Contacted’.

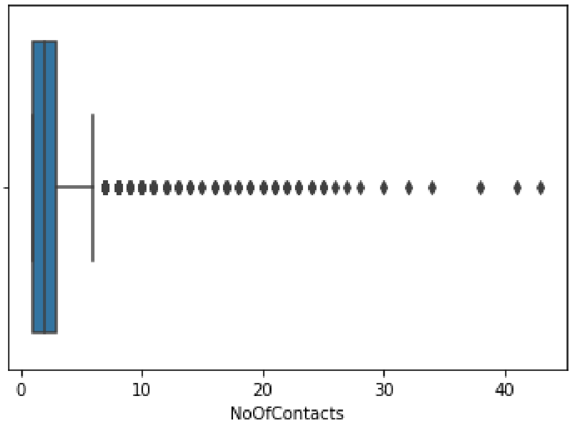
For all the numerical variables in the data set, the 5-number summary was generated to get an idea of the spread of values.

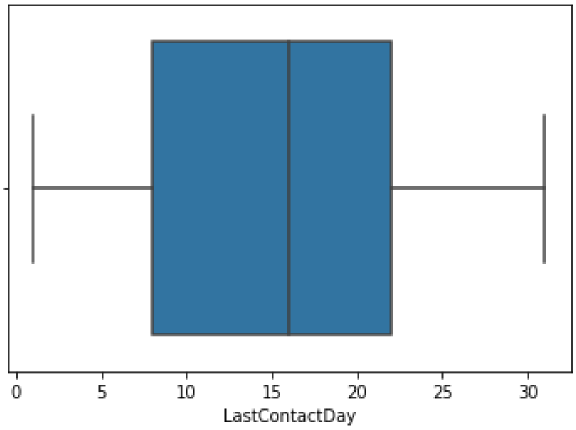


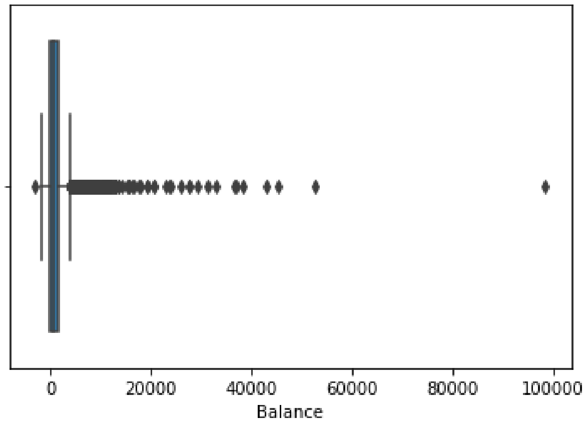
	Age	Balance	LastContactDay	NoOfContacts	DaysPassed	PrevAttempts
count	3200.000000	3200.00000	3200.000000	3200.000000	3200.000000	3200.000000
mean	41.297812	1536.73375	15.745938	2.636250	47.947812	0.706875
std	11.583601	3615.13688	8.403888	3.189375	105.929616	1.916762
min	18.000000	-3058.00000	1.000000	1.000000	-1.000000	0.000000
25%	32.000000	112.50000	8.000000	1.000000	-1.000000	0.000000
50%	39.000000	565.50000	16.000000	2.000000	-1.000000	0.000000
75%	49.000000	1626.25000	22.000000	3.000000	-1.000000	0.000000
max	92.000000	98417.00000	31.000000	43.000000	842.000000	30.000000

For a more visual and intuitive representation, boxplots were generated.

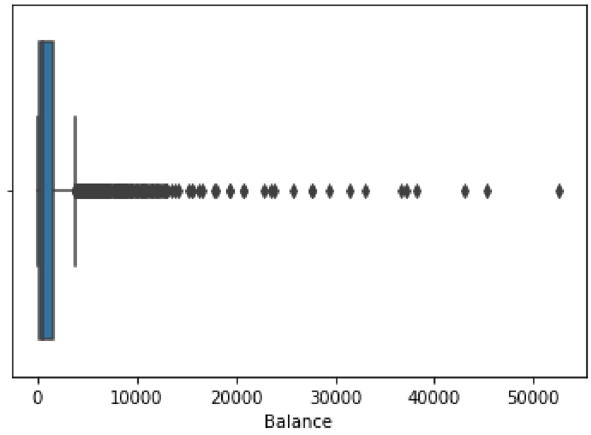




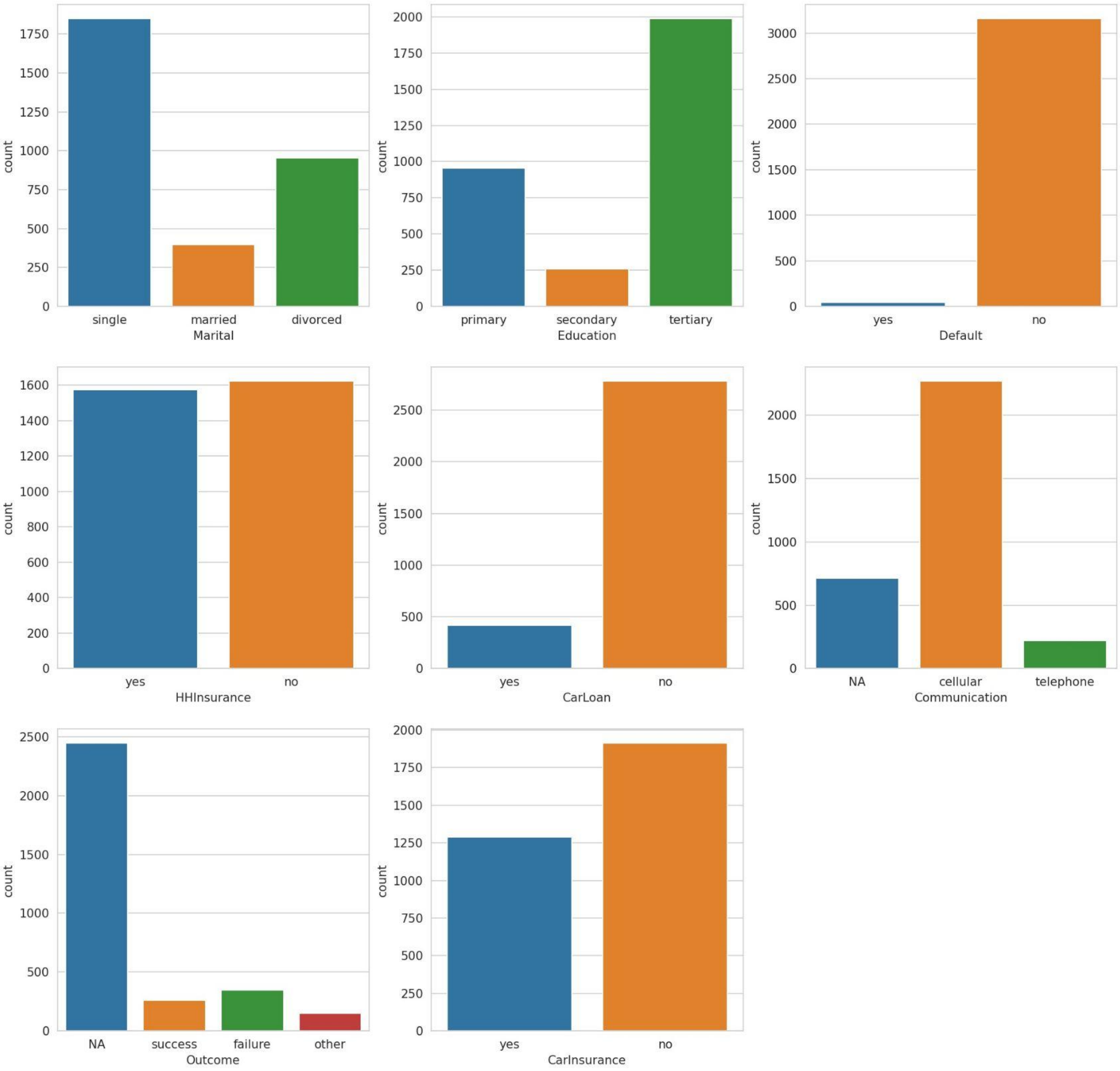


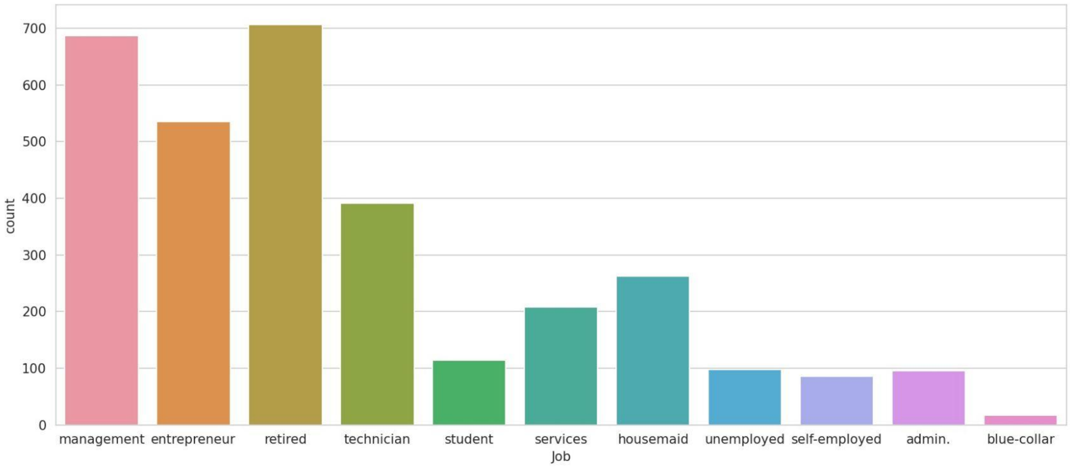


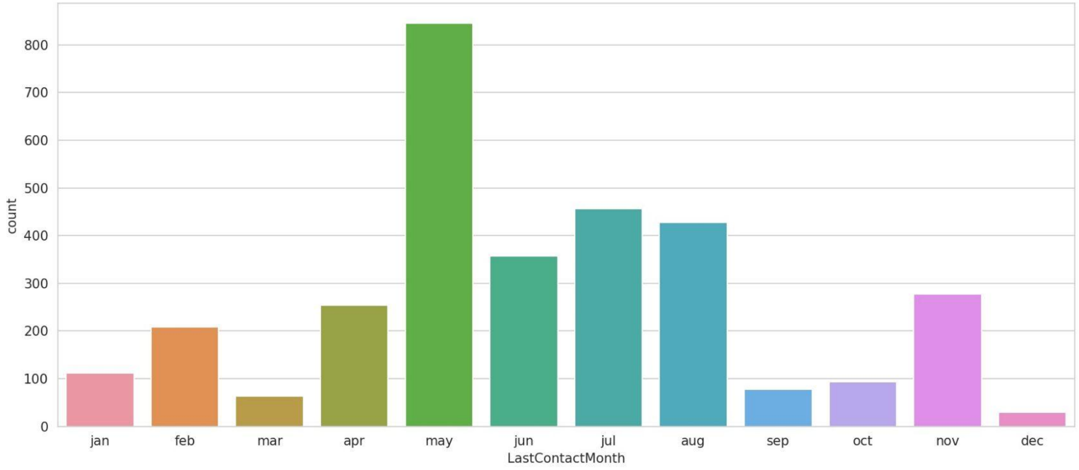
It can be observed that the Balance feature has a clear outlier. This outlier was replaced with the mean, and the box-plot before the replacement was as follows.



For all the categorical variables in the data set, the count-plot was generated to get an idea of the spread of classes.

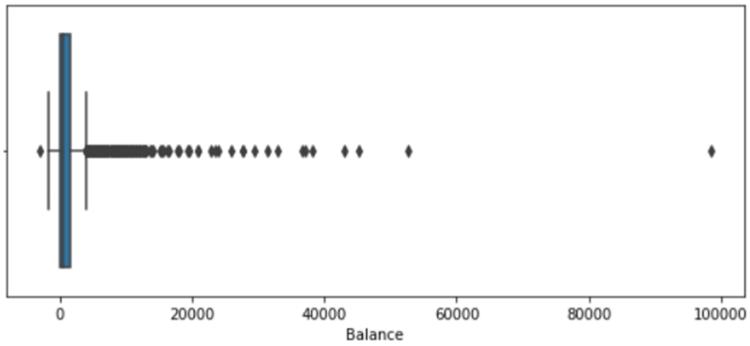

count
0
250
500
750
1000
1250
1500
1750
single
married
divorced
Marital
count
0
250
500
750
1000
1250
1500
1750
2000
primary
secondary
tertiary
Education
count
0
500
1000
1500
2000
2500
3000
yes
no
Default
count
0
200
400
600
800
1000
1200
1400
1600
yes
no
HHInsurance
count
0
500
1000
1500
2000
2500
yes
no
CarLoan
count
0
500
1000
1500
2000
NA
cellular
telephone
Communication
count
0
500
1000
1500
2000
2500
NA
success
failure
other
Outcome
count
0
250
500
750
1000
1250
1500
1750
2000
yes
no
CarInsurance


count
0
100
200
300
400
500
600
700
management
entrepreneur
retired
technician
student
services
housemaid
unemployed
self-employed
admin.
blue-collar
Job



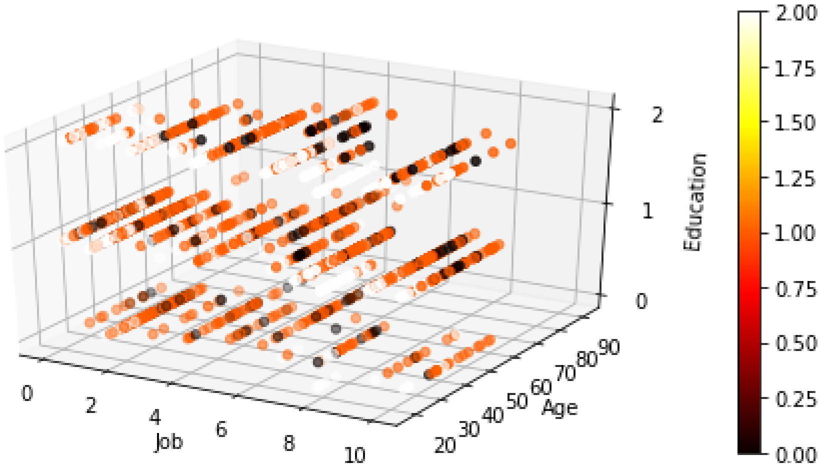
Exploratory Data Analysis

To prevent anomalies, it is important to remove outliers if there are outliers detected in the dataset. This is the box plot representation for the feature "Balance".

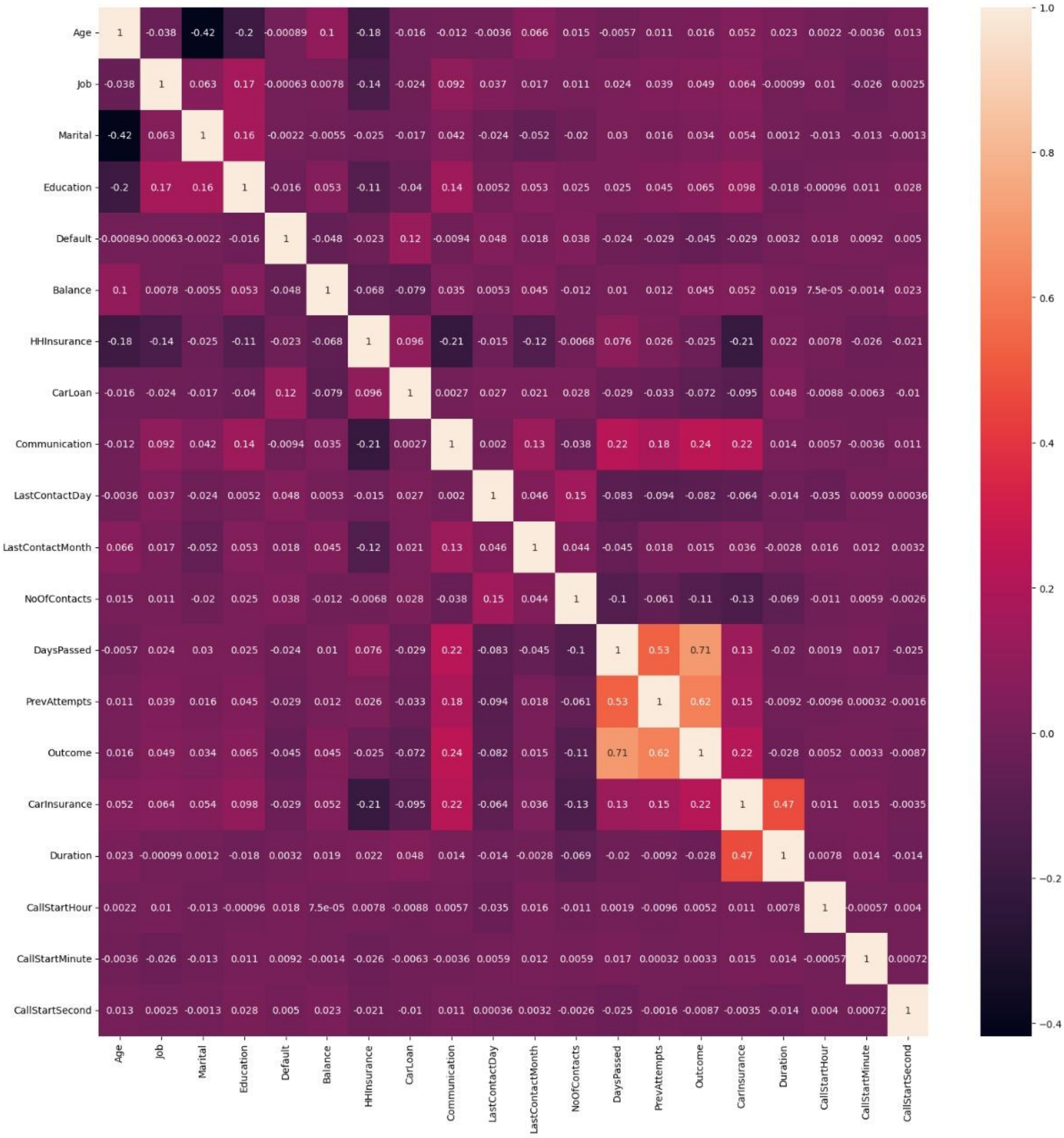


There can be identified minus values as well as very large values compared to other Balance values. There should not be that kind of data points because it may reduce the statistical significance of the data. Therefore, minus data are replaced by absolute values and the outlier data is replaced by the mean of the data except considering outliers.

In the given dataset there are missing values in Job and Education columns. To fill those missing values the KNNImputer will be used.



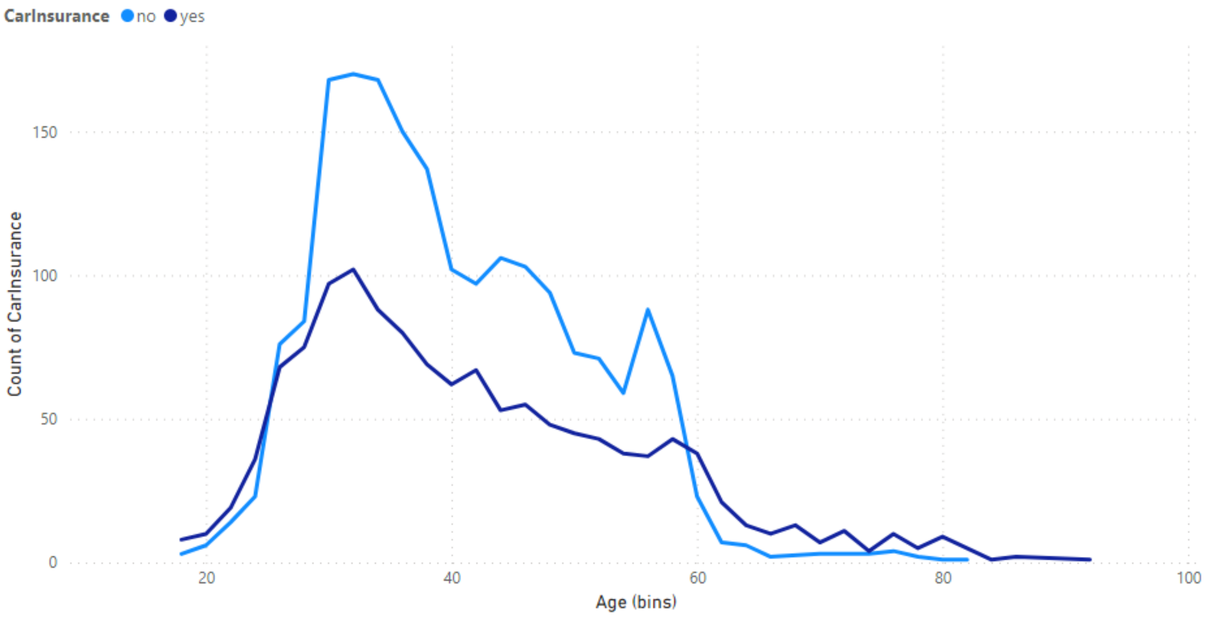
There can be identified 11 distinct values for Job, 3 distinct values for Education and 3 distinct values for Marital. The above graph shows how Job, Marital, Education and Age get spread. "Marital" is represented by colour. Missing values will be filled by the outcome of the KNNImputer.



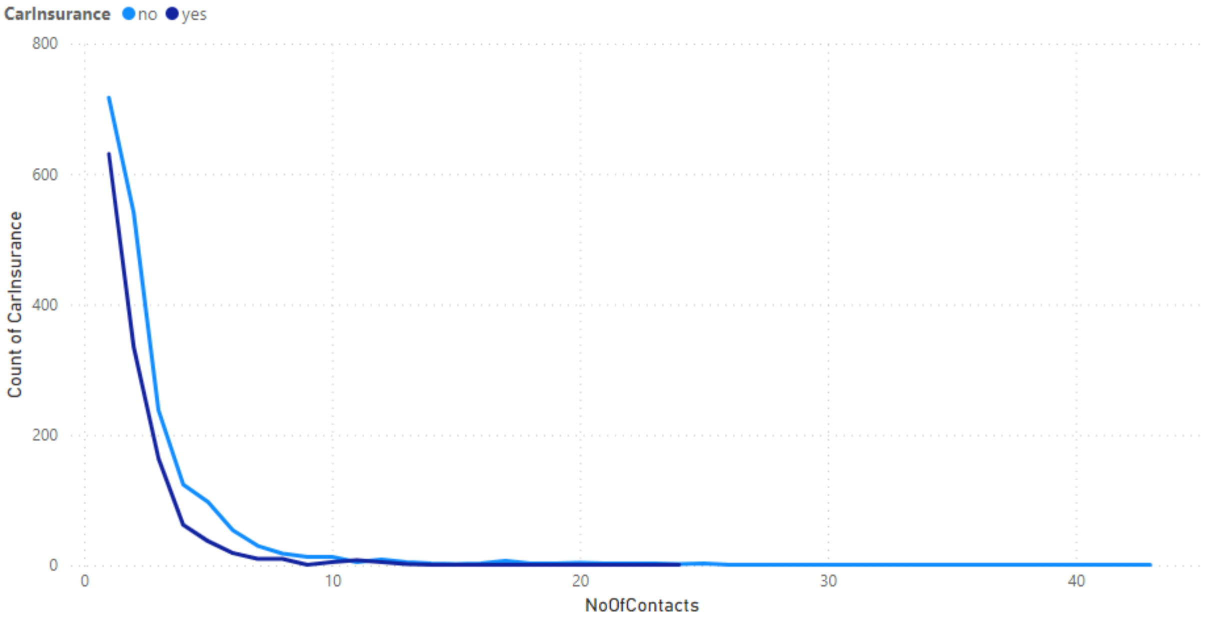
It can be observed that there are significant correlations between certain features. It is expected since the features related to contacting customers through a previous campaign would be correlated.

- ## Visualisations

Each of the predictor variables were analysed as shown in the graphs below.

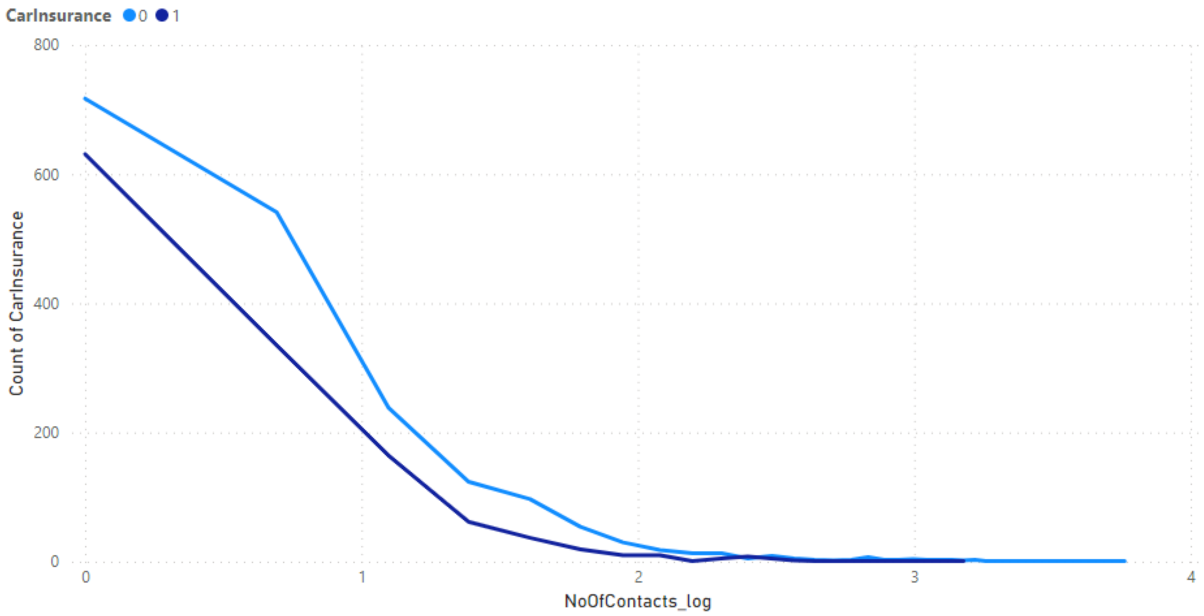


If a client's age is less than 25 or greater than 60, then that client is more likely to purchase car insurance. However, clients who are between 25 to 60 were mostly targeted for the previous campaigns. In this range, clients aren't interested in purchasing car insurance.

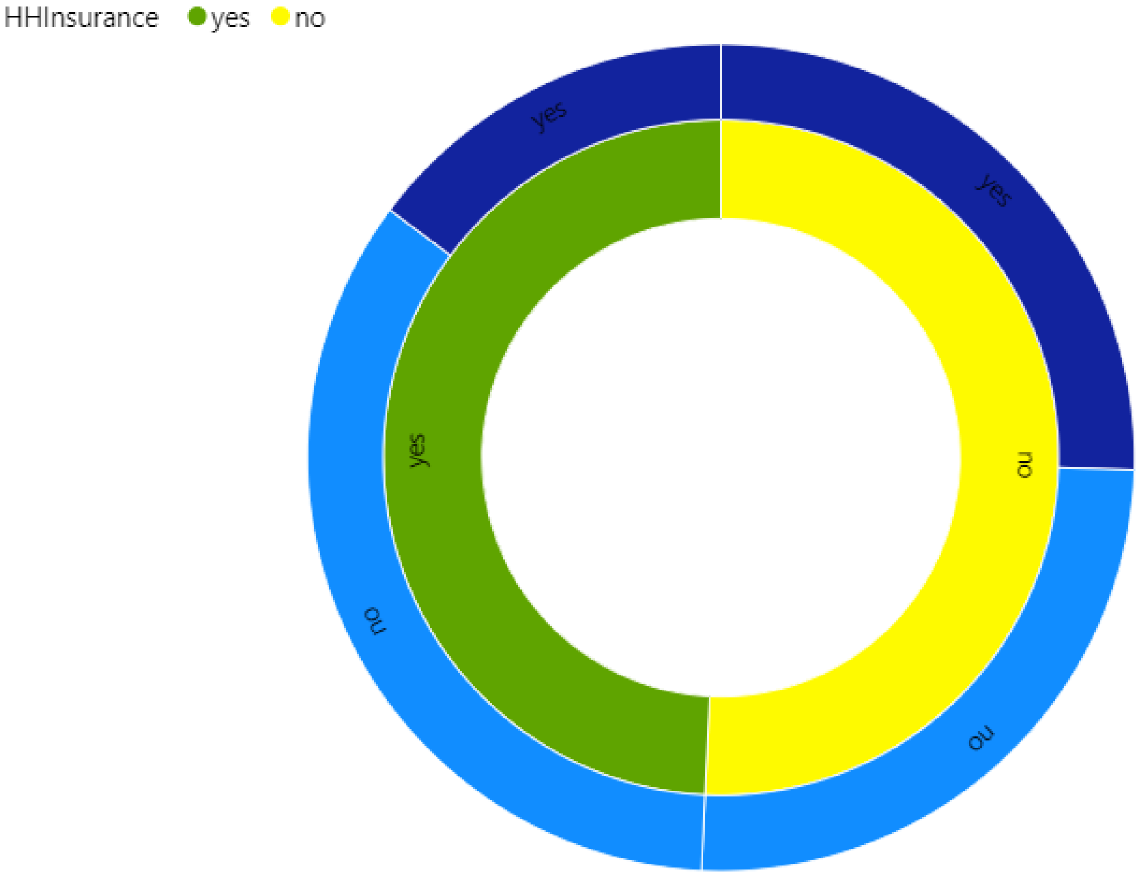


NoOfContacts refers to the number of contacts performed during the campaign for a client. It's not that much affect the number of clients who are going to purchase car insurance.

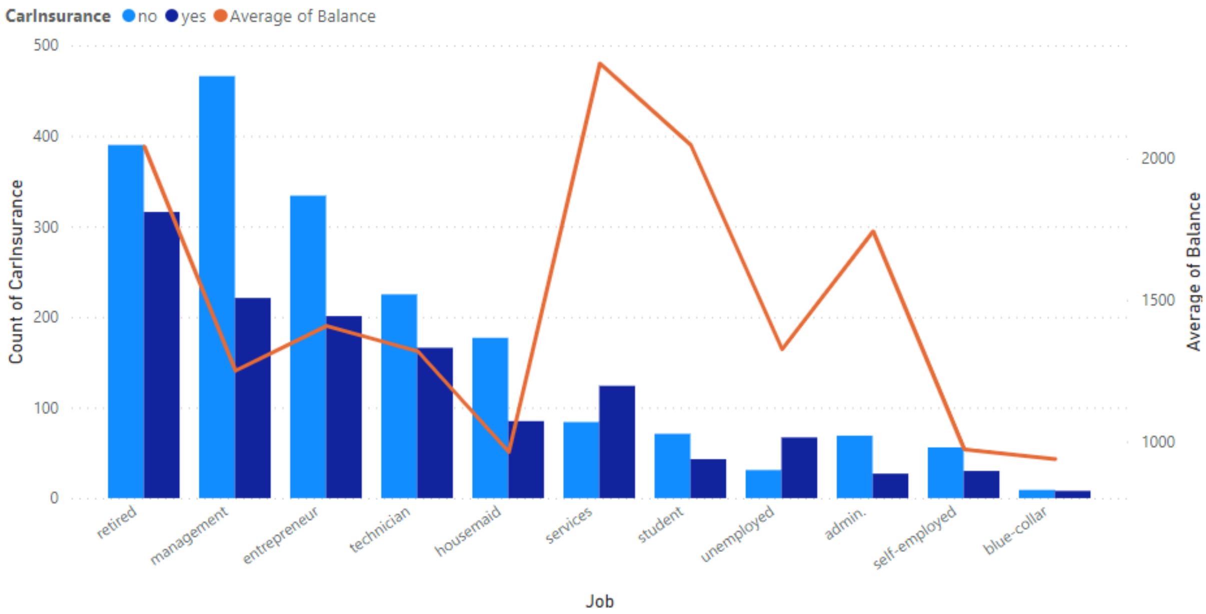
Since the NoOfContacts depicted an exponentially decreasing pattern, the natural logarithm of the variable was considered, and the below graph was generated.



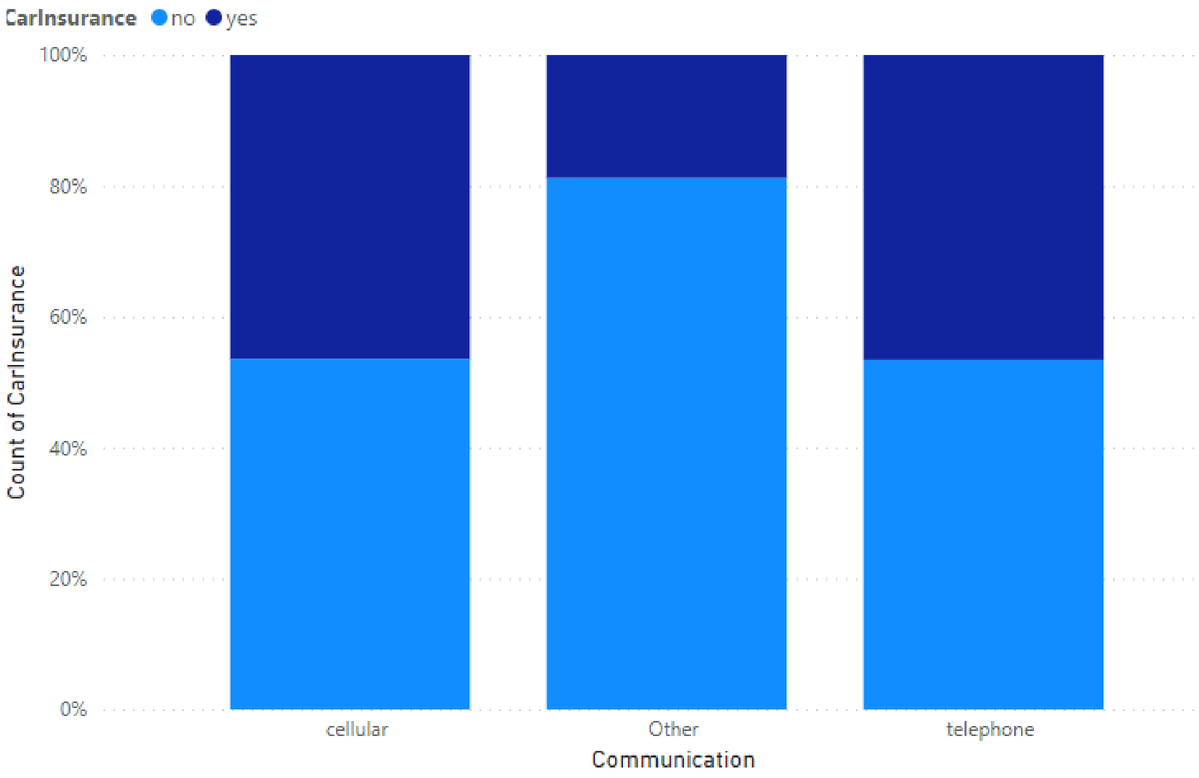
However, despite taking the log of the variable, the graph still exhibited an exponentially decreasing pattern.



The Inner layer refers to whether the household is insured by the client or not. The outer layer refers to whether Car Insurance was purchased or not. It can be observed that, people who don't insure their households are more likely to purchase car insurance than the people who insure their households.

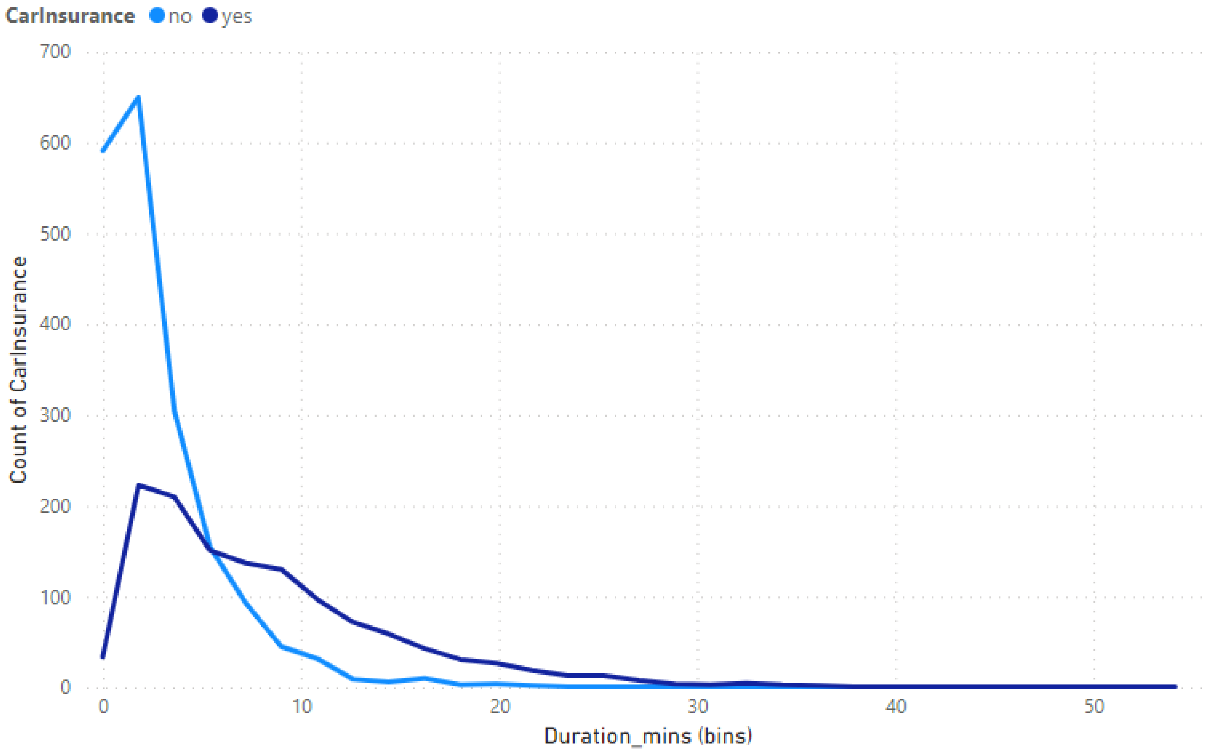


Clients with jobs in the service sector have the highest average yearly balance, and it can be observed that a higher proportion of them have agreed to purchase car insurance. It is interesting to note that students have a high average balance, which could be an indication of a possible selection bias in the dataset. That is, the company might have contacted more students with higher yearly balance. Additionally, it can be observed that clients who were unemployed were more likely to purchase car insurance.



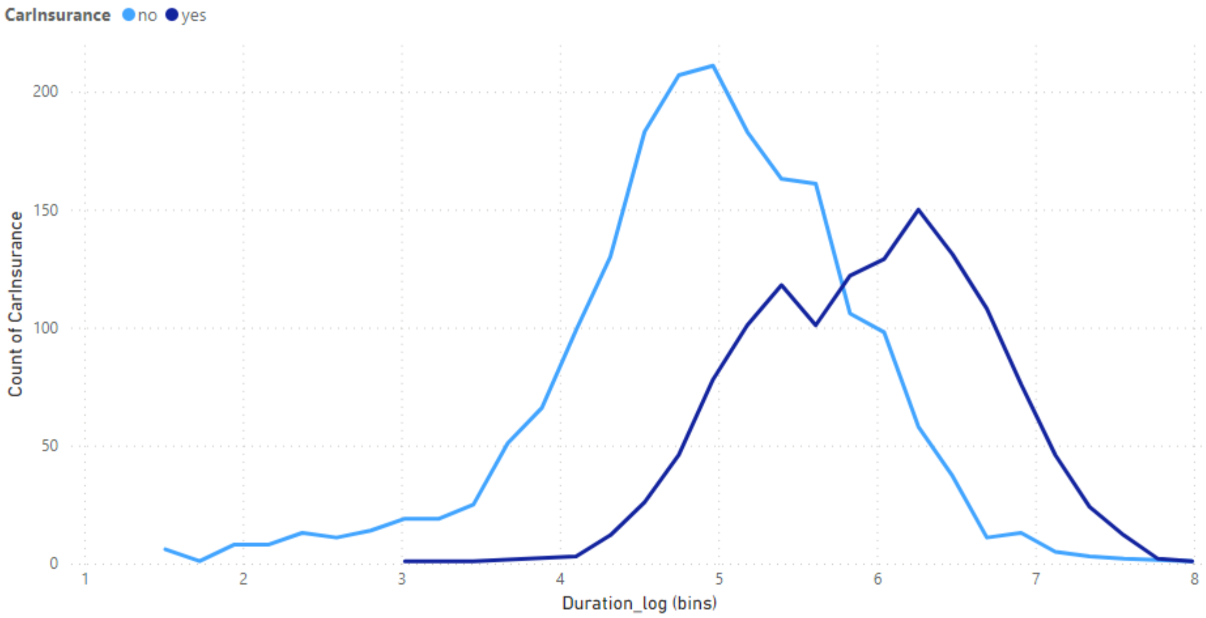
We can note that the ratio between the clients who purchased car insurance and those who didn't purchase it, is very low where the communication medium is not either cellular or telephone. However, this ratio is nearly one where the communication medium is a telephone.

Since the start time and end time of the last call in the campaign is given, a synthetic feature named 'Duration' was created to capture the call duration. The variation of this call duration with the count of car insurance purchases is given below.

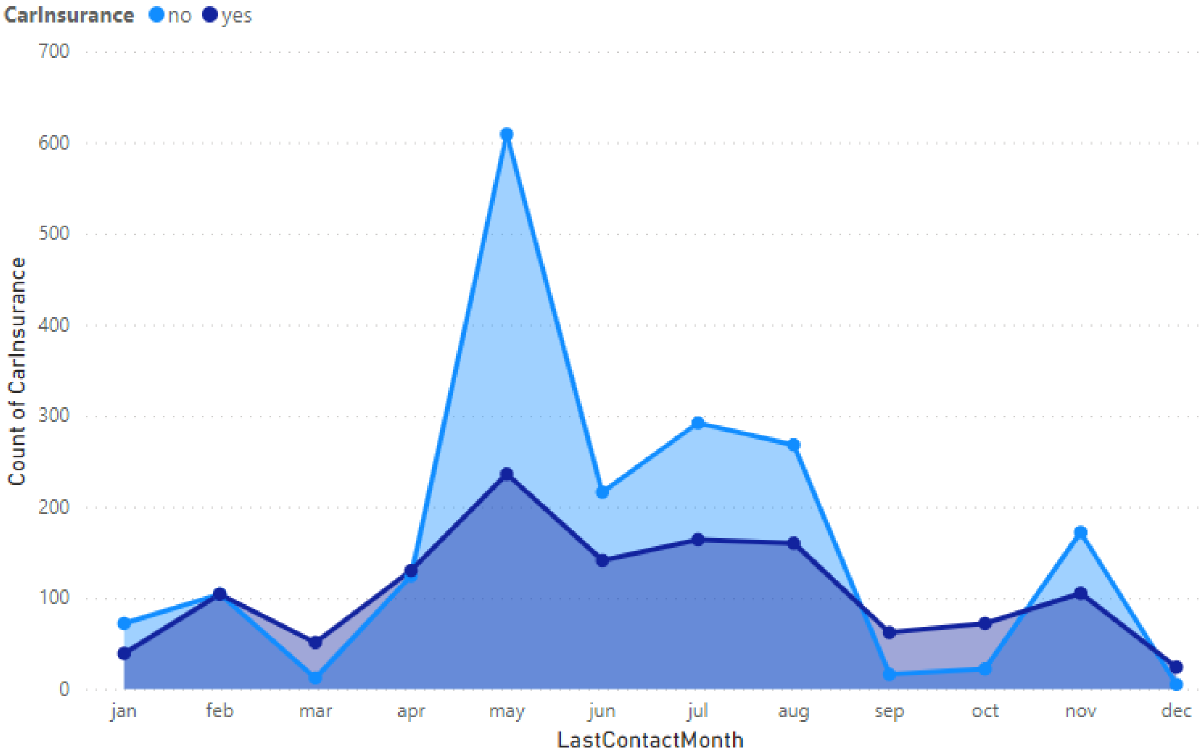


Mostly the count of car insurances ended as no when the duration is too low. But between the 5.41 mins to 28.87 mins the count of car insurances resulted as yes is larger than car insurances resulted as no which can be identified as optimum time duration to success car insurances.

Since the graph showed an exponential decay, the log transform of the values were obtained and the graph was generated.

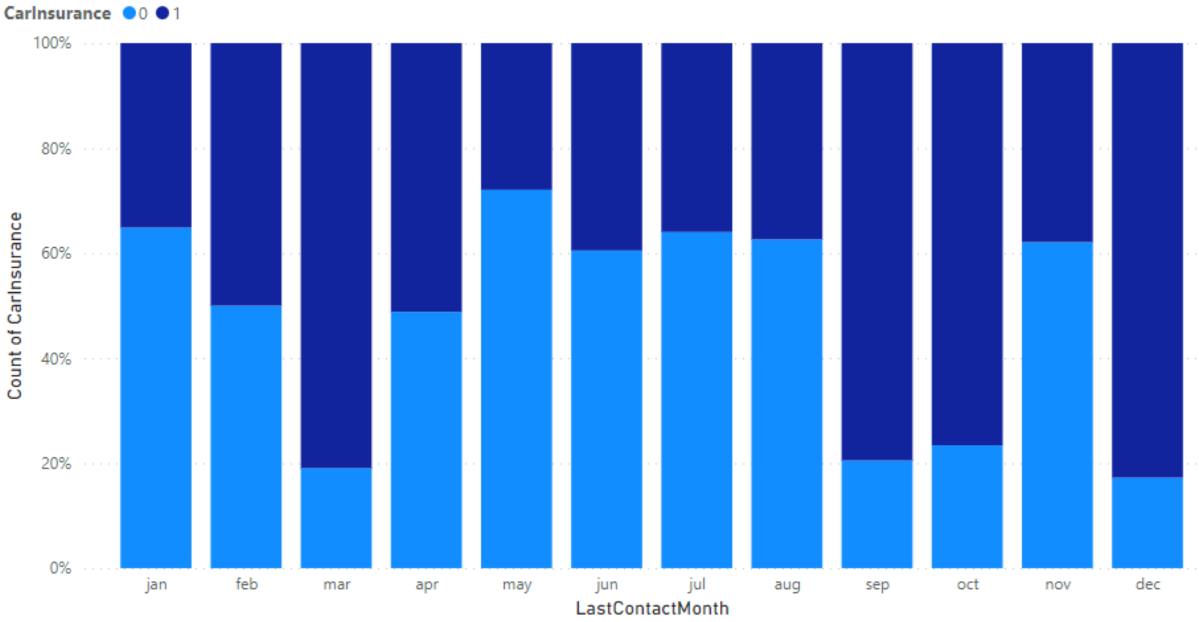


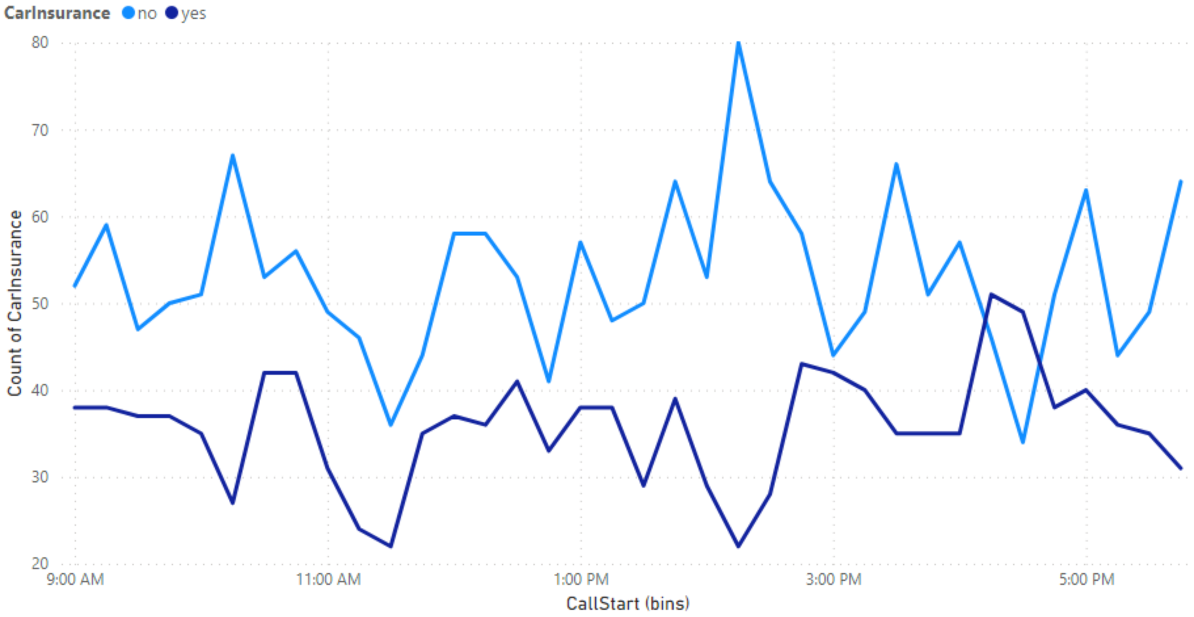
The log transformed values indicate that longer durations are relatively more likely to receive a positive response.



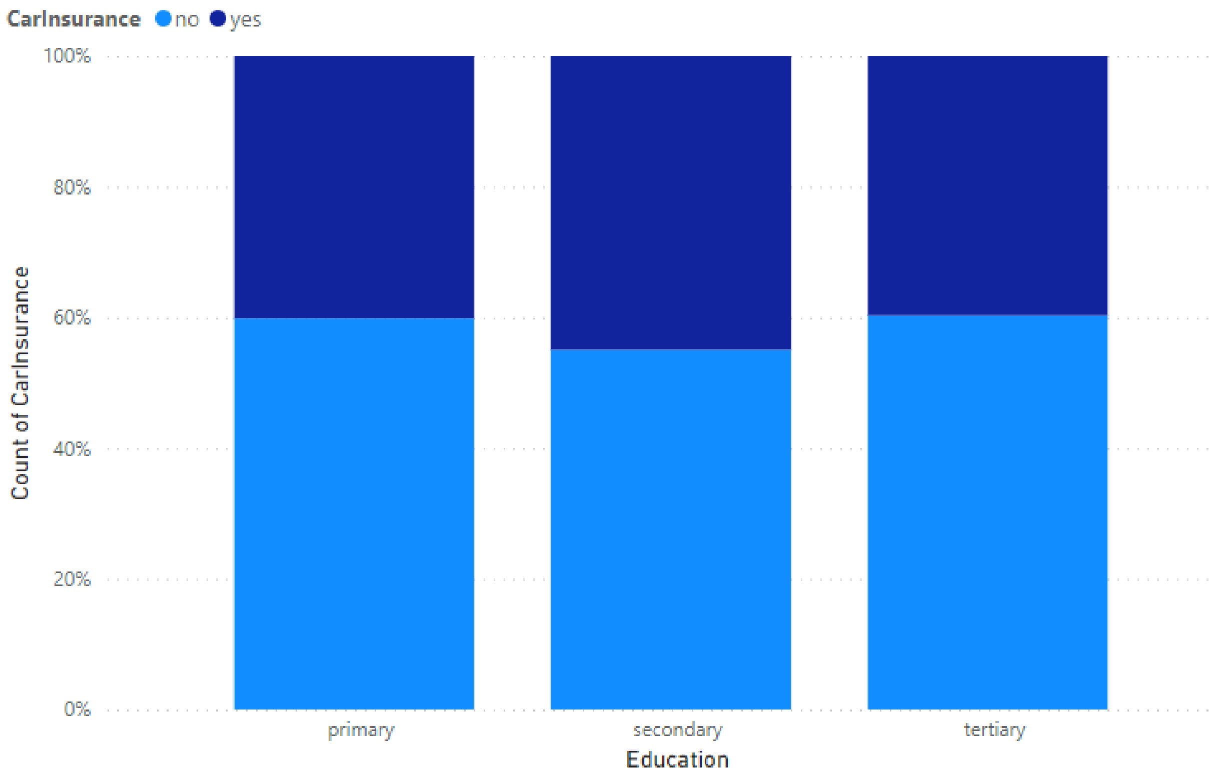
In October, September, March and December months, there can be identified more success results than the other months. The count of both outputs significantly differ in other months except April and February months.

This difference can be significantly observed when the values are normalised against the total number of calls made for the month. However it can be misleading since it doesn’t indicate that the total count of the number of calls made is different each month.



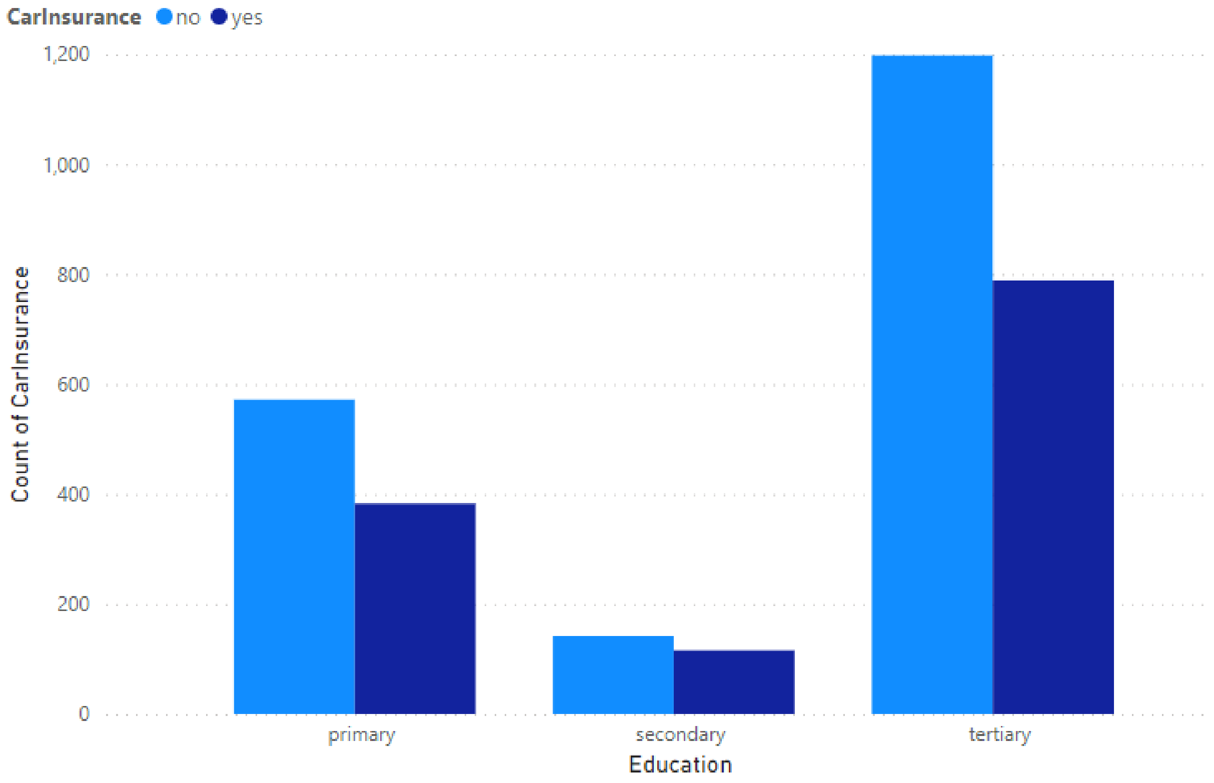


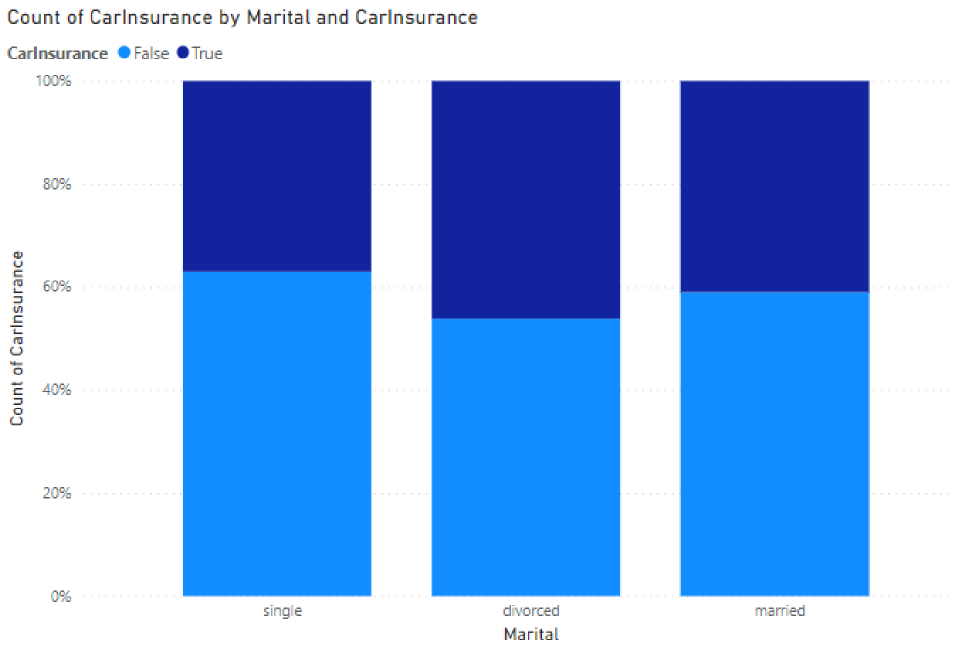
Between 4.00 pm and 4.45pm the car insurances resulted as success are higher than which were not success. Within 10.00 am to 10.30 am and 2.00 pm to 2.30 pm most of the results are ended as no comparison to success results.



The graph represents the count of car insurance results based on the education level of customers. The bar chart shows the same distribution for all levels of education. There is no significant insight given by education level.

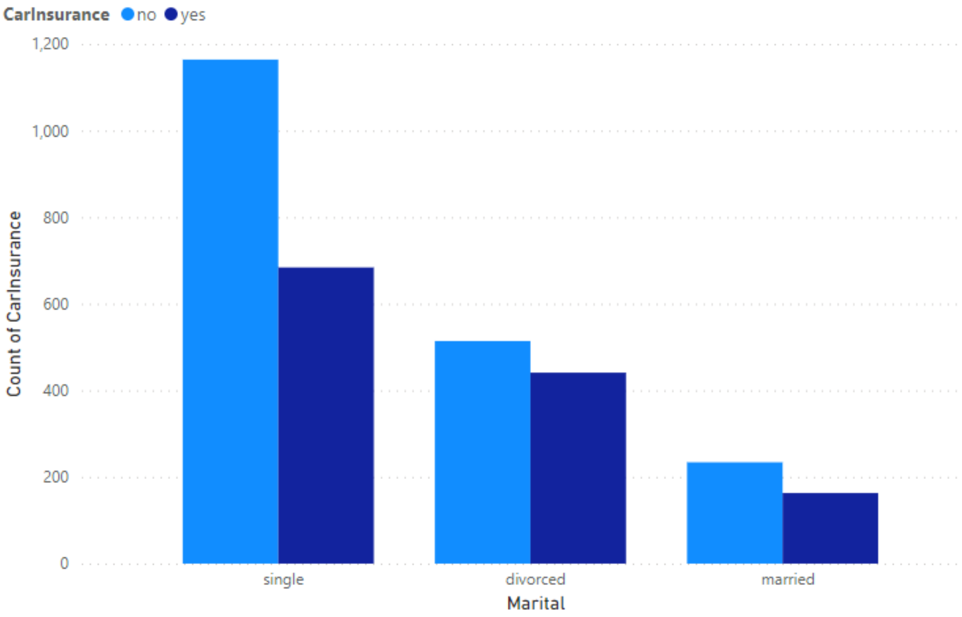
However, the graph containing normalised data doesn't take into account the differences between the total number of people contacted per category. Thus, the comparison is shown in the graph below.





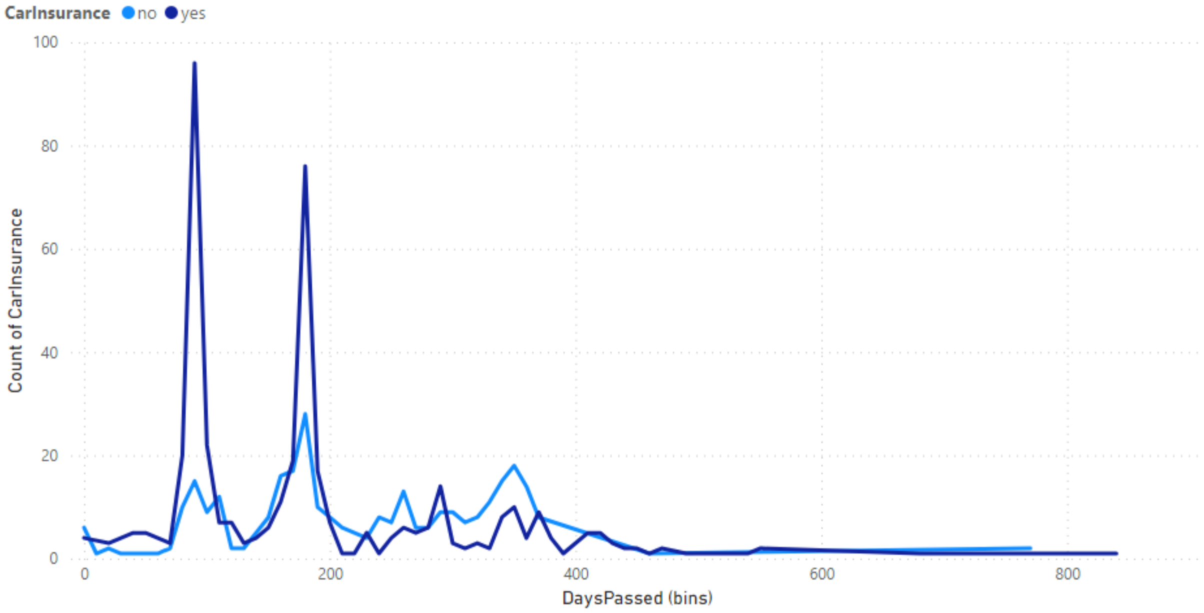
The graph represents the count of car insurance results based on marital status of customers. The bar chart shows the same distribution for all marital statuses. There is no significance result given by marital level.

However, the graph containing normalised data doesn't take into account the differences between the total number of people contacted per category. Thus, the comparison is shown in the graph below.

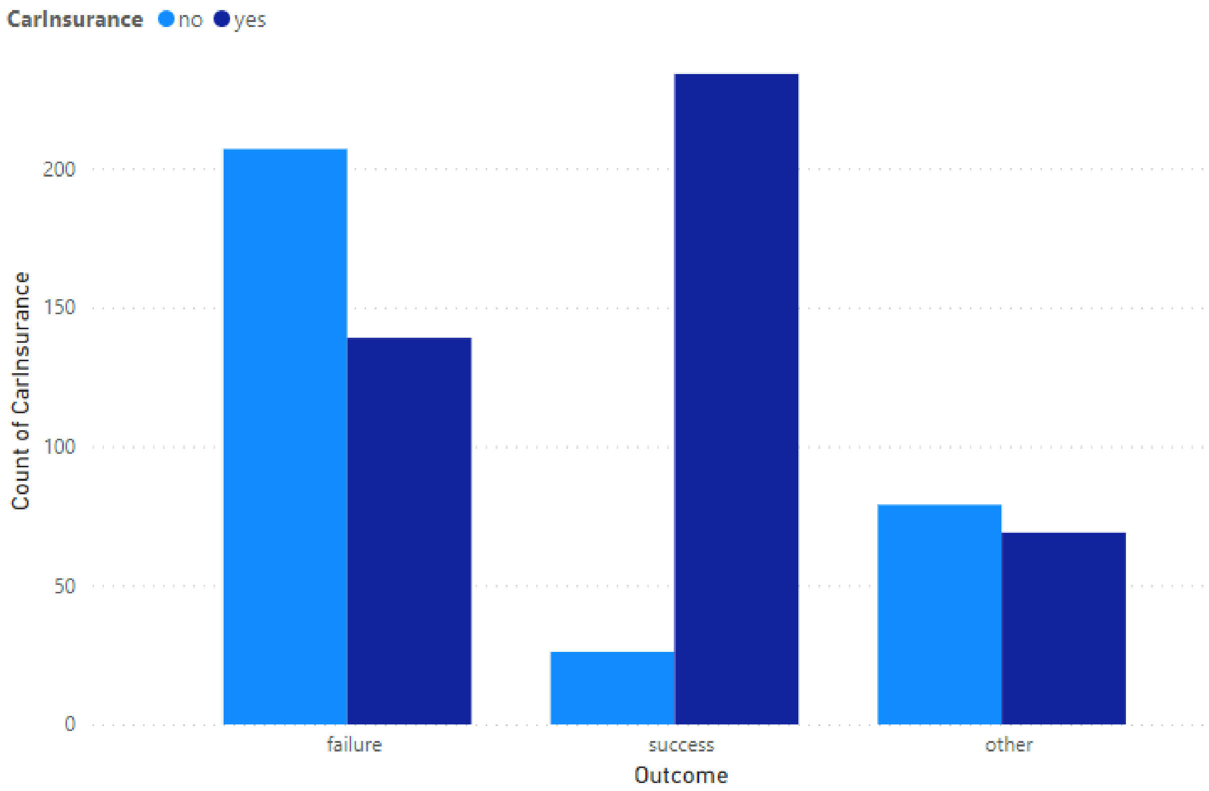


Previously contacted customers -

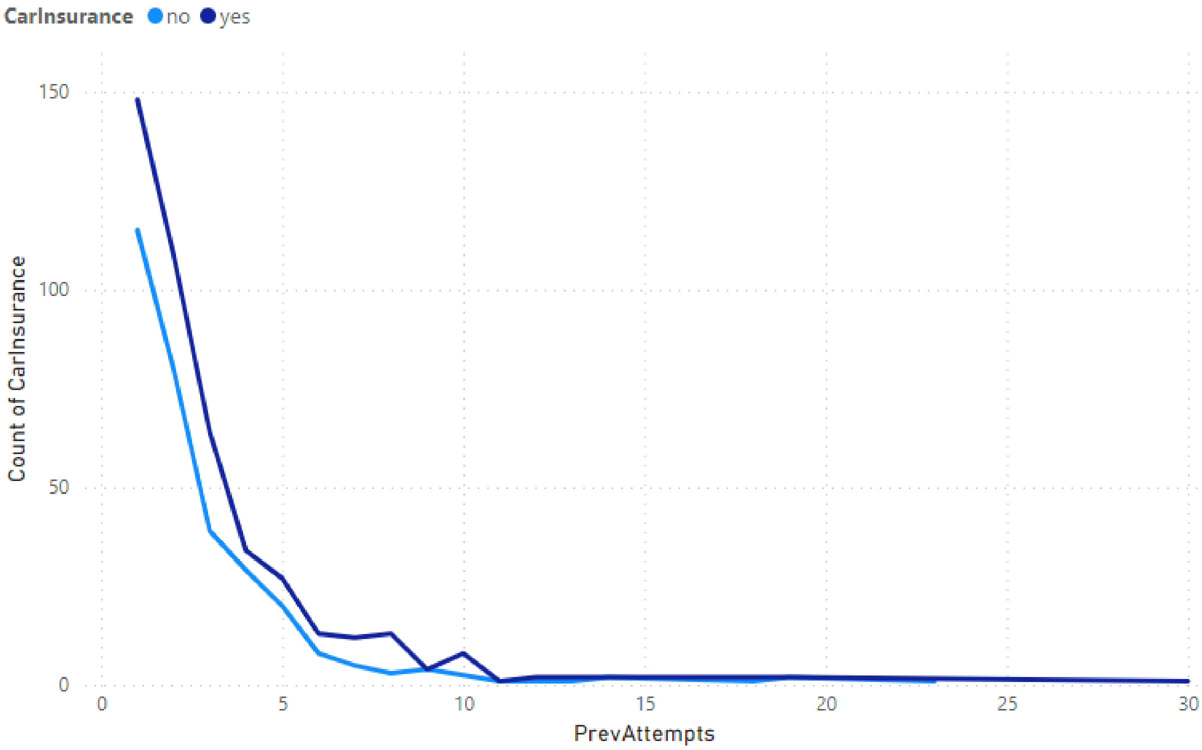
Since some of the customers have already been contacted before, separate visualisations were derived to observe trends in those customers purchasing car insurance.



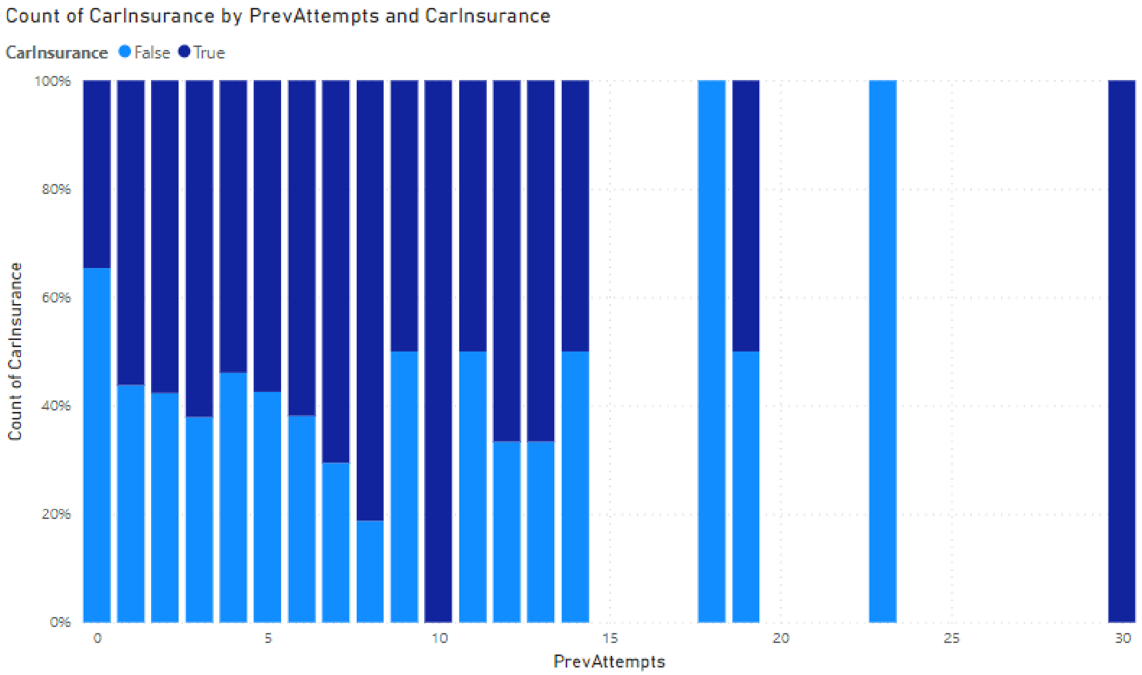
No. of days passed refers to the number of days that passed by after the client was last contacted from a previous campaign. It is evident that more people are likely to purchase car insurance if they were contacted around 90 or 180 days after the last contact of the previous campaign. However, between 100-160 days, there is a decrease in the number of positive responses.



It is clear that those who have purchased car insurance before through a previous campaign are significantly more likely to purchase car insurance through this campaign as well.



The data was then normalized and plotted, as given below.



Prev attempts represent the number of contacts performed before this campaign. There seems to be no particularly interesting trend, since it is to be expected that the total number of calls with a higher number of previous attempts would be relatively less.

Predictive Analysis

Since one of the main objectives of the webapp is to predict whether a given customer will purchase or not, models were considered for binary classification. Accuracy was considered as the primary evaluation metric.

All the variables were first converted to numericals in order to input to the machine learning model. ‘Yes’ and ‘No’ variables were encoded as 1 and 0 respectively. The CallStart attribute was broken down to CallStartHour, CallStartMinute and CallStartSecond. A synthetic attribute named ‘Duration’ was created using the difference between the Call start time and the Call end time. The CallEnd feature was dropped.

A basic logistic regression model produced a 5-fold cross validated accuracy of 0.76 with a 70:30 train:test split.

K Nearest Neighbours (KNN) model produced a 5-fold cross validated accuracy of 0.76 with the same train:test split. The features containing numerical variables were normalised and standardised and the model was tested again. KNN with normalisation produced an accuracy of 0.689 and KNN with standardisation produced an accuracy of 0.734.

Support Vector Machines (SVM) are commonly used for 2 class classification problems with high dimensionality. A SVM model with parameters C=0.1, gamma=1, kernel='linear' produced an accuracy of 0.801 on the train:test split ratio mentioned above.

Since ensemble learning models minimise any error-causing factors affecting a single predictive model, a commonly used bagging technique, RandomForestClassifier was used. With the following parameters, it produced an accuracy of 0.846. The optimum parameters were chosen using RandomSearch.

Parameters: bootstrap=False, max_depth=90, max_features='sqrt', min_samples_split=5, n_estimators=800

Other evaluation metrics were also used to get a better idea of the model performance.

- F1 score - 0.81
- Recall - 0.82
- Precision - 0.81

Since the dataset had a high number of features, Principal Component Analysis (PCA) was used as a technique of dimensionality reduction and the model was trained again and the accuracy obtained was 0.6853.

Since the target class showed a slight imbalance, upsampling was used to create a few artificial data points of the minority class. When the new upsampled data was used to train a RandomForestClassifier model with the following parameters, a 5-fold cross validated accuracy of 0.919747.

Parameters (up to now): default parameters.

Catboost, a popular gradient boosting algorithm, was also used, and produced an accuracy of 0.84.

Parameters (up to now): depth=6, iterations=100 and learning rate=0.04.

Other evaluation metrics were also used to get a better idea of the model performance.

- F1 score - 0.84
- Recall - 0.84
- Precision - 0.84

As observed in the exploratory analysis section, it was observed that there was a significant correlation between certain features. This can be attributed to customers contacted through a previous campaign having similar values for certain attributes. Since certain machine learning models may be sensitive to highly correlated data, an attempt was made to handle these highly correlated features. Thus, the dataset was broken down into 2 parts. Customers contacted through a previous campaign were included in one dataset. Customers not contacted through a previous campaign were included in the other dataset, and features containing information about previous campaigns were removed since those features would not affect these customers. Various models were trained on these datasets, and the best performance was exhibited as follows.

Dataset	Dataset dimensions	Model	Accuracy	Other evaluation metrics

Contacted through a previous campaign	774 entries
	17 predictor variables
		RandomForestClassifier (max_depth=60, max_features='sqrt', min_samples_leaf=2, n_estimators=1000)	0.8222	Precision = 0.83
				Recall = 0.82
				F1 Score = 0.82

Not contacted through a previous campaign	2446 rows
	14 predictor variables
		RandomForestClassifier(bootstrap=False, max_depth=50, min_samples_leaf=4, min_samples_split=10, n_estimators=1000)	0.8475	Precision = 0.84
				Recall = 0.83
				F1 Score = 0.83

It can be observed that there was no significant improvement in the model performance by splitting the dataset.

Conclusion

The main objective of this project is to provide business insights to speedstar vehicle insurance providers to improve their telemarketing campaign. Python libraries and Power BI were used to analyse the data and generate insights from features that affect customers' purchase of car insurance. Scikit learn library was used to create machine learning models that were trained on the given dataset in order to accurately predict whether a given customer would purchase car insurance or not. Thus, the analysis from the given data can be used to create a web application with an interactive dashboard and a powerful predictor tool, for the business to derive valuable insights.