**Technical Documentation**

**CAPSTONE PROJECT**

**ON**

**Health Insurance Cross Sell Prediction**

**Abstract:**

**Problem Statement**

Our client is an insurance company that has provided Health Insurance to its customers now they need your help in building a model to predict whether the policyholders (customers) from past year will also be interested in Vehicle Insurance provided by the company.

An insurance policy is an arrangement by which a company undertakes to provide a guarantee of compensation for specified loss, damage, illness, or death in return for the payment of a specified premium. A premium is a sum of money that the customer needs to pay regularly to an insurance company for this guarantee.

For example, you may pay a premium of Rs. 5000 each year for a health insurance cover of Rs. 200,000/- so that if, God forbid, you fall ill and need to be hospitalised in that year, the insurance provider company will bear the cost of hospitalisation etc. for upto Rs. 200,000. Now if you are wondering how can company bear such high hospitalisation cost when it charges a premium of only Rs. 5000/-, that is where the concept of probabilities comes in picture. For example, like you, there may be 100 customers who would be paying a premium of Rs. 5000 every year, but  only a few of them (say 2- 3) would get hospitalised that year and not everyone. This way everyone shares the risk of everyone else.

Just like medical insurance, there is vehicle insurance where every year customer needs to pay a premium of certain amount to insurance provider company so that in case of unfortunate accident by the vehicle, the insurance provider company will provide a compensation (called ‘sum assured’) to the customer.

Building a model to predict whether a customer would be interested in Vehicle Insurance is extremely helpful for the company because it can then accordingly plan its communication strategy to reach out to those customers and optimise its business model and revenue.

Now, in order to predict, whether the customer would be interested in Vehicle insurance, you have information about demographics (gender, age, region code type), Vehicles (Vehicle Age, Damage), Policy (Premium, sourcing channel) etc.

**Attribute Information:**

1. id: Unique ID for the customer

2. Gender: Gender of the customer

3. Age:  Age of the customer

4. Driving\_License 0: Customer does not have DL, 1: Customer already has DL

5. Region\_Code: Unique code for the region of the customer

6. Previously\_Insured: 1: Customer already has Vehicle Insurance, 0: Customer doesn't have Vehicle Insurance

7. Vehicle Age:  Age of the Vehicle

8. Vehicle\_Damage :1: Customer got his/her vehicle damaged in the past. 0: Customer didn't get his/her vehicle damaged in the past.

9. Annual\_Premium: The amount customer needs to pay as premium in the year

10. PolicySalesChannel:  Anonymized Code for the channel of outreaching to the customer ie. Different Agents, Over Mail, Over Phone, In Person, etc.

11. Vintage: Number of Days, Customer has been associated with the company

12. Response:  1: Customer is interested, 0: Customer is not interested

**Insights from Problem:**

**What is an insurance firm?**

* If a loss occurred a guarantee of compensation for specified loss, damage, illness, or death in return for the payment of a specified premium.

**What is the probability of buying insurance?**

* In the insurance industry, it refers to a situation in which people only buy insurance when they expect high risks. Buying insurance is not appropriate for all levels and types of risks. In many cases, people are better off taking actions to avoid risk, retain (accept) risk or reduce risk. Buying insurance makes the most sense when the potential loss is great and there is a significant probability of loss.

**How many people are knowledgeable about insurance policy and how many of them claim insurance?**

* Let's say about four in 10 men describe themselves as being very knowledgeable about life insurance. As in the problem statement, about 2 or 3 get hospitalised out of 100, which means 2 to 3 percent claim the insurance. This way everyone shares the risk of everyone else.

So we need to build a model to predict whether a customer would be interested in Vehicle Insurance is extremely helpful for the company because it can then accordingly plan its communication strategy to reach out to those customers and optimise business model, revenue. Now, we need to predict whether the customer would be interested in Vehicle insurance or not.

**Algorithm followed:**

* importing the required libraries and packages
* Using Colab, mount the device and open the.csv [Comma Separated file].
* Applying the concept of Data Wrangling and Data Visualization so we can analyse the dataset and retrieve necessary information.

***Step 1:***

* importing the required libraries:
* The most widely used Python package is NumPy [Numerical Python]. efficiently utilised for array calculation problems.
* Analysing tabular data is an excellent use of the Pandas library. It can be used for exploratory data analysis, including cleaning, wrangling, and manipulating data.
* With the use of the Matplotlib package, we can better visualise our tabular data in the form of [Pie Chart, Bar Graph, Line Graph, Histogram, etc.] which aids in doing the necessary analysis.

***Step 2:***

* Analysing the information sheet to identify any NULL/NAN/Missing values. As these numbers affect the correctness of the results, we wish to get rid of them. We frequently treat outliers as well; in essence, these are data points that are greatly dispersed from other data points and that would skew visual outcomes.
* Now that our dataset is clear of all ambiguity, we will go forward with the process of exploratory data analysis (EDA), which refers to the act of analysing data in order to produce visual results.

**Terminology that is frequently used:**

***Data Wrangling***

* Data wrangling is the process of organising and cleaning up large and disorganised data sets so they are simpler to access and analyse. the process of transforming and mapping data from one "raw" data type into another with the goal of improving its suitability and value for a range of downstream uses, including analytics.
* Functionalities performed under it:

Data exploration, handling NAN and missing values, removing duplicates, and data filtering.

**Data Visualization:**

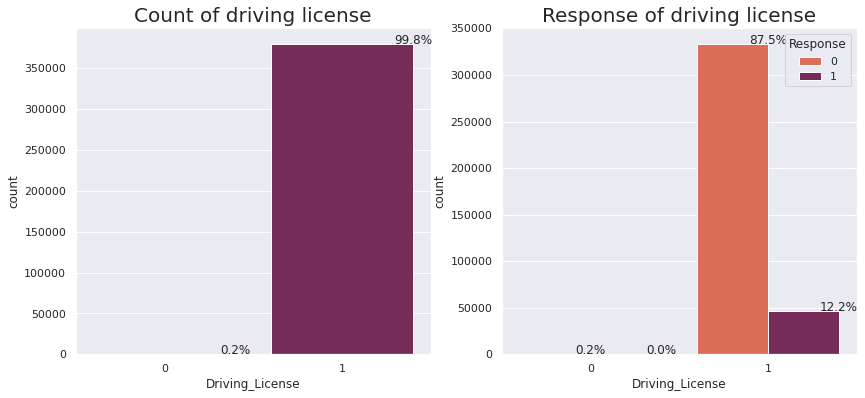
It involves converting tabular data into graphic images so that users may quickly and easily understand it.

* It is employed because visualising graphical data yields considerably better outcomes than doing so for tabular data. The verbal and statistical descriptions of the data set are closely integrated during visualisation. This identifies problem areas and directs attention to them.

**Exploratory Data Analysis**

* Exploratory data analysis is a crucial procedure that entails performing early investigations on data in order to find patterns, identify anomalies, test hypotheses, and validate assumptions with the aid of summary statistics and graphical representations.

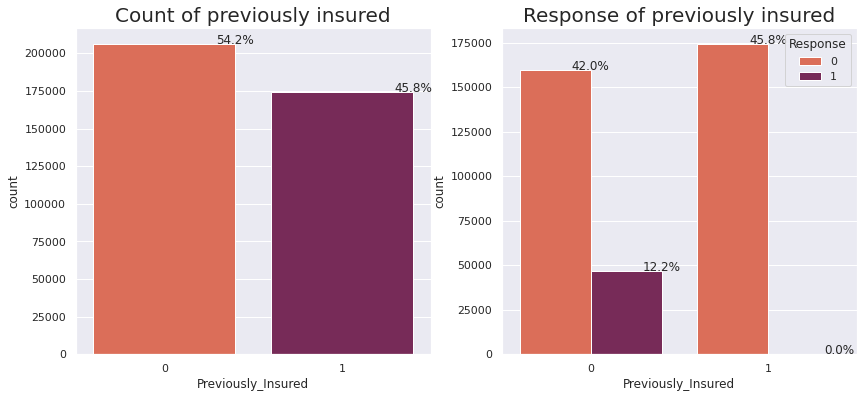
# **Driving Licence**



* 99.8% of customers have DL, whereas 0.2% do not have DL.
* Only a small percentage of people who have a DL (12.2%) are interested in buying insurance.

From this, we can conclude that almost all the people who own vehicles have DL because it's mandatory when you have a bike, and only a small percentage of people are interested in buying vehicle insurance. The possible reason might be that people who own the bike may already have vehicle insurance or insurance might be expired.

# **Previously Insured**



* 45.8% people are insured previously, in that 12.2% poeple interested to buy the vehicle insurance again(insurance might be expired or should be renewed), Which means people are aware of insurance policy and ready to pay a premium amount, for better off taking actions to avoid certain risks or reduce risk.
* So buying insurance makes the most sense when the potential loss is great and there is a significant probability of loss.

# 

# 

# 

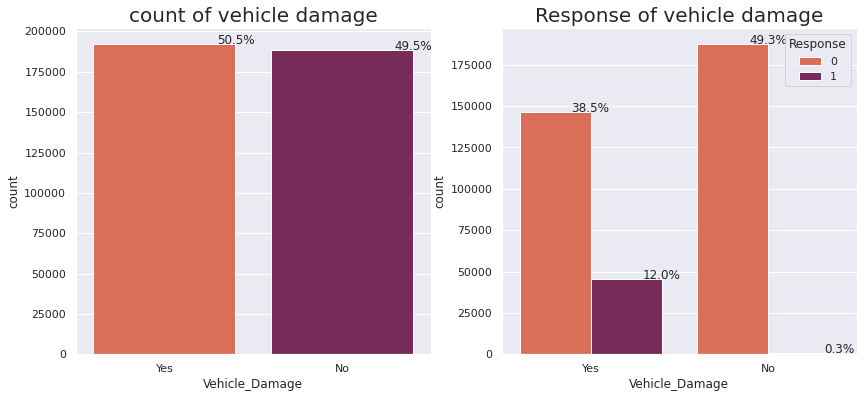
# **Vehicle Age**

# 

* Around 4.2% of vehicles are more than two years old, 52.6% are between one and two years old, and 43.2% are under one year old.
* 1.2% are interested in purchasing vehicle insurance for vehicles older than 2 years, 9.1% are interested in purchasing insurance for vehicles between 1 and 2 years old, and 1.9% are interested in purchasing insurance for vehicles older than 1 year.

As vehicle age increases most of the people are aware of insurance and interested to buy the insurance for reducing the risk

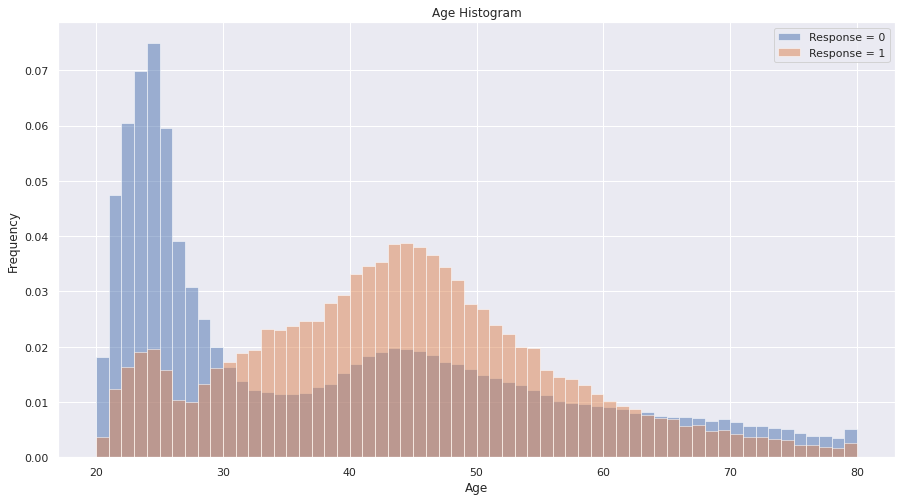
# **Vehicle Damage**



* 50.5% of the vehicles have past damage
* 12.0% of people who have had a damaged vehicle in the past want to acquire vehicle insurance

Here, 50 percent of vehicles are damaged and 50 percent are not damaged, which means people with damaged vehicles (12%) are interested in buying insurance and are aware of vehicle insurance policies and its benefits, while the rest of the people might already have purchased insurance and do not need to purchase again.

**Age**



* The dataset has more individuals with an age of 24.
* 40 to 60-year-olds had a higher likelihood of purchasing vehicle insurance.
* From the above boxplot we can see that there is no outlier in the dataset.

**Tools using which we create graphical representation/Pictorial data**

***Boxplot:***

* A box plot is a visual representation of the location, dispersion, and skewness groups of quartiles of numerical data. The image of an "outlier" is the most significant one it offers.

***Barplot :***

* A bar chart, often known as a bar graph, is a diagram that displays categorical data as rectangular bars with heights or lengths proportional to the values they stand for. You can plot the bars either vertically or horizontally. A vertical bar chart may also be referred to as a column chart.

***Countplot:***

* Utilising bars, a method is utilised to display the numbers of observations in each category bin.

***Heatmaps:***

* A heatmap is a graphical display of data in which values are represented by colours. A heatmap that displays a 2D correlation matrix between two discrete dimensions and uses coloured cells to represent data from typically a monochromatic scale is called a correlation heatmap. Heatmaps make the association between one feature (variable) and every other feature incredibly simple to understand (variable).

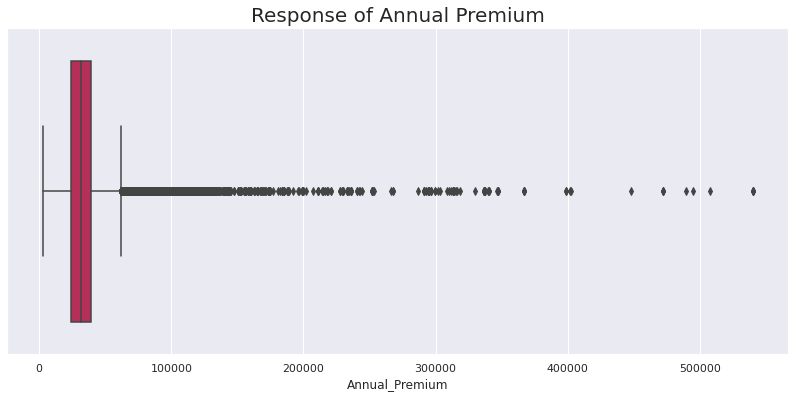
**Feature Engineering:**

* **Null, Missing and Duplicate Values Treatment**

It is an important aspect of Data Cleaning because there can be some null, missing and duplicate values in our dataset. But our dataset doesn't contain null or missing values which might tend to disturb our accuracy, if it has null or missing values then we have to drop them at the beginning of our project in order to get a better result.

* **Outliers handling**

Checking outliers in the dataset because Outliers is also something that we should be aware of. Why? Because outliers can markedly affect our models and can be a valuable source of information, providing us insights about specific behaviours. Outliers is a complex subject and it deserves more attention



We have a lot of outliers in our annual premium column, but Hence removed it.

**Feature Encoding**

We used One Hot and label Encoding to produce binary integers of 0 and 1 to encode our categorical features because categorical features that are in string format cannot be understood by the machine and needs to be converted to numerical format.

**Feature Selection:**

Feature selection is the process of reducing the number of input variables when developing a predictive model.It is desirable to reduce the number of input variables to both reduce the computational cost of modelling and, in some cases, to improve the performance of the model.

**Preparation for Model Making**

* **Imbalance technique-** One of the major issues when dealing with unbalanced datasets relates to the metrics used to evaluate their model. Using simpler metrics like accuracy score can be misleadingIn a dataset with highly unbalanced classes, the classifier will always “predict” the most common class without performing any analysis of the features and it will have a high accuracy rate, obviously not the correct one.

We have tried with undersampling, oversampling, and SMOTE. Of these, oversampling gives the best result.

* **Splitting**

train test split is a model validation procedure that allows you to simulate how a model would perform on new/unseen data.

In this particular step we splitted our data to train and test data with 30% test data.

* **Standardisation of features**

Our main motive through this step was to scale our data into a uniform format that would allow us to utilize the data in a better way while performing fitting and applying different algorithms to it.

The basic goal was to enforce a level of consistency or uniformity to certain practices or operations within the selected environment.

We used two scaler Min-max and Standard for our data but Standard scaler gives good results so we proceed further with it.

**Result of all the Models:**

Here we used different types of classification algorithms to compare which one is giving best result and accuracy. Comparing the result of all the model Decision Tree, Random Forest gives the best Result.

**Conclusion:**

* The gender variable in the dataset is spread nearly evenly. The male category is marginally larger than the female category, and the likelihood of purchasing insurance is also slightly higher. The response rate of those who are not interested in purchasing vehicle insurance is higher than that of those who are interested in buying vehicle insurance.Only 12.3% people are interested in buying vehicle insurance and 87.7% are not interested in buying vehicle insurance. So people who own a vehicle may already have vehicle insurance, or people might not be aware of insurance policy and pricing factors, which means the firm needs to come up with good marketing techniques and a pricing strategy to create awareness and offer an affordable price to the customers in order to reach out to more customers to generate more leads.
* The response rate of those who are not interested in purchasing vehicle insurance is higher than that of those who are not interested in buying vehicle insurance. 99.8% of customers have DL, whereas 0.2% do not have DL. Only a small percentage of people who have a DL (12.2%) are interested in purchasing vehicle insurance. So almost all the people who own vehicles have DL because it's mandatory when you have a bike, and only a small percentage of people are interested in buying vehicle insurance. The possible reason might be that people who own the bike may already have vehicle insurance or insurance might be expired.
* 45.8% of people are insured previously, in that 12.2% of people interested to buy the vehicle insurance again, Which means people are aware of insurance policy and ready to pay a premium amount, for better off taking actions to avoid certain risks or reduce risk. So buying insurance makes the most sense when the potential loss is great and there is a significant probability of loss.
* Around 4.2% of vehicles are more than two years old, 52.6% are between one and two years old, and 43.2% are under one year old. 1.2% are interested in purchasing vehicle insurance for vehicles older than 2 years, 9.1% are interested in purchasing insurance for vehicles between 1 and 2 years old, and 1.9% are interested in purchasing insurance for vehicles older than 1 year. As vehicle age increases most of the people are aware of insurance and interested to buy the insurance for reducing the risk.
* The dataset has more individuals with an age of 24. 40 to 60-year-olds had a higher likelihood of purchasing vehicle insurance. 9.3% of people in their middle age are interested in purchasing insurance. Almost 47% of middle-aged individuals have a driver's licence. About 21.9% of people in their teens have health insurance. Around 21.9% of persons in their teens have had insurance previously. So most teenagers have insurance and are aware of their policy. So the target audience might be middle-aged people and teenagers to generate more leads for insurance companies.
* By using the interquartile range, we eliminated outliers and dealt with null data. We split the dataset into train and test splits after feature encoding three columns. Further, we applied 6 machine learning algorithms to see which customers might be interested in purchasing vehicle insurance and we also used hyperparameter tuning for three models to discover which model gives the best results. Vehicle damage and annual premium are the two most significant features seen in decision trees, while vintage and annual premium are seen in random forests. With 93% and 92% ROC AUC scores, Decision Tree and Random Forest outperform all other models.

***Project by***

Ravi Kumar