[**Health Insurance Cross Sell Prediction**](https://github.com/bysubanji/Health_insurance_cross_sell_prediction)  
**Harshjyot Singh Chawla**

**Ajinkya Dakhale**

**Suvir Kapse**

**Data science trainees,**

**AlmaBetter, Bangalore**

**Abstract:**

Cross-selling is the practice of offering a consumer something in addition to the good or service they initially sought for. Before releasing their new products to the general public, the majority of businesses in the globe cross-sell them to their current clientele. By identifying the correct kind of clients to market the product to and by streamlining their communication strategy, this aids the business in cutting time and costs.

Any insurance agency can increase revenue through cross-selling without having to start from zero. Similar to medical insurance, car insurance requires the client to pay an annual premium to the insurance provider so that, in the event of an unfortunate accident involving the vehicle, the insurance provider will provide the customer with compensation (referred to as "sum assured").

In accordance with the provisions of an insurance policy, a corporation undertakes to give a guarantee of reimbursement for a particular loss, damage, illness, or death in exchange for the payment of a particular premium. A number of elements are crucial in attracting customers for any insurance coverage. Age, gender, area code, vehicle damage, vehicle age, annual

premium, and insurance source channel are among the demographic information shown here.

Any dataset from an insurance plan can be utilised with the model to forecast cross-selling. Our data set is based on a database of health insurance customers. This experiment can assist in determining the variables that may influence the cross-selling of insurance plans to current clients. After Exploratory Data Analysis, we tried to build multiple machine learning algorithms which contributed in predicting whether a customer would be interested in Vehicle Insurance.

***Keywords: machine learning, Supervised machine learning.Cross-selling, Predictive Model building, Decision Tree, Logistic Regression, Random Forest, XGB Classifier***

**Introduction:**  
A new marketing technique called cross-selling makes use of data analysis to identify potential clients with different wants who can be satisfied by the sales of numerous connected services or goods. The "suggested items" on Amazon and the "similar searches" on Google are both examples of this sales strategy. The value of each customer can be maximised by cross-selling, but there are other advantages for your business as well. When done correctly, cross-selling is also a favourite among customers since it introduces them to new products and services that can improve their lives.

When a new product is introduced, a business always seeks to collect feedback from its current consumers about the product and also tries to sell them the new product. Additionally, in order to optimise the communication strategy plan and cut time and costs when cross-selling a product, the corporation needs to know which customers are likely to buy the new product.

Advantages of cross selling include:

### Helps Customers Feel Understood

### Builds Loyalty

### Increases Earnings

### Greater Convenience

Without having to spend money on lead generation, you can increase your revenue by cross-selling insurance. By keeping abreast of developments in your clients' life that might need more recent or extensive coverage, you can also strengthen your relationship with them. Client retention may increase as a result. The sum of money that a customer is required to constantly pay to an insurance provider in exchange for this assurance is referred to as a premium. Customers of vehicle insurance must pay an annual premium to the insurance provider firm in order to be rewarded (referred to as the "sum assured") in the sad event that their vehicle is to blame for an accident. This process is similar to that of paying for medical insurance.

This dataset includes data from an insurance company's health insurance customers from the previous calendar year. The insurance provider has chosen to provide auto insurance as their new service. Now, the business needs a model to determine whether those policyholders will also be interested in the company's vehicle insurance. Following receipt of your model data, your business can use digital marketing and automation to expand its auto insurance agency through cross-selling and boost its revenue by funding a reproductive marketing campaign.

**Problem Statement:**

An insurance policy is a contract whereby a business agrees to guarantee financial compensation in the event of a specified loss, damage, disease, or death in exchange for the payment of a specified premium. The amount of profit that the consumer typically needs to give an insurance firm in exchange for this assurance is known as the premium. Building a model to forecast a customer's interest in Vehicle Insurance is very beneficial for the business because it allows it to design its communication strategy to reach out to those clients in the most effective way possible and maximise its business model and revenue.

Our client is an insurance provider that previously offered health insurance to its clients. Now, the company needs our assistance in developing a model to predict whether current policyholders from the previous year will also be interested in the vehicle insurance the provider offers.

The objective of this project is to develop a prediction model employing statistically significant variables from the provided data set using machine learning algorithms like Logistic Regression, XGB Classifier, and Decision Tree.

**Data information:**

Any good machine learning model must include data. No matter how effective your machine learning models are, you cannot produce a dependable high-performance model from the prediction model without sufficient rich data.

A dataset is provided that contains information about demographics (gender, age, region code type), Vehicles (Vehicle Age, Damage), Policy (Premium, sourcing channel) etc. in connection to an individual who is interested in vehicle insurance. We have 381109 data points available.

| **Feature Name** | **Type** | **Description** |
| --- | --- | --- |
| id | (continuous) | Unique identifier for the Customer. |
| Age | (continuous) | Age of the Customer. |
| Gender | (dichotomous) | Gender of the Customer |
| Driving\_License | (dichotomous) | 0 for customer not having DL, 1 for customer having DL. |
| Region\_Code | (nominal) | Unique code for the region of the customer. |
| Previously\_Insured | (dichotomous) | 0 for customer not having vehicle insurance, 1 for customer having vehicle insurance. |
| Vehicle\_Age | (nominal) | Age of the vehicle. |
| Vehicle\_Damage | (dichotomous) | Customer got his/her vehicle damaged in the past. 0: Customer didn't get his/her vehicle damaged in the past. |
| Annual\_Premium | (continuous) | The amount customer needs to pay as premium in the year. |
| Policy\_Sales\_Channel | (nominal) | Anonymized Code for the channel of outreaching to the customer i.e. Different Agents, Over Mail, Over Phone, In Person, etc. |
| Vintage | (continuous) | Number of Days, Customer has been associated with the company. |
| **Response** (Dependent Feature) | (dichotomous) | 1 for Customer is interested, 0 for Customer is not interested. |

**Steps involved:**

**Exploratory Data Analysis:**

Exploratory Data Analysis (EDA) is a method for examining datasets to highlight their key features, frequently using visual techniques. We can comprehend the data's statistical components, such as mean, mode, median, etc.

We can identify outliers, duplicates in the data collection, missing values, and null values using EDA. In the dataset, we can identify correlations between features. In EDA, we can apply different data visualisation techniques to data and summarize the data.

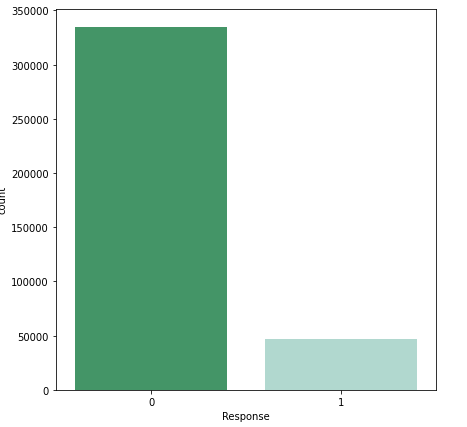
Two categories typically cross-classify exploratory data analysis. The first distinction is between graphical and non-graphical methods. Additionally, every technique is multivariate, univariate or bivariate analysis.

**We can infer from the data that there are 381109 rows and 12 columns.**

**UNIVARIATE ANALYSIS:**

Univariate analysis is the process of comparing and examining the dependency of a single predictor and a response variable. One feature is examined at a time in a univariate analysis. When we examine a feature alone, we disregard other feature in the dataset at time time and focus only on the distribution of its values.

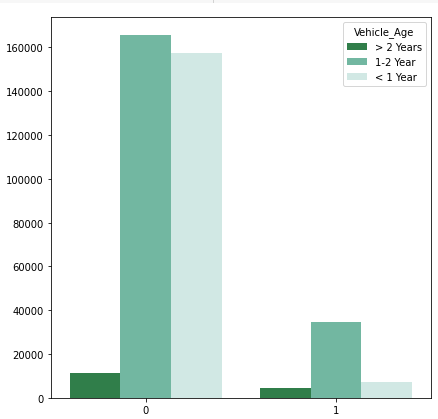
In our analysis we have examined each 11 independent variable and 1 dependent variable (Response) individually with the help of count plot, pie chart, distribution plot etc so that we can gather information about them.



**BIVARIATE ANALYSIS:**

Bivariate analysis is the process of analysing data by considering two variables or columns from a dataset. We have categorical and numerical data of two different types. We compare the Response, our dependent variable, to every independent variable. There are two different ways we can analyse it.

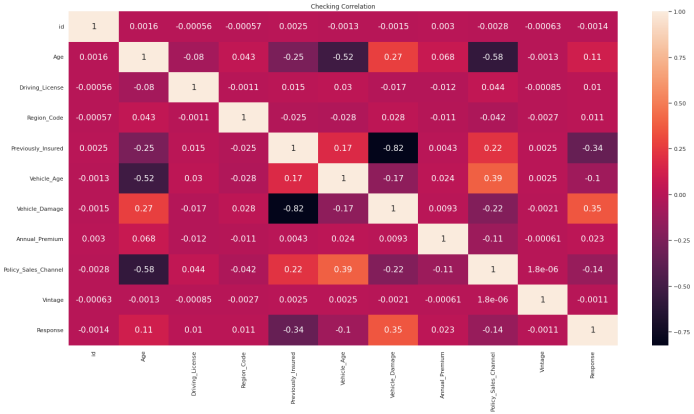
* Pie Chart
* Bar Plot



**MULTIVARIATE ANALYSIS:**

The analysis of three or more variables is done using multivariate analysis. As opposed to bivariate analysis, this enables us to examine correlations (i.e., how one variable changes in relation to another) and attempt to predict future behaviour more correctly.

The most common way of plotting multivariate data is to build a Heat map,



From heat map we can conclude that there is no such strong correlation among features.

### Feature Engineering:

### Feature engineering is the process of selecting, manipulating, and transforming raw data into features that may be used in supervised learning. To make machine learning effective on new tasks, it may be necessary to develop and train better features. A "feature," as you may know, is any quantifiable input that may be used in a predictive model. The practice of using statistical or machine learning approaches to convert raw observations into desired characteristics is known as feature engineering.

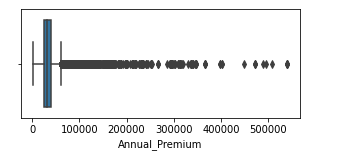
**Null, Missing and Duplicate Value Treatment:**

Given that our dataset may contain some null, missing, and duplicate values, this component of data cleaning is crucial. However, our dataset doesn't have any null or missing values that could potentially affect its accuracy; if it did, we would need to remove them at the beginning of the project in order to provide better results.

**Outliers handling:**

A data point in a data set that is far from all other observations is known as an outlier. A data point that is not distributed normally over the entire dataset.

Any machine learning project must complete this work because if not done correctly, a machine learning technique could produce subpar outcomes. We discovered Annual Premium feature during EDA that have excessive values that could result in inaccurate predictions.



We transformed Annual Premium to normal distribution with the help of power transformer.

**Encoding of categorical columns:**

Because categorical features that are in string format cannot be read by the machine and must be translated to numerical format, we used One Hot and label Encoding to produce binary numbers of 0 and 1 to encode our categorical features.

**Feature Selection:**

For feature selection we have used variance inflation factor (VIF) method.

In a regression model, the VIF calculates the degree and strength of correlation between the explanatory variables... The range for VIF values should be:

* For variables that can be used (0-10)
* For variable that can be considered to drop (above 15)

After performing VIF, we noticed that variable Driving\_License has a very high VIF value and the fact that people without driving license are few so we can drop the variable Driving\_License.

**Handling imbalance data:**

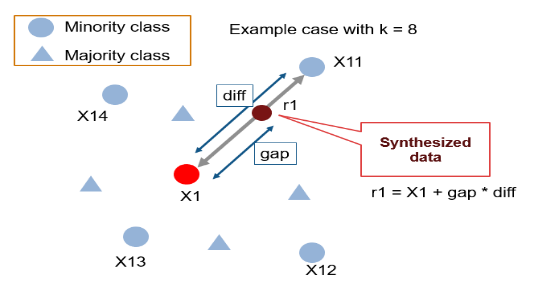
The dataset is unbalanced and gives a higher weight to classes that appear most frequently, a model could perform poorly. SMOTE is a technique that has been used to handle this kind of data set (Synthetic minority oversampling technique)

As many synthetic examples for the minority class as necessary can be made using this method.

**Working:**

* SMOTE finds the k closest minority class neighbours of the minority class instance it has chosen at random, instance a. Next, a line segment in the feature space is formed by joining a and b to form the synthetic instance by randomly selecting one of the k nearest neighbours, b. The two selected instances, a and b, are convexly combined to create the synthetic instances.

As many synthetic examples of the minority class as needed can be produced using this procedure. It suggests using SMOTE to oversample the minority class in order to balance the class distribution, followed by using random under-sampling to reduce the number of examples in the majority class



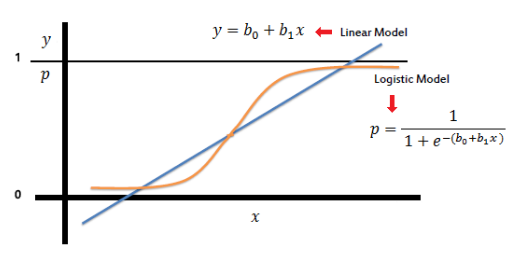
**Model implementation:**  
After we performed one hot encoding and label encoding on respective features and performing SMOTE we moved next  
step i.e. applying machine learning model to predict a class for a new observation.  
In this project we have used three models-  
1. Logistic Regression  
2. Decision Tree model  
3. Random Forest model

4.XGB Classifier model

**1. Logistic Regression**:

**Definition**:

• The logistic regression "Supervised machine learning" approach can be used to model the probability of a specific class or occurrence. When the result is binary or dichotomous and the data can be divided linearly, it is used.



logistic regression results in a logistic curve that can only have values between 0 and 1. Similar to a linear regression, a logistic regression builds its curve using the natural logarithm of the target variable's "odds" rather than the probability. Additionally, the predictors don't necessarily need to be evenly distributed among groups or have a normal distribution... This means that logistic regression is frequently used to address problems with binary categorization.

**Working:**

* Based on one or more predictor variables, logistic regression is used to predict the class (or category) of individuals (x). It is used to simulate a binary result, or a variable with only two possible values, such as 0 or 1, yes or no, or diseased or not.

### 2.Decision Trees:

### Definition:

A decision tree uses a tree structure to develop classification or regression models. It incrementally develops an associated decision tree while segmenting a dataset into smaller and smaller sections. The outcome is a tree containing leaf nodes and decision nodes. The root node is the topmost decision node in a tree and corresponds to the best predictor. Decision trees can manage categorical and numerical data.

It is one of the most well-liked and effective methods for supervised learning. Both classification and regression tasks are performed using decision trees, a non-parametric supervised learning method. The objective is to learn straightforward decision rules derived from the data attributes in order to build a model that predicts the value of a target variable.

### C:\Users\admin\Desktop\index.png

#### Working:

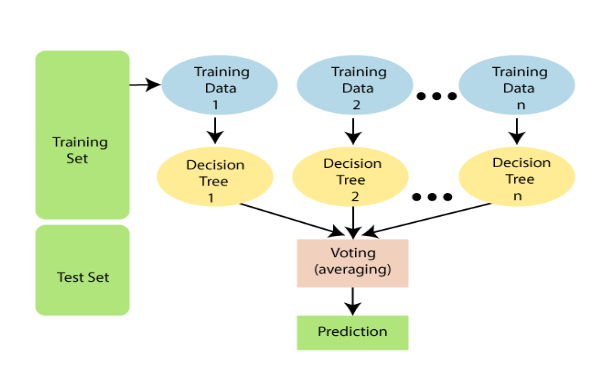
• The variables that serve as nodes in the decision trees are established by calculating the Information Gain of each independent variable in the dataset in order to produce the finest decision trees possible.

•We must first define entropy, a term frequently used in information theory to describe the degree of impurity in a group of samples, before we can accurately characterise information gain.

**3.Random Forests:**

**Definition**:

The decision trees used in the random forest classification technique are numerous. Random Forest, as the name implies, is a classifier that uses a number of decision trees on different subsets of the provided dataset and averages them to increase the dataset's predictive accuracy. Instead of relying on a single decision tree, the random forest uses forecasts from each tree and calculates the final output based on the majority votes of predictions. It uses bagging and feature randomness to construct each individual tree in an effort to create an uncorrelated forest of trees whose forecast by committee is more precise than that of any individual tree.



#### Algorithm:

First, N decision trees are combined to generate the random forest, and then predictions are made for each tree that was produced in the first phase.

Working of Random Forest can be explained with these steps:

Step 1: From the training data set grab K data points at random.

Step 2: Now build a decision tree taking into consideration the data points.

Step 3: Select N for the size of the decision trees you wish to construct.

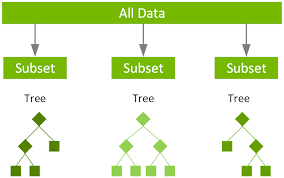
Step 4: Repeat steps 1 and 2.

Step 5: Assign new data points to the category that receives the majority of votes by finding each decision tree's predictions for the new data points.

**4. XGBOOST Classifier**:

**Definition**:

Gradient-boosted decision tree (GBDT) machine learning library called Extreme Gradient Boosting (XGBoost) is networked and scalable. It features parallel tree boosting and is the best machine learning library for problems with regression, classification, and ranking.

****

**Working:**

It is based on the theory that when past models are coupled with the best feasible next model, the overall prediction error is minimised. Random Forest operates in two phases. The first is the creation of the random forest by mixing N decision trees. The main idea is to specify the expected outcomes for this following model in order to lower inaccuracy.

**Hyper-parameter Tuning:**

Now we performed hyper-parameter tuning on the models which are performing well.

**Random Forest Classifier:**

We used Bayes search Cv for hyperparameter tuning. We used max\_depth, min\_samples\_leaf, min\_samples\_split, n\_estimators and max\_features as our hyperparameter to tune our model. After tuning it gives accuracy around 0.855315, recall around 0.895048, Precision around 0.829152 and F1 score around 0.860841.

**XGBoost Classifier:**

We used Bayes search Cv for hyper parameter tuning. We used max\_depth, min\_samples\_leaf, min\_samples\_split, n\_estimators and max\_features as our hyperparameter to tune our model. After tuning it gives accuracy around 0.906026, recall around 0.915862, Precision around 0.898188 and F1 score around 0.906939 which are better.

# Conclusion:

* Due to the Response variable's value 1 being much lower than its value 0, the provided dataset is an imbalance dataset.
* Compared to their female counterparts, male consumers own a little bit more vehicles and have a higher likelihood to get insurance.
* Customers between the ages of 30 and 60 are the most likely to get insurance   whereas Vehicle insurance is not interesting to anyone under the age of 30. The lack of involvement, a lack of knowledge about insurance, and possibly the lack of expensive vehicles are potential causes.
* Customers with driving licences are more likely to purchase insurance
* Compared to consumers with vehicles less than one year old, those with vehicles between one and two years old are more interested in purchasing insurance.
* Due to their personal experience with the costs associated with vehicle repairs, customers with vehicle damage are more likely to purchase insurance.
* The variable such as Age, Previously\_insured, Annual\_premium is more affecting the target variable.
* We used different types of algorithms to train our model like, Logistic Regression, Random Forest model, Decision tree and XGB Classifier. And Also we tuned the parameters of XGBClassifier and Random Forest model.Comparing the model on the basis of precision, recall, accuracy, F1 score we can see that the XGBClassifier model performs better. Even comparing the ROC curve XGB Classifier performed better because curves closer to the top-left corner indicate better performance.

**REFERENCES:**

1. [*https://towardsdatascience.com/*](https://towardsdatascience.com/)
2. GeeksforGeeks
3. Analytics Vidhya
4. [*https://stackoverflow.com/*](https://stackoverflow.com/)