Capstone Project-3

HEALTH INSURANCE CROSS SELL PREDICTION

**Abstract**

Predicting whether a customer would be interested in buying Vehicle Insurance so that the company can then accordingly plan its communication strategy to reach out to those customers and optimise its business model and revenue. 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. There are multiple factors that play a major role in capturing customers for any insurance policy. Here we have information about demographics such as age, gender, region code, and vehicle damage, vehicle age, annual premium, policy sourcing channel. Based on the previous trend, this data analysis and prediction with machine learning models can help us understand what are the reasons for news popularity on social media and obtain the best classification model.

# Problem Statement

Our client is an Insurance company that has provided Health Insurance to its customers. Now they need the help in building a model to predict whether the policyholders (customers) from the 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.

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 optimize its business model and revenue.

# Data Description

# We have a dataset which contains information about demographics (gender, age, region code type), Vehicles (Vehicle Age, Damage), Policy (Premium, sourcing channel) etc. related to a person who is interested in vehicle insurance. We have 381109 data points available.

| **Feature Name** | **Type** | **Description** |
| --- | --- | --- |
| id | (continous) | Unique identifier for the Customer. |
| Age | (continous) | 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 | (continous) | 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 ie. Different Agents, Over Mail, Over Phone, In Person, etc. |
| Vintage | (continous) | 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. |

# Data Wrangling

# After loading our dataset, we observed that our dataset has 381109 rows and 12 columns. We applied a null check and found that our data set has no null values. Further, we treated the outliers in our dataset using a quantile method.

# Univariate Analysis:

# From above fig we can see that the data is highly imbalanced.

# From the above distribution of age we can see that most of the customers age is between 21 to 25 years.There are few Customers above the age of 60 years.

# From the distribution plot we can infer that the annual premimum variable is right skewed

# For the boxplot above we can see that there's a lot of outliers in the annual premium.

# 

# Bivariate analysis

# People ages between from 31 to 50 are more likely to respond.

# while Young people below 30 are not interested in vehicle insurance.

# Male category is slightly greater than that of female and chances of buying the insurance is also little high

# Customers with vechicle age 1-2 years are more likely to interested as compared to the other two

# Customers with with Vehicle\_Age <1 years have very less chance of buying Insurance

# People who response have slightly higher annual premium

## 4 Normalization

## After outlier treatment, we observed that the values in the numeric columns were of different scales, so we applied the min-max scaler technique for feature scaling and normalization of data.

## 5  EDA

## In Exploratory Data Analysis, firstly we explored the 4 numerical features: Age, Policy\_Sales\_Channel, Region\_Code, Vintage. Further, we categorized age as youngAge, middleAge, and oldAge and also categorized policy\_sales\_channel and region\_code. From here we observed that customers belonging to the youngAge group are less interested in taking vehicle insurance. Similarly, Region\_C, Channel\_A have the highest number of customers who are not interested in insurance. From the vehicle\_Damage feature, we were able to conclude that customers with vehicle damage are more likely to take vehicle insurance. Similarly, the Annual Premium for customers with vehicle damage history is higher.

# Encoding categorical values

## We used one-hot encoding for converting the categorical columns such as 'Gender', 'Previously\_Insured','Vehicle\_Age','Vehicle\_Damage', 'Age\_Group', 'Policy\_Sales\_Channel\_Categorical', 'Region\_Code\_Categorical' into numerical values so that our model can understand and extract valuable information from these columns.

# Feature Selection

## At first, we obtained the correlation between numeric features through Kendall’s Rank Correlation to understand their relation. We had two numerical features, i.e. Annual\_Premium and Vintage. For categorical features, we tried to see the feature importance through Mutual Information. It measures how much one random variable tells us about another.

# Model Fitting

## For modeling, we tried the various classification algorithms like

## Logistic Regression-

## Decision Trees are non-parametric supervised learning methods, capable of finding complex non-linear relationships in the data. Decision trees are a type of algorithm that uses a tree-like system of conditional control statements to create the machine learning model. A decision tree observes features of an object and trains a model in the structure of a tree to predict data in the future to produce output. For classification trees, it is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. Logistic regression is not performing well on this dataset as in confusion matrix model is predicting positive responses but with positive responses it is predicting negative responses in high numbers too.

## RandomForest Classifier-

## A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is controlled with the Parameter if bootsrap=True otherwise whole data set build in each tree. Here , random forest is performing better as in the confusion matrix the model now is much better with predicting positive responses.

## XGBoost-

## XGBoost comes under boosting and is known as extra gradient boosting.

## GBM first calculates the model using X and Y then after the prediction is obtain.

## It will again calculates the model based on residual of previous model

## loss function will give more weightage to error of previous model. and this process continuous until MSE gets minimizes.

## From the confusion matrix we see that the model is a bit better with predicting positive responses.

## Comparing the Model

Further, we applied Machine Learning Algorithms to determine whether a customer would be interested in Vehicle Insurance.For the logistic regression we got an accuracy of 78% and for the XGBClassifier we got the aacuracy of 79% whereas,.We are getting the highest accuracy of about 91% and ROC\_AUC score of 92% with random forest So, From this we can conclude that random forest is the best models as compare to the other models

# Conclusion

* from loading our dataset, we initially checked for null values and duplicates. There were no null values and duplicates so treatment of such was not required.
* Through Exploratory Data Analysis, we observed that customers belonging to young Age are more interested in vehicle response .while Young people below 30 are not interested in vehicle insurance. We observed that customers having vehicles older than 2 years are more likely to be interested in vehicle insurance. Similarly, customers having damaged vehicles are more likely to be interested in vehicle insurance.
* The variable such as Age, Previously\_insured, Annual\_premium are more affecting the target variable.
* For Feature Selection, we applied the Mutual Information technique. Here we observed that Previously\_Insured is the most important feature and has the highest impact on the dependent feature and there is no correlation between the two.
* We observed that the target variable was highly imbalanced.So this issue was solved by using Random Over Sample resampling technique.
* we applied feature scaling techniques to normalize our data to bring all features on the same scale and make it easier to process by ML algorithms.
* Further, we applied Machine Learning Algorithms to determine whether a customer would be interested in Vehicle Insurance.For the logistic regression we got an accuracy of 78% and for the XGBClassifier we got the aacuracy of 79% whereas,.We are getting the highest accuracy of about 91% and ROC\_AUC score of 92% with random forest So, From this we can conclude that random forest is the best models as compare to the other models.