Capstone Project-3

HEALTH INSURANCE CROSS SELL PREDICTION

BY:- ASHWANI BIST

**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 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).

The dataset contains data on demographics (gender, age, region code type), vehicles (vehicle age, damage), policies (premium, sourcing channel), and so on. We predicted that the customer who has medical insurance from the company will or will not be interested in purchasing a vehicle insurance policy based on this feature. This model is extremely beneficial to the company because it allows it to plan its communication strategy to reach out to those customers and optimize its business model and revenue accordingly.

# 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 | (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 ie. 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. |

# 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:

# The data is highly imbalanced.

# From the distribution of age, we can conclude 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 premium variable is right skewed

# From the boxplot we conclude 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 vehicle 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

We applied different Machine Learning Models to our data set and see how each of them performs. Firstly, We will tune the hyper-parameters of those models and then we will compare and choose the best model among them, based on Elapsed Time and Evaluation Metrics of the best parameters.

List of **Machine Learning Models** we are going to train and evaluate our data set on:

* Decision Tree
* Gaussian Naive Bayes
* AdaBoost Classifier
* Bagging Classifier
* LightGBM
* Logistic Regression

### ****Hyper-Parameter Tuning Methods:****

We have tried different hyper-parameter tuning methods. Every method gave the same result but **GridSearchCV** and **RandomizedSearchCV** took a huge amount of time to train the models. **HalvingRandomizedSearchCV** took the least time to train the models and predict the output. That's why we highly **recommend** you to keep the Tuning\_Method as Halving\_Randomized\_Search\_CV from the drop-down menu below.

We have also added some results of the model tuning with GridSearchCV and RandomizedSearchCV, just for performance comparison.

#### **Tuning Methods:**

* HalvingRandomizedSearchCV
* GridSearchCV
* RandomizedSearchCV

### ****Evaluation Metrics:****

* Accuracy Score
* Precision
* Recall
* F1 Score
* ROC AUC Score
* Log Loss

### ****Plots:****

At the end of every model's hyper-parameter tuning, there is one **ROC Curve** which shows the ROC Scores and **Parallel Coordinates Plot** which shows all the combinations of hyper-parameters used for tuning the model to get the best parameters.

## ****Best Model****

From all the above models that we tried to train and predict the output, we can conclude that **Bagging Classifier** is the best model for our data set. The best parameter of this model is {'n\_estimators': 200}. Its Accuracy Score is 0.85, Precision is 0.31, Recall is 0.15, F1\_Score is 0.20, ROC\_AUC\_Score is 0.55 and Log\_Loss is 4.98. Its Elapsed time is 03 minutes and 21 seconds.

We can see that we have other models with higher Accuracy Score than Bagging Classifier. But the problem with those models is, their Precision and Recall values are zero which means True Positives are zero. That means those models are unable to predict correct output if any customer is ready to take vehicle insurance. And as we all know, classification accuracy alone can be misleading if you have an unequal number of observations in each class. This is exactly the case with our data set.

Hence, ***Bagging Classifier*** is the ***best model*** for our data set

# Conclusion

Starting 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. Before data processing, 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.

Through **Exploratory Data Analysis**, we categorized Age as YoungAge, MiddleAge, and OldAge, then we categorized the Region\_Code as Region\_A, Region\_B, Region\_C. We categorized the Policy\_Sales\_Channel into channel\_A, channel\_B, channel\_C. Further, we observed that customers belonging to youngAge are more interested in vehicle response. 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.

For **Feature Selection**, we used Kendall's rank correlation coefficient for numerical features and for categorical features, 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 numeric features

Further, we applied **Machine Learning Algorithms** to determine whether a customer would be interested in Vehicle Insurance. For the Naive Bayes algorithm, we got an accuracy score of 68% and after hyperparameter tuning, the accuracy score increased to 72%. Similarly, for Decision Tree Classifier, AdaBoost, BaggingClassifier, LightGBM accuracy score was obtained around 82%-87%. So, we selected our **best model** as the model with an accuracy score of **85%** considering precision and recall as we have an unequal number of observations in each class in our dataset, so accuracy alone can be misleading.

References: -

Data science for business: what you think about data mining

• Hands-On Exploratory Data Analysis with Python Perform EDA techniques to understand, summarize

• https://bunker2.zlibcdn.com/dtoken/01c5fc197 a94283bfb0c0943bd5b2d0c