**Health Insurance Cross Sell Prediction**

Manda Shilpa

**Data science trainee,**

**Great Lakes Institute, Bangalore**

**Abstract:**

Insurance companies offer a variety of insurance services and generate thousands of sales each day, in this industry, claims do occur, and many customers choose to use such services.

Our study can shed light on the likelihood that a customer who purchases health insurance will also buy auto insurance. To gain insights into our customers, we will do data analysis and prediction using machine learning algorithms while taking into consideration historical trends and if they are planning to buy or use such a service, and what types of customers are interested, etc.

***Keywords****- Data analysis, machine learning, Classification regression, xgbooster, logistic regression, Xgb, Random forest classifier, XGBclassifier.*

**Problem Statement**

Consumers of insurance companies' subscribed health plans make up the dataset.

For this project, the data were examined and cleansed. Based on each consumer's unique data and other information, we will be predicting the consumers who may choose to opt for purchasing the Vehicle insurance of the same company.

The task is to build a model that predicts Health insurance owners' who will be interested in buying Vehicle Insurance and our Dataset consists of Gender, Age, Driving license, vehicle age and premium details, and many other variables.

**Data details :**

* id - A unique id for each customer.
* Gender - Gender details of the health insurance owner.
* Age - Age details of the health insurance owner.
* Driving\_License - Whether the customer has a driving license or Not.
* Region\_Code - Region with code details of the health insurance owner.
* Previously Insured - Whether the customer previously Insured or Not.
* Vehicle\_Age - Age of vehicle of the health insurance owner.
* Vehicle\_Damage - Whether the customer Vehicle Damaged or Not.
* Annual Premium - Annual Premium amount details of a Customer.
* Policy\_Sales\_Channel - Policy Sales Channel shows us, the number of the sales channel.
* Response - Response of the customer to buying vehicle insurance.

**INTRODUCTION:**

The insurance sector is large and offers many different types of plans, Schemes, services provided to the consumer based on their interest, vehicle, type of vehicle, etc.

These businesses struggle with issues like comprehending the customer, purchases made in a particular area, or the charges of their service. The number of purchased customers, interest in their different products, premium region, etc are decisive factors for these companies to understand the demand forecast the business and gain opportunities.

Vehicle insurance means, protecting the car against any kind of accident that can result in losses or damage to cars.

Examining the number of people who purchase health insurance, the cost of that insurance, and other factors details of health insurance owners will help predict these businesses to improve their services and obtain money and vital resources that will aid in enhancing customer service and introducing new products to the market.

Different factors affect vehicle insurance; and the major reasons are Age of customers or vehicle age, gender, area, etc.

**BUSINESS OVERVIEW**

To understand the objective of the project, it is necessary to understand the business.

Companies that provide life insurance charge you premiums, and some of the money they get from your premiums is invested, in addition to taking use of cancelled or expired policies and managing other insurance products, like along with managing other insurance products and utilizing cancelled or expired policies.

These firms mostly focus on sales or subscribers to their insurance policy. Businesses must comprehend the features that affect their industry as well as the causes that are affecting them.

Understanding the proportion of clients interested in auto insurance will be made easier by analyzing the insurance subscribers.

Businesses can forecast customer behavior and make appropriate plans by using these criteria, so to keep their consumers, businesses need to have an effective data analysis model.

**VEHICLE INSURANCE:**

Businesses profit from the sale or subscription of insurance and the associated premium, whereby clients utilized their service, to comprehend health insurance, and to give customers dynamic offerings

Companies must understand the necessary information like gender, age, vehicle, or responses, etc.

The primary goal of this investigation is to identify health insurance owner specifics and comprehend data insights.

While performing the analysis, the most important thing is to identify the customers interested in purchasing auto insurance while taking numerous features into consideration such as gender, age, vehicle, age, etc.

Creating a model to predict who among people with health insurance will purchase vehicle insurance.

For Service-based companies, it is crucial to predict their business and make appropriate plans.

Here in this project, by developing the prediction model, we will be able to predict the health insurance consumers who are considering purchasing auto insurance.

**Characteristics affecting the purchase of insurance**

* Age of customer
* Vehicle type and age
* Sales channel
* Premium, etc.

**APPROACH:**

**EXPLORATORY DATA ANALYSIS:**

Exploratory data analysis is a technique where we attempt to understand the potential of the provided data in order to gain some insights from it and utilize such data.

Using the exploratory data analysis we understand the important elements or characteristics such as Avg, mean Std deviations, etc also to confirm the missing values or null values and outliers.

Exploratory data analysis is a process of verifying the available data set to determine patterns, anomalies, test hypotheses, and check assumptions using statistical measures.

Python is simple to comprehend and provides insights when used in the exploratory data analysis process and visually comparison between the variables.

**ANALYSIS OF DATA:**

Analysis of data is an essential step, which deals with descriptive statistics and analysis of the data.

This stage comprises compiling the information and identifying the subtle connections and effects between the various datasets, it supports developing and foreseeing the models, assessing them, and describing the characteristics in great detail.

**SOURCING OF DATA**

The method of finding data and storing it in our systems or frameworks is known as data sourcing. There are many different ways to see the data, and each one needs to be configured correctly.

An experienced and authorized member of the relevant organization should manage the information. There are numerous tools available to locate the data, collect it, and store it.

**DATA PROCESSING & CLEANING:**

Pre-processing is crucial to enhancing the quality of data because raw data can occasionally contain noise, null values, inconsistent format, and values.

Then, at that time, only we are able to carry out the Data cleaning successfully, which is vital to track down the abnormalities (such as Null Values, incorrect format & Header Anomalies/Outliers, etc)

After that, cleaning the data to gain some insights is simple.

**DATA TRANSFORMATION :**

This method involves enhancing data to increase performance and clarity. Sometimes rows and values in data are duplicated. For the dataset to be of higher quality, duplicate values should be removed.

**MISSING VALUES:**

In datasets, there are numerous causes of missing values, including mistakes or data handling mistakes.

We don't have any Null values.

Therefore, before starting any analysis, we should take care of any missing values in the dataset.

**DROPPING OF MISSING VALUES:**

To remove the missing values, to eliminate the missing values, we can employ a variety of strategies. For example, IsNull() and not-null() functions from the panda's library to determine null values and remove using dropna.

Also handling of outliers based on their impact on our dataset.

By doing this, we can proceed to exploratory analysis after removing the missing values from the dataset.

**ANALYSIS OF DATA :**

**Univariate analysis**

We investigated information of a solitary variable/segment from a dataset, otherwise called Univariate Analysis. In Univariate analysis, we take each component in turn. When we independently disassemble a component, the generally essential adage of this examination is to find the circulation of its values(range) and disregard different highlights in the dataset

The univariate analysis is the least complex type of information examination. A single type of factor should make up the information before we conduct our analysis. The univariate analysis is mostly used to take information, compile the data and look for patterns among the attributes. It does not handle the relationships or causes between the attributes.

In graphical methods, we utilized various plots, for example, pie diagrams and so forth.

**Bivariate analysis**

Analyzing data is done by taking two variables/columns into account from a dataset, known as bivariate analysis. Graphical strategies we utilized various plots, for example, box plots and so forth.

**IDENTIFYING CORRELATION** :

Correlation assists with building up the relations between the factors and gives significant experiences, for example, which variables have a significant impact on one another and what each factor means for the other. Also Customers' responses regarding their decision to purchase auto insurance help the business plan appropriately. Additionally, it helps with applying good judgment, expectations, or predictions and is very helpful for future or visionary forecasts of organizational items.

**GRAPHICAL REPRESENTATION OF THE RESULTS:**

This step depicts the investigated dataset to the ideal interest group as perception techniques like diagrams, synopsis tables, plots, drawings, etc.

**Encoding of categorical columns**

We utilized Label Encoding to deliver binary numbers of 0 and 1 to encode our unmitigated highlights because absolute elements that are in string design can't be perceived by the machine and should be changed over to the mathematical organization.

Label encoding is the process of transforming labels into a numeric form so that they may be read by machines. The operation of those labels can then be better determined by machine learning techniques. It is a crucial step in the supervised learning preparation of the structured dataset.

**Feature Selection**

In these means, we utilized calculations like the ExtraTreesClassifier It employs a meta estimator to fit a number of randomly selected decision trees to various variable subsamples.

ExtraTreesClassifier is an ensemble learning method fundamentally based on decision trees. ExtraTreesClassifier, like RandomForest, randomizes certain decisions and subsets of data to minimize over-learning from the data and overfitting.

**Standardization of features**

Our main goal in moving forward was to get our data into a standardised structure that would enable us to use it better while doing fitting and applying various calculations to it.

The main goal was to make sure that specific procedures or methods operated consistently throughout a chosen setting

and removed the least important variables.

**Model fitting**

For modeling, we tested various algorithms like:

The issue can be categorized as Binary, such as whether a consumer buys vehicle insurance or not.

The given data consists of more than 300000 information or variables.

It's better to stay away from SVM Classifier because it takes more time to train as the dataset increases.

So we have used below classification algorithms:

* Logistic Regression
* Random forest classifier
* Xgbclassifier

**RESAMPLING OF DATA**

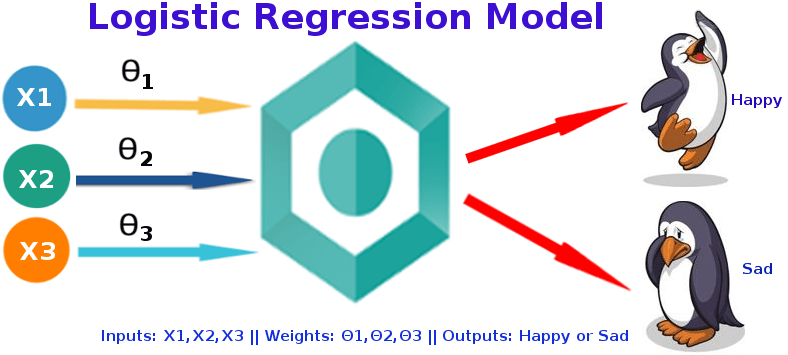
With the use of resampling, we will resolve the class imbalance. There is a class imbalance when the observations in one class are higher than the observations in other classes. We can see that there is a colossal distinction between the informational index.

**ALGORITHMS**

**Logistic regression**

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. There are only two viable classes since the goal or dependent variable is dichotomous in nature.

Mathematically, a logistic regression model predicts P(Y=1) as a function of X. It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, Diabetes prediction, cancer detection, etc.



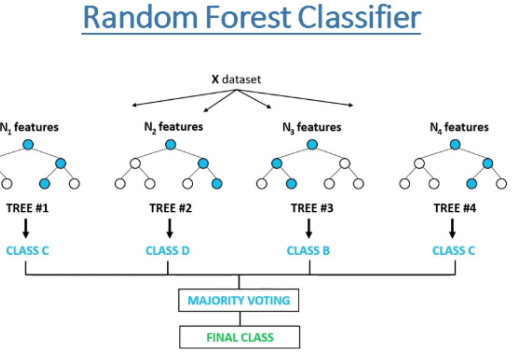
**Random forest classifier**

A random forest is a supervised machine learning algorithm that is constructed from decision tree algorithms. This algorithm is used to forecast behavior and results in a variety of industries, including banking and e-commerce.

A random forest algorithm consists of many decision trees. The random forest method, which was trained via bagging or bootstrap aggregation, creates the "forest." The accuracy of machine learning algorithms is increased by bagging, an ensemble meta-algorithm.

The (random forest) algorithm establishes the outcome based on the predictions of the decision trees. It makes predictions by averaging or averaging out the results from different trees. The accuracy of the result grows as the number of trees increases.

The decision tree algorithm's shortcomings are eliminated with a random forest. It improves precision and lowers dataset overfitting. It produces forecasts without requiring a lot of package configurations.



**XGB Classifier**

XGboost is the most widely used algorithm in machine learning, whether the problem is a classification or a regression problem. When compared to all other machine learning algorithms, it is renowned for its strong performance.

Let's say you have a variety of data sets, each containing a wide range of unique attributes. But some of the features hold few similarities to their operations, output goals, etc. In simpler terms, we might state that there may appear to be a number of data subsets with overlapping characteristics. Classifying those data subsets based on feature similarities is what you should do if you need to find these kinds of feature subsets of data in order to get the outcome you want.

But this type of classification will be automated. To initiate that, Based on the variables and observations of the considered dataset, all of the data must be coached. Once the data has been trained, it becomes dynamic and can classify or categorize all future new datasets using the knowledge gained from prior training.

Now XGboost owns the ability to manage both types of situations, whether you need to go with regression or classification modeling. So, we can consider XGboost both as a classification and regression algorithm.



**Performance of models**

To assess the effectiveness of your regression model, some statistical metrics are used.

Here we will discuss four of the most popular metrics. They are-

* Confusion matrix
* Precision or Recall
* Accuracy
* ROC

A model can be evaluated by various metrics such as:

**Confusion Matrix-**

The classification models' accuracy in classifying examples into distinct groups is summarized in a table called the confusion matrix. One axis of the confusion matrix is the label that the model predicted, and the other axis is the actual label.

**Precision/Recall-**

Precision is the proportion of right sure forecasts to the general number of positive expectations: TP/TP+FP

The recall is the proportion of right certain expectations to the general number of positive models in the set: TP/FN+TP

**Accuracy**

The accuracy is calculated by dividing the total number of classified examples by the number of cases that were successfully categorized. In terms of the confusion matrix, it is given by: TP+TN/TP+TN+FP+FN

**ROC Curve**

ROC curves use a mixture of the true positive rate and the false positive rate to build up a summary picture of the classification performance.

**Hyperparameter tuning:**

Hyperparameters are sets of details that are utilized to control the path of learning an algorithm. Their descriptions have an effect on the model parameters, which are changed as a result of the new hyperparameters. This set of values affects the performance, stability, and interpretation of a model. Each algorithm has a unique set of hyperparameters that can be changed depending on the business issue. Hyperparameters change how a model comprehends how to start the training method and how to create outputs after parameters.

**CONCLUSIONS:**

1. Customers aged between 30 to 60 are more likely to buy insurance.
2. Youngsters under 30 are not intrigued by vehicle insurance. The lack of involvement, a lack of knowledge about insurance, and possibly the lack of expensive vehicles are potential causes.
3. Consumers with 1-2-year-old vehicles are more interested as compared to others.
4. Consumers with less than 1-year-old Vehicles have very less chance of buying Insurance
5. Customers with Driving License have a higher chance of buying Insurance.
6. Customers with Vehicle\_Damage are likely to buy insurance.
7. Males are slightly more notable than females in this regard, and their odds of buying insurance are also not particularly great.
8. The variable such as Age, Previously\_insured, Annual\_premium is more affecting the target variable.
9. Comparing the ROC curve we can see that the Random Forest model performs better. Because better performance is shown by curves that are closer to the top-left corner.

**REFERENCES:** <https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/>

<https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>

<https://www.sciencedirect.com/topics/computer-science/logistic-regression#:~:text=Logistic%20regression%20is%20a%20process,outcome%20given%20an%20input%20variable.&text=Logistic%20regression%20is%20a%20useful,fits%20best%20into%20a%20category.>