**Health Insurance Sell Prediction.**

# **ABSTRACT.**

Our client is an insurance company that has provided Health Insurance to its customers now they need our 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 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.

## **INTRODUCTION**

Insurance is an agreement by which an individual obtains protection against any losses from an insurance company against the risks of damage, financial losses, damage, illness, or death in return for the payment of a specified premium. In this project, we have an insurance details dataset which contains a total of *381109 rows* and *12 features*. We have a categorical dependent variable *Response* which represents whether a customer is interested in vehicle insurance or not. As an initial step, we checked for the null and duplicate values in our dataset. As there were no null and duplicate values present in our dataset, so data cleaning was not required. Further, we *normalized* the numerical columns for bringing them on the same scale.

In **Exploratory Data Analysis**, we categorized the Age as *YoungAge,* *MiddleAge, OldAge*.Then we categorized *Region\_Code* and *Policy\_Sales\_Channel* to extract some valuable information from these features. We explored the independent features using some plots.

For **Feature selection**, we used Kendall's rank correlation coefficient for numerical features and for categorical features, we applied the Mutual Information technique.

For **Model prediction**, we used supervised machine learning algorithms like Logistic Regression, Random Forest, XGB Classifer. Then applied hyperparameter tuning techniques to obtain better accuracy and to avoid overfitting.

## **DATA SET.**

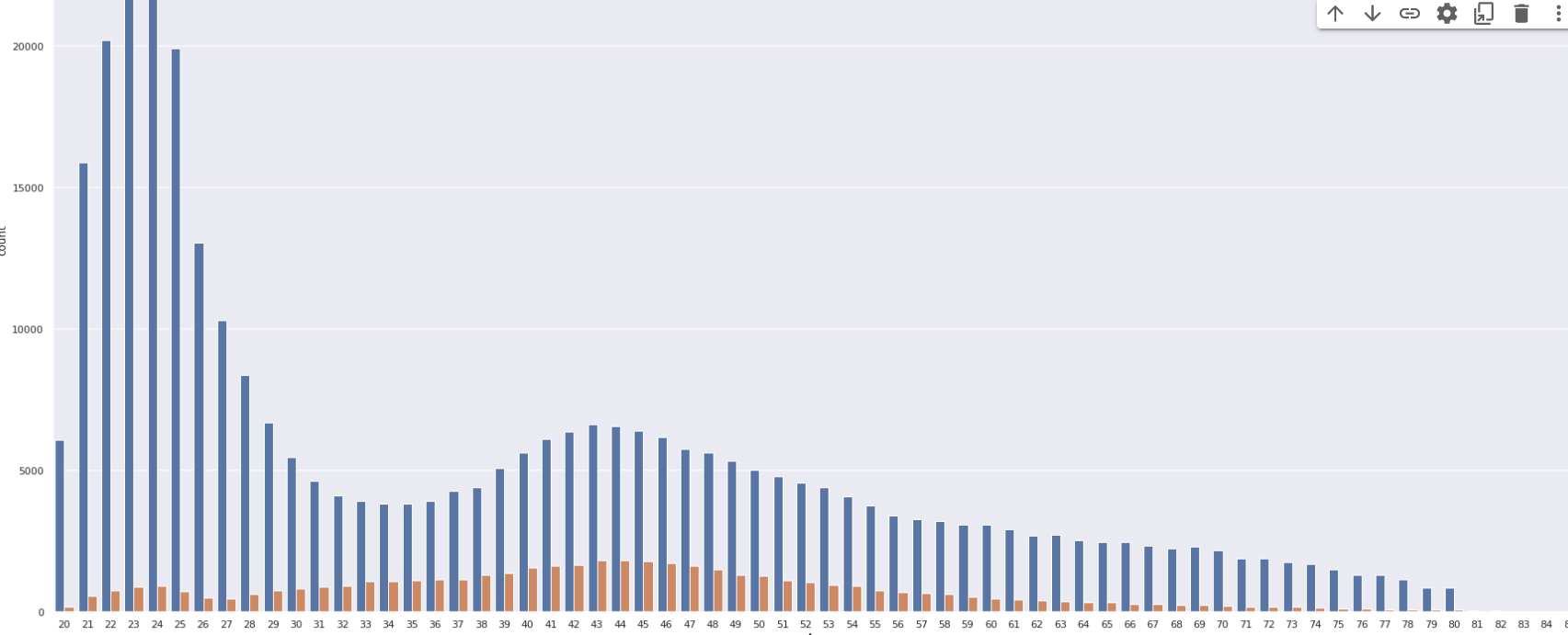
1. ID : Unique ID for the customer.
2. Gender : Gender of the customer.
3. Age : Age of the customer.
4. Driving License 0 : Customer does not have DL, 1 : Customer already has DL.
5. Region Code : Unique code for the region of the customer.
6. Previously Insured : 1 : Customer already has Vehicle Insurance, 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. Policy Sales Channel : 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.

## **DATA VISUALIZATION.**

* UNIVARIATE ANALYSIS.
* BIVARATE ANALYSIS.
* FINDING THE OUTLIERS.
* FINDING THE NULL VALUES AND FILLING IT WITH MIN/MAX SCALER.
* RESPONSE ANALYSIS.
* COUNT OF MALE AND FEMALE .
* RESPONSE OF MALE AND FEMALE.



## 



## **PREPARE DATA**

* SPILTTING THE DATASET INTO TRAINING AND TESTING SET.
* FEATURE SELECTION.
* HANDLING IMBALANCE DATA.

## **MODEL SELECTION.**

**LOGISTIC REGRESSION.**

1. **Logistic Regression:**

Logistic Regression is actually a classification algorithm that was given the name regression due to the fact that the mathematical formulation is very similar to linear regression.

The function used in Logistic Regression is sigmoid function or the logistic function given by:

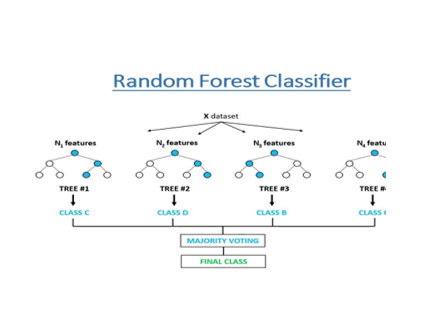
f(x)= 1/1+e ^(-x).

The optimization algorithm used is: Maximum Log Likelihood. We mostly take log likelihood in Logistic.

### 

## **RANDOM FOREST**

Random Forest is a bagging type of Decision Tree Algorithm that creates a number of decision trees from a randomly selected subset of the training set, collects the labels from these subsets and then averages the final prediction depending on the most number of times a label has been predicted out of all.



**XGB CLASSIFIER.**

To understand XGBoost we have to know gradient boosting beforehand.

* **Gradient Boosting-**

Gradient boosted trees consider the special case where the simple model is a decision trees.

****

In this case, there are going to be 2 kinds of parameters P: the weights at each leaf, w, and the number of leaves T in each tree (so that in the above example, T=3 and w=[2, 0.1, -1]).

When building a decision tree, a challenge is to decide how to split a current leaf. For instance, in the above image, how could I add another layer to the (age > 15) leaf? A ‘greedy’ way to do this is to consider every possible split on the remaining features (so, gender and occupation), and calculate the new loss for each split; you could then pick the tree which most reduces your loss.

**XGBoost** is one of the fastest implementations of gradient boosting. trees. It does this by tackling one of the major inefficiencies of gradient boosted trees: considering the potential loss for all possible splits to create a new branch (especially if you consider the case where there are thousands of features, and therefore thousands of possible splits). XGBoost tackles this inefficiency by looking at the distribution of features across all data points in a leaf and using this information to reduce the search space of possible feature splits.

## **CONCLUSION.**

* Customers of age between 30 to 60 are more likely to buy insurance.
* Customers with Driving License have higher chance of buying Insurance.
* Customers with Vehicle Damage are likely to buy insurance.
* The variable such as Age, Previously insured , Annual premium are more affecting the target variable.
* comparing ROC curve we can see that Random Forest model preform better. Because curves closer to the top-left corner, it indicate a better performance.