Capstone Project - 3

HEALTH INSURANCE CROSS SELL PREDICTION

Team members :

Rohit Meshram

Rahul Gayakwad

Rutuja Hingankar

Prashik Ingle

Narayan Borde

Content

- **Introduction**
- **Problem statement**
- **Data summary**
- **Exploratory Data Analysis (EDA)**
- **Feature Engineering & Selection**
- **Building and Evaluating Model**
- **Conclusion**

Introduction

- 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.
- 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.

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.
- 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.

Dataset Description

id : Unique identifier for the Customer.

Gender : Age of the Customer.

Age : Gender of the Customer

Driving License :0 for customer not having DL, 1 for customer having DL.

Region Code :Unique code for the region of the customer.

Previously Insured:0 for customer not having vehicle insurance, 1 for customer having vehicle insurance.

Vehicle Age :Age of the vehicle.

Vehicle Damage :Customer got his/her vehicle damaged in the past. 0 : Customer didn't get his/her vehicle damaged in past

Annual Premium :The amount customer needs to pay as premium in the year.

Policy Sales Channel : Anonymized Code for the channel of outreaching to the customer ie. Different Agents, Over Mail

Vintage :Number of Days, Customer has been associated with the company.

Response :1 for Customer is interested, 0 for Customer is not interested.

Import Libraries

- Library is a collection of related modules. It contains bundles of code that can be used repeatedly in different programs. It makes python programming simpler and convenient for the programmer.

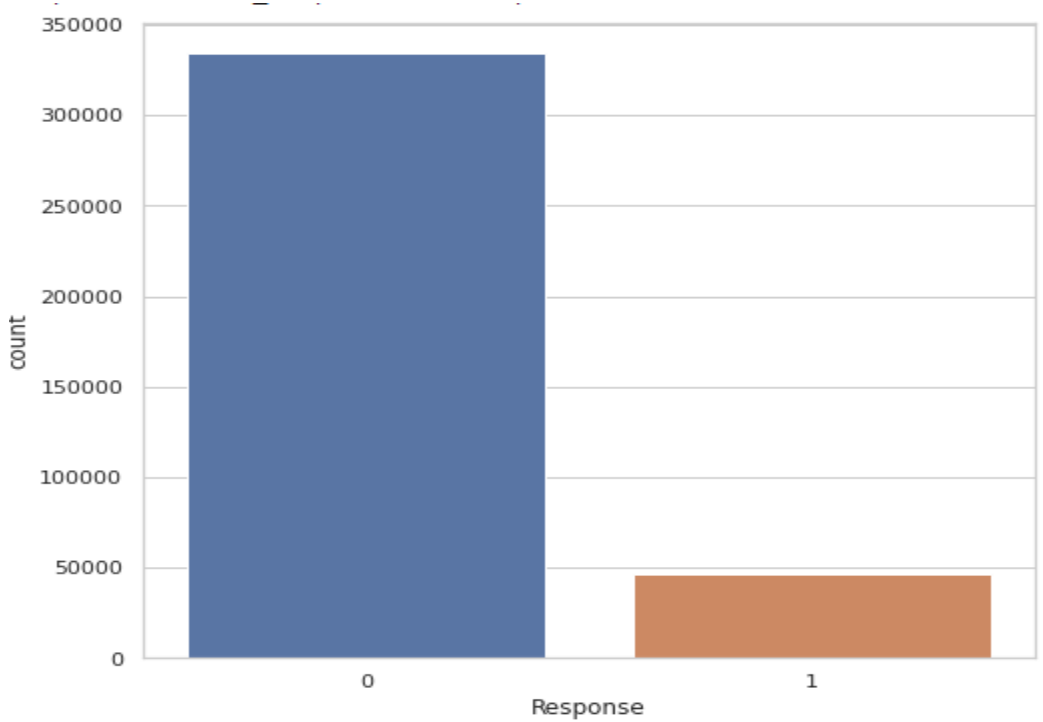
- **Importing Requrie Libraries**

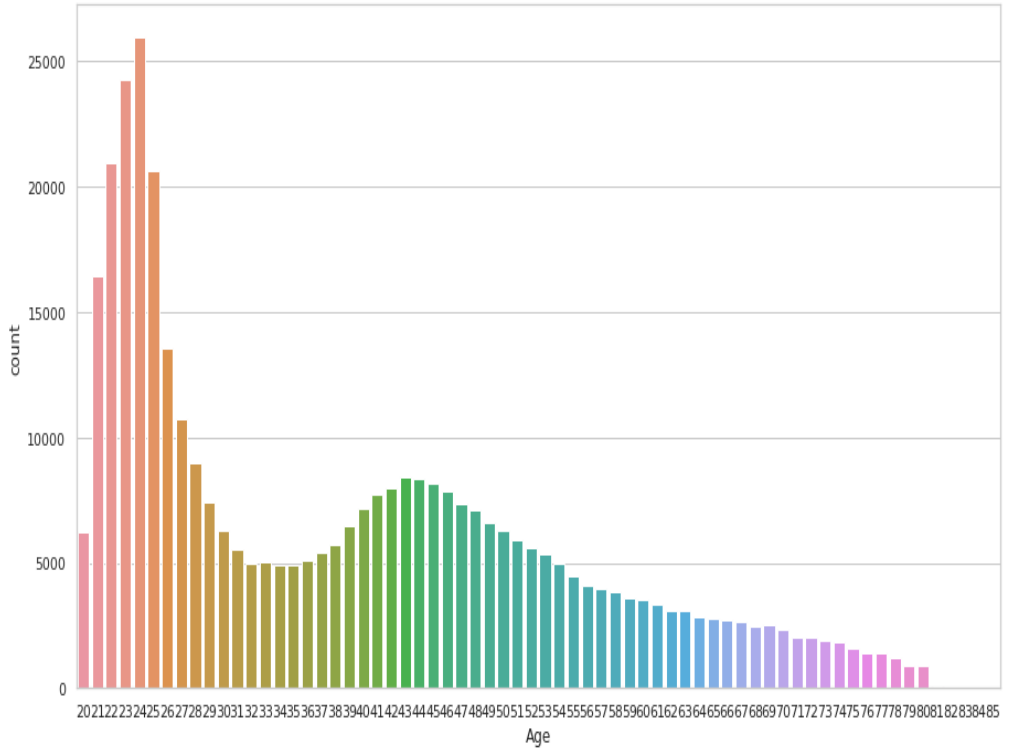
```
[ ]   1 import pandas as pd
      2 import numpy as np
      3 import seaborn as sns
      4 import matplotlib.pyplot as plt
```

BASIC EXPLORATION

- The dataset contains 381109 rows and 12 columns.
- Outliers present in some features.
- No null values present
- Fill any numerical NaNs with mode()
- Id, Age, Driving_License, Previously_Insured, Vintage and Response are having integer value.
- Response , Annual_Premium and Policy_Sales_Channel are having float values
- Drop duplicate value
- Changing categorical value to numerical values

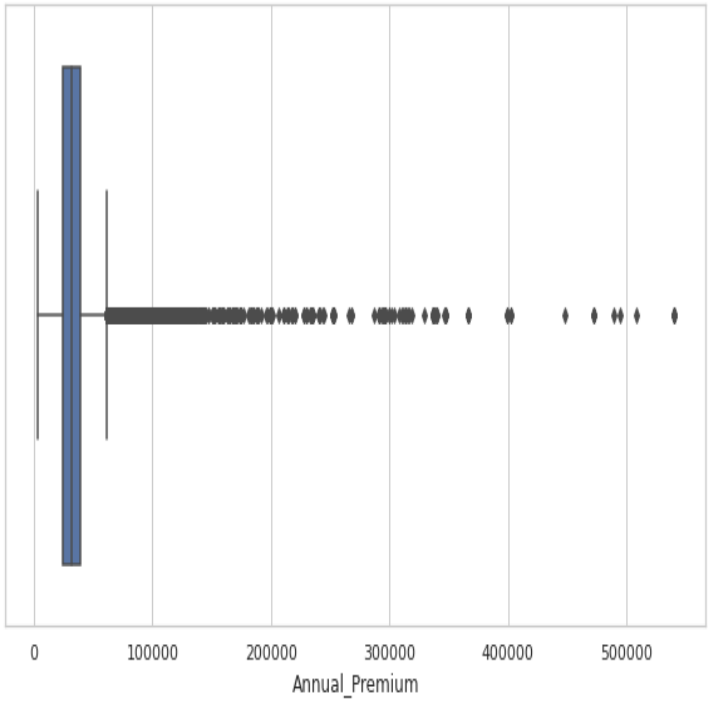
Univariate Analysis



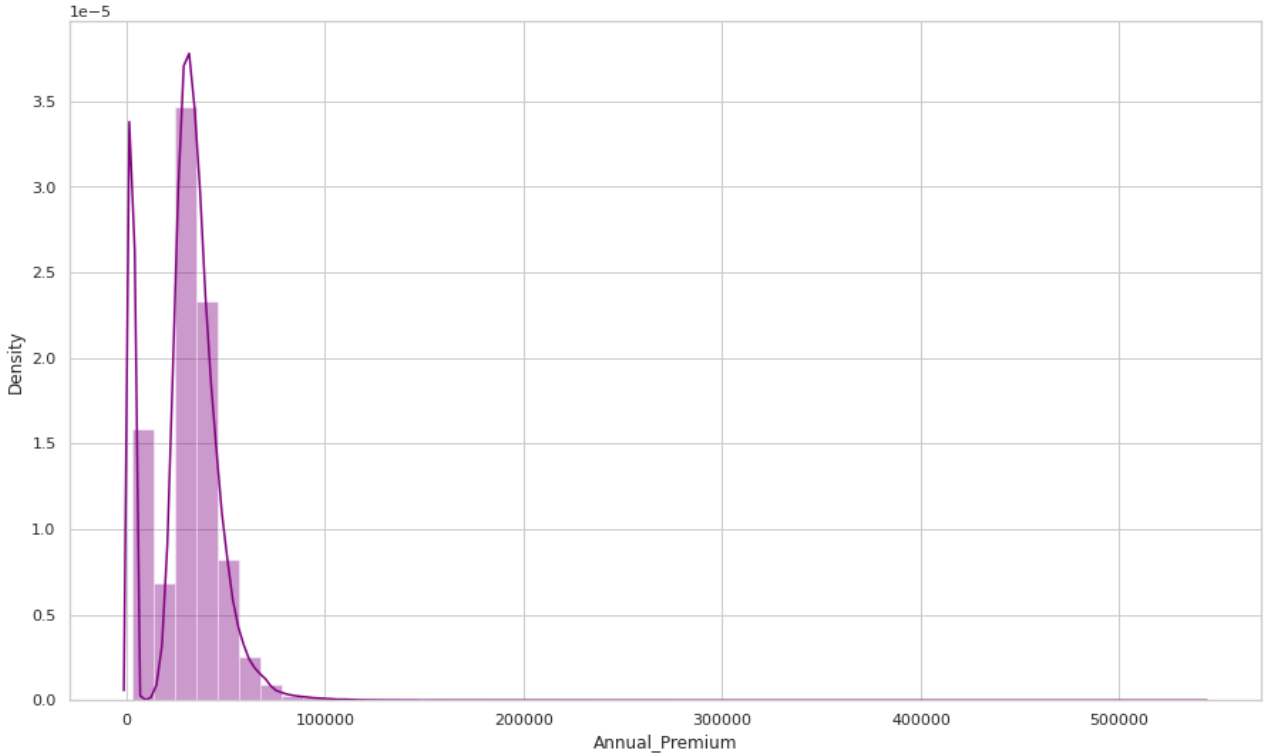


From above fig of response we can see that the data is highly imbalanced.
From the above fig of distribution of age we can see that most of the customers age is between 21 to 25 years.There are few Customers above the age of 60 years.

Univariate Analysis

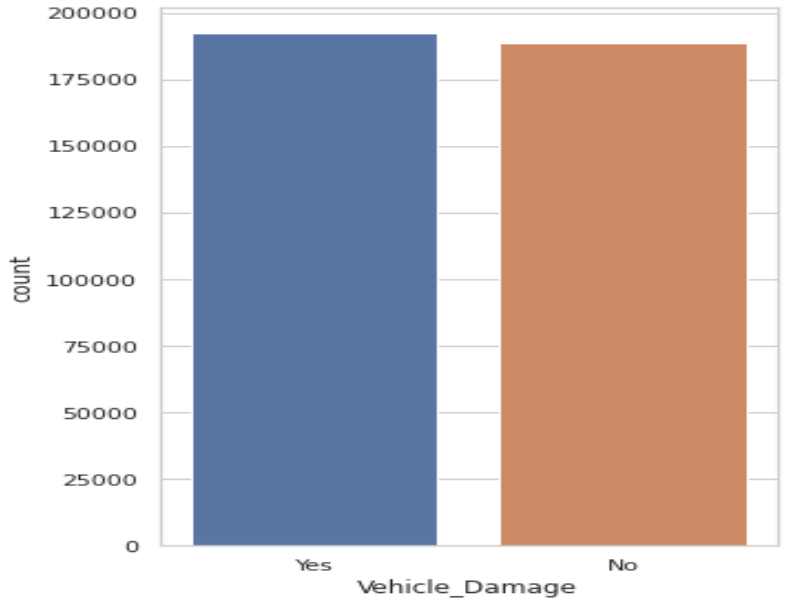


For the boxplot above we can see that there's a lot of outliers in the annual premium.

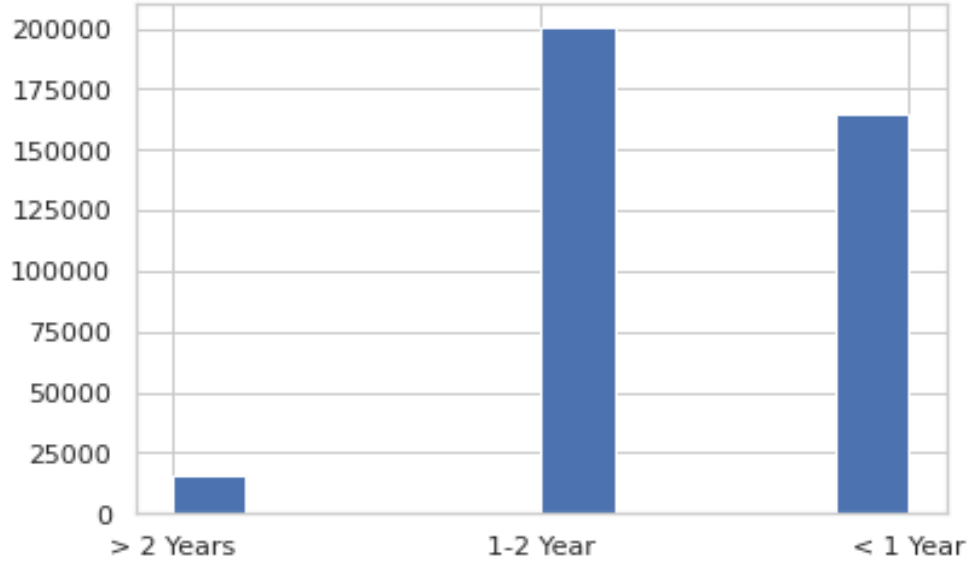


From the distribution plot we can infer that the annual premimum variable is right skewed

Univariate Analysis

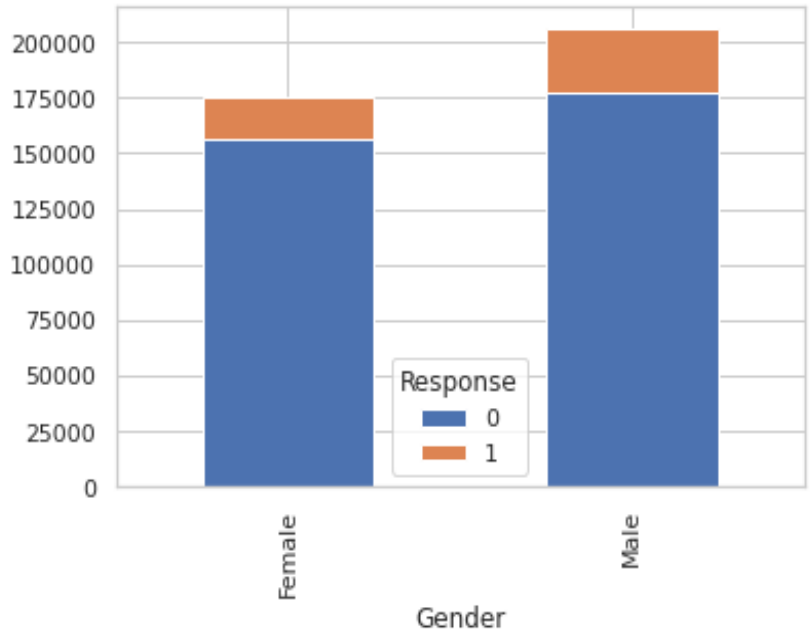


- Customers with Vehicle_Damage are likely to buy insurance

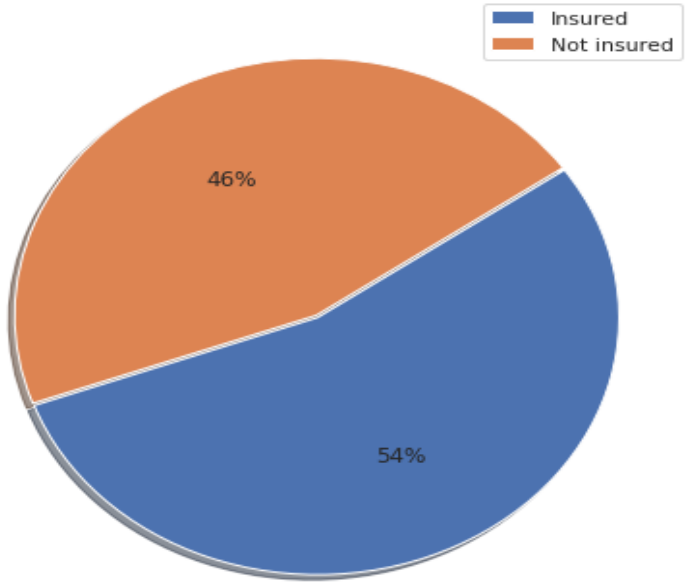


- From the above plot we can see that most of the people are having vehicle age between 1 or 2 years and very few people are having vehicle age more than 2 years.

Data analysis

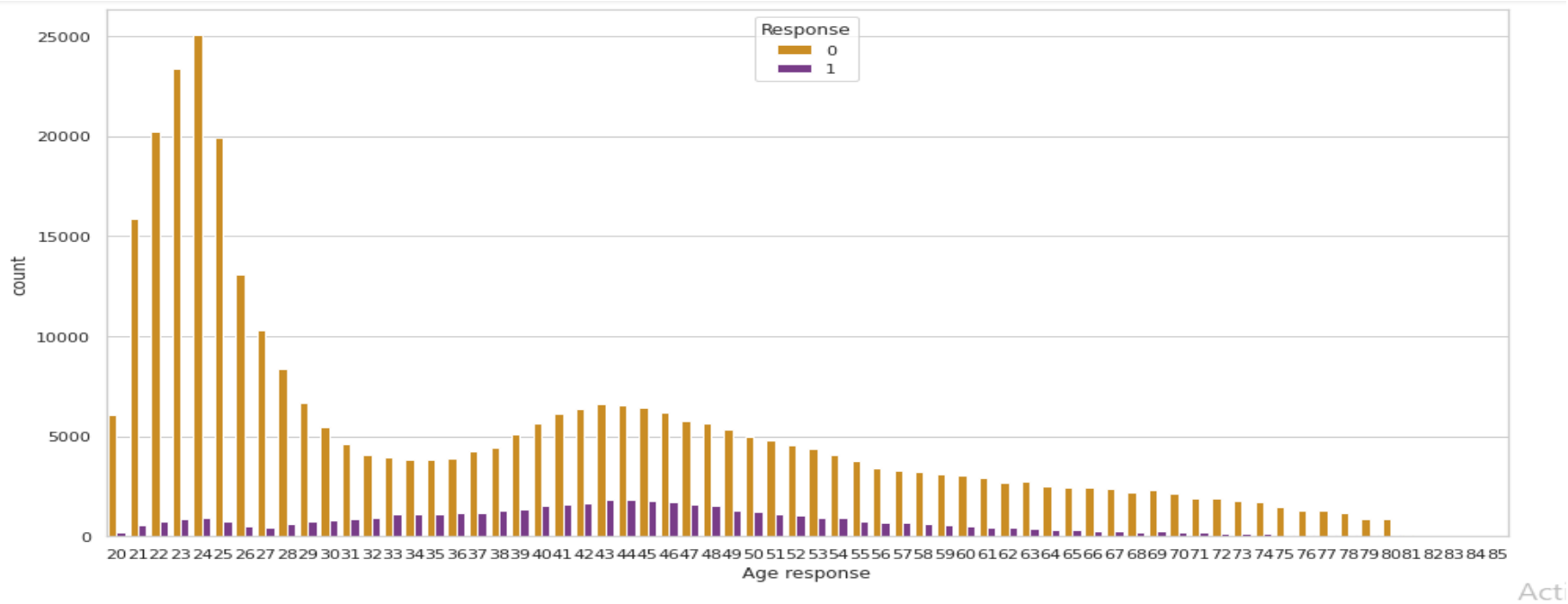


- Male category is slightly greater than that of female and chances of buying the insurance is also little high.



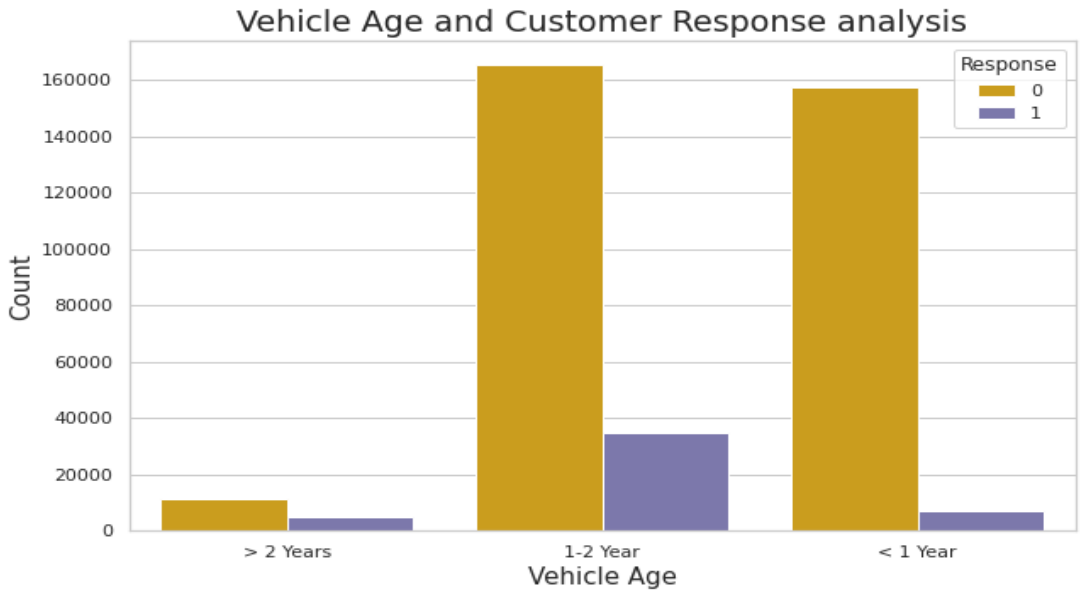
- 54% customer are previously insured ahe 46% customer are are not insured yet.
- Customer who are not perviosly insured are likely to be inetrested

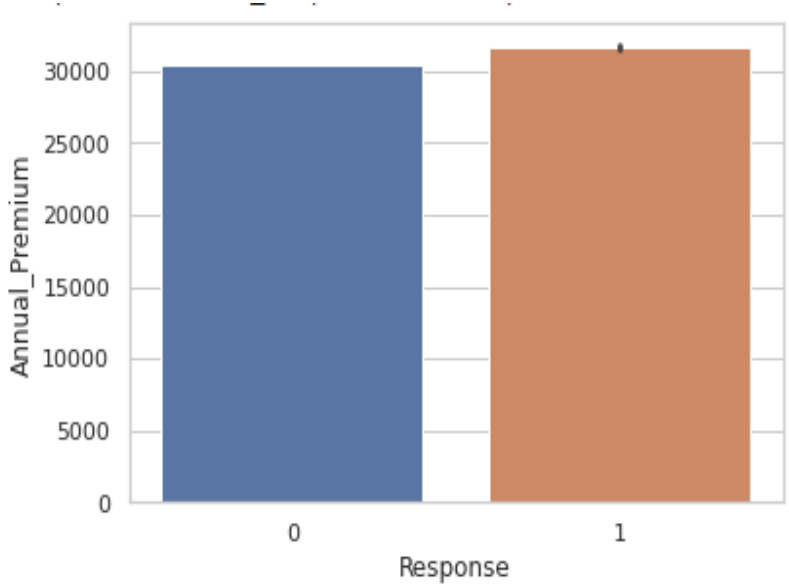
Bivariate analysis



- People ages between from 31 to 50 are more likely to respond.
- while Young people below 30 are not interested in vehicle insurance.

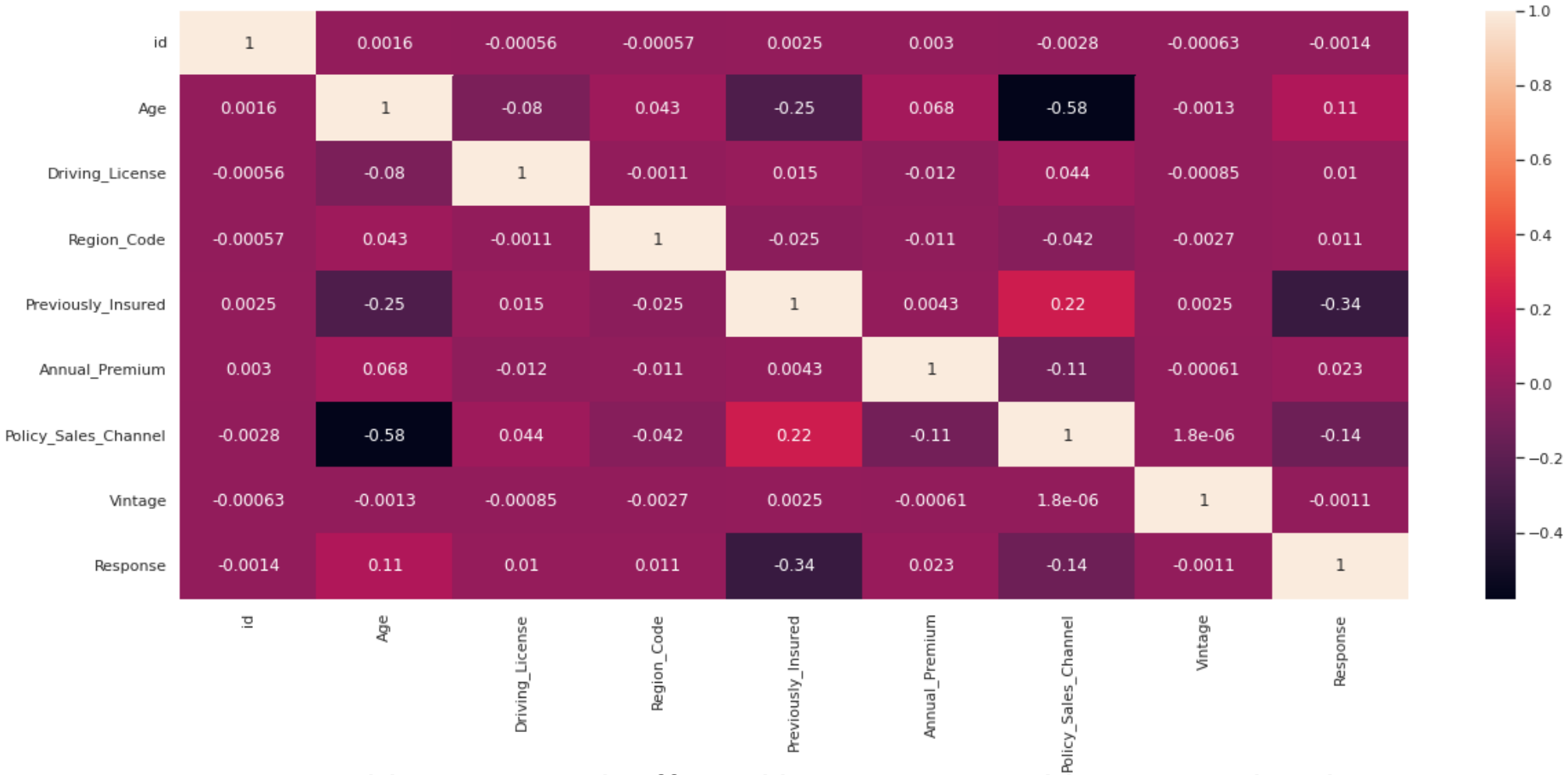
Bivariate analysis





- 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

Correlation



- Target variable is not much affected by Vintage variable. we can drop least correlated variable

Model Building

From the above data analysis we clearly saw that there is a huge difference between the data set.
Standard ML techniques such as Decision Tree and Logistic Regression have a bias towards the majority class, and they tend to ignore the minority class. So solving this issue we use Random Over Sampling technique.

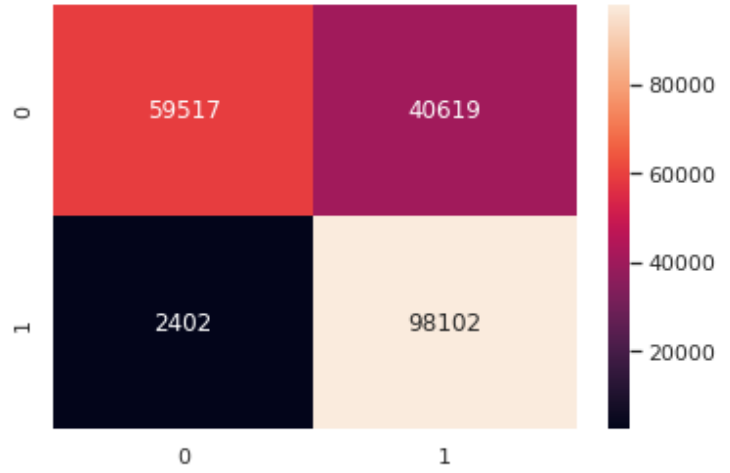
After Random Over Sampling Of Minor Class Total Samples are : 668798
Original dataset shape Counter({0: 334399, 1: 46710})
Resampled dataset shape Counter({1: 334399, 0: 334399})

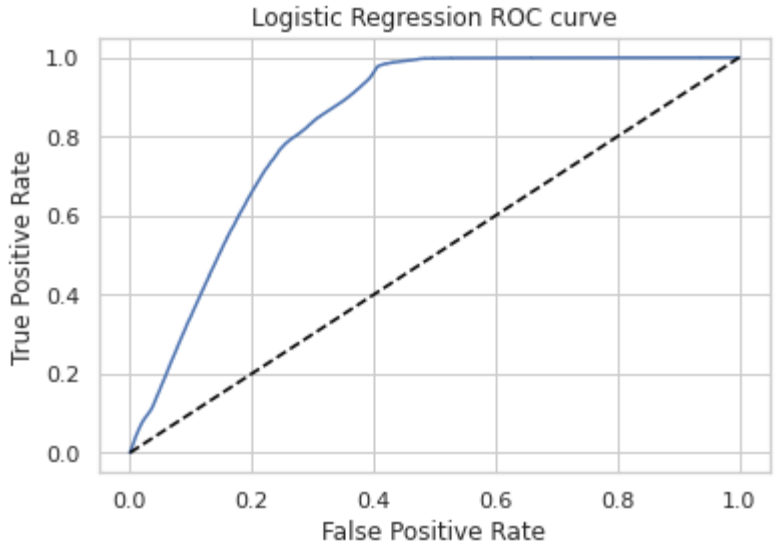
For modeling, we tried the various classification algorithms like:

- Logistic Regression
- RandomForest Classifier
- XGBoost

Logistic Regression

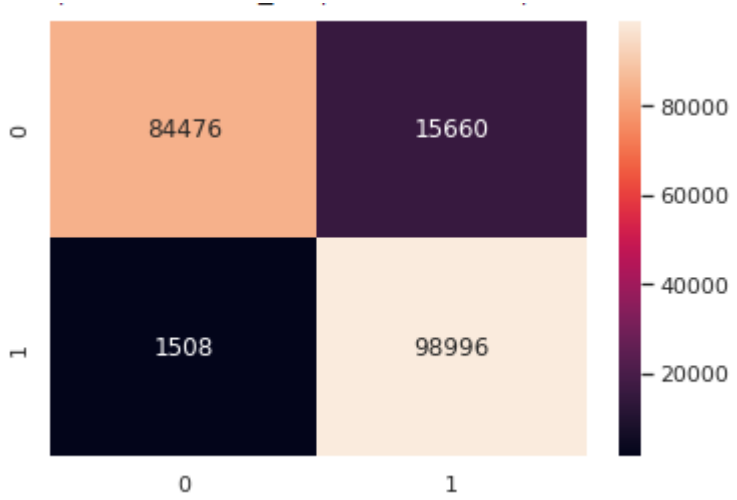
- Logistic regression is named for the function used at the core of the method, the logistic function.
- The logistic function, also called the sigmoid function, was developed by statisticians to describe properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment. It's an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits.
- A way to evaluate the results is by the confusion matrix, which shows the correct and incorrect predictions for each class
- Logistic regression is not performing well on this dataset as shown in confusion matrix model is predicting positive responses but with positive responses it is predicting negative reponses in high numbers too.

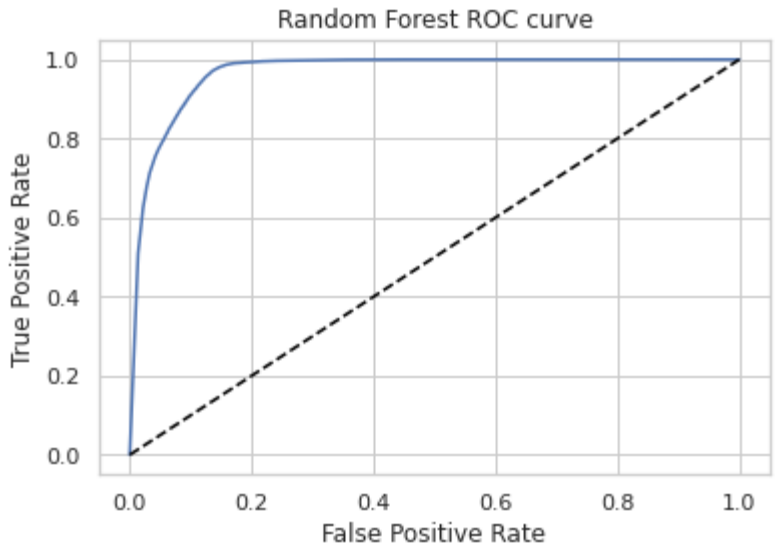




RandomForest Classifier

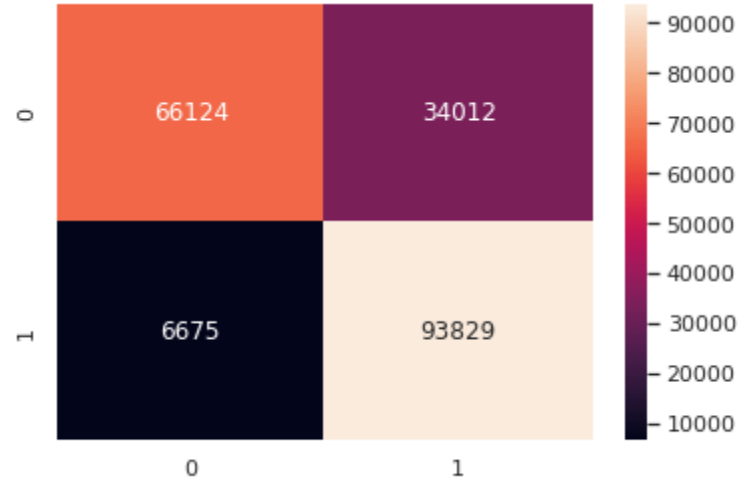
- A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses
- averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is controlled with the
- Parameter if bootsrap=True otherwise whole data set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

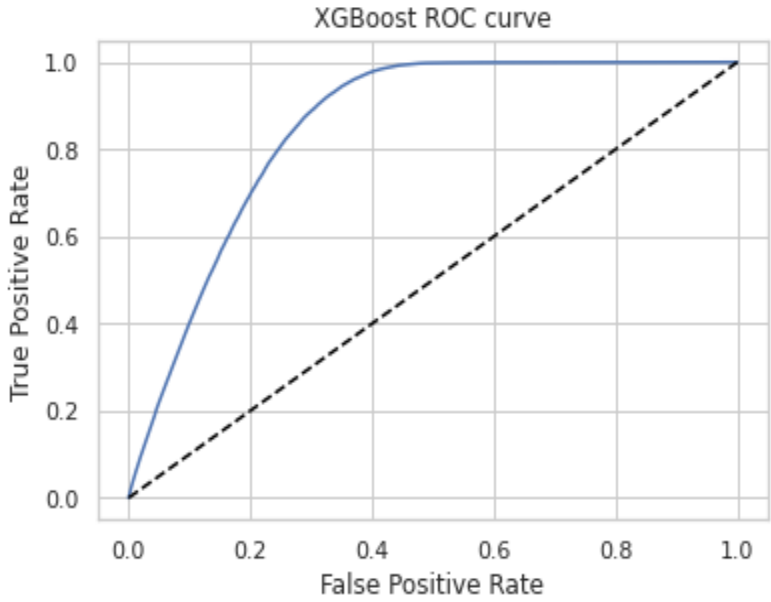




XGBoost

- 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.





XGBoost is just an extension of GBM with following advantages.

- Regularization
- Parallel Processing
- High Flexibility
- Handles Missing values
- Tree pruning
- Buitin cross validation
- Continuous on existing model

Comparing the Model

	Accuracy	Recall	Precision	f1_score	ROC_AUC
Logistic Regression	0.785581	0.976100	0.707189	0.820165	0.834198
Randomforest	0.914434	0.984996	0.863418	0.920208	0.922940
XGBClassifier	0.797214	0.933585	0.733951	0.821818	0.821130

Conclusion:

- 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 afecting the target variable.
- 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 intere sted in Vehicle Insurance.For the logistic regression we got an accuracy of 78% and for the XGBClas sifier we got the aacuracy of 79% whereas,.We are getting the highest accuracy of about 91% and R OC_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.

Thank you !