EDA CAPSTONE PROJECT

TOPIC -: Health Insurance Cross Sell Prediction

TEAM MEMBERS

VISHAKHA KUMARI - ABHINAV AKOTKAR - SHIVAM KUMAR

Content

Introduction

1. Prepare the problem
2. Summarize Data
3. Data visualizations
4. Correlation Matrix
5. Prepare Data
6. Feature Selection
7. Handling Imbalanced data
8. Model Selection
   1. Logistic Regression
   2. Random Forest
   3. XGBClassifier
9. Comparing the model
10. Conclusion

Introduction

- **Predict Health Insurance Owners' who will be interested in Vehicle Insurance**

Your 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 customers from past year will also be interested in Vehicle Insurance provided by the company.

- **Objective**

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

The data given to us for Health Insurance

Cross Sell Prediction was contains 12 column

each column showing some aspect and

381109 rows.

SHAPE OF DATA

Range Index: 381109 entries,

0 to 381108 Data

columns (total 12 columns)

id
Gender
Age
Driving License
Region Code
Previously Insured
Vehicle Age
Vehicle Damage
Annual Premium
Policy Sales Channel
Vintage
Response

Summarize Data

check datatypes, shape,null values

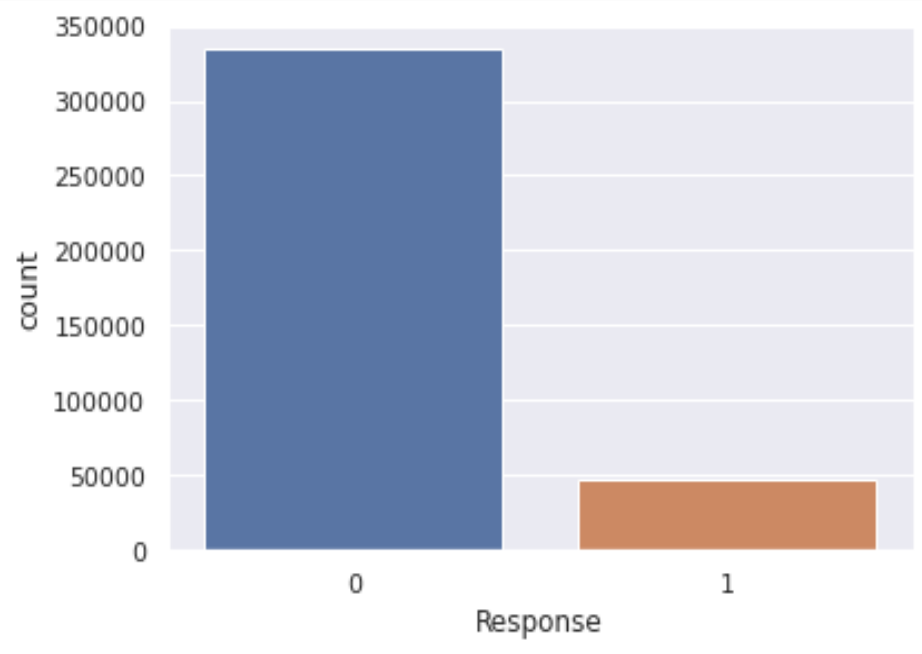
Column	Non-Null Count	Dtype
id	381109 non-null	int64
Gender	381109 non-null	object
Age	381109 non-null	int64
Driving License	381109 non-null	int64
Region Code	381109 non-null	float64
Previously insured	381109 non-null	int64
Vehicle Age	381109 non-null	object
Vehicle Damage	381109 non-null	object
Annual Premium	381109 non-null	float64
Policy Sales Channel	381109 non-null	float64
Vintage	381109 non-null	int64
Response	381109 non-null	int64

id	0
Gender	0
Age	0
Driving License	0
Region Code	0
Previously insured	0
Vehicle Age	0
Vehicle Damage	0
Annual Premium	0
Policy Sales Channel	0
Vintage	0
Response	0

Data visualizations

<matplotlib.axes._subplots.AxesSubplot at 0x7f85e67d4b10>

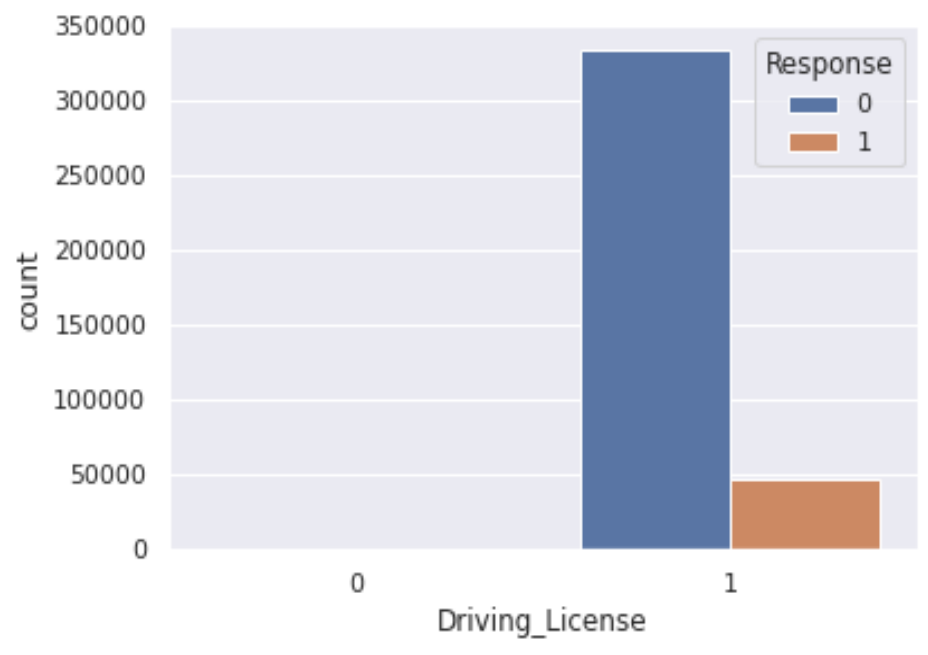
Response



The data is highly imbalanced

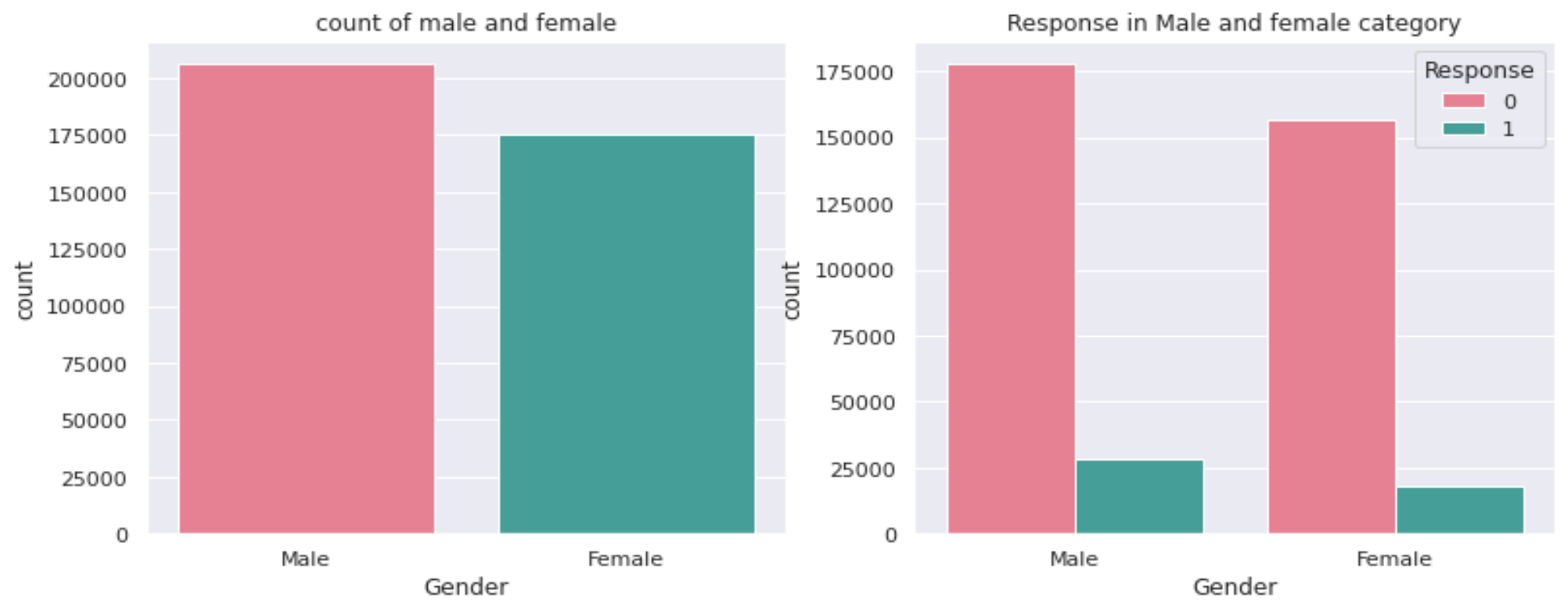
<matplotlib.axes._subplots.AxesSubplot at 0x7f85e5654cd0>

Driving License



Customers who are interested in Vehicle Insurance almost all have driving license

Data visualizations on Gender

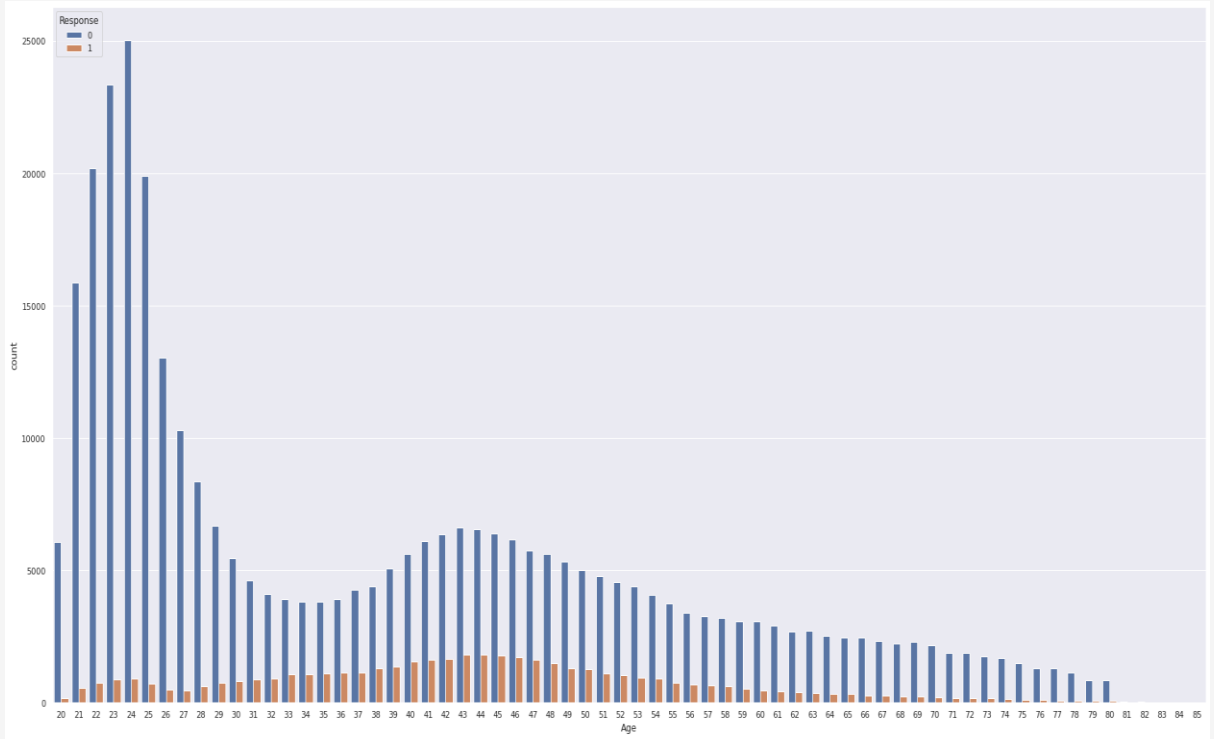


The gender variable in the dataset is almost equally distributed.

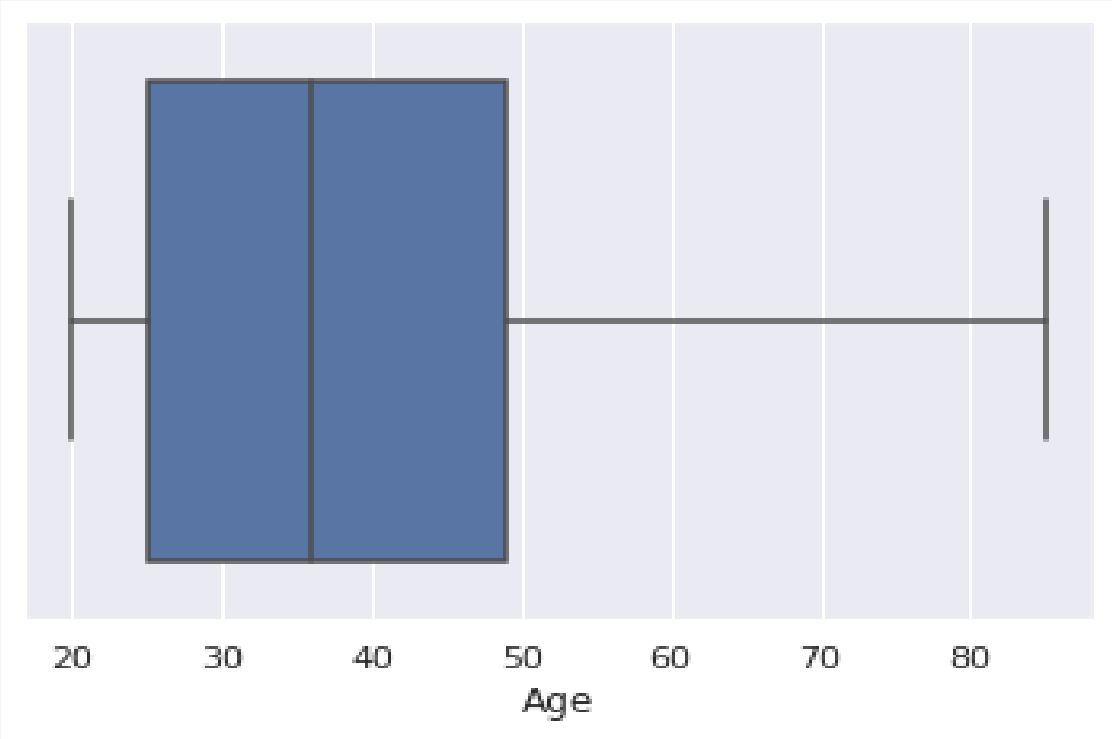
Male category is slightly greater than that of female and chances of buying the insurance is also little high.

Data visualizations between Age & Response

<matplotlib.axes._subplots.AxesSubplot at 0x7f85e5f5c9d0>



<matplotlib.axes._subplots.AxesSubplot at 0x7f85e56af1d0>



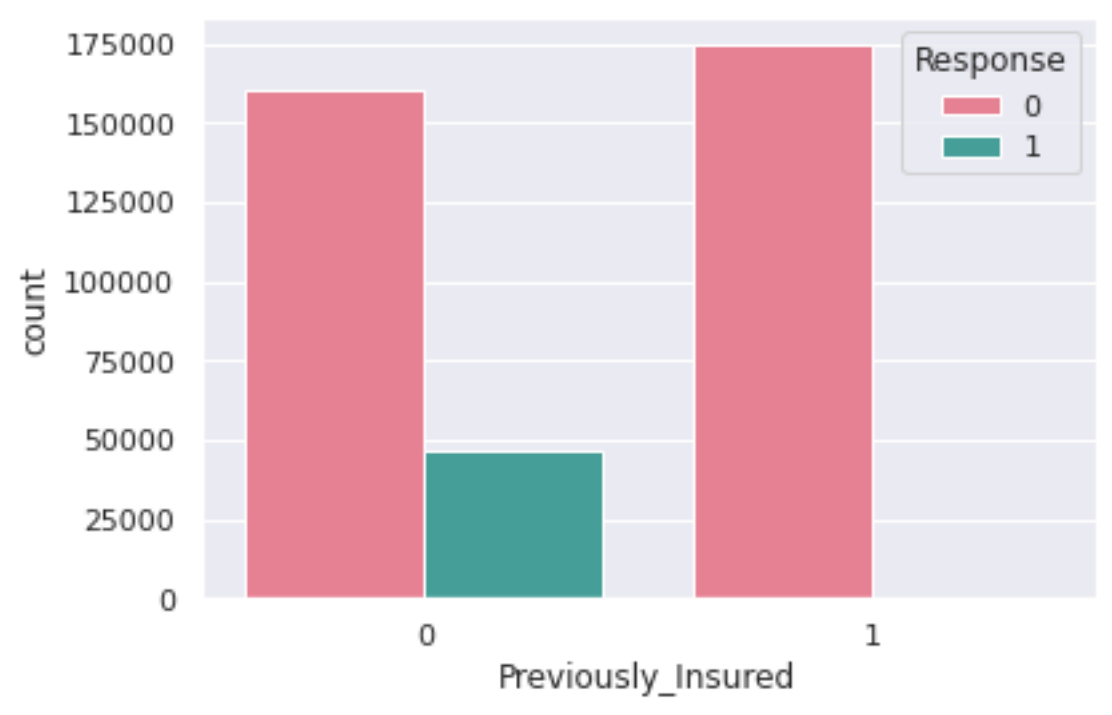
>>Young people below 30 are not interested in vehicle insurance. Reasons could be lack of experience, less maturity level and they don't have expensive vehicles yet.

>>People aged between 30-60 are more likely to be interested.

>>From the boxplot we can see that there no outlier in the data.

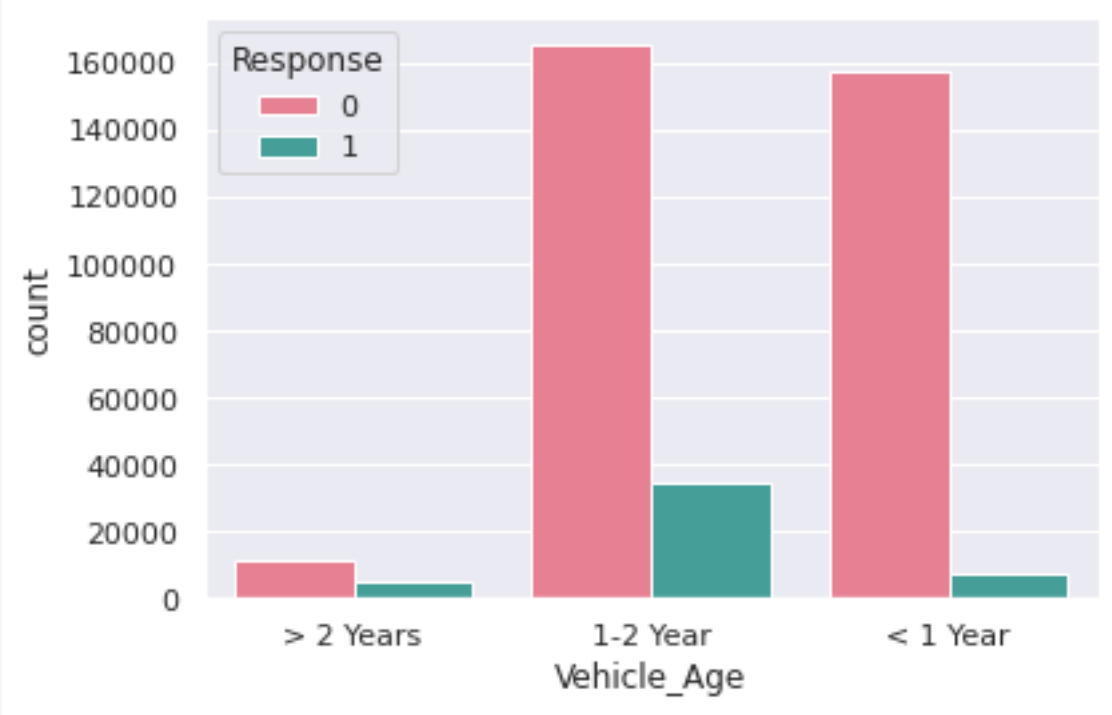
Data visualizations

Previously Insured & Response



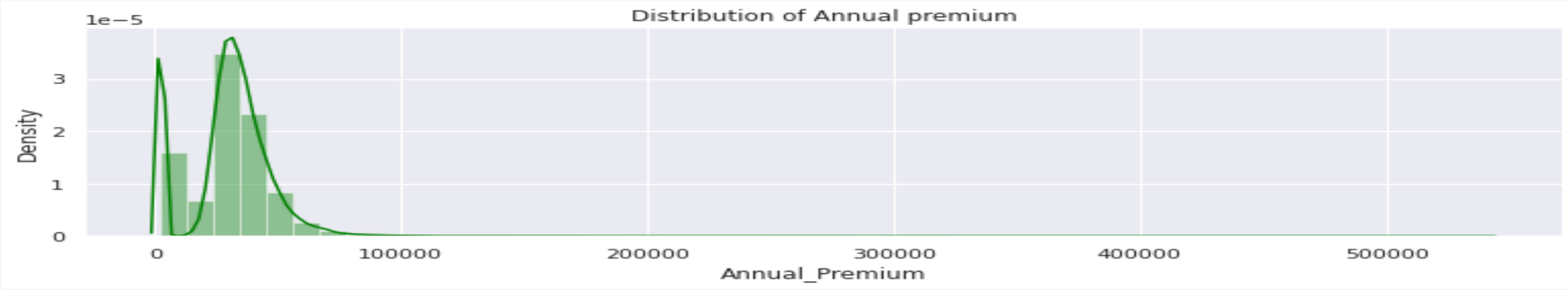
Customer who are not previously insured are likely to be interested.

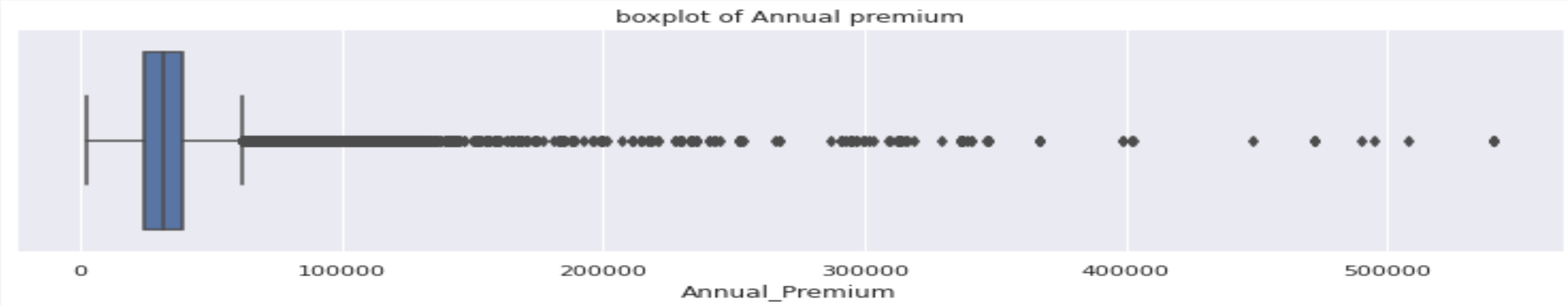
Vehicle Age Vs Response



- Customers with vehicle age 1-2 years are more likely to interested as compared to the other two
- Customers with Vehicle Age <1 years have very less chance of buying Insurance

Data visualizations on Annual Premium

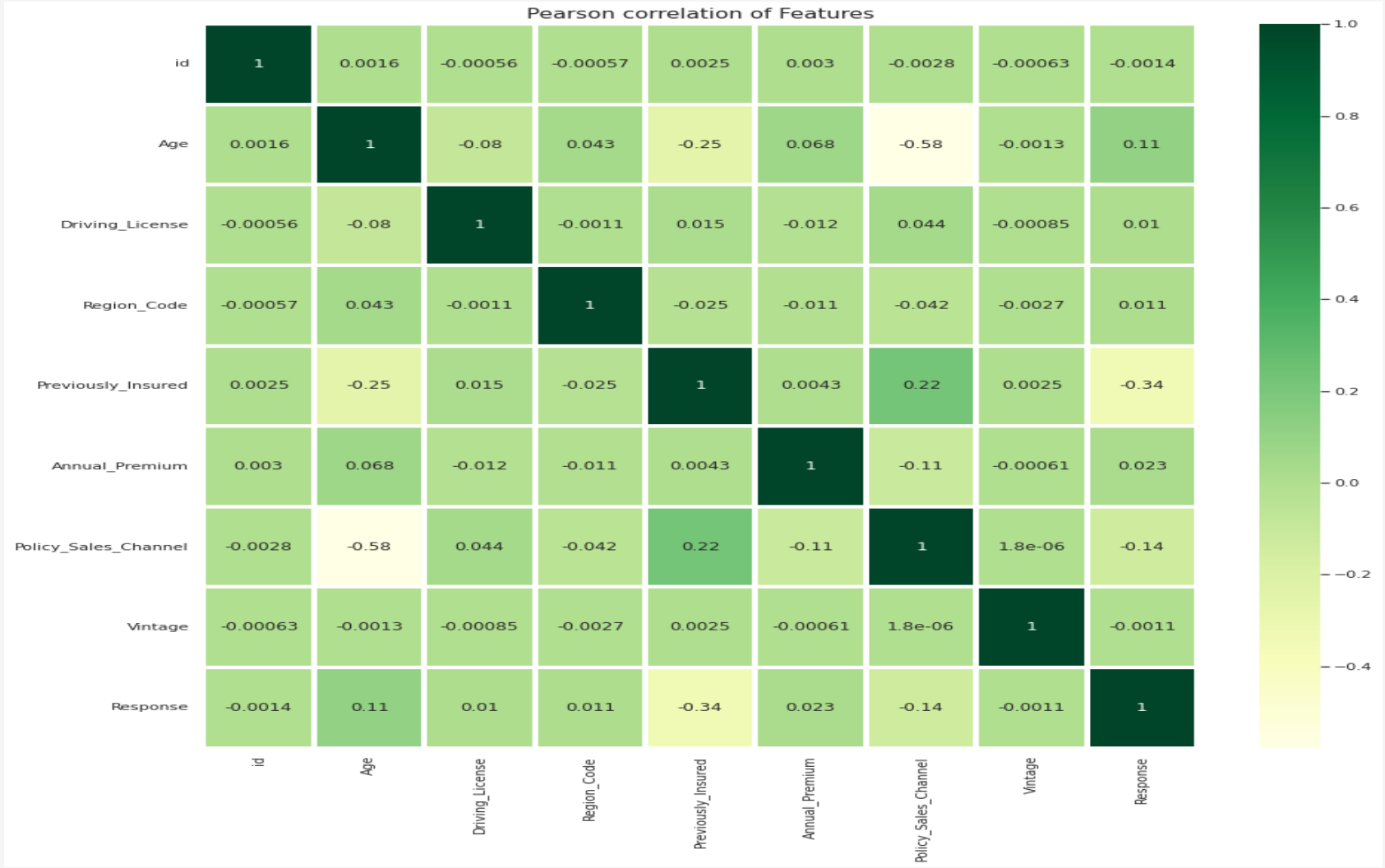




From the distribution plot we can infer that the annual premimum variable is right skewed *From the boxplot we can observe lot of outliers in the variable

Correlation Matrix

Text(0.5, 1.05, 'Pearson correlation of Features')



Target variable is

not much affected

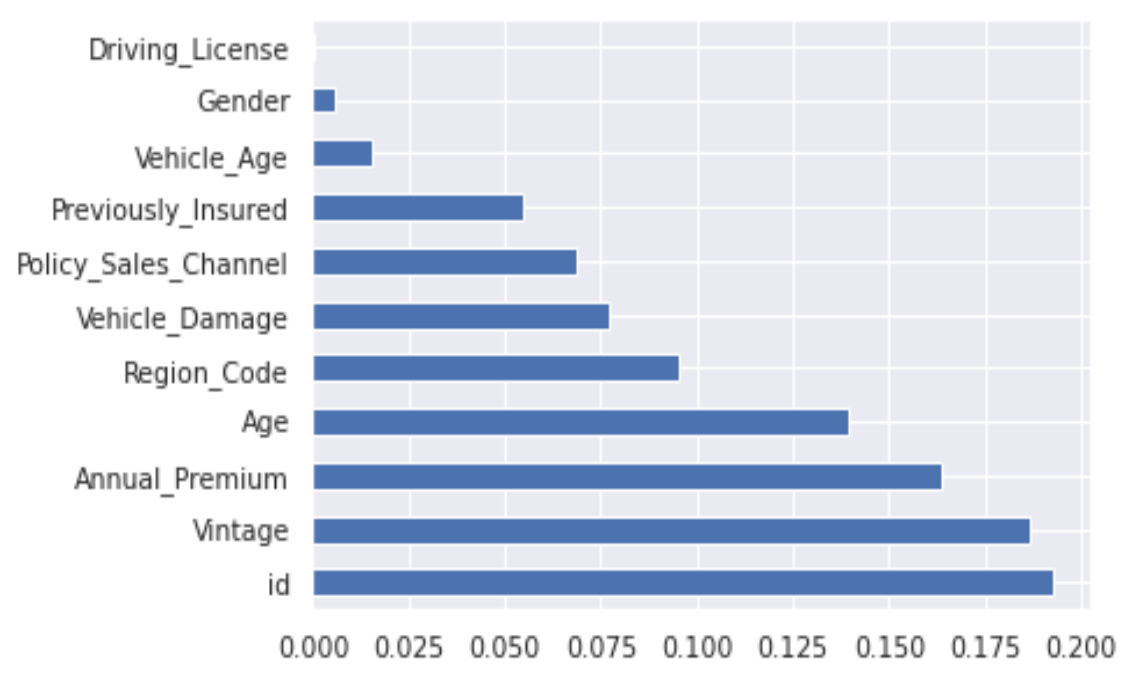
by Vintage variable.

we can drop least

correlated variable.

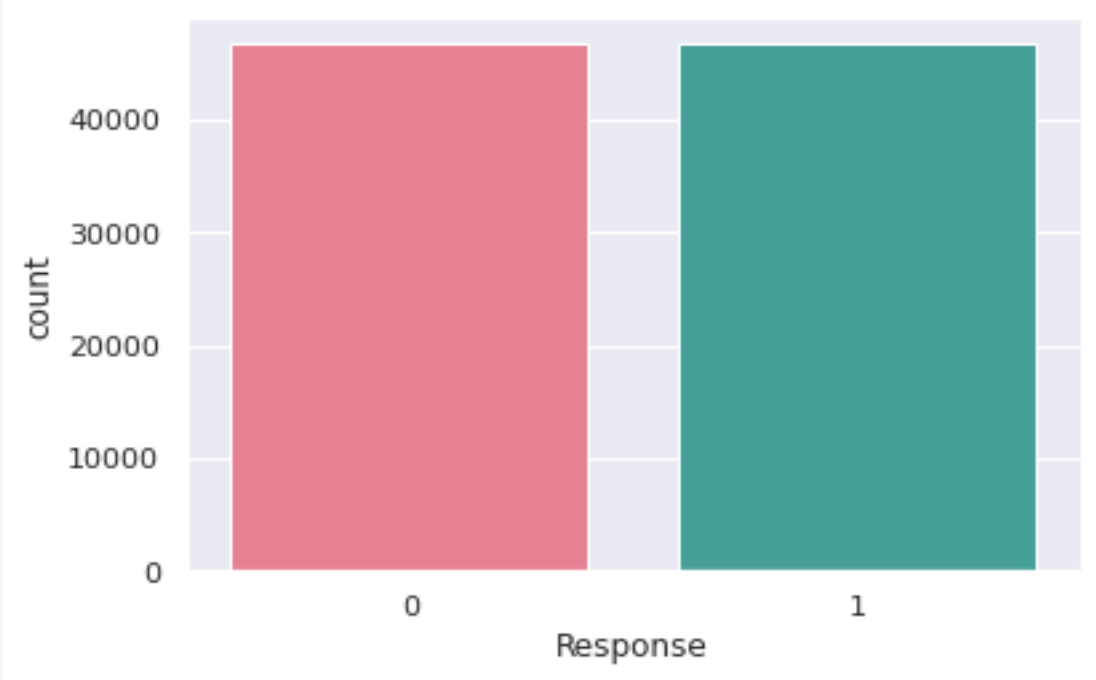
Feature Selection & Handling Imbalance Data

Feature Selection



We can remove less important features from the data set

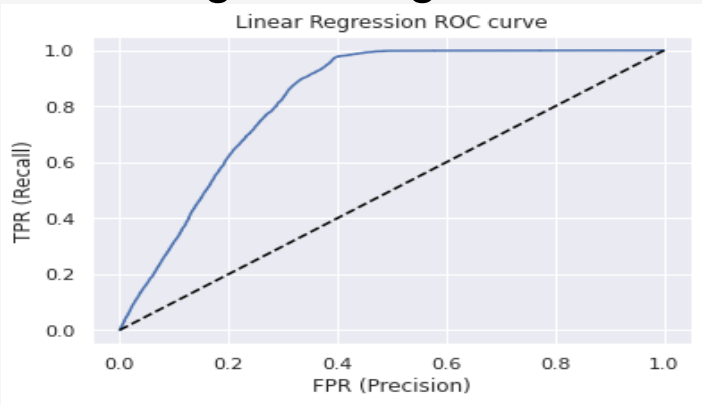
Handling Imbalanced data

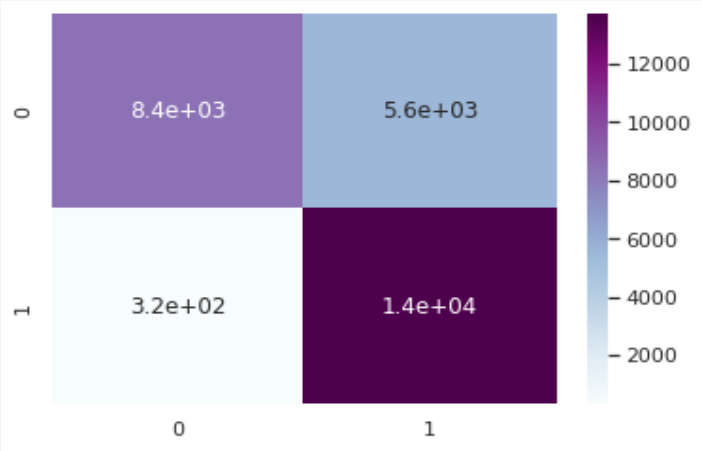


When observation in one class is higher than the observation in other classes then there exists a class imbalance. We can clearly see that there is a huge difference between the data set. Solving this issue we use resampling technique.

Model Selection

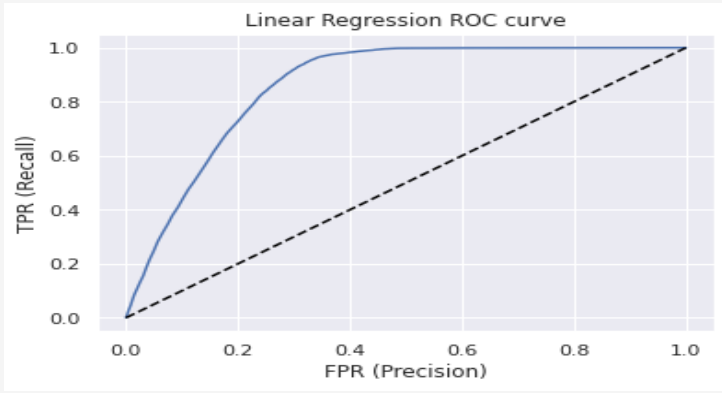
Logistic Regression

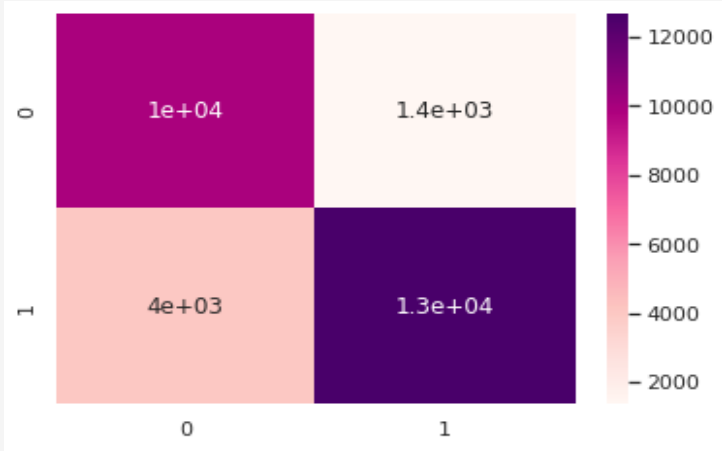




Accuracy : 0.7894098337258261
ROC_AUC Score: 0.8369216600457233

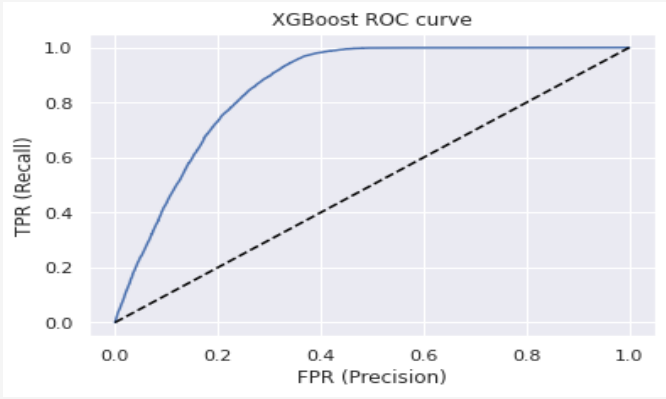
Random Forest Classifier

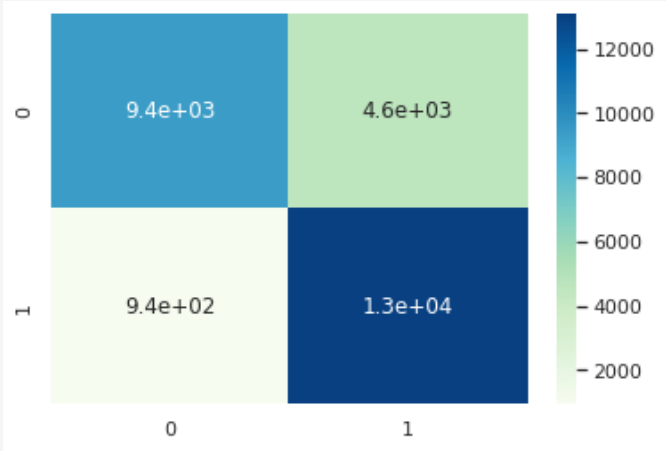




Accuracy : 0.8062156568900307
ROC_AUC Score: 0.8173930079254514

XGBClassifie





Accuracy : 0.8018982373510312
ROC_AUCScore:0.8242774302254495

Comparing the model

	Accuracy	Recall	Prec sion	f1_score	ROC_AUC
Logistic regression	0.789410	0 .976931	0.710969	0.822997	0.836922
Randomforest	0.807893	0.903026	0.759234	0.824911	0.819463
XGBClassifier	0.801898	0.933143	0.739658	0.825211	0.824277

The ML model for the problem statement was created using python with the help of the dataset, and the ML model created with Random Forest and XGBClassifier models performed better than Logistics Regression model. Thus, for the given problem, the models created by Random Forest and XGBClassifier.


Thank You