Capstone Project

Health Insurance Cross Sell Prediction

By : Mohd Atif Ansari

Health Insurance Cross Sell prediction

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.

The data is about an Insurance company that has provided Health Insurance to its customers in past year and is now interested in providing Vehicle Insurance to its policy holders.

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

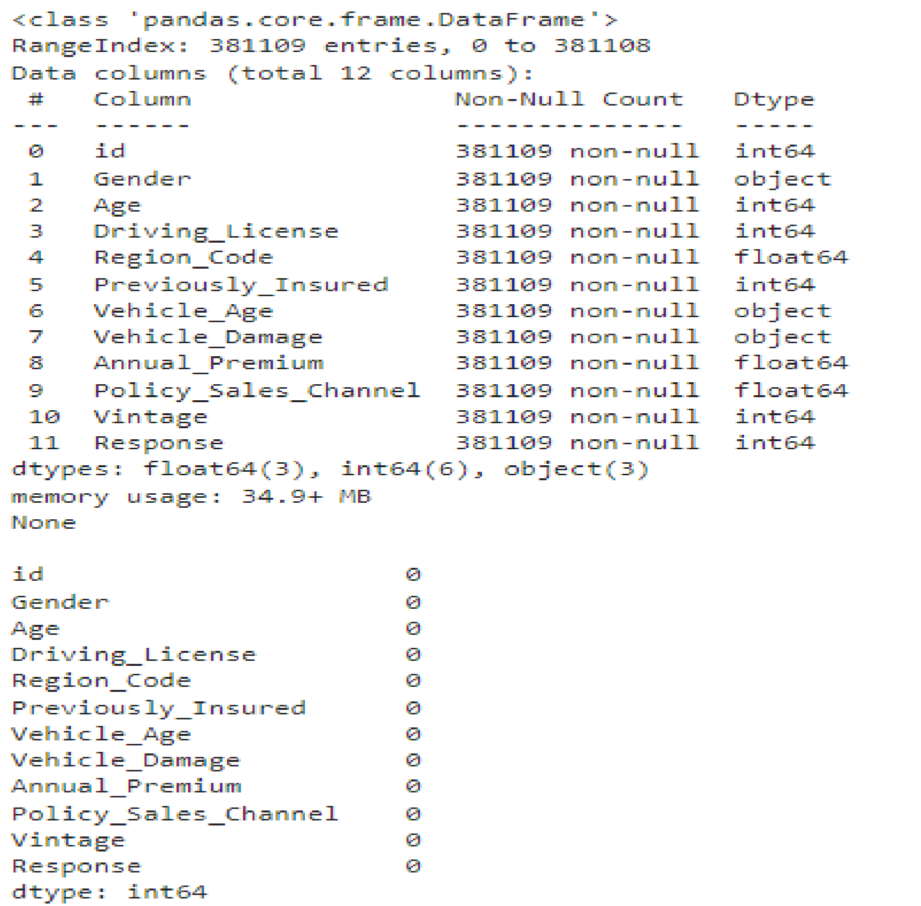
Attribute Information:

- id : Unique ID for the customer
- Gender : Gender of the customer
- Age : Age of the customer
- Driving_License 0 : Customer does not have DL, 1 : Customer already has DL
- Region_Code : Unique code for the region of the customer
- Previously_Insured : 1 : Customer already has Vehicle Insurance, 0 : Customer doesn't have Vehicle Insurance
- Vehicle_Age : Age of the Vehicle
- Vehicle_Damage :1 : Customer got his/her vehicle damaged in the past. 0 : Customer didn't get his/her vehicle damaged in the past.
- Annual_Premium : The amount customer needs to pay as premium in the year
- PolicySalesChannel : Anonymized Code for the channel of outreaching to the customer ie. Different Agents, Over Mail, Over Phone, In Person, etc.
- Vintage : Number of Days, Customer has been associated with the company
- Response : 1 : Customer is interested, 0 : Customer is not interested

Objective:

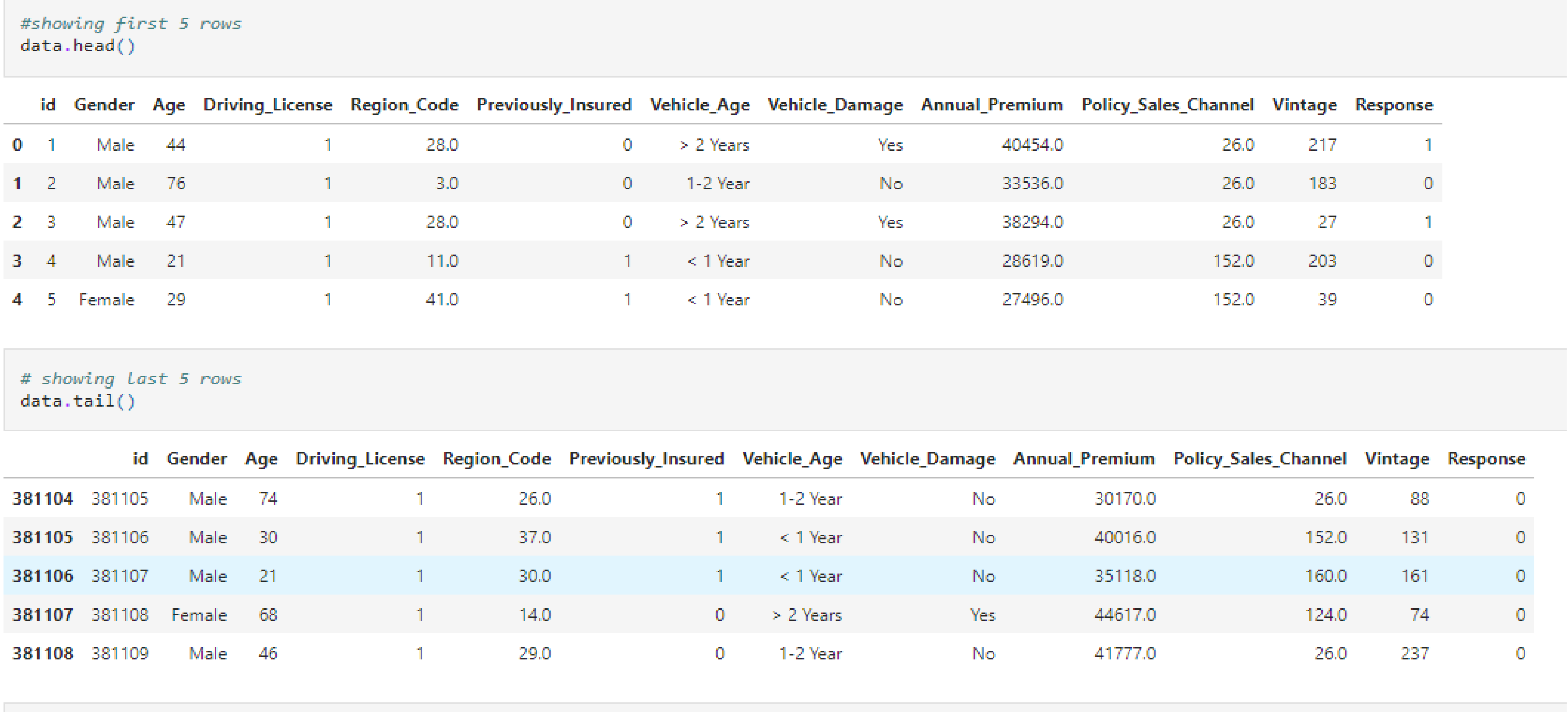
- Understand what is Cross-sell using Vehicle insurance data.
- Learn how to build a model for cross-sell prediction.
- To predict if an insurance policy holder would be interested to buy a vehicle insurance as well. Building a model to predict whether a customer would be interested in Vehicle Insurance 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.
- The aim of this project is to leverage the machine learning algorithms such as Logistic Regression and Random Forest to create a predictive model using statistically significant variables from the given data set.
- Model accuracy will be assessed using different techniques such as ROC (Receiver operating characteristic), AUC (Area under the ROC curve) and Confusion Matrix.

Datatype:



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 381109 entries, 0 to 381108
Data columns (total 12 columns):
 #   Column                Non-Null Count   Dtype  
---  ------                --------------   -----  
 0   id                    381109 non-null  int64  
 1   Gender                381109 non-null  object 
 2   Age                   381109 non-null  int64  
 3   Driving_License       381109 non-null  int64  
 4   Region_Code           381109 non-null  float64
 5   Previously_Insured    381109 non-null  int64  
 6   Vehicle_Age           381109 non-null  object 
 7   Vehicle_Damage        381109 non-null  object 
 8   Annual_Premium        381109 non-null  float64
 9   Policy_Sales_Channel  381109 non-null  float64
 10  Vintage               381109 non-null  int64  
 11  Response              381109 non-null  int64  
dtypes: float64(3), int64(6), object(3)
memory usage: 34.9+ MB
None

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
dtype: int64
```

Sample Data:



```
#showing first 5 rows
data.head()
```

	id	Gender	Age	Driving_License	Region_Code	Previously_Insured	Vehicle_Age	Vehicle_Damage	Annual_Premium	Policy_Sales_Channel	Vintage	Response
0	1	Male	44	1	28.0	0	> 2 Years	Yes	40454.0	26.0	217	1
1	2	Male	76	1	3.0	0	1-2 Year	No	33536.0	26.0	183	0
2	3	Male	47	1	28.0	0	> 2 Years	Yes	38294.0	26.0	27	1
3	4	Male	21	1	11.0	1	< 1 Year	No	28619.0	152.0	203	0
4	5	Female	29	1	41.0	1	< 1 Year	No	27496.0	152.0	39	0

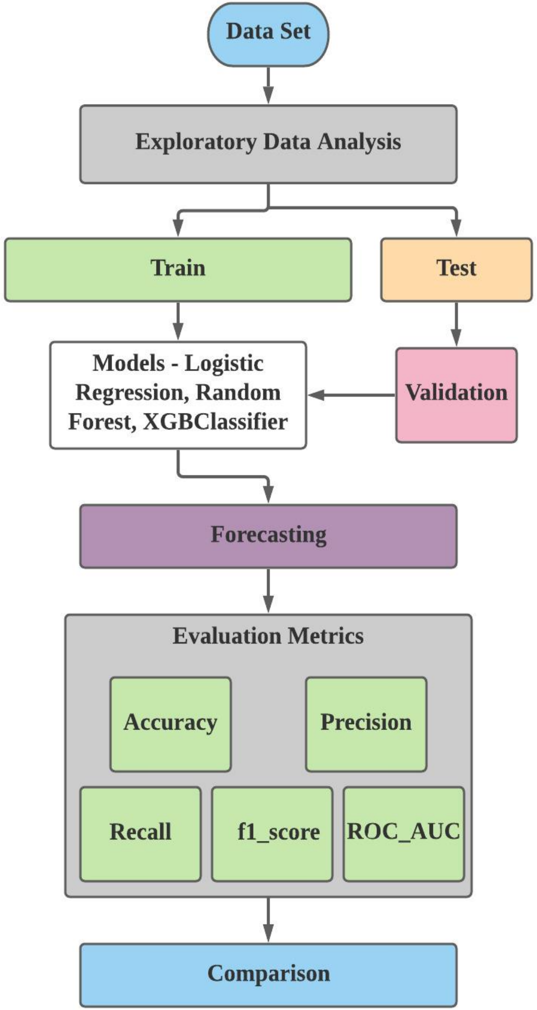
```
# showing last 5 rows
data.tail()
```

	id	Gender	Age	Driving_License	Region_Code	Previously_Insured	Vehicle_Age	Vehicle_Damage	Annual_Premium	Policy_Sales_Channel	Vintage	Response
381104	381105	Male	74	1	26.0	1	1-2 Year	No	30170.0	26.0	88	0
381105	381106	Male	30	1	37.0	1	< 1 Year	No	40016.0	152.0	131	0
381106	381107	Male	21	1	30.0	1	< 1 Year	No	35118.0	160.0	161	0
381107	381108	Female	68	1	14.0	0	> 2 Years	Yes	44617.0	124.0	74	0
381108	381109	Male	46	1	29.0	0	1-2 Year	No	41777.0	26.0	237	0

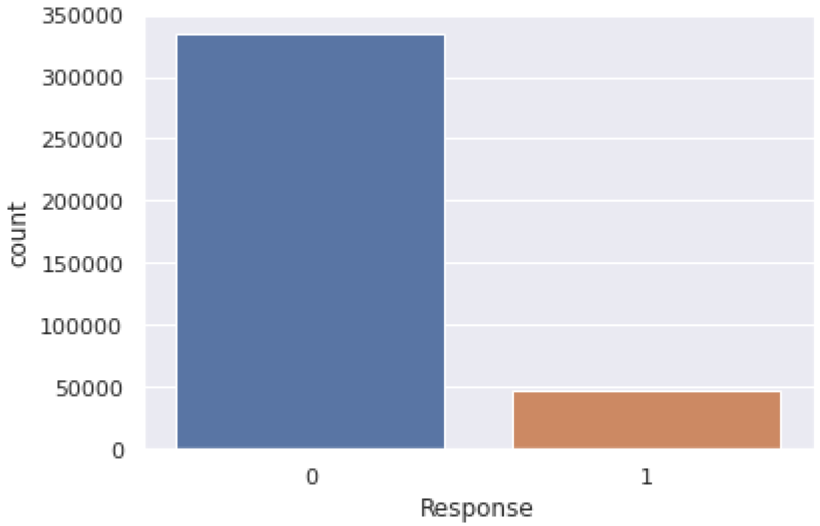
Data Preprocessing & Implementation

- **Data processing-1:** In first part we have to remove unnecessary features. Since there were many column with all null values.
- **Data processing-2:** we have manually go through each features select from part 1, and encoded the categorical features.
- **EDA:** In this part we do some exploratory data analysis (EDA) on the features selected in part-1 and part-2 to see the trend.
- **Split the data:** we have to split the data into two parts train and test.
- **Create the model:** Finally, in the last part but not the last part we creates models and function, and import some libraries it’s not the easy task. Its also an iterative process. We show how to start with simple models and then add complexity for better performance.

Flow Chart:


Data Set
Exploratory Data Analysis
Train
Test
Models - Logistic Regression, Random Forest, XGBClassifier
Validation
Forecasting
Evaluation Metrics
Accuracy
Precision
Recall
f1_score
ROC_AUC
Comparison

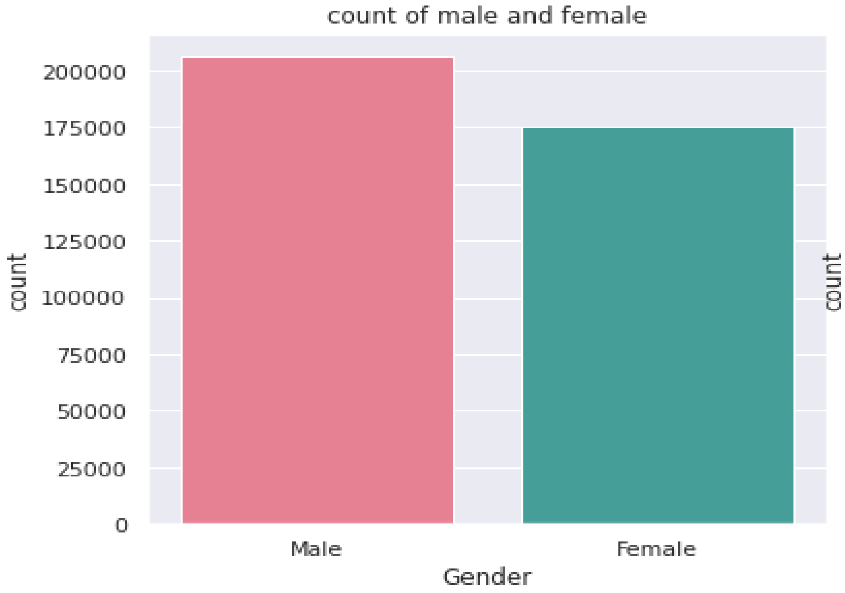
Target variable

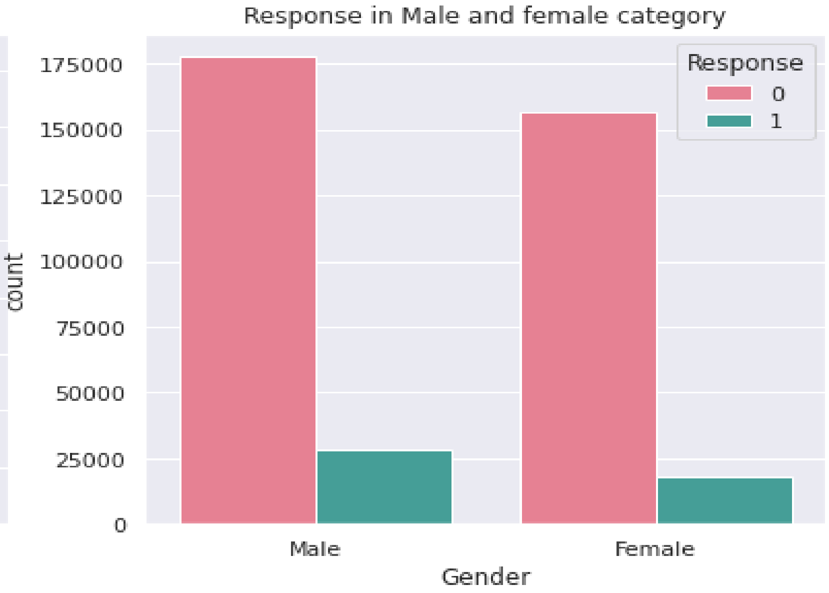


0: indicate Customer is not interested
1: indicate Customer is interested

As you can see on the graph, there are very few interested customers whose stats are less than 50000 and which customers are not interested those is above 300000.

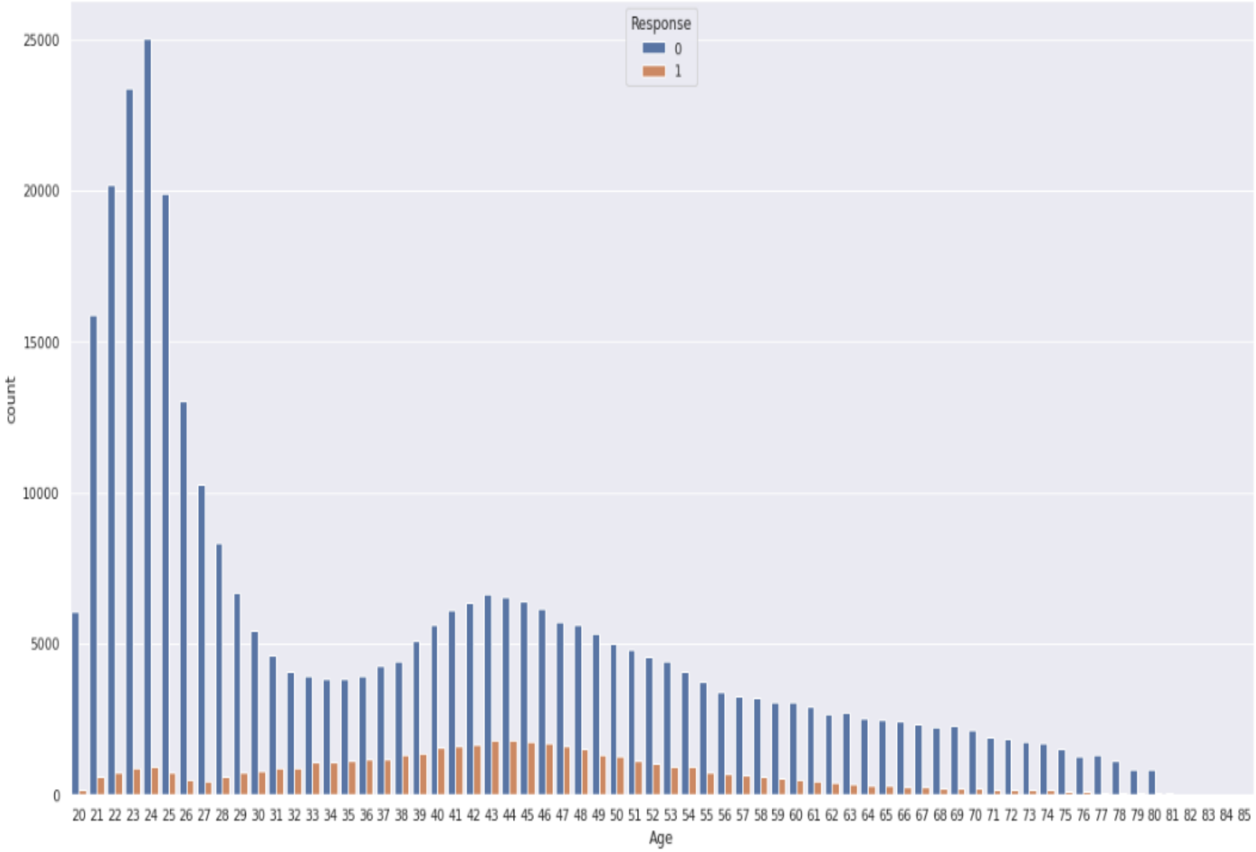
Gender


count of male and female


Response in Male and female category

- As you can see above the graph:
- The number of male is greater than 200000 and The number of female is close to 175000.
- The number of male is interested which is greater than 25000 and The number of female is interested which is below 25000.
- Male category is slightly greater than that of female and chances of buying the insurance is also little high

Age vs response

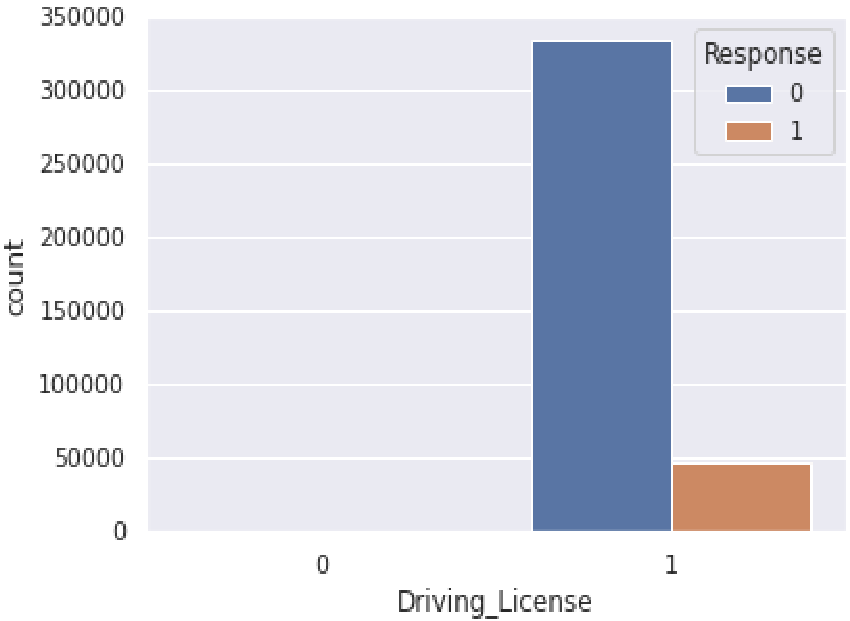


- Young people below 30 are not interested in vehicle insurance. Reasons could be lack of experience, less maturity level and they don't have expensive vehicles yet.

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

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

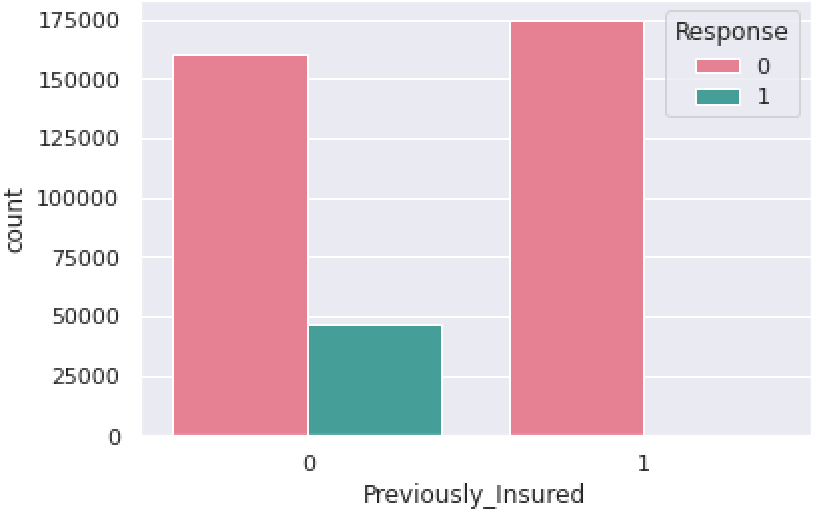
Driving



0: indicate Customer is not interested
1: indicate Customer is interested

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

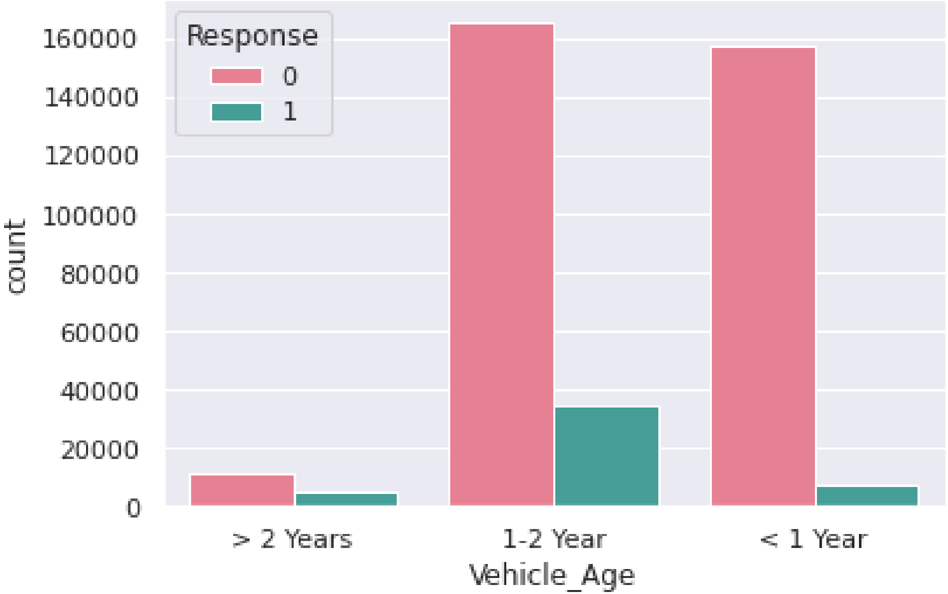
Previously vs response



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

- 1 indicate no one interested and the value of this is null.

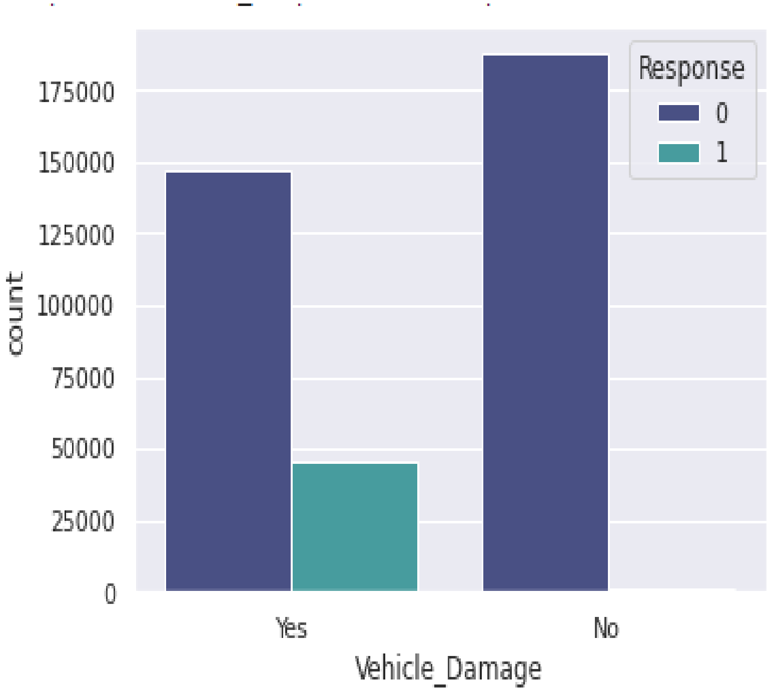
Vehicle Age vs Response



- 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.

Vehicle Damage vs Response

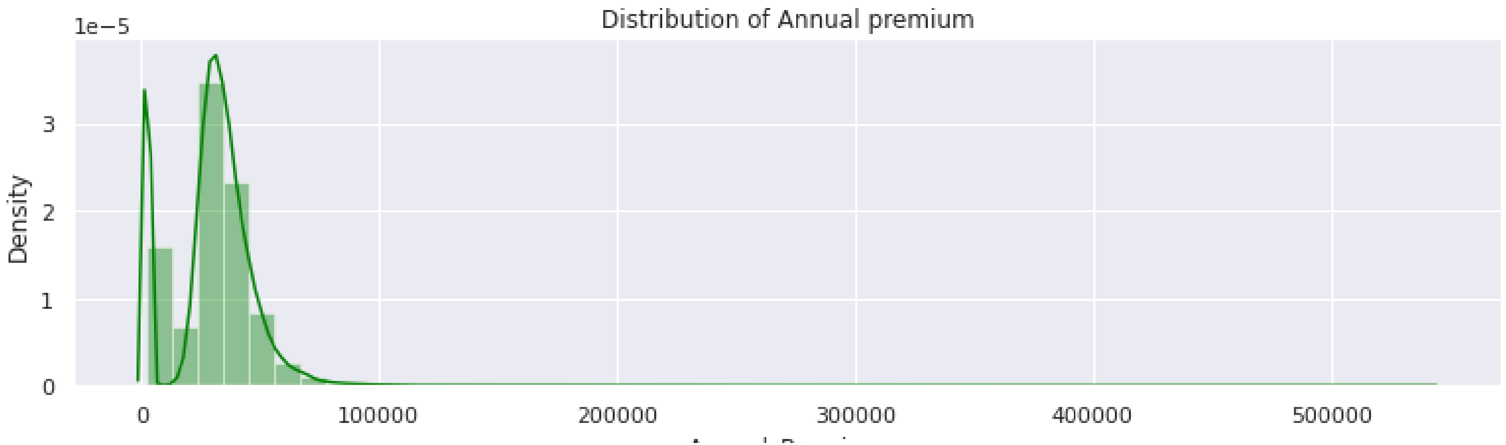


- From the above plot, we can infer that if the vehicle has been damaged previously then the customer will be more interested in buying the insurance as they know the cost.

- It is also important to look at the target column, as it will tell us whether the problem is a balanced problem or an imbalanced problem. This will define our approach further.

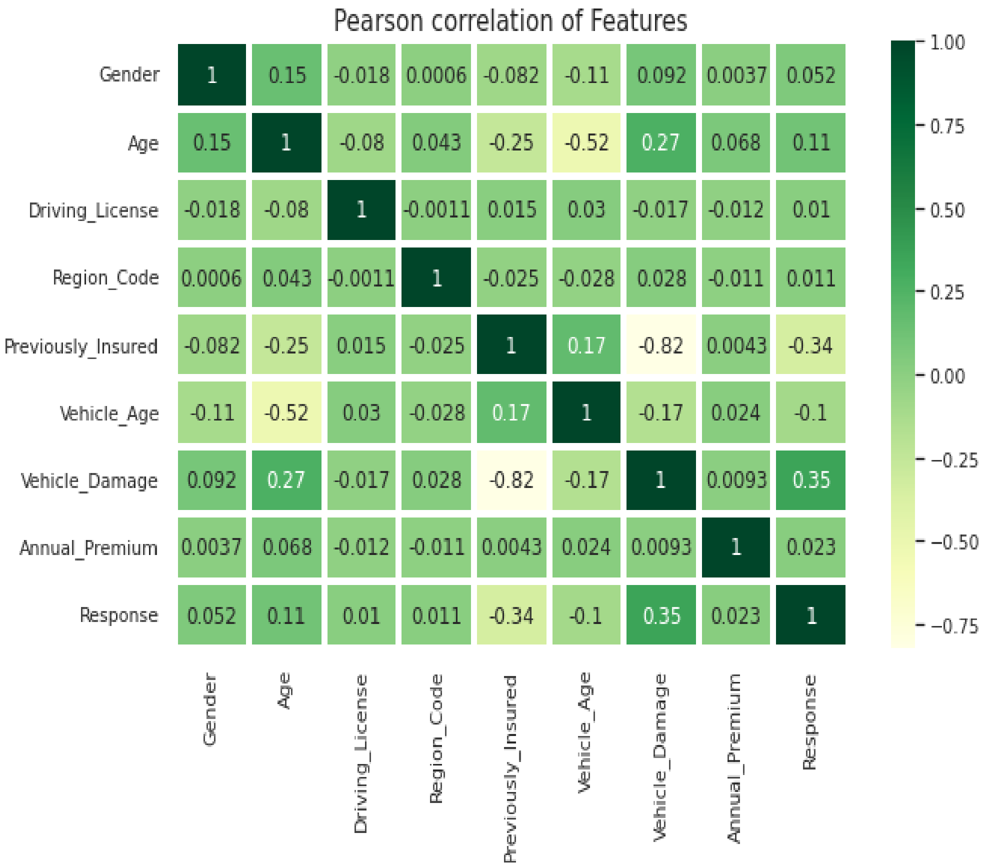
- The given problem is an imbalance problem as the Response variable with the value 1 is significantly lower than the value zero.

Annual



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

Correlation plot:



Pearson correlation of Features

	Gender	Age	Driving_License	Region_Code	Previously_Insured	Vehicle_Age	Vehicle_Damage	Annual_Premium	Response
Gender	1	0.15	-0.018	0.0006	-0.082	-0.11	0.092	0.0037	0.052
Age	0.15	1	-0.08	0.043	-0.25	-0.52	0.27	0.068	0.11
Driving_License	-0.018	-0.08	1	-0.0011	0.015	0.03	-0.017	-0.012	0.01
Region_Code	0.0006	0.043	-0.0011	1	-0.025	-0.028	0.028	-0.011	0.011
Previously_Insured	-0.082	-0.25	0.015	-0.025	1	0.17	-0.82	0.0043	-0.34
Vehicle_Age	-0.11	-0.52	0.03	-0.028	0.17	1	-0.17	0.024	-0.1
Vehicle_Damage	0.092	0.27	-0.017	0.028	-0.82	-0.17	1	0.0093	0.35
Annual_Premium	0.0037	0.068	-0.012	-0.011	0.0043	0.024	0.0093	1	0.023
Response	0.052	0.11	0.01	0.011	-0.34	-0.1	0.35	0.023	1

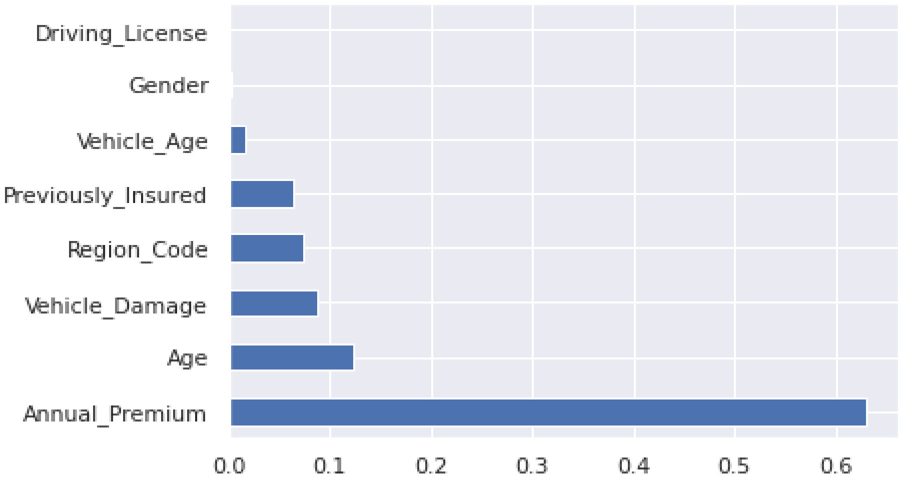
- Heat maps are a great tool for visualizing complex statistical data.

- By this graph we will find out correlation all about the data. Which column is useful and which is not useful for us. We can easily find out by this graph.

- Target variable is not much affected by id , Vintage variable, Policy sales channel. we can drop least correlated variable.

- As we can see the heatmap graph Vehicle Damage column is more correlated with target variable.

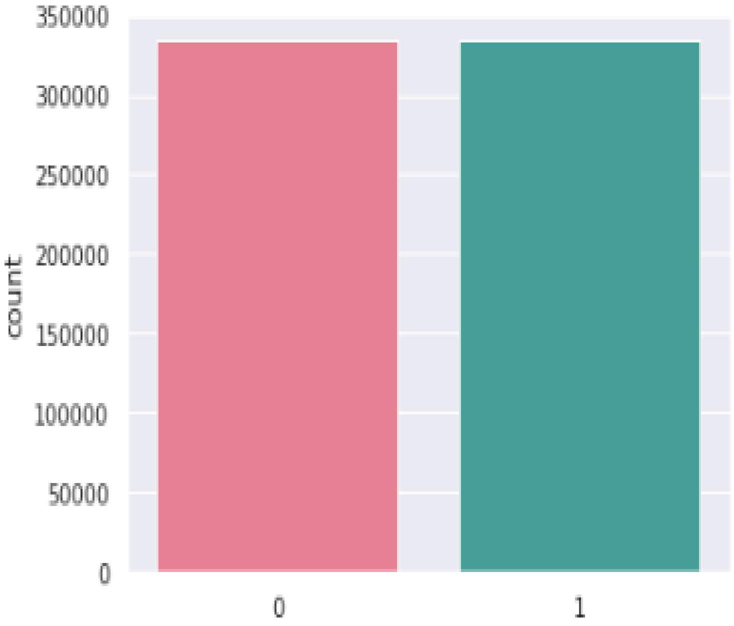
Feature Selection



- The features you use influence more than everything else the result. No algorithm alone, to my knowledge, can supplement the information gain given by correct **feature engineering**.

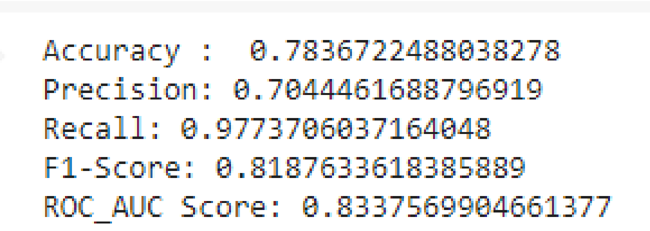
- We can remove less important features from the data set like Driving license, Gender.

Handling imbalance Data

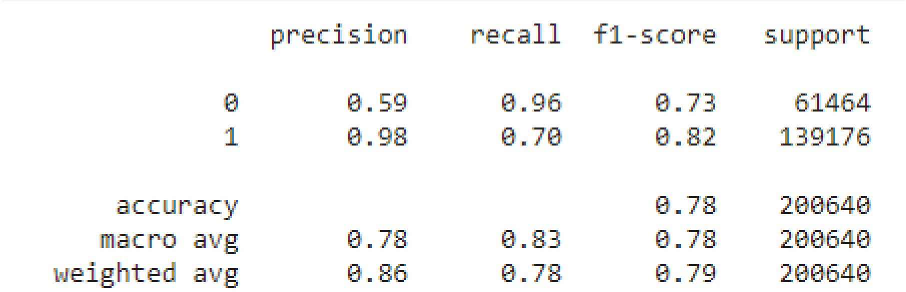


- As from the distribution of target variables in the EDA section, we know it is an imbalance problem. The imbalance datasets could have their own challenge
- For example, a disease prediction model may have an accuracy of 99% but it is of no use if it can not classify a patient successfully.
- So to handle such a problem, we can resample the data. In the following code, we will be using under sampling.
- Under sampling is the method where we will be reducing the occurrence of the majority class up to a given point.

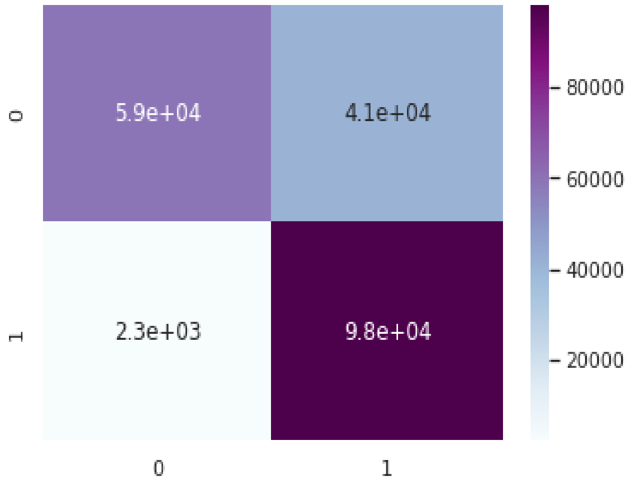
Logistic regression

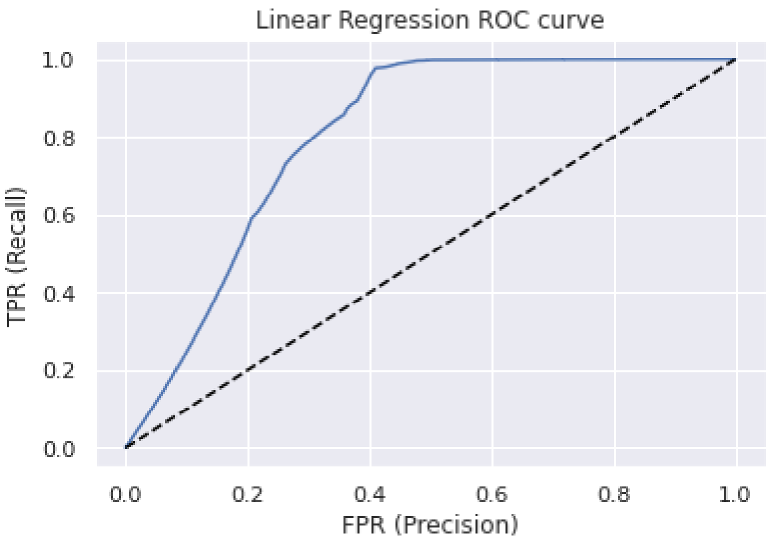


```
Accuracy :  0.7836722488038278
Precision: 0.704461688796919
Recall: 0.9773706037164048
F1-Score: 0.8187633618385889
ROC_AUC Score: 0.8337569904661377
```

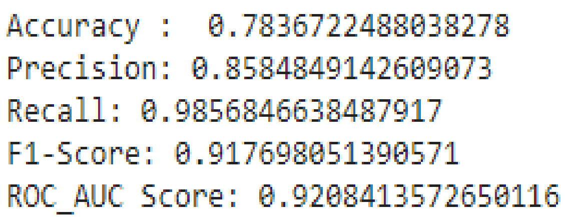


```
              precision    recall  f1-score   support

           0       0.59      0.96      0.73     61464
           1       0.98      0.70      0.82    139176

    accuracy                           0.78    200640
   macro avg       0.78      0.83      0.78    200640
weighted avg       0.86      0.78      0.79    200640
```

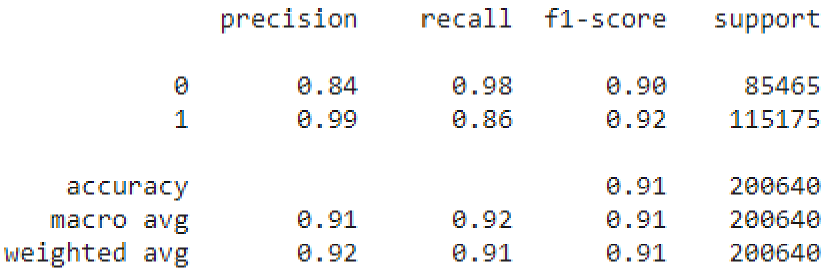




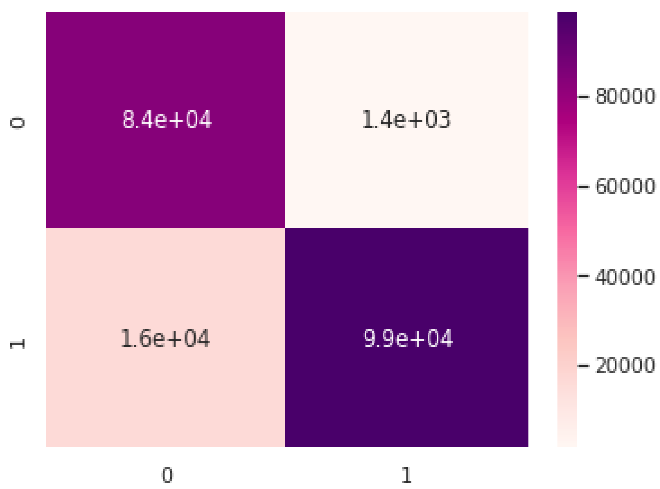
Random forest Classifier

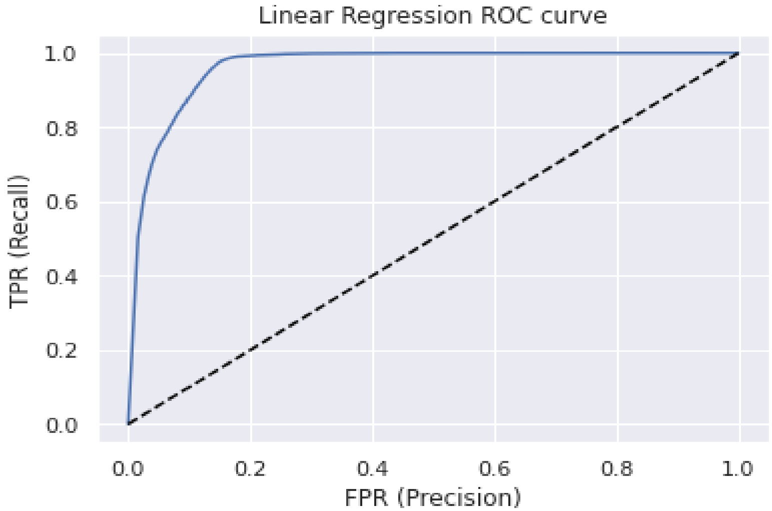


```
Accuracy :  0.7836722488038278
Precision: 0.8584849142609073
Recall: 0.9856846638487917
F1-Score: 0.917698051390571
ROC_AUC Score: 0.9208413572650116
```

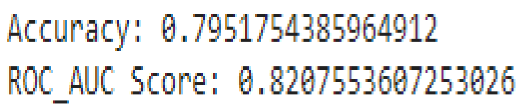


```
              precision    recall  f1-score   support

           0       0.84      0.98      0.90     85465
           1       0.99      0.86      0.92    115175

    accuracy                           0.91    200640
   macro avg       0.91      0.92      0.91    200640
weighted avg       0.92      0.91      0.91    200640
```

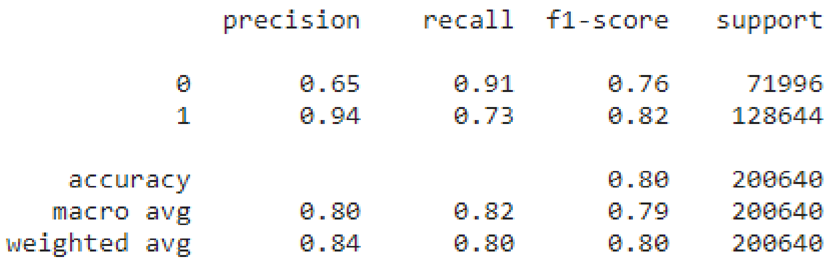




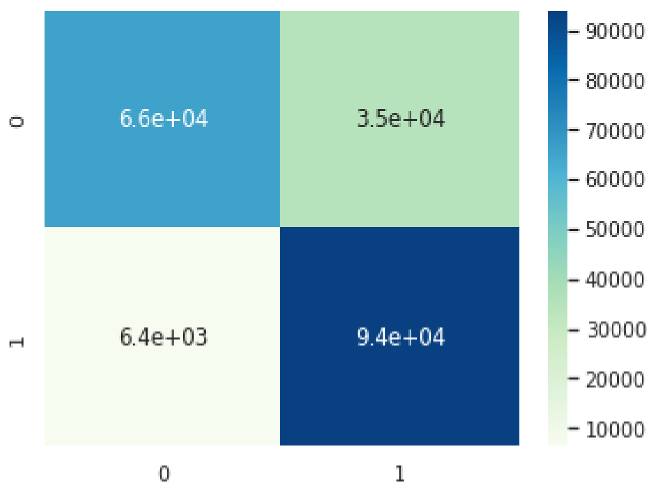
XGB Classifier

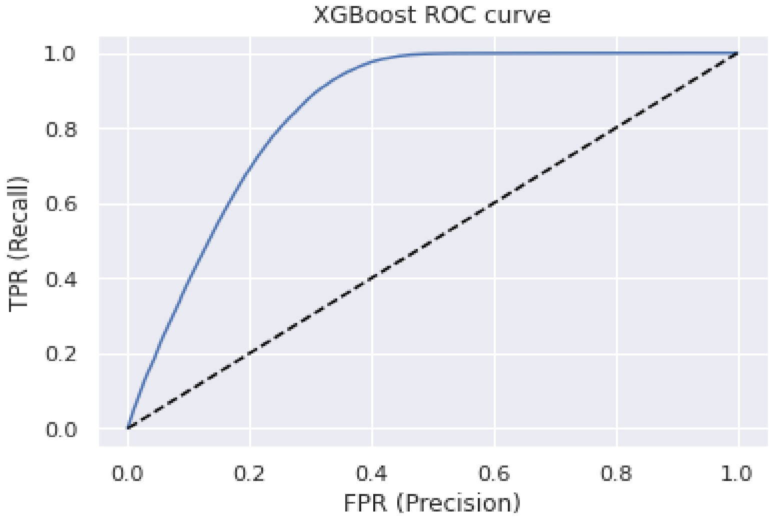


```
Accuracy: 0.7951754385964912
ROC_AUC Score: 0.8207553607253026
```

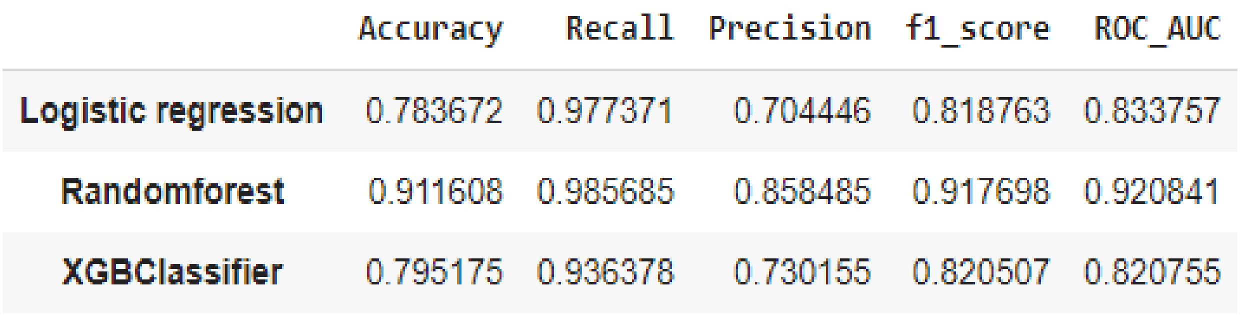


```
              precision    recall  f1-score   support

           0       0.65      0.91      0.76     71996
           1       0.94      0.73      0.82    128644

    accuracy                           0.80    200640
   macro avg       0.80      0.82      0.79    200640
weighted avg       0.84      0.80      0.80    200640
```





Comparing Model



	Accuracy	Recall	Precision	f1_score	ROC_AUC
Logistic regression	0.783672	0.977371	0.704446	0.818763	0.833757
Randomforest	0.911608	0.985685	0.858485	0.917698	0.920841
XGBClassifier	0.795175	0.936378	0.730155	0.820507	0.820755

- For this problem we have create 3 models i.e. Logistic Regression, Random Forest and XGB classifier.
- 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.
- 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.

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 afecting 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.

Reference:

1) https://www.almabetter.com/

2) https://www.wikipedia.org

3) https://www.kaggle.com/

4) https://github.com/