Capstone Project

Health Insurance Cross Sell Prediction
ML Supervised Classification

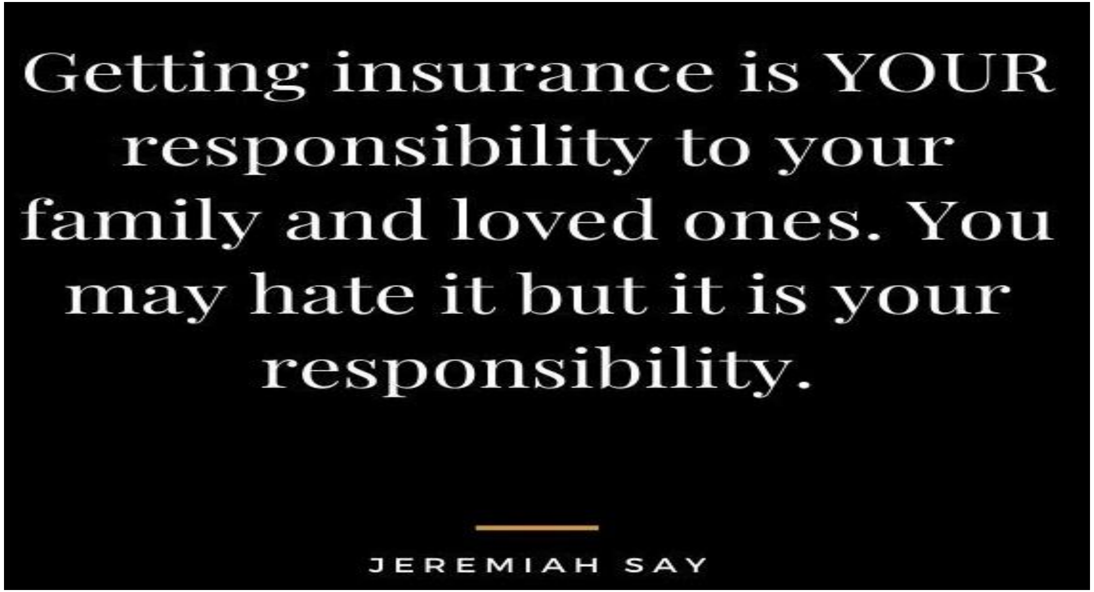
Name:
Anirban Patra

Table Of Contents

- Problem Statement
- Data Summary
- Data Cleaning
- Data Visualization
- Feature Selection
- Model Selection *(Implemented Various Classification Algorithms)*
- Hyperparameter tuning
- Conclusion

understanding insurance

- *What is an insurance policy ?*
- *How it works ?*



Problem Statements

To Predict whether a customer will buy insurance(vehicle) or not.

Data Summary

- *id :-* Unique ID for the customer
- *Gender :-* customers' gender
- *Age :-* Age of the customer
- *Driving_License :-* Customer is having driving license or not
- *Region_Code :-* Unique code for the region
- *Previously_Insured :-* Whether the customer has insured previously or not
- *Vehicle_Age :-* Age of the Vehicle
- *Vehicle_Damage :-* Is the customer got 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 :-* The customer is interested or not

Basic Data Exploration

- ***The dataset has 381109 observations and 12 features(columns).***
- ***Three categorical features `Gender' , 'Vehicle_Age' , 'Vehicle_Damage'***
- ***No Missing Values.***
- ***No Duplicate values.***

```
<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: 93.9 MB
```

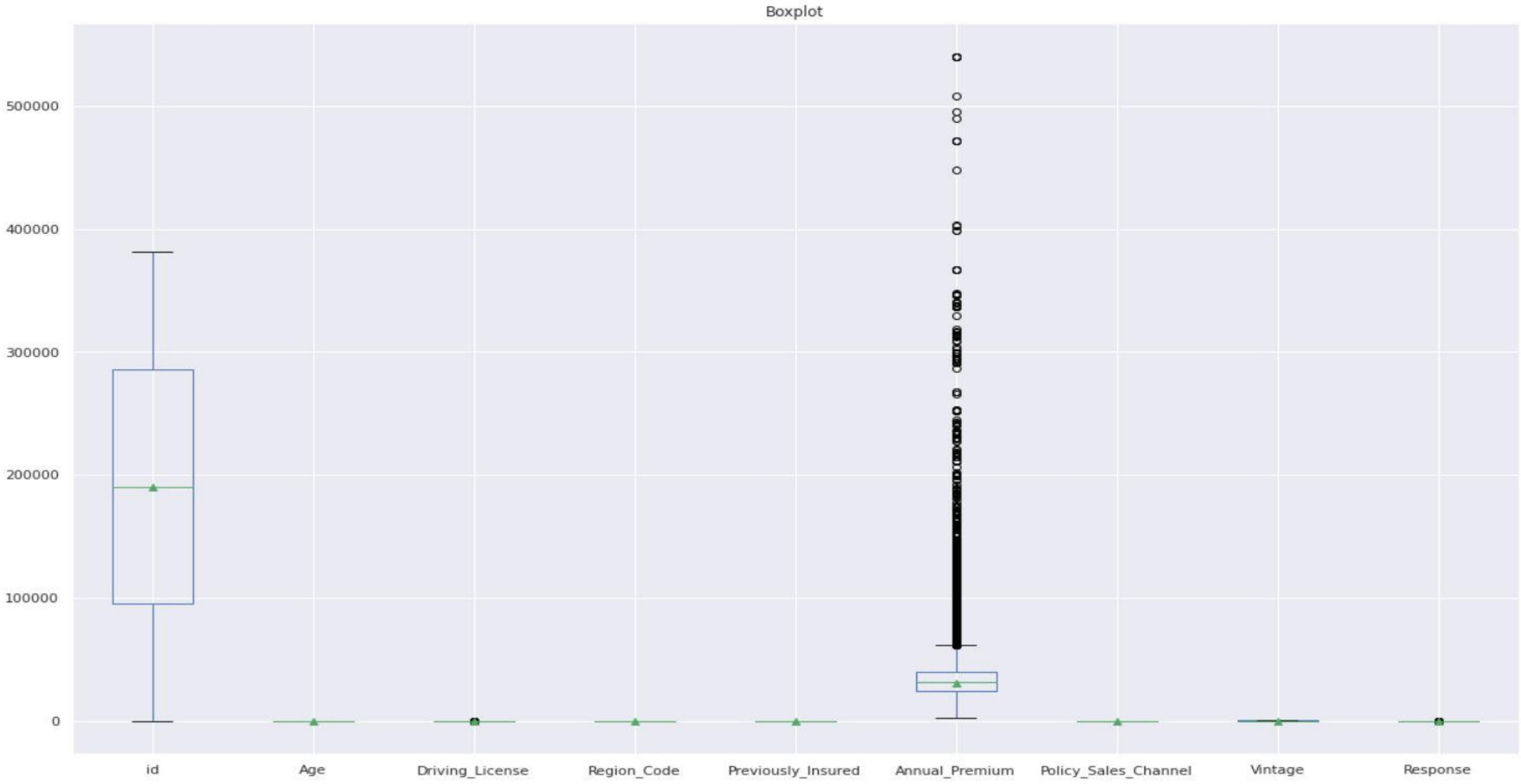
Data Info:

Name specifies the column name,dtypes stand for data types,missing denotes missing values,first value gives an instance of the first value and the same for second value.
Vehicle_age column is not in a integer,we will chnage it later.

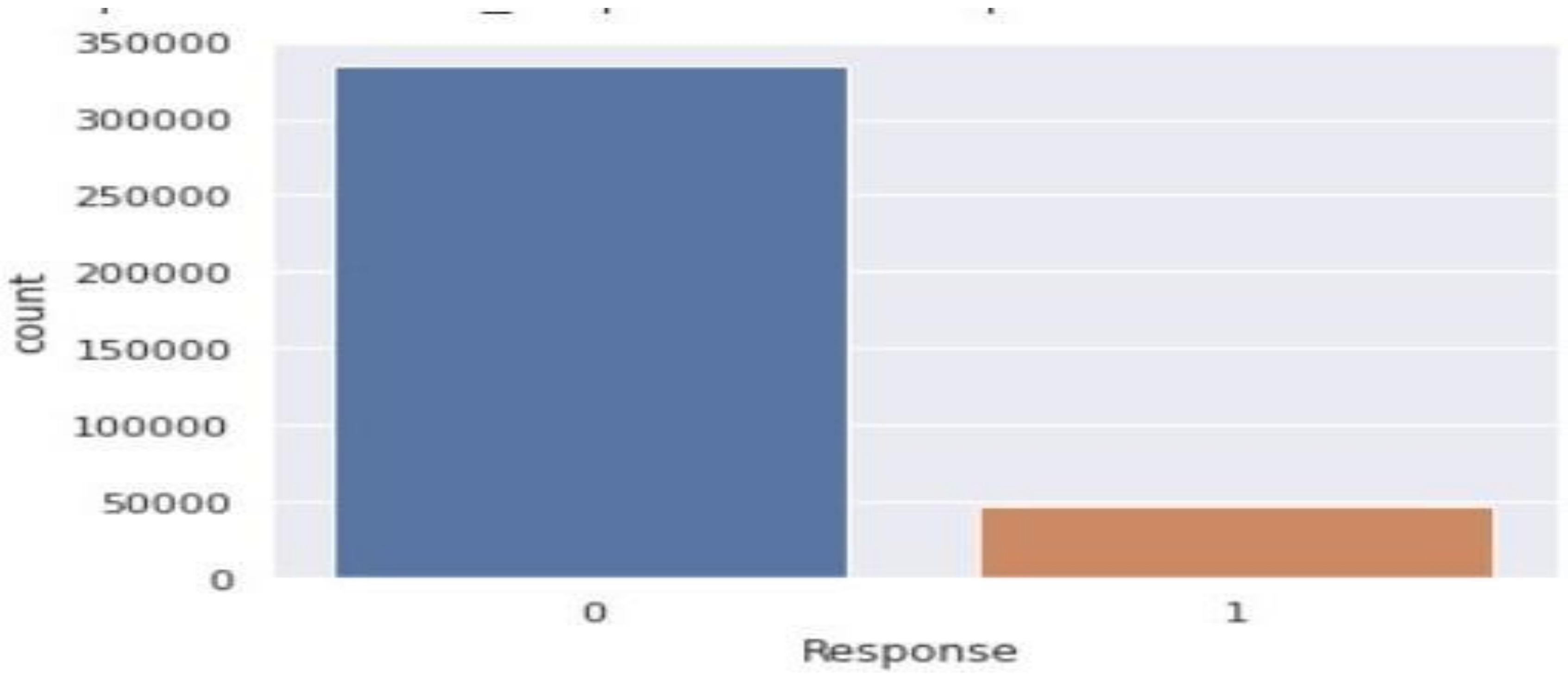
Dataset Shape: (381109, 12)

	Name	dtypes	Missing	Uniques	First Value	Second Value
0	id	int64	0	381109	1	2
1	Gender	object	0	2	Male	Male
2	Age	int64	0	66	44	76
3	Driving_License	int64	0	2	1	1
4	Region_Code	float64	0	53	28.0	3.0
5	Previously_Insured	int64	0	2	0	0
6	Vehicle_Age	object	0	3	> 2 Years	1-2 Year
7	Vehicle_Damage	object	0	2	Yes	No
8	Annual_Premium	float64	0	48838	40454.0	33536.0
9	Policy_Sales_Channel	float64	0	155	26.0	26.0
10	Vintage	int64	0	290	217	183
11	Response	int64	0	2	1	0

Outliers in the features

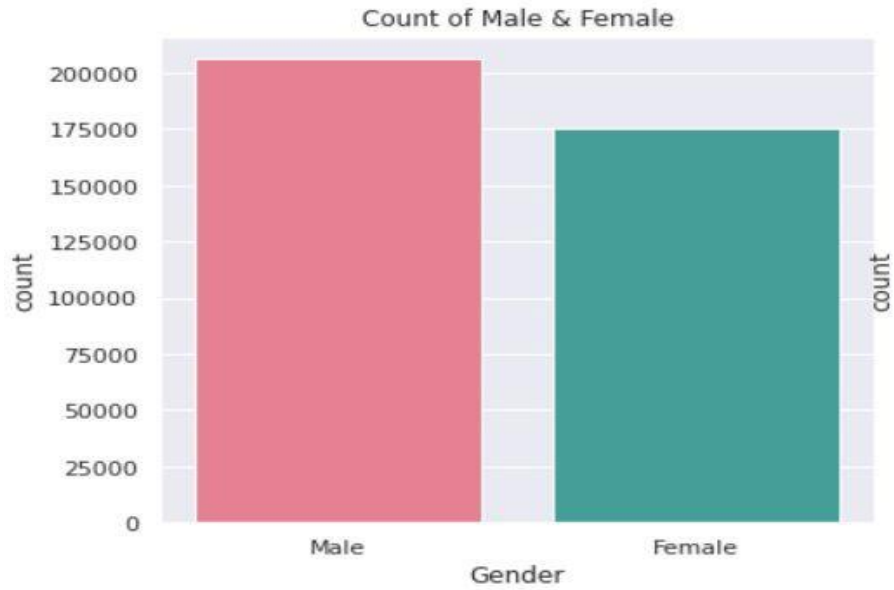


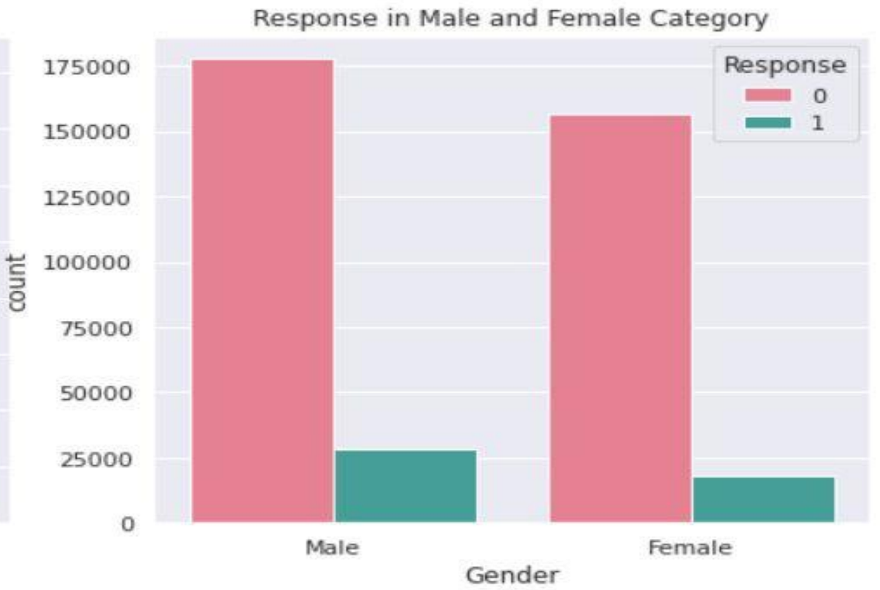
Target Column countplot



The data is **highly imbalanced**.

Gender

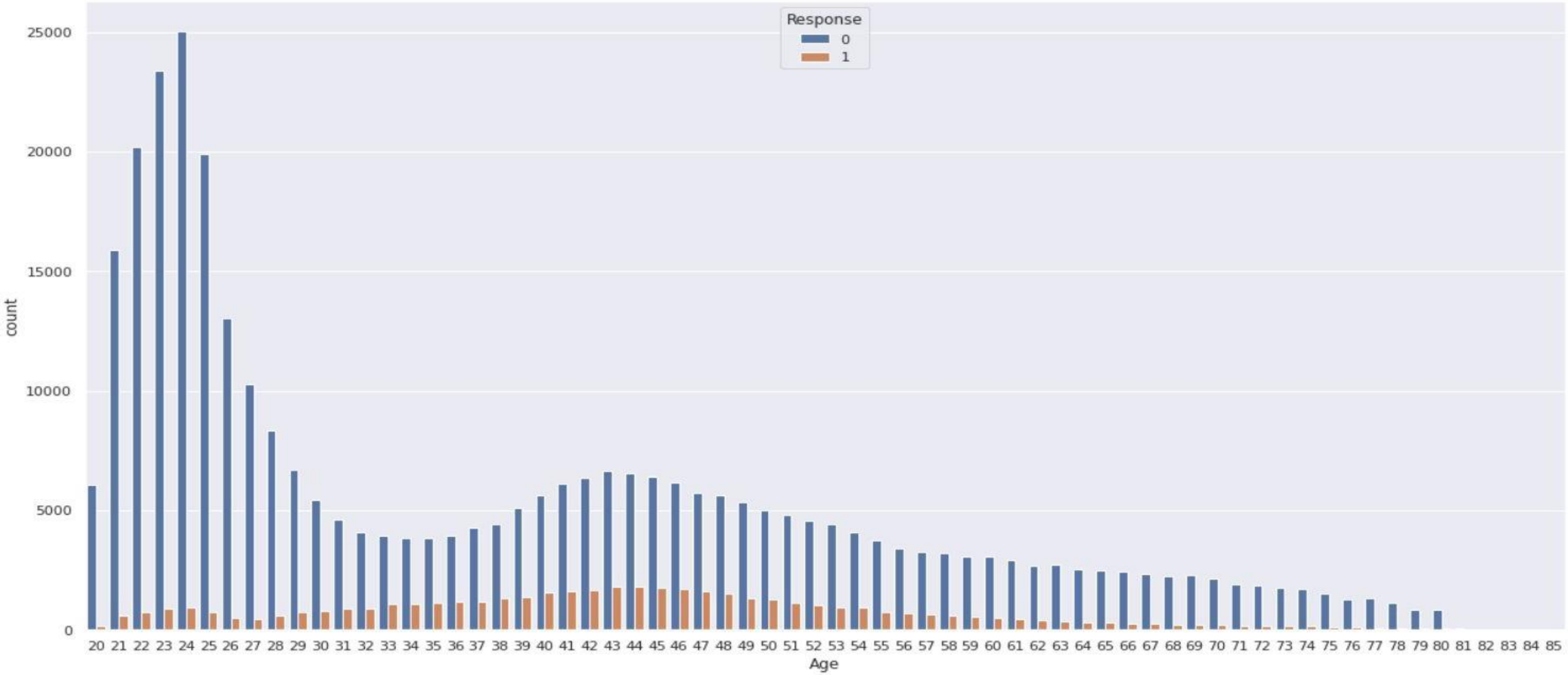




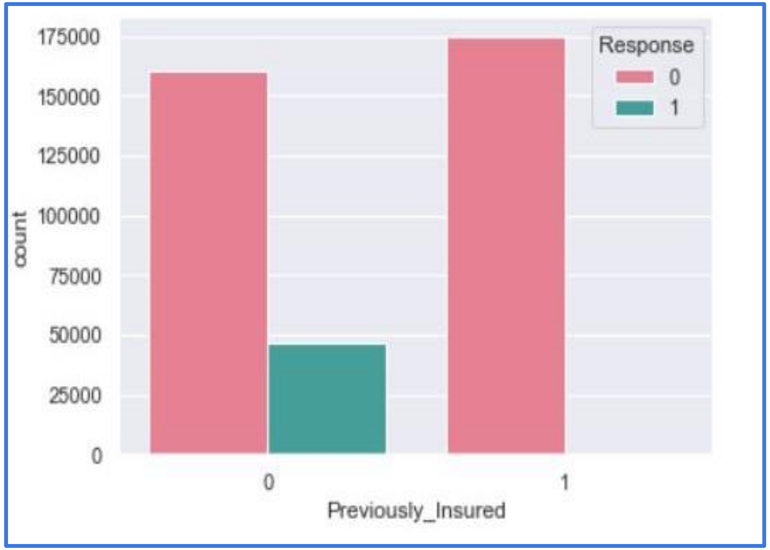
- From the 1st graph , I can say that ,The gender variable ratio in the dataset is almost equal, male category is slightly more than female and also the chances of buying insurance is also little high than female.The number of male is greater than 200000 and The number of female is close to 175000.

- From **2nd graph we can say that ,** 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

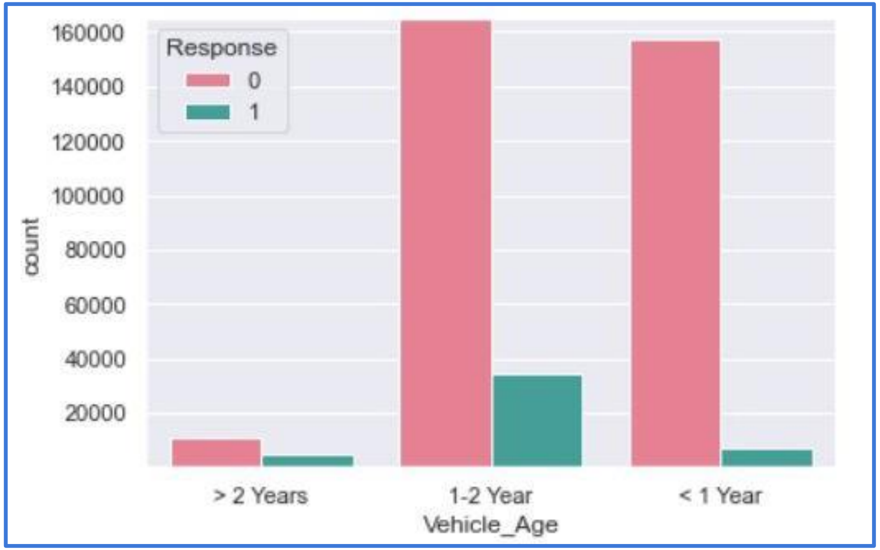
Age countplot



Data Visualization



We can conclude that those who have not insurance some of them are taking insurance



From the above graph, we can clearly say that if the vehicle age is in between 1 to 2 year they are taking more insurance than others.

Correlation Matrix

As you can see the target variable(Response) is not highly correlated with any dependent variable.

Data Cleaning & Preparation

	Vehicle_Age	Vehicle_Damage	Gender
0	> 2 Years	Yes	Male
1	1-2 Year	No	Male
2	> 2 Years	Yes	Male

Converted

Categorical Columns

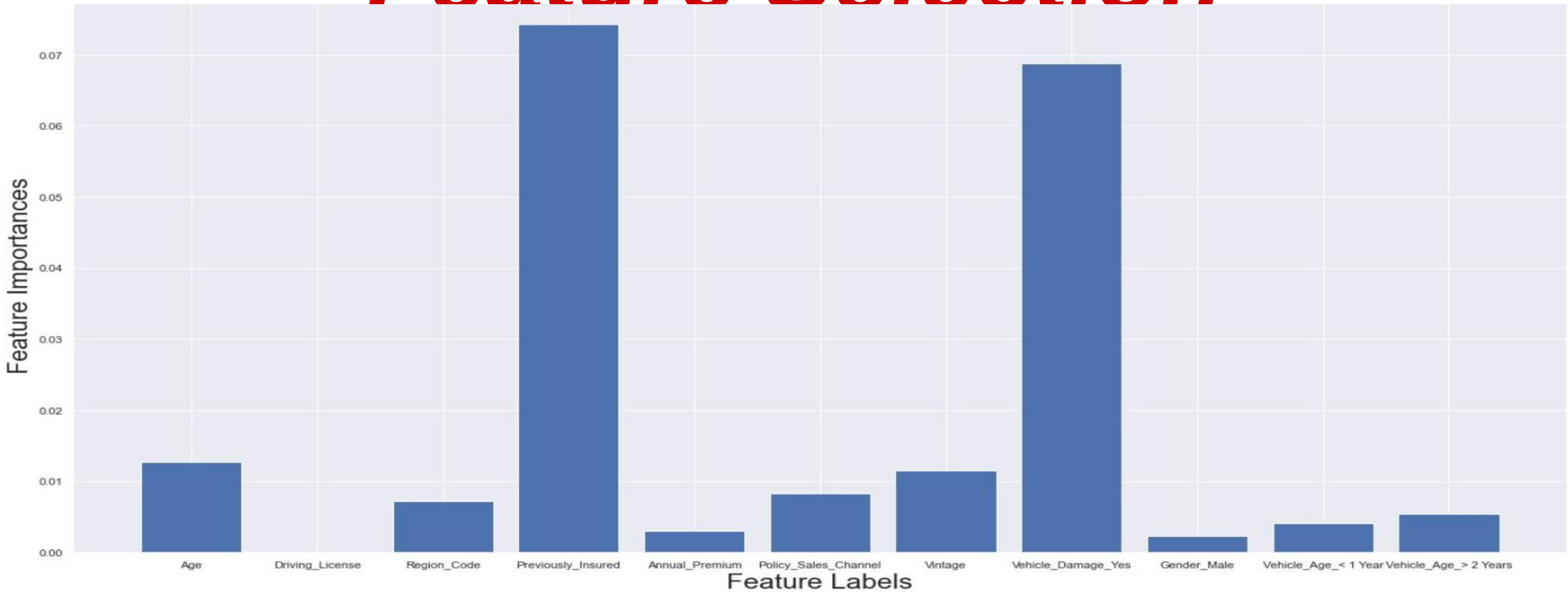
	Vehicle_Age	Vehicle_Damage	Gender
0	2	1	1
1	0	0	1
2	2	1	1

	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

All the features in **new_df** is numerical

	id	Age	Driving_License	Region_Code	Previously_Insured	Annual_Premium	Policy_Sales_Channel	Vintage	Response	Vehicle_Age	Vehicle_Damage	Gender
0	1	44	1	28.0	0	40454.0	26.0	217	1	2	1	1
1	2	76	1	3.0	0	33536.0	26.0	183	0	0	0	1

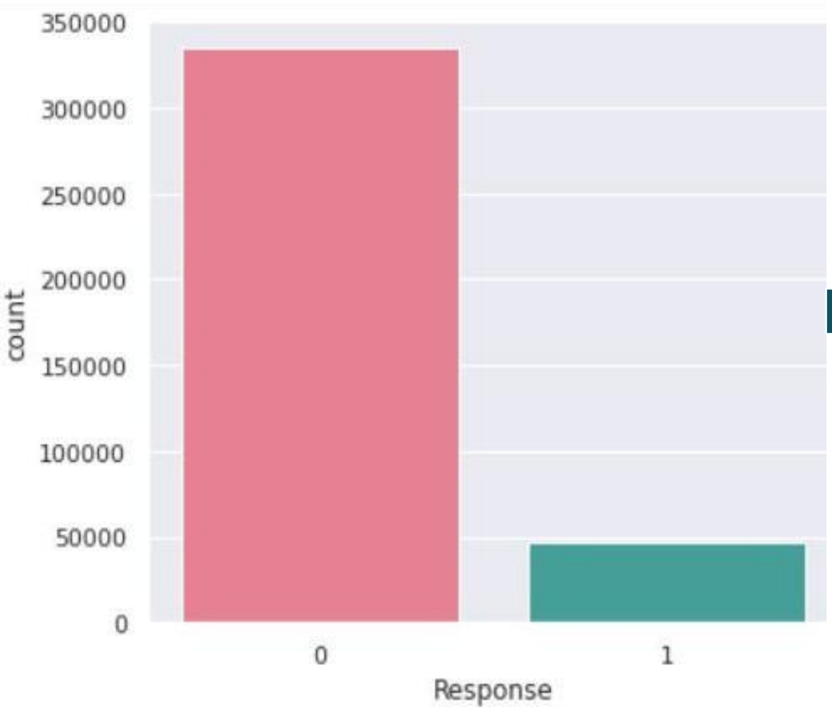
Feature Selection



- **previously insured and vehicle damage** is contributing most
- But, **Driving license , Annual Premium , Gender** contributing least
- **Dropping** these columns :- Driving_License' , 'Vehicle_Age_> 2 Years', Gender_Male

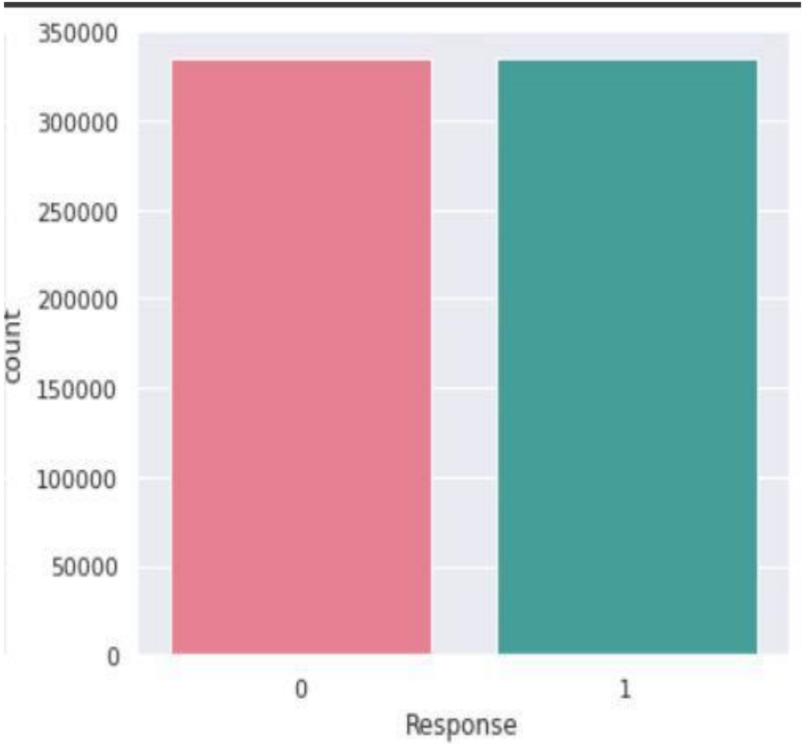
Data Preparation (part1)

Using **RandomOverSampler** to resample because the data is highly imbalanced



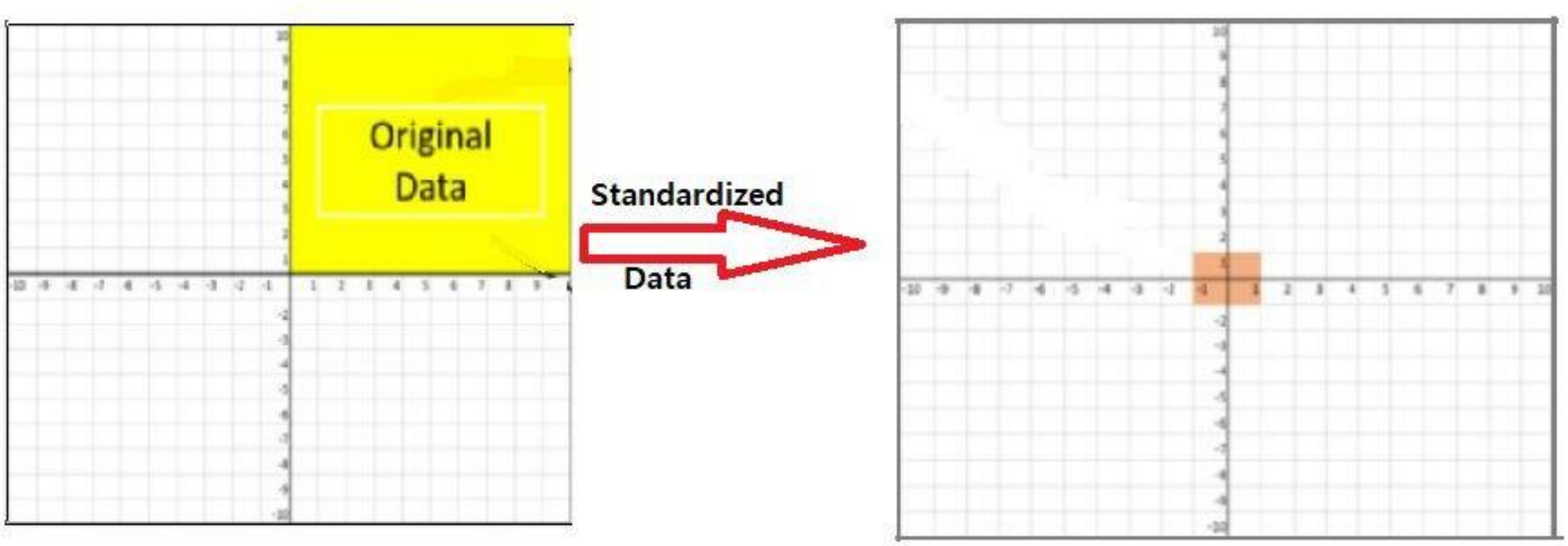
After using

RandomOverSampler



Data Preparation (part2)

- **Label Encoding**
- **Train Test Split (** test_size = 0.2 , random_state = 1**)**
- **StandardScaler**



Model Selection

- This problem can be identified as Binary Classification (whether customer opts for vehicle insurance or not)
- Dataset has more than 300k records
- Cannot go with SVM Classifier as it takes more time to train as dataset increase

Models we will be using here are:

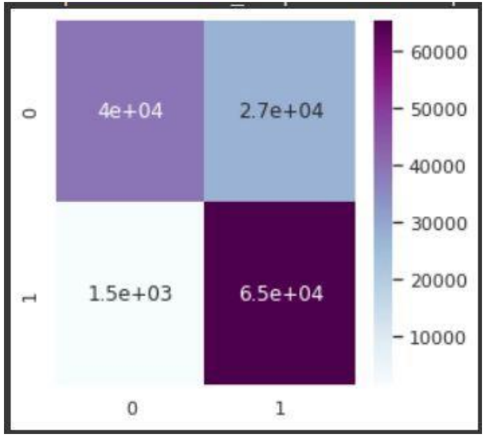
1. Logistic Regression
2. RandomForestClassifier
3. XGBClassifier
4. KNN-Classifier

1.Logistic Regression

Classification Report

	precision	recall	f1-score	support
0	0.59	0.96	0.73	41010
1	0.98	0.71	0.82	92750
accuracy			0.78	133760
macro avg	0.78	0.83	0.78	133760
weighted avg	0.86	0.78	0.79	133760

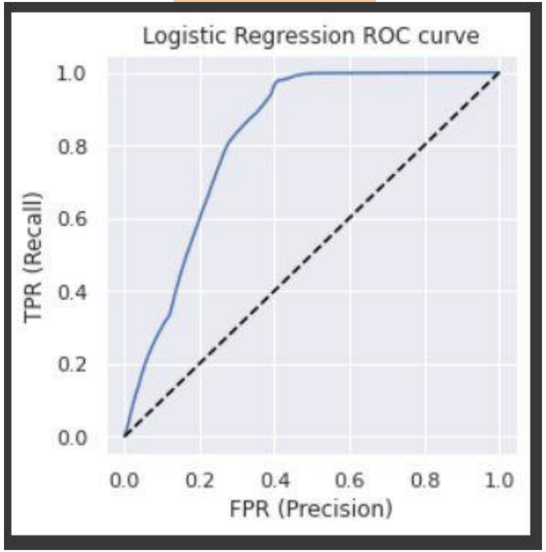
Confusion Matrix



Test Dataset details

Accuracy : 0.784
Precision: 0.705
Recall: 0.978
F1-Score: 0.819
ROC_AUC Score: 0.834

ROC Curve

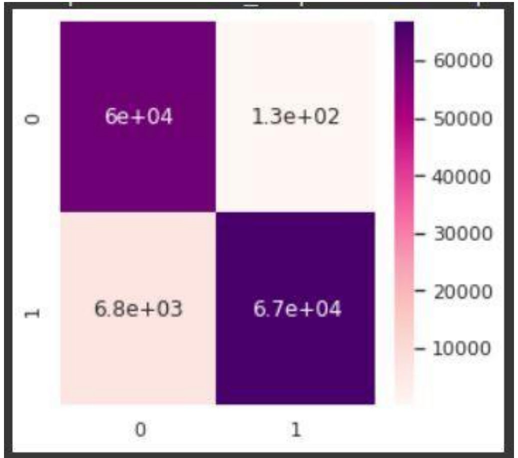


2. RandomForestClassifier

Classification Report

	precision	recall	f1-score	support
0	0.90	1.00	0.95	60119
1	1.00	0.91	0.95	73641
accuracy			0.95	133760
macro avg	0.95	0.95	0.95	133760
weighted avg	0.95	0.95	0.95	133760

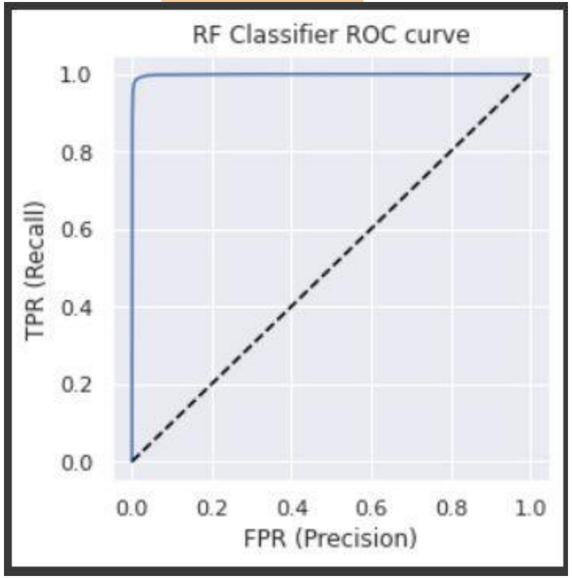
Confusion Matrix



Test Dataset details

```
Accuracy :  0.948
Precision: 0.908
Recall: 0.998
F1-Score: 0.951
ROC_AUC Score: 0.834
```

ROC Curve

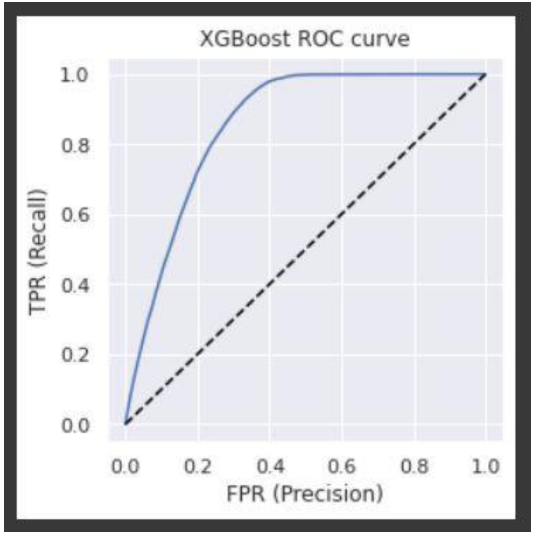


3.XGBClassifier

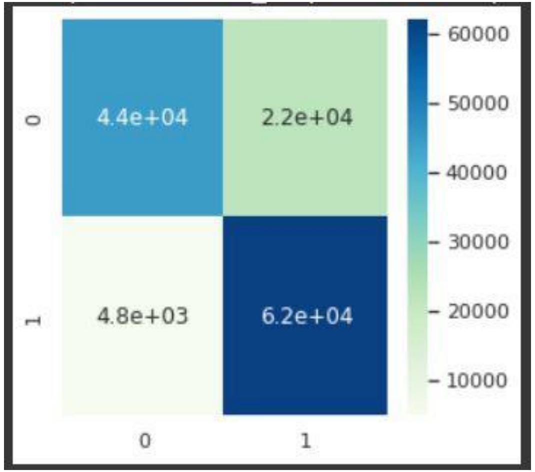
Classification Report

	precision	recall	f1-score	support
0	0.67	0.90	0.77	49255
1	0.93	0.74	0.82	84505
accuracy			0.80	133760
macro avg	0.80	0.82	0.79	133760
weighted avg	0.83	0.80	0.80	133760

ROC Curve



Confusion Matrix



Test Dataset details

Accuracy : 0.797
Precision: 0.735
Recall: 0.928
F1-Score: 0.821
ROC_AUC Score: 0.819

4.KNNClassifier

Classification Report

	precision	recall	f1-score	support
0	0.97	0.76	0.85	66847
1	0.80	0.98	0.88	66913
accuracy			0.87	133760
macro avg	0.88	0.87	0.86	133760
weighted avg	0.88	0.87	0.86	133760

Let’s compare those models

	Accuracy	Recall	Precision	f1_score	ROC_AUC
Logistic regression	0.784412	0.977583	0.705261	0.819388	0.866401
RandomForest	0.947884	0.998087	0.907060	0.950399	0.952466
XGBClassifier	0.796920	0.928474	0.735187	0.820603	0.819010
KNeighborsClassifier	0.866455	0.975491	0.800935	0.879637	0.866401

Will select RandomForest as final model
as it has the best scores

Hyperparameter Tuning

	Accuracy	Recall	Precision	f1_score	ROC_AUC
RandomForest	0.947967	0.998266	0.907063	0.950482	0.952567
RandomForest(Using Hyper.)	0.962051	0.965027	0.911644	0.952364	0.954106

- Used **GirdSearchCV**
- Best hyperparametes values:
    - criterion: gini
    - max_depth: 50
    - min_samples_split : 2
    - n_estimators: 10

Conclusion

The ML model for the problem statement was created using python with the help of the dataset(contains more than 300k observations)and RandomForestClassifier performed best among those three models (Logistic Reg. , XGBClassifier , RandomForestClassifier).

Thus, for the given problem, the model created by Random Forest is preferred.

NOTES:

1.Customers of age between 30 to 60 are more likely to buy insurance.

2. Customers with Driving License have higher chance of buying Insurance.

3. Customers with Vehicle_Damage are likely to buy insurance.

4.The variable such as Previously_insured , Vehcile_Damage are more affecting the target variable.

5 The variable such as Driving_License , Gender are not affecting the target variable.

6. comparing ROC curve we can see that Random Forest model perform better. Because curves closer to the top-left corner, it indicate a better performance.


THANK YOU