Capstone Project - 3

Project Title - Health Insurance cross sell prediction

By
Sapana Pawar

Points for discussion

- Problem Statement
- Introduction
- Data Summary
- EDA and Data visualization
- Feature Engineering
- Handling Imbalance data
- Model Implementation
- Model Evaluation
- Hyperparameter Tuning
- Conclusion

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.

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.

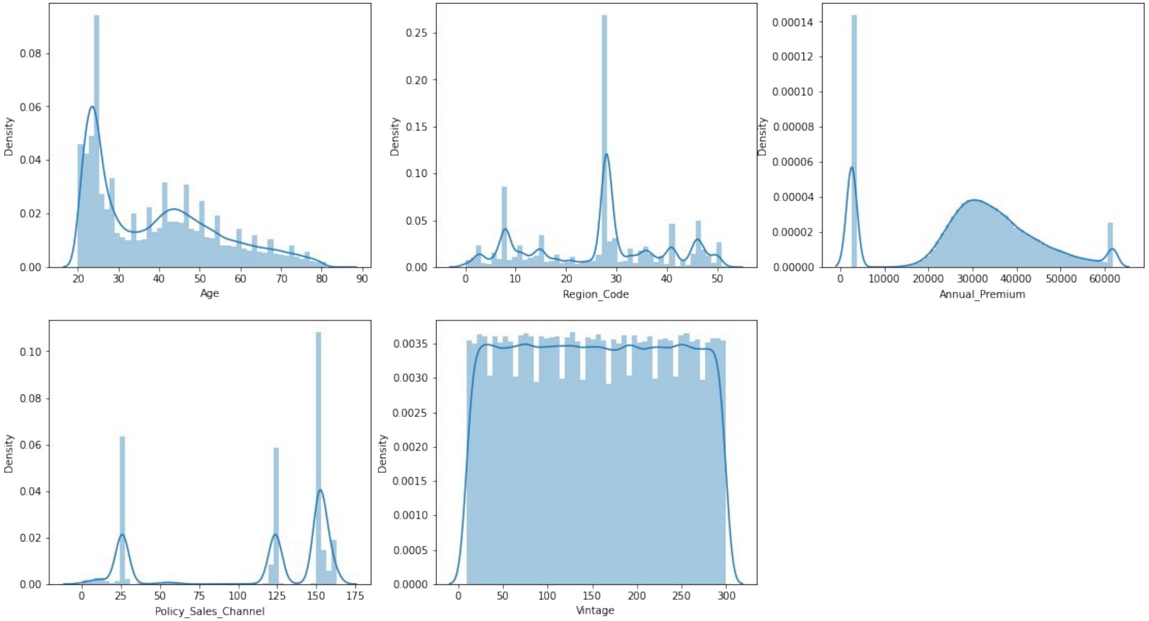
Building a model to predict whether a customer would be interested in Vehicle Insurance is extremely helpful for the company because it can then accordingly plan its communication strategy to reach out to those customers and optimise its business model and revenue.

Data Summary

- 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

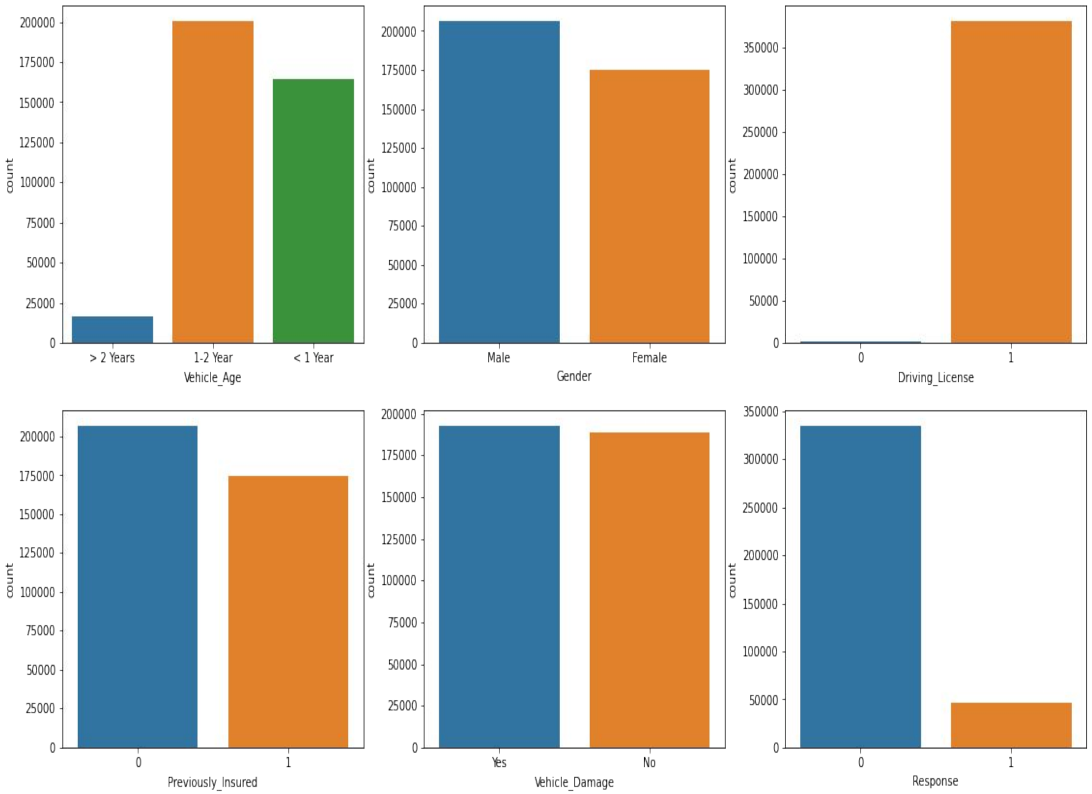
EDA and Data Visualization

Univariate Analysis Of Continuous Variables



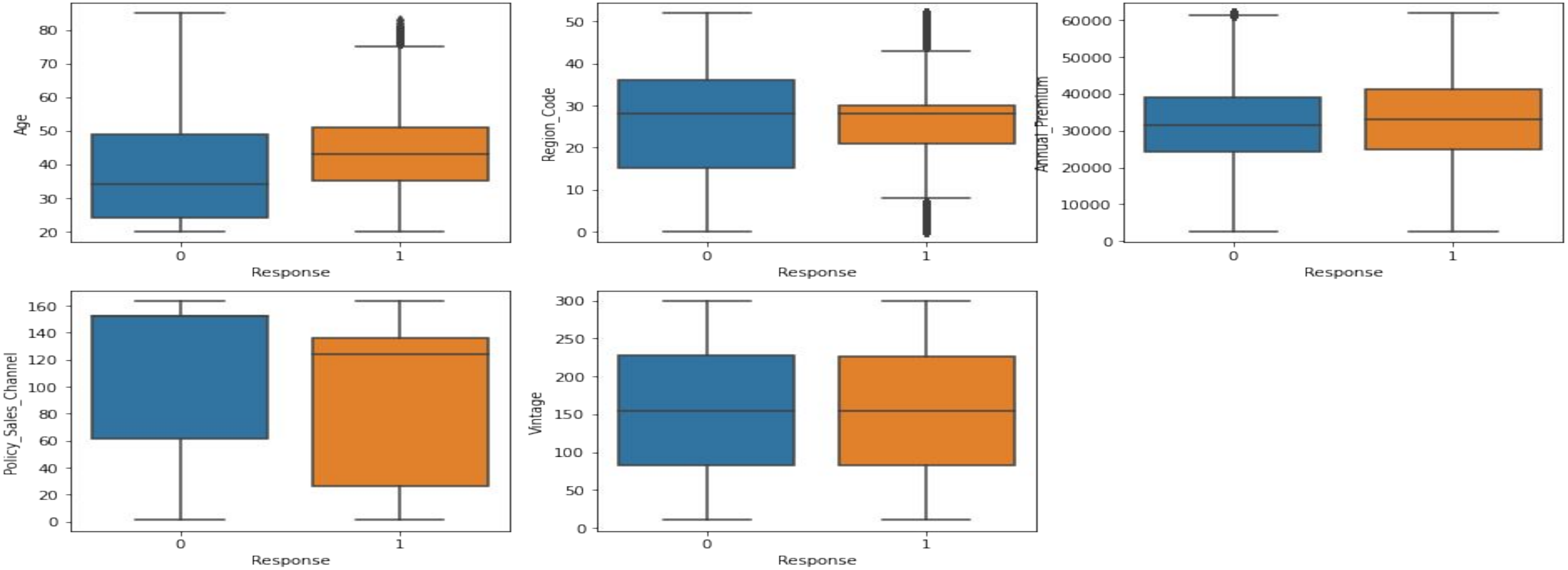
- Variable ‘Age’ is highly skewed towards right.
- ‘Region Code’ is randomly distributed.
- ‘Annual Premium’ normally distributed with little right skewed.
- The variable ‘Policy Sales Channel’ is randomly distributed with hueness.
- “Vintage’ is uniformly distributed.

Univariate Analysis Of Categorical Variables



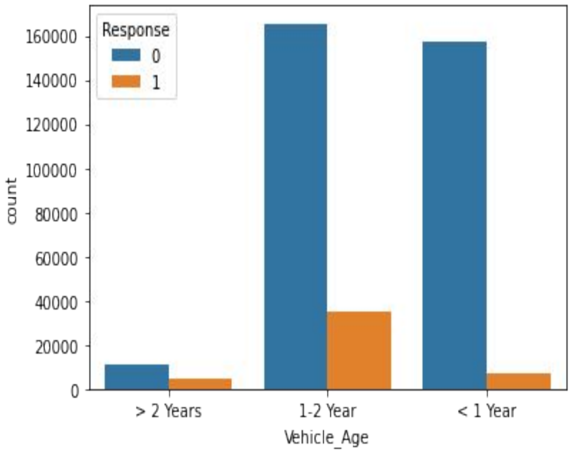
- most of the vehicle taken in this study is 1-2 years old.
- Count of male and female customers are quite same.
- Almost all people have driving license.
- There are more number of people who don’t have previous insurance policy than the number of previously insured.
- Count of customer with and without vehicle damage is quite same.
- Most of the customers seems to not interested in vehicle insurance policy.

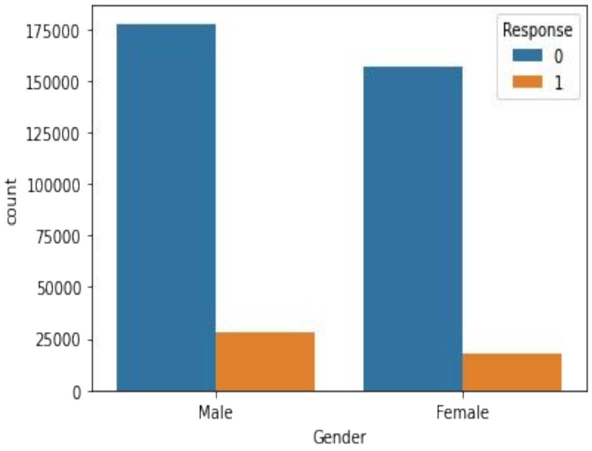
Bivariate Analysis Of Continuous Variables with Target Variable

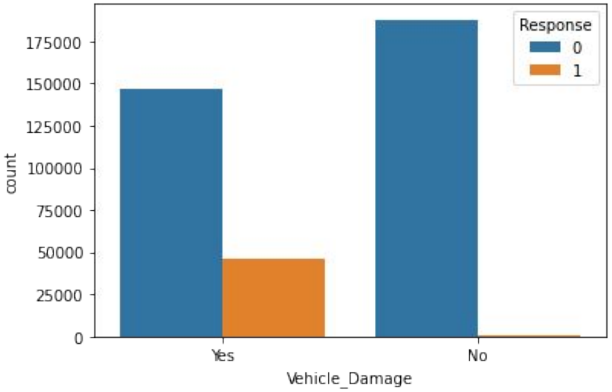


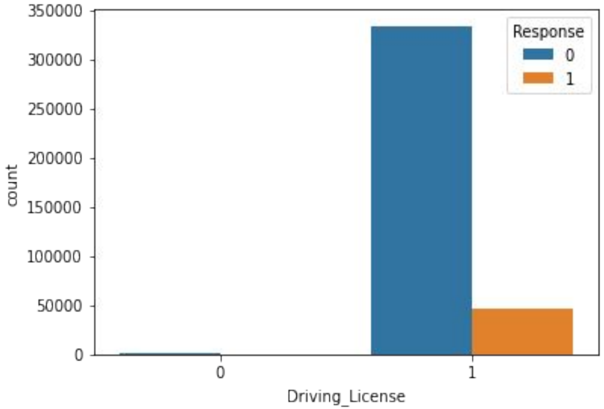
- Elder customers are interested in vehicle insurance
- region code does not affect on the vehicle insurance.

Bivariate Analysis Of Categorical Variables with Target Variable



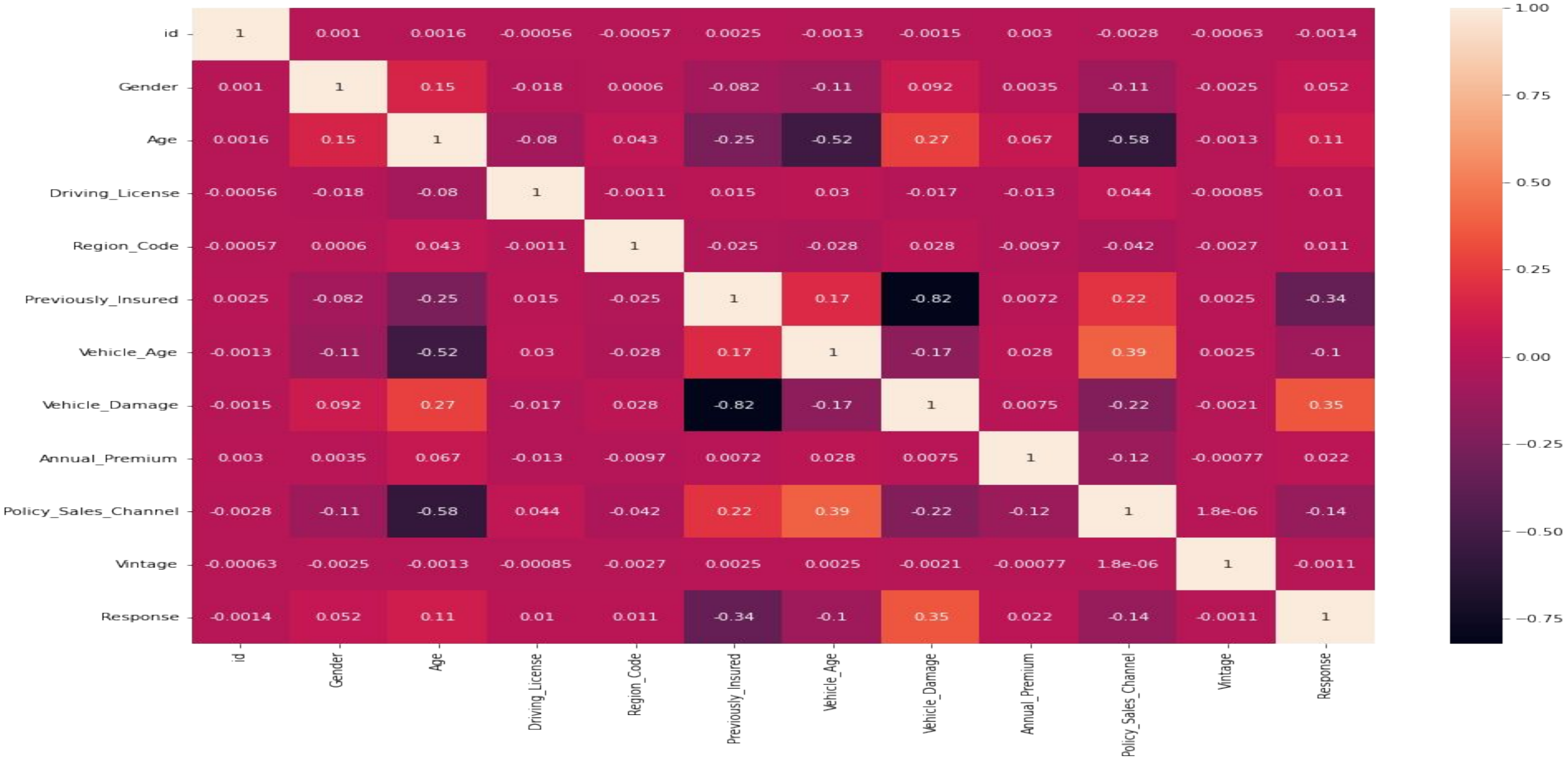






- Customers with vehicle age 1-2 years are more interested in insurance.
- Chances of buying insurance is little high if customer is male.
- Almost all customers who have driving licence are seems to be interested in vehicle insurance.
- Customers with vehicle damage are more interested in Vehicle Insurance.

Correlation Heatmap

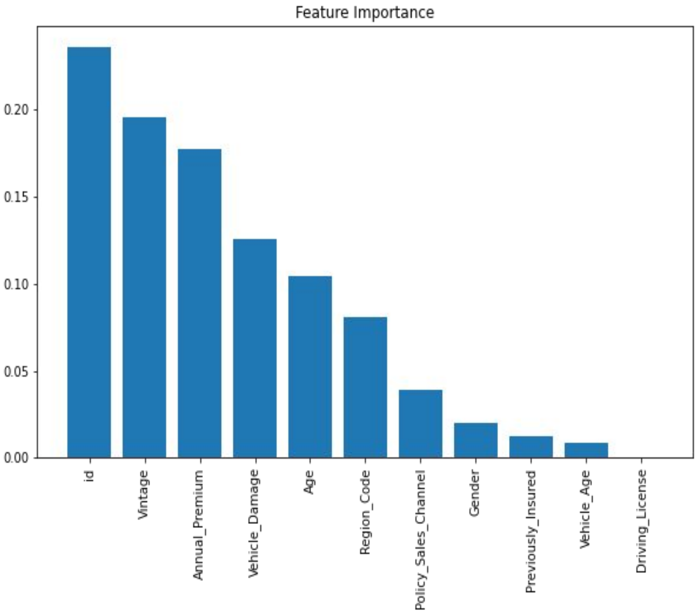


	id	Gender	Age	Driving_License	Region_Code	Previously_Insured	Vehicle_Age	Vehicle_Damage	Annual_Premium	Policy_Sales_Channel	Vintage	Response
id	1	0.001	0.0016	-0.00056	-0.00057	0.0025	-0.0013	-0.0015	0.003	-0.0028	-0.00063	-0.0014
Gender	0.001	1	0.15	-0.018	0.0006	-0.082	-0.11	0.092	0.0035	-0.11	-0.0025	0.052
Age	0.0016	0.15	1	-0.08	0.043	-0.25	-0.52	0.27	0.067	-0.58	-0.0013	0.11
Driving_License	-0.00056	-0.018	-0.08	1	-0.0011	0.015	0.03	-0.017	-0.013	0.044	-0.00085	0.01
Region_Code	-0.00057	0.0006	0.043	-0.0011	1	-0.025	-0.028	0.028	-0.0097	-0.042	-0.0027	0.011
Previously_Insured	0.0025	-0.082	-0.25	0.015	-0.025	1	0.17	-0.82	0.0072	0.22	0.0025	-0.34
Vehicle_Age	-0.0013	-0.11	-0.52	0.03	-0.028	0.17	1	-0.17	0.028	0.39	0.0025	-0.1
Vehicle_Damage	-0.0015	0.092	0.27	-0.017	0.028	-0.82	-0.17	1	0.0075	-0.22	-0.0021	0.35
Annual_Premium	0.003	0.0035	0.067	-0.013	-0.0097	0.0072	0.028	0.0075	1	-0.12	-0.00077	0.022
Policy_Sales_Channel	-0.0028	-0.11	-0.58	0.044	-0.042	0.22	0.39	-0.22	-0.12	1	1.8e-06	-0.14
Vintage	-0.00063	-0.0025	-0.0013	-0.00085	-0.0027	0.0025	0.0025	-0.0021	-0.00077	1.8e-06	1	-0.0011
Response	-0.0014	0.052	0.11	0.01	0.011	-0.34	-0.1	0.35	0.022	-0.14	-0.0011	1

- Vehicle_Damage is highly correlated with Previously_Insured with correlation of -0.82
- Policy_Sales_Channel is correlated with Age with correlation with -0.58

Feature Engineering

1. Feature Selection



- The most important features from the dataset are vintage, annual premium and vehicle damage.

2. Label Encoding

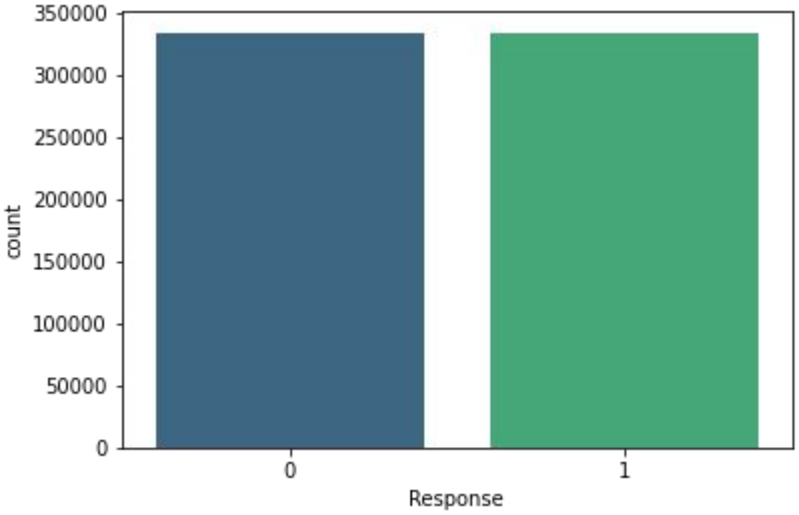
- Label Encoding refers to converting the labels into a numeric form so as to convert them into the machine-readable form.
- Performed Label Encoding on 'Vehicle_Age', 'Gender' and 'Vehicle_Damage'.

Handling Imbalance Data

As from the distribution of target variables in the EDA section, we know that our data is highly imbalanced.

So to handle such a problem, we have balance the data. For this problem I have used Random Oversampling technique.

- After using Random oversampling technique now our data is balanced.



Model Implementation and Evaluation

Model Name	Accuracy	Precision	Recall	F-1 Score	AUC
1.Logistic Regression	0.78	0.59	0.96	0.73	0.82
2. Random Forest Classifier	**0.94**	**0.90**	**0.99**	**0.94**	**0.99**
3.XGBoost Classifier	0.80	0.66	0.91	0.77	0.85
4.Naive Baye’s Classifier	0.78	0.59	0.96	0.73	0.82

- Random forest classifier model has given better results as compared to other algorithms.

Hyperparameter Tuning

- Hyperparameter Tuning is choosing a set of optimal parameters for learning algorithms.

- Tuned Hyperparameters

```
parameters = {'max_depth':[50, 100, None],
              'max_leaf_nodes':[500,1000,
None],
             'n_estimators': [50, 100, 200]}
```

Conclusion

Conclusion:

After Hyperparameter tuning on Random Forest model using Random Search CV we can see the slight change in the accuracy i. e.

- Accuracy is increased from 0.9450 to 0.9464
- Precision is increased from 0.9031 to 0.9055
- Recall is increased from 0.9967 to 0.9970
- F1-Score is increased from 0.9477 to 0.9489

This results can said to be good for this dataset.

Thank You