Capstone Project

Project Title

HEALTH INSURANCE CROSS SELL PREDICTION

By- Arkopravo Pradhan

CONTENTS OF THE PRESENTATION

1. 1. Problem Statement
2. 2. Data Summary
3. 3. EDA
4. 4. Data Preprocessing
5. 5. Model Evaluation
6. 6. Conclusion
7. 7.Future Work

Problem Statement

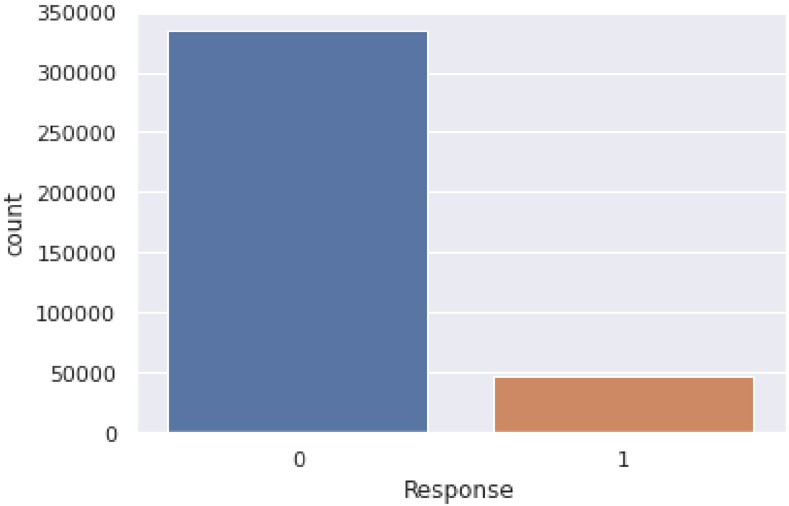
- An insurance policy is an arrangement by which a company undertakes to provide a guarantee of compensation for specified loss, damage, illness, or death in return for the payment of a specified premium. A premium is a sum of money that the customer. Just like medical insurance, there is vehicle insurance 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. So we tried our best to build a model to predict whether a customer would be interested in Vehicle Insurance or not.

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 i.e. 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

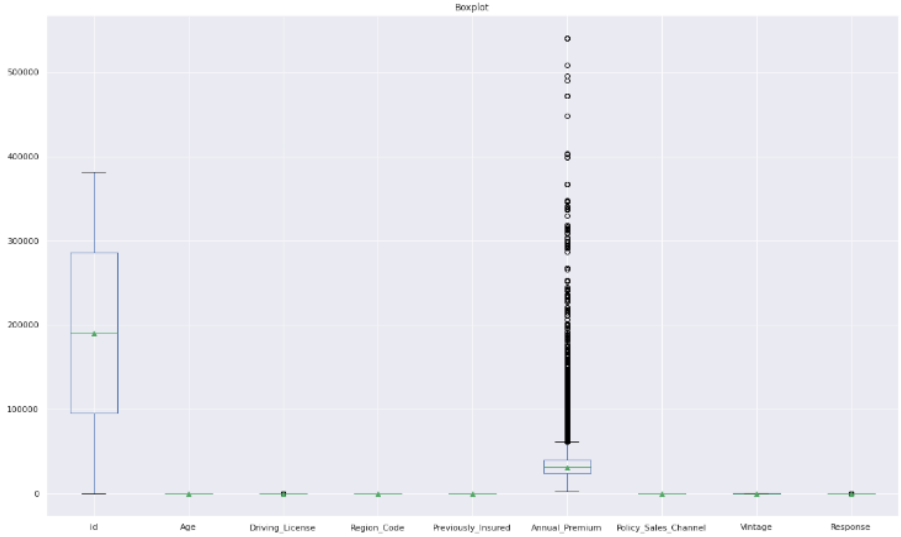
Exploratory Analysis and Visualization

The data is highly imbalanced. As we can see in above graph, there are very few interested customers whose stats are less than 50000 and those above 300000 are not interested



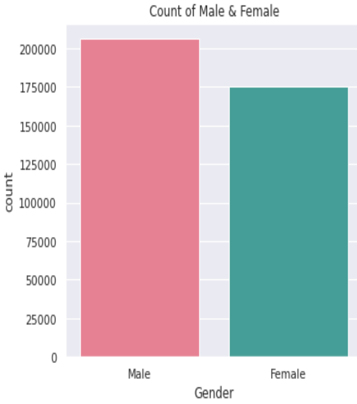
Exploratory Analysis and Visualization

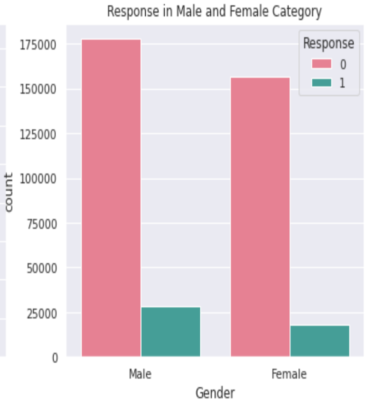
- As we can see
- 1. Annual_Premium has the highest outliers present in this dataset
- 2. Driving_License has very less outliers.
- 3. Response has very less outliers.



Exploratory Analysis and Visualization

- 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. 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.

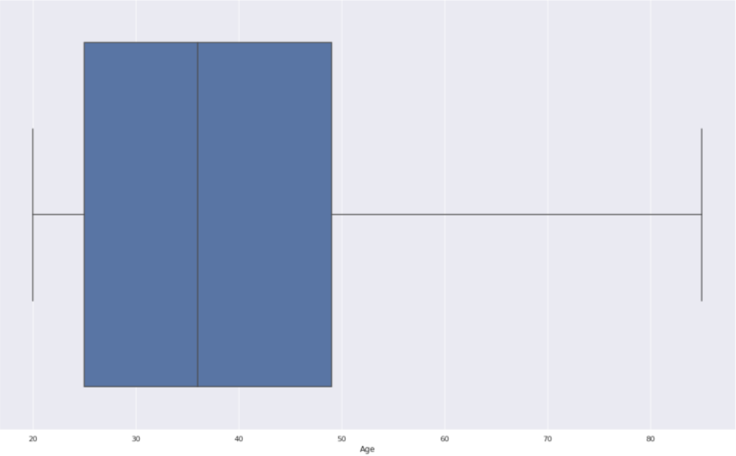

Count of Male & Female

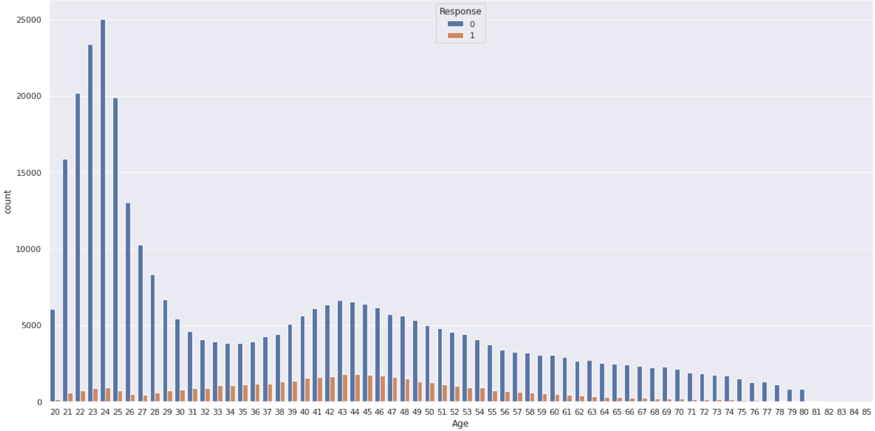

Response in Male and Female Category

Exploratory Analysis and Visualization

- Results:

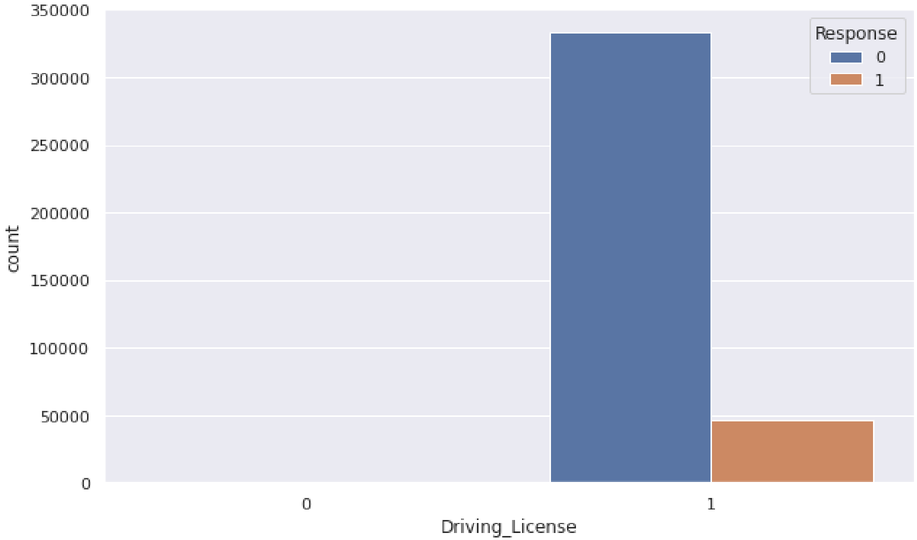
- Young people below 30 are not interested in vehicle insurance. Reasons could be lack of experience, less maturity level and they don't have expensive vehicles yet.
- People aged between 30-60 are more likely to be interested.
- From the boxplot we can see that there no outlier in the data as you can see there is no outliers present in Age





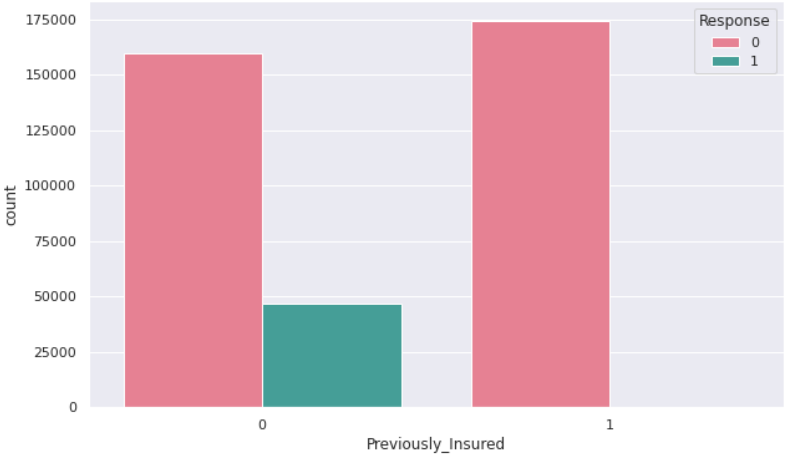
Exploratory Analysis and Visualization

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



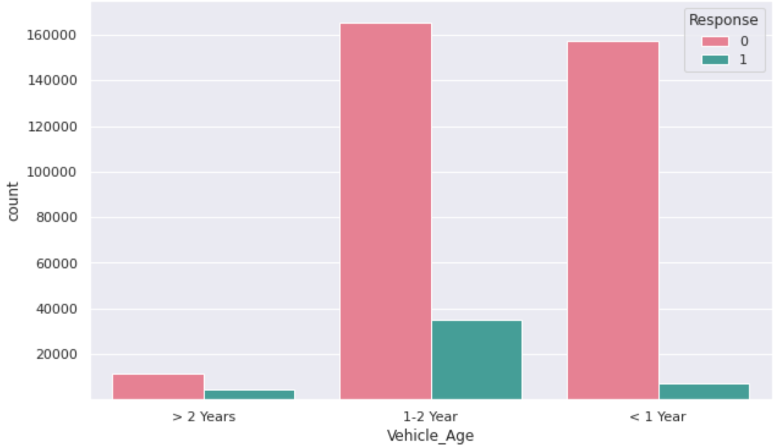
Exploratory Analysis and Visualization

- we can see that those who have not insurance some of them are taking insurance



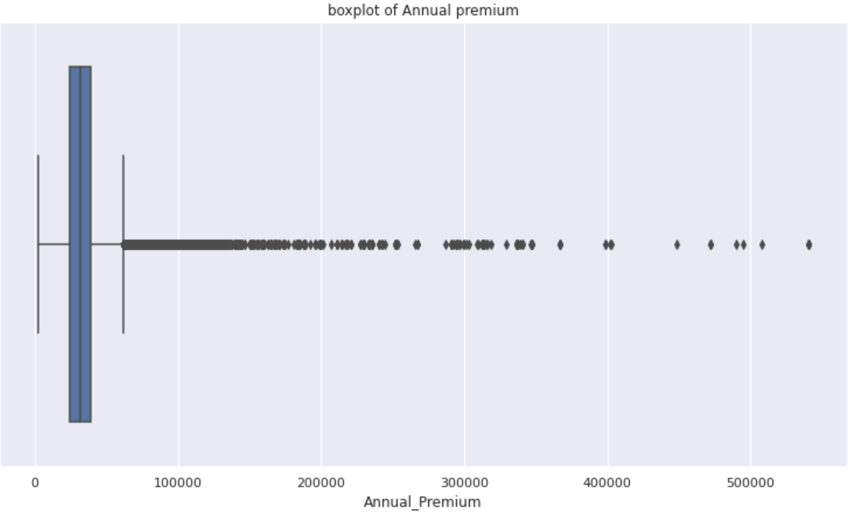
Exploratory Analysis and Visualization

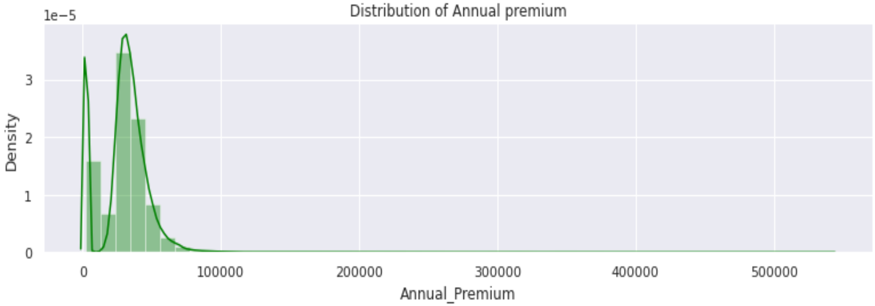
- From seeing this graph we can say that if the vehicle's age is in between 1 to 2 years ,those vehicle owners are more likely to buy insurance
- No of customers with Vehicle_Age >2 is more than the no of customers whose Vehicle_Age< 1



Exploratory Analysis and Visualization

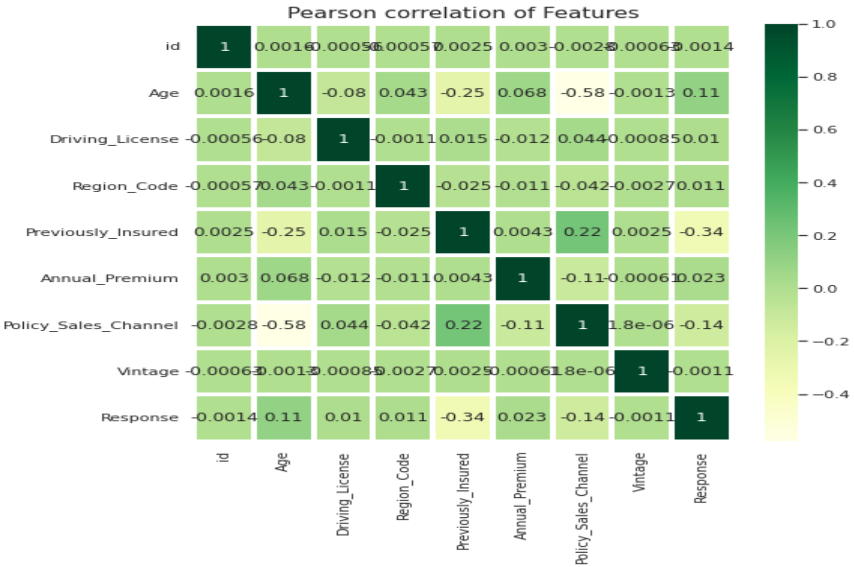
- From the distribution plot we can infer that the annual premium variable is right skewed.
- As you can see that in the column Annual_premium there are many outliers present



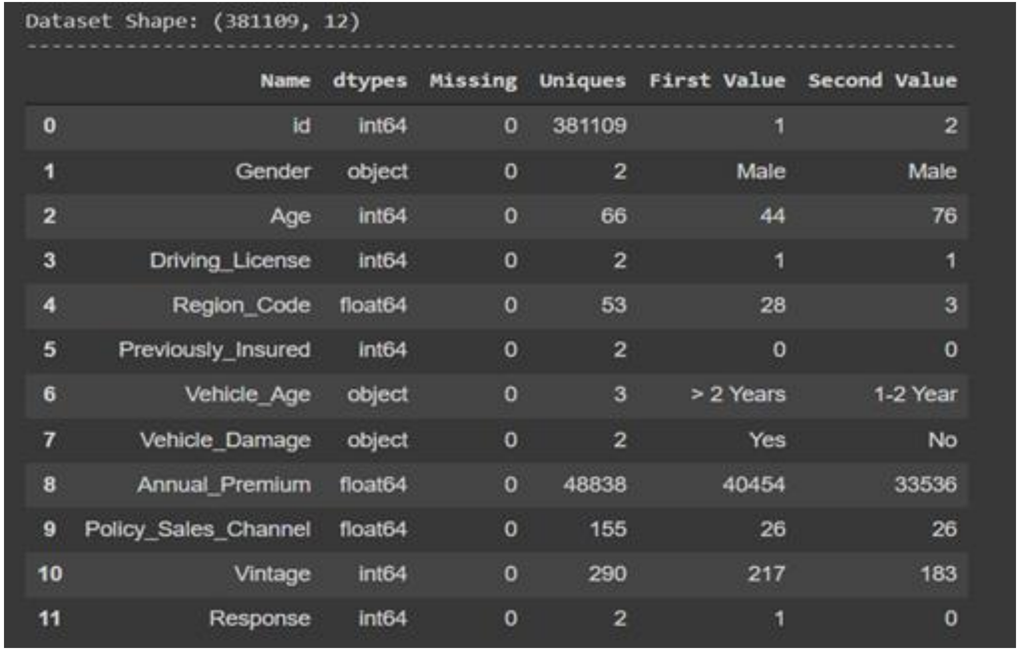


Exploratory Analysis and Visualization

- **Target variable (Response)** is not much affected by Vintage variable. we can drop least correlated variable.



Data Preprocessing

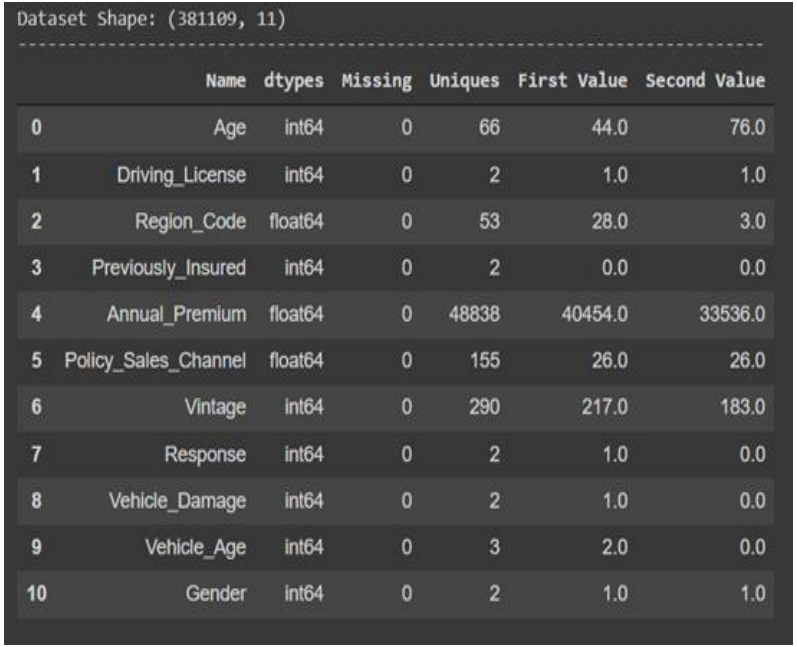


```
Dataset Shape: (381109, 12)
--------------------------------------------------------------------
```

	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	3
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	33536
9	Policy_Sales_Channel	float64	0	155	26	26
10	Vintage	int64	0	290	217	183
11	Response	int64	0	2	1	0

Data Preprocessing

- We tried to remove the duplicated rows of the dataset but there are no duplicate rows
- **Label encoding:** Label Encoding refers to converting the labels into a numeric form so as to convert them into the machine-readable form. Machine learning algorithms can then decide in a better way how those labels must be operated. It is an important pre-processing step for the structured dataset in supervised learning. We used level encoding to convert the categorical column –vehicle_damge, vehicle_age and gender to numeric.



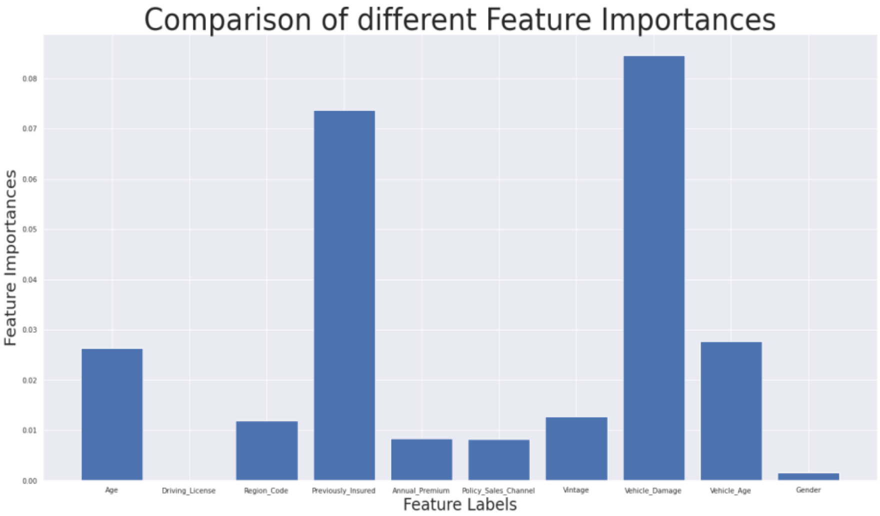
```
Dataset Shape: (381109, 11)
--------------------------------------------------------------------------------
```

	Name	dtypes	Missing	Uniques	First Value	Second Value
0	Age	int64	0	66	44.0	76.0
1	Driving_License	int64	0	2	1.0	1.0
2	Region_Code	float64	0	53	28.0	3.0
3	Previously_Insured	int64	0	2	0.0	0.0
4	Annual_Premium	float64	0	48838	40454.0	33536.0
5	Policy_Sales_Channel	float64	0	155	26.0	26.0
6	Vintage	int64	0	290	217.0	183.0
7	Response	int64	0	2	1.0	0.0
8	Vehicle_Damage	int64	0	2	1.0	0.0
9	Vehicle_Age	int64	0	3	2.0	0.0
10	Gender	int64	0	2	1.0	1.0

Data Preprocessing

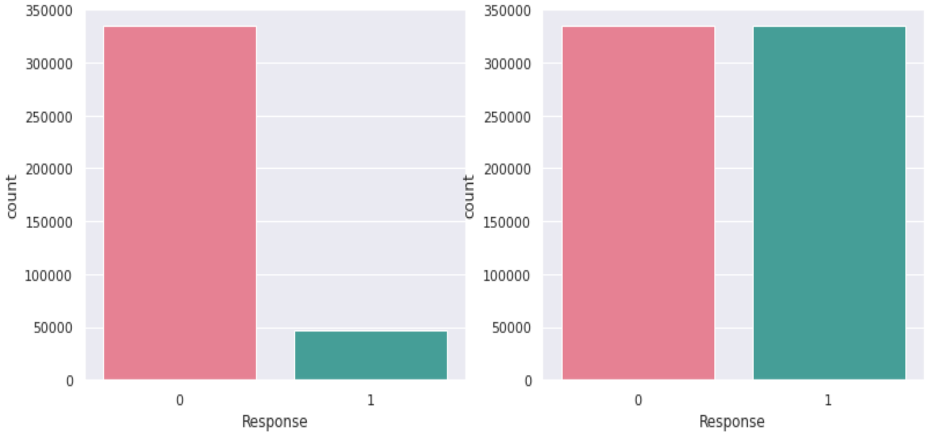
Feature Selection:

- ❖ Extra Tree Classifier
- ❖ We can remove less important features from the data set
- ❖ Driving_License, Gender is contributing very less that's why I'm removing those columns.



Data Preprocessing

- **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 **RandomOverSampler for**
- resampling technique



Model Evaluation

Logistic Regression:

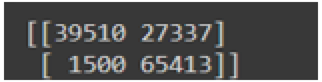
Accuracy : 0.784

Precision: 0.705

Recall: 0.978

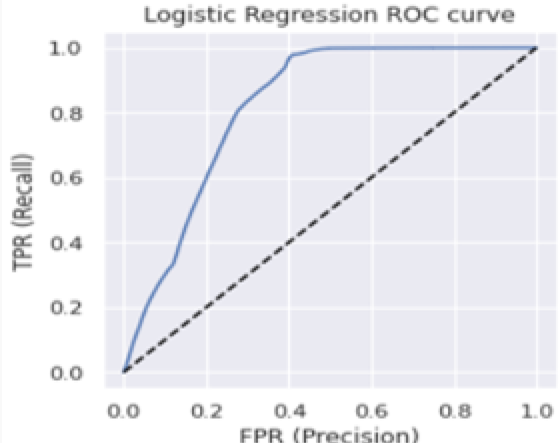
F1-Score: 0.819

ROC_AUC Score: 0.834



```
[[39510 27337]
 [ 1500 65413]]
```

Figure 1: Confusion Matrix (LOGR)



RandomForest Classifier

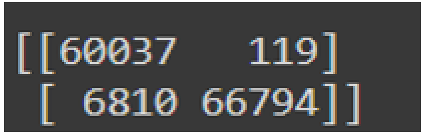
Accuracy : 0.948

Precision: 0.907

Recall: 0.998

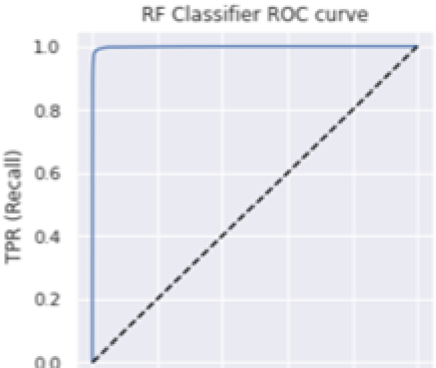
F1-Score: 0.951

ROC_AUC Score: 0.834



```
[[60037   119]
 [ 6810 66794]]
```

Figure 2: Confusion Matrix (RFC)



Model Evaluation

XGBClassifier

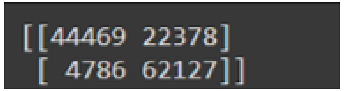
Accuracy : 0.797

Precision: 0.735

Recall: 0.928

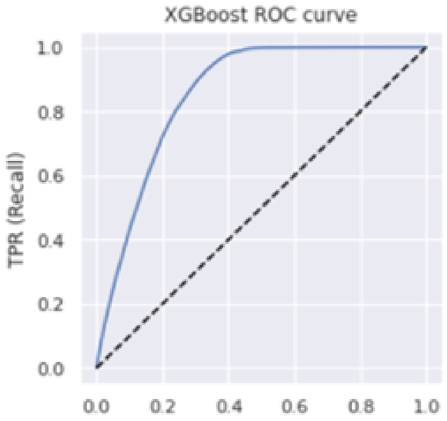
F1-Score: 0.821

ROC_AUC Score: 0.819



```
[[44469 22378]
 [ 4786 62127]]
```

Figure 3: Confusion Matrix (XGBC)



Best Model: RandomForestClassifier is giving highest accuracy , that's why by using GridSearchCV I will set Hyperparameters value

<u>Best Parameters:</u> {'criterion': 'gini', 'max_depth': 50, 'min_samples_split': 2, 'n_estimators': 10}

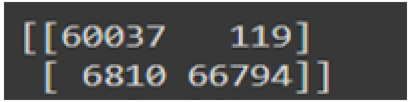
Accuracy : 0.951

Precision: 0.912

Recall: 0.997

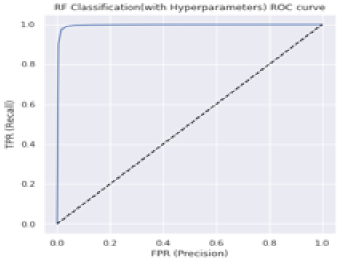
F1-Score: 0.953

ROC_AUC Score: 0.834



```
[[60037   119]
 [ 6810 66794]]
```

Figure 4: Confusion Matrix



As we can see after using Hyperparameters Accuracy , Precision , f1_score , ROC_AUC Increased(tiny change) , Recall decresed. But the change is very low , If u wish then we can ignore it also.

Conclusion

- In this problem statement RandomForest Classifier with the GridSearchCv is the best model for prediction. The important insights of the dataset and the model are Customers of age between 30 to 60 are more likely to buy insurance.
- 1. Customers with Driving License have higher chance of buying Insurance.
- 2. Customers with Vehicle_Damage are likely to buy insurance.
- 3. The variable such as Previously_insured , Vehcile_Damage are more affecting the target variable.
- 4. The variable such as Driving_License , Gender are not affecting the target variable.
- 5. 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.

Future Work

- We can use any deep learning models as a classifier for this insurance prediction

Thank You!