Capstone Project - 3

Health Insurance Cross Sell Prediction

Team Members

Bindu Kovvada
Manoj Patil M
Gulzar
Saksham Tripathi
Deepak Kumar Gautam

Table of content

- Importing the Dataset and Relevant Libraries
- Data Inspection
- Exploratory Data Analysis
- Visualization
- Feature Selection
- Feature Engineering
- Train and Test split
- Model Training
- Model Testing
- Hyper-parameter Tuning
- Cross Validation
- Conclusion (with key and improvement points)

Problem Statement

Our client is an Insurance company that has provided Health Insurance to its customers, and now they need a model to predict whether the policy-holders (customers) will also be interested in Vehicle Insurance provided by the company or not.

Objective: our main objective here is to build a model which can predict based on given data whether the policyholders or customers from past year will also be interested in Vehicle Insurance provided by the company.

About Domain

What is Insurance Policy ?

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’.

What is Premium ?

A ‘Premium’ is a sum of money that the customer needs to pay regularly to an Insurance company for this guarantee.

What is cross-selling ?

Cross-selling is a sales technique involving the selling of an additional product or service to an existing customer.

Data set information

Columns Used:

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

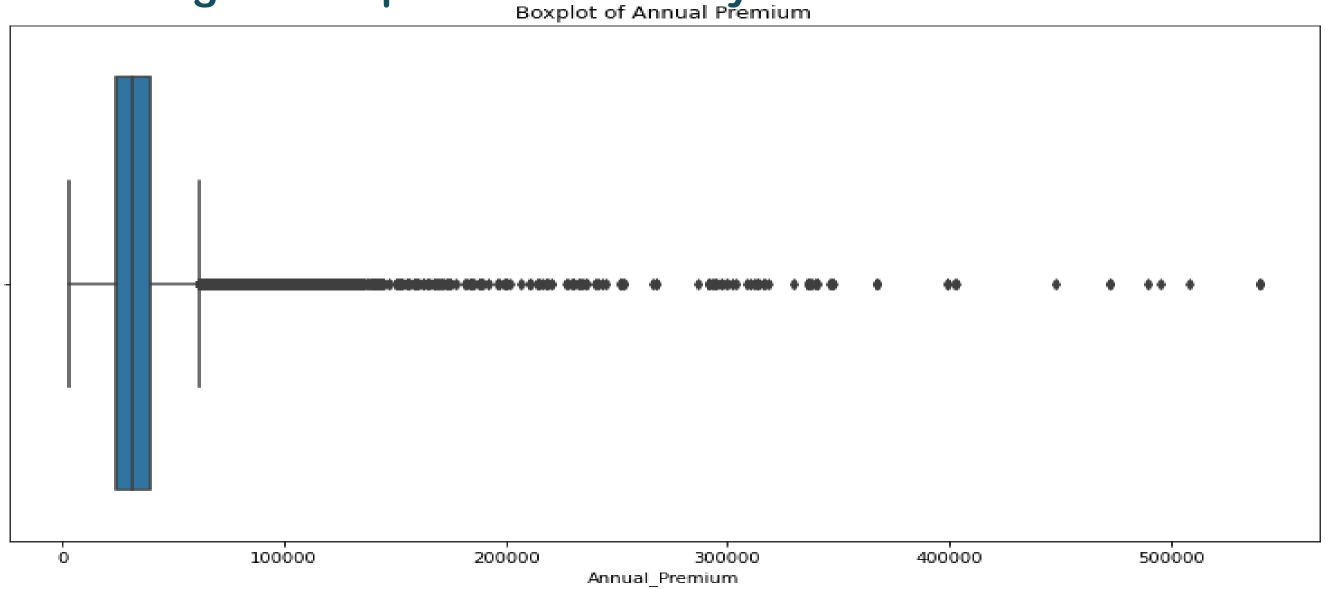
Data Inspection

From the statistics part of our data we can observe that :

- There are 12 columns and 381109 rows in our dataset.
- We have counts of 325634 for not intrested customer and have 45155 for intrested customer, as we can see our data is imbalanced and we have to use some sampling techniques for balancing.
- Id feature doesn’t have any use for our model making so we have to remove it from dataset.
- We have three type of features- Integer, float and Object in our dataset.
- There are only three categorical columns in the dataset Gender, Vehicle-age and Vehicle-damage, We have to encode them.
- There are no null or missing values.
- We can observe from the age feature that the oldest insured client is 85 and the youngest is 20.
- In Annual Premium we have some outlier.
- We doesn’t have any duplicated values as well in our dataset.

Exploratory Data Analysis

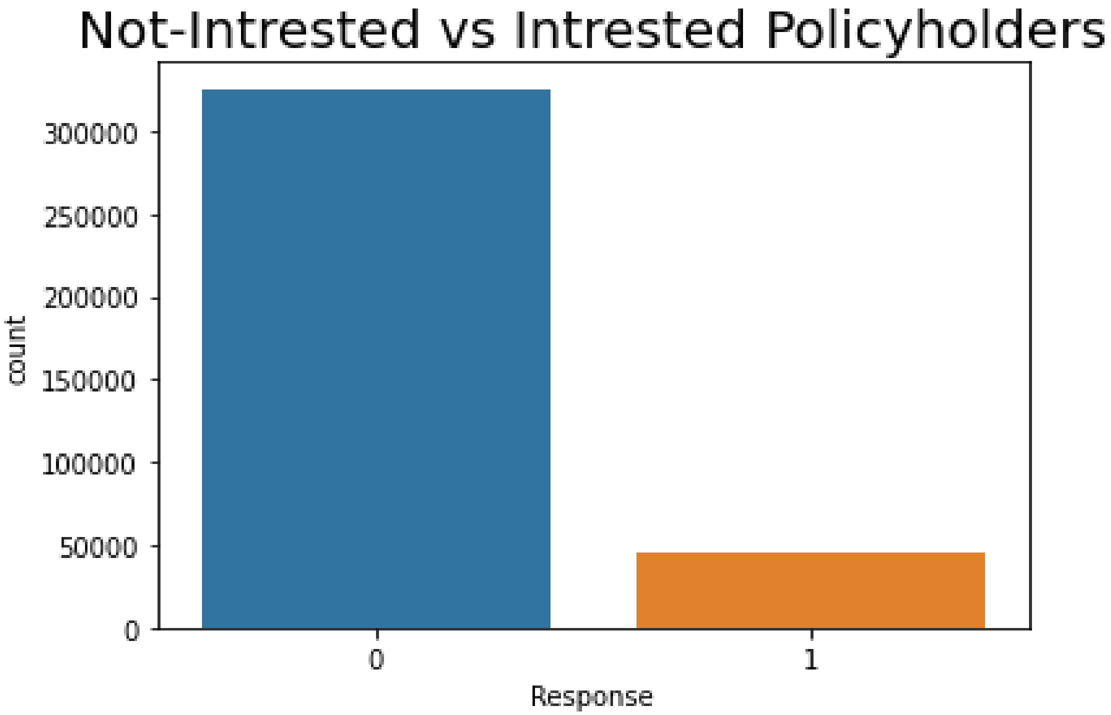
Checking Outlier presence in every features of our dataset



<u>Observation:</u>

After checking all the features for outlier presence we've got only one feature which is Annual_Premium. But it is general to have outlier for annual premium as according to company's policy scheme so we didn't remove outlier in this case.

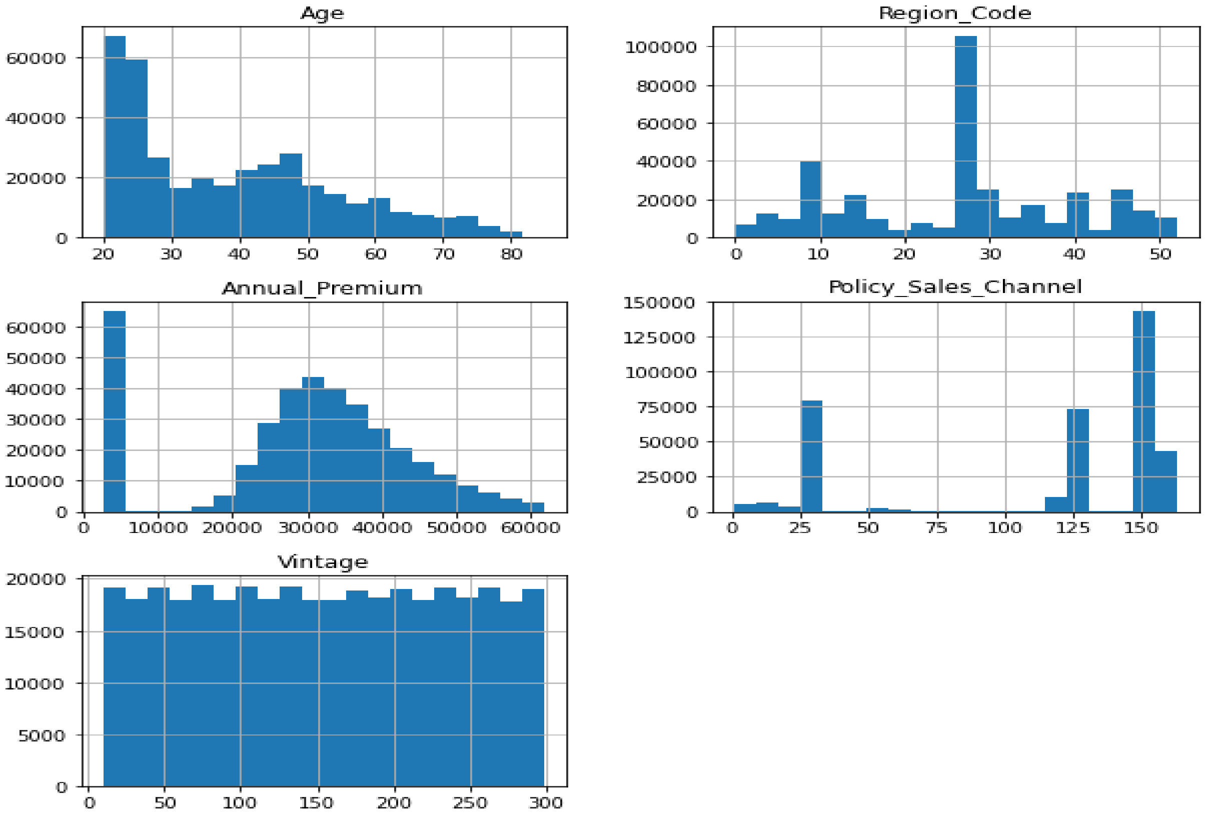
Checking the Balanceness of dataset



Observation

As we can see in this count-plot that our data is unbalanced so we have to make it balanced. For that we used Over-sampling, Under-sampling and SMOTE as well but Oversampling gives promising results than others, So we will proceed further with this technique only.

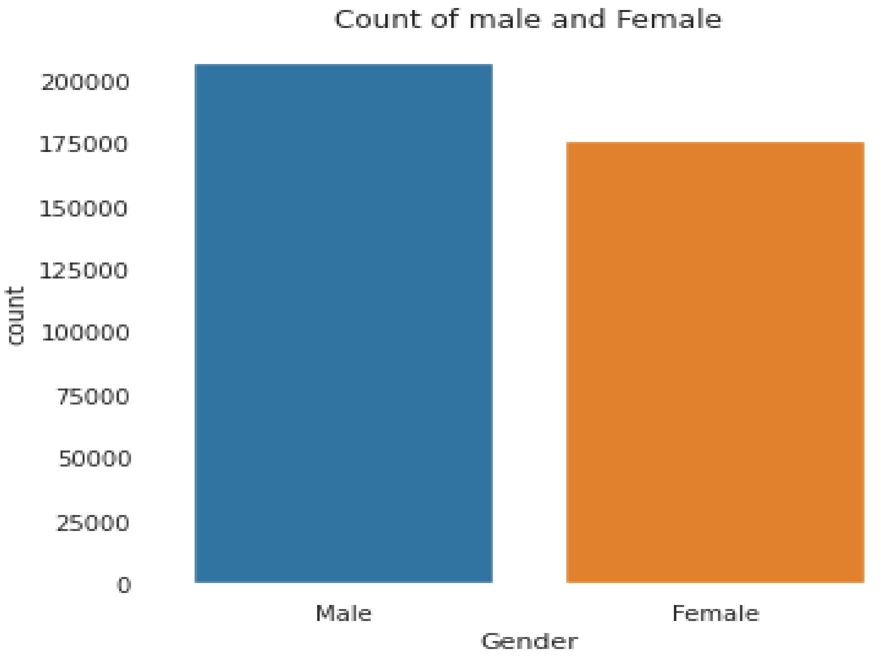
Checking distribution of numerical features

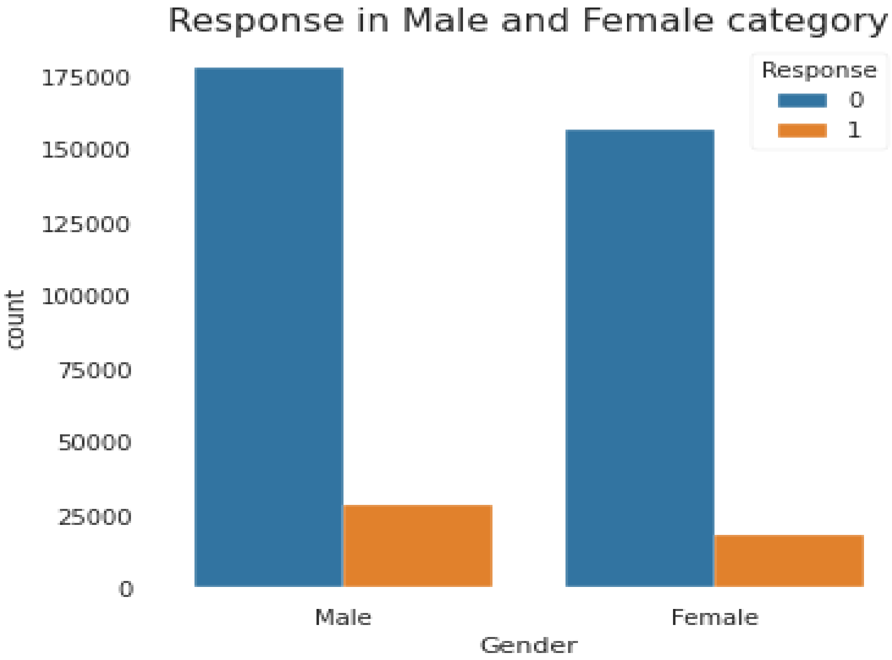


As we can see no feature are normally distributed in Numerical features

Visualization

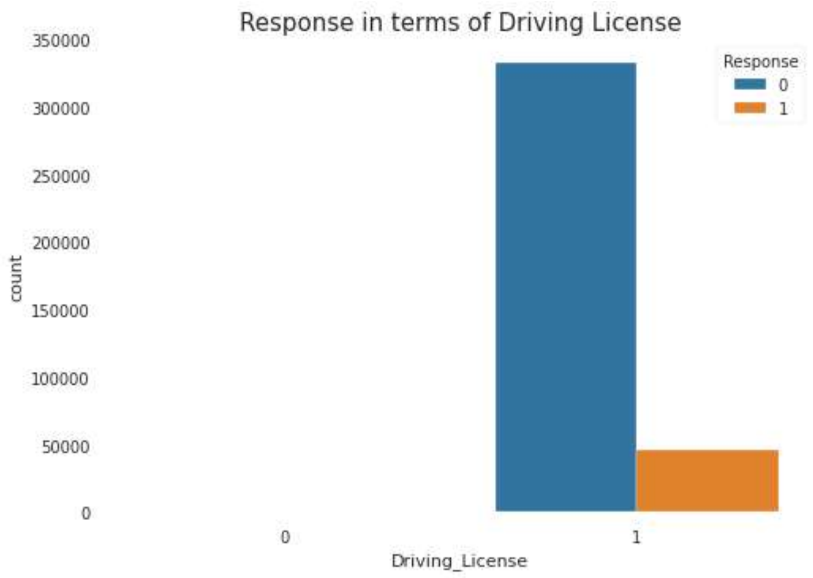
As we can see from below plots we have more number of men than women so we have a gender-gap here.
And as a results males have more intrested than female in their vehicle insurance so we have to target woman more for increasing conversion of women.

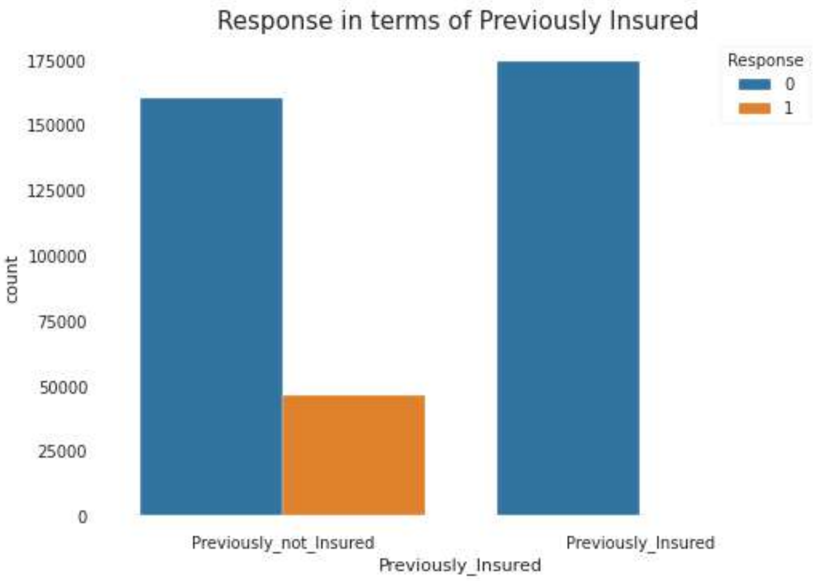




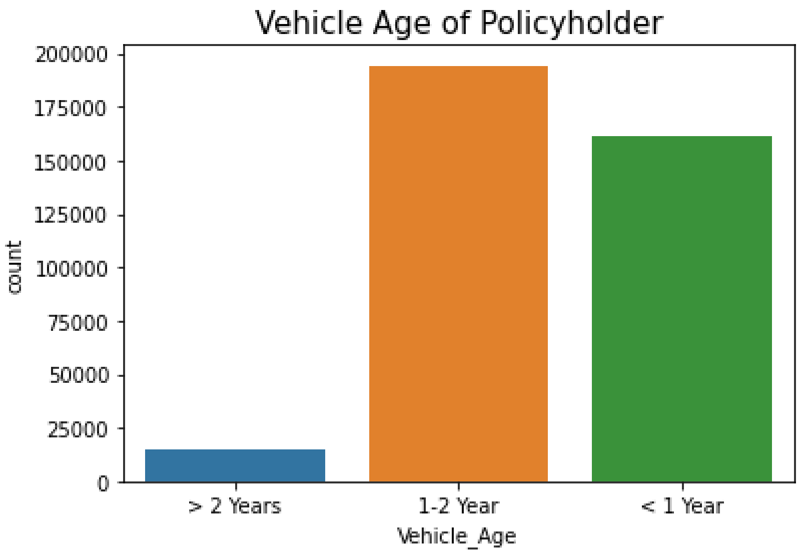
Visualization Continue….

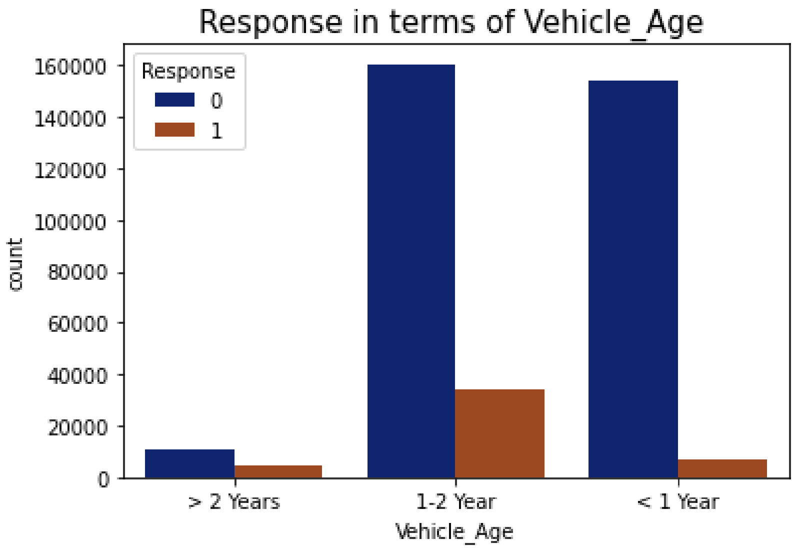
From these below plots we can observe that policyholders who doesn’t have license are not intrested for any vehicle insurance. And those who previously insured their vehicle are also not intrested in any vehicle insurance.





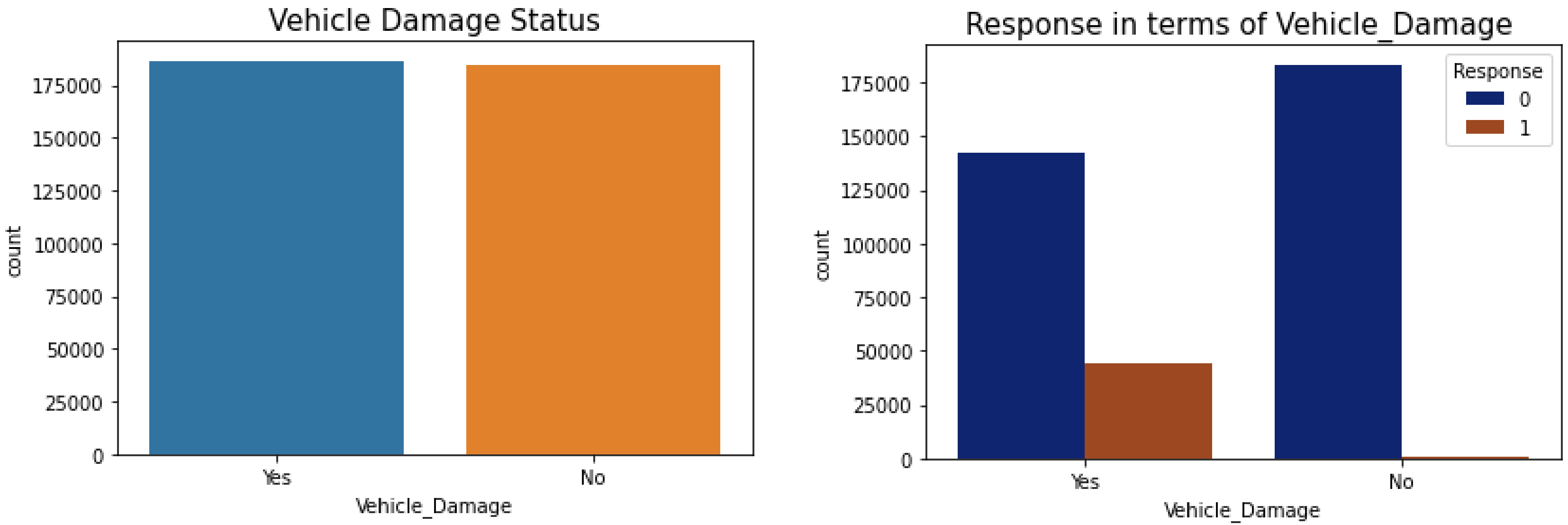
Visualization Continue….





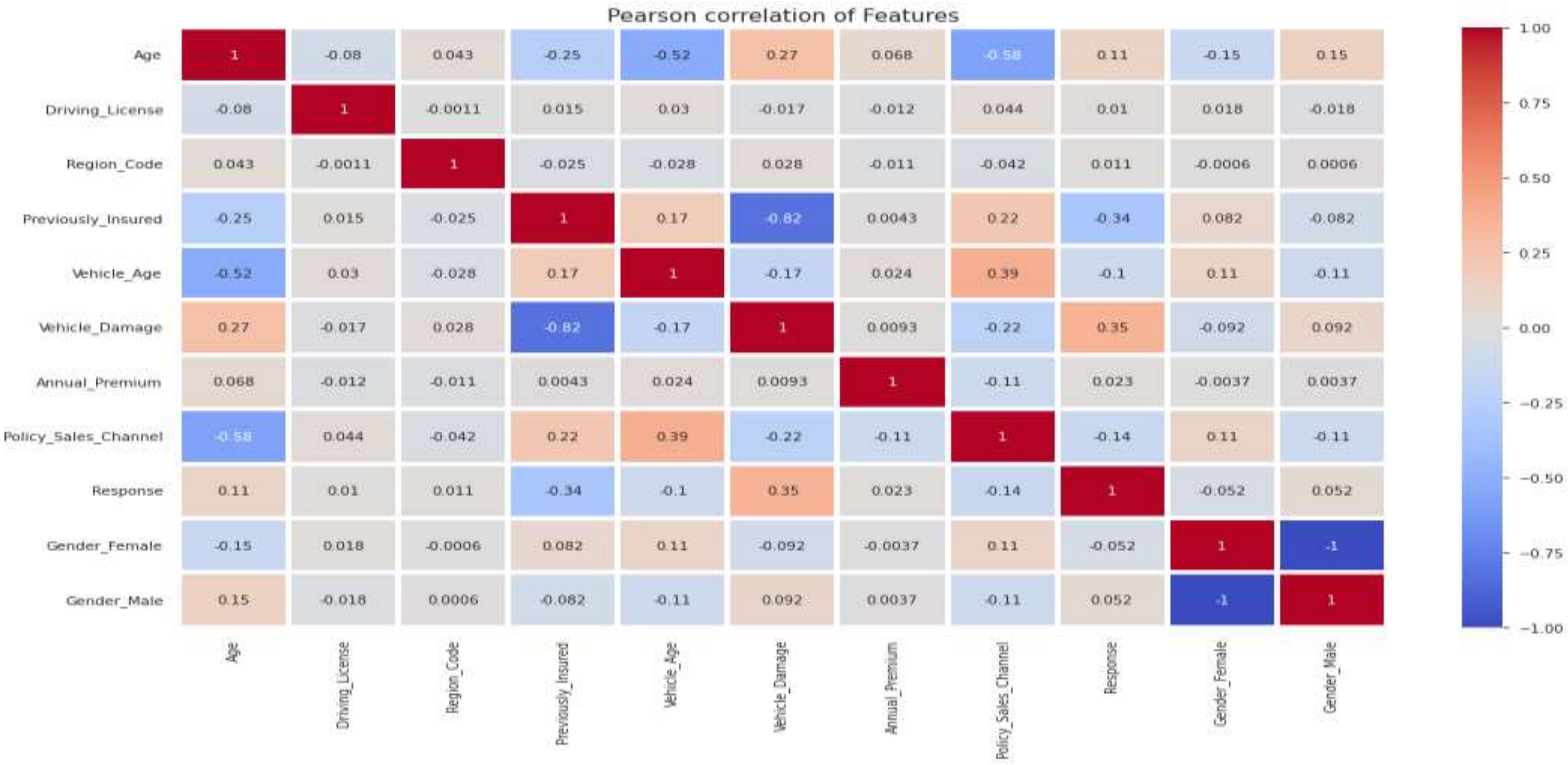
We can clearly see from above plots that we have a lot of policyholders whose vehicle's age is between 1-2 years and this category is also giving good amount of positive response than other policyholders.

Visualization Continue....



We can observe here that we have nearly equal ratio of damaged and non-damaged vehicle's policyholders and we can also see here that customers whose vehicle is damaged in recent years are more likely to buy vehicle insurance.

Correlation of features

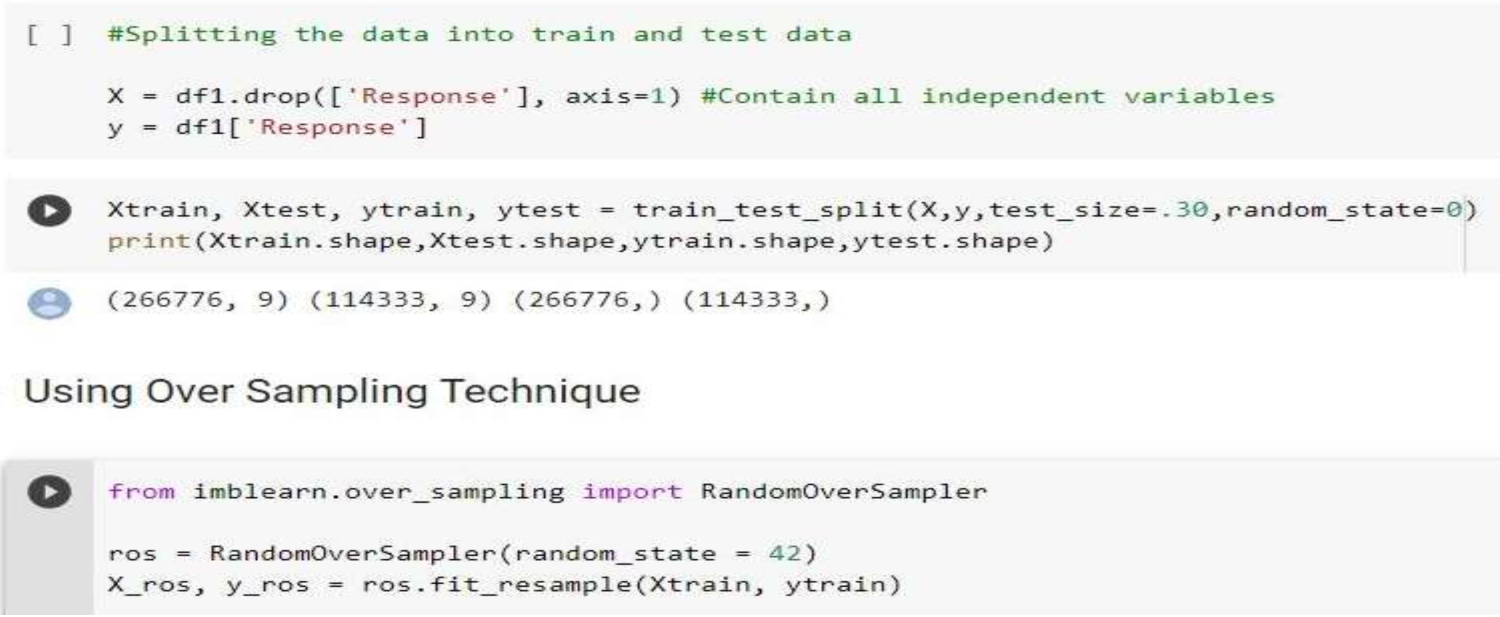


	Age	Driving_License	Region_Code	Previously_Insured	Vehicle_Age	Vehicle_Damage	Annual_Premium	Policy_Sales_Channel	Response	Gender_Female	Gender_Male
Age	1	-0.08	0.043	-0.25	-0.52	0.27	0.068	-0.58	0.11	-0.15	0.15
Driving_License	-0.08	1	-0.0011	0.015	0.03	-0.017	-0.012	0.044	0.01	0.018	-0.018
Region_Code	0.043	-0.0011	1	-0.025	-0.028	0.028	-0.011	-0.042	0.011	-0.0006	0.0006
Previously_Insured	-0.25	0.015	-0.025	1	0.17	-0.82	0.0043	0.22	-0.34	0.082	-0.082
Vehicle_Age	-0.52	0.03	-0.028	0.17	1	-0.17	0.024	0.39	-0.1	0.11	-0.11
Vehicle_Damage	0.27	-0.017	0.028	-0.82	-0.17	1	0.0093	-0.22	0.35	-0.092	0.092
Annual_Premium	0.068	-0.012	-0.011	0.0043	0.024	0.0093	1	-0.11	0.023	-0.0037	0.0037
Policy_Sales_Channel	-0.58	0.044	-0.042	0.22	0.39	-0.22	-0.11	1	-0.14	0.11	-0.11
Response	0.11	0.01	0.011	-0.34	-0.1	0.35	0.023	-0.14	1	-0.052	0.052
Gender_Female	-0.15	0.018	-0.0006	0.082	0.11	-0.092	-0.0037	0.11	-0.052	1	-1
Gender_Male	0.15	-0.018	0.0006	-0.082	-0.11	0.092	0.0037	-0.11	0.052	-1	1

Observations based on correlation plot

- As we can see that no features have highly positive correlation with each others.
- Gender_male and gender female have highly negative correlation so we can remove one of them.
- We have some a small correlation of vehicle_damage with our target feature.
- We have some negative correlation of our target feature with previously_insured, vehicle_age and policy_sales_channel.
- Vehicle_damaged and previously_insured features have highest negative correlation.

Test and Train split



```
[ ]  #Splitting the data into train and test data

     X = df1.drop(['Response'], axis=1) #Contain all independent variables
     y = df1['Response']
```

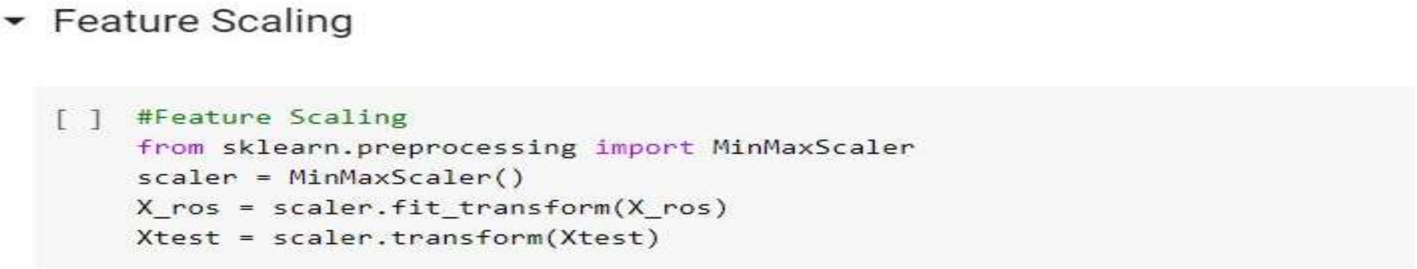
```
Xtrain, Xtest, ytrain, ytest = train_test_split(X,y,test_size=.30,random_state=0)
print(Xtrain.shape,Xtest.shape,ytrain.shape,ytest.shape)
```

```
(266776, 9) (114333, 9) (266776,) (114333,)
```

Using Over Sampling Technique

```
from imblearn.over_sampling import RandomOverSampler

ros = RandomOverSampler(random_state = 42)
X_ros, y_ros = ros.fit_resample(Xtrain, ytrain)
```

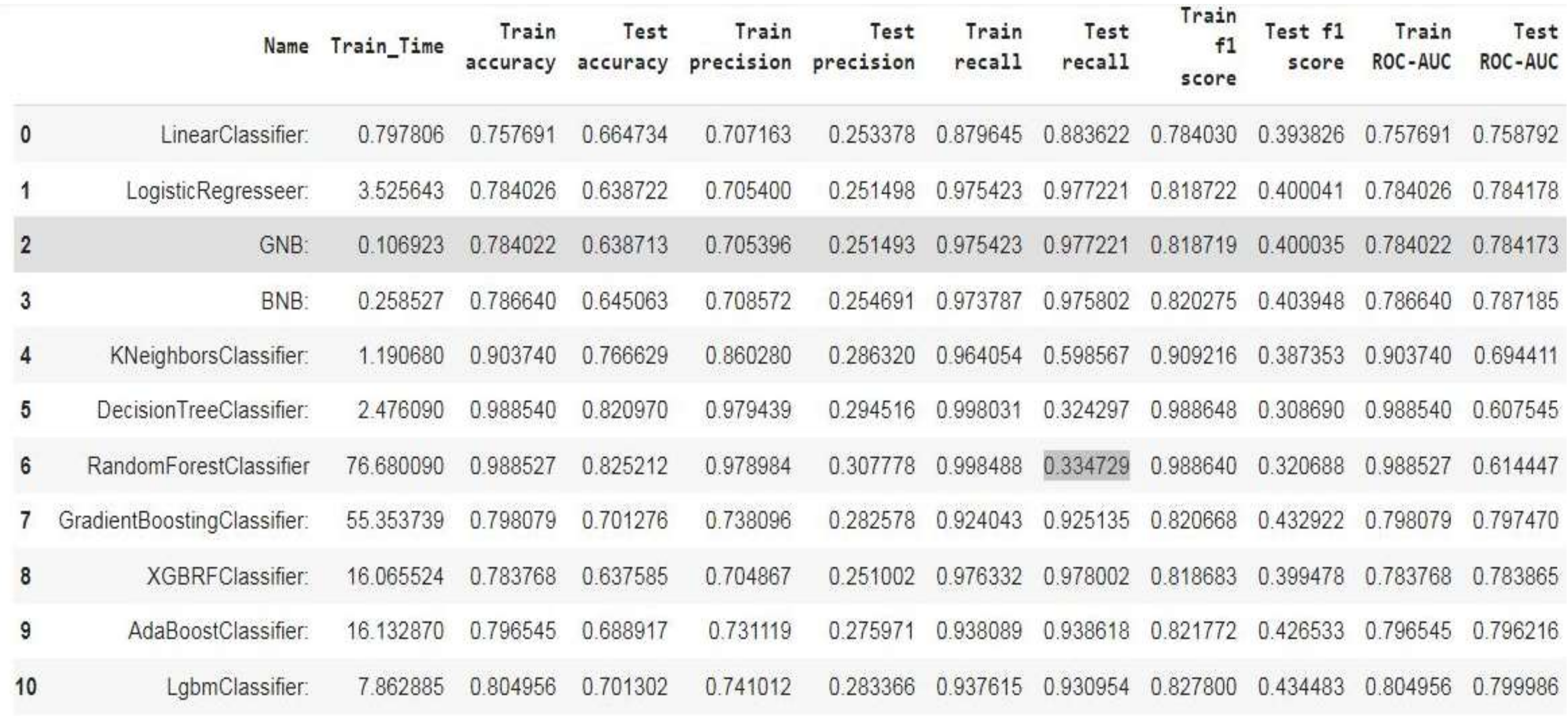


Feature Scaling

```
[ ]  #Feature Scaling
     from sklearn.preprocessing import MinMaxScaler
     scaler = MinMaxScaler()
     X_ros = scaler.fit_transform(X_ros)
     Xtest = scaler.transform(Xtest)
```

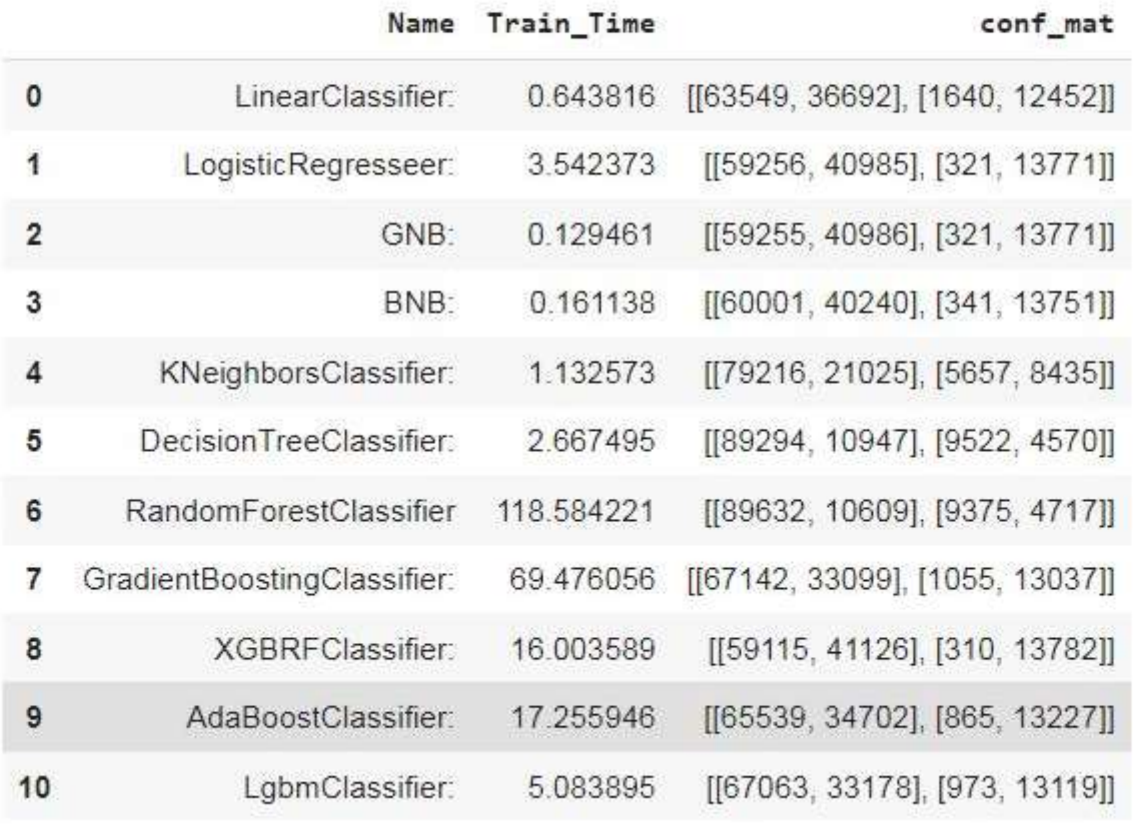
- Scaled down the train varible which makes easy for a model to learn.

Fitting the multiple models



	Name	Train_Time	Train accuracy	Test accuracy	Train precision	Test precision	Train recall	Test recall	Train f1 score	Test f1 score	Train ROC-AUC	Test ROC-AUC
0	LinearClassifier:	0.797806	0.757691	0.664734	0.707163	0.253378	0.879645	0.883622	0.784030	0.393826	0.757691	0.758792
1	LogisticRegresseer:	3.525643	0.784026	0.638722	0.705400	0.251498	0.975423	0.977221	0.818722	0.400041	0.784026	0.784178
2	GNB:	0.106923	0.784022	0.638713	0.705396	0.251493	0.975423	0.977221	0.818719	0.400035	0.784022	0.784173
3	BNB:	0.258527	0.786640	0.645063	0.708572	0.254691	0.973787	0.975802	0.820275	0.403948	0.786640	0.787185
4	KNeighborsClassifier:	1.190680	0.903740	0.766629	0.860280	0.286320	0.964054	0.598567	0.909216	0.387353	0.903740	0.694411
5	DecisionTreeClassifier:	2.476090	0.988540	0.820970	0.979439	0.294516	0.998031	0.324297	0.988648	0.308690	0.988540	0.607545
6	RandomForestClassifier	76.680090	0.988527	0.825212	0.978984	0.307778	0.998488	0.334729	0.988640	0.320688	0.988527	0.614447
7	GradientBoostingClassifier:	55.353739	0.798079	0.701276	0.738096	0.282578	0.924043	0.925135	0.820668	0.432922	0.798079	0.797470
8	XGBRFClassifier:	16.065524	0.783768	0.637585	0.704867	0.251002	0.976332	0.978002	0.818683	0.399478	0.783768	0.783865
9	AdaBoostClassifier:	16.132870	0.796545	0.688917	0.731119	0.275971	0.938089	0.938618	0.821772	0.426533	0.796545	0.796216
10	LgbmClassifier:	7.862885	0.804956	0.701302	0.741012	0.283366	0.937615	0.930954	0.827800	0.434483	0.804956	0.799986

Comparison of Confusion Matrix



	Name	Train_Time	conf_mat
0	LinearClassifier:	0.643816	[[63549, 36692], [1640, 12452]]
1	LogisticRegresseer:	3.542373	[[59256, 40985], [321, 13771]]
2	GNB:	0.129461	[[59255, 40986], [321, 13771]]
3	BNB:	0.161138	[[60001, 40240], [341, 13751]]
4	KNeighborsClassifier:	1.132573	[[79216, 21025], [5657, 8435]]
5	DecisionTreeClassifier:	2.667495	[[89294, 10947], [9522, 4570]]
6	RandomForestClassifier	118.584221	[[89632, 10609], [9375, 4717]]
7	GradientBoostingClassifier:	69.476056	[[67142, 33099], [1055, 13037]]
8	XGBRFClassifier:	16.003589	[[59115, 41126], [310, 13782]]
9	AdaBoostClassifier:	17.255946	[[65539, 34702], [865, 13227]]
10	LgbmClassifier:	5.083895	[[67063, 33178], [973, 13119]]

As we can see here that Random forest is giving lesser FN as lesser FN is important in our case it means we only make 10k (out of 114k) wrong prediction of Positive response which is least in comparison to others models.

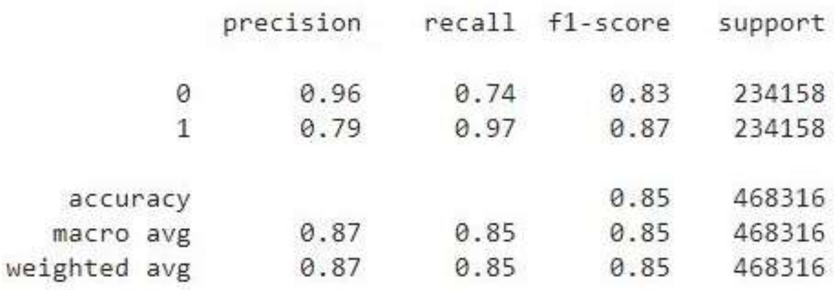
Observation

- From previous results we can see LinearClassifier is not performing good at all.
- XGBClassifier is worst than any other models as it predict max 41k wrong prediction of positive response.
- GNB(Gaussion) and BNB(Bernoulli) also doesn’t performing well.
- RandomForest Classifier is giving promising results than other models.

- **By comparing these models, RandomForest Classifier is performing well till yet now let’s do some hyper-parameter tuning for top models and compare the results further.**

Compare Classification Report of top 2 models

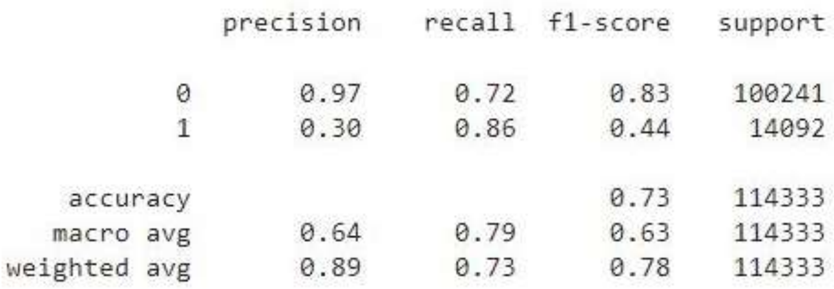
Random Forest Report-



```
              precision    recall  f1-score   support

           0       0.96      0.74      0.83    234158
           1       0.79      0.97      0.87    234158

    accuracy                           0.85    468316
   macro avg       0.87      0.85      0.85    468316
weighted avg       0.87      0.85      0.85    468316
```

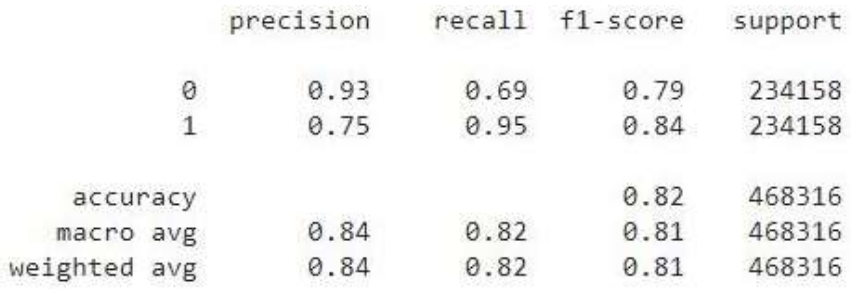


```
              precision    recall  f1-score   support

           0       0.97      0.72      0.83    100241
           1       0.30      0.86      0.44     14092

    accuracy                           0.73    114333
   macro avg       0.64      0.79      0.63    114333
weighted avg       0.89      0.73      0.78    114333
```

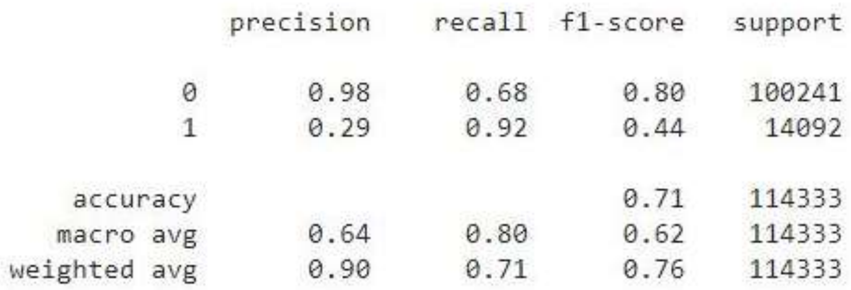
Lgbm Report -



```
              precision    recall  f1-score   support

           0       0.93      0.69      0.79    234158
           1       0.75      0.95      0.84    234158

    accuracy                           0.82    468316
   macro avg       0.84      0.82      0.81    468316
weighted avg       0.84      0.82      0.81    468316
```

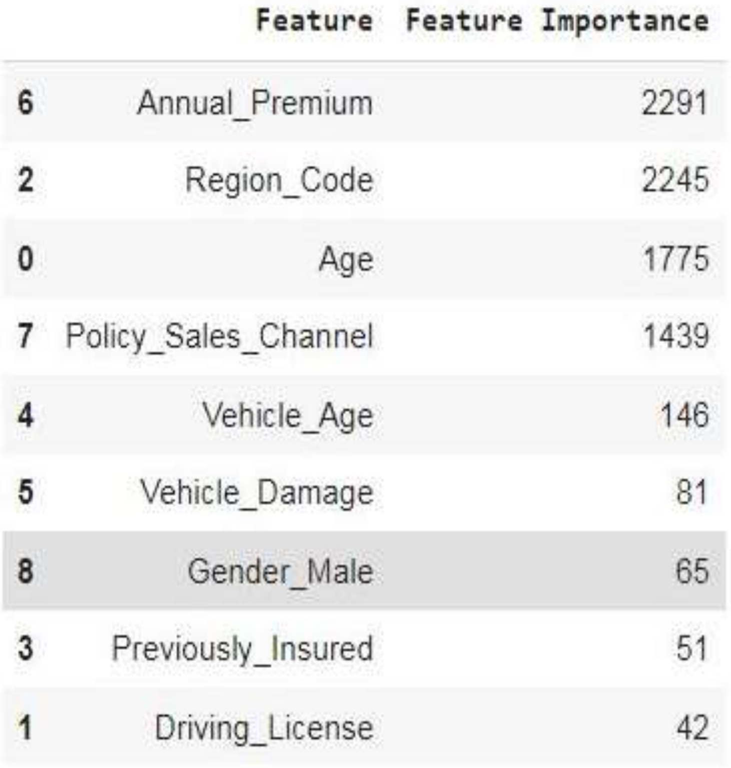


```
              precision    recall  f1-score   support

           0       0.98      0.68      0.80    100241
           1       0.29      0.92      0.44     14092

    accuracy                           0.71    114333
   macro avg       0.64      0.80      0.62    114333
weighted avg       0.90      0.71      0.76    114333
```

Feature Importance



	Feature	Feature Importance
6	Annual_Premium	2291
2	Region_Code	2245
0	Age	1775
7	Policy_Sales_Channel	1439
4	Vehicle_Age	146
5	Vehicle_Damage	81
8	Gender_Male	65
3	Previously_Insured	51
1	Driving_License	42

We can see the feature importance of our dataset via this table itself in decreasing order. The more the weightage the more that feature is important to our final models. So Annual_premium, Region_code, Age, Policy_sales_channel and Vehicle_age are our top five features.

Conclusion

Our client is an insurance firm that has supplied Health Insurance to its customers. They now need assistance in developing a model to predict whether the policyholders (customers) from the previous year will be interested in the company's Vehicle Insurance.

Building a model to predict if a client is interested in Vehicle Insurance is extremely beneficial to the company because on the basis of that they can plan communication strategy to reach out to those customers and optimize its business model and revenue accordingly.

Starting with loading the data so far we have done EDA , null values treatment, dropping unnecessory columns, outliers handling, visualization, knowing the distribution, feature engineering, Applying some sampling technique(OS), model making, finalizing our best model and then we do some hyperparameter tuning also.

The Lgbm Classifier was the best model when compared with rest all models for this data set. For all the models This Classifier worked the best because it has highest recall in comparison to other models which is important to us in this project..

It gives 0.95 recall on train and around 0.92 recall on test data for positive response which can be good enough.

Key Points

- As a feature, id are extremely undervalued.
- Customers of age between 30 to 60 are more likely to buy insurance.
- Customers with Vehicle_Damage are likely to buy insurance.
- Customers with Driving License have higher chance of buying Insurance.
- The variable such as Annual_premium, Region_code and Age are three most important feature which is more affecting the target variable.
- We can say that we have less number of policyholders who has vehicle older than 2 years so we have to focus more on other two category.
- Customer who already secured their vehicle, are clearly not intrested in our company's vehicle insurance scheme.
- We can see that **LGBM model perform better** for this dataset.

Improvement points:

- By using a marketing and advertising approach, we can reduce the gender gap.
- We can clearly see that we have a larger number of consumers without vehicle insurance, therefore we can easily target them directly with our campaign.
- Since there are less policy holders with vehicles older than two years, we must pay more attention to the other two categories (1-2 years and >1 year). Because most sales agencies that offer vehicle insurance for the first year are actually our target and we can give them the best incentives to reduce competition in the market.
- As we saw that we have nearly equal policy holders for both vehicle damage status, so we can target those policy holders whose vehicles are damaged in the past.
- We have to focus more on previously not insured vehicle on our compaign because they are more prone to buy vehicle insurance.

THANK YOU