Capstone Project-III

Classification-HEALTH INSURANCE CROSS SELL Prediction

By- Ajay Tiwari

CONTENTS

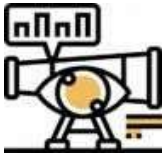
1. Problem Statement
2. Data observations & cleaning
3. EDA
4. Preparing dataset for modelling
5. Data Modelling
6. Challenges
7. Conclusion

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.
- 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.
- 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.

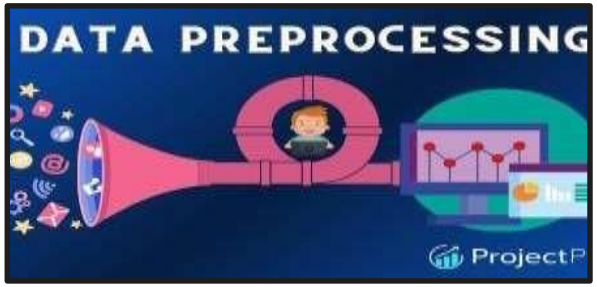
DATA OBSERVATIONS & CLEANING

1) The shape of our dataset is 381109 rows and 12 columns
2) Datatype of Date is given as object which we need to change that to Date Time
3) From the statistical information we can see that it is not a normal distribution
   as mean and 50% values are having a lot of difference
5) There are no duplicates present
6) There are no null values present



```
<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
```

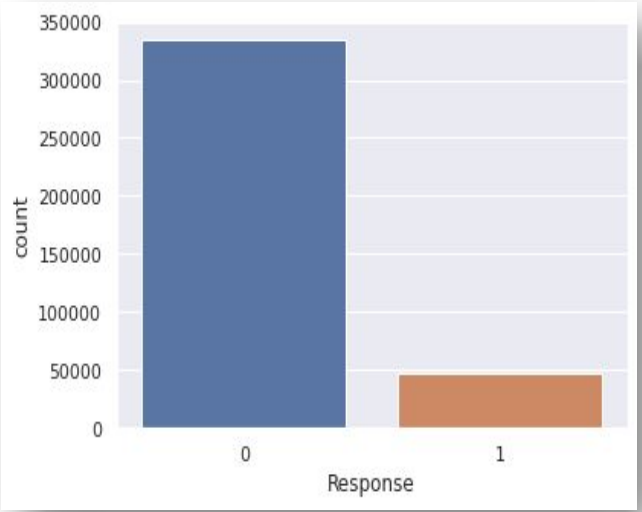
EDA



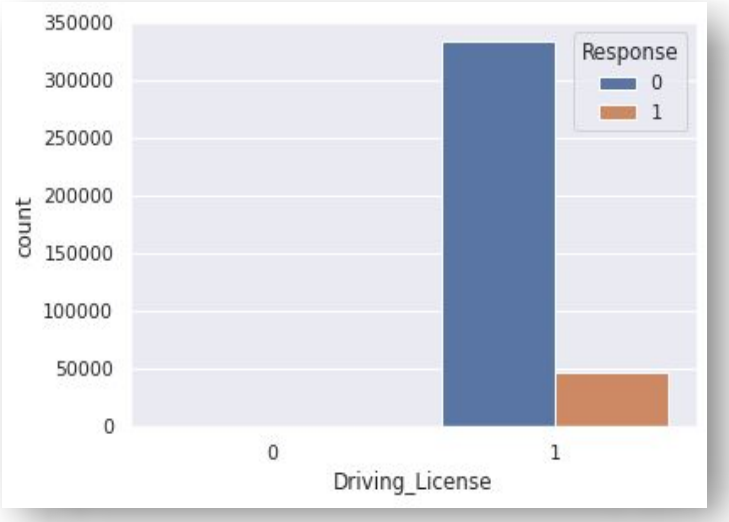
In Exploratory Data Analysis we try to find some insights from the data using visualizations. It consists of few steps:

EDA (Visualizations)

0: indicate Customers are not interested
1: indicate Customers are 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



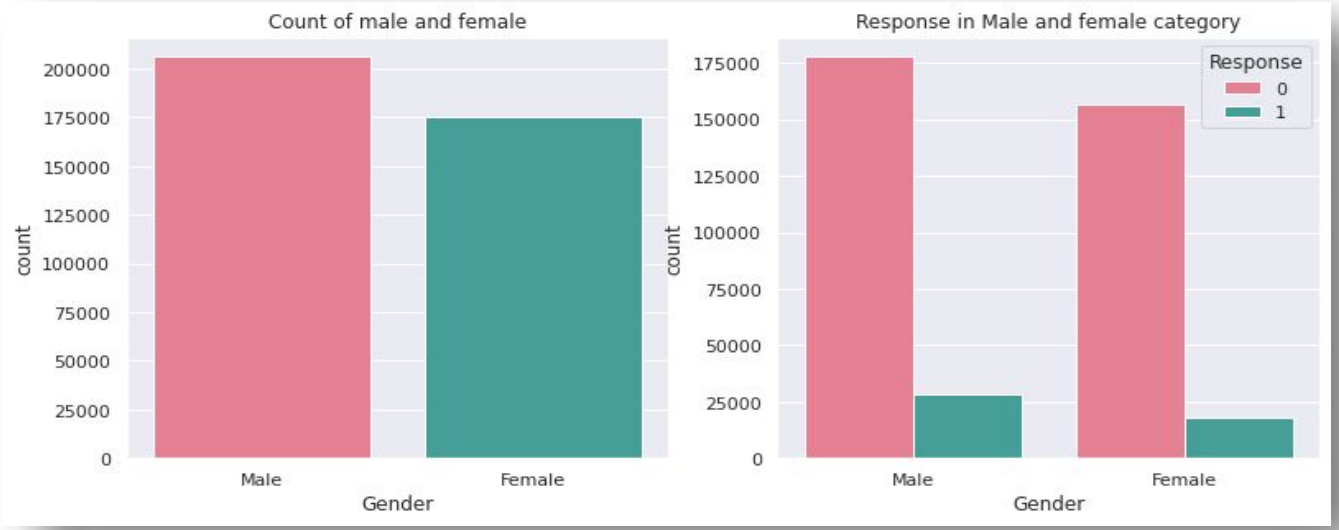
0: indicate Customers are not interested 1: indicate Customers are interested
Customers who are interested in Vehicle Insurance almost all have driving license



EDA (Visualizations)

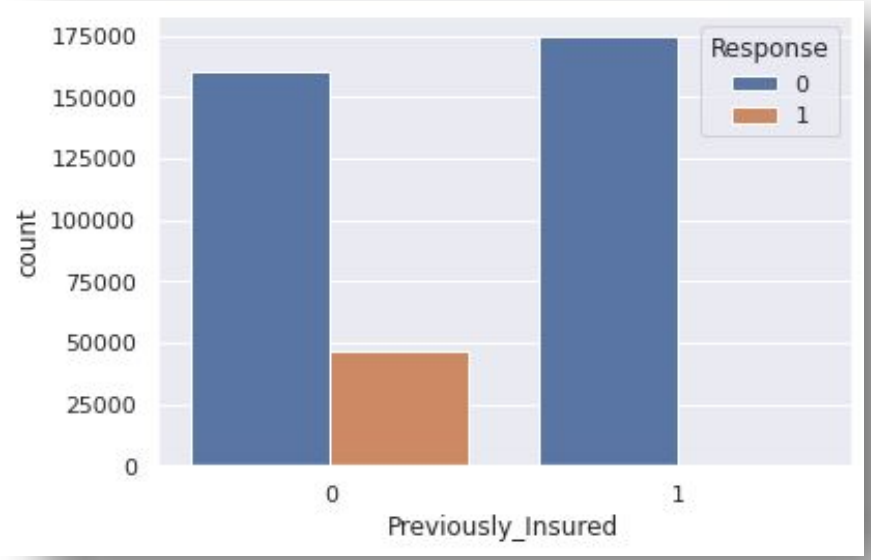
As you can see below the graph:

- 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

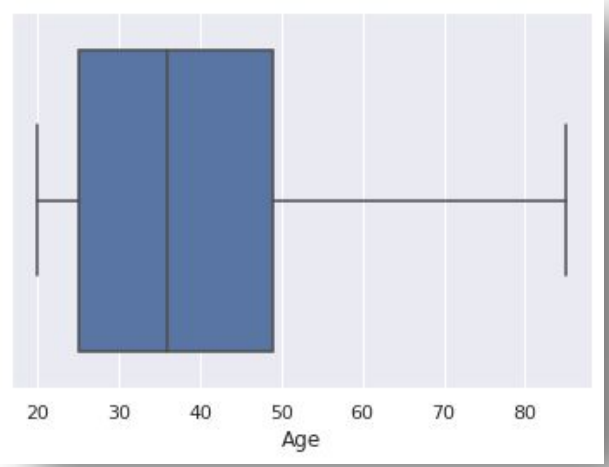


EDA (Visualizations)

- Customer who are not previously insured are likely to be interested.
  - 1 indicate no one interested and the value of this is null.

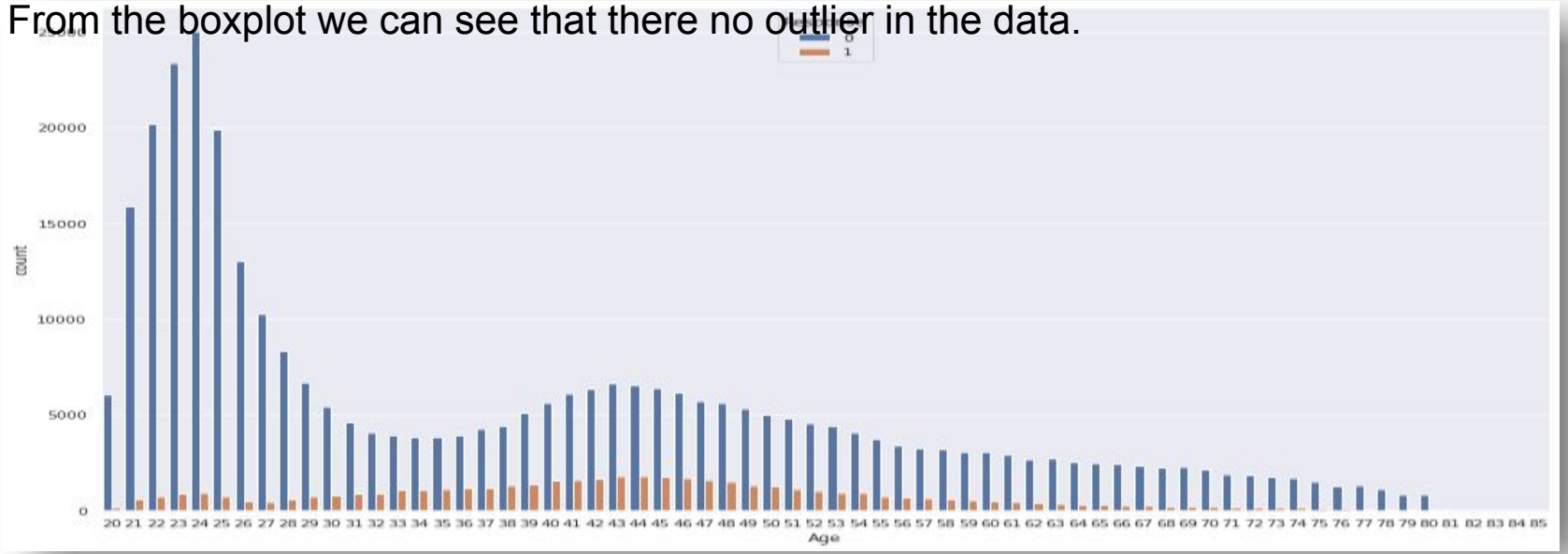


- There is no outliers.

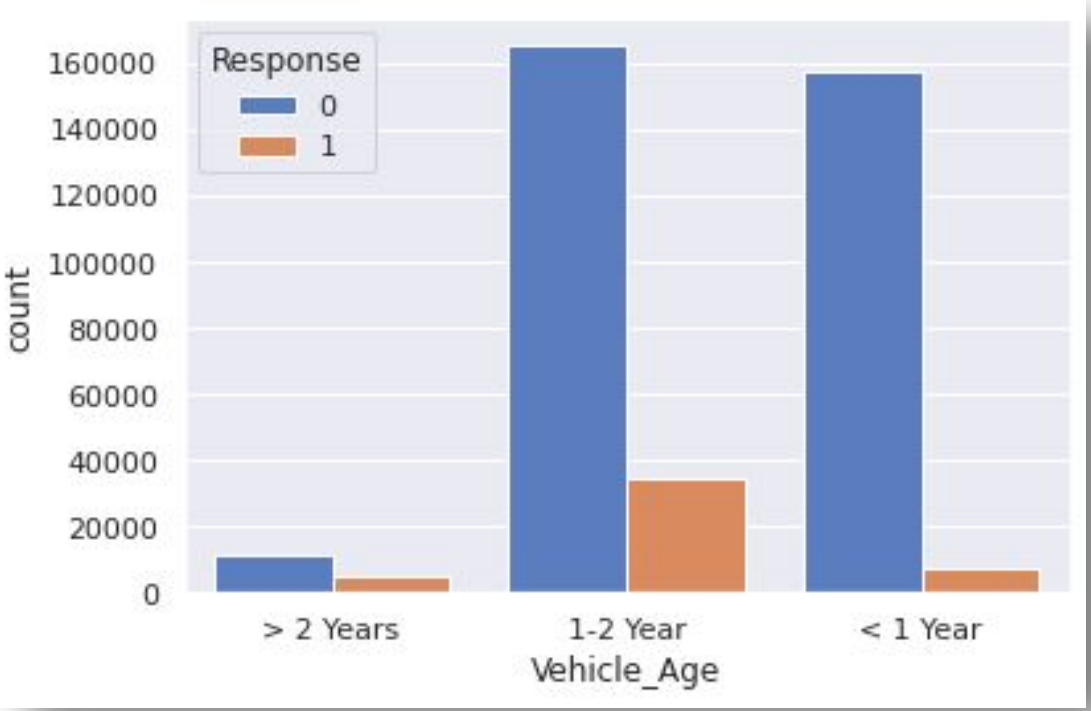


EDA (Visualizations)

- 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.

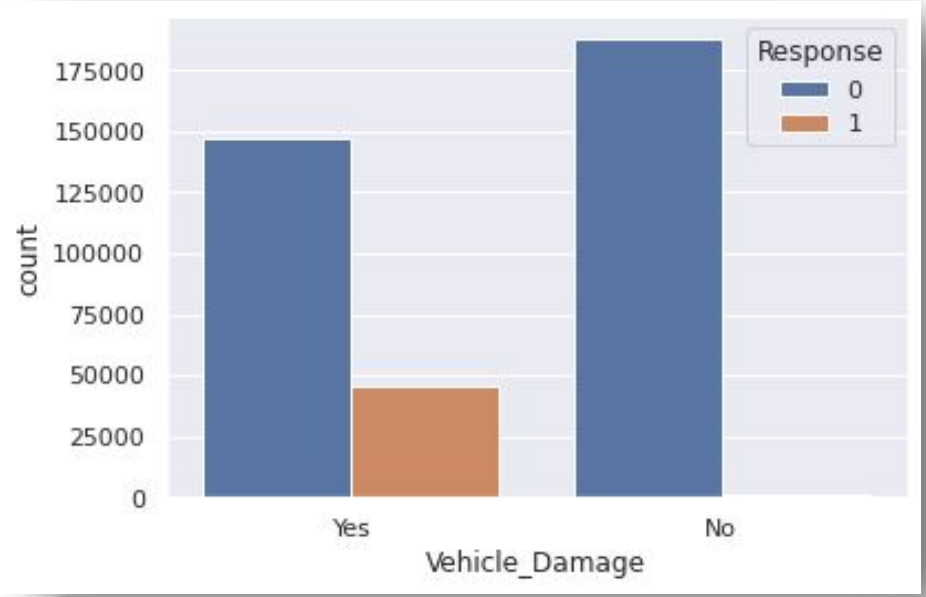


EDA (Visualizations)



- Customers with vehicle age 1- 2 years are more likely to interested as compared to the other two.
- Customers with Vehicle Age <1 years have very less chance of buying Insurance.

EDA (Visualizations)

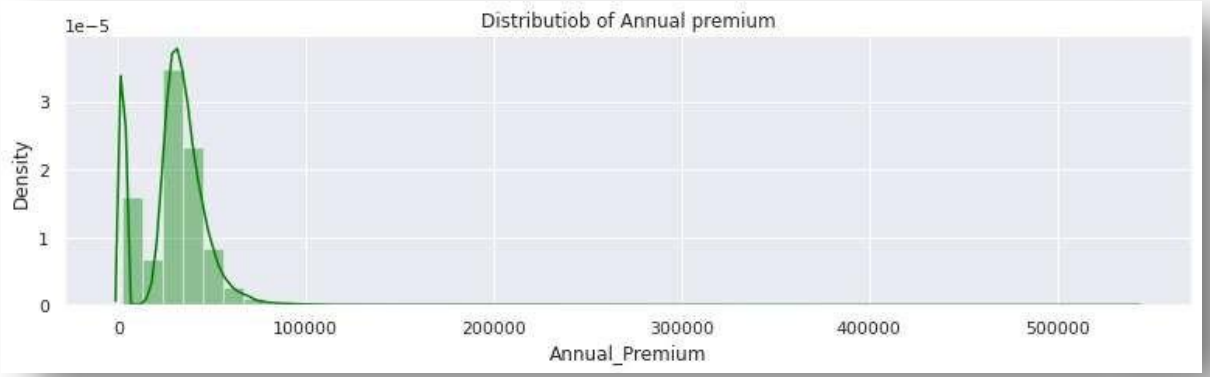


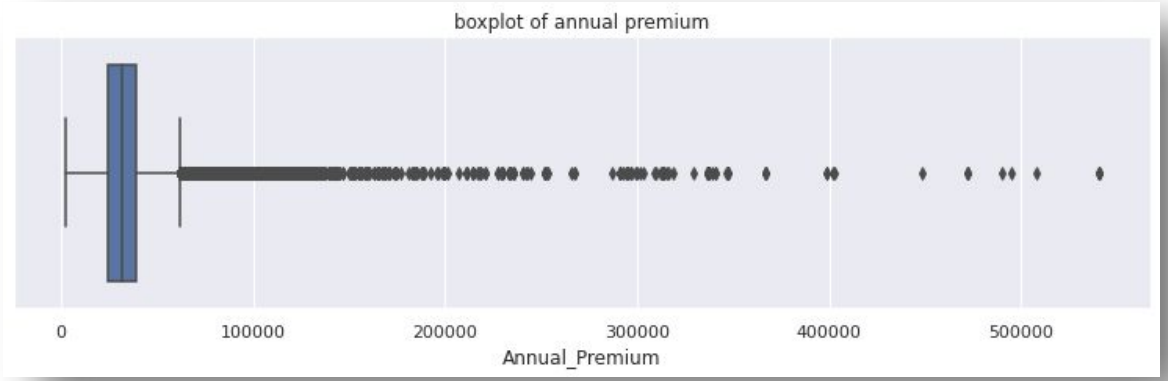
•From the above plot, we can infer that if the vehicle has been damaged previously then the customer will be more interested in buying the insurance as they know the cost.

• 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

EDA (Visualizations)





- From distribution plot we can see annual premium is rightly skewed.
- From boxplot we can see there are too much outliers.

EDA (Visualizations)

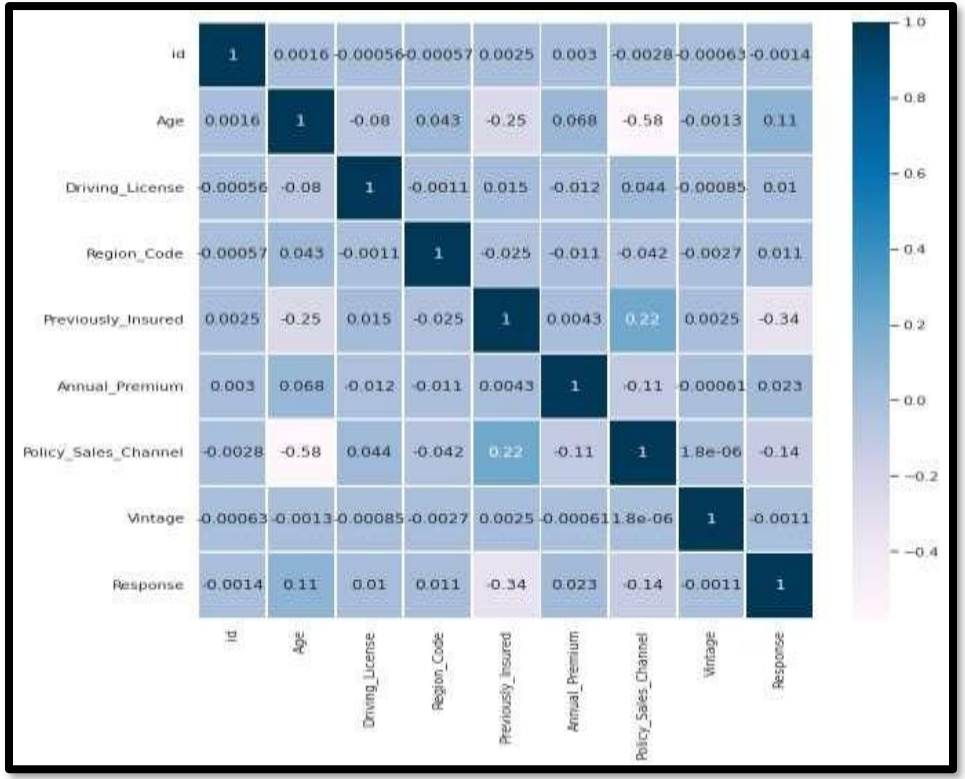
Heat maps

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

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

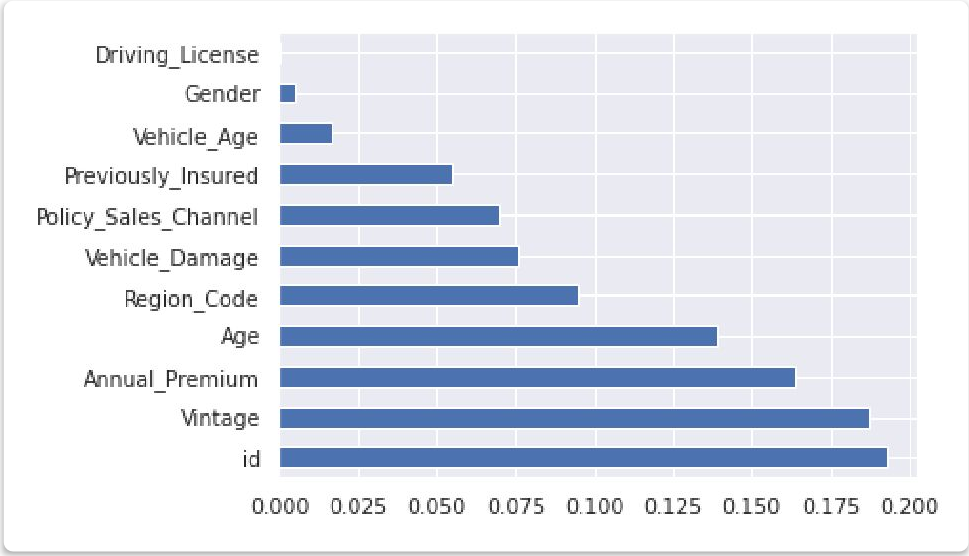
- Target variable is not much affected by id , Vintage variable, Policy sales channel. we can drop least correlated variable

- As we can see the heat map graph Vehicle Damage column is more



Preparing dataset for modelling

Feature selection

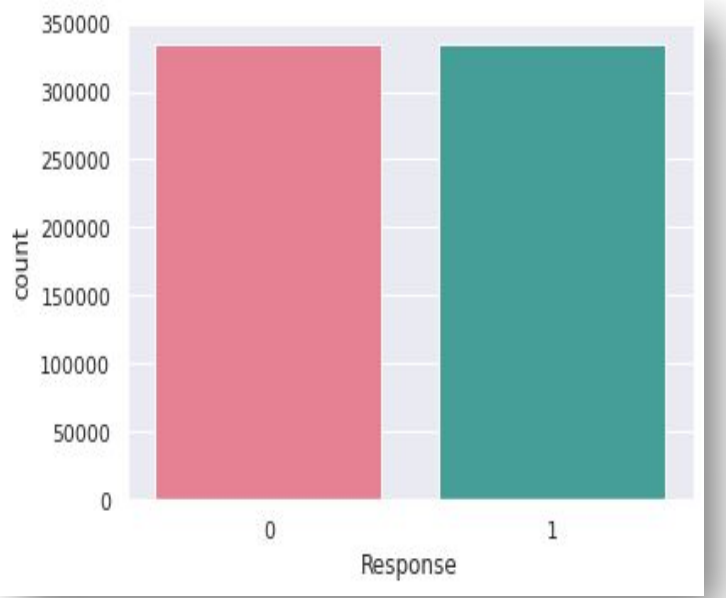


- The features you use influence more than everything else the result. No algorithm alone, to my knowledge, can supplement the information gain given by correct feature engineering.

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

Preparing dataset for modelling

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 sampling.

- Oversampling methods duplicate or create new synthetic examples in the minority class

Data Modelling

Model selection

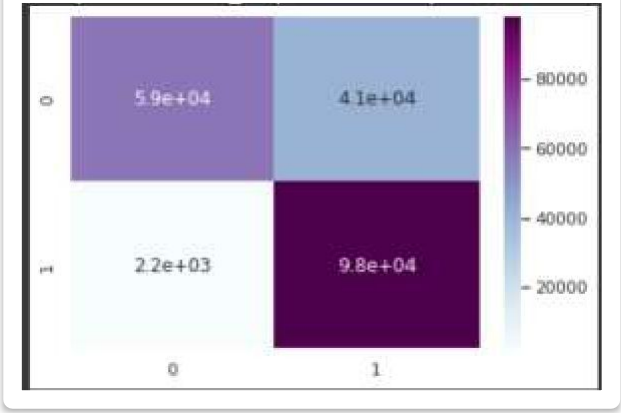
The models in use are:

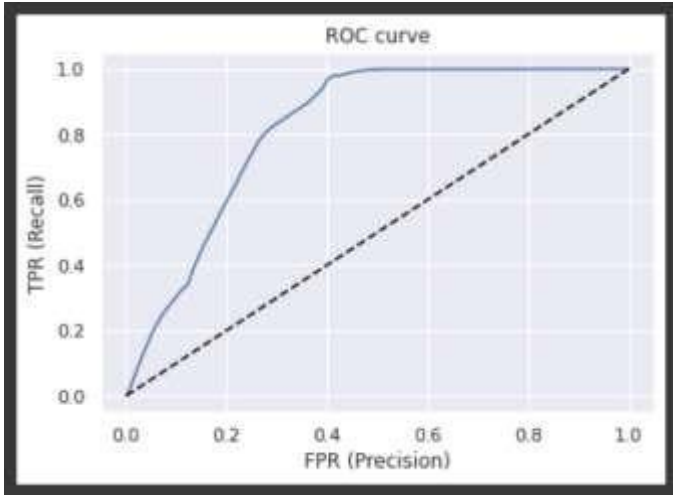
- Logistic Regression Model
- Random Forest Classifier Model
- XGBoost Classifier Model

Data Modelling

Logistic Regression

	precision	recall	f1-score	support
0	0.59	0.96	0.73	61444
1	0.98	0.70	0.82	139196
accuracy			0.78	200640
macro avg	0.78	0.83	0.78	200640
weighted avg	0.86	0.78	0.79	200640



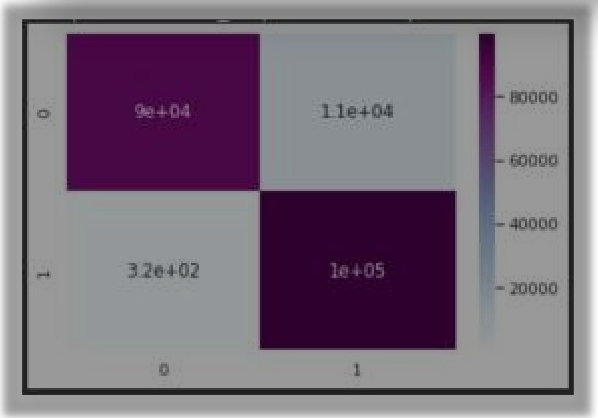


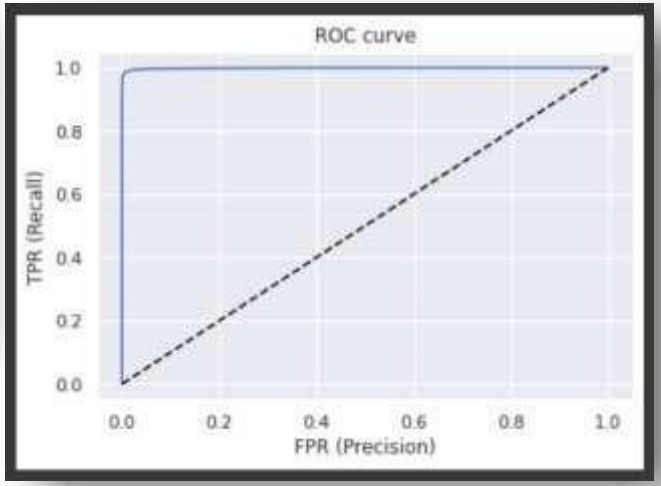
Accuracy : 0.7837719298245615
ROC_AUC Score: 0.833934962505732

Data Modelling

Random forest Classifier

	precision	recall	f1-score	support
0	0.89	1.00	0.94	89817
1	1.00	0.90	0.95	110823
accuracy			0.94	200640
macro avg	0.94	0.95	0.94	200640
weighted avg	0.95	0.94	0.94	200640



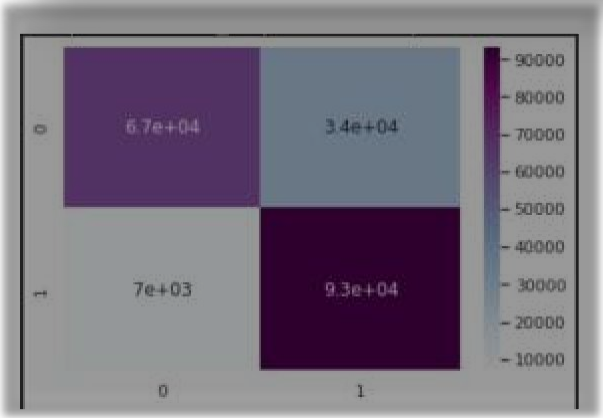


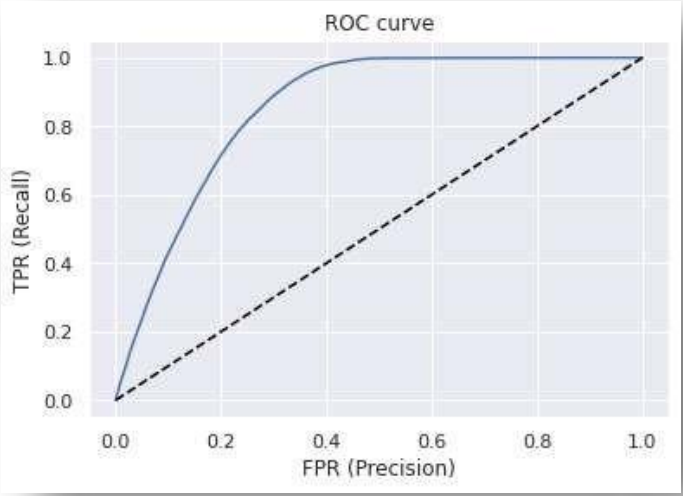

```
Accuracy :  0.9444527511961722
ROC_AUC Score: 0.9493826256806724
```

Data Modelling

XGB Classifier

	precision	recall	f1-score	support
0	0.66	0.91	0.77	73674
1	0.93	0.74	0.82	126966
accuracy			0.80	200640
macro avg	0.80	0.82	0.79	200640
weighted avg	0.83	0.80	0.80	200640





Accuracy Score: 0.7973484848484849
ROC_AUC Score: 0.8197333207905508

Model Metrics Dataset

	Accuracy	Recall	Precision	F1Score	ROC_AUC
Logistic regression	0.783936	0.977899	0.704559	0.819025	0.834229
Randomforest	0.945200	0.997099	0.903328	0.947900	0.950046
XGBClassifier	0.797732	0.932321	0.734568	0.821714	0.820989

- 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 perform better. Because curves closer to the top-left corner, it indicate a better performance.

Challenges

- Identify the highly imbalance dataset and manage it carefully.
- Ensure model is treating all groups fairly.
- Before deploy model set prediction thresholds.
- After deploying understand model behavior on real data.
- Tough to make prediction analysis to end users.

Conclusion

- 87.7% customers responded as No for buying a vehicle insurance. It clearly shows that most of the customers are not interested in buying a vehicle insurance.
- Males are 30% more likely to respond as yes for vehicle insurance than females. So company could focus more on targeting male customers and do more promotions targeted towards the female customers.
- Most of the customers have driving license and out of them 12% are likely to respond as yes for vehicle insurance.
- There is no point in reaching out to customers who already have vehicle insurance as almost all of them responded negatively for buying another insurance.
- 22% of customers responded positively who don't have previous insurance. So, company should focus more such customers as conversion possibility is higher is such cases.
- Company should focus on customers whose vehicle is more than 2 years old, as 30% of times they are interested in buying an insurance, which is huge compared to other features.
- Customers with vehicle age less than an year are least interested in insurance as while buying the vehicle people often buy 1 year insurance. Company shouldn't spend more time on these customers as just 4% of times they are likely to say Yes for a vehicle insurance.
- Customers who damaged their vehicles in past are more sensitive towards buying a vehicle insurance. In fact 24% of times they responded positively based on this dataset.

Conclusion(Contd.)

- Customers who haven't damaged their vehicle in past, almost all of the times they respond as No for insurance. In order to increase the customer base company could focus on conveying importance of a vehicle insurance to such customers.

- Number of days customer associated with company has no impact on response by customers. Company should try building rapport, trust with old customers and could offer them extra perks while buying new products.

- Based on our data, customers in the age group 31 to 60 have very high positive response rate compared to the younger and older customers. We also saw that customers in this age group are more likely to damage their vehicle and people with damaged vehicles are more likely to buy a vehicle insurance. So, this is a very good filter for company to target customers with high conversion rate.

- Vehicle damage, previously insured, policy sales channel, age etc. are the most important features for predicting the response.

- Driving license, gender, vintage etc. features have no significant impact on predicting the response. This dataset is the clear case of imbalance and we applied oversampling techniques such as SMOTE to help us improve the training data and hence the model prediction.

- Logistic regression has highest recall for test data. So, if company needs a very high recall rate, i.e. lowest False Negatives then they may consider using logistic regression for prediction.

- In test data we weren't able to maintain the high precision for 1 but recall, which is most important parameter in this cross sell prediction is above 85% for both Random Forest and XGBoost along with the 70% recall for 0s.

- Random forest and XGBoost after hyper parameter tuning have highest f1 score of 44% on test data and their test recall is also very high.

References

I. Stack overflow
II. GeeksforGeeks
III. Almabetter
IV. Analytics Vidhya
V. Towards data science
VI. github

THANK YOU