Capstone Project 3

Health Insurance Cross Sell Prediction

Team Members

Mohammed Saif Khan

Kaustubh Kulkarni

Content

- Problem Statement
- Data Description
- Approach
- Exploratory Data Analysis
- Feature engineering
- Model Implementation
- 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 optimise its business model and revenue.

Data Description

We have been given a dataset :

1. **Health Insurance Cross Sell Prediction.csv** – Insurance related data including customer response.

Columns in the Dataset :

- 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
- Policy_Sales_Channel - 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

Approach

1) Data Exploration And Cleaning

2) Exploratory Data Analysis

- Hypotheses Testing
- Univariate Plots
- Bivariate Plots

3) Feature Engineering

4) Modelling

- Logistic Regression Model
- Random Forest Classifier Model
- Hyper Parameter Tuning
- XGBoost Classifier

5) Conclusion

Exploratory Data Analysis

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

- Hypotheses Testing
- Univariate Plots
- Bivariate Plots
  - ❑ Response Percentage Plots

Hypotheses

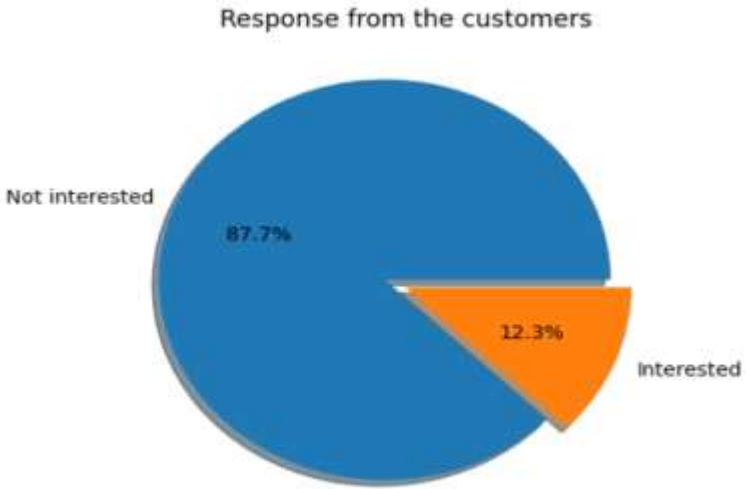
After initial exploration of the data we came up with the following hypotheses.

- As age increases the annual premium should also increase.
- If customer has previous insurance then they are less likely to buy another one.
- If vehicle age is <1 year then they are less likely to respond positively, as while buying the vehicle people often buy 1 year insrance.
- If person has damaged vehicle then they are more likely to buy an insurance.
- Longer the customer is associated with company, they are more likely to respond as yes.

Univariate Plots

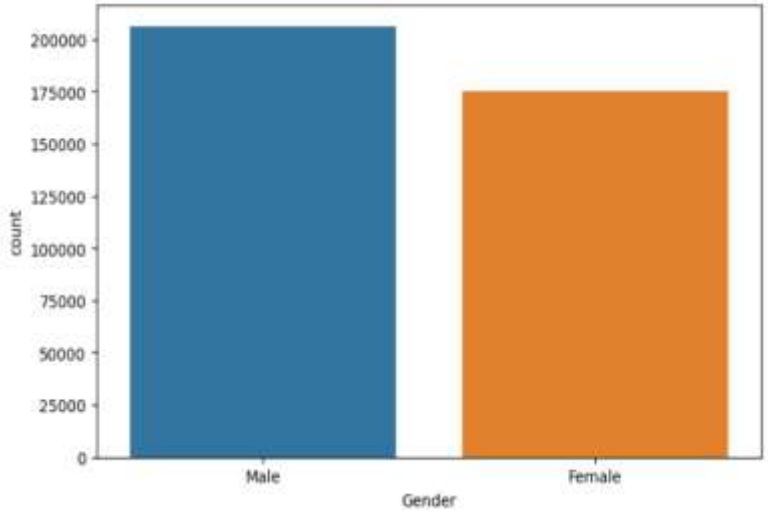
Response from Customers:

This is the response from the customers in the dataset, looks like majority of the people are not interested in buying vehicle insurance from the company and this is the clear case of imbalance dataset.

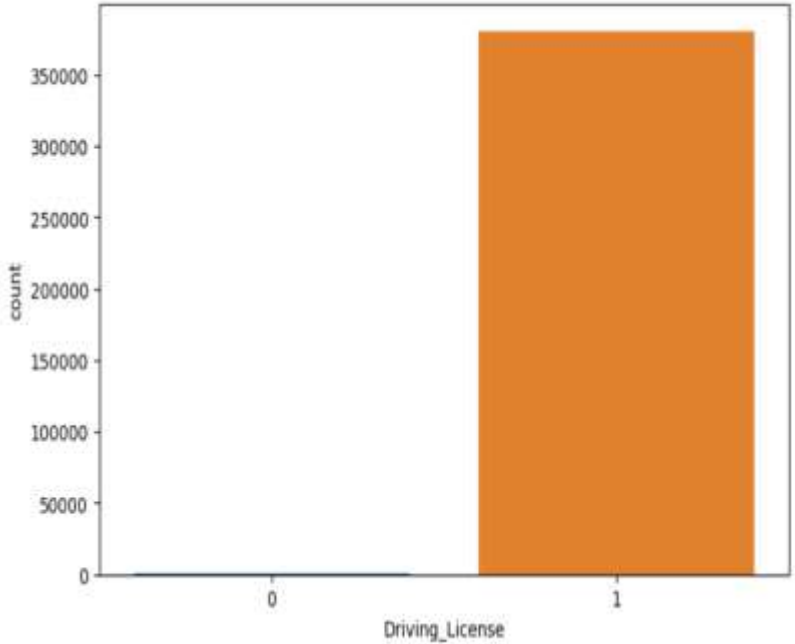


Countplots for Categorical Feature:

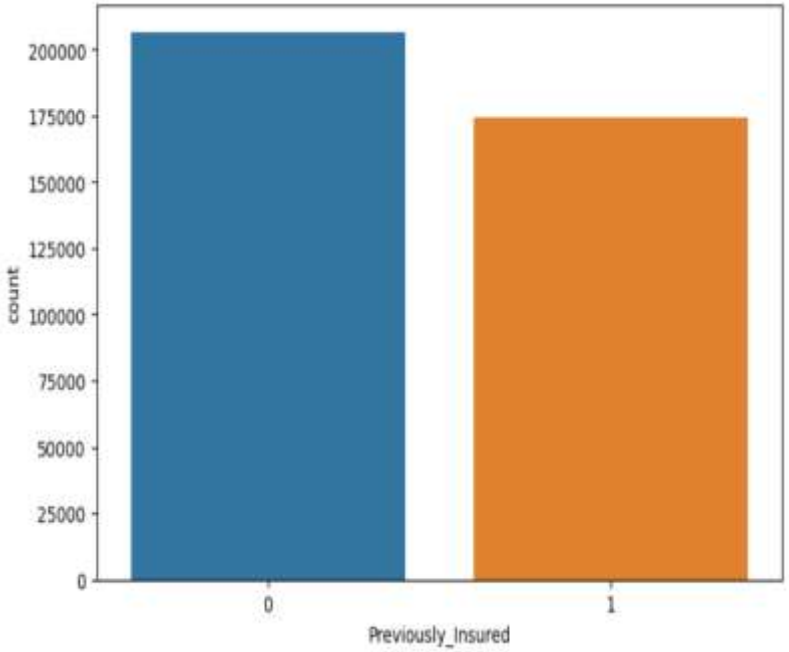
In ‘Gender’ column, there is slight difference between number of male and female customers in the dataset.



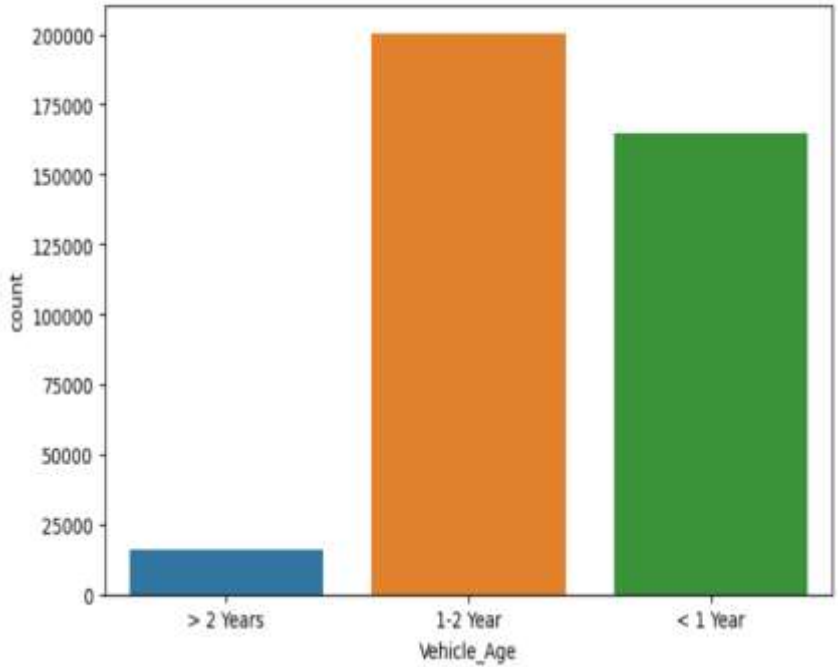
Almost all of the customers have driving license.



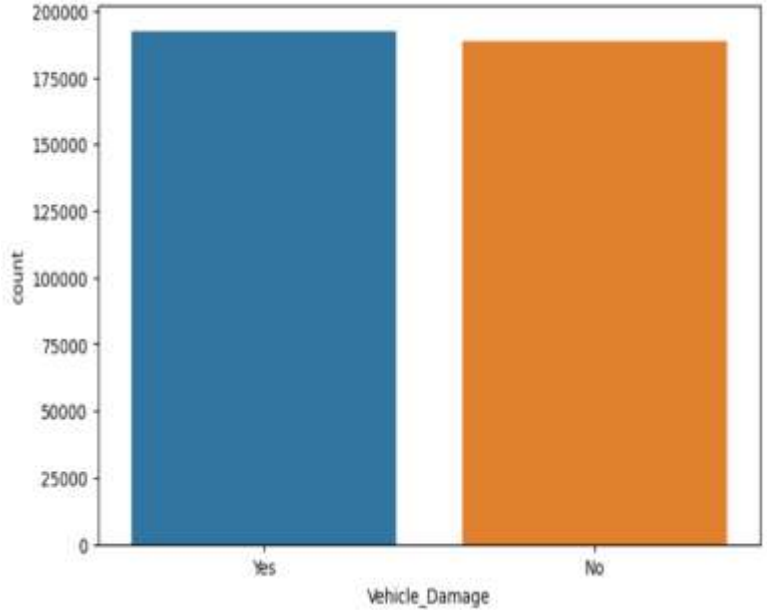
Number of Customers without previous vehicle insurance is slightly higher than with insurance.



There is very less number of customers with Vehicle age greater than 2 years.

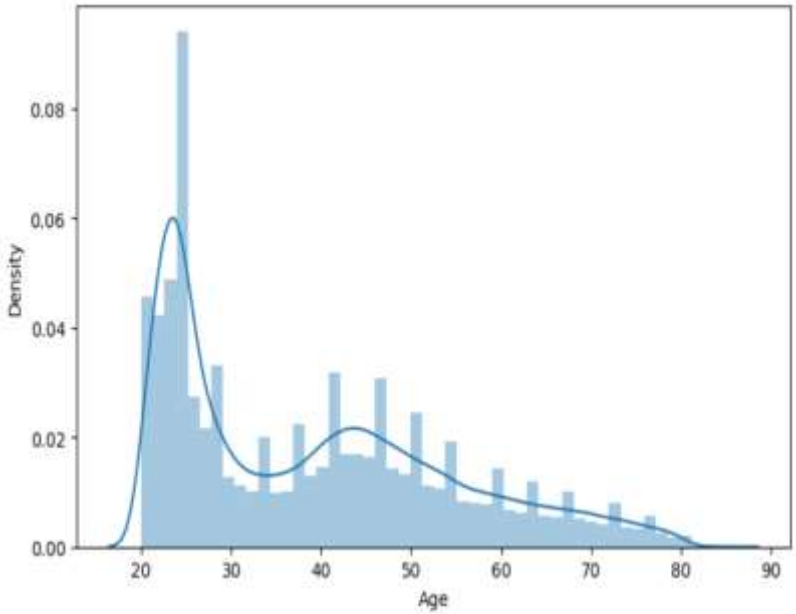


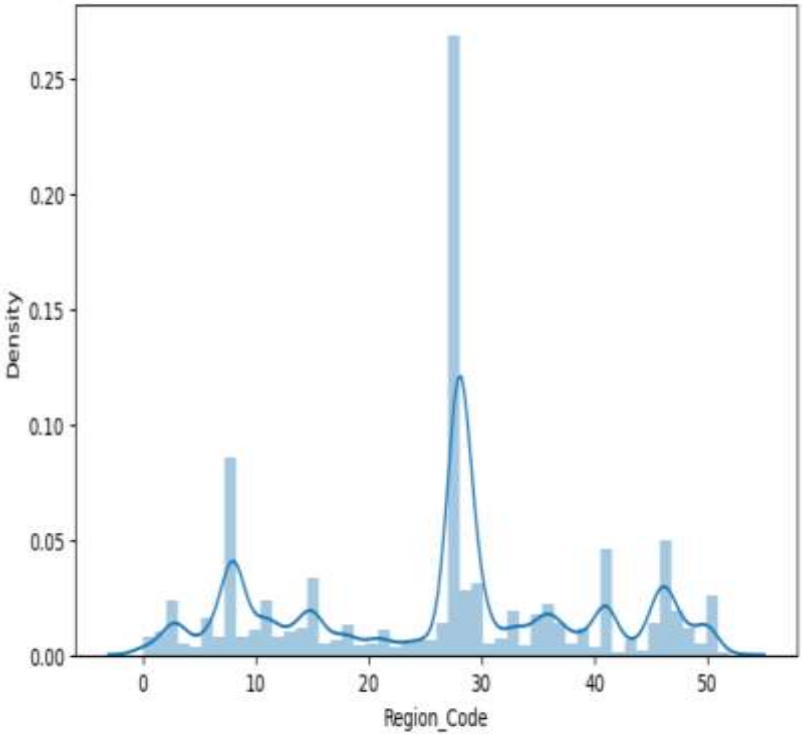
Number of customers with and without damaged vehicle is almost equal.



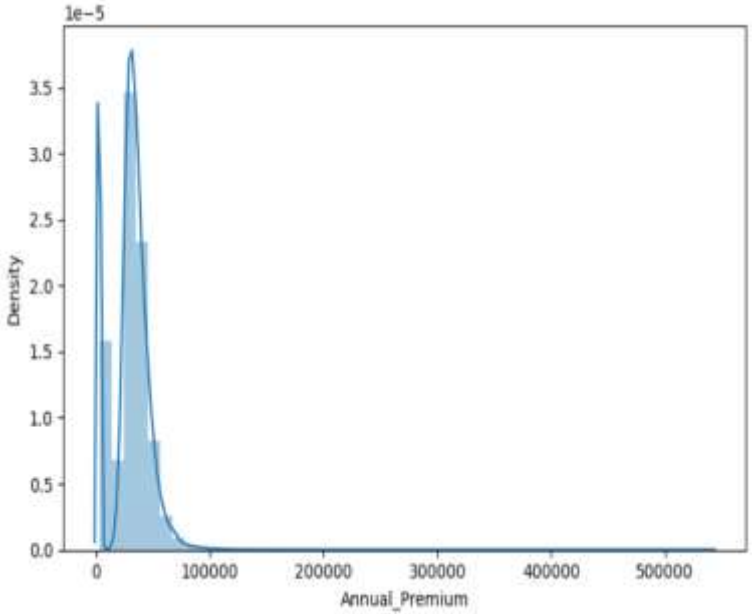
Distplots for Numerical Features:

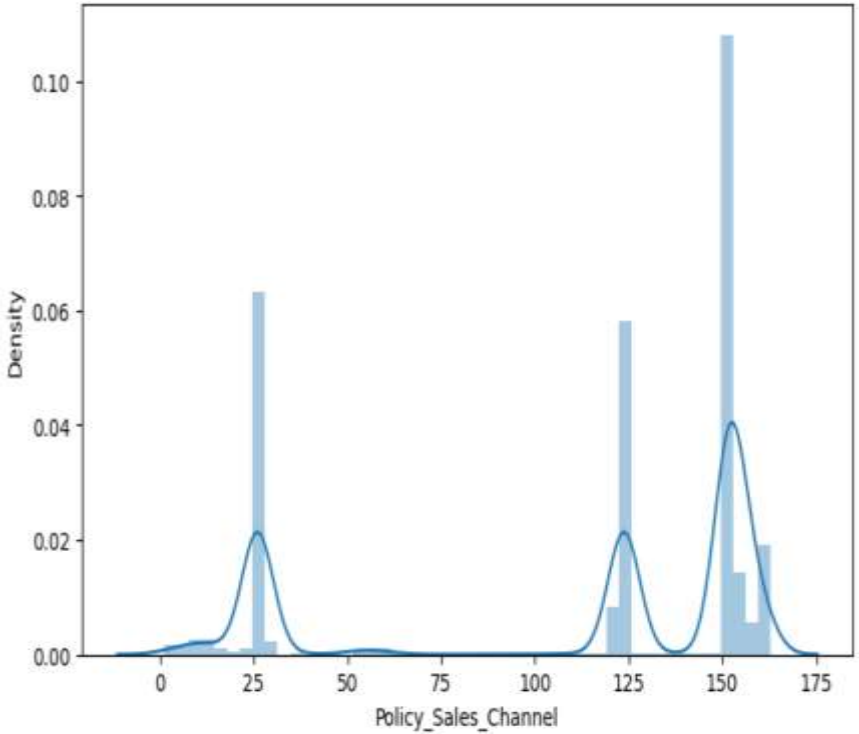
There is large number of customers from age 20 to 30 and then 40 to 50 age group.



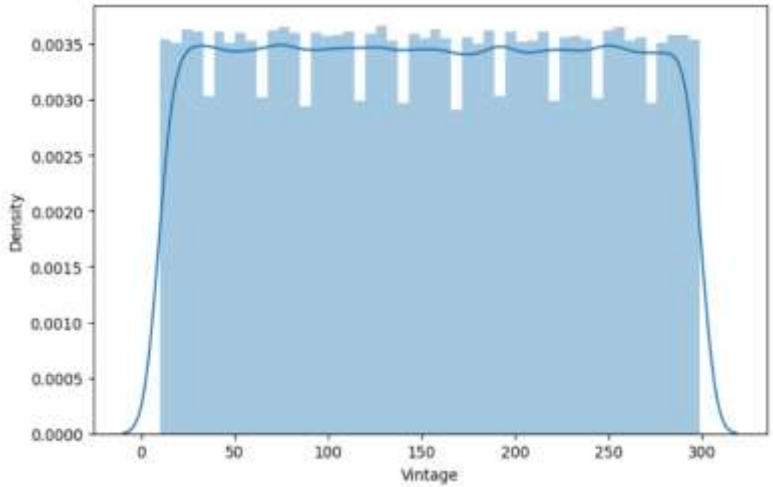


Annual premium is right skewed and has very long tail, which means there are possible outliers in this column.



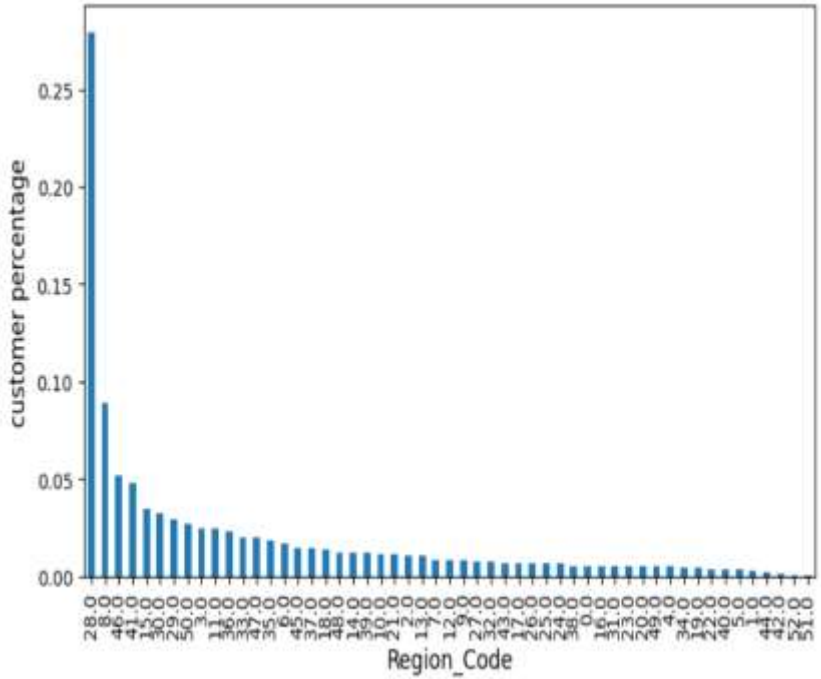


Vintage column has uniform distribution and no inference can be made based on distplot.



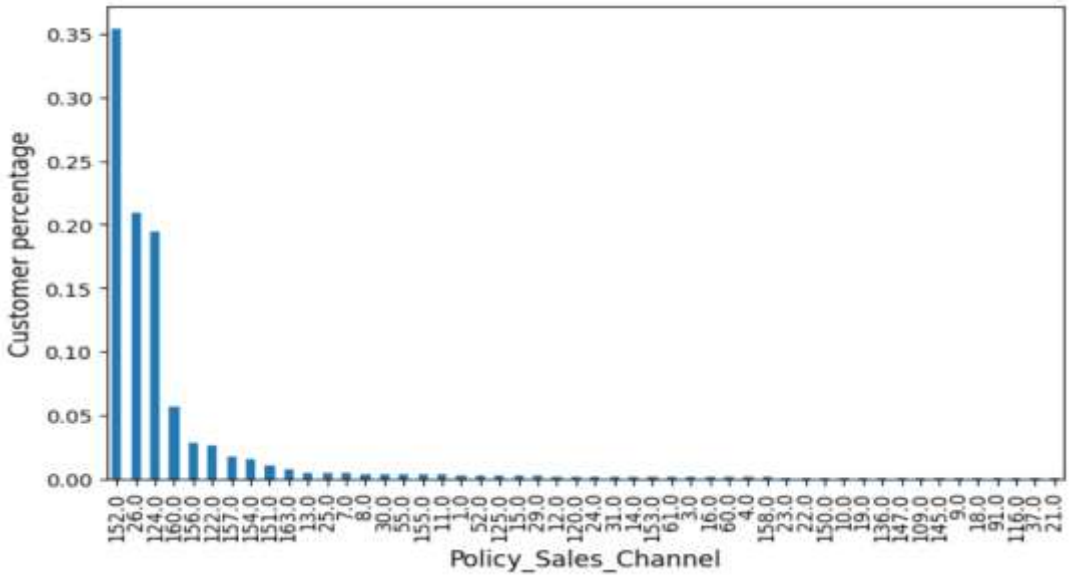
Customer Percentage in each Region:

Most of the customers (30%) are from region code 28 followed by region code 8.



Customer Percentage in each Policy Sales Channel:

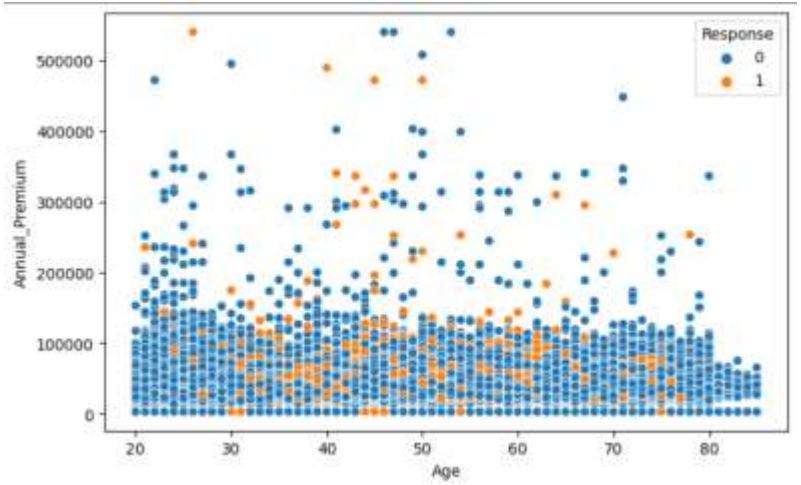
Majority of customers(35%) were connected to the company by sales channel 152, followed by 26 and 124.



Bivariate Plots

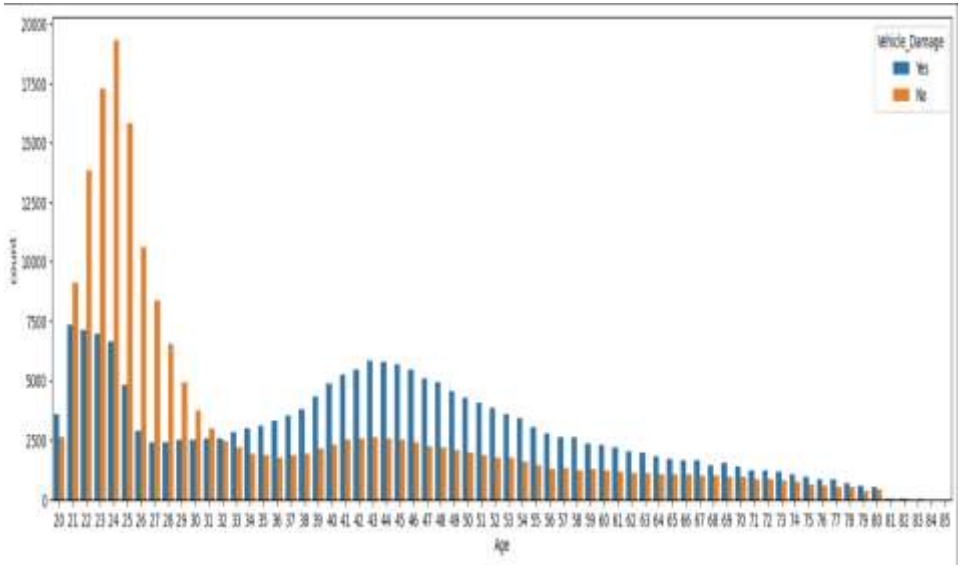
Age wise Annual Premium:

Usually, as age increases annual premium also goes up but that doesn't seem to be the case here. There is no significant relation found between age and annual premium, hence our **1st hypothesis** "As age increases the annual premium should also increase" turns out to be **false** here.



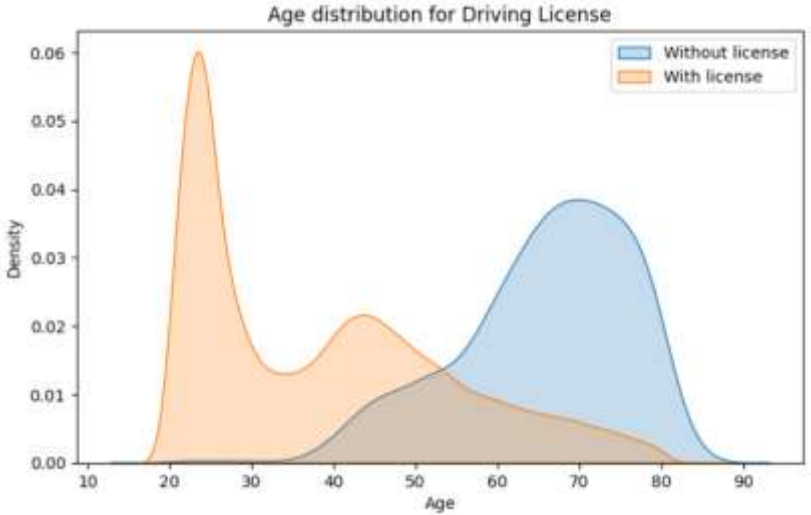
Age wise Vehicle Damage:

The proportion of vehicles that are damaged vs non-damaged is lower in young customers i.e. 20-30 age group but, as age increases the proportion of damaged vehicles is more than non -damaged ones.



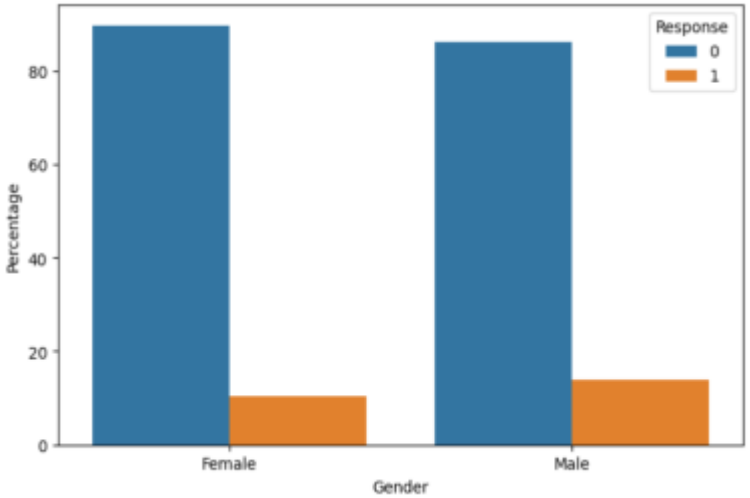
Age Distribution of Customers with or without License:

Though number of customers without driving license is very low i.e. just 812 customers, above density plot shows most of the customers without license are from age group 50 to 85.

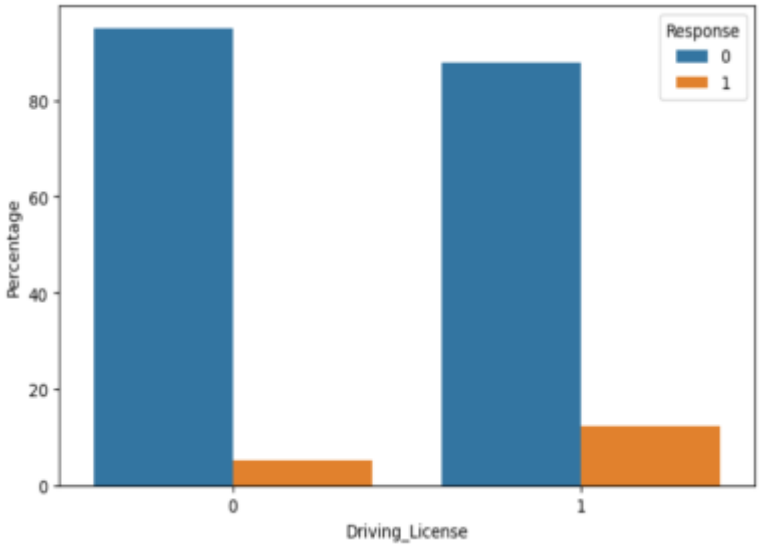


Response Percentage plots for Categorical Features:

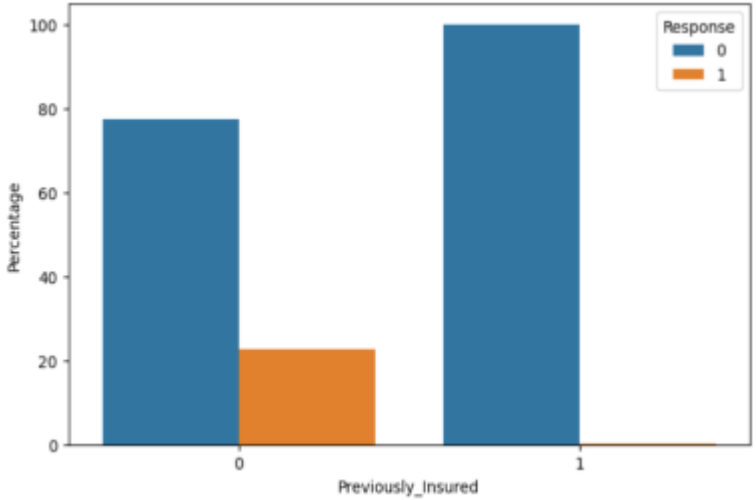
Positive response in male category is almost 14% than that of female category which is 10%. This means males are 30% more likely to respond as yes than females.



As we saw earlier majority of customers have driving license and from the percentage plot above we can say that, customers who have driving license are more likely to respond as yes, comparatively.

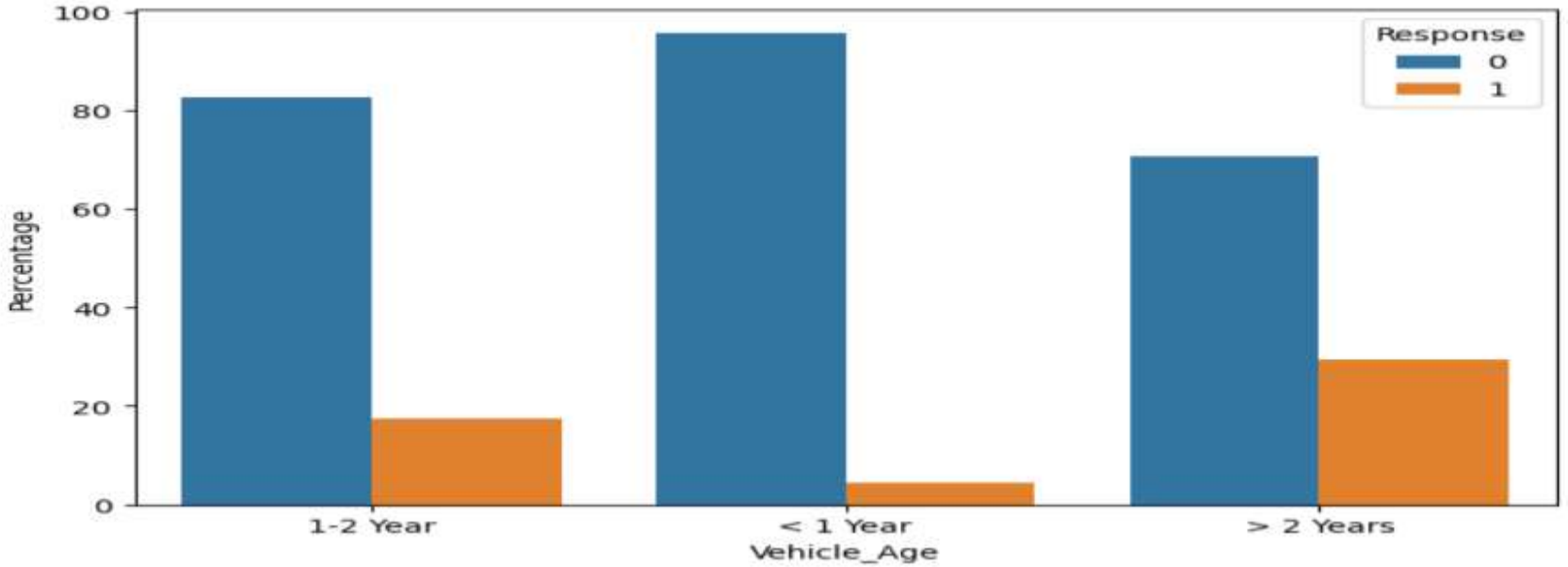


Out of customers who don't have previous insurance, 22% of them responded positively. Whereas customers with previous vehicle insurance are obviously not interested in buying another one. This also makes our **2nd hypothesis true**.

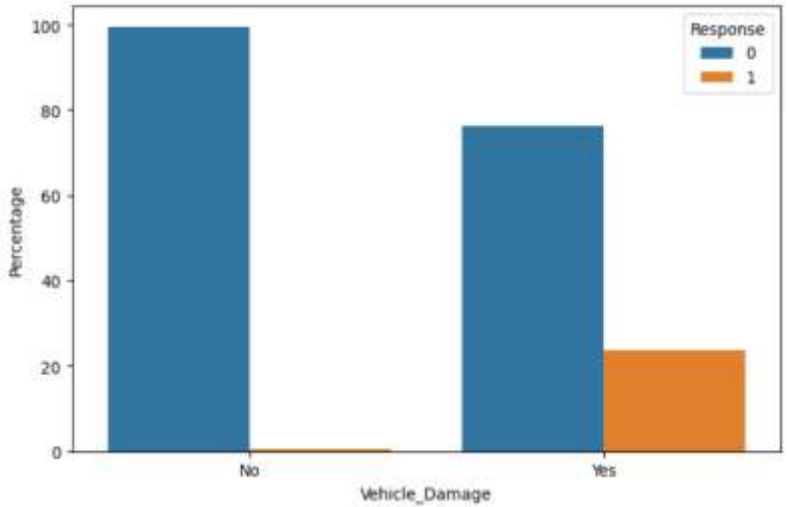


There is very less number of customers with vehicle age >2 years but almost 30% of times they are likely to buy an insurance. For vehicle age 1-2 years 17% respond positively.

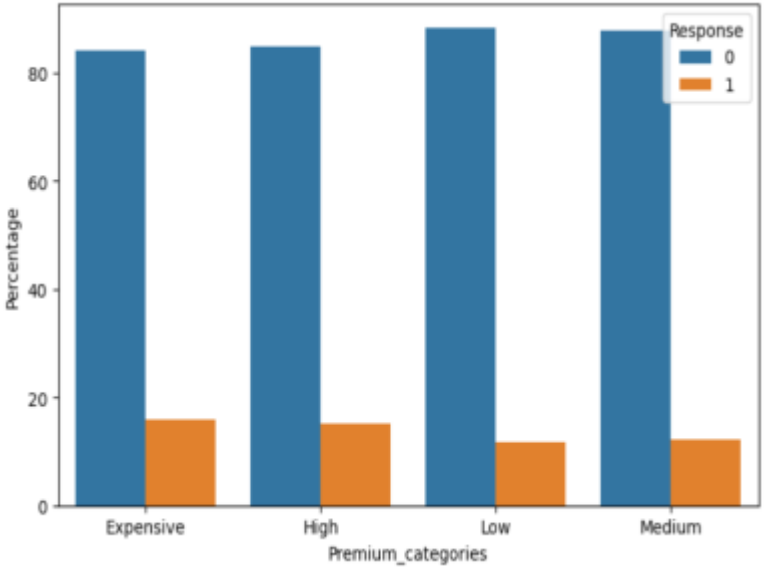
We hypothesized correctly! Customers with vehicle age <1 year are least interested in insurance, 95% of times they respond negatively. Hence our 3rd hypothesis turns out to be true.



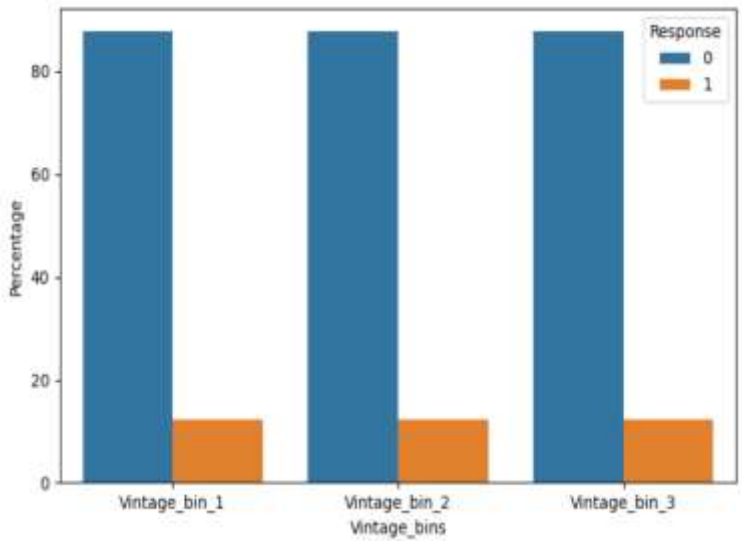
Customers who damaged their vehicles in past, almost 24% of times they respond as Yes for vehicle insurance. And customers without vehicle damage, almost all of the times they respond as No. With that our 4th hypothesis is also true.



Response Percentage plots for Premium Categories column -
Premium categories 'Expensive' and 'High' have slightly high positive response rate, but there is no significant different in response across all of the premium categories.

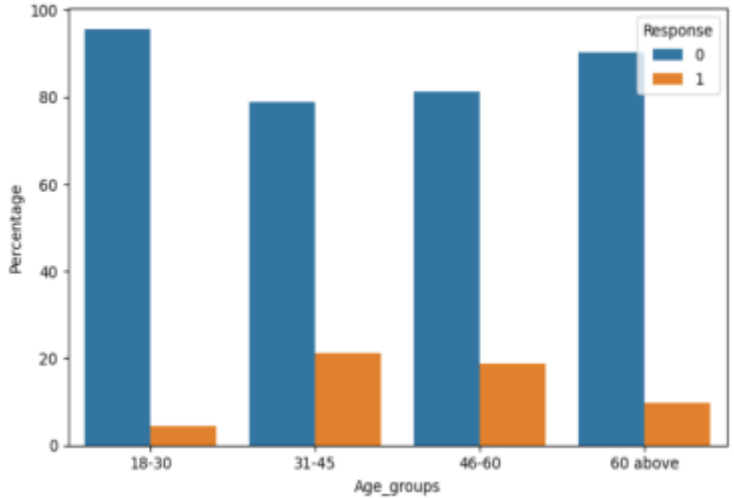


Response Percentage plots for Vintage bins column - We divided vintage column into 3 bins with 100 days interval. As we can see in above plot there is no difference in response rate based on number of days customer associated with company. This makes our **5th hypothesis** "Longer the customer is associated with company, they are more likely to respond as yes" **false**.



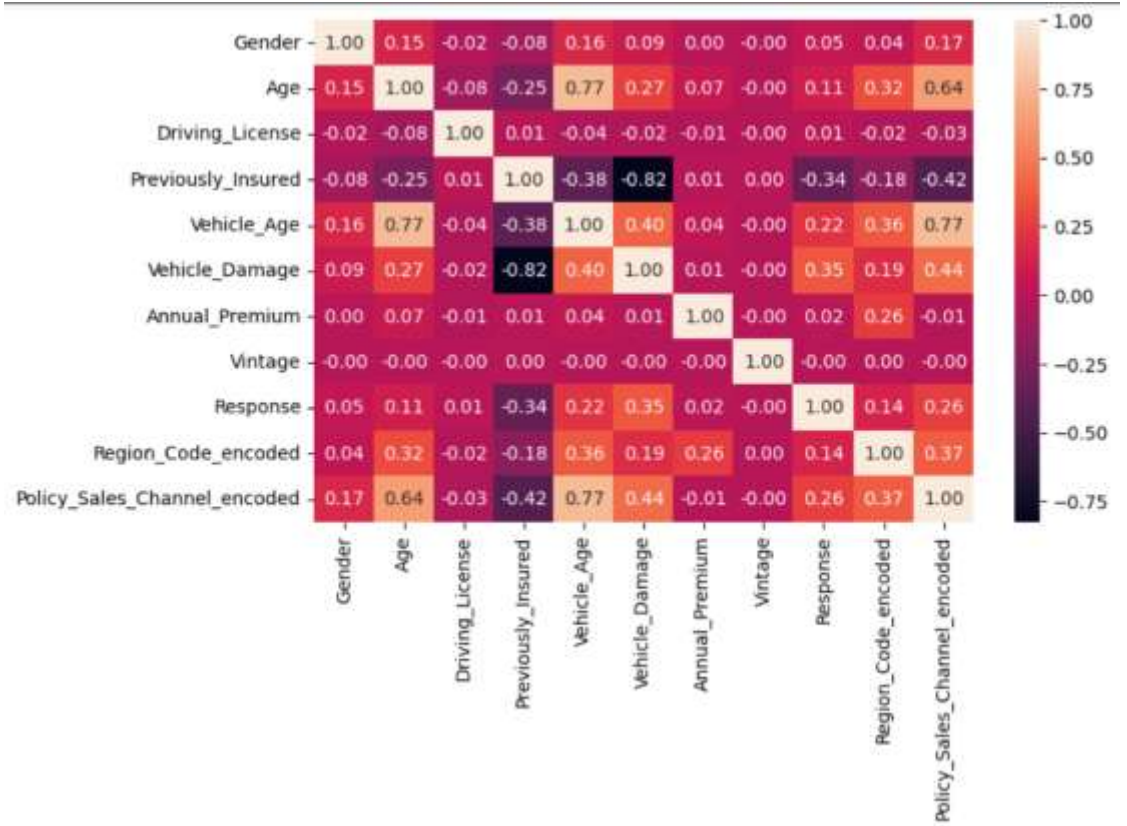
Response Percentage plots for Age_groups column -

We can see that customers in the age group 31 to 45 and 46 to 60 have very high positive response rate compare to other two categories. We also know from earlier graph (*Countplot for age based on vehicle damage*) that customers in these age groups have high vehicle damage rate and also, people with damaged vehicles are more likely to buy a vehicle insurance.



Correlation Heatmap

Target column Response has no direct correlation with any of the features.



Heatmap

Feature Engineering

In feature Engineering we performed the following steps:

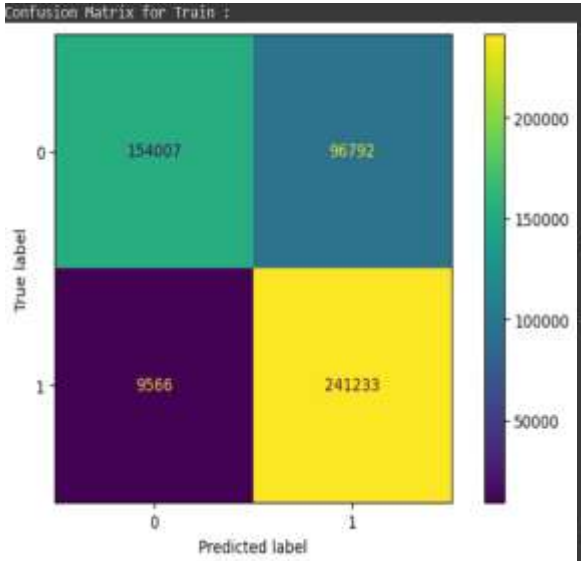
- Label Encoding on Gender, Vehicle_Damage and Vehicle_Age columns.
- Target Mean Encoding on Region_code and Policy_Sales_Channel columns.
- We replaced outliers with the upper limit values for that column.

Model Implementation

The models in use are:

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

Logistic Regression

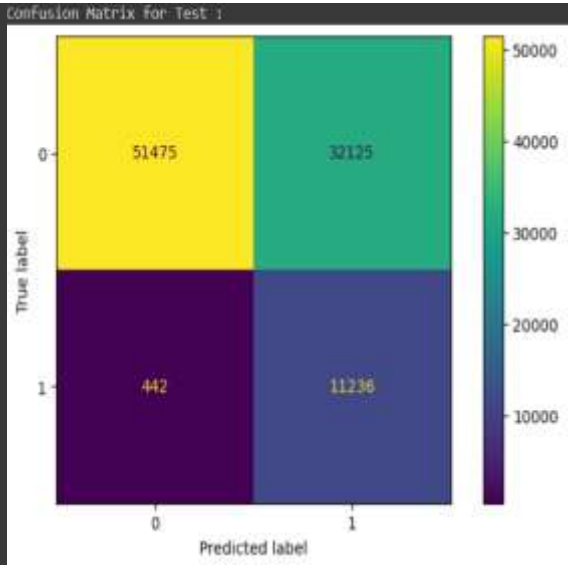


Training Classification Report :

	precision	recall	f1-score	support
0	0.94	0.61	0.74	250799
1	0.71	0.96	0.82	250799
accuracy			0.79	501598
macro avg	0.83	0.79	0.78	501598
weighted avg	0.83	0.79	0.78	501598

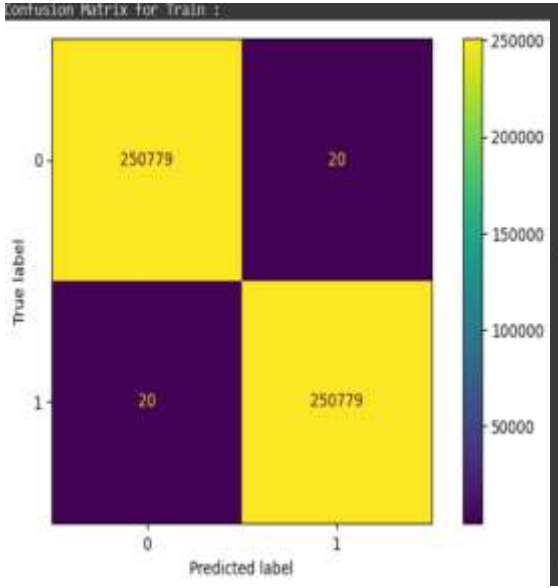
Testing Classification Report :

	precision	recall	f1-score	support
0	0.99	0.62	0.76	83600
1	0.26	0.96	0.41	11678
accuracy			0.66	95278
macro avg	0.63	0.79	0.58	95278
weighted avg	0.90	0.66	0.72	95278



	Model Name	Train accuracy	Test accuracy	Train precision	Test precision	Train recall	Test recall	Train f1 score	Test f1 score
0	LogisticRegression HP_tuning	0.787962	0.65819	0.713654	0.259127	0.961858	0.962151	0.819372	0.408292

Random Forest Classifier

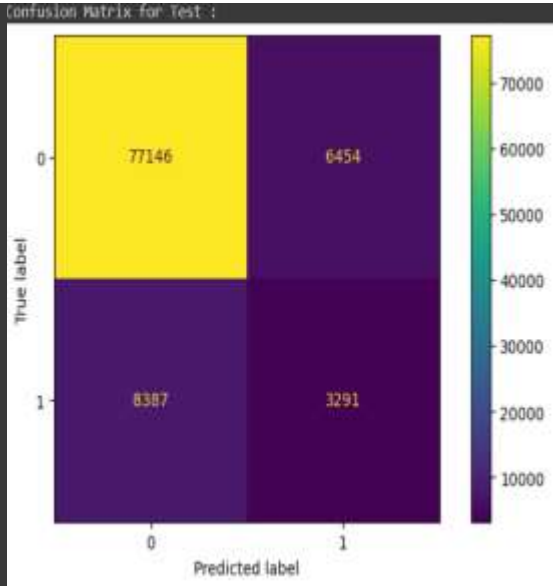


Training Classification Report :

	precision	recall	f1-score	support
0	1.00	1.00	1.00	250799
1	1.00	1.00	1.00	250799
accuracy			1.00	501598
macro avg	1.00	1.00	1.00	501598
weighted avg	1.00	1.00	1.00	501598

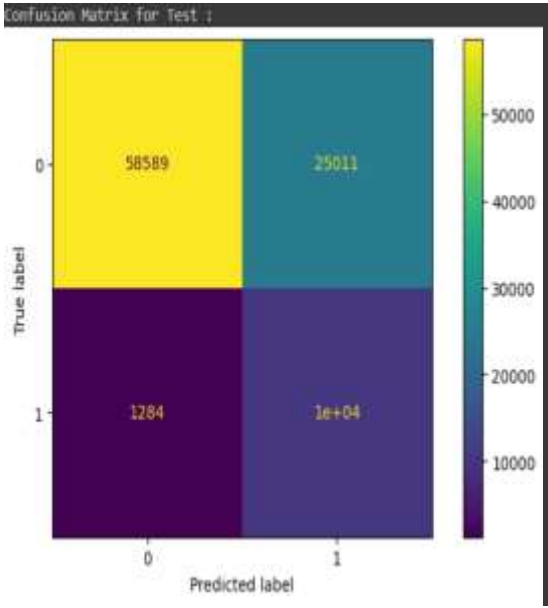
Testing Classification Report :

	precision	recall	f1-score	support
0	0.90	0.92	0.91	83600
1	0.34	0.28	0.31	11678
accuracy			0.84	95278
macro avg	0.62	0.60	0.61	95278
weighted avg	0.83	0.84	0.84	95278



	Model Name	Train accuracy	Test accuracy	Train precision	Test precision	Train recall	Test recall	Train f1 score	Test f1 score
0	RandomForestClassifier	0.99992	0.844235	0.99992	0.337712	0.99992	0.281812	0.99992	0.30724

Random Forest after tuning

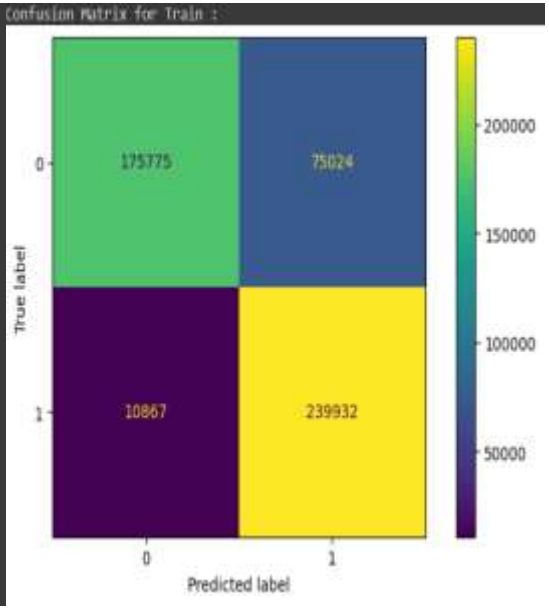


Training Classification Report :

	precision	recall	f1-score	support
0	0.94	0.70	0.80	250799
1	0.76	0.96	0.85	250799
accuracy			0.83	501598
macro avg	0.85	0.83	0.83	501598
weighted avg	0.85	0.83	0.83	501598

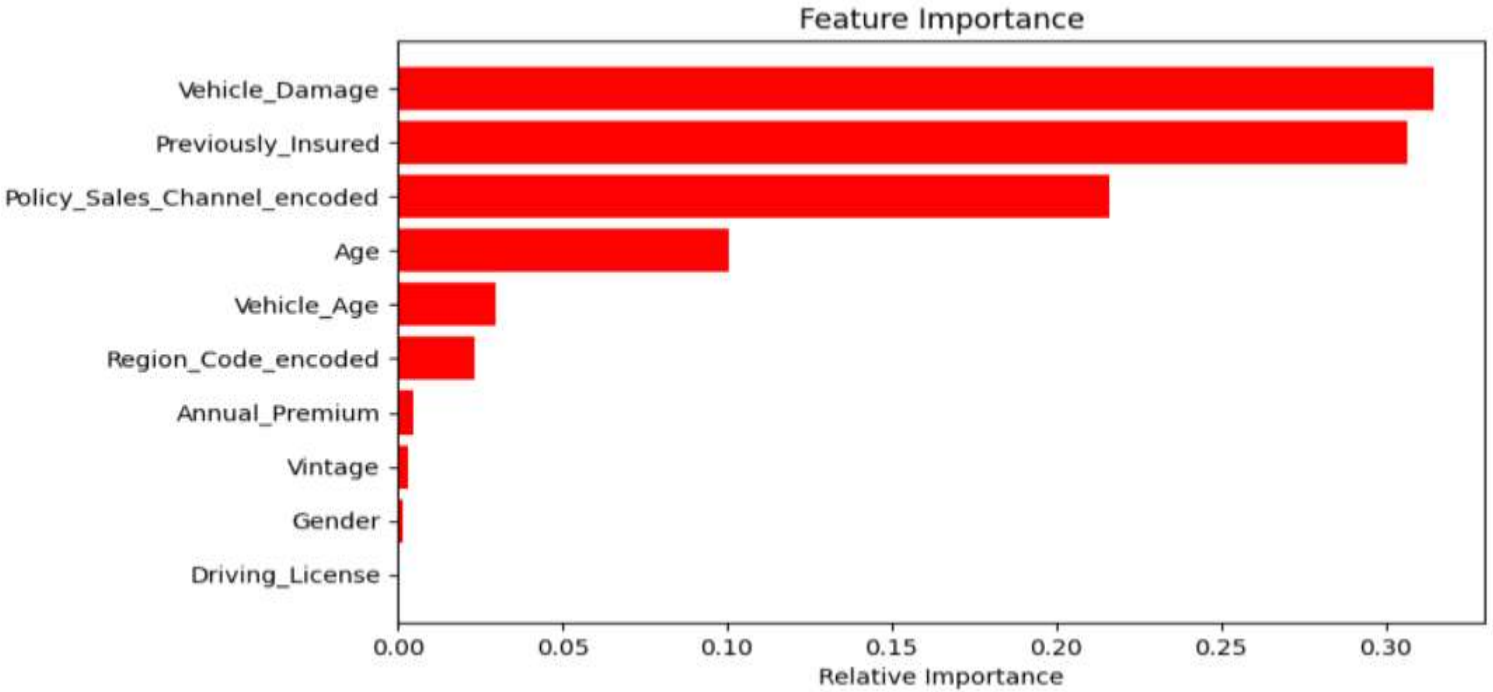
Testing Classification Report :

	precision	recall	f1-score	support
0	0.98	0.70	0.82	83600
1	0.29	0.89	0.44	11678
accuracy			0.72	95278
macro avg	0.64	0.80	0.63	95278
weighted avg	0.89	0.72	0.77	95278



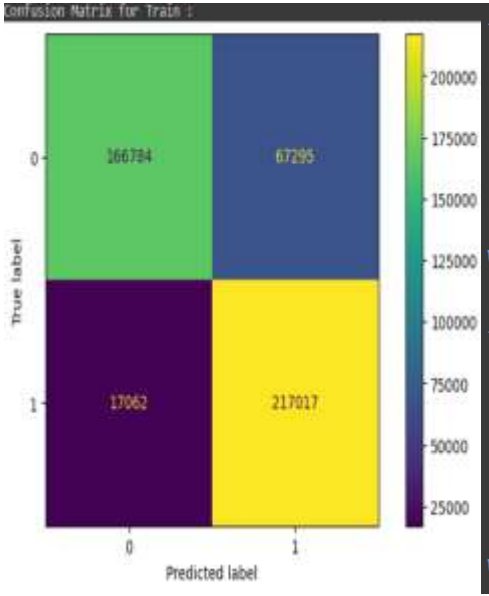
	Model Name	Train accuracy	Test accuracy	Train precision	Test precision	Train recall	Test recall	Train f1 score	Test f1 score
0	RandomForestClassifier HP_tuning	0.828765	0.724018	0.761795	0.293574	0.95667	0.89005	0.848183	0.441518

Feature Importance



Features like Vehicle_Damage, Previously_Insured, Policy_Sales_Channel_encoded and Age have more importance.

XGBoost Classifier

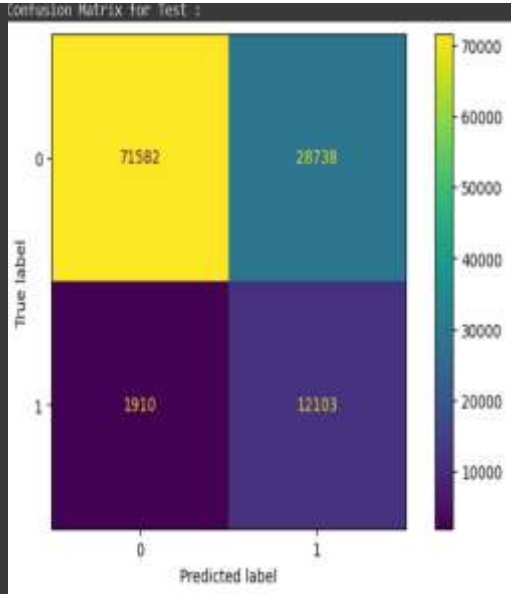


Training Classification Report :

	precision	recall	f1-score	support
0	0.91	0.71	0.80	234079
1	0.76	0.93	0.84	234079
accuracy			0.82	468158
macro avg	0.84	0.82	0.82	468158
weighted avg	0.84	0.82	0.82	468158

Testing Classification Report :

	precision	recall	f1-score	support
0	0.97	0.71	0.82	100320
1	0.30	0.86	0.44	14013
accuracy			0.73	114333
macro avg	0.64	0.79	0.63	114333
weighted avg	0.89	0.73	0.78	114333



	Model Name	Train accuracy	Test accuracy	Train precision	Test precision	Train recall	Test recall	Train f1 score	Test f1 score
0	XGBClassifier HP_tuning	0.819811	0.731941	0.763306	0.296344	0.92711	0.863698	0.837271	0.44128

Model Metrics Dataset

From the model metrics we can say that the Hyper parameter Tuned Random Forest Classifier is the best Model from all the other models.

	Model Name	Train accuracy	Test accuracy	Train precision	Test precision	Train recall	Test recall	Train f1 score	Test f1 score
0	LogisticRegression HP_tuning	0.787962	0.658190	0.713654	0.259127	0.961858	0.962151	0.819372	0.408292
1	RandomForestClassifier	0.999920	0.844235	0.999920	0.337712	0.999920	0.281812	0.999920	0.307240
2	RandomForestClassifier HP_tuning	0.828765	0.724018	0.761795	0.293574	0.956670	0.890050	0.848183	0.441518
3	XGBClassifier HP_tuning	0.819811	0.731941	0.763306	0.296344	0.927110	0.863698	0.837271	0.441280

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. Infact 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 postive 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. featues 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.

Conclusion (Contd.)

- 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 hyperparameter tuning have highest f1 score of 44% on test data and their test recall is also very high.

THANK YOU!