Capstone Project
Supervised ML - Classification
Health Insurance Cross Sell Prediction

Completed By :

Ankush Kumar

Points to Discuss:

- Introduction and Problem Statement
- Data Exploration
- Analysis Methodology
- Data Processing and Exploratory Data Analysis(EDA)
- Modeling
- Hyper Parameter Tuning
- Final Modeling and Results
- Conclusions and Summary

Introduction and Problem Statement

Introduction:

An insurance policy is a contract whereby a business agrees to guarantee compensation in the event of a specific loss, damage, illness, or death in exchange for the payment of a predetermined premium.

The amount of money that the customer must consistently pay to an insurance company in exchange for this guarantee is known as a premium. Vehicle insurance works similarly to medical insurance in that customers must pay an annual premium to the insurance provider company in order for them to be compensated (referred to as "sum assured") in the event that their vehicle is responsible for an unfortunate accident

Problem Statement:

Determining whether a person with an insurance policy would be interested in purchasing car insurance as well. The company can then plan its communication strategy to reach out to those customers and optimise its business model and revenue by building a model to predict whether a customer would be interested in Vehicle Insurance.

Data Exploration:

Dataset fil format: CSV(Comma Separated) file is used

Name of the data-source file : TRAIN-HEALTH INSURANCE CROSS SELL PREDICTION.csv

Number of columns : 12

Number of rows : 381109

Number of Numerical columns are: 9

Number of Categorical columns are: 3

Missing Values in the dataset

```
print(data.isnull().sum())
```

```
id                    0
Gender                0
Age                   0
Driving_License       0
Region_Code           0
Previously_Insured    0
Vehicle_Age           0
Vehicle_Damage        0
Annual_Premium        0
Policy_Sales_Channel  0
Vintage               0
Response              0
dtype: int64
```

No missing value in any feature of our data.

Dataset Summary

Feature Columns / Variables:

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

Feature Columns / Variables:

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

Target Column/Variable:

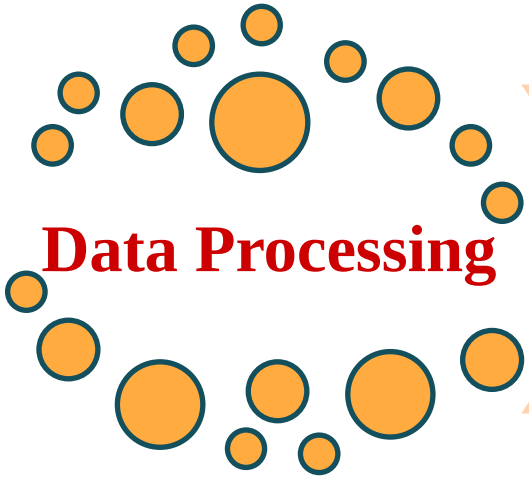
Response : 1 : Customer is interested, 0 : Customer is not interested.

Unique Values and Five Point Statistical Summary

	Column_name	Number_of_Unique_Values
0	Gender	2
1	Driving_License	2
2	Previously_Insured	2
3	Vehicle_Damage	2
4	Response	2
5	Vehicle_Age	3
6	Region_Code	53
7	Age	66
8	Policy_Sales_Channel	155
9	Vintage	290
10	Annual_Premium	48838
11	id	381109

	count	mean	std	min	25%	50%	75%	max
id	381109.0	190555.000000	110016.836208	1.0	95278.0	190555.0	285832.0	381109.0
Age	381109.0	38.822584	15.511611	20.0	25.0	36.0	49.0	85.0
Driving_License	381109.0	0.997869	0.046110	0.0	1.0	1.0	1.0	1.0
Region_Code	381109.0	26.388807	13.229888	0.0	15.0	28.0	35.0	52.0
Previously_Insured	381109.0	0.458210	0.498251	0.0	0.0	0.0	1.0	1.0
Annual_Premium	381109.0	30564.389581	17213.155057	2630.0	24405.0	31669.0	39400.0	540165.0
Policy_Sales_Channel	381109.0	112.034295	54.203995	1.0	29.0	133.0	152.0	163.0
Vintage	381109.0	154.347397	83.671304	10.0	82.0	154.0	227.0	299.0
Response	381109.0	0.122563	0.327936	0.0	0.0	0.0	0.0	1.0

Analysis Methodology



- Exploratory Data Analysis (EDA)
- Outlier Analysis
- Correlation Analysis

Feature Transformation

- Label Encoding
- Standardizing Features

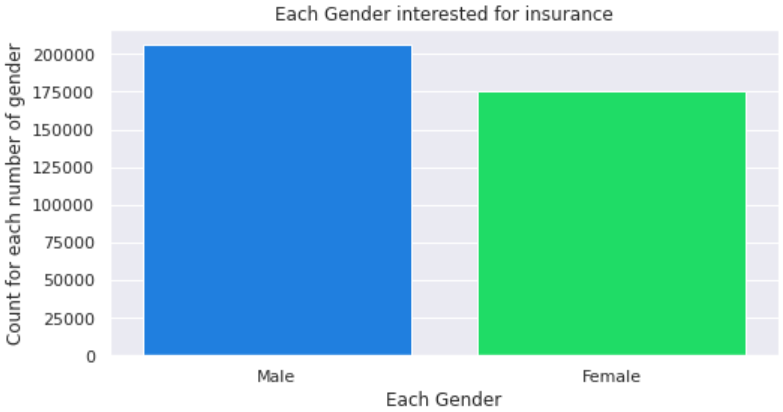
Modeling

- Hyperparameter Tuning
- Final Modeling

Exploratory Data Analysis (EDA)

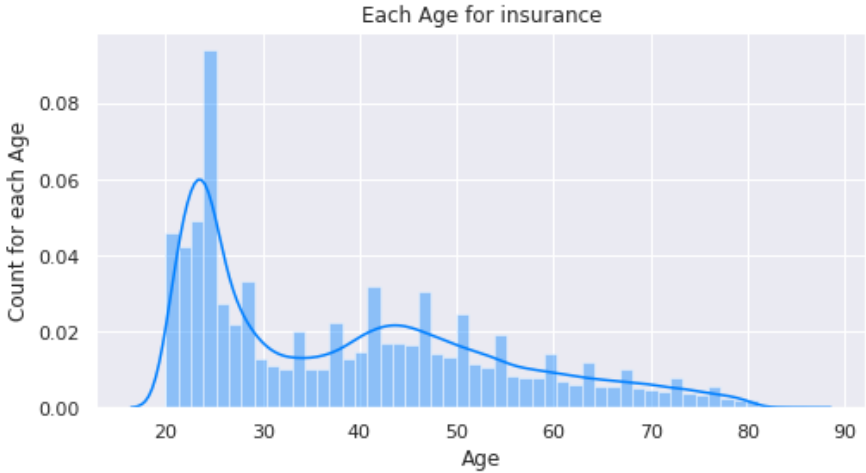
Univariate Analysis

Checked Gender Column



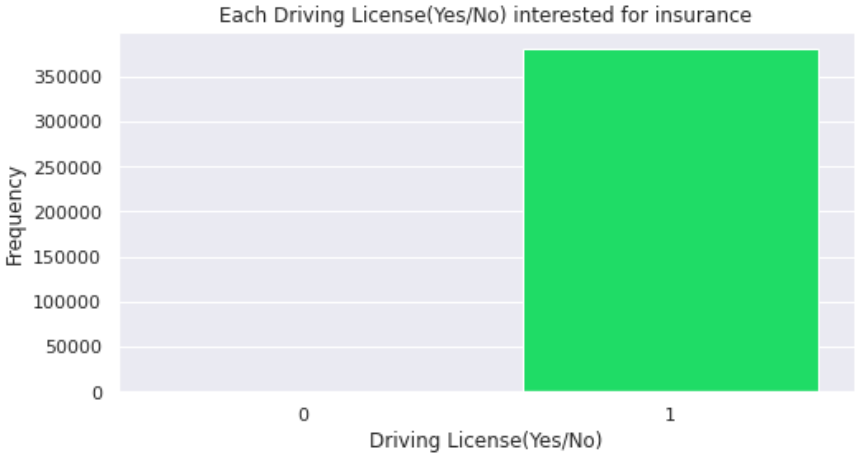
- Males are more interested in buying a insurance rather than females

Checked Gender Column



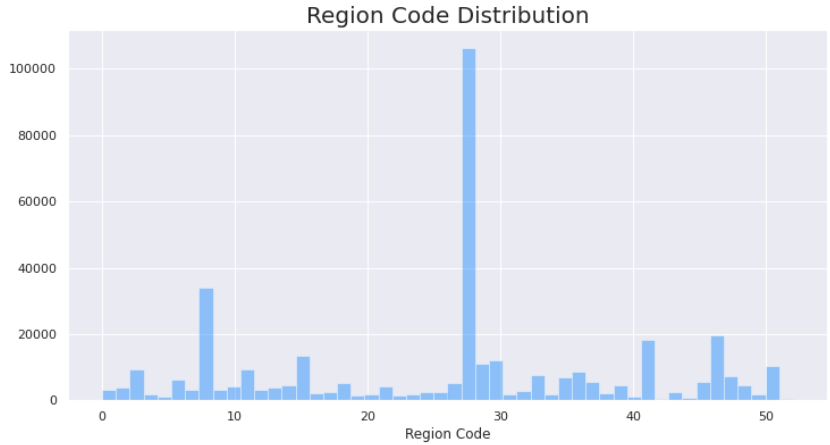
- Age column is not following proper uniform distribution as it is skewed to the right and also it indicates that the people of older age are more wiling to buy an insurance

Analysing Driving License Column



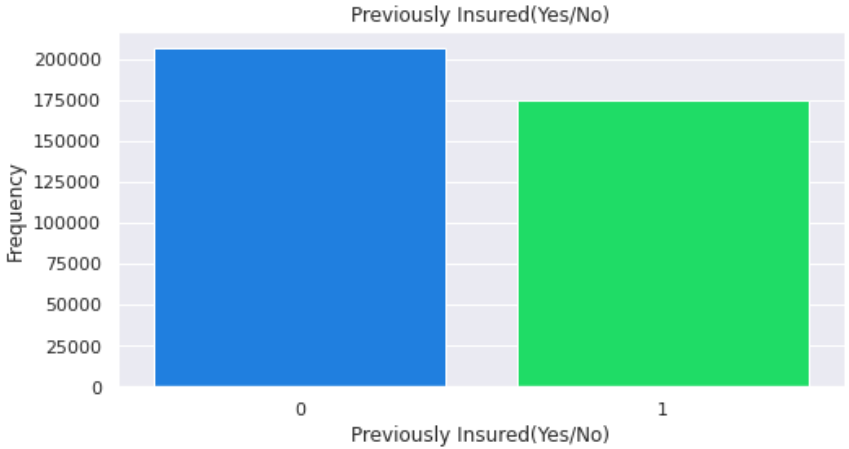
If there is a license holder than it will be more likely to buy an insurance.
Also there is very less count of People who do not have driving license.

Analyzing Region Code Column



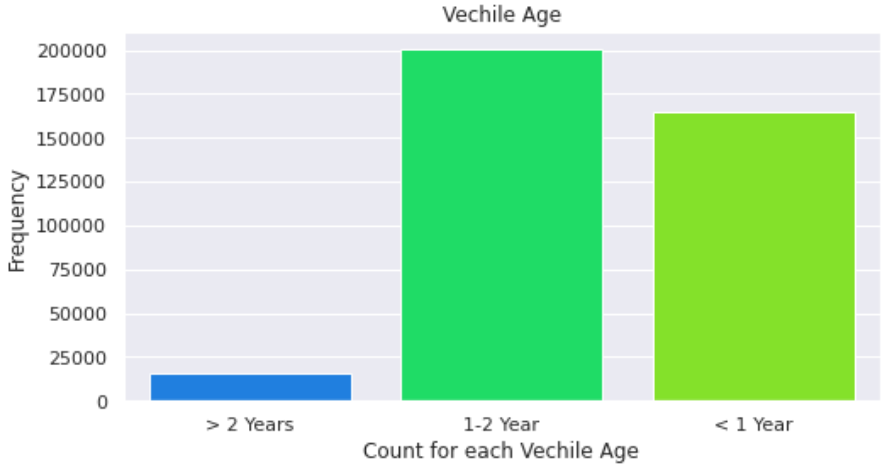
Area code 28 has the most health insurance clients and region code 51 has the fewest.
As this Region_Code column seems like an categorical column as it has different types of region codes so we will treat it into data_preprocessing part.

Analyzing Previously Insured



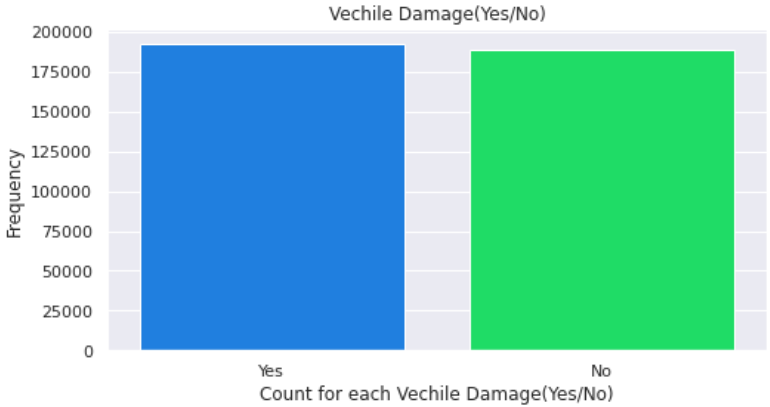
There is almost no effect on insurance buy if a person is already having a insuarance or not
But if we consider more about it then the person who doesn't have any insaurance previously than the person will more like to get an insaurance

Analyzing Vehicle Age



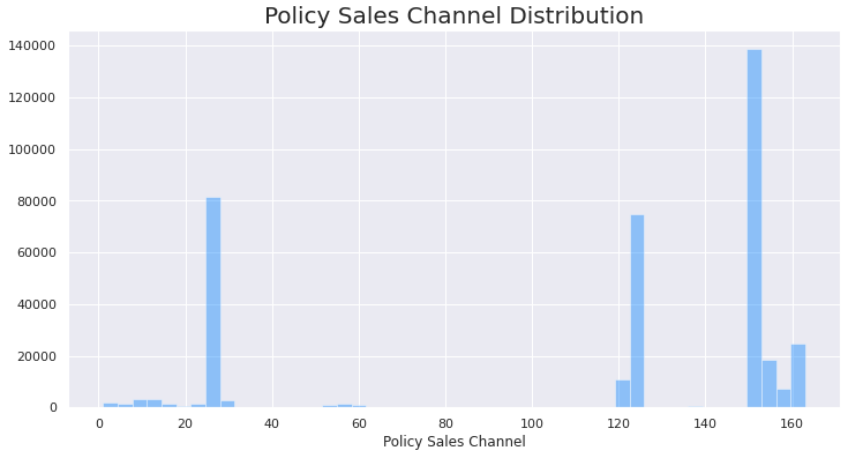
We can see that vechel age between 1-2 years have the most number of counts and vechile age less than 2 years has the least count.

Analyzing Vehicle Column



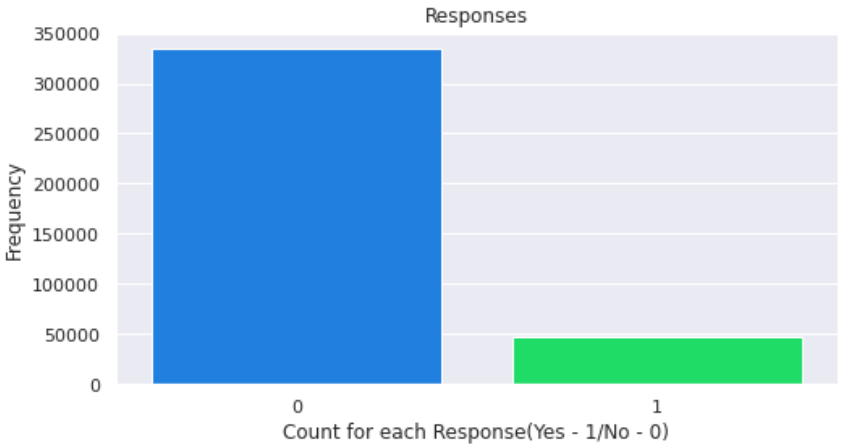
Here this count plot is shoeing that there is almost equal count for both the Vehicle property based on damage.

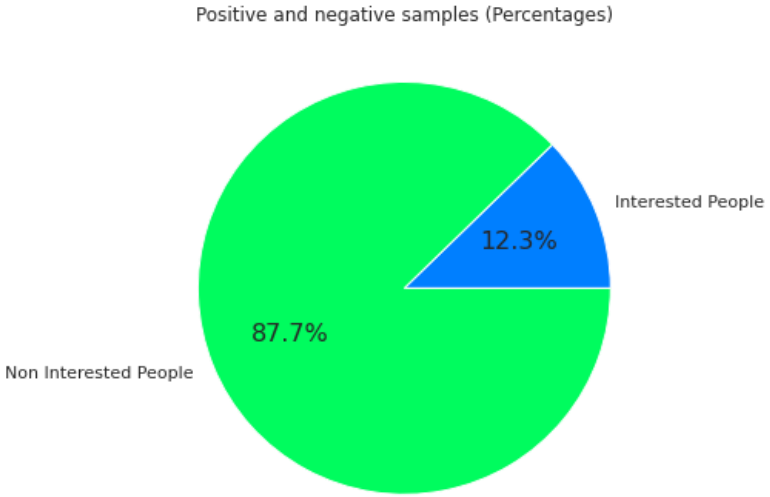
Analyzing Policy Sales Channel Column



Two channel clusters make up the majority of the policy sales channel. At 80, they are primarily divided

Analyzing Target Variable Response Code





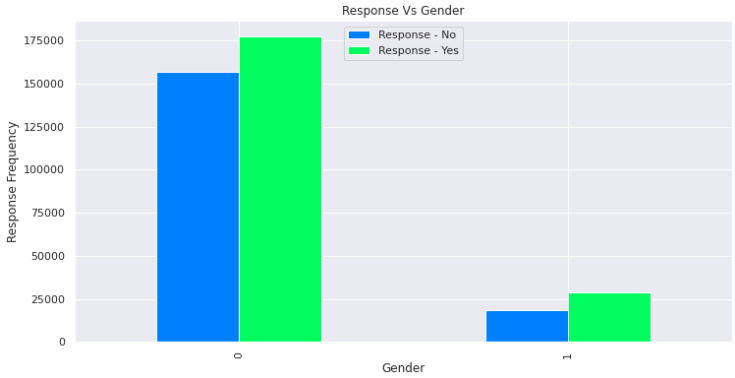
Only 12% are interested in purchasing vehicle insurance.

This is also showing the data imbalance as our response as yes is very less i.e. 12% which will lead our model to a biased model.

We will use Resampling technique to over come this data imbalance.

Bivariate and Multivariate Analysis

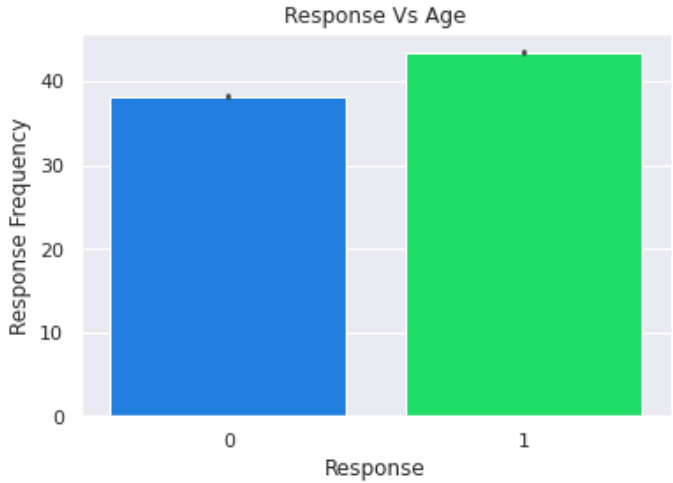
Comparing Gender and Response Column.



Compared to women, men are more likely to be interested in car insurance.

Male respondents make up 61% of the respondents who expressed interest.

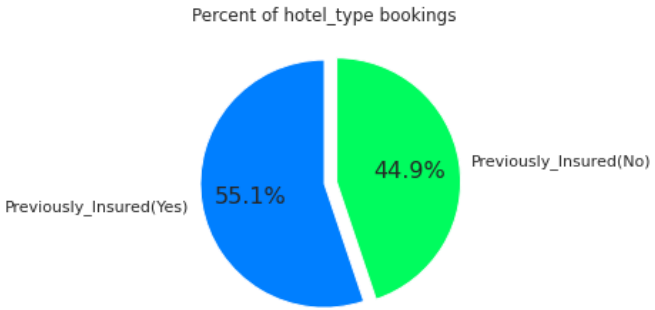
Comparing Age and Response Column.



The average age of customers who are interested in vehicle insurance is 43 years old, while customers who are not interested are 38 years old.

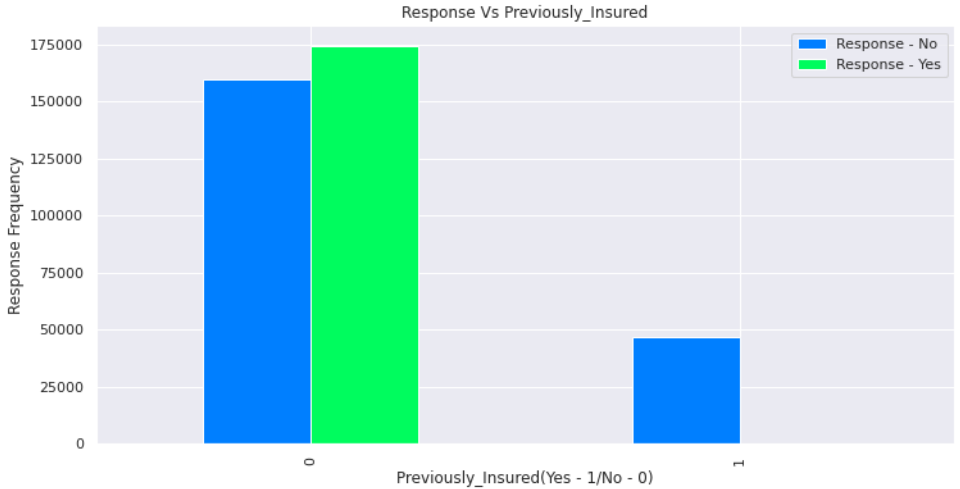
Indicating that younger customers are not interested in vehicle insurance.

Comparing Age and Previously Insured Column.



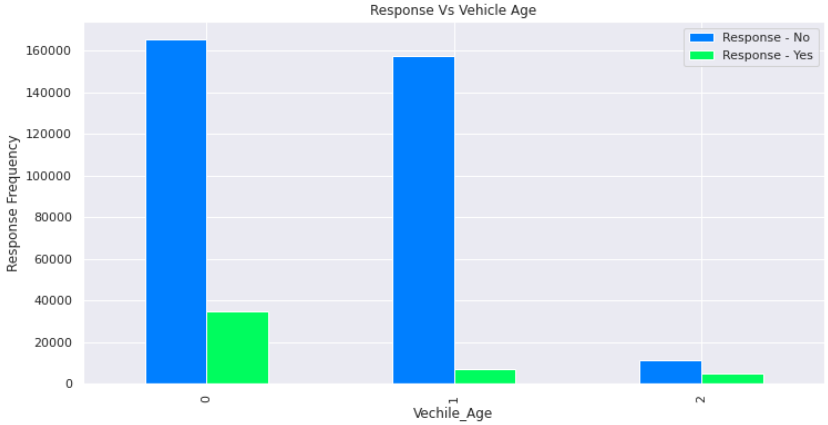
The average age of customers with auto insurance is 34.5 while the average age of customers without auto insurance is 42.4. It is indicating that younger customers are more likely to have auto insurance than older ones.

Comparing Response and Previously Insured Column.



Almost all customers who currently have a car insurance are not interested in getting another one, compared to all customers who do not.
And roughly 25% of them are interested in buying car insurance.

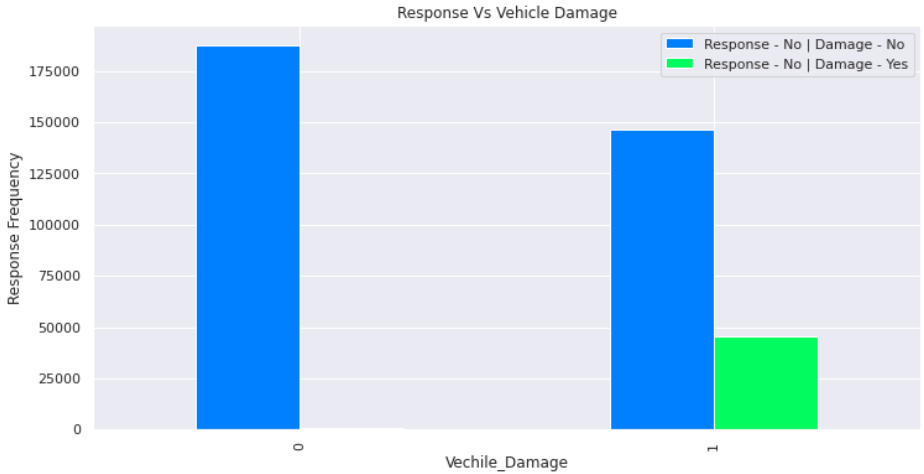
Comparing Response and Vehicle Age Column.



Age range of vehicles Customers with newer cars are less likely to be interested in purchasing auto insurance.

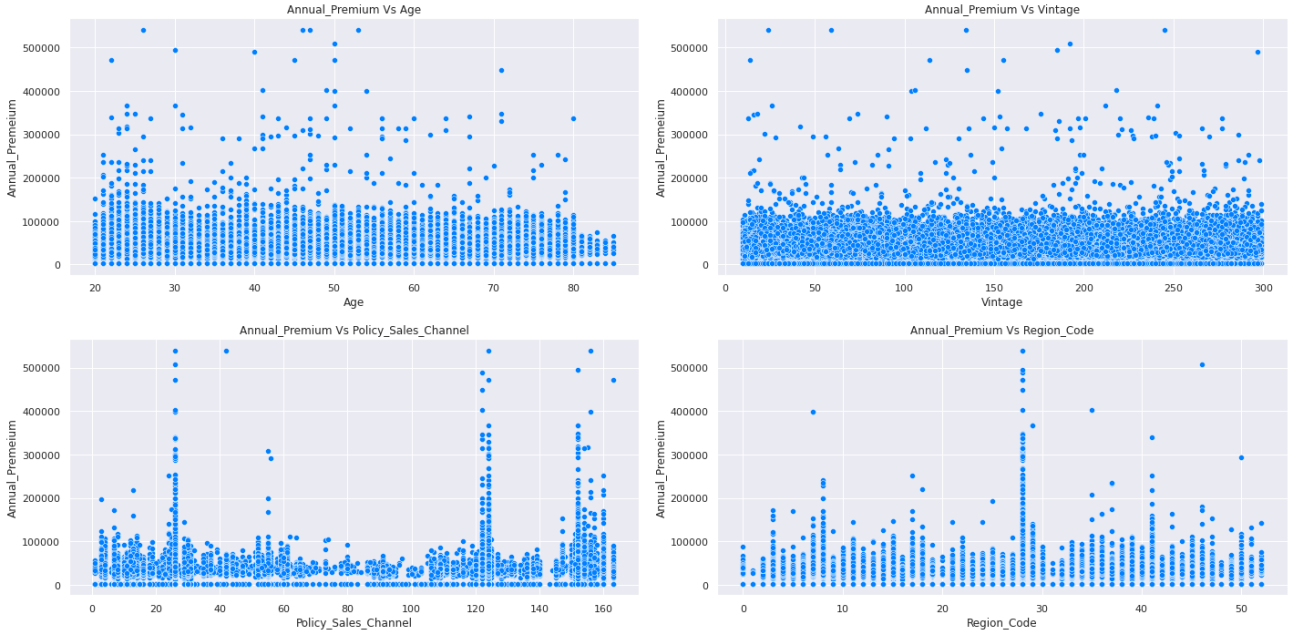
Customers with vehicles older than two years are more likely to be interested in purchasing auto insurance.

Comparing Response and Vehicle Damage Column.



Customers who own a newer car are more likely to get their vehicle insured, this could indicate that insurance companies should collaborate with dealerships to offer a vehicle and insurance bundle.

Comparing Annual_Premium with different columns and extract the relationship between them



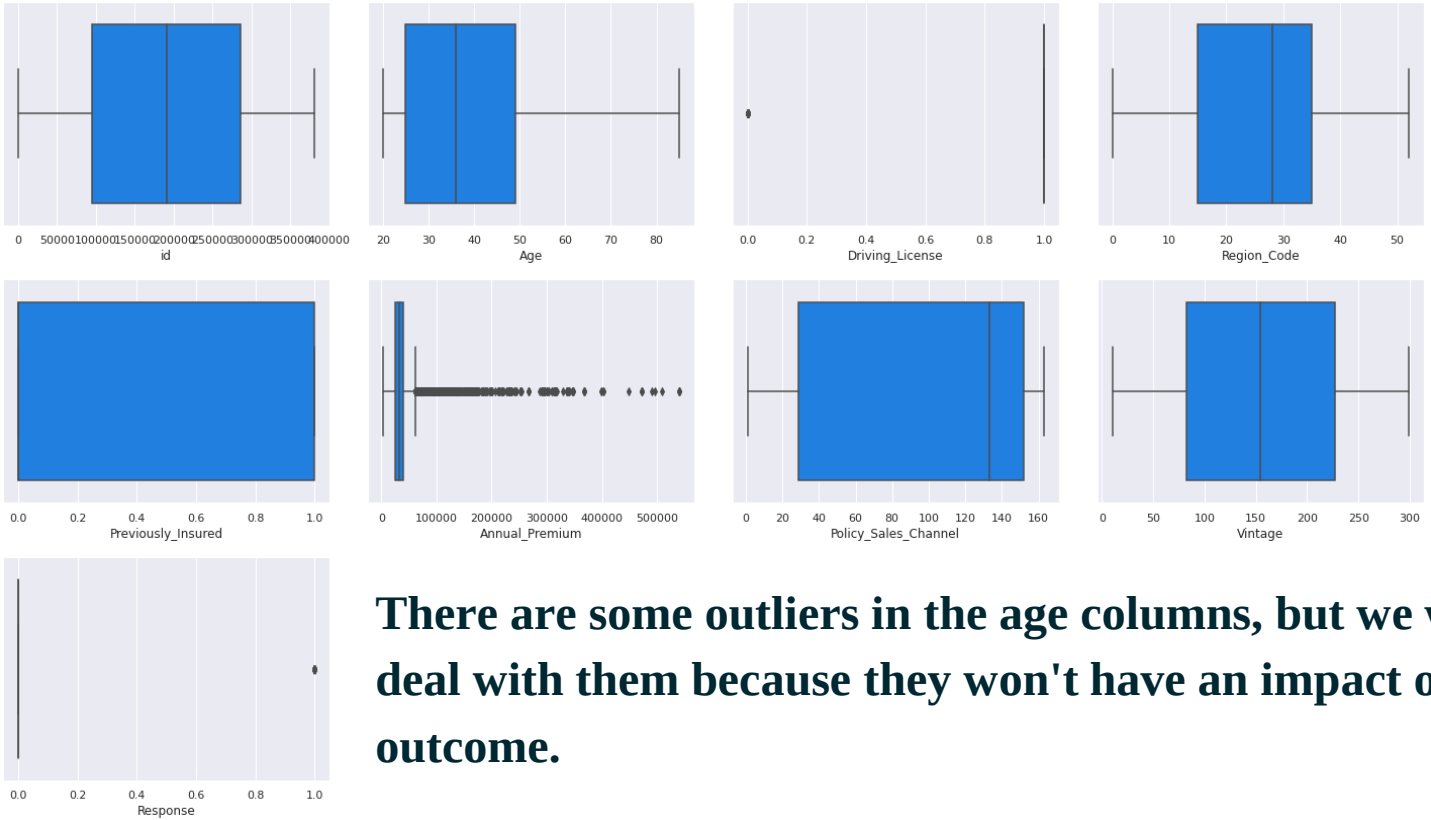
we can see that there is almost same annual premium which is below 200000 in all age group but the premiums of higher values are done by the age group of 40 -60.

And there is not that much of difference in vintage almost every age group has customers of same rate.

Policy channel 26, 124, 152, 156, 160 has the most number of annual premium collections.

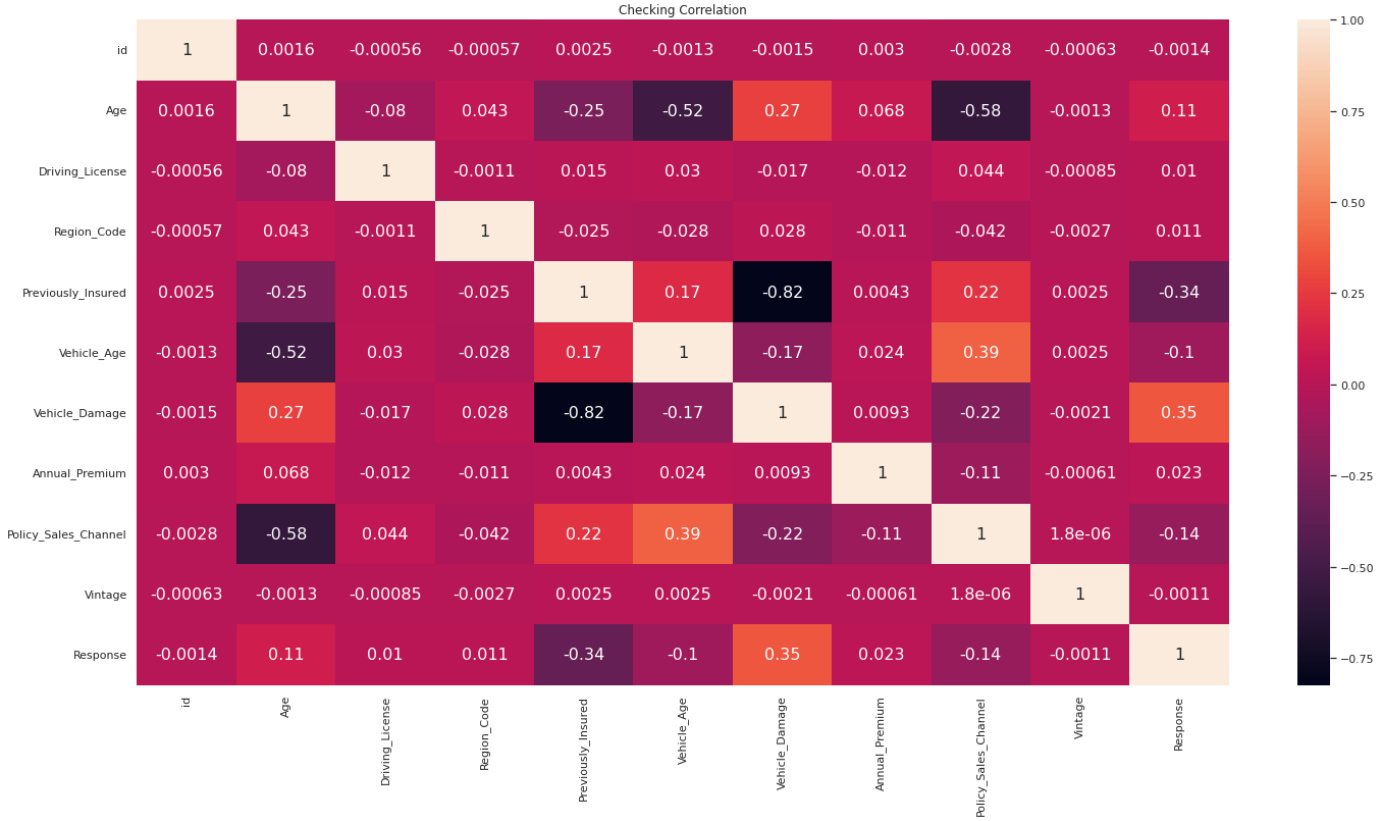
Region code 8, 28, 46 has the customers for the premium collections.

Outlier Analysis



There are some outliers in the age columns, but we won't deal with them because they won't have an impact on the outcome.

Correlation Analysis – Using Heatmap



As we can see that there is no correlated features

Feature Transformation

We have many features that have some categorical values in them like '**Gender'**, **'Vehicle_Age' and 'Vehicle_Damage'**

We have applied LabelEncoder from Scikit Learn and transformed these features to numeric format.

LabelEncoder:

Sklearn is a very efficient tool for encoding categorical feature levels into numeric values. LabelEncoder encodes labels with values ranging from 0 to n classes-1, where n denotes the number of distinct labels.

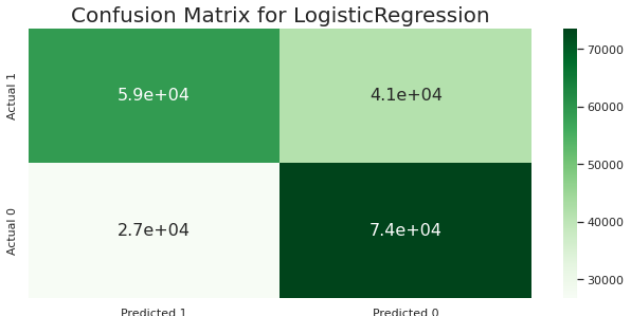
We have used **Synthetic Minority Oversampling Technique (SMOTE)** from imblearn and Balanced our imbalanced class data to save our modelig to be biased

Synthetic Minority Oversampling Technique (SMOTE):

A statistical method for expanding the number of cases in our dataset in a balanced manner. The component creates new instances from existing minority cases that you provide as input.

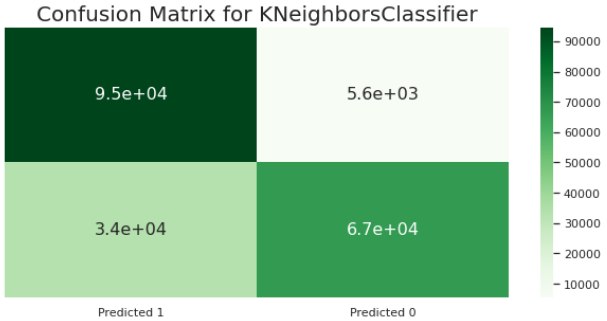
Modeling

Logistic Regression



	Accuracy	Recall	Precision	f1_score
LogisticRegression	0.659714	0.586325	0.686927	0.632652
KNeighborsClassifier	0.803813	0.943823	0.737242	0.827839
DecisionTreeClassifier	0.877487	0.886040	0.871038	0.878475
RandomForestClassifier	0.835123	0.930938	0.781127	0.849478
XGBOOST	0.893760	0.890518	0.896238	0.893369

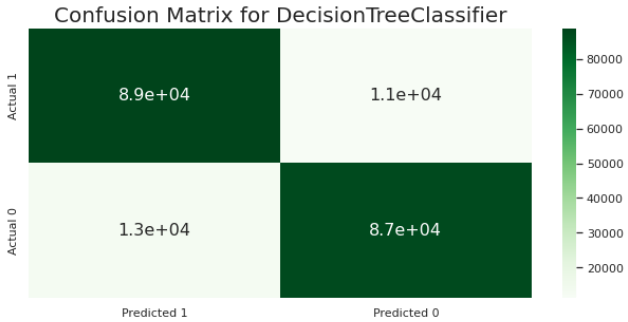
KNN Classifier



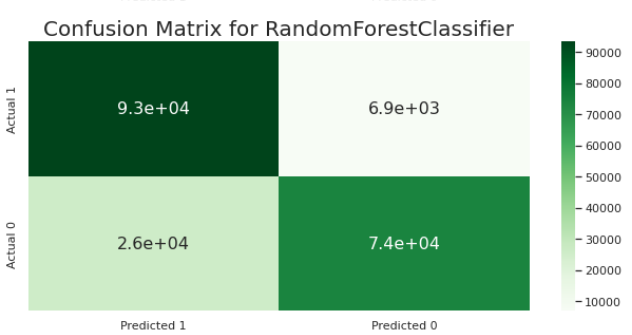
- As we can see that **recall rate of KNN is the highest among all the models** which is **more than 94%**.
- Logistic regression is not performing up to the mark.

Modeling

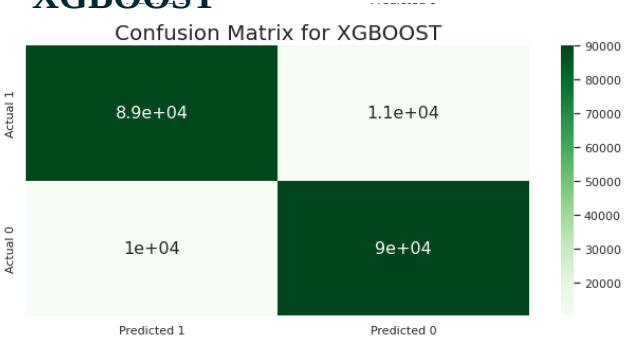
Decision Tree



Random Forest



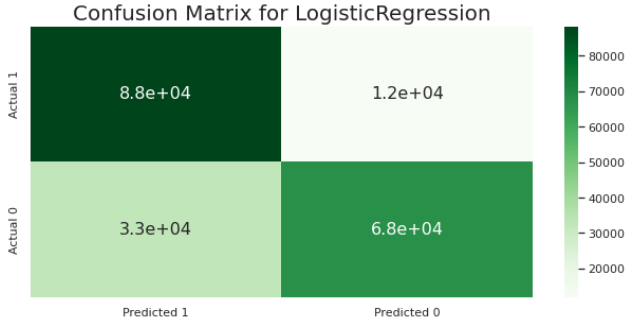
XGBOOST



- **Random forest is doing a great job which is near to KNN in terms of Recall score.**
- **Decision Tree is also doing a descent job** but still its **recall value is less than 90%.**

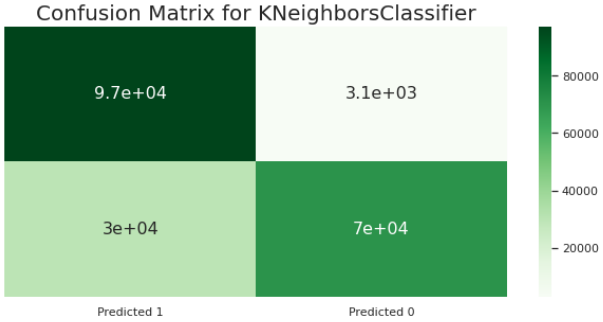
Modeling with Hyperparameter Tuning

Logistic Regression



	Accuracy	Recall	Precision	f1_score
LogisticRegression	0.777352	0.880625	0.729744	0.798116
KNeighborsClassifier	0.835506	0.969284	0.764593	0.854856
DecisionTreeClassifier	0.675434	0.981869	0.608654	0.751474
RandomForestClassifier	0.793072	0.908409	0.738021	0.814398
XGBOOST	0.819936	0.950126	0.753738	0.840614

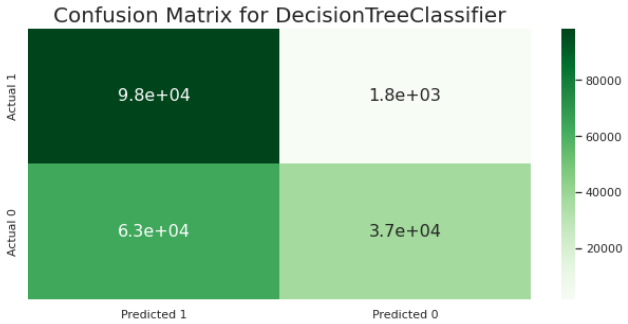
KNN Classifier



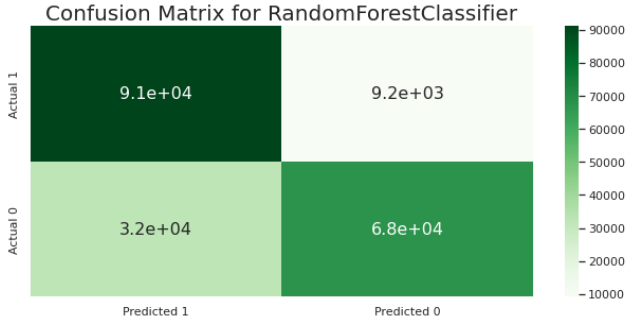
KNN's recall rate is now slightly higher than before, but Decision Tree, after hyper parameter tuning, is doing an excellent job and gives us a recall score of 98 percent. And yet, logistic regression is still underperforming.

Modeling with Hyperparameter Tuning

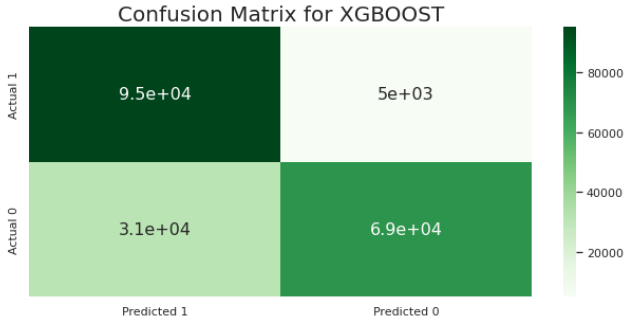
Decision Tree



Random Forest



XGBOOST



After hyperparameter tuning, the recall rate of Random Forest drops from 93 percent to 90 percent.

When we compare KNN and XGBOOST classifiers, we can see that KNN still superior to XGBOOST.

Conclusions

- **We can see that after hyper parameter tuning, the False Negative rate of Logistic Regression is 11970, which is the highest among all models, and the False Positive rate of Random Forest is 9184, which is also high.**
- **When we compare the false positive rates of KNN and XGBOOST, we get 3080 and 5001 respectively, which is lower than the Logistic and Random Forest models but higher than the Decision Tree Decision Tree Classifier, which has the lowest False Positive Rate of 1818 of all the models.**
- **However, if we consider good recall, accuracy, and a low false positive rate, KNN does a great job with (96 percent recall, 83 percent accuracy, and a 3080 false positive rate) if all parameters are considered, our final model should be KNN Classifier.**
- **However, if we are only concerned with Recall Rate, we should use Decision Tree Regressor.**
- **As we improve the KNN Classifier Model with more hyper parameter tuning, we will get more desired results that will be more useful.**

Conclusion Cont.

- **We can conclude from our preliminary analysis of feature distribution, outlier detection, and overall, EDA analysis that: Customers who have never had a vehicle damaged are only 0.5 percent interested in vehicle insurance.**

- **Vehicles less than a year old are more likely to have insurance, with 66 percent insured, while vehicles older than a year but less than two years old are insured in 33 percent.**

THANK YOU