Capstone Project

Health insurance cross sell prediction

Puneet Subanji

Content

1. Data Summary
2. Analysis Of Data
3. Challenges
4. Conclusion

Problem statement

Predict Health insurance owners' who will be interested in buying Vehicle Insurance :
Our client is an Insurance company that has supplied Health Insurance to its customers now they require help in building a model to predict whether the consumers from the past year will also be interested in Vehicle Insurance provided by the company.

Python Modules/Packages/Libraries.

Let us begin our analysis by loading the above mentioned Python Modules/Packages/Libraries.

```
# importing libraries
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from imblearn.over_sampling import RandomOverSampler
from collections import Counter
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import
precision_score,recall_score,accuracy_score,f1_score,confusion_matrix,roc_auc_sc
from sklearn.ensemble import RandomForestClassifier
from xgboost import XGBClassifier
```

Data Summary

Data set name:-**Health insurance cross sell prediction.csv** - the training set (contains 381109 Insurance records)
The dataset is based on the Health insurance data made available.

Shape :

- Rows: **381109**
- Columns: **12**

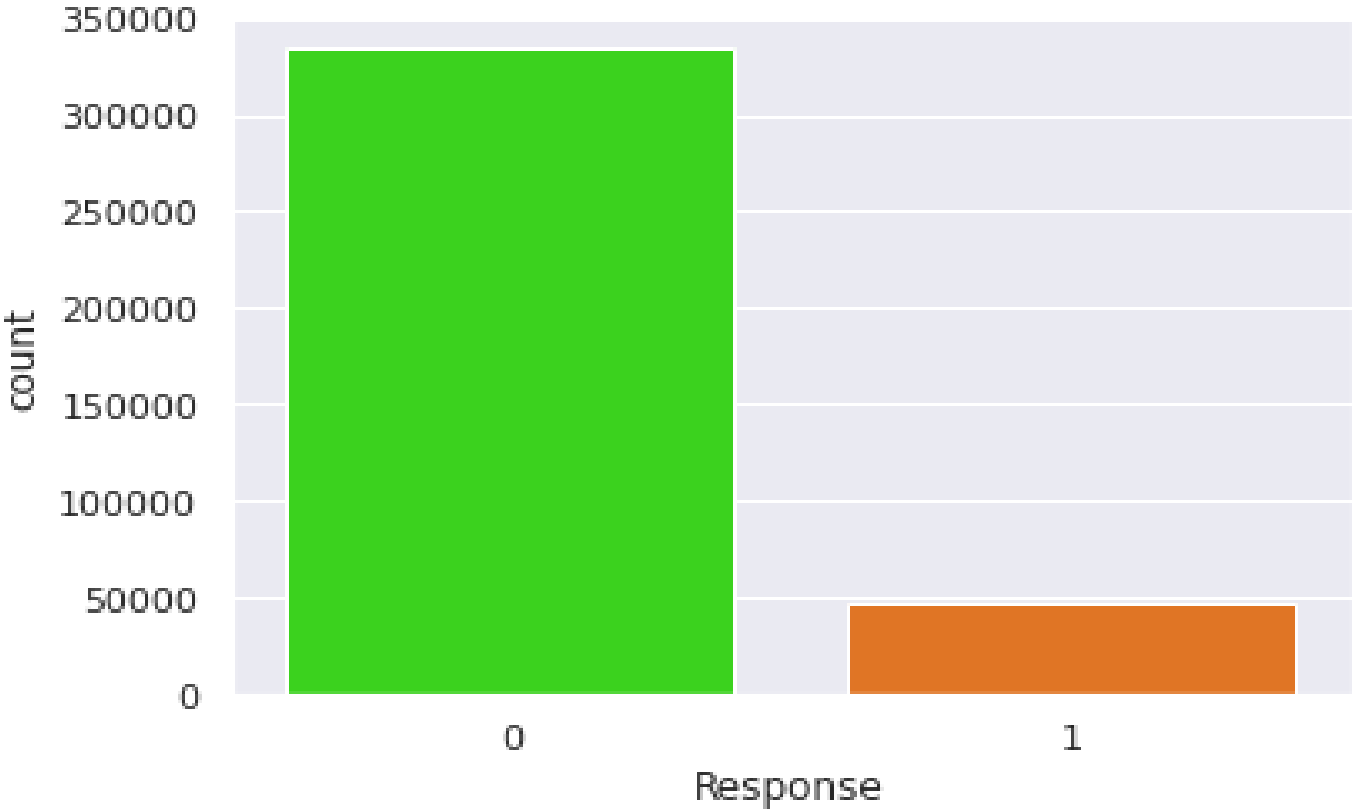
Important Columns:

['id Gender Age
Driving_License Region_Code Previously_Insured
Vehicle_Age Vehicle_Damage Annual_Premium
Policy_Sales_Channel Vintage Response']

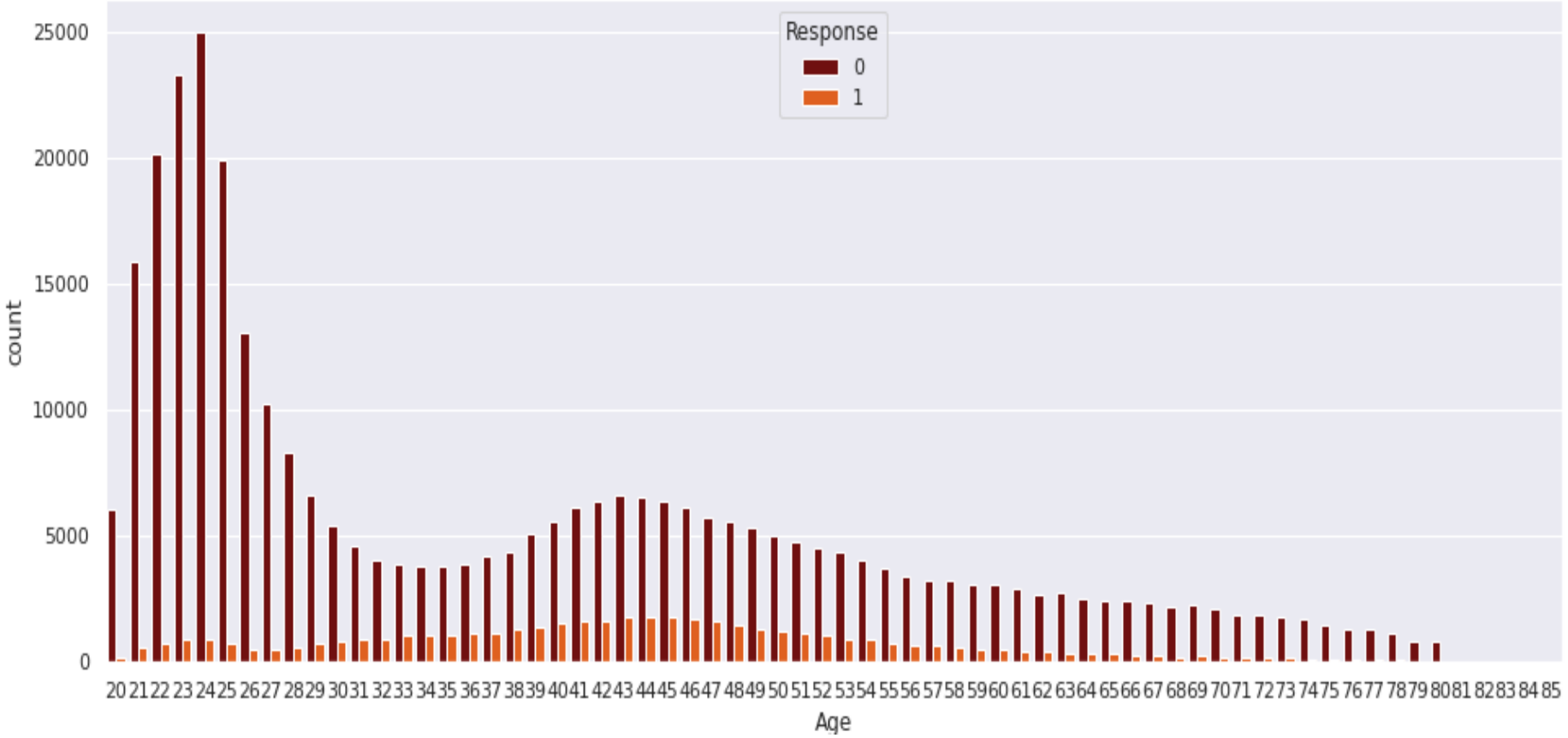
DATA Description:

- id : Unique ID of the existing Health insurance customer
- Gender : Gender details of the health insurance owner.
- Age : Age details of the health insurance owner.
- Driving_License : Whether the customer has a driving license or Not.
- Region_Code : Region with code details of the health insurance owner.
- Previously_Insured : Whether the customer previously_Insured or Not.
- Vehicle_Age : Age of vehicle of the health insurance owner.
- Vehicle_Damage : Whether the customer Vehicle Damaged or Not.
- Annual_Premium : Annual Premium amount details of a Customer.
- Policy_Sales_Channel : Policy Sales Channel shows us,the number of the sales channel.
- Vintage : vintage details of year and car.
- Response : Response of the customer to buying vehicle insurance.

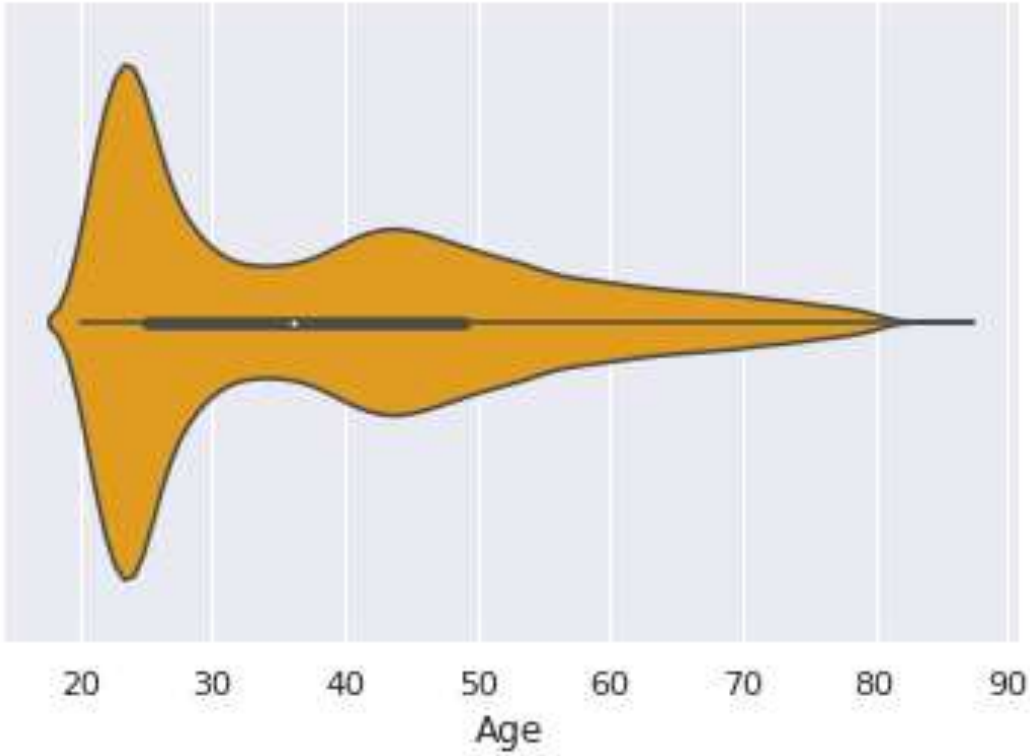
Analysis of Response and Count



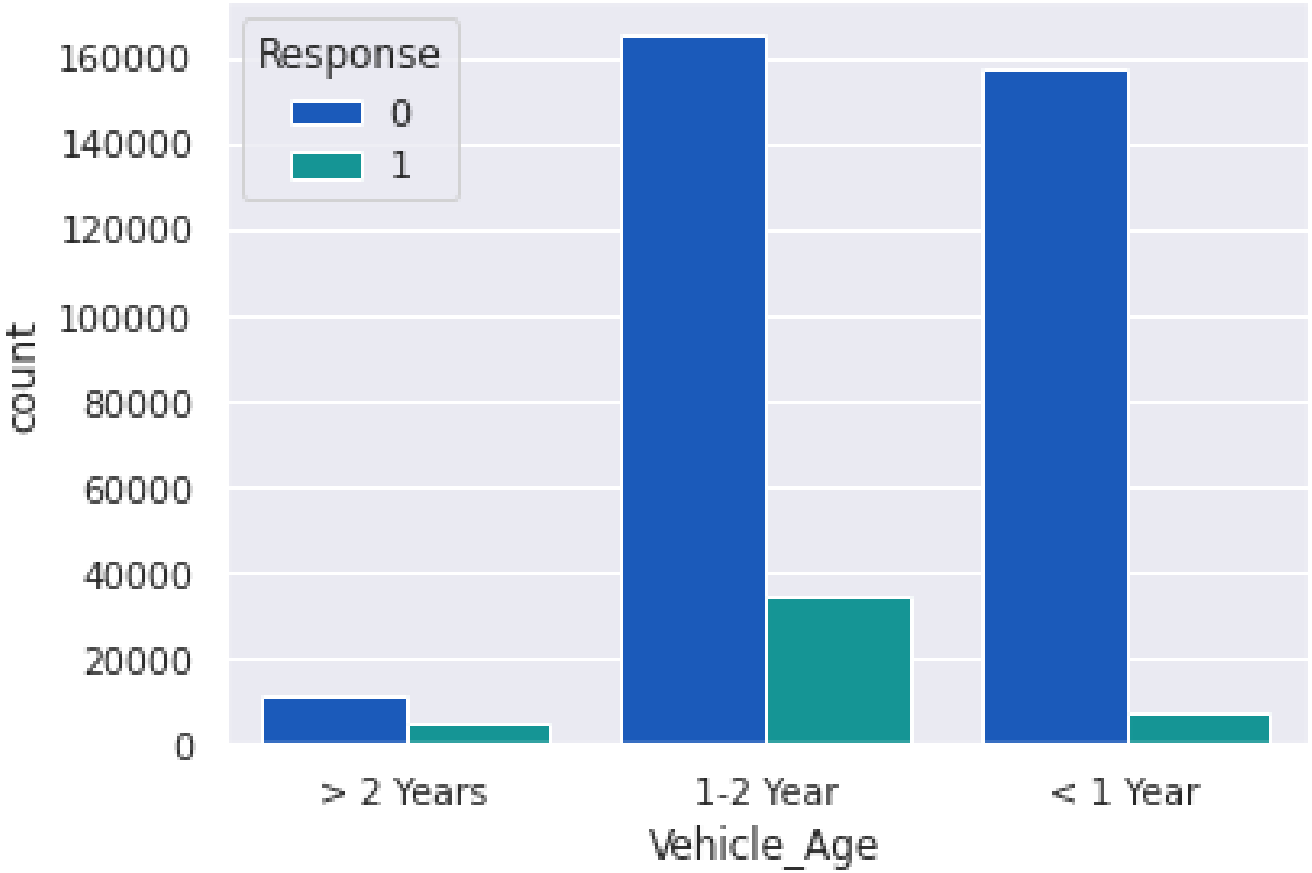
Analysis of Age vs Response



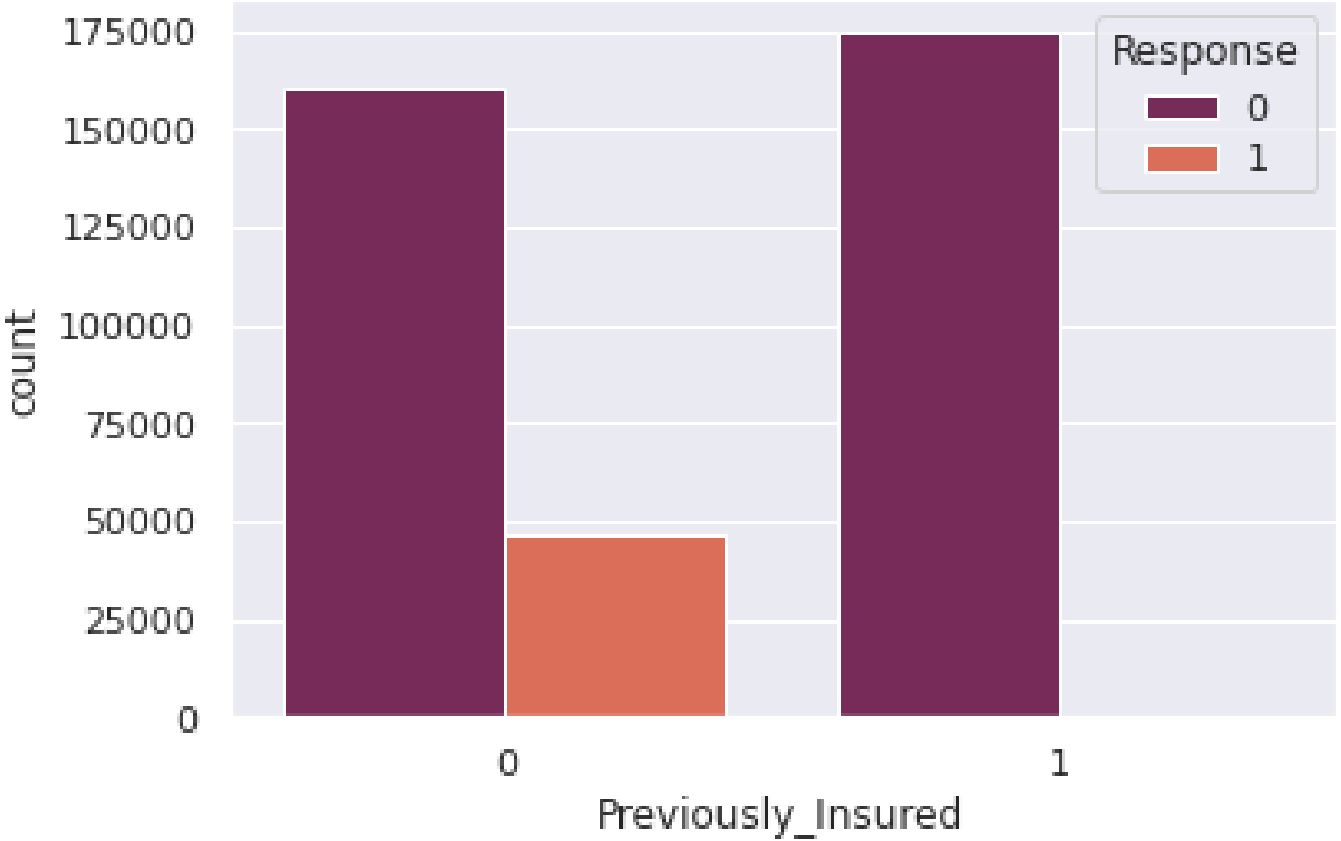
Analysis of Age



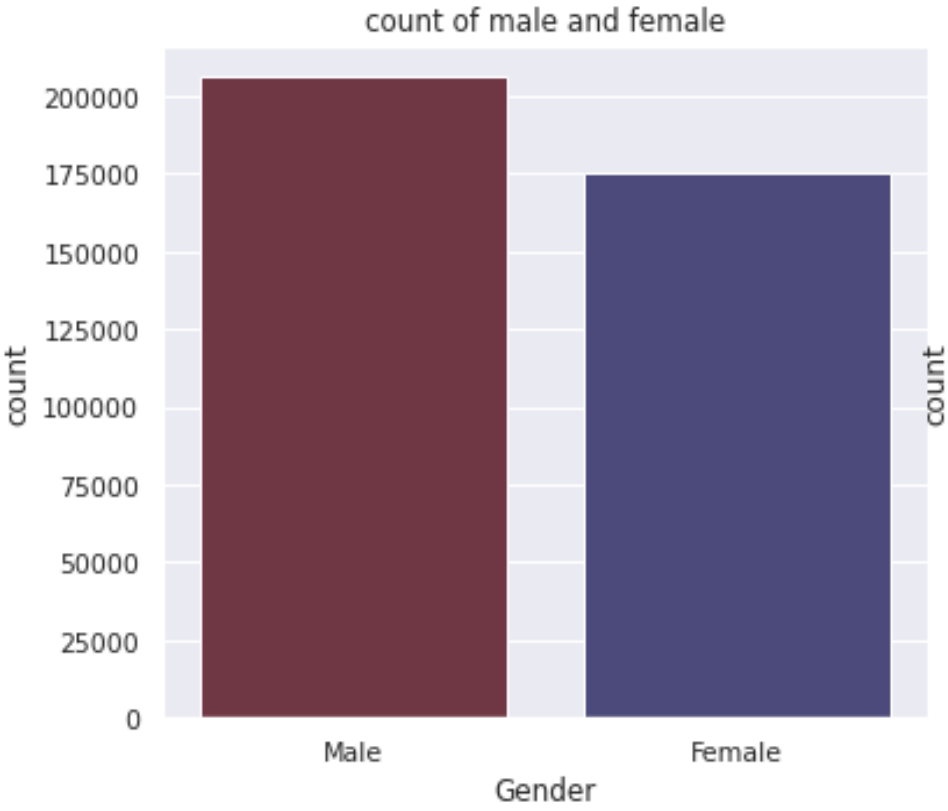
Analysis of Vehicle_Age w.r.t Response

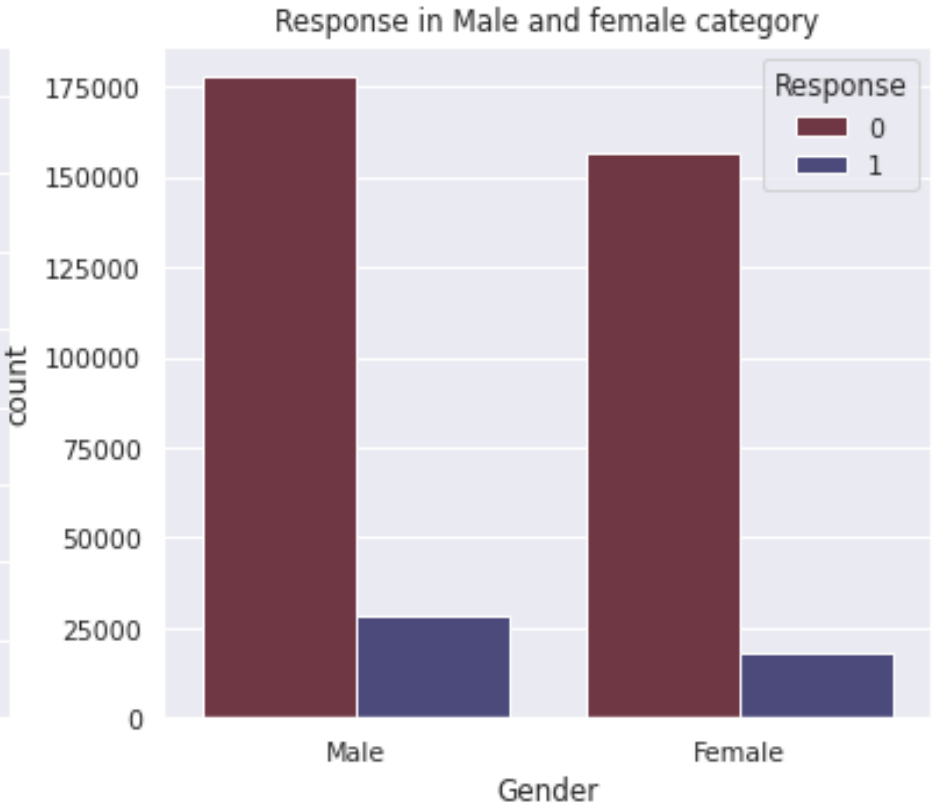


Analysis :Previously_Insured w.r.t Response

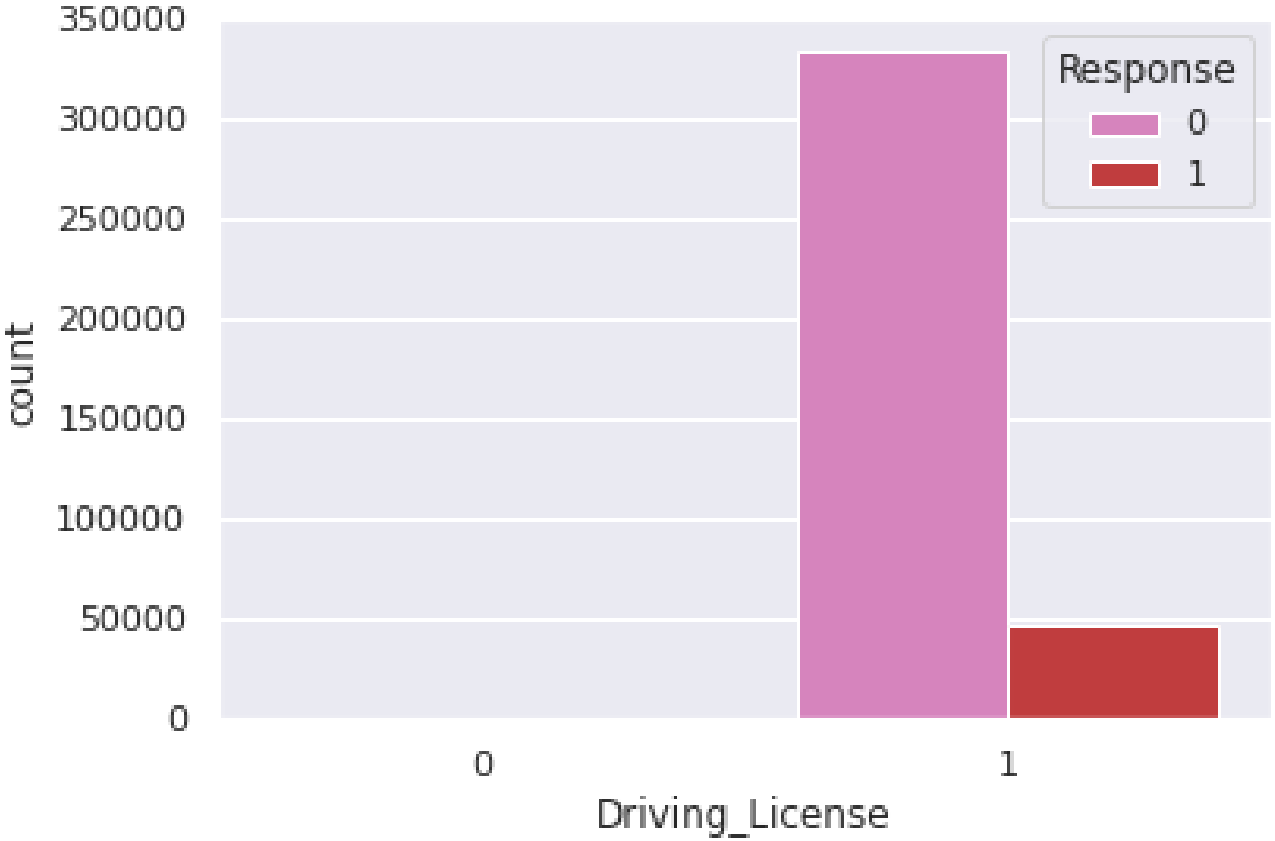


Analysis based on Gender

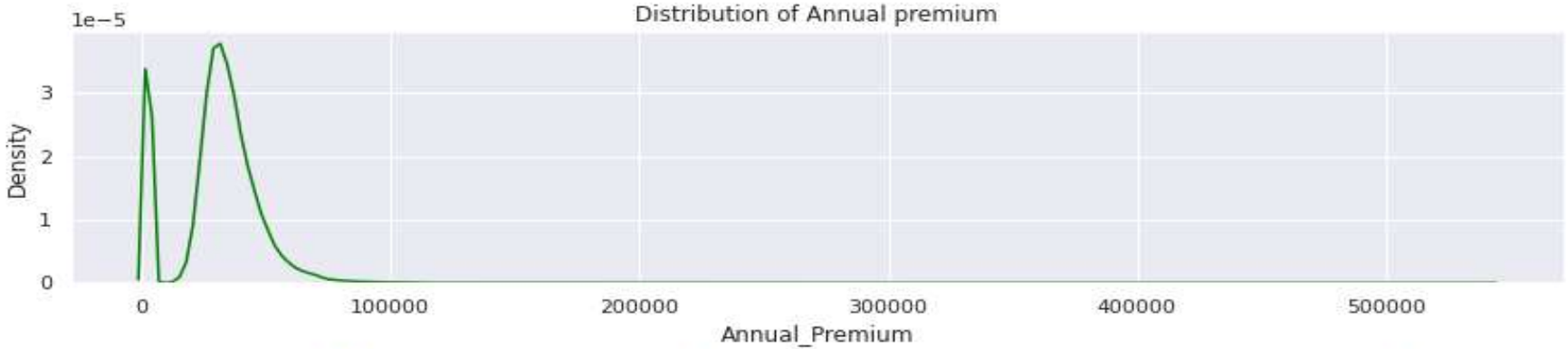




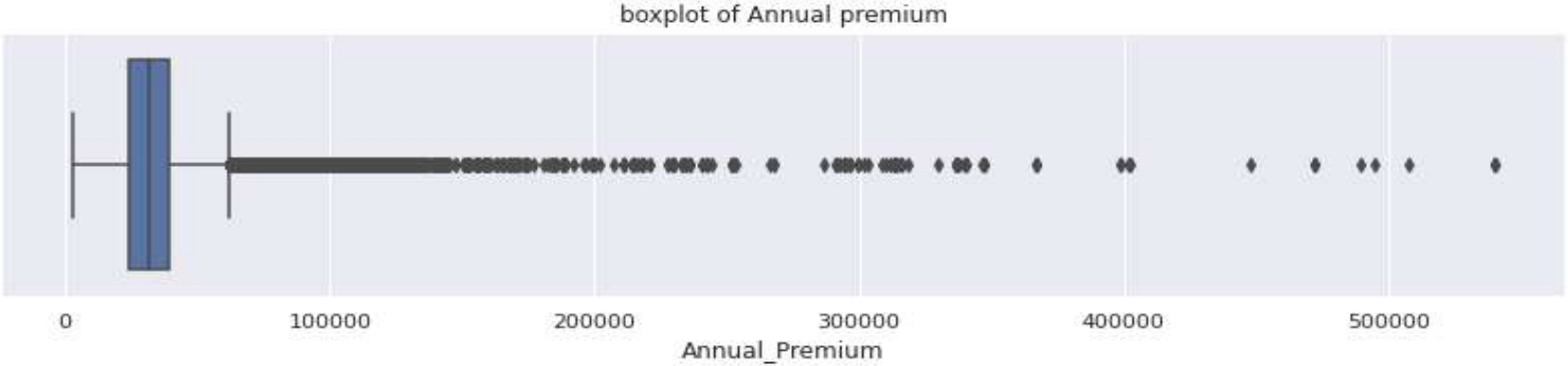
Analysis on Driving License



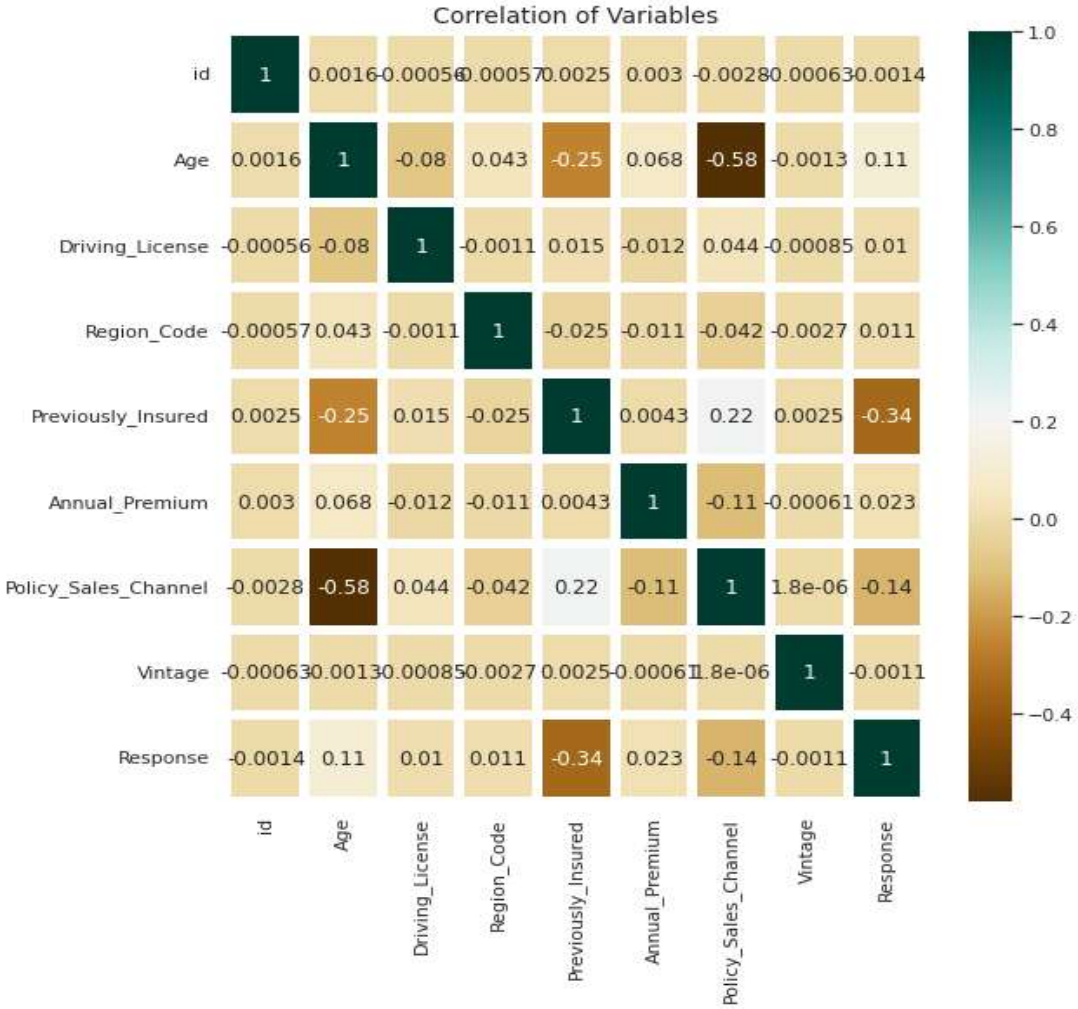
Analysis based on Annual_Premium



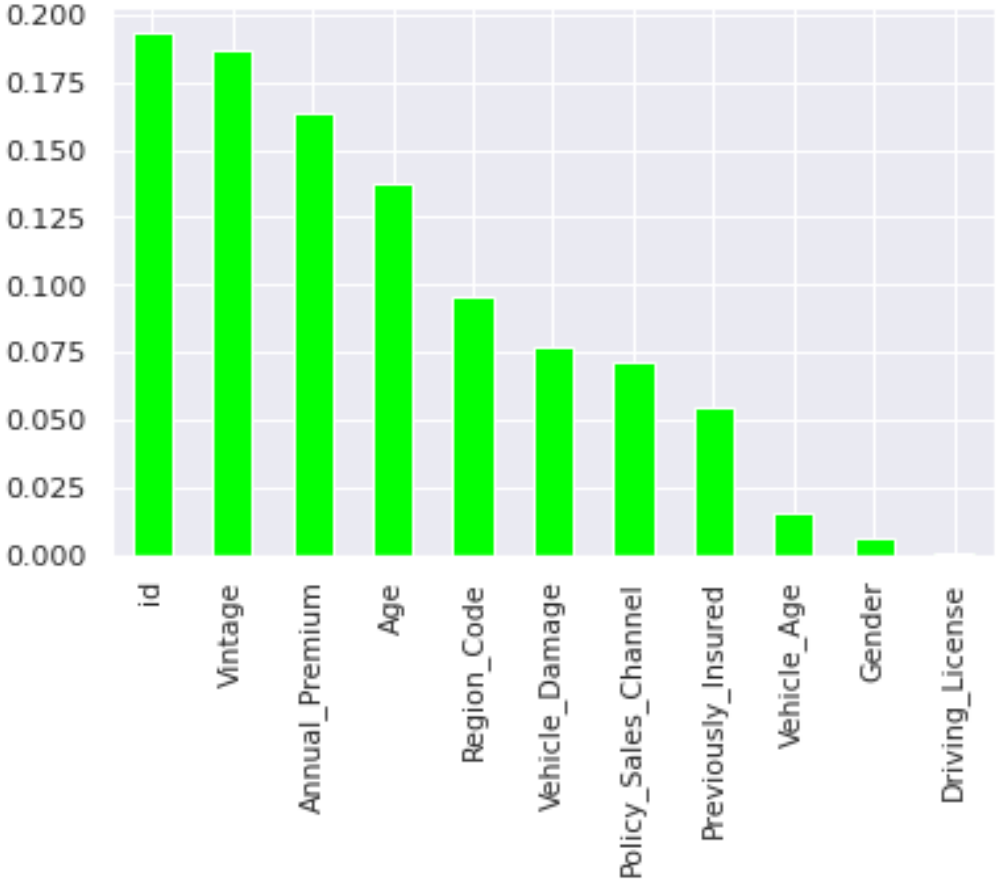

```
/usr/local/lib/python3.8/dist-packages/seaborn/_decorators.py:36: FutureWarning: Pass the following variable as
  warnings.warn(
```



Defining Correlation



Feature Selection



Model selection and fitting

- The Problem can be identified as Binary Classification such as whether consumer purchases vehicle insurance or not.
- Data contains more than 300000 information or variables.
- It’s smarter way to stay away from SVM Classifier because takes more time to train as the dataset increases.

Fitting of Models

1. Logistic regression model
2. Random forest classifier
3. XGBoost Classifier

Challenges

1. The data set was challenging to comprehend and manipulate.
2. Due to the Binary nature of the problem, finding and fitting the appropriate model for the data might be difficult.
3. Identifying the key factors and deciding best fit models after getting results of models.
4. We were unable to do visualisation and fitting of model easily due to bulk and running models were taking lot of time.

CONCLUSION

1. More customers between the ages of 30 and 60 are likely to purchase insurance.
2. Vehicle insurance is not interesting to anyone under the age of 30. The lack of involvement could be a factor, they may not yet have expensive vehicles and have little knowledge about insurance.
3. Consumers with 1-2-year-old vehicles are more interested as compared to others.
4. Customers who own vehicles that are less than 1 year old have very little chance of purchasing insurance.
5. Customers with driver license are more likely to get insurance.
6. Vehicle damage customers are more likely to purchase insurance.
7. The male category is slightly more notable than the female category, and chances of buying the insurance are likewise minimally high.
8. The variable such as Age, Previously_insured, Annual_premium is more affecting the target variable.
9. We can observe from a comparison of the ROC curve that the Random Forest model performs better. Because better performance is shown by curves that are closer to the top-left corner.