Capstone Project - 3

Health Insurance Cross Sell Prediction (Supervised ML Classification)

Team Members

Nihal Habeeb

Parvez Makandar

Content

- Problem Description
- Objective
- Data Summary
- Exploratory Data Analysis
- Models
- Model Selection
- Feature Importance
- Conclusion

What is cross selling?

Offering a product that is similar or compatible to the product that the customer already purchased, in a way that can potentially add value to the customer experience (maybe by offering discounts).

It can:

- build strong relationship with the customer base (customer loyalty).
- increase customer satisfaction.
- boost the revenue of the company.

Why is customer loyalty so important?

- Acquiring new customers can be really expensive, it's definitely more expensive than retaining current customers. The more competitive the industry gets, the more harder it is to retain customers.
- A customer with strong relationship can lead to further cross selling and more revenue.

A company needs to stand out by providing more value to the customer than the competitor. Cross selling can achieve this:

- If rightly done, it can satisfy customers’ needs further and,
- Build a deeper relationship with them.

But it can go wrong…

If the customer is not interested in the product, it can lead to:

- Customer being annoyed, thus risking the quality of the customer relationship
- Marketing efforts going to waste

Problem Description

- Our client is an Insurance company that provides Health Insurance.
- We want to predict if a health insurance policy holder would be interested in the vehicle insurance.
- This makes sure the cross selling strategy does not affect the customer relationship negatively.
- It insures the efficiency of marketing and communication efforts.

Objective:

- Understand the application of machine learning models in developing cross selling strategy.
- Leverage machine learning algorithms such as Logistic Regression, Decision Tree and Random Forest to create classification models that classify a customer as ‘Interested’ or ‘Not Interested’ (in vehicle insurance) based on features available in existing data of customers.
- Evaluate model performance using different metrics such as Accuracy, F1 score, ROC (Receiver operating characteristic) curve etc.

Data Summary

We have information on existing customers such as their gender, age, vehicle age, annual premium and so on. We also know whether these customers were interested in the vehicle insurance or not. We are going to use this information to build models to predict the interest of a customer based on the available features

Features:

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

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

Response (target): 1 : Customer is interested, 0 : Customer is not interested

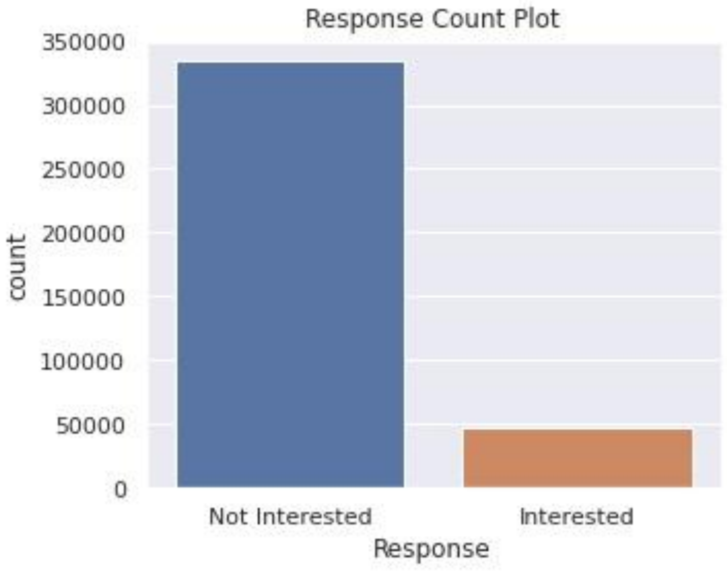
Exploratory Data Analysis

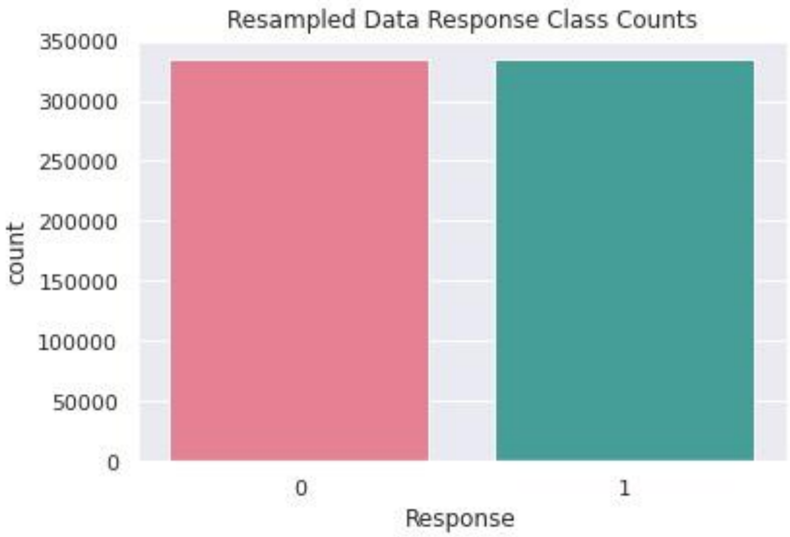
We explored the features in detail:

- Used count plots for categorical variables.
- Studied the distribution of response across different categories.
- Distribution plots of numerical variables.

Exploratory Data Analysis

Target Variable

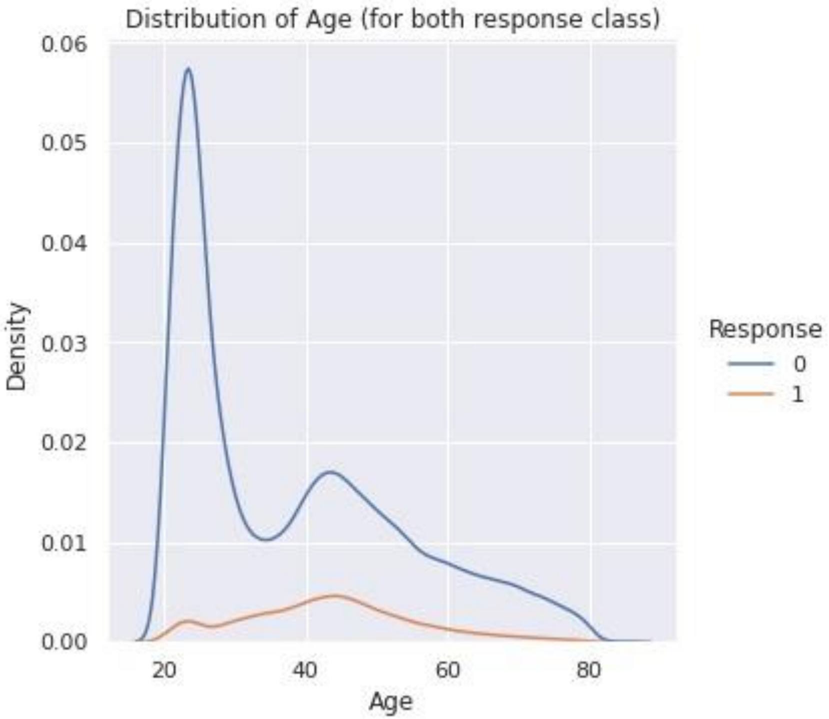


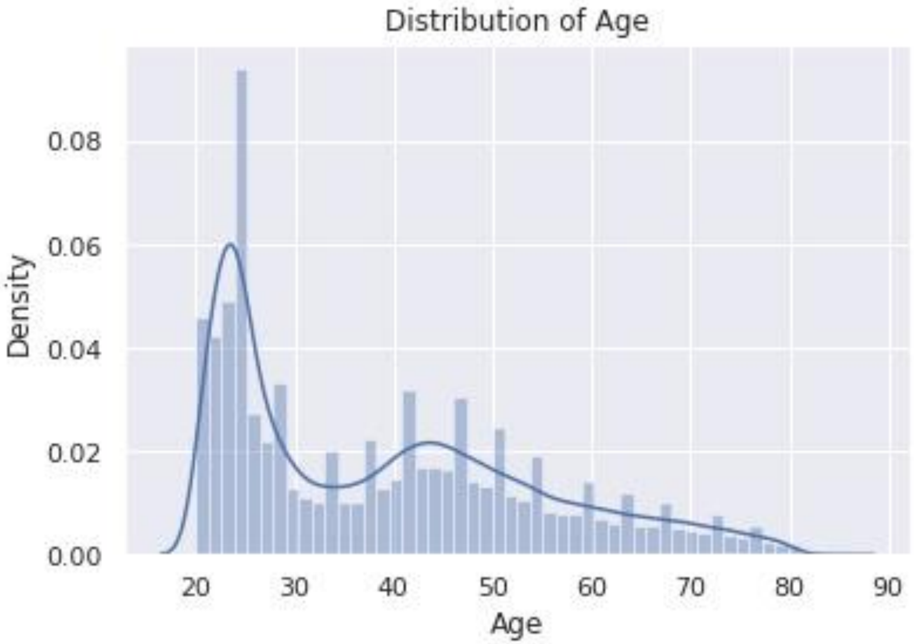


We could observe that the target variable class is highly imbalanced.

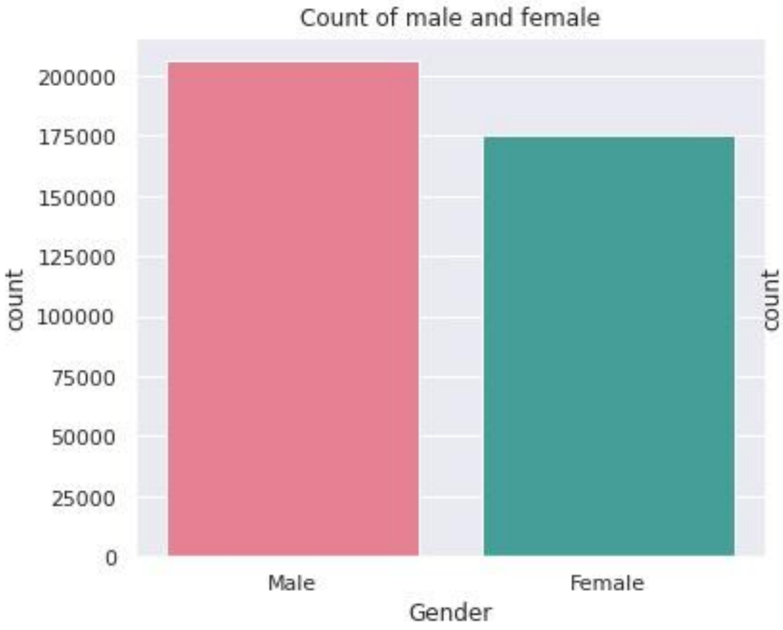
We will use SMOTE (Synthetic Minority Oversampling Technique) to balance the classes.

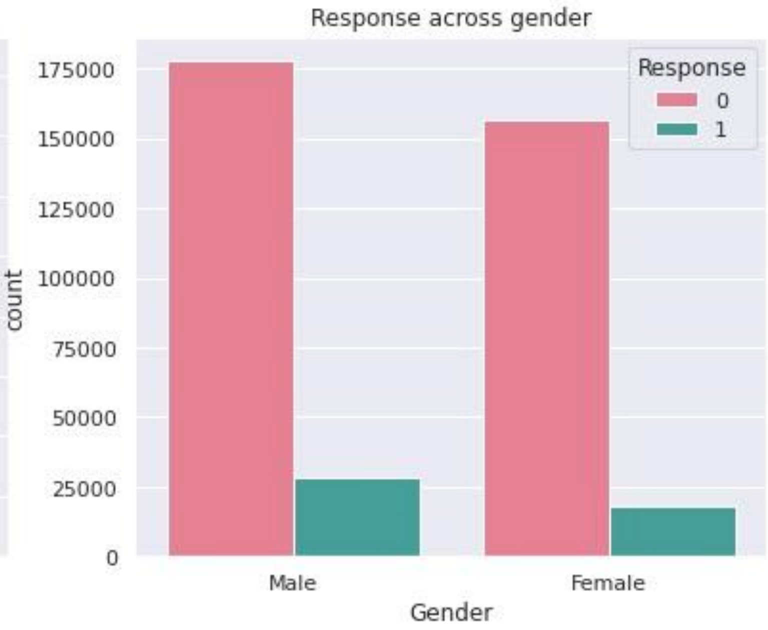
Age


Distribution of Age (for both response class)
Density
0.00
0.01
0.02
0.03
0.04
0.05
0.06
20
40
60
80
Age
Response
0
1

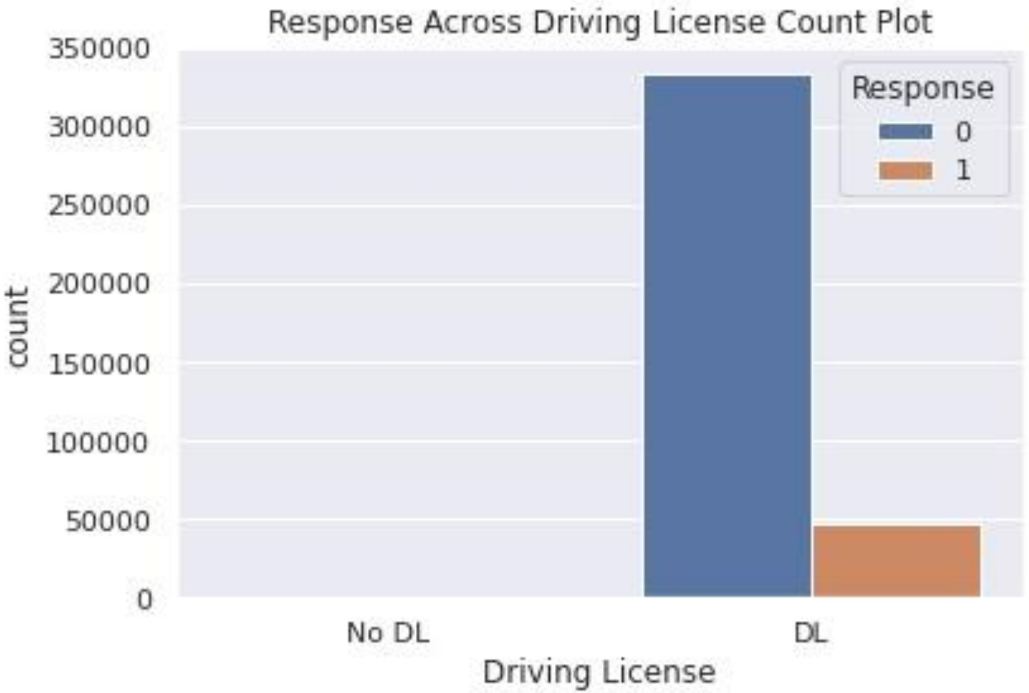

Distribution of Age
Density
0.00
0.02
0.04
0.06
0.08
20
30
40
50
60
70
80
90
Age

Gender

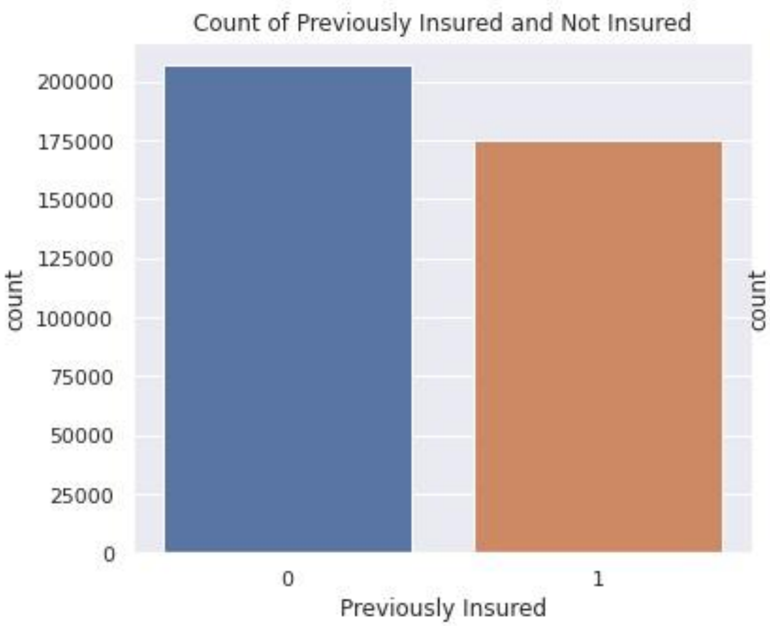

Count of male and female
count
200000
175000
150000
125000
100000
75000
50000
25000
0
Male
Female
Gender

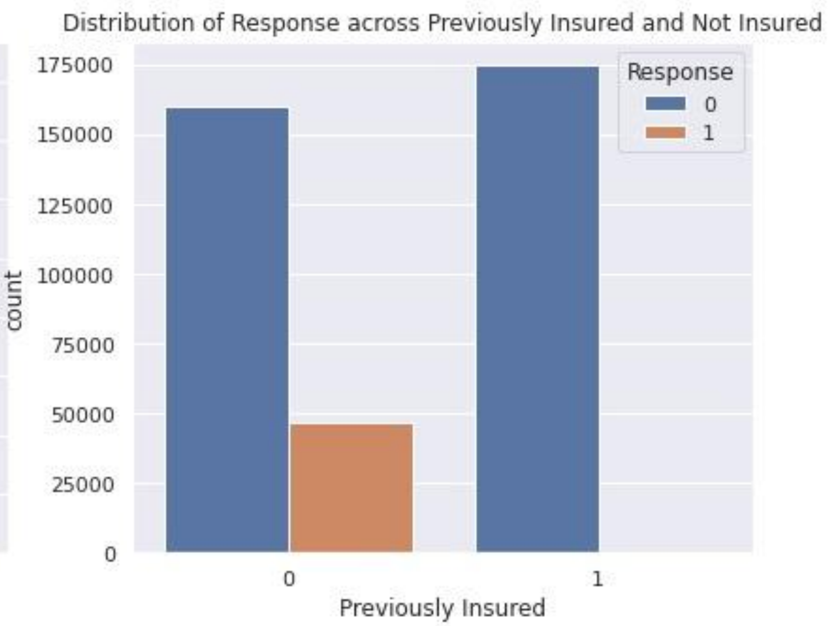

Response across gender
count
175000
150000
125000
100000
75000
50000
25000
0
Response
0
1
Male
Female
Gender

Driving License

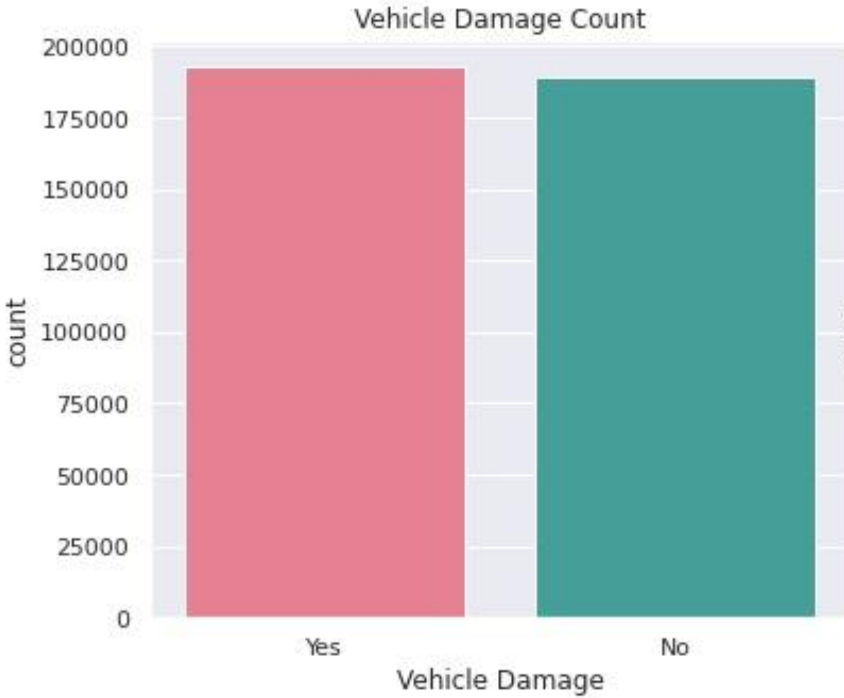


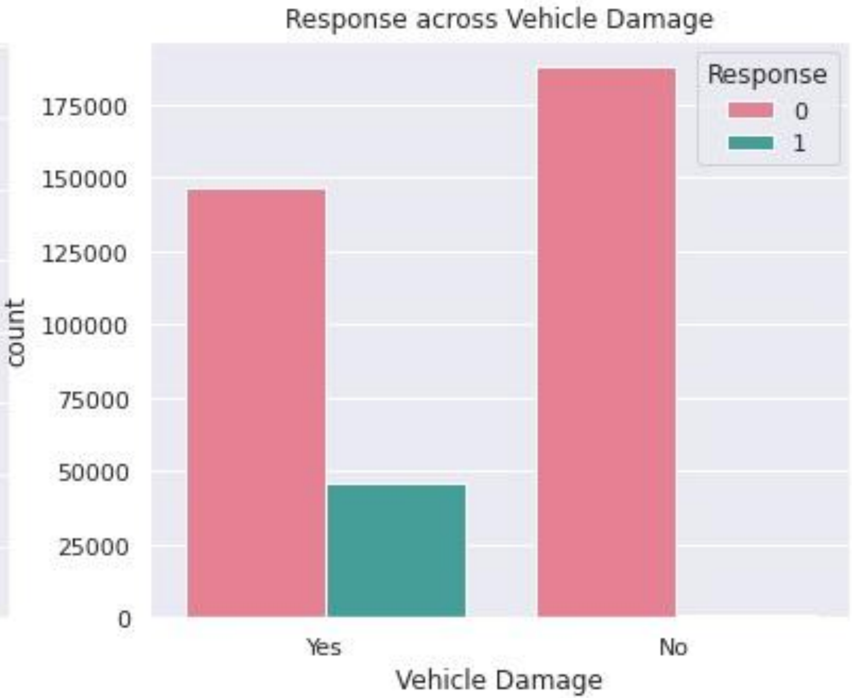
Previously Insured


Count of Previously Insured and Not Insured

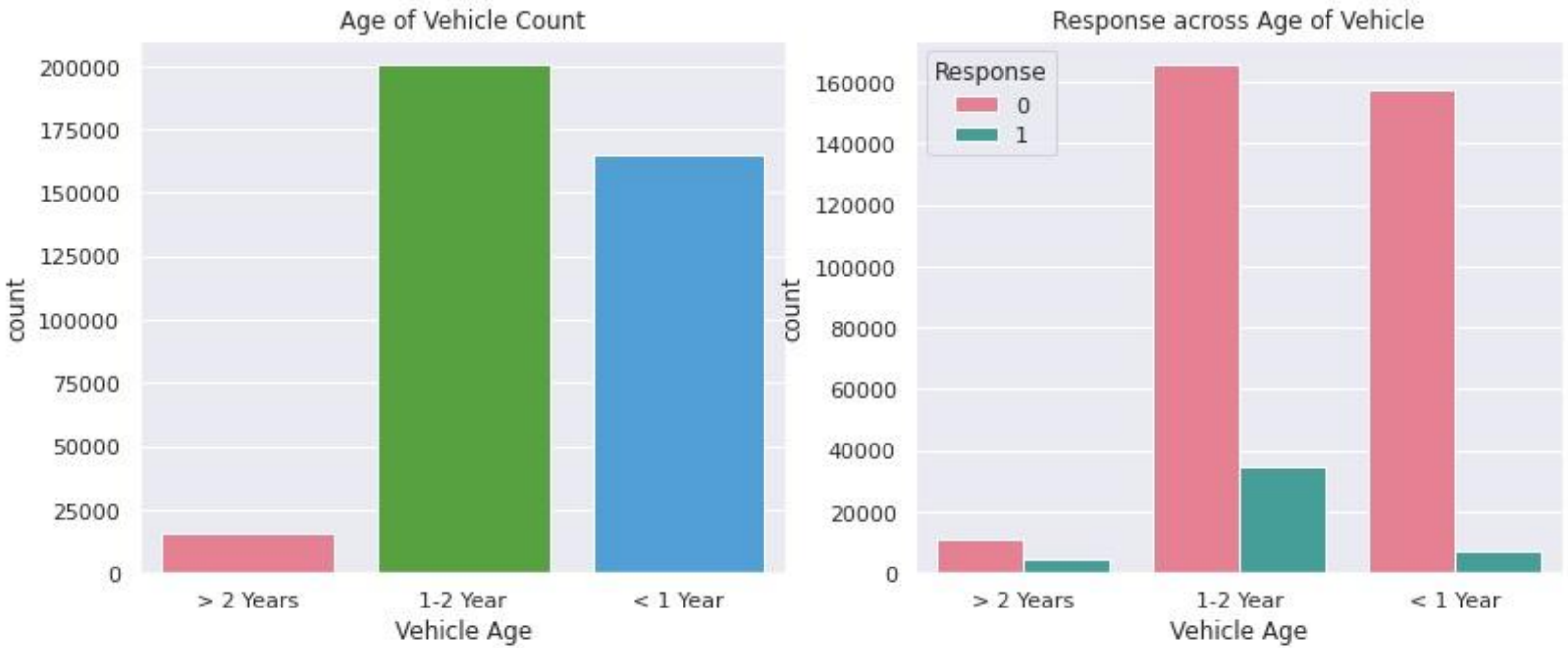

Distribution of Response across Previously Insured and Not Insured

Vehicle Damage

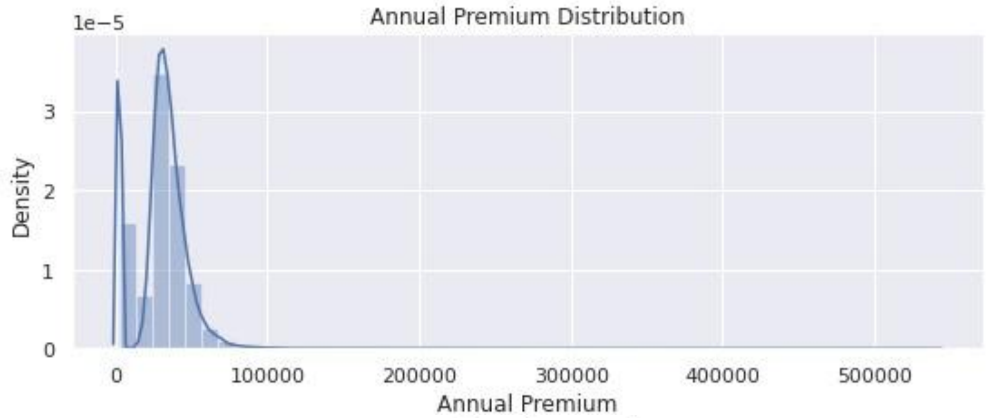

Vehicle Damage Count

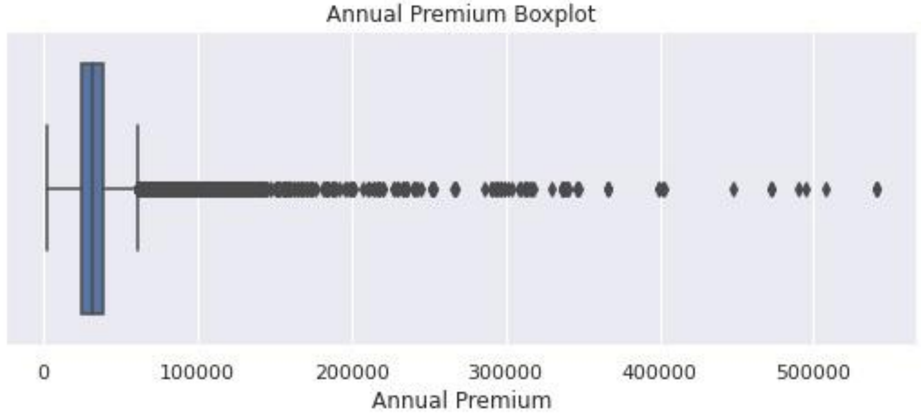

Response across Vehicle Damage

Vehicle Age


Age of Vehicle Count
count
200000
175000
150000
125000
100000
75000
50000
25000
0
> 2 Years
1-2 Year
< 1 Year
Vehicle Age
Response across Age of Vehicle
Response
0
1
count
160000
140000
120000
100000
80000
60000
40000
20000
0
> 2 Years
1-2 Year
< 1 Year
Vehicle Age

Annual Premium

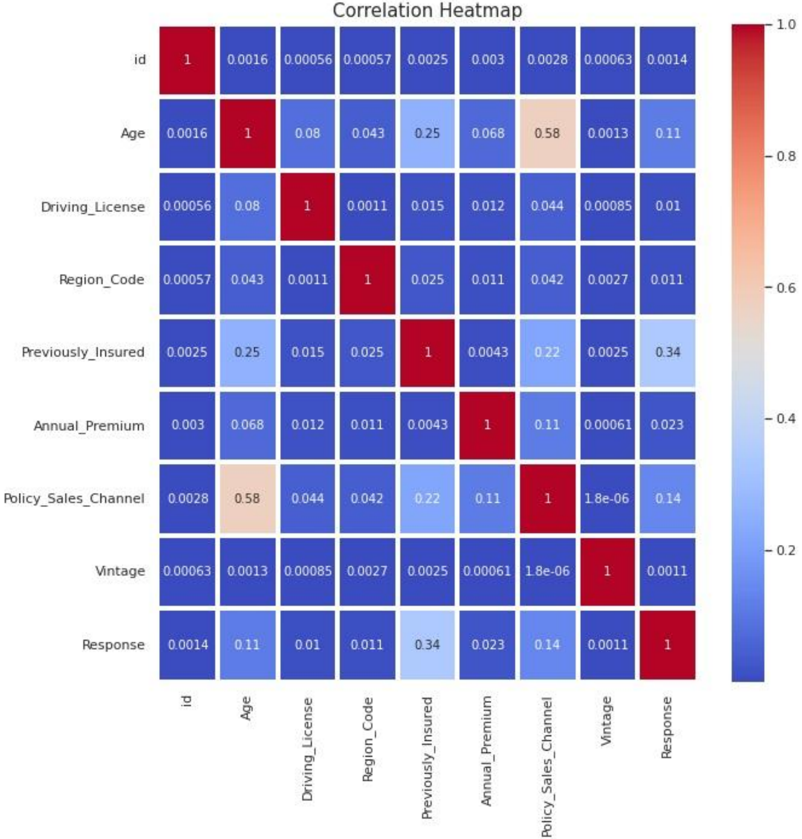




The distribution is skewed.

There are a lot of outliers in Annual Premium. But the models we use are not impacted by outliers.

Multicollinearity



Correlation Heatmap

	id	Age	Driving_License	Region_Code	Previously_Insured	Annual_Premium	Policy_Sales_Channel	Vintage	Response
id	1	0.0016	0.00056	0.00057	0.0025	0.003	0.0028	0.00063	0.0014
Age	0.0016	1	0.08	0.043	0.25	0.068	0.58	0.0013	0.11
Driving_License	0.00056	0.08	1	0.0011	0.015	0.012	0.044	0.00085	0.01
Region_Code	0.00057	0.043	0.0011	1	0.025	0.011	0.042	0.0027	0.011
Previously_Insured	0.0025	0.25	0.015	0.025	1	0.0043	0.22	0.0025	0.34
Annual_Premium	0.003	0.068	0.012	0.011	0.0043	1	0.11	0.00061	0.023
Policy_Sales_Channel	0.0028	0.58	0.044	0.042	0.22	0.11	1	1.8e-06	0.14
Vintage	0.00063	0.0013	0.00085	0.0027	0.0025	0.00061	1.8e-06	1	0.0011
Response	0.0014	0.11	0.01	0.011	0.34	0.023	0.14	0.0011	1

Multicollinearity can affect logistic regression models.

There is no perfect multicollinearity between any features.

Feature Engineering

- Encoding
  - Gender and Vehicle damage (Label Encoder)
  - Vehicle age (One hot encoding)
- Handled class imbalance (using SMOTE)

SMOTE

- Synthetic Minority Oversampling TEchnique
- New sample is created between a minority and a neighbor

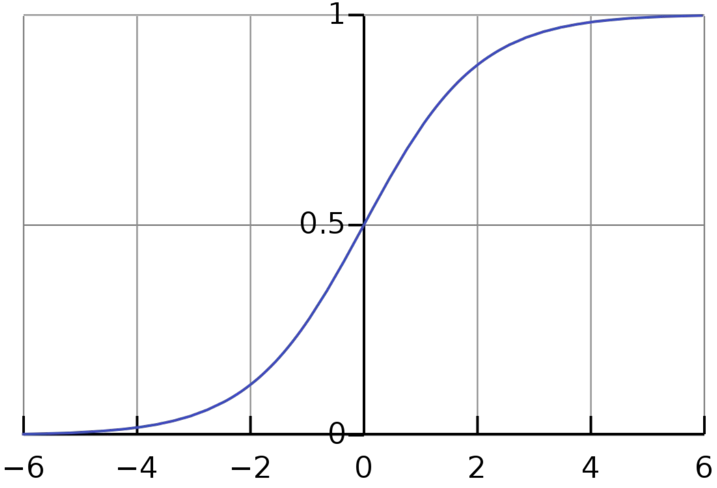
Models

Logistic Regression

$$\log(p/(1-p)) = \omega_0 + \omega_1 x_1 + \ldots + \omega_p x_p$$

$$= \omega^T x$$

$$p = \exp(\omega^T x)/(1+\exp(\omega^T x))$$



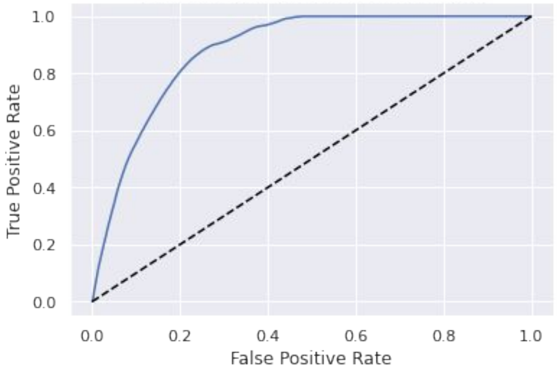
Logistic Regression

Testing Data Performance
Accuracy : 0.81
Precision: 0.77
Recall: 0.90
F1-Score: 0.83
Area Under the ROC Curve: 0.88

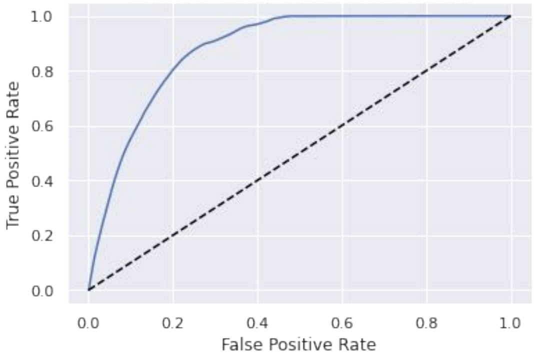
Training Data Performance
Accuracy : 0.81
Precision: 0.77
Recall: 0.90
F1-Score: 0.83
Area Under the ROC Curve: 0.88

Logistic regression

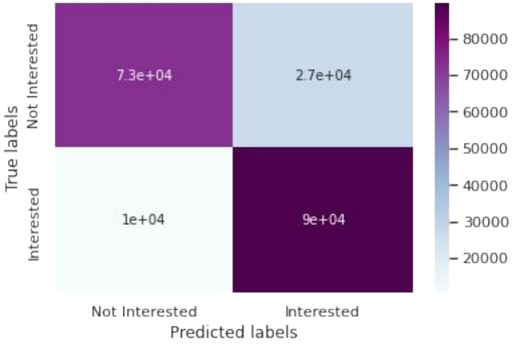
Logistic Regression ROC Curve (Testing)



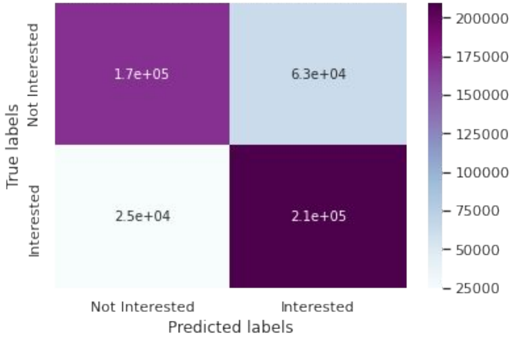
Logistic Regression ROC Curve (Training)



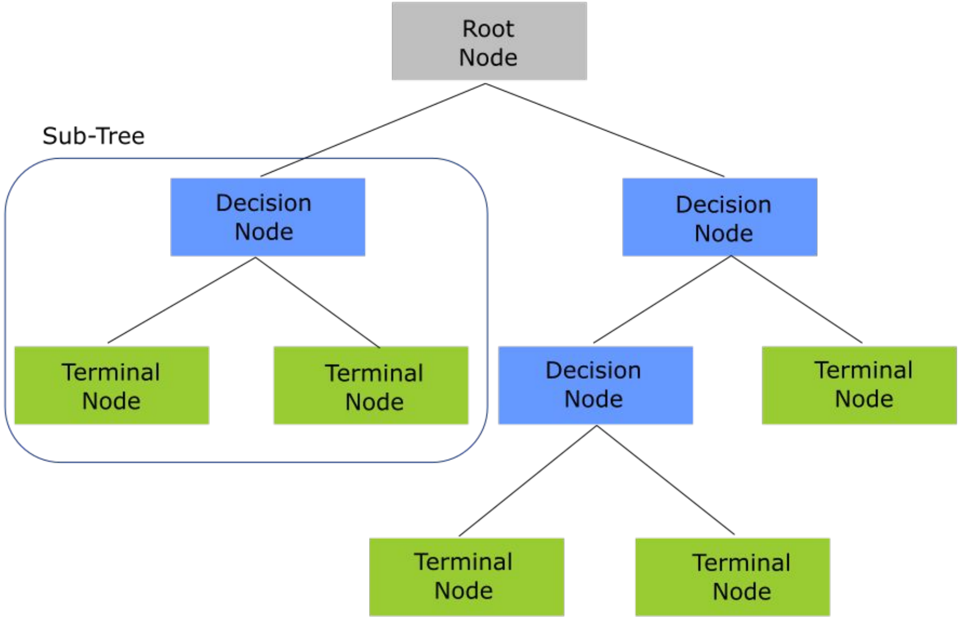
Logistic Regression Confusion Matrix (Testing)



Logistic Regression Confusion Matrix (Training)



Decision Tree Classifier

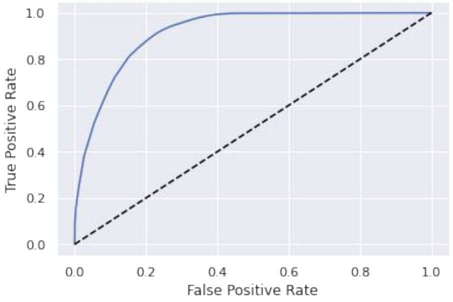


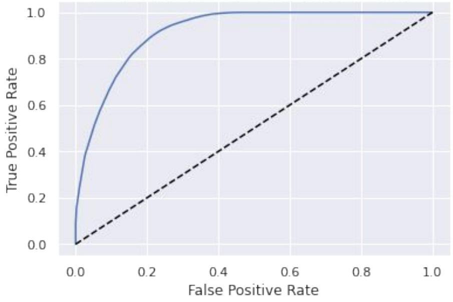
Decision Tree Classifier

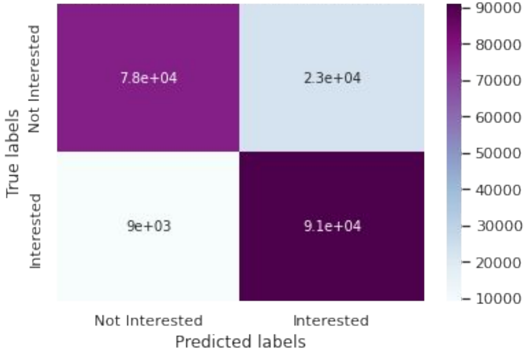
Testing Data Performance
Accuracy : 0.84
Precision: 0.80
Recall: 0.91
F1-Score: 0.85
Area Under the ROC Curve: 0.92

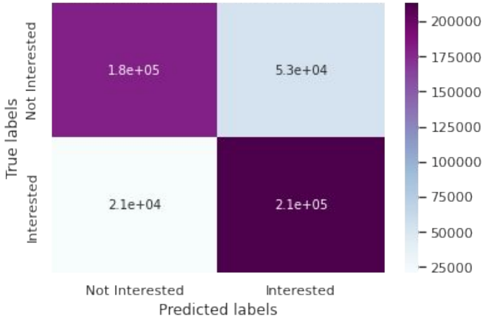
Training Data Performance
Accuracy : 0.84
Precision: 0.80
Recall: 0.91
F1-Score: 0.85
Area Under the ROC Curve: 0.92

Decision Tree Classifier

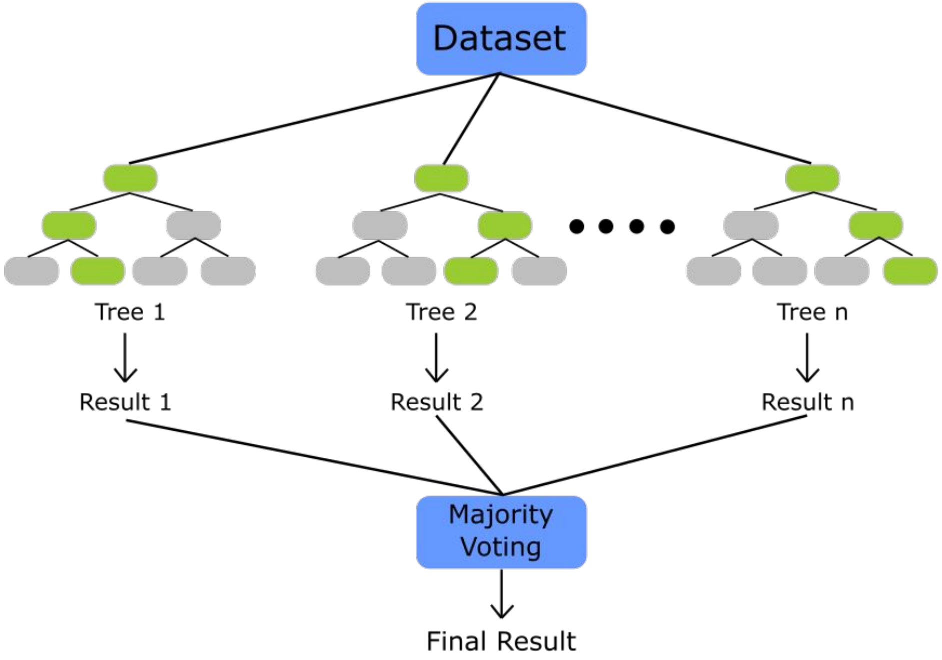

Decision Tree Classifier ROC Curve (Testing)


Decision Tree Classifier ROC Curve (Training)


Decision Tree Classifier Confusion Matrix (Testing)


Decision Tree Classifier Confusion Matrix (Training)

Random Forest Classifier



Random Forest Classifier

Testing Data Performance

Accuracy : 0.84

Precision: 0.78

Recall: 0.93

F1-Score: 0.85

Area Under the ROC Curve: 0.92

Training Data Performance

Accuracy : 0.84

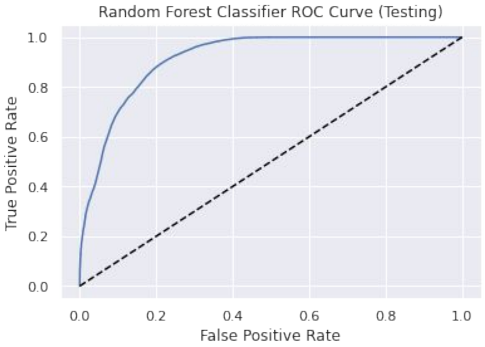
Precision: 0.79

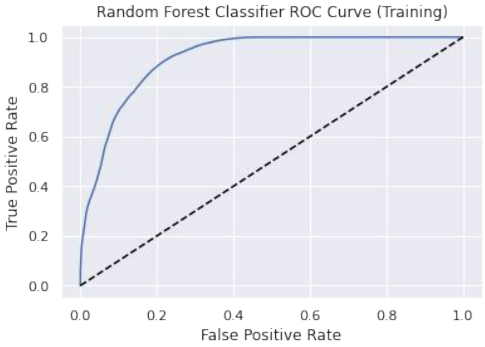
Recall: 0.93

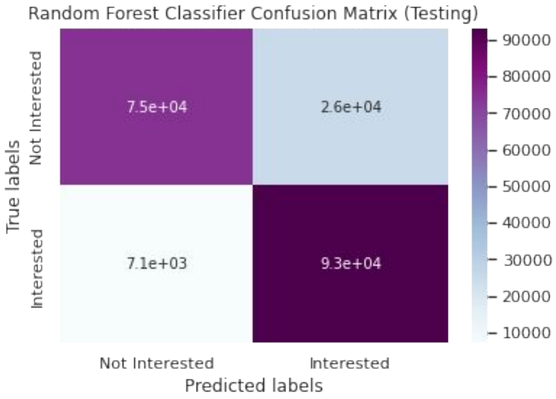
F1-Score: 0.85

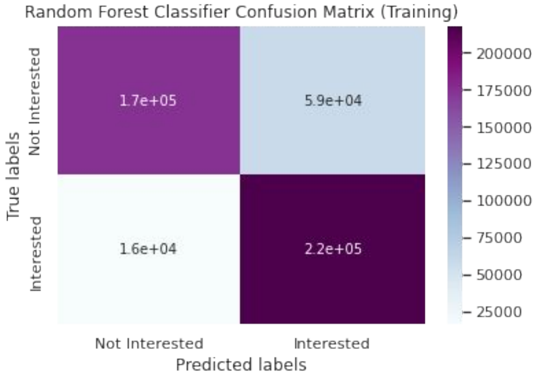
Area Under the ROC Curve: 0.92

Random Forest Classifier









XGBoost Classifier

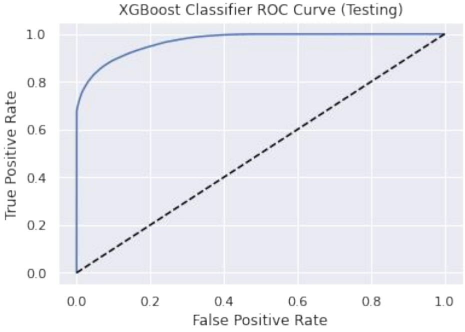
- XGboost stands for Extreme Gradient Boost.
- In Boosting, trees are build sequentially, with each tree “learning” from the previous tree and updating the error.
- Gradient boosting involves improving (or boosting) a single weak model by combining it with a number of other weak models to generate a collectively strong model.
- XGBoost is an implementation of gradient boosting.

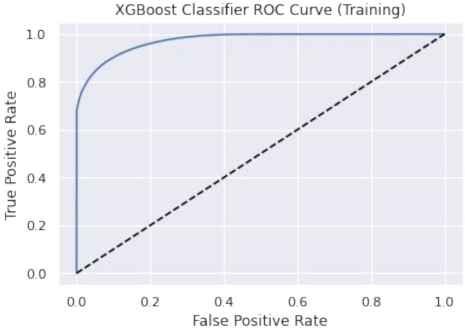
XGBoost Classifier

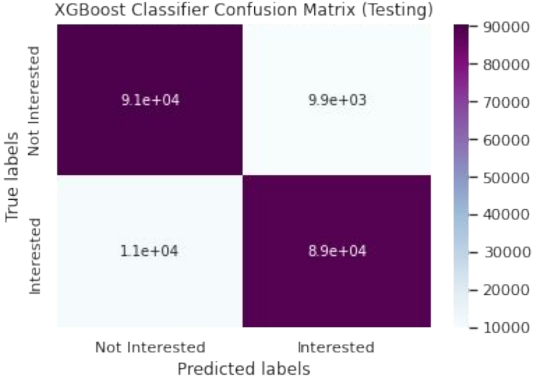
Testing Data Performance
Accuracy : 0.90
Precision: 0.90
Recall: 0.89
F1-Score: 0.89
Area Under the ROC Curve: 0.97

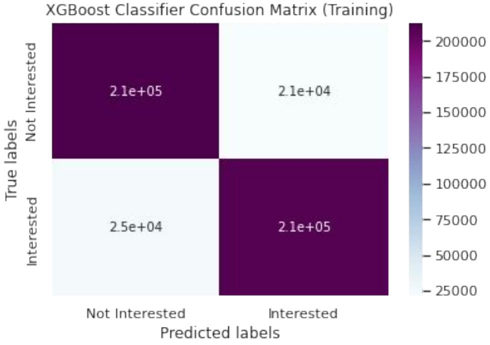
Training Data Performance
Accuracy : 0.90
Precision: 0.90
Recall: 0.89
F1-Score: 0.90
Area Under the ROC Curve: 0.97

XGBoost Classifier


XGBoost Classifier ROC Curve (Testing)


XGBoost Classifier ROC Curve (Training)

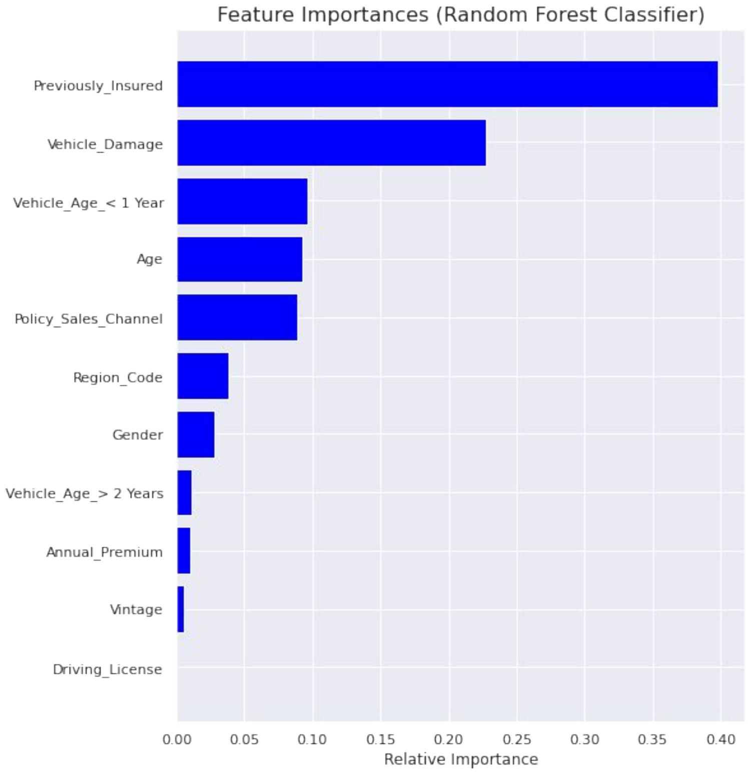

XGBoost Classifier Confusion Matrix (Testing)


XGBoost Classifier Confusion Matrix (Training)

Model Comparison

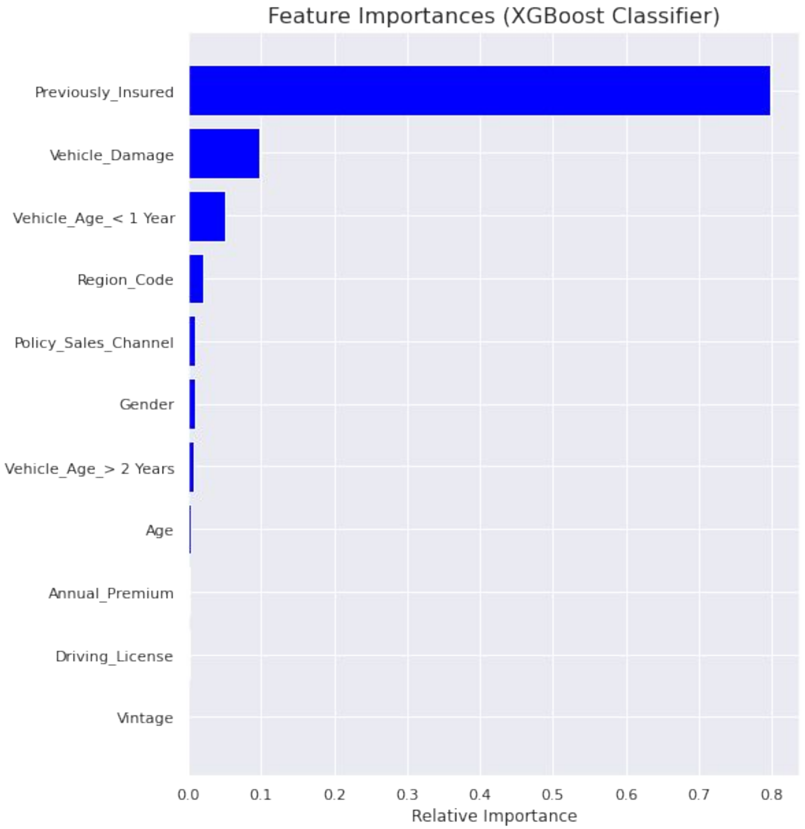
Model	Accuracy	Recall	Precision	F1 score	ROC AUC Score
Logistic Regression	0.81	0.90	0.77	0.82	0.88
Decision Tree	0.84	0.91	0.80	0.85	0.92
Random Forest	0.84	0.93	0.78	0.85	0.92
XGBoost	0.90	0.89	0.90	0.90	0.97

Feature Importances (Random Forest)



Previously Insured followed by Vehicle Damage are the most important features for Random Forest model.

Feature Importances (XGBoost)



Previously Insured followed by Vehicle Damage are the most important features for XGBoost model as well.

We can see differences between both the models in features such as Age and Gender.

Conclusion

- Customers aged between 30 to 50 are more interested in the vehicle insurance compared to the youngsters.
- The chance of customers without Driving License being interested in buying insurance is very low.
- There is very low chance that a person who is previously insured is interested in the insurance.
- Most of the customers pay annual premium below 100,000.
- Number of men with driving license is higher than women in the data. This further results in the number of interested customers being higher in men than women.
- Most of the customers whose vehicle wasn't damaged before are not interested in the insurance

Conclusion

- There is no perfect multicollinearity between any of the independent variables.
- XGBoost Classifier stands out slightly with 90% accuracy and area under ROC curve of 0.97.
- It is followed by Random forest and decision tree classifier with almost similar performance (in terms of accuracy and ROC score).
- The logistic regression model is the worst of them all (by a small margin).
- The are some differences in the feature importances of Random Forest model and XGBoost model. But 'Previously Insured' followed by 'Vehicle Damage' are the most important features for both.

Thank You!!