Capstone Project

Health Insurance Cross-Sell Prediction Technical Documentation

Jyoti Patel
Agam Singh
Data science trainees
AlmaBetter

Table of Content:-

Abstract :

The Health Insurance Cross-Sell Prediction project aims to help an insurance company predict whether its existing health insurance policyholders from the past year would also be interested in purchasing vehicle insurance. This predictive analysis will enable the company to target potential customers more effectively, optimize communication strategies, and enhance revenue generation. The dataset contains valuable information on customer demographics, vehicle-related details,

and policy information. By employing machine learning algorithms and feature selection techniques, the project seeks to build a robust classification model for accurate predictions.

Introduction:

Insurance serves as a protective measure against potential financial losses, damages, illnesses, or death, with customers paying a specified premium to the insurance company. The primary objective of this project is to develop a model that can predict customer interest in vehicle insurance based on historical health insurance data. Understanding customer behavior and preferences is crucial for any insurance provider to enhance customer satisfaction and improve business operations. By utilizing data analytics and machine learning, the project aims to facilitate data-driven decision-making and customer targeting.

Problem Statement:

Our client is an Insurance company that has provided Health Insurance to its customers. Now they need the help in building a model to predict whether the policyholders (customers) from the 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.

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 Description:

We have a dataset which contains information about demographics (gender, age, region code type), Vehicles (Vehicle Age, Damage), Policy (Premium, sourcing channel) etc. related to a person who is interested in vehicle insurance.The dataset consists of 381109 rows and 12 columns. There are no null values or duplicates present in the dataset.

The columns present in the dataset are:

1.id : Unique ID for the customer

2.Gender : Gender of the customer

3.Age : Age of the customer

4.Driving_License : 0 - Customer does not have DL, 1 - Customer already has DL

5.Region_Code : Unique code for the region of the customer

6.Previously_Insured : 1 : Customer already has Vehicle Insurance, 0 : Customer doesn't have Vehicle Insurance

7.Vehicle_Age : Age of the Vehicle

8.Vehicle_Damage : 1 : Customer got his/her vehicle damaged in the past. 0 : Customer didn't get his/her vehicle damaged in the past.

9.Annual_Premium : The amount customer needs to pay as premium in the year

10.PolicySalesChannel : Anonymized Code for the channel of outreaching to the customer ie. Different Agents, Over Mail, Over Phone, In Person, etc.

11.Vintage : Number of Days, Customer has been associated with the company

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

Steps involved:

1.Data Cleaning and Preprocessing:

The first step involved thorough data cleaning to identify and handle any missing values or duplicates. As the dataset was free of such issues, data normalization was performed to bring numerical features to a consistent scale, ensuring fair comparisons during analysis.

2.Exploratory Data Analysis (EDA):

EDA was carried out to gain insights into the data distribution, correlations, and trends. Age was categorized into different groups to better understand customer demographics. Additionally, features like Region_Code and Policy_Sales_Channel were analyzed to extract meaningful information.

3.Feature Selection:

To identify the most influential features, both numerical and categorical, feature selection techniques like Kendall's rank correlation coefficient and Mutual Information were employed. These techniques helped us identify the key variables that impact customer interest in vehicle insurance.

4.Model Prediction:

The project utilized various supervised machine learning algorithms to build predictive models. Decision Tree Classifier, AdaBoost, LightGBM, Bagging Regressor, Naive Bayes, and Logistic Regression were among the models evaluated.

Hyperparameter tuning was applied to improve the accuracy of the models and mitigate overfitting.

Exploratory Data Analysis:

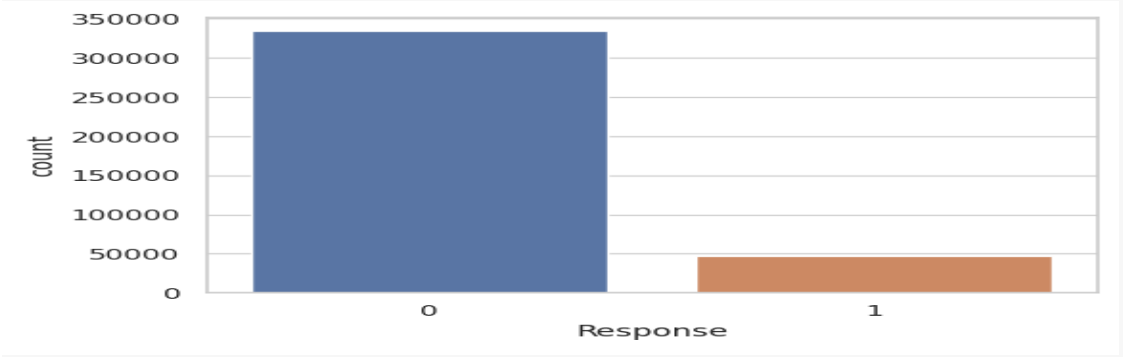
A) Data Cleaning: During the data cleaning phase, we encountered several object values in our dataset. To address this, we performed a conversion by mapping these object values into integer format. Specifically, the columns with the names "Gender," "Vehicle_Age," and "Vehicle_Damage" were transformed from their original object datatype into integer representations. This conversion allowed us to handle the data more effectively and prepare it for further analysis and modeling.

B) Null values Treatment: Our dataset does not contain null values which tend to affect our accuracy. If we had null values, we could drop them or input them with mean or median depending on the situation.

C) Data Visualization:

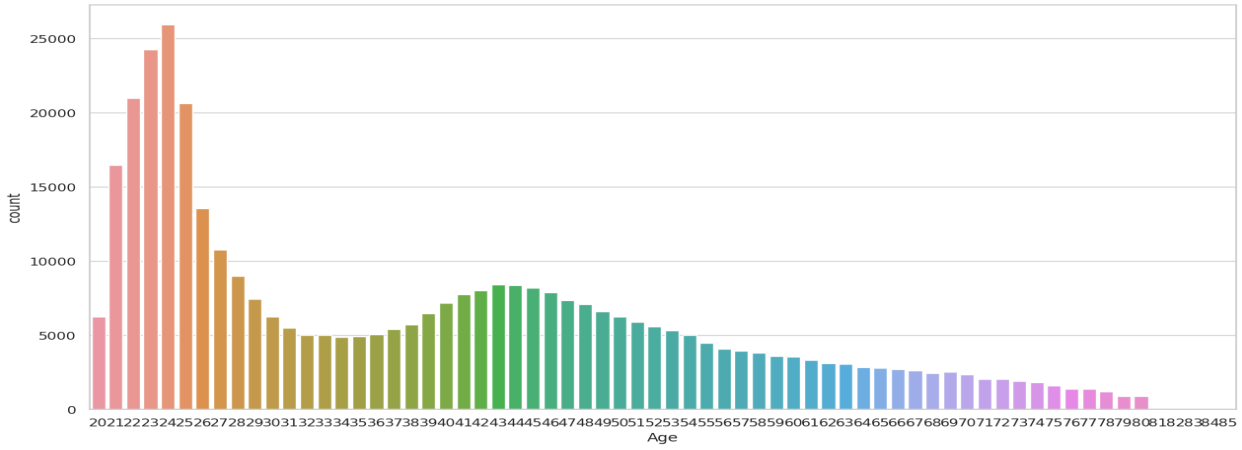
1. Univariate Analysis:

➢ Dependent variable 'Response'



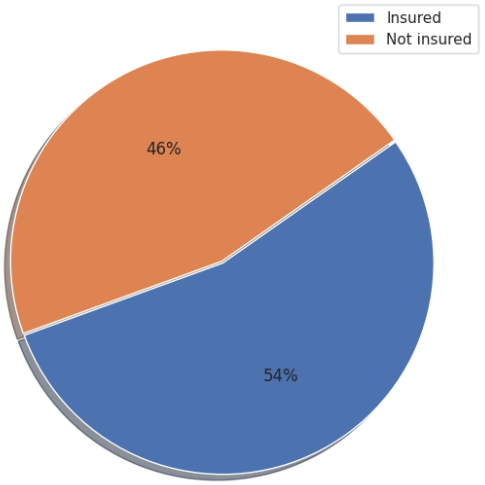
- From above fig we can see that the data is highly imbalanced.

➢ Distribution of Age



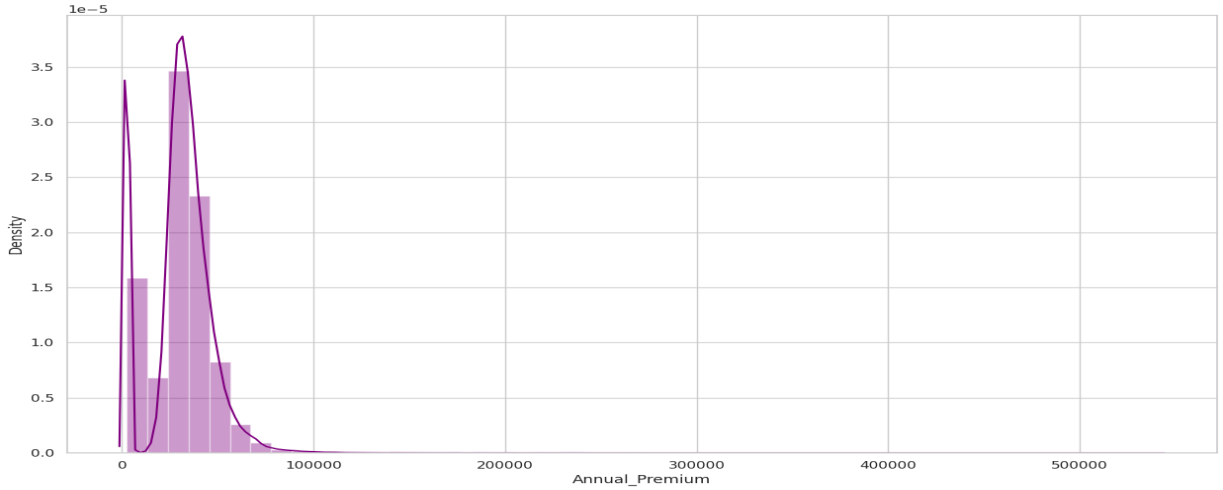
- From the above distribution of age we can see that most of the customers age is between 21 to 25 years.There are few Customers above the age of 60 years.

➢ Previously_Insured('Insured','Not insured')

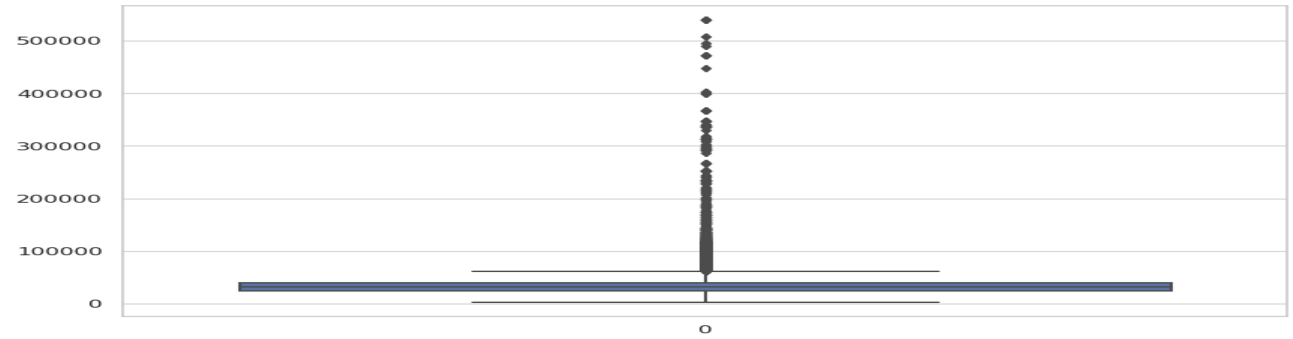


- 54% customer are previously insured ahe 46% customer are are not insured yet. Customer who are not perviosly insured are likely to be inetrested.

➢ Annual_Premium Density

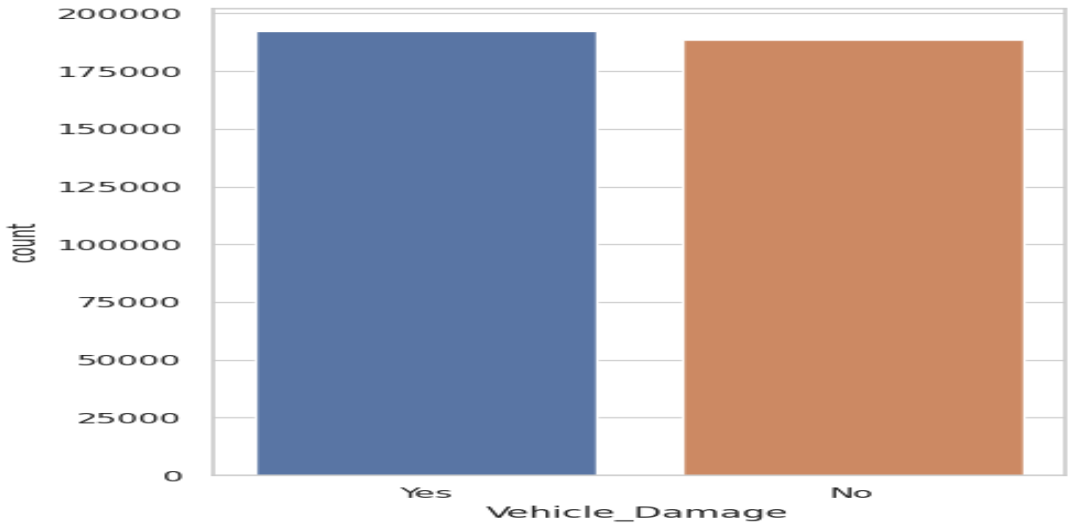


- From the distribution plot we can infer that the annual premimum variable is right skewed.



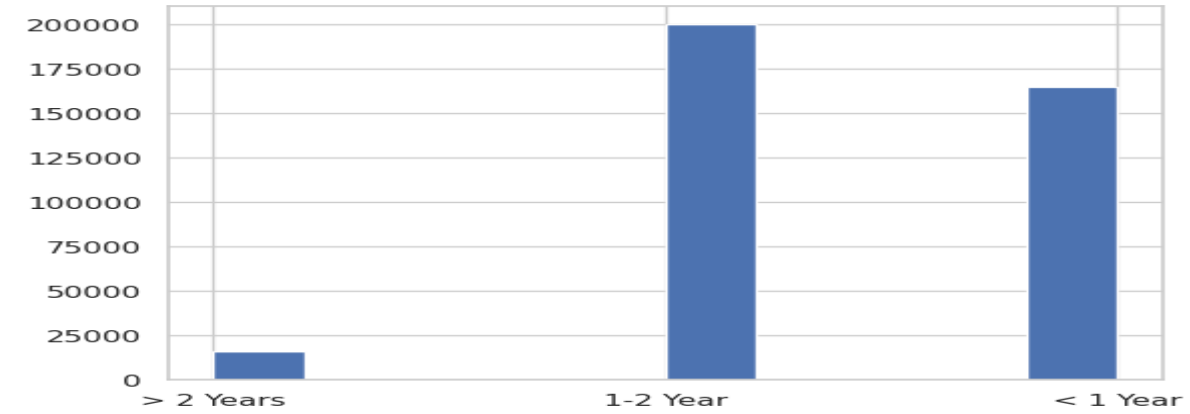
- For the boxplot above we can see that there's a lot of outliers in the annual premium.

➢ Vehicle_Damage count



- Customers with Vehicle_Damage are likely to buy insurance

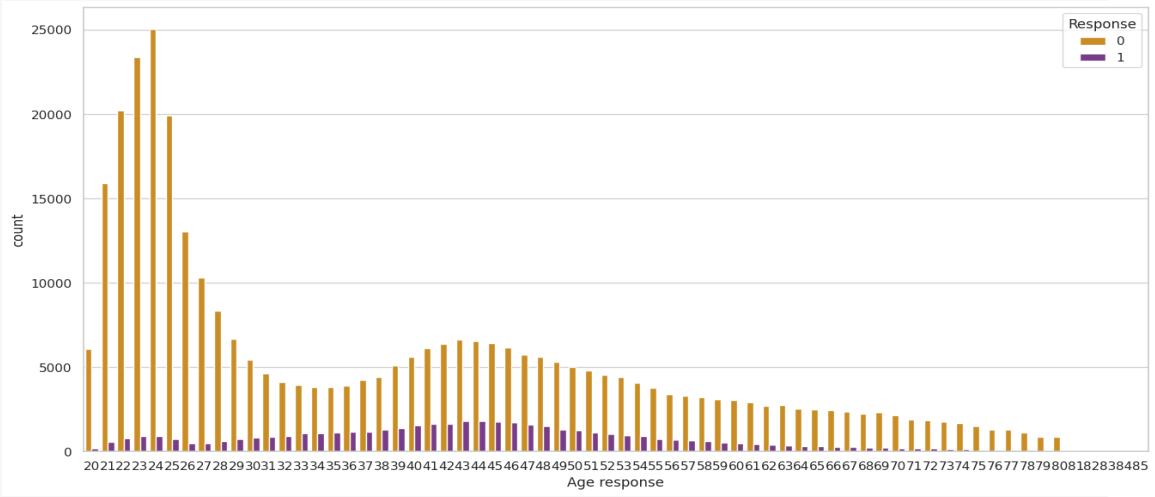
➢ Vehicle_Age



- From the above plot we can see that most of the people are having vehicle age between 1 or 2 years and very few people are having vehicle age more than 2 years.

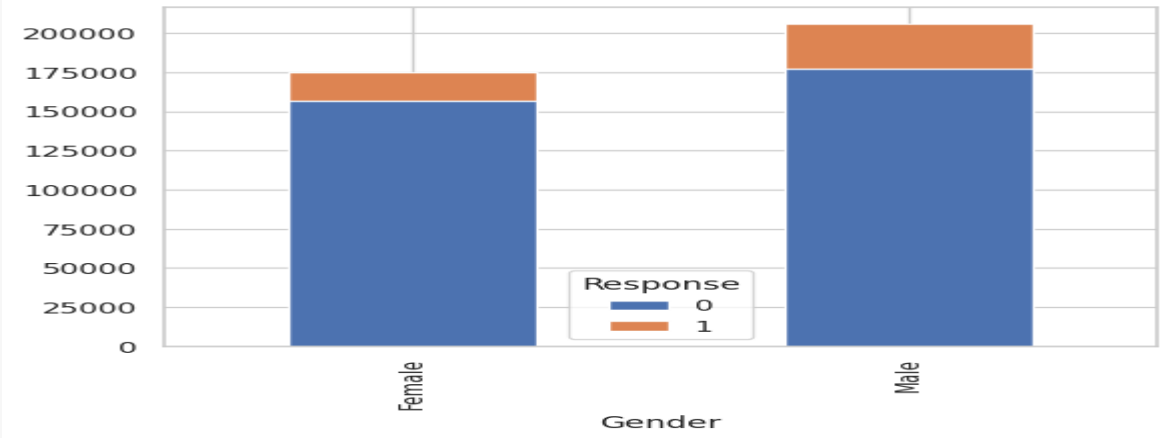
2.Bivariate analysis:

➢ Age VS Response



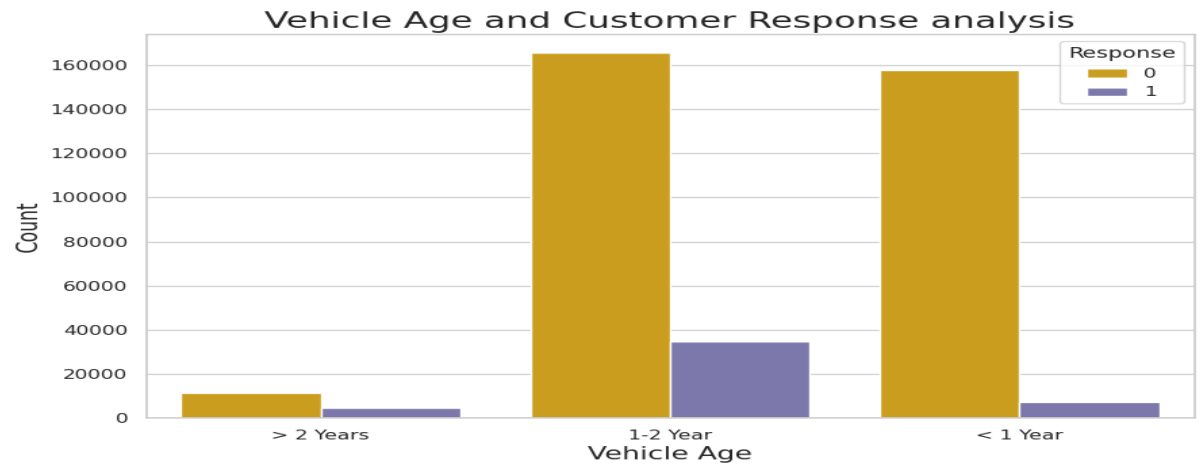
- People ages between from 31 to 50 are more likely to respond.
- while Young people below 30 are not interested in vehicle insurance.

➢ Gender vs Response



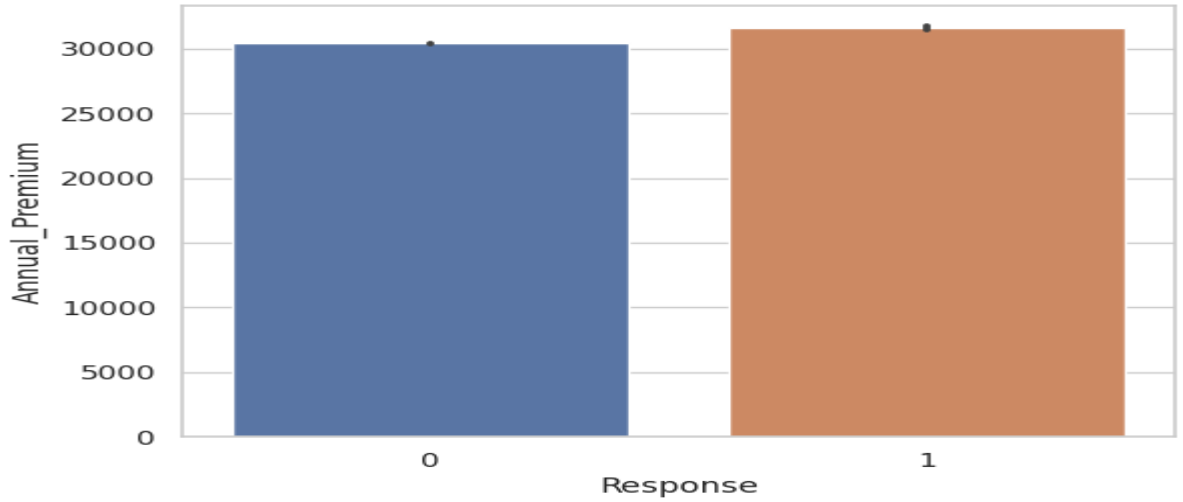
- Male category is slightly greater than that of female and chances of buying the insurance is also little high

➢ Vehicle Age and Customer Response analysis



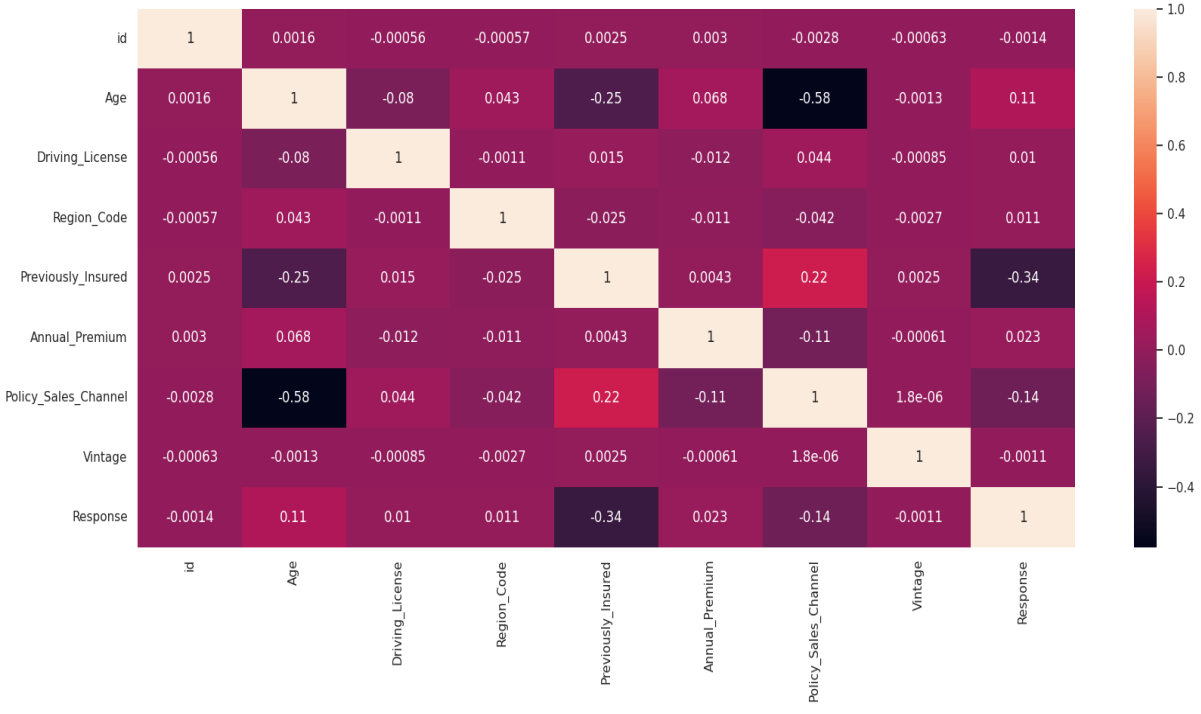
- Customers with vechicle age 1-2 years are more likely to interested as compared to the other two
- Customers with with Vehicle_Age <1 years have very less chance of buying Insurance

➢ Annual_Premium Vs Response analysis



- People who response have slightly higher annual premium

➢ Correlation Analysis



- Target variable is not much affected by Vintage variable. we can drop least correlated variable.

Encoding categorical values

We used one-hot encoding for converting the categorical columns such as 'Gender', 'Previously_Insured','Vehicle_Age','Vehicle_Damage', 'Age_Group', 'Policy_Sales_Channel_Categorical', 'Region_Code_Categorical' into numerical values so that our model can understand and extract valuable information from these columns.

Feature Selection

At first, we obtained the correlation between numeric features through Kendall's Rank Correlation to understand their relation. We had two numerical features, i.e. Annual_Premium and Vintage. For categorical features, we tried to see the feature importance through Mutual Information. It measures how much one random variable tells us about another.

Model Fitting

For modeling, we tried the various classification algorithms like

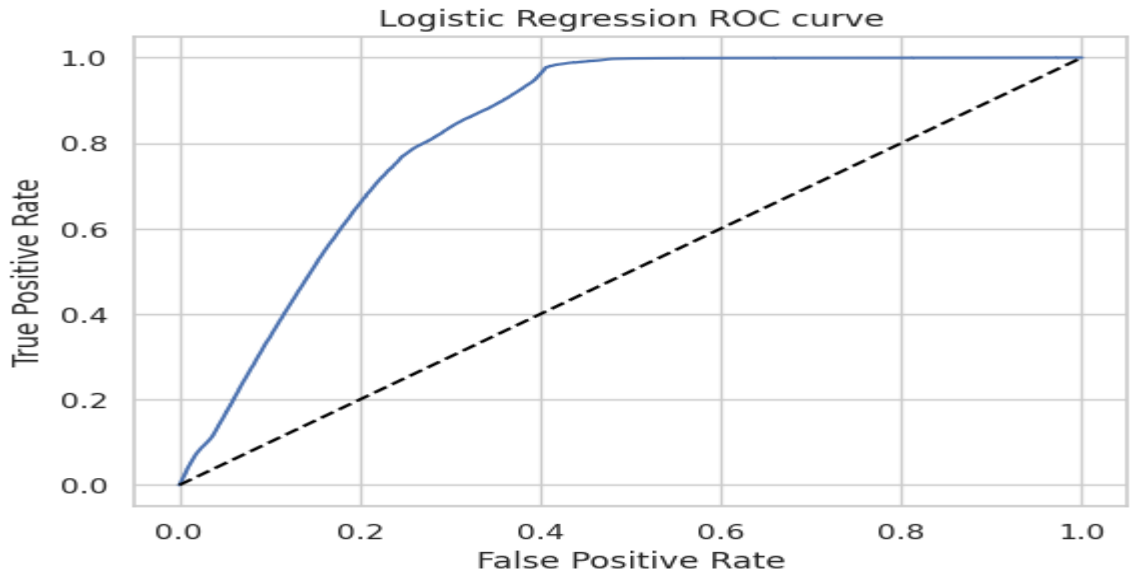
1.Logistic Regression-

Decision Trees are non-parametric supervised learning methods, capable of finding complex non-linear relationships in the data. Decision trees are a type of algorithm that uses a tree-like system of conditional control statements to create the machine learning model. A decision tree observes features of an object and trains a model in the structure of a tree to predict data in the future to produce output. For classification trees, it is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. Logistic regression is not performing well on this dataset as in confusion matrix model is predicting positive responses but with positive responses it is predicting negative responses in high numbers too.

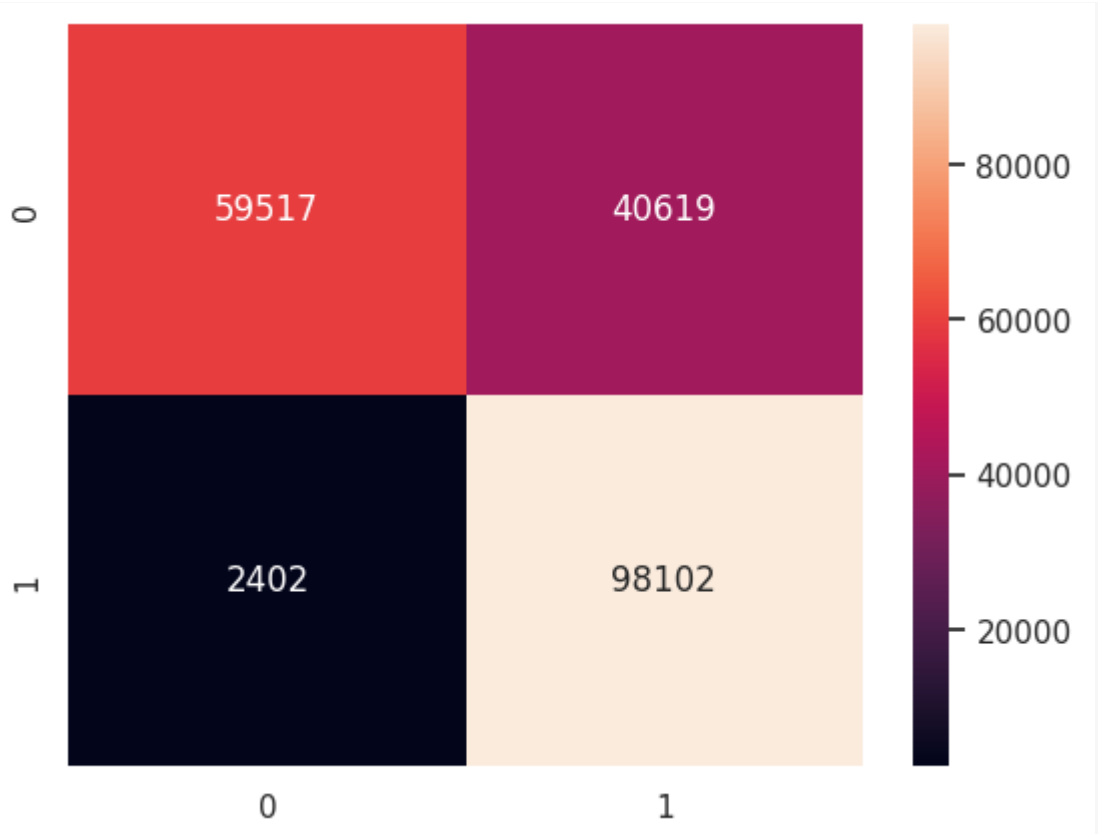
❖ Model Evaluation

recall_score : 0.976100453713285
precision_score : 0.7071892503658422
f1_score : 0.8201651165221027
accuracy_score : 0.7855811403508772
ROC_AUC Score: 0.8341983171030103

❖ Logistic Regression ROC curve



❖ confusion_matrix

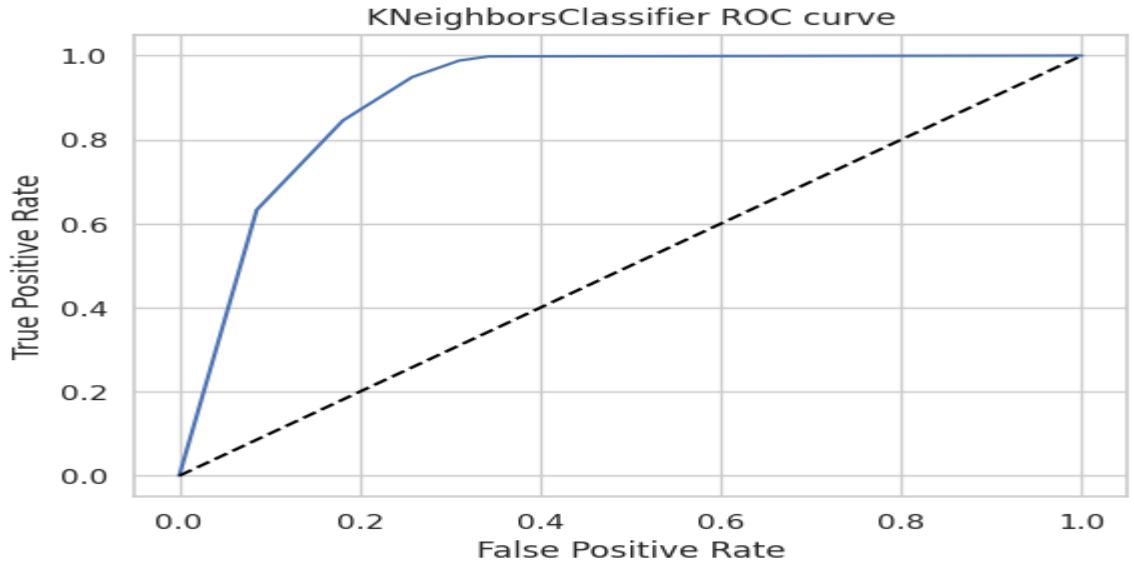


2. KNeighborsClassifier

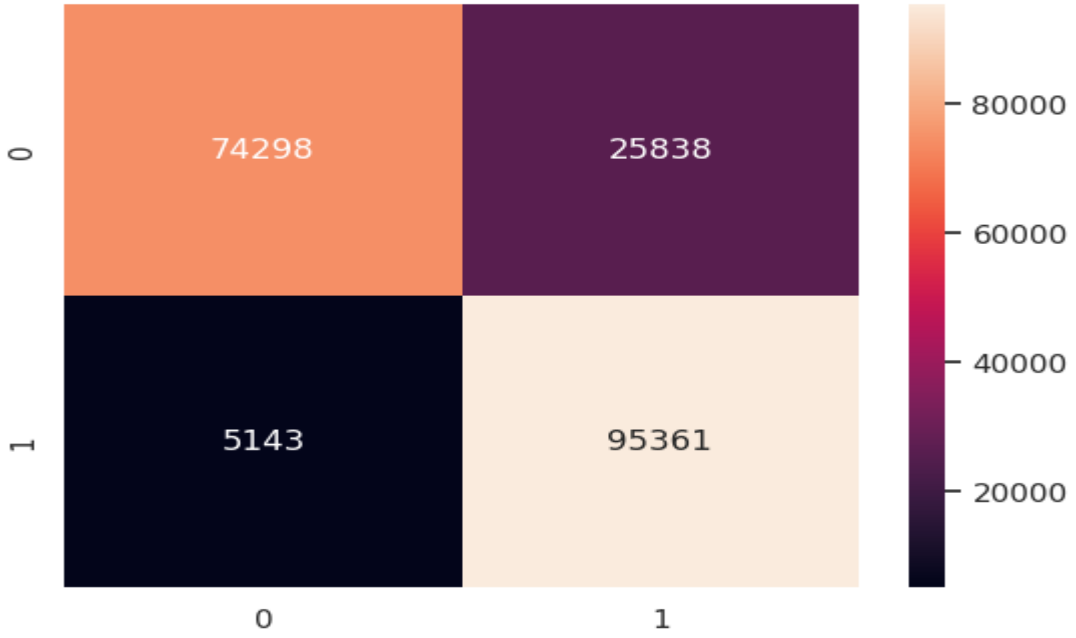
❖ Model Evaluation

recall_score : 0.9488279073469713
precision_score : 0.7868134225529914
f1_score : 0.8602589951421497
accuracy_score : 0.8455891148325358
ROC_AUC Score: 0.8610367763562403

❖ KNeighborsClassifier ROC curve



❖ confusion_matrix



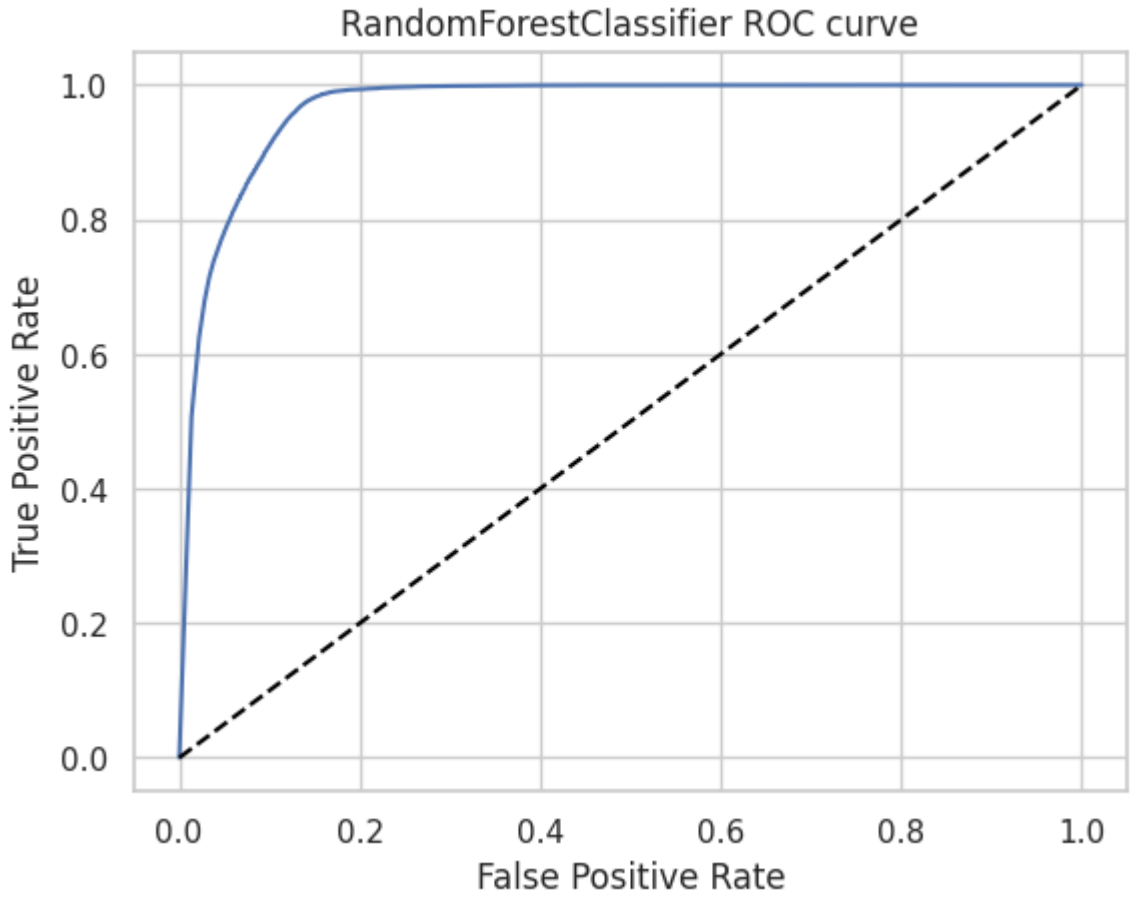
3. RandomForest Classifier-

A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is controlled with the Parameter if bootsrap=True otherwise whole data set build in each tree. Here , random forest is performing better as in the confusion matrix the model now is much better with predicting positive responses.

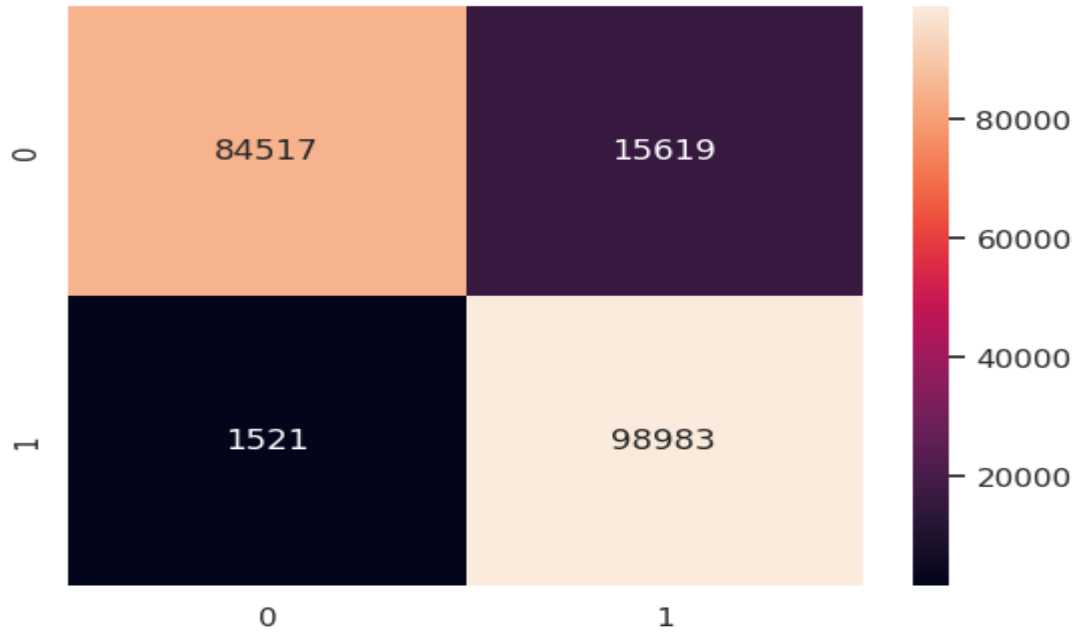
- Model Evaluation

recall_score : 0.9848662739791451
precision_score : 0.8637109300012216
f1_score : 0.9203183546716502
accuracy_score : 0.91457336523126
ROC_AUC Score: 0.9230163474014105

- RandomForestClassifier ROC curve



❖ confusion_matrix



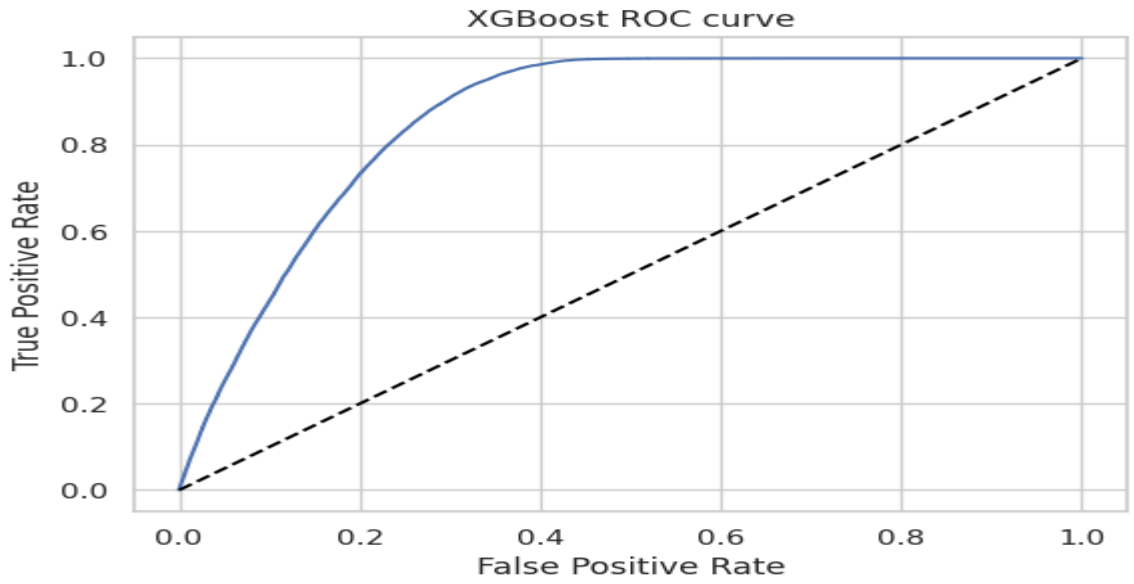
3. XGBoost-

- XGBoost comes under boosting and is known as extra gradient boosting.
- GBM first calculates the model using X and Y then after the prediction is obtain.
- It will again calculates the model based on residual of previous model
- loss function will give more weightage to error of previous model. and this process continuous until MSE gets minimizes.
- From the confusion matrix we see that the model is a bit better with predicting positive responses.

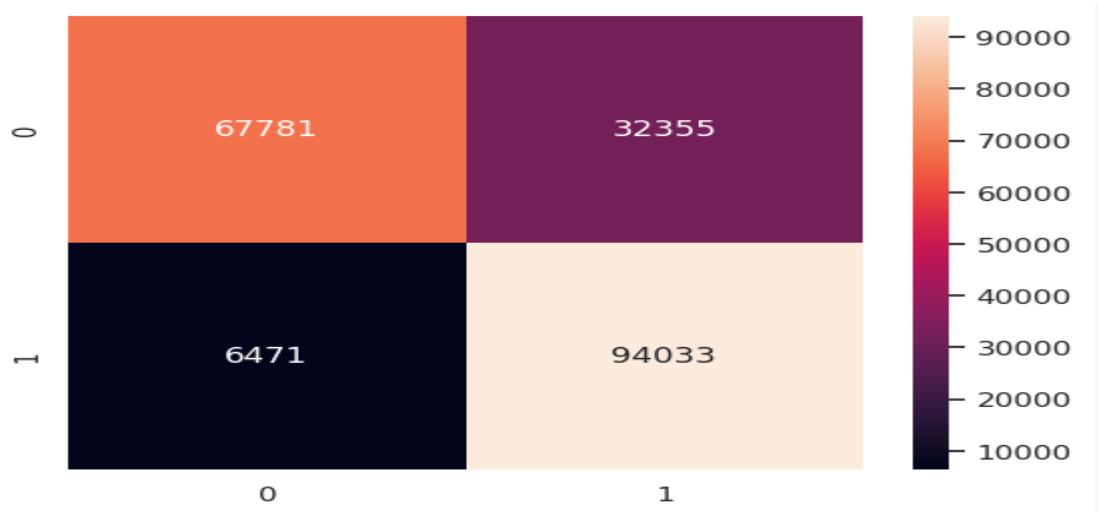
❖ Model Evaluation

recall_score : 0.935614502905357
precision_score : 0.744002595183087
f1_score : 0.8288789379969324
accuracy_score : 0.8064892344497607
ROC_AUC Score: 0.8284267137419502

❖ XGBoost ROC curve



❖ confusion_matrix



Comparing the Model

	Accuracy	Recall	Precision	f1_score	ROC_AUC
Logistic Regression	0.785581	0.976100	0.707189	0.820165	0.834198
KNeighbors	0.785581	0.976100	0.707189	0.820165	0.834198
Randomforest	0.914573	0.984866	0.863711	0.920318	0.923016
XGBClassifier	0.806489	0.935615	0.744003	0.828879	0.828427

From the above results ,we can see that the Random Forest model has the highest

accuracy with (91%),

recall(98%),

precision(86%),

f1 score(92%),

and ROC-AUC (92%)

among the four models.

The Random Forest model outperforms all other model so random forest would be our the best choice.

Conclusion

Throughout the course of this project, we embarked on a data-driven journey to predict customer interest in vehicle insurance among existing health insurance policyholders. We started by meticulously loading and preprocessing our dataset, ensuring the elimination of any null values and duplicates.

During Exploratory Data Analysis (EDA), we discovered intriguing patterns, revealing that customers belonging to the "YoungAge" group exhibited higher interest in vehicle insurance. Conversely, individuals below 30 years of age showed less interest in such insurance. Furthermore, customers with vehicles older than 2 years and those with damaged vehicles displayed a stronger inclination towards vehicle insurance. Among the various features, we found that "Age," "Previously_Insured," and "Annual_premium" significantly influenced the target variable.

To select the most relevant features for our predictive models, we employed the Mutual Information technique, identifying "Previously_Insured" as the most impactful feature without any correlation to the dependent variable. Additionally, we

encountered a class imbalance in the target variable and effectively addressed it using the Random Over Sample resampling technique.

To enhance the models' performance, we standardized the data through feature scaling, allowing the machine learning algorithms to operate seamlessly across all features.

For our predictive modeling, we evaluated multiple algorithms, such as Logistic Regression, RandomForest, XGBClassifier, and KNeighbors.

The Random Forest model demonstrated an impressive accuracy of 91%, signifying its ability to accurately classify customers interested in vehicle insurance. Additionally, it achieved a remarkable recall of 98%, indicating its proficiency in identifying true positive instances, a crucial aspect in correctly identifying potential customers. The model's precision of 86% underscores its capability to minimize false positive classifications, ensuring targeted and relevant communication strategies. Moreover, the F1 score of 92%, representing a harmonious balance between precision and recall, highlights the model's overall efficacy in handling imbalanced classes.Moreover, the model achieved an outstanding ROC-AUC score of 92%, a vital metric that measures the model's ability to discriminate between positive and negative instances. This high score signifies the model's strong capability in distinguishing interested customers from non-interested ones.

In light of these outcomes, we confidently conclude that the RandomForest model is the most suitable choice for this classification task. Its high accuracy and robust ROC_AUC score indicate its efficacy in identifying potential customers interested in vehicle insurance among existing health insurance policyholders. As the insurance company endeavors to optimize its communication strategies and revenue generation, the RandomForest model will be instrumental in driving more personalized marketing efforts, enhancing customer satisfaction, and ensuring continued business success. By leveraging data insights and advanced machine learning techniques, the insurance company is poised to stay ahead in the competitive market landscape, cementing its position as a customer-centric and reliable insurance provider.