Capstone Project

Health Insurance Cross-Sell Prediction

TEAM NAME : New World

Team Members

1. Agam Singh 2.Jyoti Patel

About Health Insurance Cross-Sell Prediction

- Health Insurance Cross-Sell Prediction is a data-driven approach that uses machine learning and analytics to predict whether an existing health insurance customer is likely to be interested in purchasing additional insurance products. By analyzing historical customer data, demographics, past purchase behavior, and other relevant variables, the predictive model can identify potential cross-selling opportunities.
- Health Insurance Cross-Sell Prediction employs machine learning and customer data analysis to forecast whether existing health insurance customers are likely to purchase additional insurance products. By identifying potential cross-selling opportunities, insurers can tailor personalized offers, optimize marketing efforts, and enhance customer satisfaction while prioritizing data privacy and security.

CONTENTS

1. Objective
2. Introduction
3. Problem Statement
4. Data Description
5. Exploratory Data Analysis
6. Encoding categorical values
7. Feature Selection
8. Model Fitting

Objective

The main objective of this project is to develop a predictive model that accurately determines customers' interest in Vehicle Insurance. This model will empower the company to strategically plan its communication approach, enabling targeted outreach to potential customers and optimizing the business model and revenue streams effectively.

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 SUMMARY

Name of the Dataset—----------Health Insurance Cross Sell Prediction

Number of variables/Columns —---------12

Number of observations/Row—----------381109

Duplicate rows —----------0 (0.0%)

Total size in memory—---------- 20.4 MB

VARIABLE DATA TYPE

Data Type	Columns
Numeric – int64	**0 id** **2 Age** **3 Driving_License** **5 Previously_Insured** **10 Vintage** **11 Response**
Numeric – float64	**4 Region_Code** **8 Annual_Premium** **9 Policy_Sales_Channel**
String - object	**1 Gender** **6 Vehicle_Age** **7 Vehicle_Damage**

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

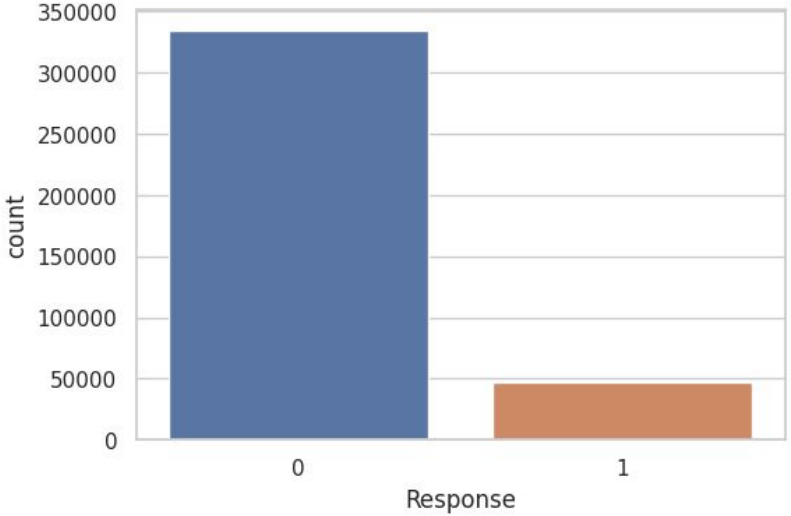
11.Vintage : Number of Days, Customer has been associated with the company

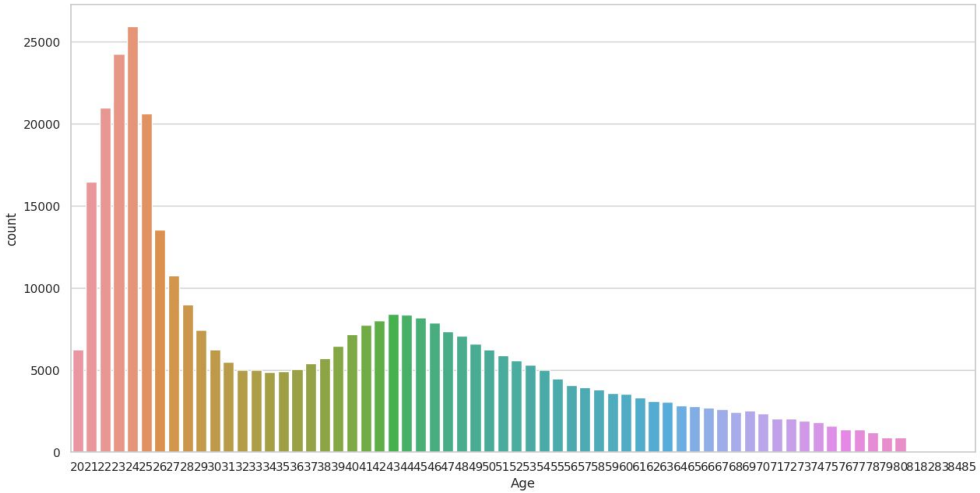
12.Response : 1 : Customer is interested, 0 : Customer is not interested

STEPS INVOLVED

1) Exploratory Data Analysis
2) Null values Treatment
3) Data Exploration
4) Data Visualization
5) Standardization of features
6) Modeling
7) Conclusion

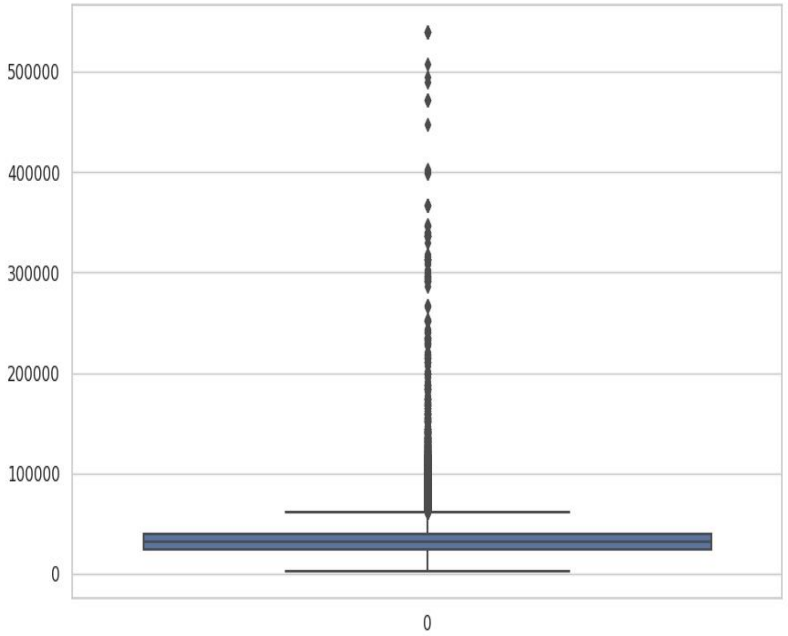
Univariate Analysis

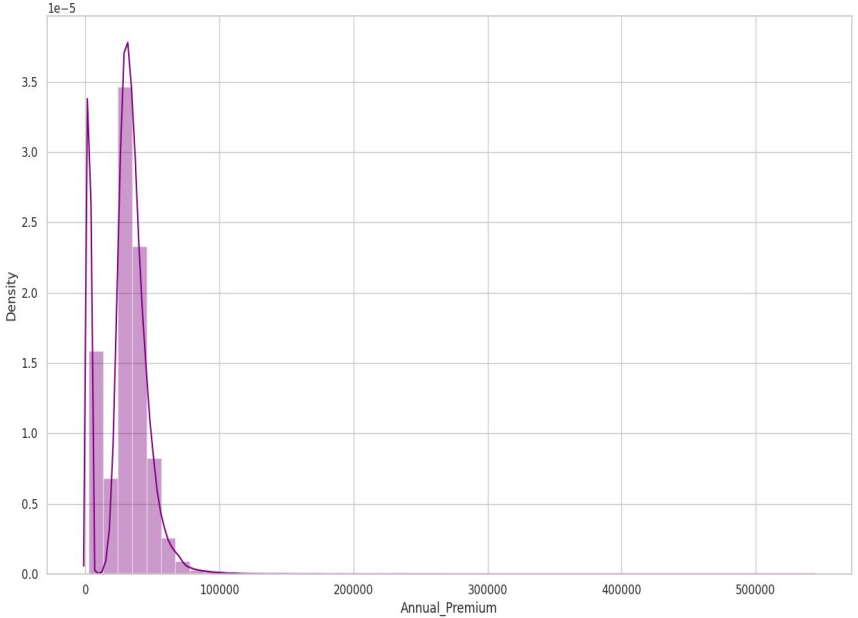




From above fig of response we can see that the data is highly imbalanced. From the above fig of distribution of age we can see that most of the customers age is between 21 to 25 years.There are few Customers above the age of 60 years.

Univariate Analysis

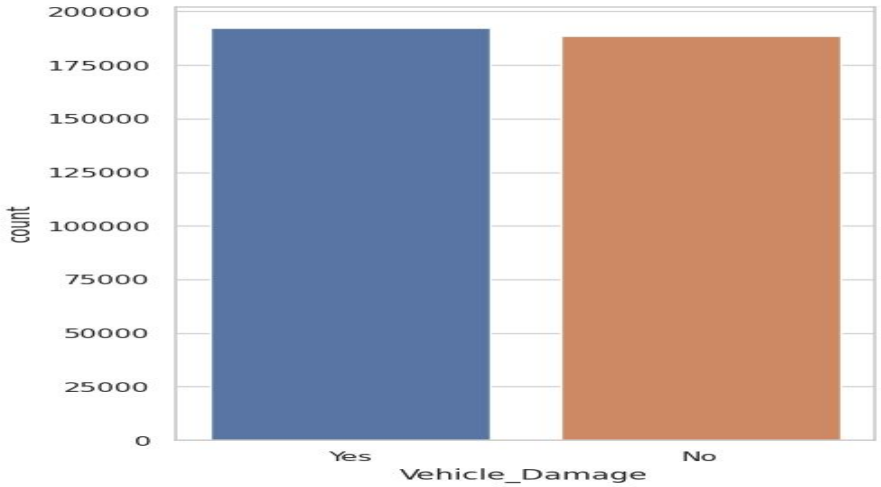


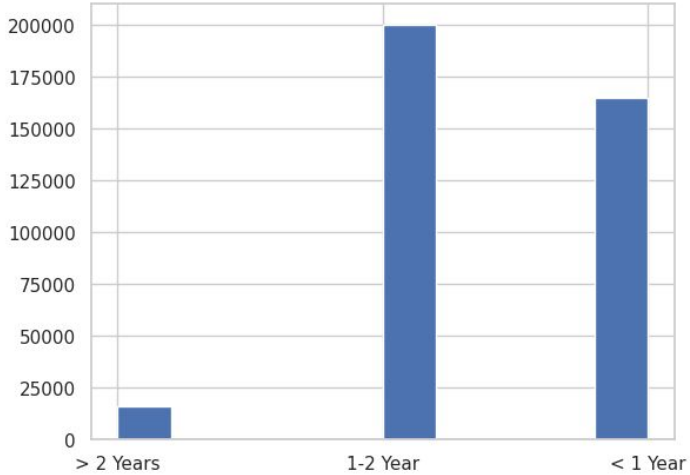


- For the boxplot above we can see that there's a lot of outliers in the

•From the distribution plot we can infer that the annual

Univariate Analysis

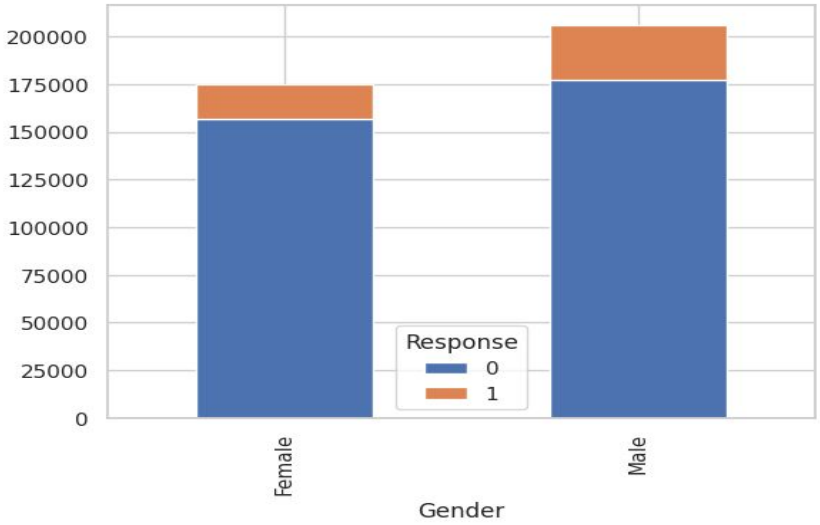


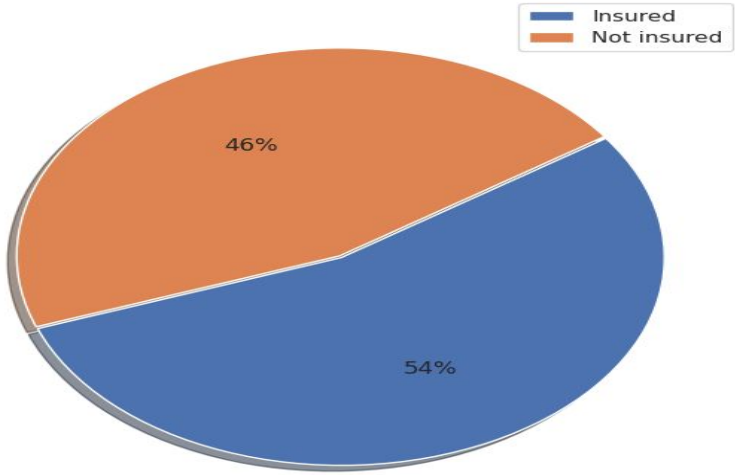


- Customers with Vehicle_Damage are likely to buy insurance

- 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.

Data analysis

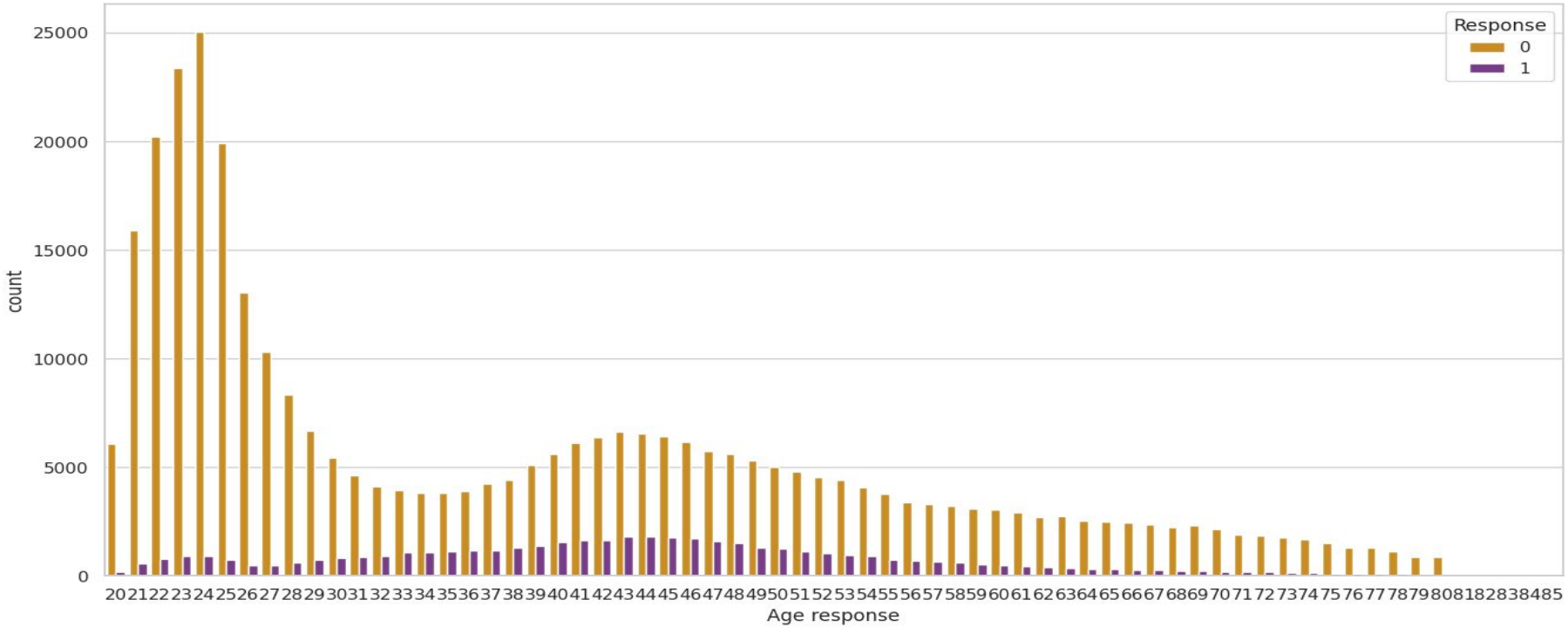




• Male category is slightly greater than that of female and chances of buying the insurance is also little high.

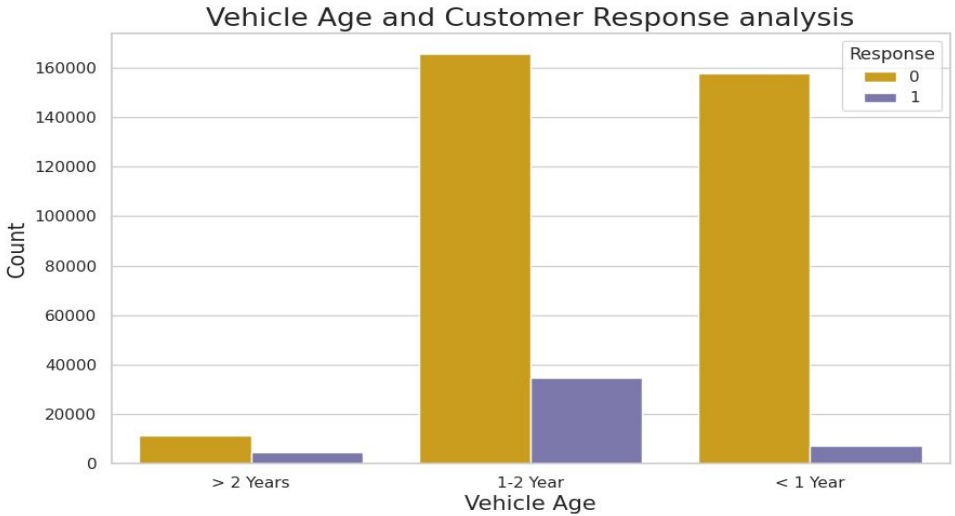
• 54% customer are previously insured the 46% customer are are not insured yet.
• Customer who are not previously insured are likely to be interested.

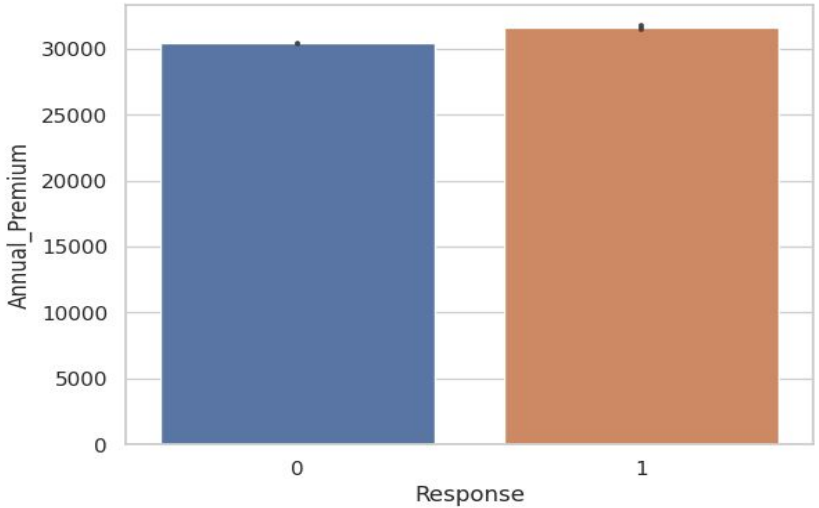
Bivariate analysis



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

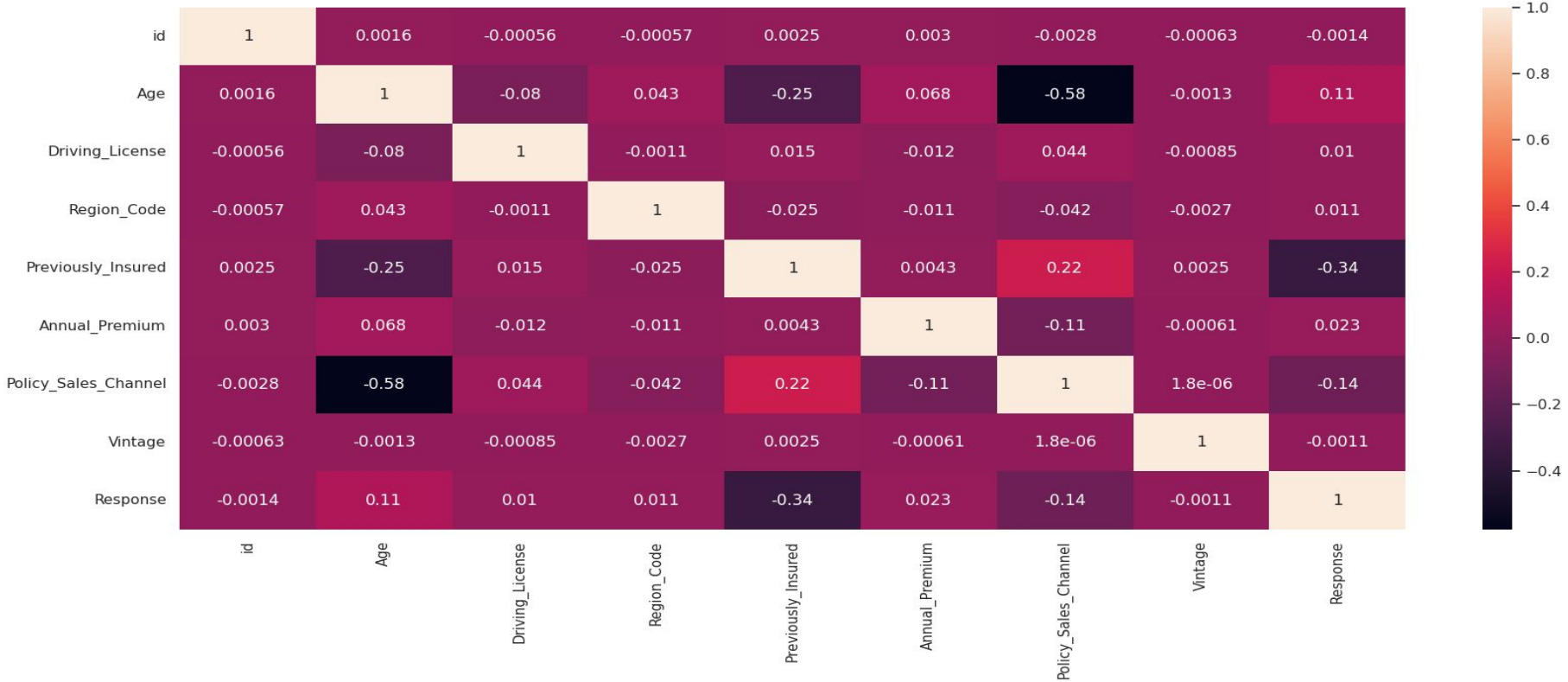
Bivariate analysis





- Customers with vehicle age 1-2 years are more likely to interested as compared to the other two
- Customers with Vehicle_Age <1 years have very less chance of buying Insurance
- People who response have slightly higher annual premium

Correlation



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

Model Building

From the above data analysis we clearly saw that there is a huge difference between the data set.

Standard ML techniques such as Decision Tree and Logistic Regression have a bias towards the majority class, and they tend to ignore the minority class. So solving this issue we use Random Over Sampling technique.

After Random Over Sampling Of Minor Class Total Samples are : 668798

Original dataset shape Counter({0: 334399, 1: 46710})

Resampled dataset shape Counter({1: 334399, 0: 334399})

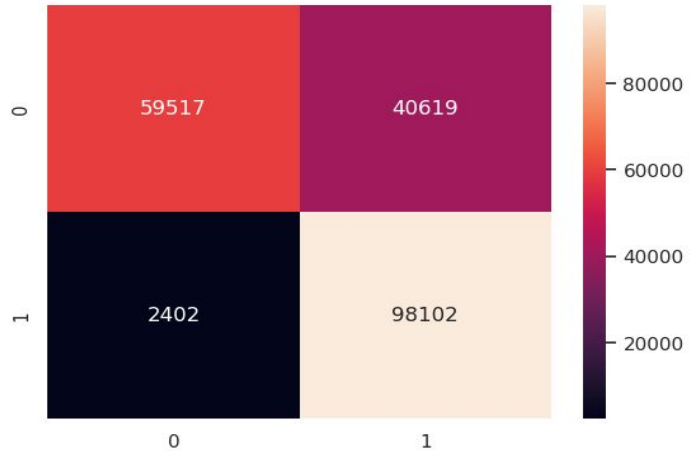
For modeling, we tried the various classification algorithms like:

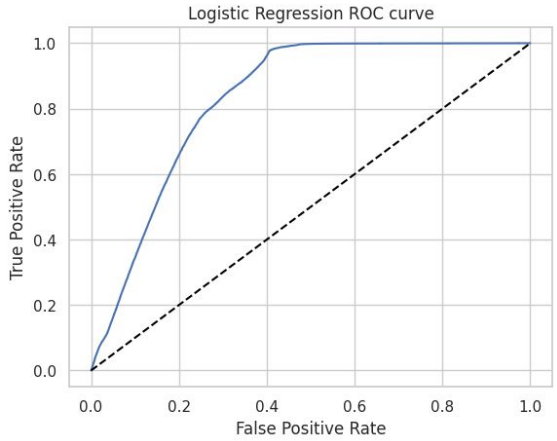
- **Logistic Regression**
- **KNeighborsClassifier**
- **Random Forest Classifier**
- **XGBoost**

Logistic Regression

The Logistic Regression model displays a moderate performance on the dataset, achieving an accuracy of 78.56%. It shows a relatively high recall score of 97.61%, indicating its capability to correctly identify positive responses. However, the precision score of 70.72% and f1 score of 82.02% suggest a trade-off between correctness and recall. The model's ROC_AUC score of 83.42% indicates its ability to distinguish between positive and negative samples. While the Logistic Regression model exhibits certain strengths, further improvements may be needed to enhance its precision without compromising recall for better overall performance.

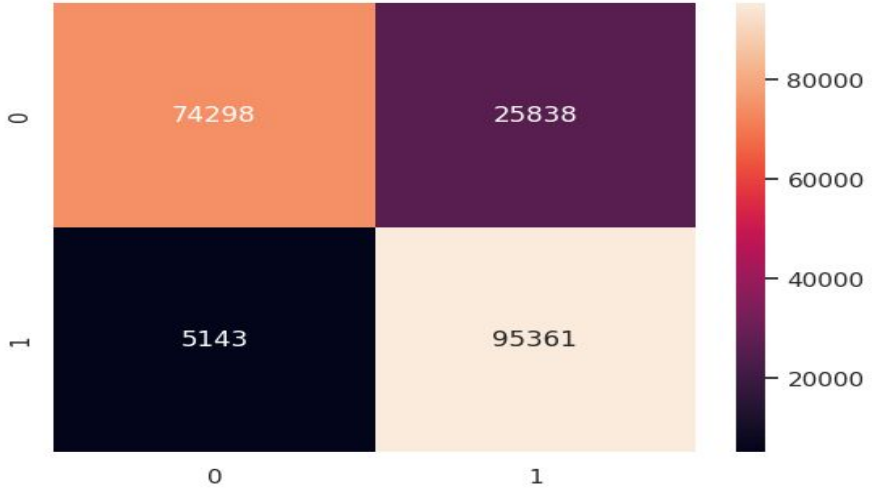
The logistic function, also called the sigmoid function, was developed by statisticians to describe properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment. It's an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits.

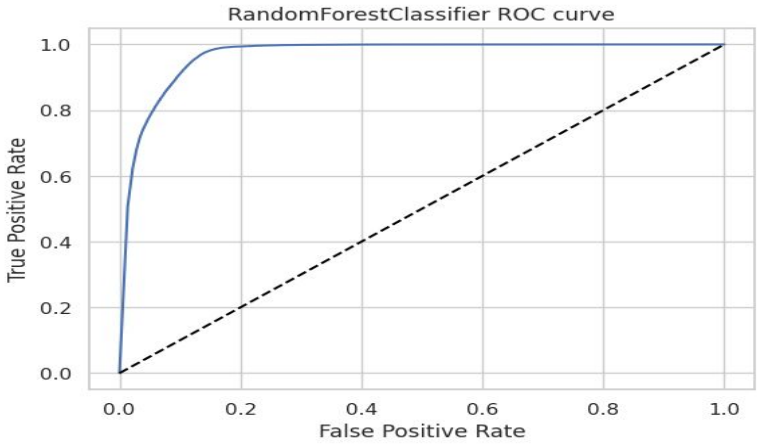




KNeighborsClassifier

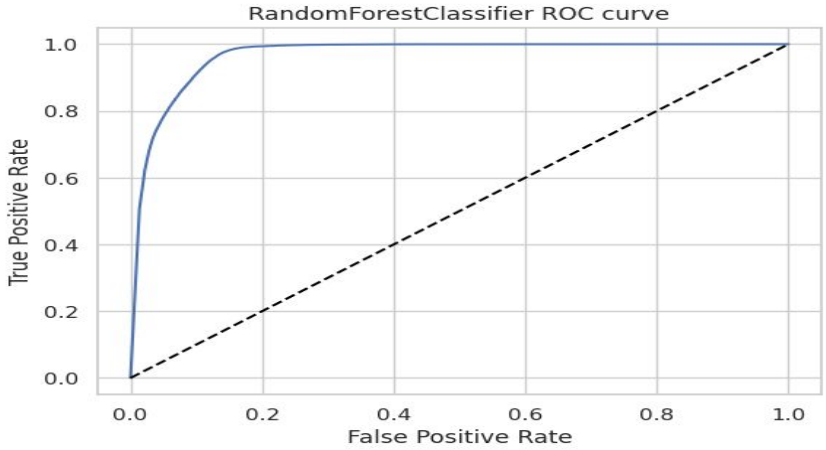
- The K-Nearest Neighbors (KNN) model demonstrates promising performance on the dataset, achieving an accuracy of 84.56%. With a high recall score of 94.88%, the model effectively identifies positive responses. The precision score of 78.68% and f1 score of 0.8603 reflect a reasonable balance between correctness and recall. Additionally, the ROC_AUC score of 0.8610 indicates the model's capability to distinguish between positive and negative samples. Overall, the KNN model shows potential, and further fine-tuning could optimize its performance for better results.





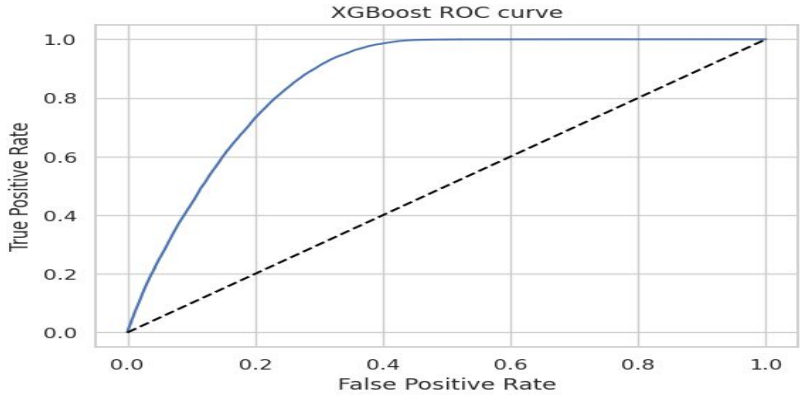
Random Forest Classifier

- The Random Forest model demonstrates excellent performance on the dataset, achieving an accuracy of 91.46%. It excels in correctly identifying positive responses, as indicated by an impressive recall score of 98.49%. The precision score of 86.37% and f1 score of 92.03% highlight a good balance between correctness and recall. Moreover, the model's ROC_AUC score of 92.30% signifies its strong ability to distinguish between positive and negative samples. Overall, the Random Forest model showcases exceptional potential and effectiveness, making it a reliable choice for this classification task.



XGBoost

The XGBoost model demonstrates a commendable performance on the dataset, achieving an accuracy of 80.65%. It exhibits a relatively high recall score of 93.56%, indicating its effectiveness in correctly identifying positive responses. The precision score of 74.40% and f1 score of 82.89% strike a good balance between correctness and recall. The model's ROC_AUC score of 82.84% highlights its ability to distinguish between positive and negative samples. Overall, the XGBoost model showcases strong potential and reliability, making it a suitable choice for this classification task. Further fine-tuning could potentially optimize its performance for even better results.



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 this project, we used data to predict customer interest in vehicle insurance among existing health insurance policyholders. We analyzed the dataset, identified significant patterns, and selected essential features. Among the various algorithms, the Random Forest model stood out, achieving an accuracy of 91% and a remarkable recall of 98%. Its precision of 86% and F1 score of 92% demonstrated a balanced approach in classifying potential customers accurately. With an outstanding ROC-AUC score of 92%, the model excelled in distinguishing interested customers from non-interested ones. Based on these results, we confidently recommend the Random Forest model for personalized marketing efforts, improving customer satisfaction, and driving revenue growth. By leveraging data insights and advanced machine learning, the insurance company can stay ahead in the competitive market and solidify its position as a customer-centric and reliable provider.