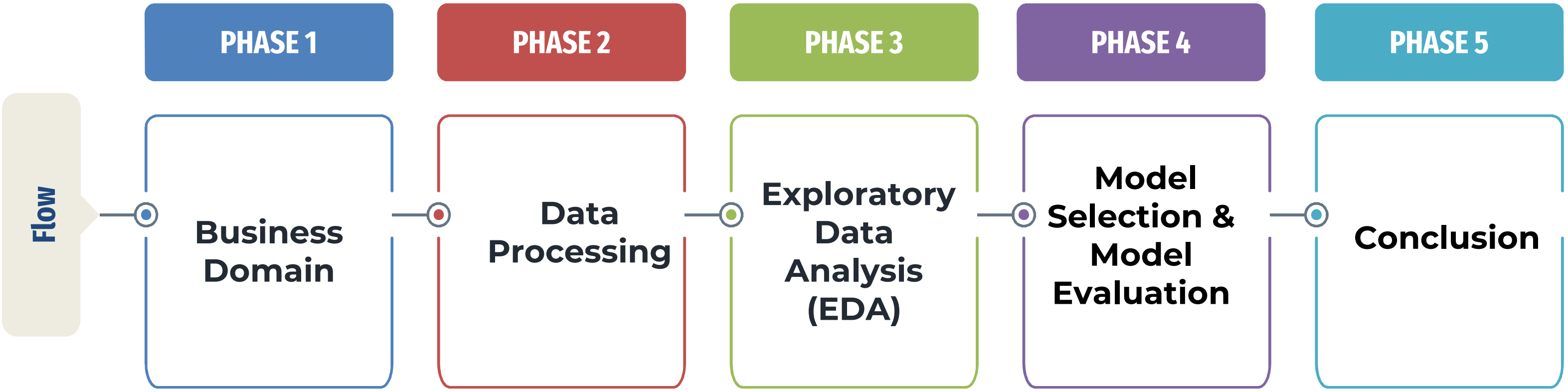
# Capstone Project – 3

## Health Insurance Cross Sell Prediction

**Presented By:**

Debashish Das
Lucky Jain
Vivek katolkar

# Data Methodology :



Flow

PHASE 1

Business Domain

PHASE 2

Data Processing

PHASE 3

Exploratory Data Analysis (EDA)

PHASE 4

Model Selection & Model Evaluation

PHASE 5

Conclusion

# Problem Statement :

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

# Key Level Highlight :

## Objectives

Build a system that can help company to predict potential customer who interested on having vehicle insurance and identify best feature to improve positive responses from customers toward vehicle insurance.

## Business Metrics

- **Response rate**

## Goals

To predict customers who own health insurance will be interested in subscribing to vehicle insurance.

# Overview Our Dataset :

| Variable | Gender | Age | Driving License | Region Code | Previously Insured | Vehicle Age | Vehicle Damage | Annual Premium | Policy Sales Channel | Vintage | Response |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Type | Categorical | Numerical | Categorical | Categorical | Categorical | Categorical | Categorical | Numerical | Categorical | Numerical | Categorical |
| Values | Male<br>Female | 20<br>…<br>85 | Yes<br>No | 0 to 52 | Yes<br>No | < 1 year<br>1-2 years<br>> 2 years | Yes<br>No | 2,630<br>…<br>540,165 | 1<br>…<br>163 | 10<br>…<br>299 | Yes<br>No |
| Description | Gender | Age | Does customer have a driver's license? | Customer region | Does customer have vehicle insurance? | Age of customer's vehicle | Does customer's vehicle have damage? | Health insurance annual premium | Sales channel customer belongs to | How long customer has been associated with company? | Is customer interested in vehicle insurance? |

# Data Cleaning:

**The data set consists of around 381109 entries and 12 columns.**

**Out of this, there were 269 duplicated entries present, which have been dropped.**

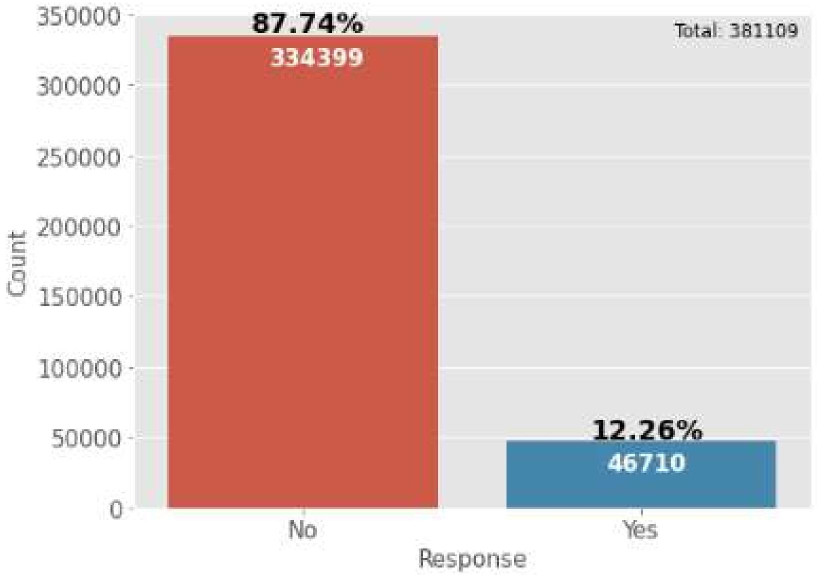
**The dataset doesn't have any feature with Null Values!!!**

# PRE-PROCESSING

1. Data Collection
2. Data Cleansing
3. Feature Engineering
4. Feature Encoding

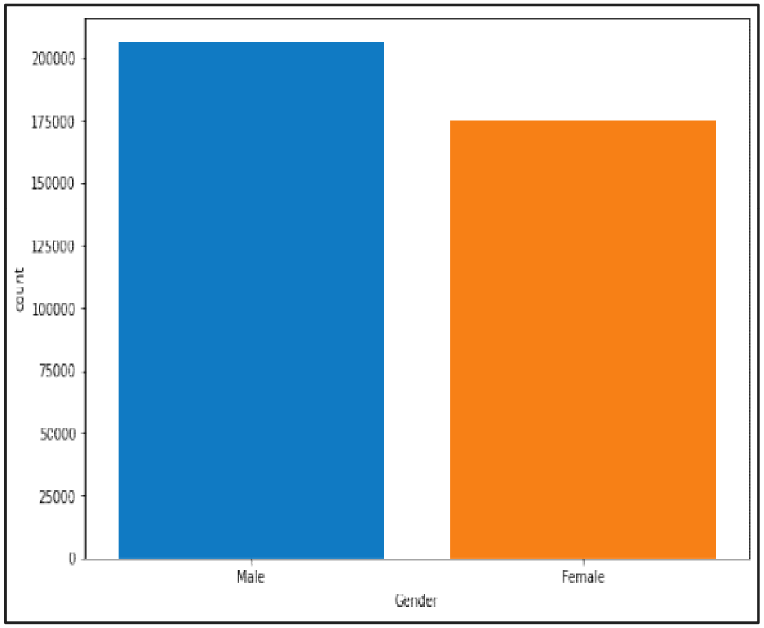
# Exploratory Data Analysis

# Target Variable :

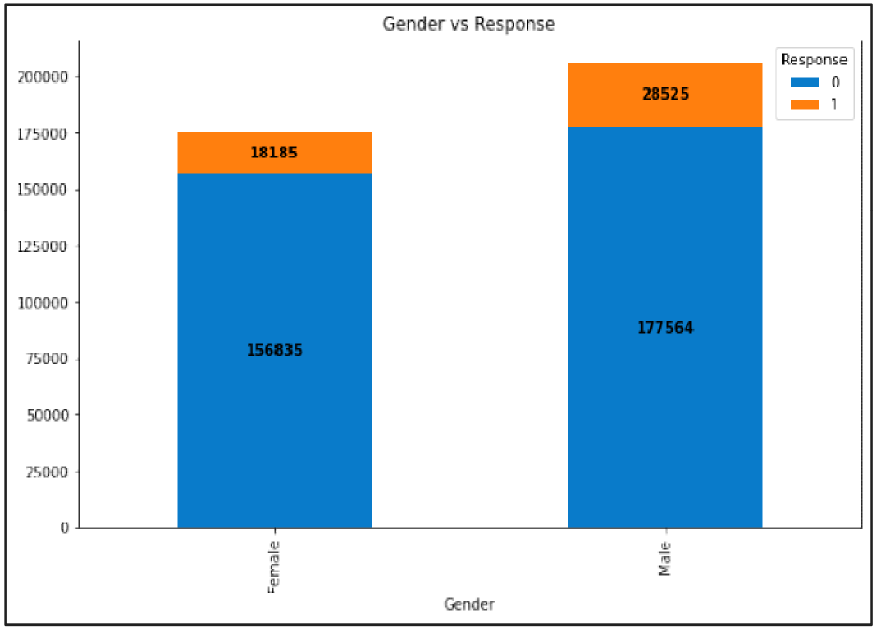


- **Customers who responds = 46710 (12.26%)**
- **Customers who aren’t responds = 334339 (87.74%)**

## The Male Policyholders are slightly more than the Female policyholders.

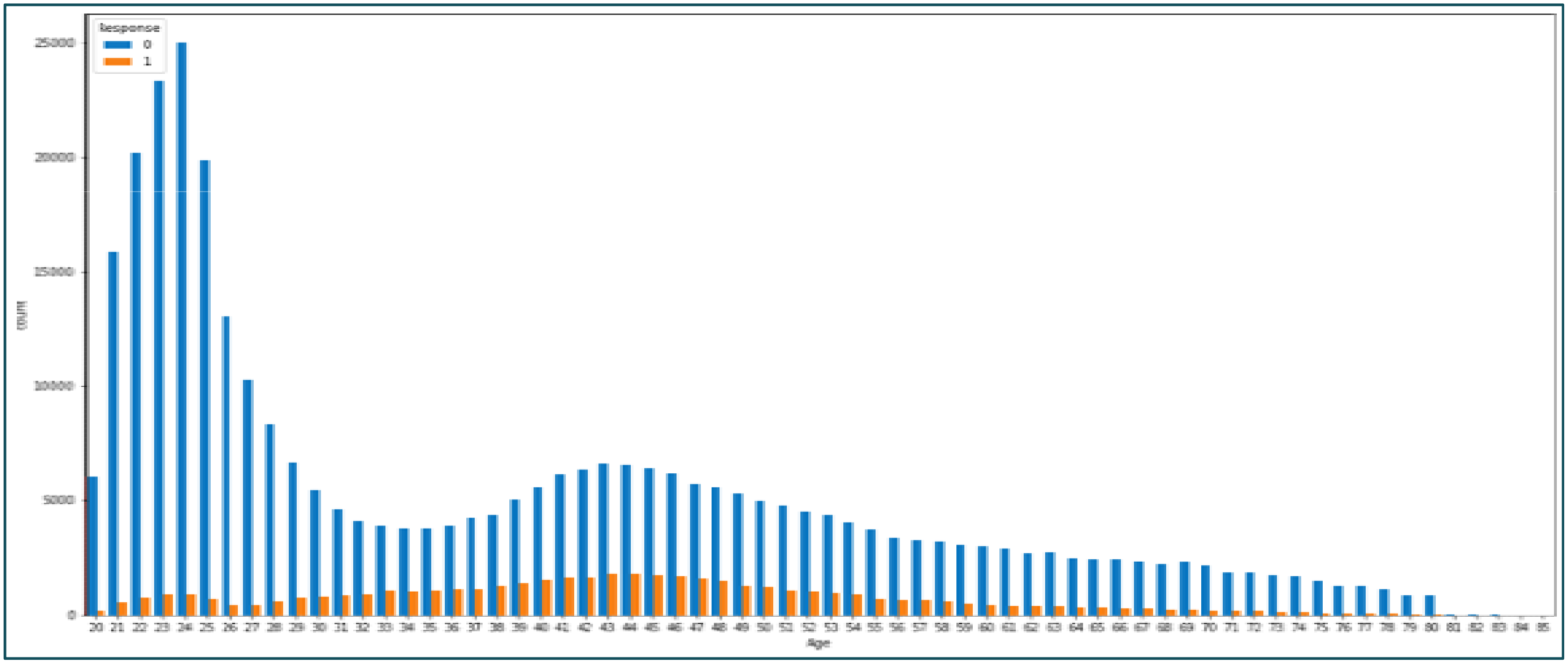


## Male customers' proportion is higher in both types of response than Female customers.

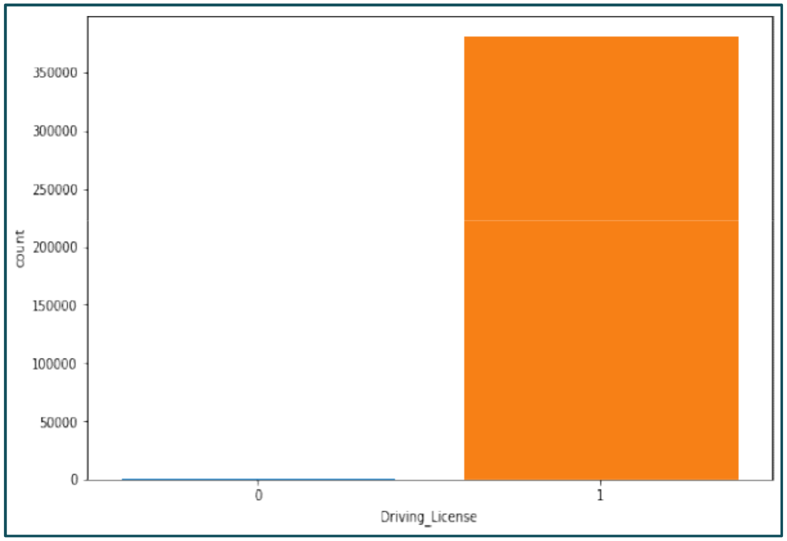


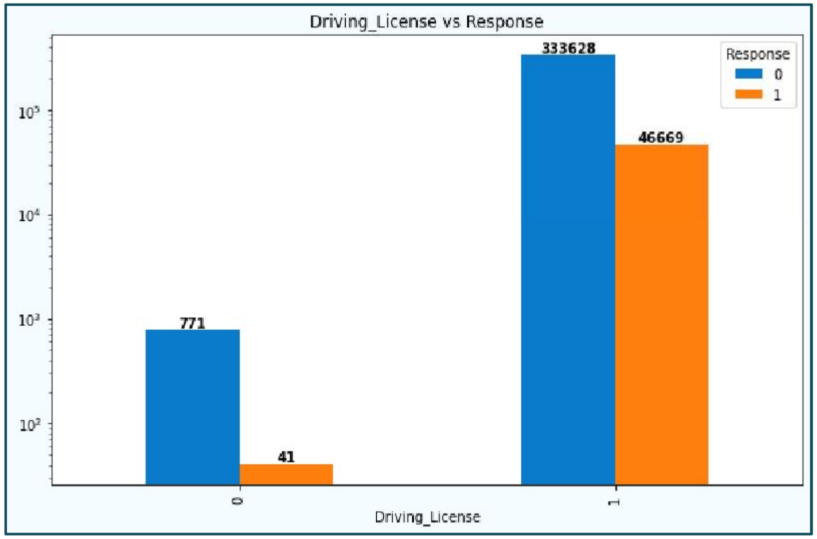
## The Age of policyholders ranges from 20 to 85.

## People aged between 30-57 are more likely to be interested in the insurance policy.

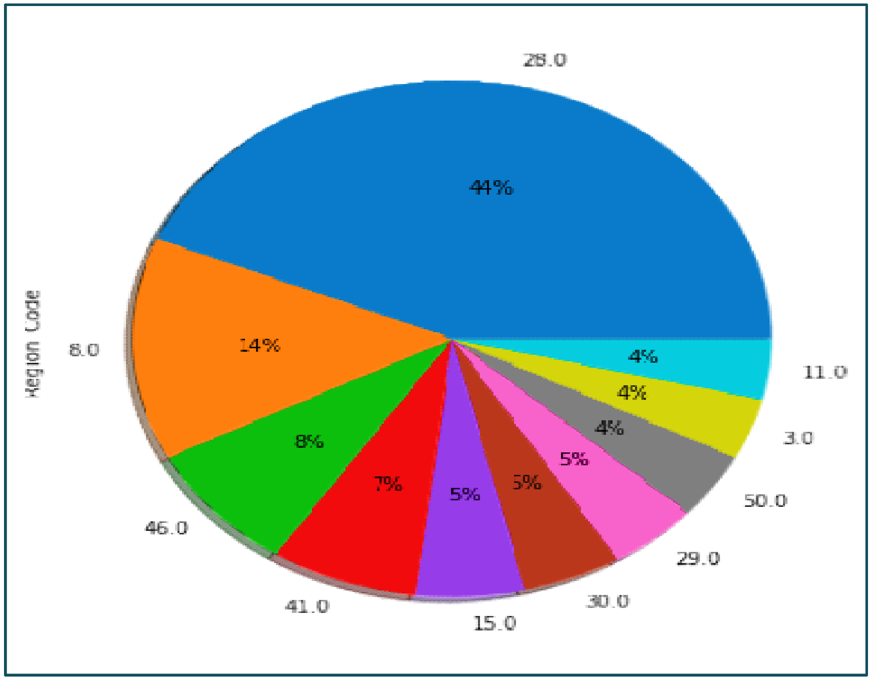


- Maximum policyholders acquire a driving license.
- From the customers who have D.L., only 12. 3 % of them  are interested in insurance policy.



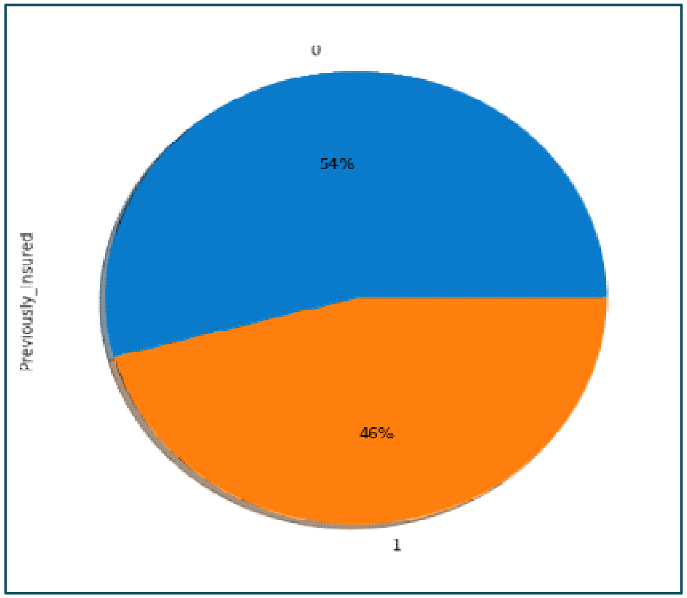


- From the above plot,we can observe that 41 customers who do not have a driving licence are also interested in the insurance policy.



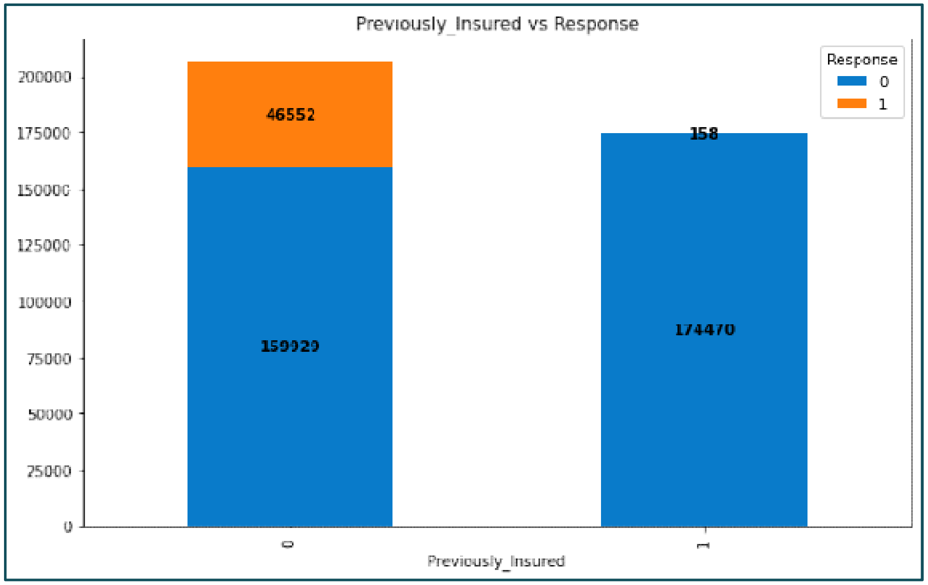
- Region code 28 is the largest contributor to people using insurance.

Around 54% of the customers does not have the vehicle insurance while 46% of the customers already have it.



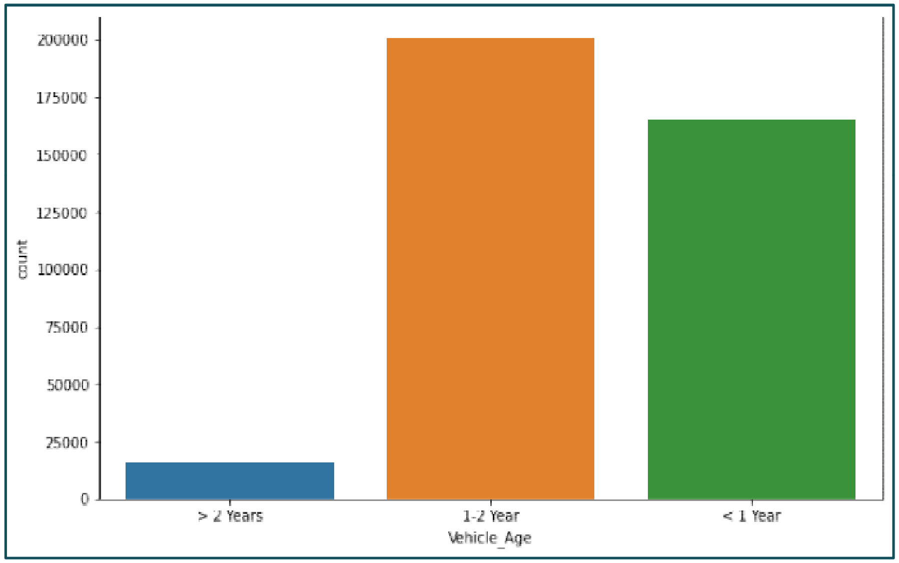
**Of the customers who were previously not insured, 46552 of them are interested in the policy, while the majority of them are not interested.**

**And also, among the customers who were previously insured, the majority of them are not interested in the policy.**



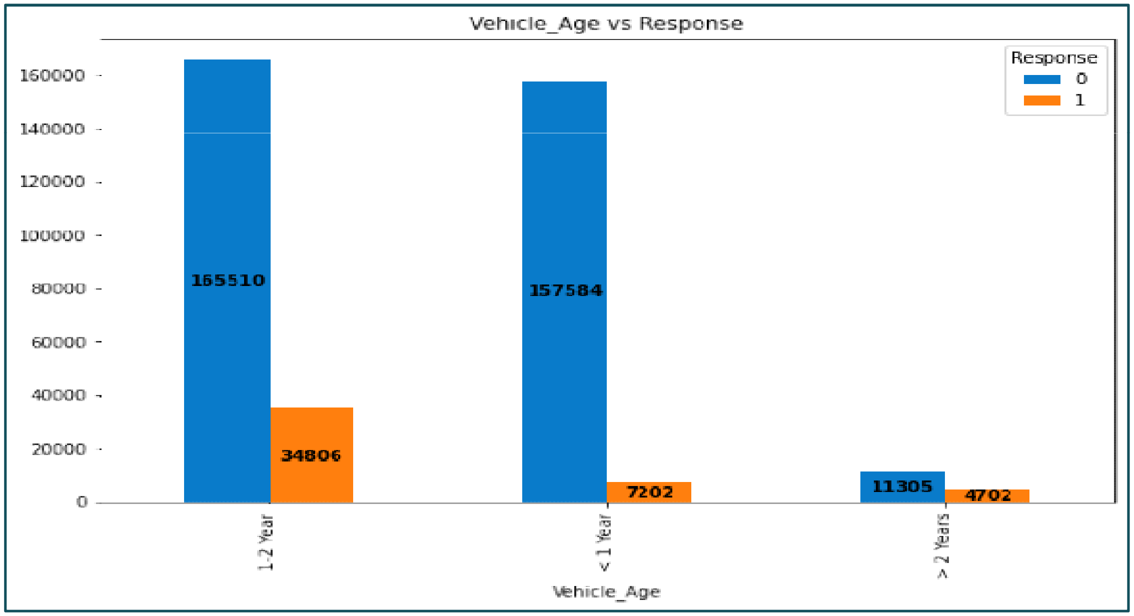
## Most of the customers have vehicles that are 1-2 years old.

## Very few customers have a vehicle more than 2 years old.



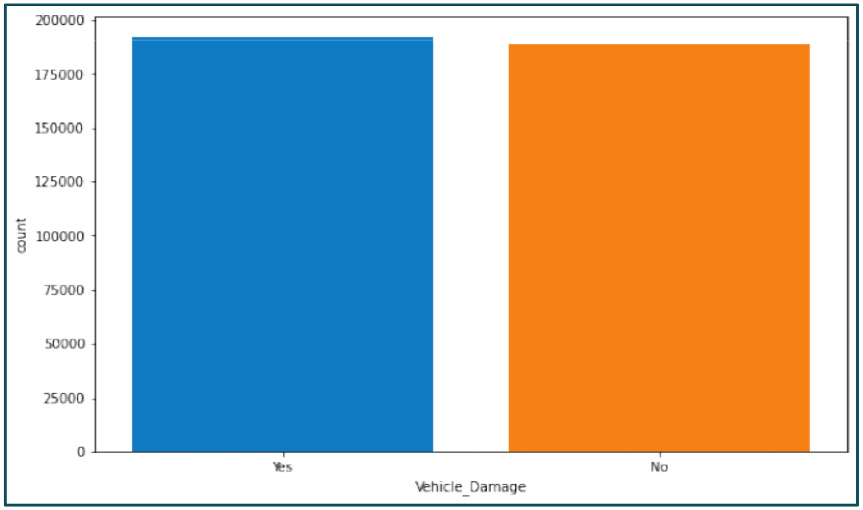
**The majority of customers interested in the insurance policy have vehicles that are 1 to 2 years old, followed by those with vehicles that are less than 1 year old.**

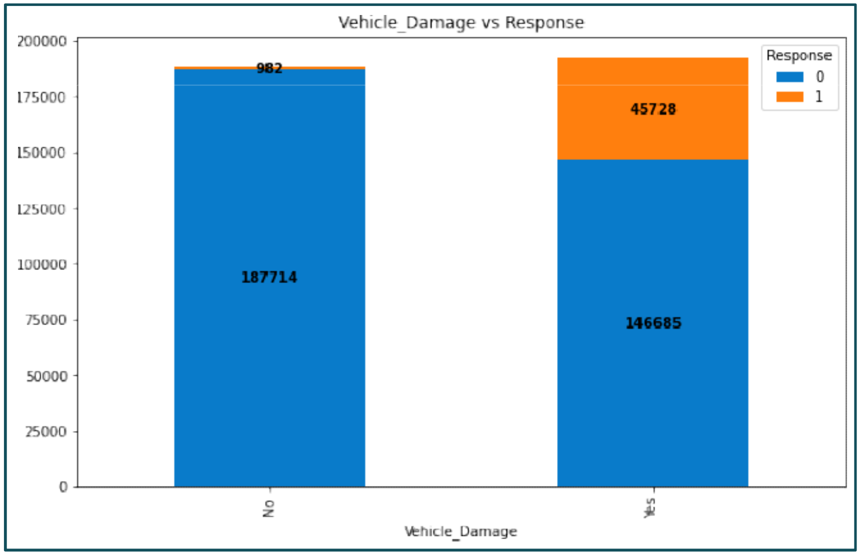
**Very few customers are interested in the policy if they have more than two-year-old vehicles.**



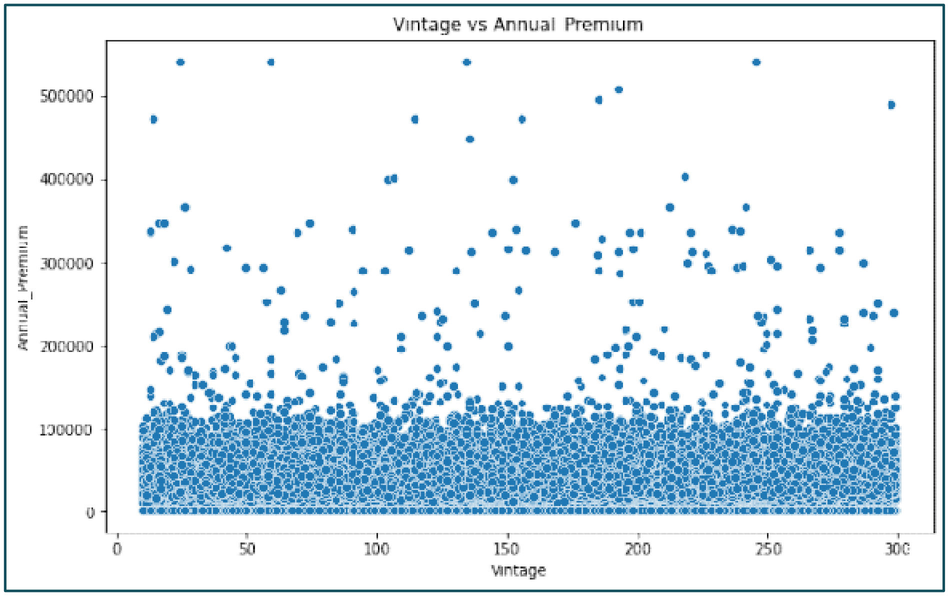
**The plot shows that the number of customers who damaged their vehicles and the ones who didn't are almost equal.**

**If we observe the number of customers who are interested in the insurance policy, then the maximum number of them are those who have had vehicle damage in the past.**





**It seems that the number of days the customer is associated with the company does not affect the amount of Annual Premium.**



# Interpretation and Insight From EDA :

- The **proportion of customer not responding** at **87.7%**. It shows that, customers are not interested about the idea of having vehicle insurance.
- Majority **age** customers **who interested in subscribing vehicle insurance at region 28** are between **30 - 57 years** old.
- **Region 28** is the **largest contributor** to people **using insurance** and also have **largest positive respond** than other regions, **but** have **largest road accident** than other regions.

# Feature Engineering :

- The binary features ‘Gender’ and ‘Vehicle Damage’ are encoded in the form of 0 and 1 for the response No and Yes respectively.

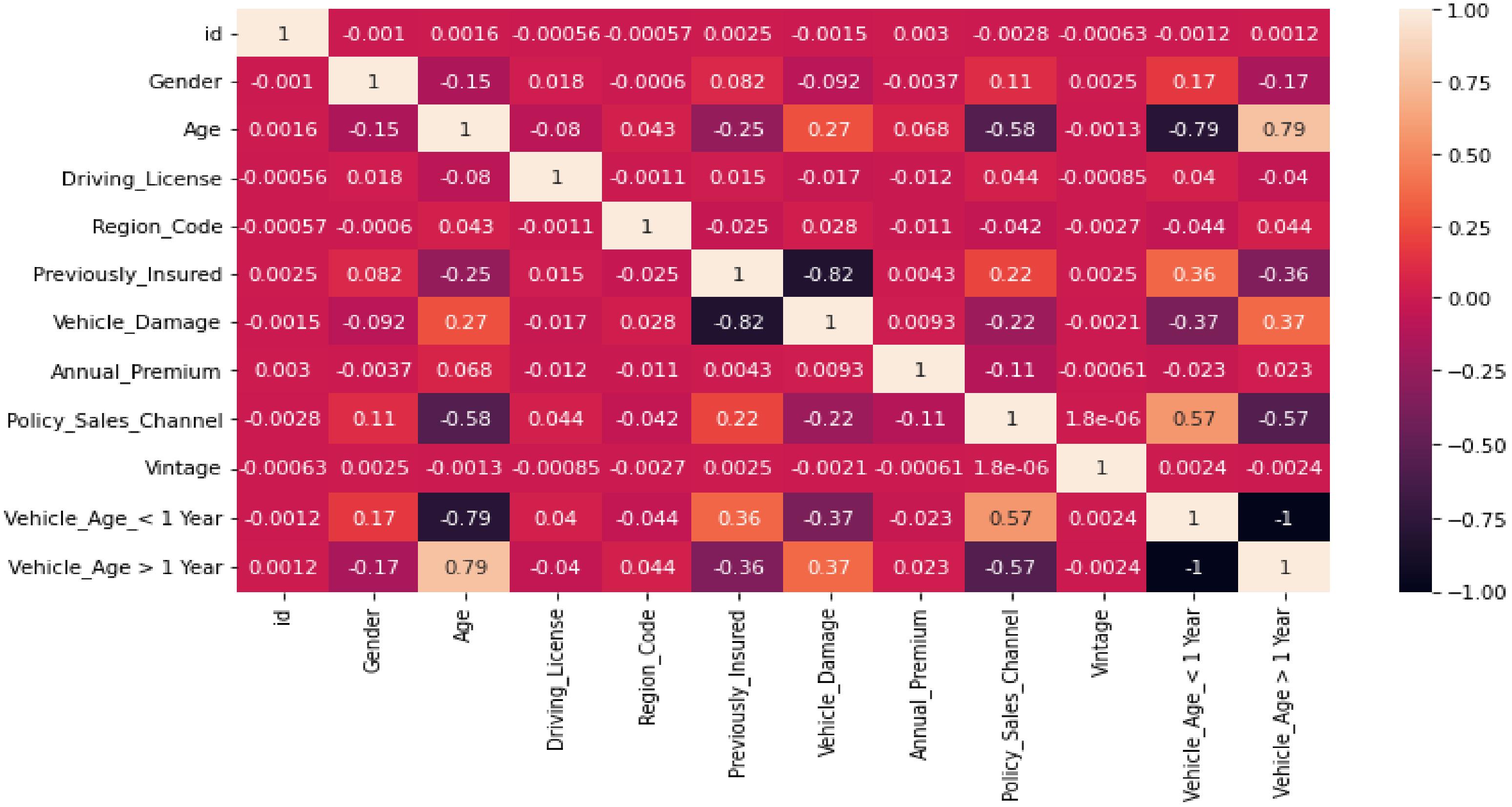
- One Hot Encoding is performed on the ‘Vehicle_Age’ Feature.

- The columns ‘Vehicle_Age_1-2 Year’ and ‘Vehicle_Age_> 2 Years’ have been merged as follows:

  x ['Vehicle_Age > 1 Year'] = x ['Vehicle_Age_1-2 Year'] + x ['Vehicle_Age_> 2 Years']

- The multicollinearity from the features is removed by keeping the VIF value as low as possible.

# Feature Engineering :



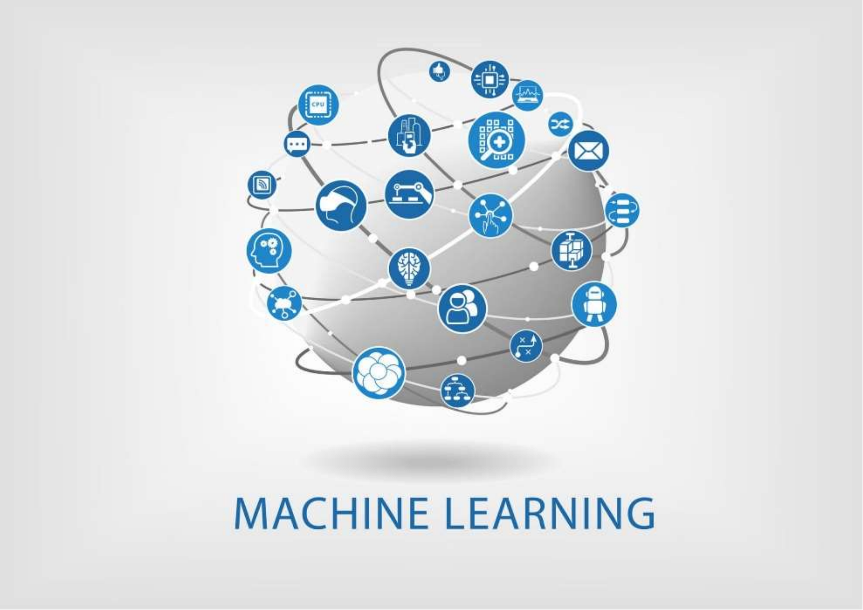
| | id | Gender | Age | Driving_License | Region_Code | Previously_Insured | Vehicle_Damage | Annual_Premium | Policy_Sales_Channel | Vintage | Vehicle_Age_< 1 Year | Vehicle_Age_> 1 Year |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| id | 1 | -0.001 | 0.0016 | -0.00056 | -0.00057 | 0.0025 | -0.0015 | 0.003 | -0.0028 | -0.00063 | -0.0012 | 0.0012 |
| Gender | -0.001 | 1 | -0.15 | 0.018 | -0.0006 | 0.082 | -0.092 | -0.0037 | 0.11 | 0.0025 | 0.17 | -0.17 |
| Age | 0.0016 | -0.15 | 1 | -0.08 | 0.043 | -0.25 | 0.27 | 0.068 | -0.58 | -0.0013 | -0.79 | 0.79 |
| Driving_License | -0.00056 | 0.018 | -0.08 | 1 | -0.0011 | 0.015 | -0.017 | -0.012 | 0.044 | -0.00085 | 0.04 | -0.04 |
| Region_Code | -0.00057 | -0.0006 | 0.043 | -0.0011 | 1 | -0.025 | 0.028 | -0.011 | -0.042 | -0.0027 | -0.044 | 0.044 |
| Previously_Insured | 0.0025 | 0.082 | -0.25 | 0.015 | -0.025 | 1 | -0.82 | 0.0043 | 0.22 | 0.0025 | 0.36 | -0.36 |
| Vehicle_Damage | -0.0015 | -0.092 | 0.27 | -0.017 | 0.028 | -0.82 | 1 | 0.0093 | -0.22 | -0.0021 | -0.37 | 0.37 |
| Annual_Premium | 0.003 | -0.0037 | 0.068 | -0.012 | -0.011 | 0.0043 | 0.0093 | 1 | -0.11 | -0.00061 | -0.023 | 0.023 |
| Policy_Sales_Channel | -0.0028 | 0.11 | -0.58 | 0.044 | -0.042 | 0.22 | -0.22 | -0.11 | 1 | 1.8e-06 | 0.57 | -0.57 |
| Vintage | -0.00063 | 0.0025 | -0.0013 | -0.00085 | -0.0027 | 0.0025 | -0.0021 | -0.00061 | 1.8e-06 | 1 | 0.0024 | -0.0024 |
| Vehicle_Age_< 1 Year | -0.0012 | 0.17 | -0.79 | 0.04 | -0.044 | 0.36 | -0.37 | -0.023 | 0.57 | 0.0024 | 1 | -1 |
| Vehicle_Age_> 1 Year | 0.0012 | -0.17 | 0.79 | -0.04 | 0.044 | -0.36 | 0.37 | 0.023 | -0.57 | -0.0024 | -1 | 1 |

# Feature Engineering :

- The data has been scaled to improve the model performance using MinMaxScaler.

- Some of the features in our dataset are highly imbalanced, hence to avoid this error, the dataset is balanced using technique called SMOTE(Synthetic Minority Oversampling Technique)

# Machine Learning



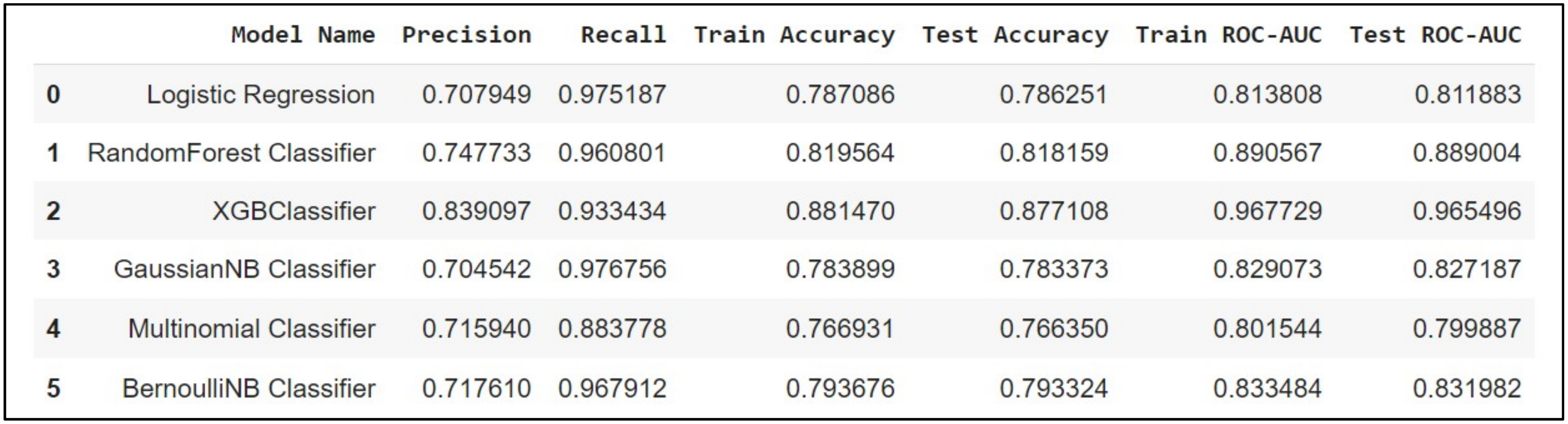
## ML Models Used :

1. Logistic Regression
2. Random Forest Classifier
3. XGBoost Classifier
4. Naïve-Bayes Classifier

## Hyper-Parameter Tuning metod used:

1. GridSearch CV

## Results obtained after Training the Dataset :



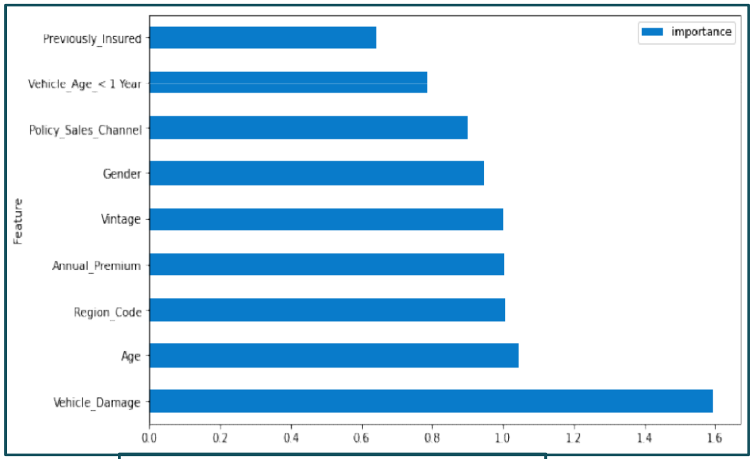
| | Model Name | Precision | Recall | Train Accuracy | Test Accuracy | Train ROC-AUC | Test ROC-AUC |
|---|---|---|---|---|---|---|---|
| 0 | Logistic Regression | 0.707949 | 0.975187 | 0.787086 | 0.786251 | 0.813808 | 0.811883 |
| 1 | RandomForest Classifier | 0.747733 | 0.960801 | 0.819564 | 0.818159 | 0.890567 | 0.889004 |
| 2 | XGBClassifier | 0.839097 | 0.933434 | 0.881470 | 0.877108 | 0.967729 | 0.965496 |
| 3 | GaussianNB Classifier | 0.704542 | 0.976756 | 0.783899 | 0.783373 | 0.829073 | 0.827187 |
| 4 | Multinomial Classifier | 0.715940 | 0.883778 | 0.766931 | 0.766350 | 0.801544 | 0.799887 |
| 5 | BernoulliNB Classifier | 0.717610 | 0.967912 | 0.793676 | 0.793324 | 0.833484 | 0.831982 |

**After training the models and comparing the results, it can be said that the XGBoost Classifier model has performed better than the other models.**

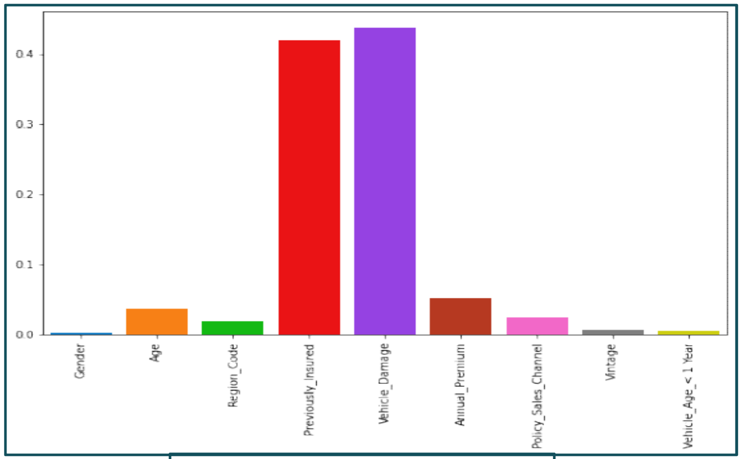
# Feature Importance:

**Most important feature according to Logistic Regression Model is Vehicle Damage followed by Age**

**Previously Insured and Vehicle Damage are most important features according to XGB Model.**



Logistic Regression



XGBoost Classifier

# Conclusion:

After loading the dataset, cleaning the data, performing EDA, Feature Engineering and after feature selection, Models are built.In terms Training and Testing Accuracy and ROC-AUC score, XGBoost Classifier gave the best results.

- Vehicle_damage and Previously_Insured came out as the most important features for the model.
- We will choose in this case is XGBoost Machine Learning. This model is able to predict the label of the target customer who is response and not response at an recall model with value 93% and Roc-AUC score at 96%.
- This means our model can improve our response rate for predicting customers who interested in subscribing vehicle insurance

# Thank You