Capstone Project – 3

Health Insurance Cross Sell Prediction

Presented By:
Aehteshaam Shaikh

Problem Statement :

Our client is an Insurance company that has provided Health Insurance to its customers.

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.

Objective :

Just like medical insurance, there is vehicle insurance where every year customer needs to pay a premium of certain amount to insurance provider company so that in case of unfortunate accident by the vehicle, the insurance provider company will provide a compensation (called ‘sum assured’) to the customer.

The objective of this project is to build a Model to predict whether the policyholders from the past year will also be interested in vehicle insurance provided by the company.

Data Summary:

Sr	Column Name	Description	Datatype
1	id	Unique ID for the customer	Numerical
2	Gender	Gender of the customer	Categorical
3	Age	Age of the customer	Numerical
4	Driving_License	0 : Customer does not have DL 1 : Customer already has DL	Binary
5	Region_Code	Unique code for the region of the customer	Numerical
6	Previously_Insured	1 : Customer already has Vehicle Insurance, 0 : Customer doesn't have Vehicle Insurance	Binary
7	Vehicle_Age	Age of the Vehicle	Numerical

Data Summary:

Sr	Column Name	Description	Datatype
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.	Binary
9	**Annual_Premium**	The amount customer needs to pay as premium in the year	Numerical
10	**PolicySalesChannel**	Anonymized Code for the channel of outreaching to the customer ie. Different Agents, Over Mail, Over Phone, In Person, etc.	Numerical
11	**Vintage**	Number of Days, Customer has been associated with the company	Numerical
12	**Response (Target)**	1 : Customer is interested, 0 : Customer is not interested	Binary

Data Cleaning:

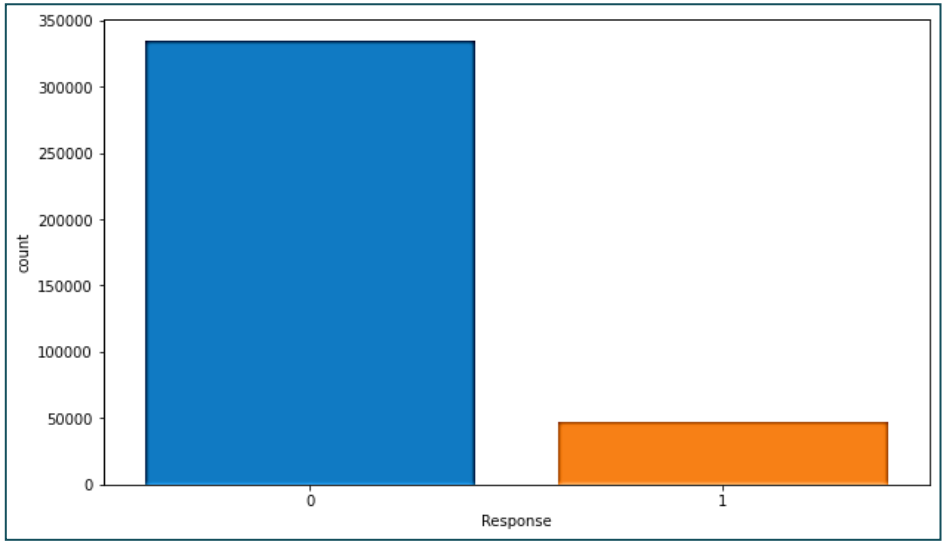
The data set consists of around 381109 entries and 12 columns.

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

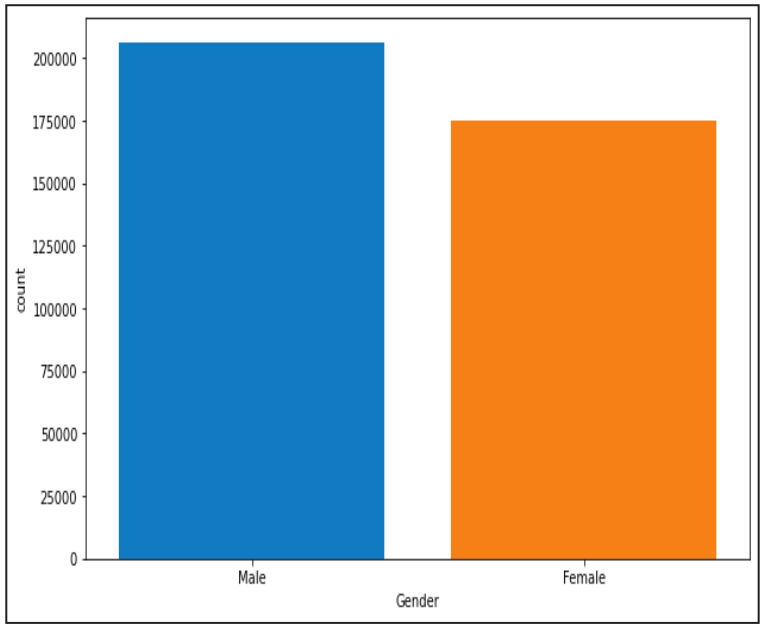
The dataset doesn't have any feature with Null Values!!!

Exploratory Data Analysis

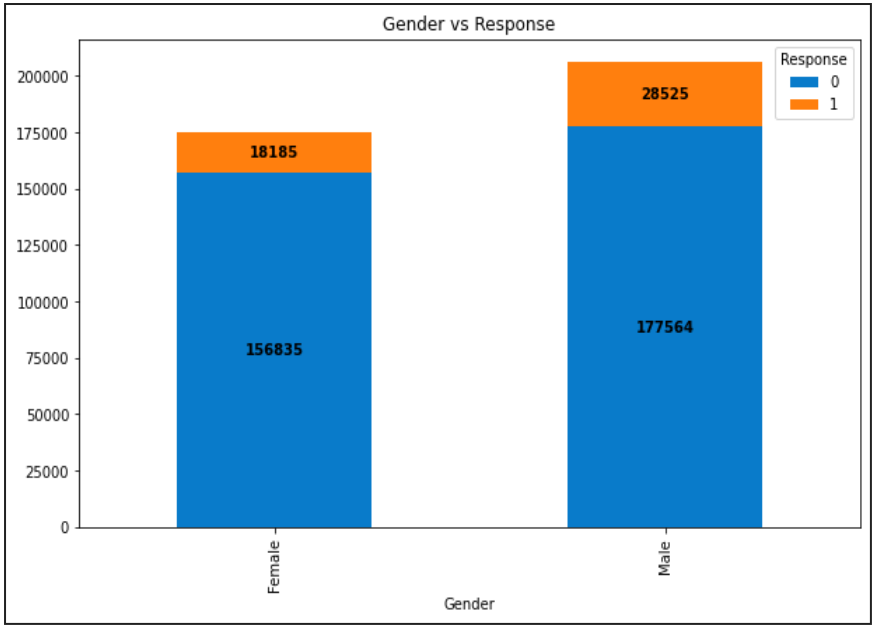
- **The Target variable is highly imbalance with maximum customers having no interest in insurance policy.**



The Male Policyholders are slightly more than the Female policyholders.

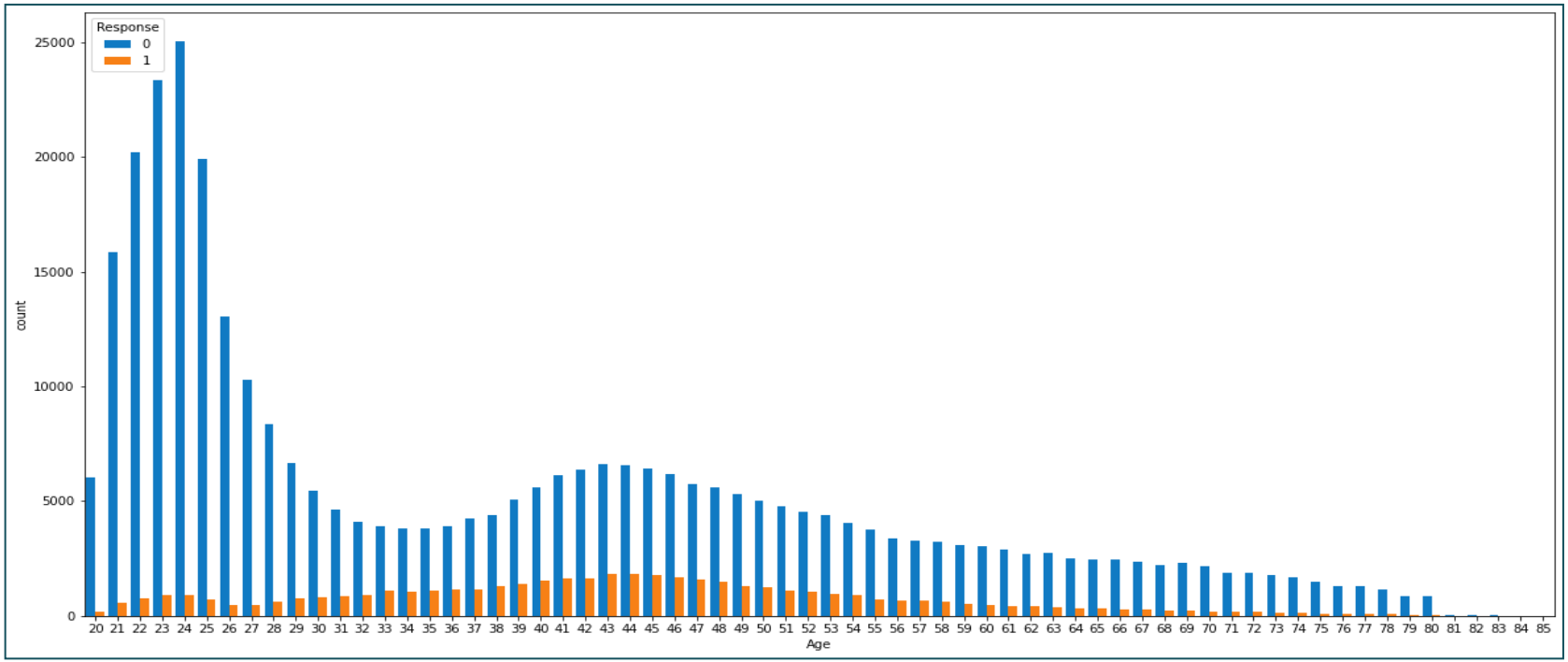


Male customers' proportion is higher in both types of response than Female customers.

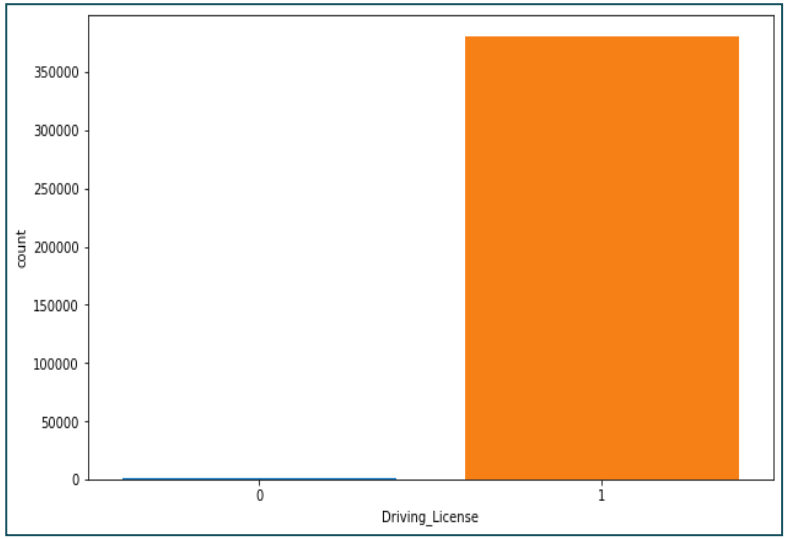


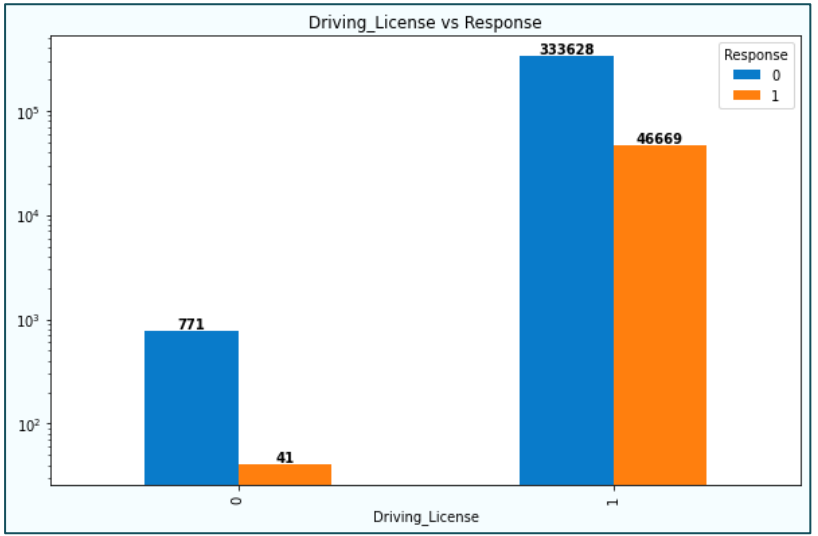
The Age of policyholders ranges from 20 to 85.

People aged between 30-60 are more likely to be interested in the insurance policy.



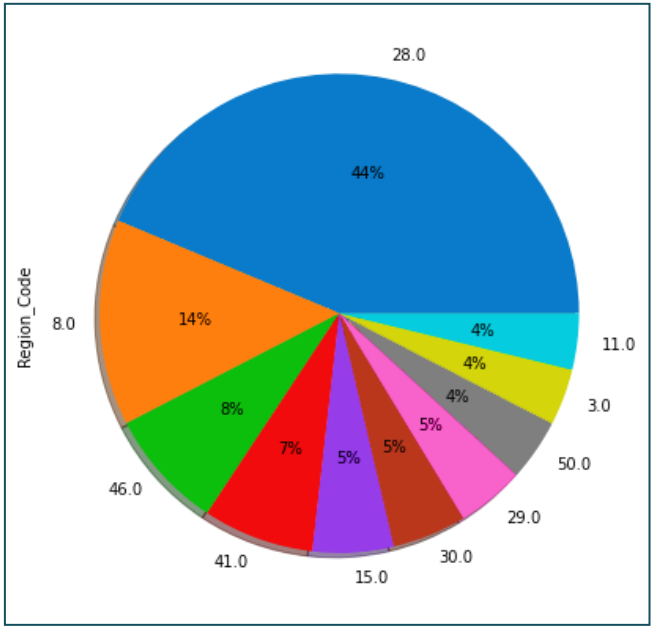
- Maximum policyholders acquire a driving license.
- From the customers who have D.L., only 12.3% of them are interested in insurance policy.



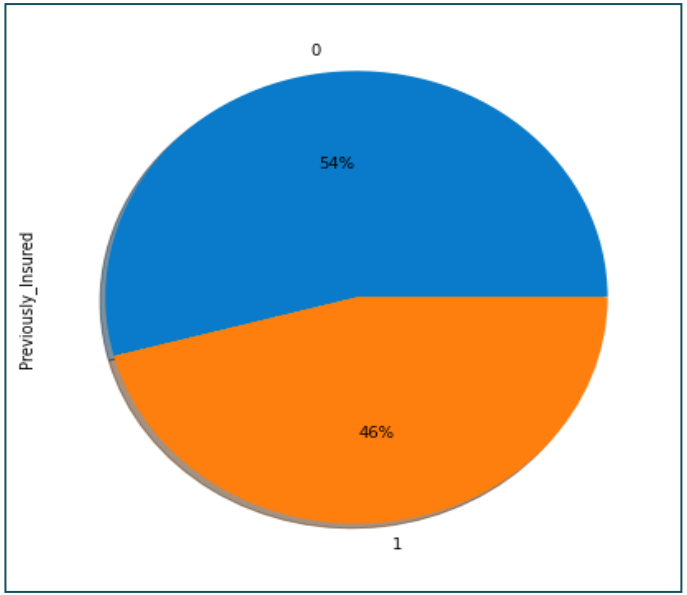


- From the above plot,we can observe that 41 customers who do not have a driving licence are also interested in the insurance policy.

Most of the policyholders belong to the region that has region code 28.

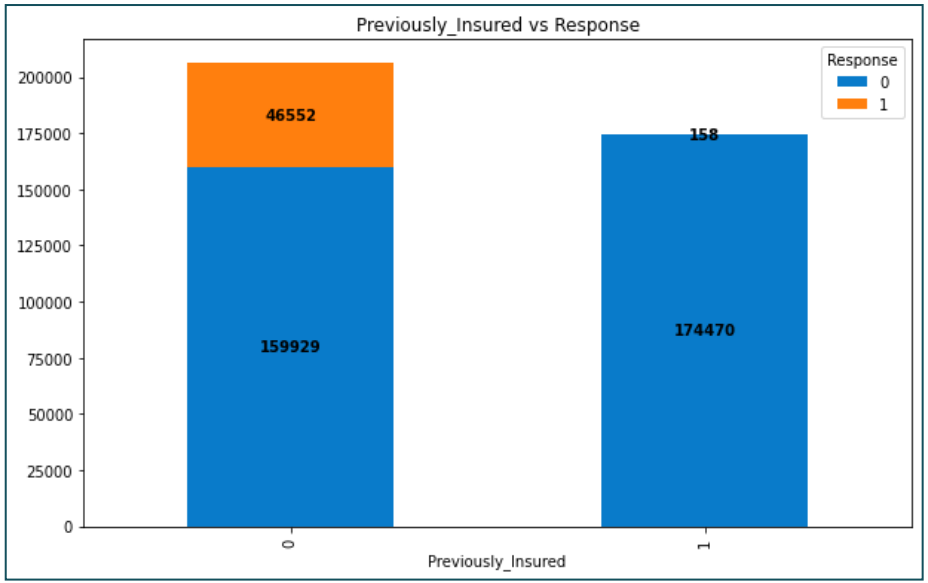


Around 54% of the customers does not have the vehicle insurance while 46% of the customers already have it.



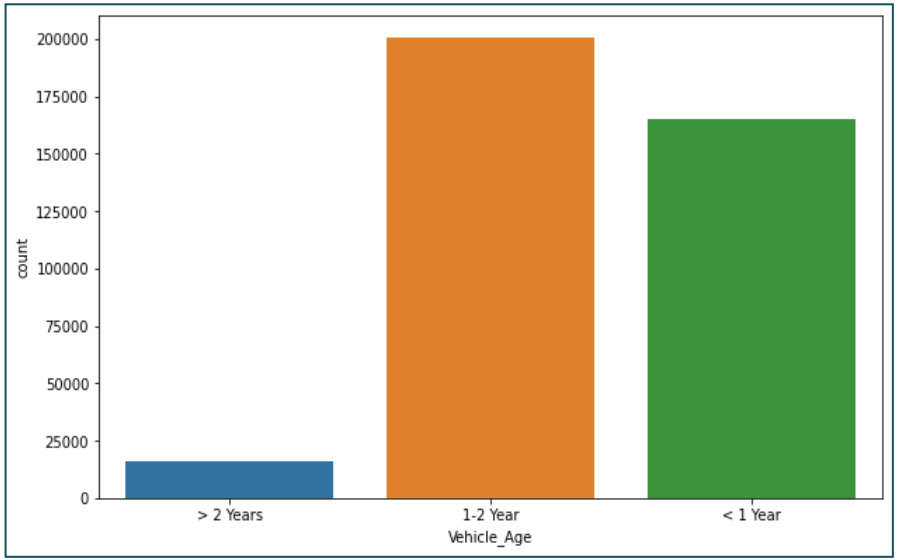
Of the customers who were previously not insured, 46552 of them are interested in the policy, while the majority of them are not interested.

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



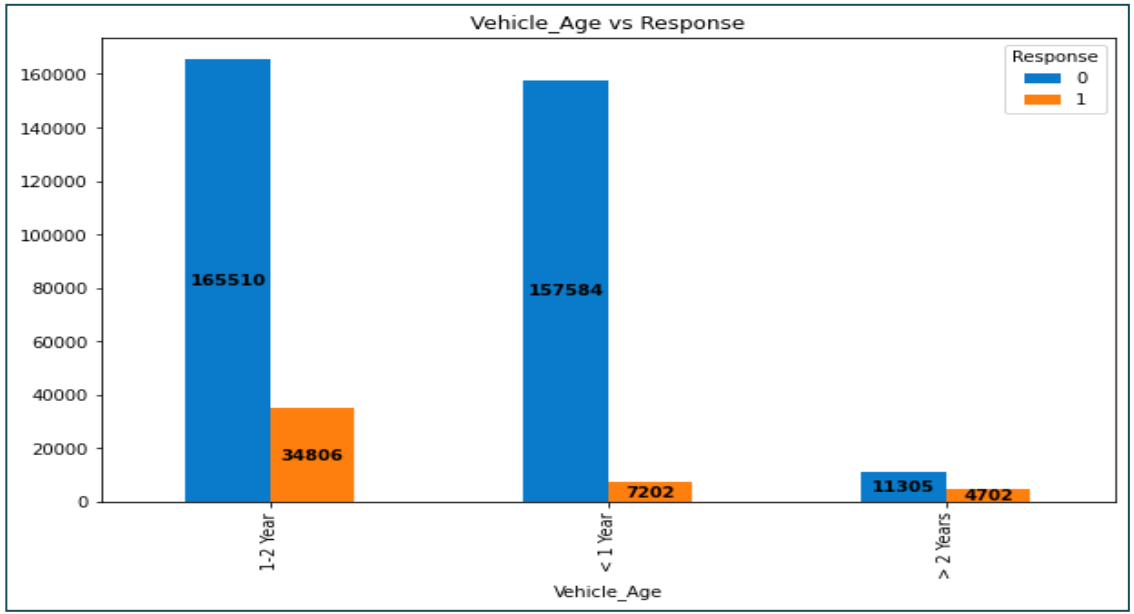
Most of the customers have vehicles that are 1-2 years old.

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



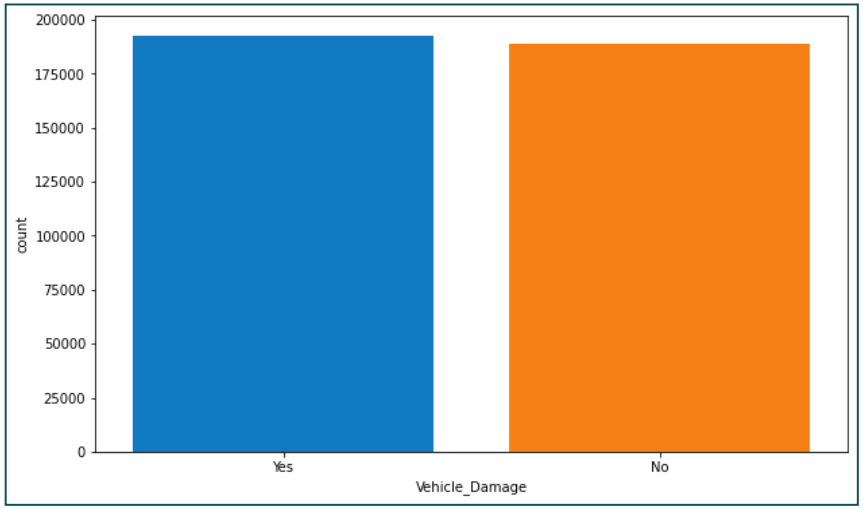
The majority of customers interested in the insurance policy have vehicles that are 1 to 2 years old, followed by those with vehicles that are less than 1 year old.

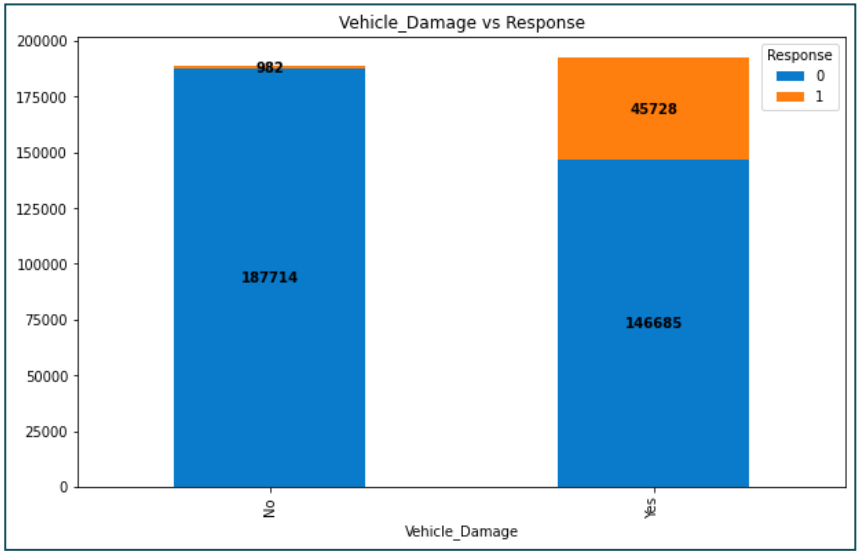
Very few customers are interested in the policy if they have more than two-year-old vehicles.



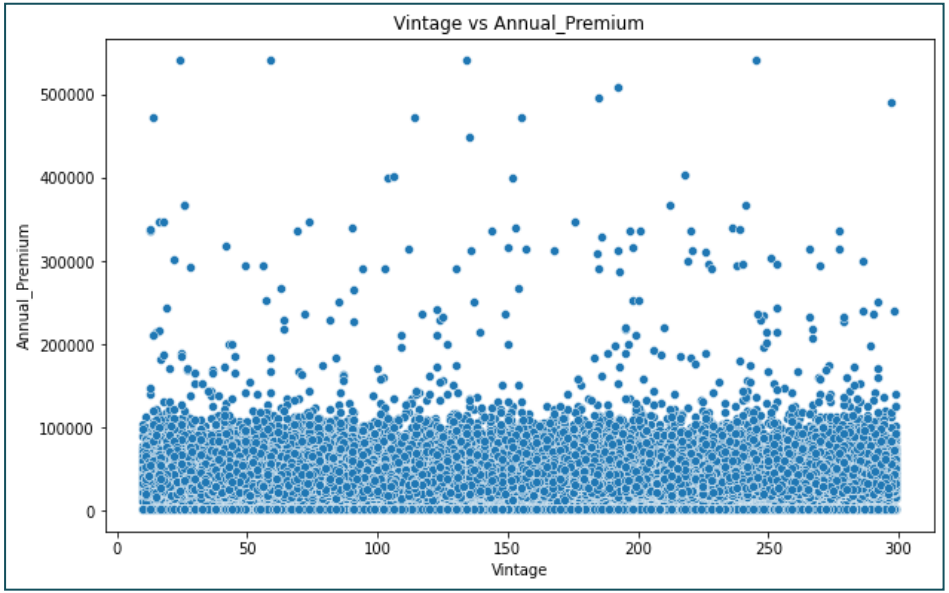
The plot shows that the number of customers who damaged their vehicles and the ones who didn't are almost equal.

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.



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 :

- 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)

AI

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 :

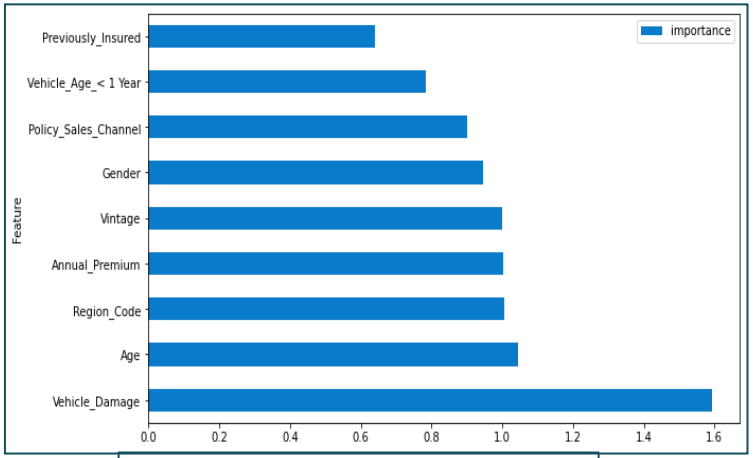
	Model Name	Precision	Recall	Train Accuracy	Test Accuracy	Train ROC-AUC	Test ROC-AUC
0	Logistic Regression	0.711132	0.972396	0.787801	0.788220	0.813540	0.814463
1	RandomForest Classifier	0.740192	0.964140	0.812439	0.812437	0.882025	0.880653
2	XGBClassifier	0.818496	0.939432	0.868654	0.865250	0.959739	0.957338
3	GaussianNB Classifier	0.705790	0.976651	0.783852	0.784277	0.826447	0.826927
4	Multinomial Classifier	0.716952	0.884194	0.766813	0.767032	0.800557	0.801394
5	BernoulliNB Classifier	0.713756	0.970485	0.789837	0.790165	0.828541	0.827154

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

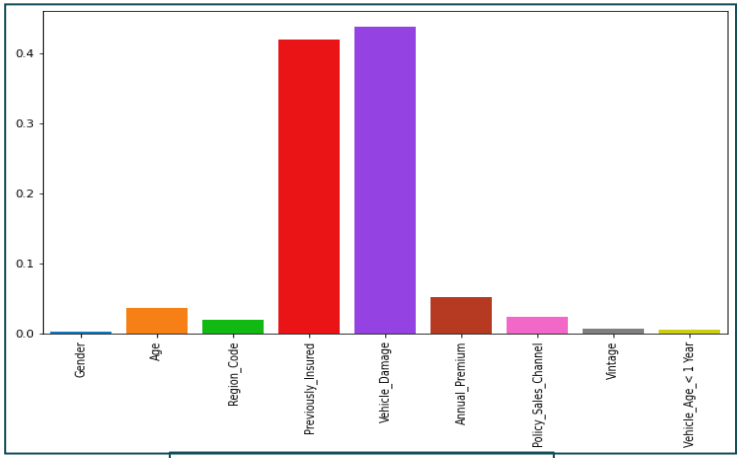
Feature Importance:

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.



Random Forest Regressor



XGBoost Regressor

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.

that the column avg views by sp

Thank You