Capstone Project

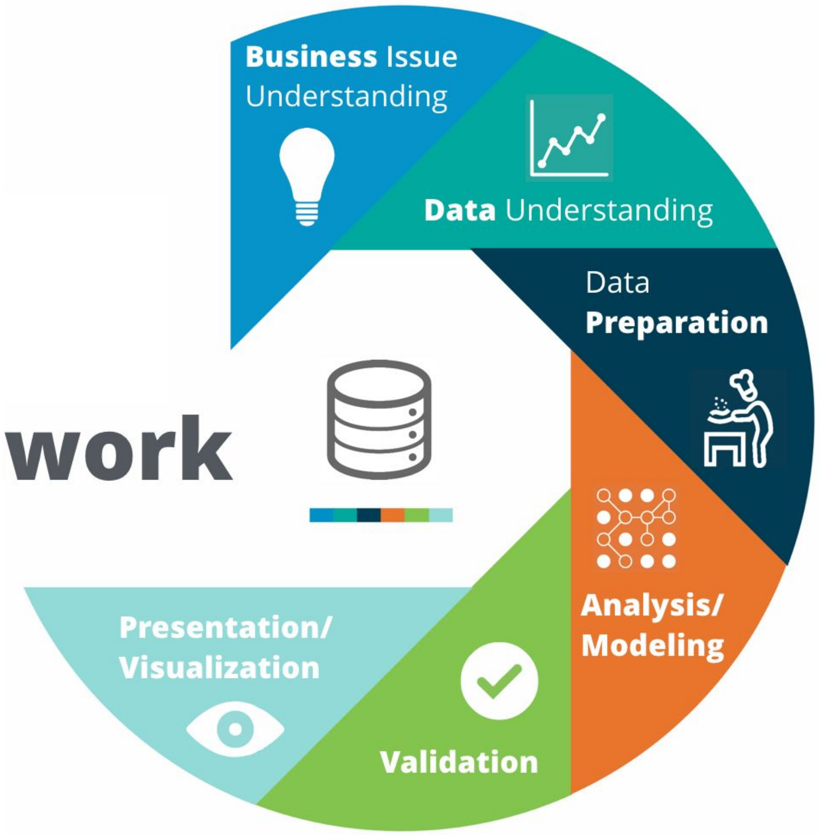
HEALTH INSURANCE CROSS SELL PREDICTION

Team Power

Hariom Bhardwaj
Mayank Kumar
Shivam Mishra
Saifuddin Raja
Sarvesh Kumar Yadav

CRISP-DM Framework :

Cross Industry Standard Process for Data Mining



Problem Statement

- Build a model to predict whether a customer would be interested in Vehicle Insurance.

- It is extremely helpful for the company because it can then accordingly plan its communication strategy to reach out to those customers and optimise its business model and revenue.

Exploratory Data Analysis

Understanding the Data

DATASET NAME: Health Insurance Cross-Sell Data

SHAPE:

- Data Points (Rows) : 381,109
- Features (Columns) : 12

TARGET VARIABLE:

- 'Response'

MISSING DATA CHECK:

- No missing, incorrect or invalid Data to Handle.

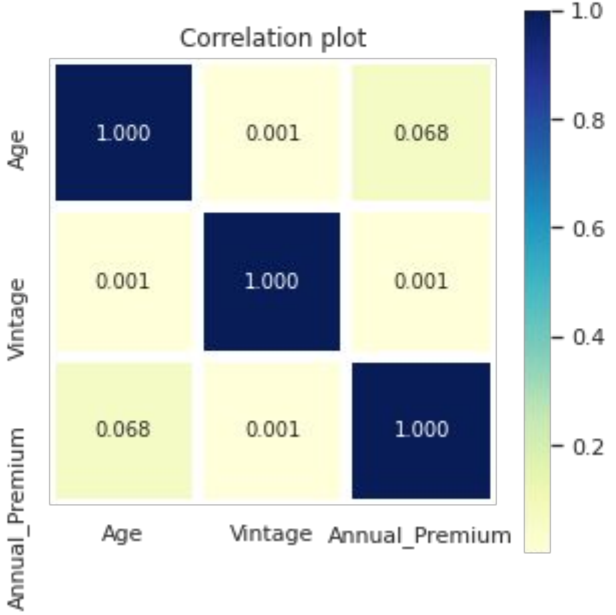
FEATURE:

```
1. id                  : Unique ID for the customer
2. Gender              : Gender of the customer
3. Age                 : Age of the customer
4. Driving_License     : whether Customer has DL
5. Region_Code         : Unique code for the region of the customer
6. Previously_Insured  : Whether Customer already has Vehicle Insurance
7. Vehicle_Age         : Age of the Vehicle
8. Vehicle_Damage      : Whether Customer got his/her vehicle damaged in the past
9. Annual_Premium      : The amount customer needs to pay as premium in the year
10. PolicySalesChannel : Code for the channel of outreaching to the customer
11. Vintage            : Number of Days, Customer has been associated with the co
12. Response           : Whether Customer is interested
```

Numerical Features

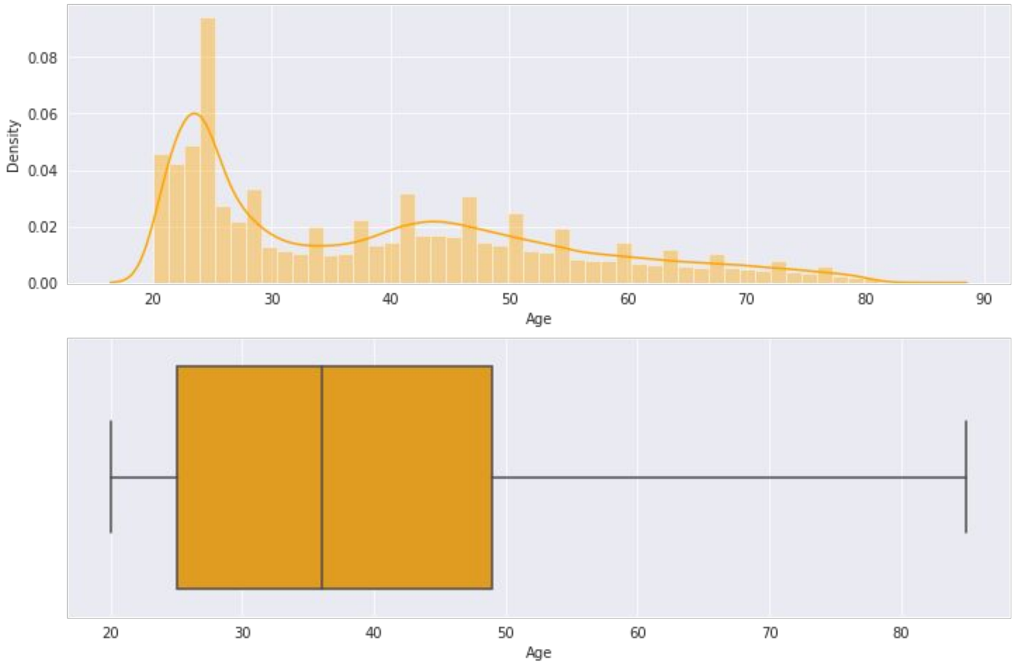
The Numerical (continuous) features of the data set include the :

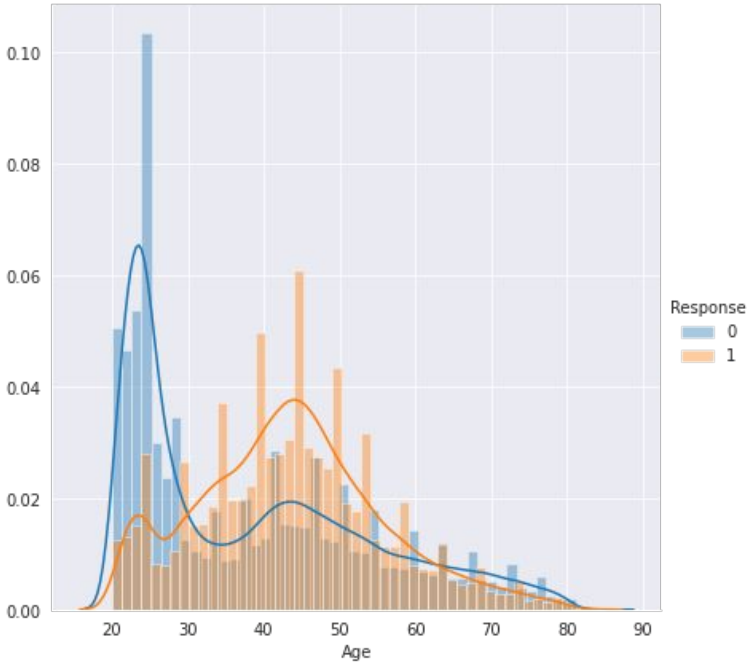
Age of the Customer,
The number of Days he has been a Customer,
And the Premium he pays annually



Feature	min	10%	25%	50%	Mean	75%	95%	99%	max
Age	20	22	25	36	38.82	49	69	77	85
Vintage	10	38	82	154	154.35	227	285	297	299
Annual_Premium	2630	2630	24405	31669	30564	39400	55176	72963	540165

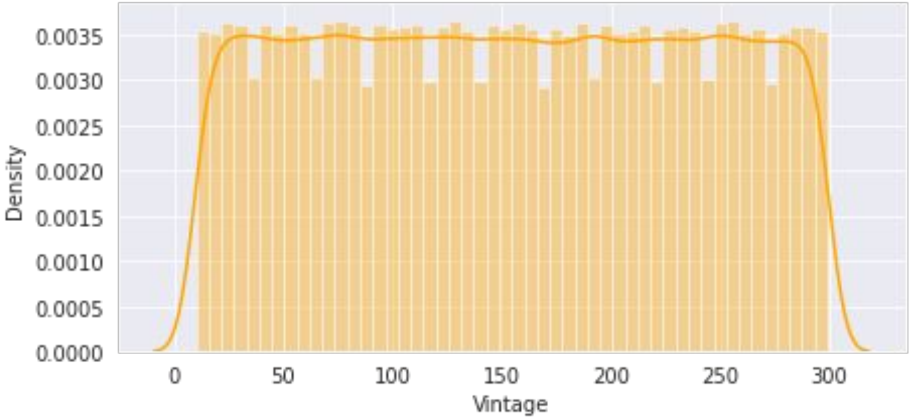
Age Distribution and its effect on Target Variable : Response





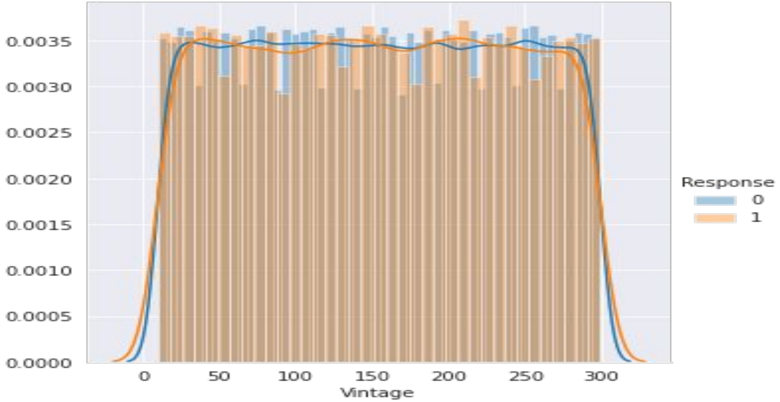
Distribution of Vintage

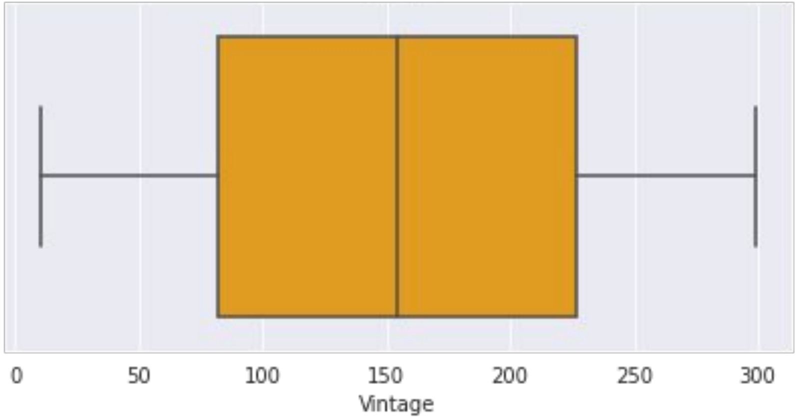
The Feature Vintage has very less information
and is Uniformly Distributed , With no skew .
Also, the Values are uniformly mixed ,
in both the classes of the target variable response .



This Feature potentially contribute to Over Fitting ,
Or it can also contain hidden information

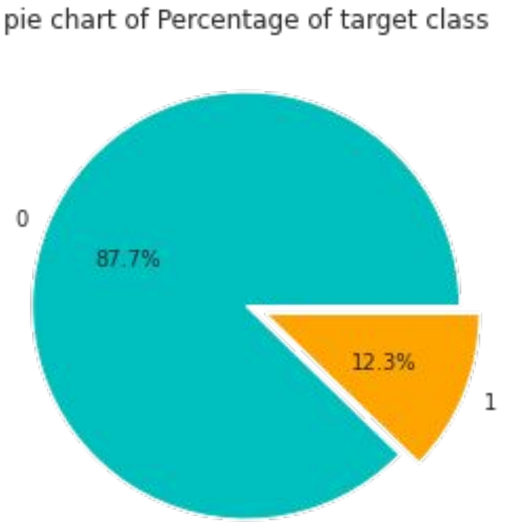
we need to analyse the feature_importances for this feature and decide whether to retain it or not





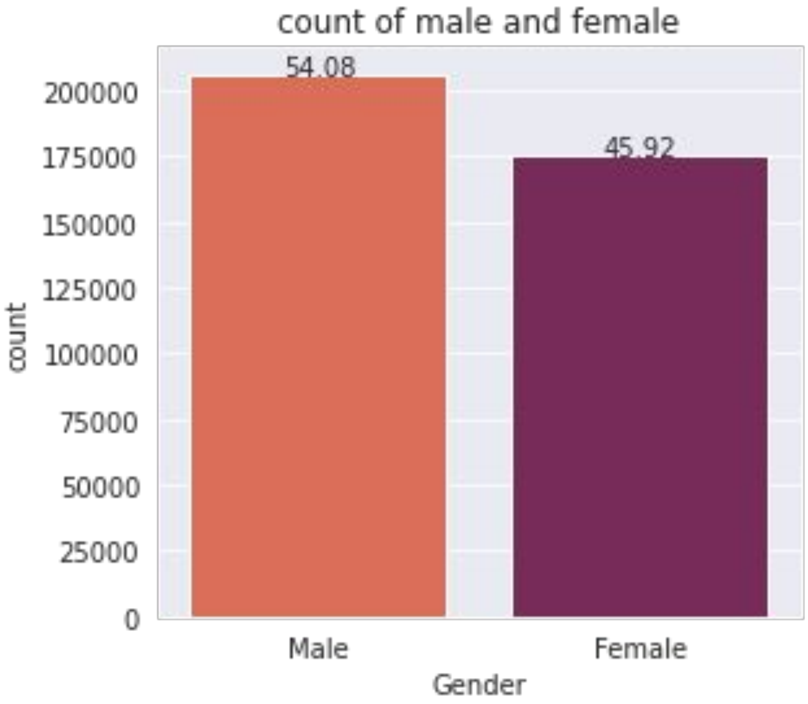
Categorical Features

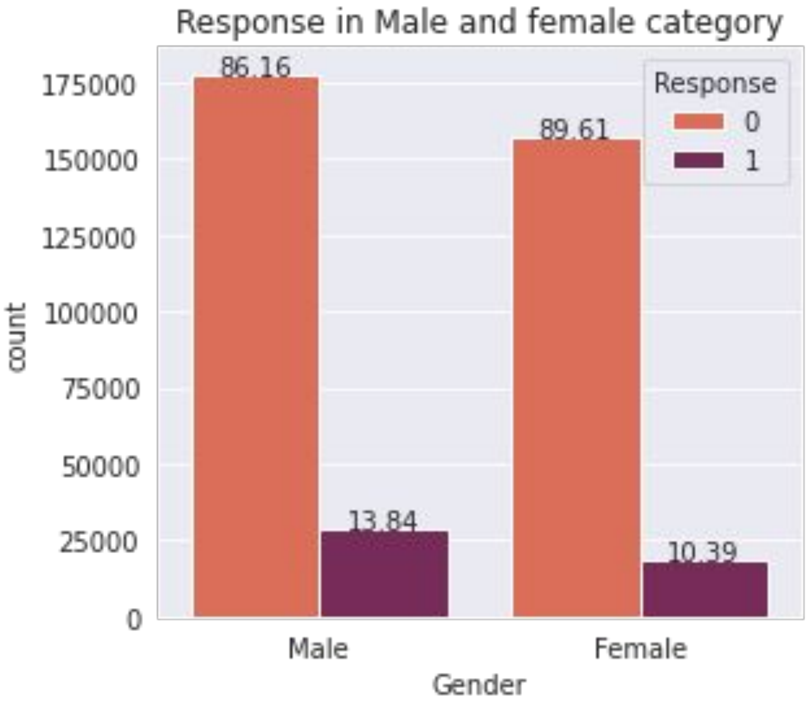
Features	# Categories	Top	% Frequency
Region_Code	53	28	28%
Policy_Sales_Channel	155	152	35%
Vehicle_Age	3	1-2 Year	53%
Gender	2	Male	54%
Driving_License	2	1	100%
Previously_Insured	2	0	54%
Vehicle_Damage	2	Yes	50%



Gender Distribution and its effect on Target Variable : Response

- The gender variable in the dataset is almost equally distributed
- Response in Male category is 13% than that of female category which is 10%.

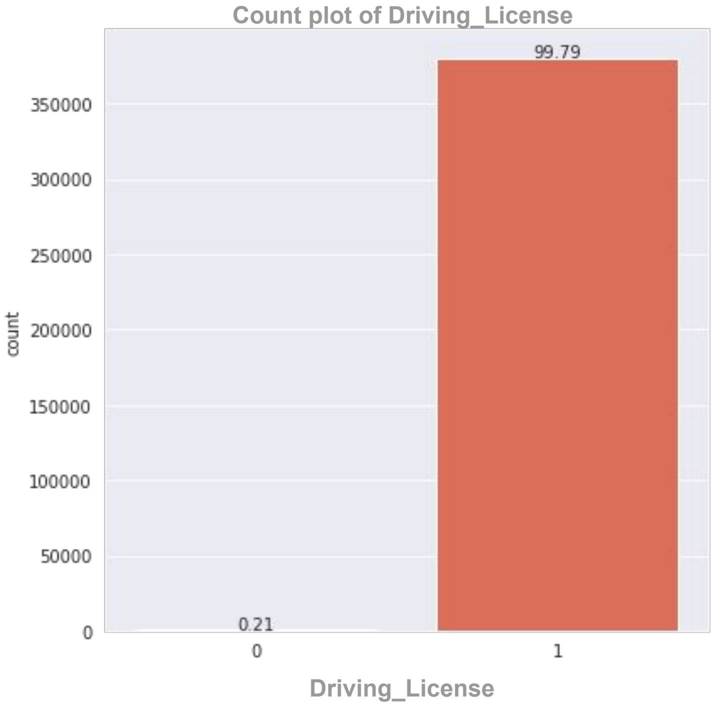




Driving License Distribution and its effect on Target Variable : Response

Driving license seems to be less important feature :

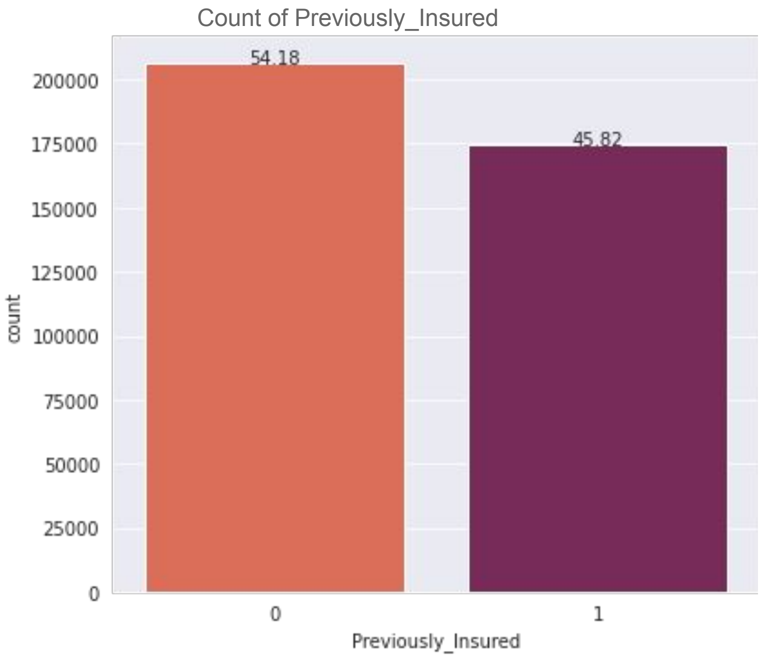
- Customers who have the DL are 99%
- Customers who are interested in Vehicle Insurance almost all have driving licence

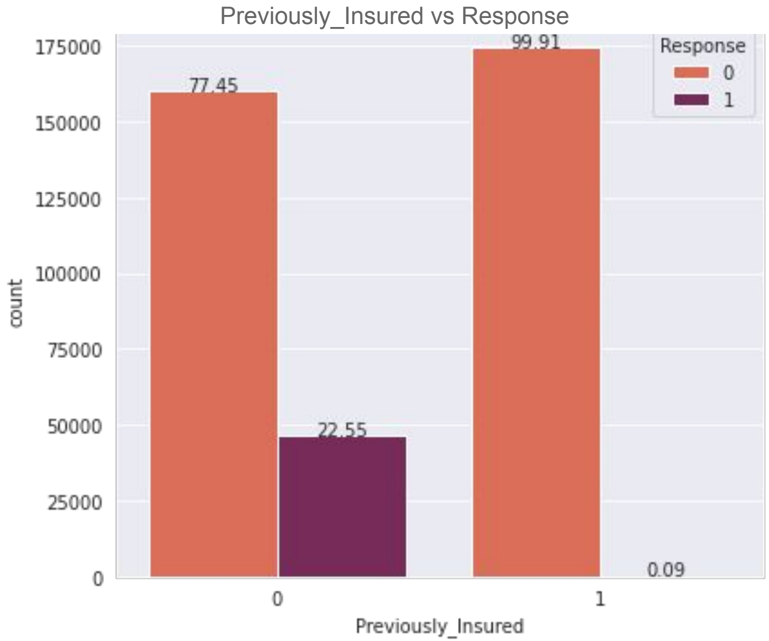


Previously_Insured Distribution and its effect on Target : Response

Customers who were previously insured tend not to be interested.

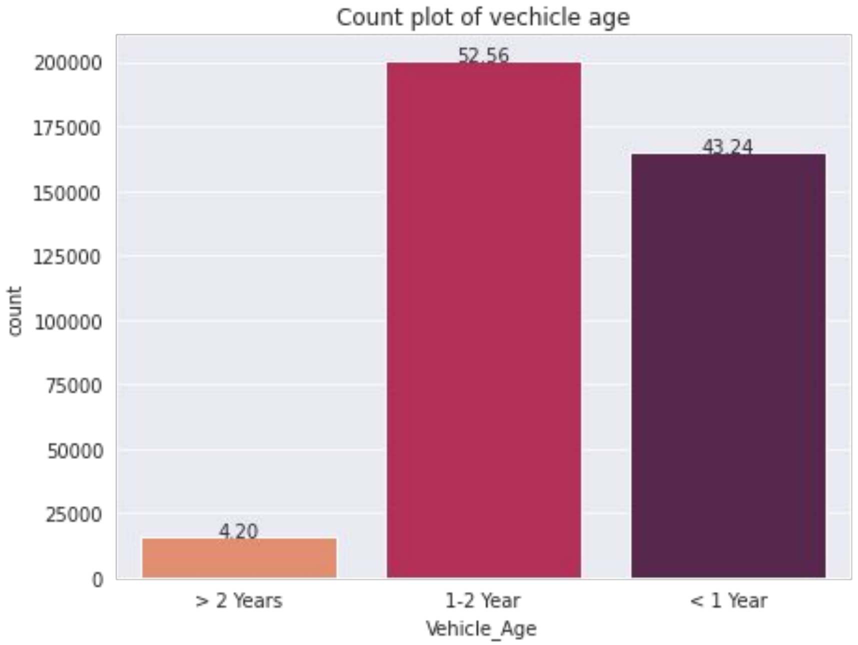
- We can think that the reason for this is that their previous insurance agreement has not expired yet
- Or maybe they are unsatisfied with previously purchased insurance services

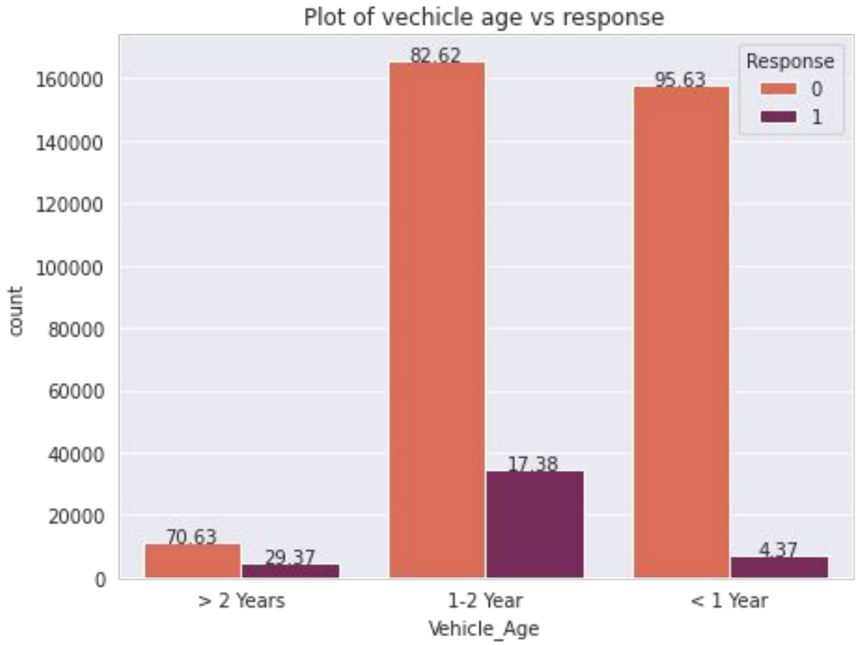




Vehicle Age Distribution and its effect on Target : Response

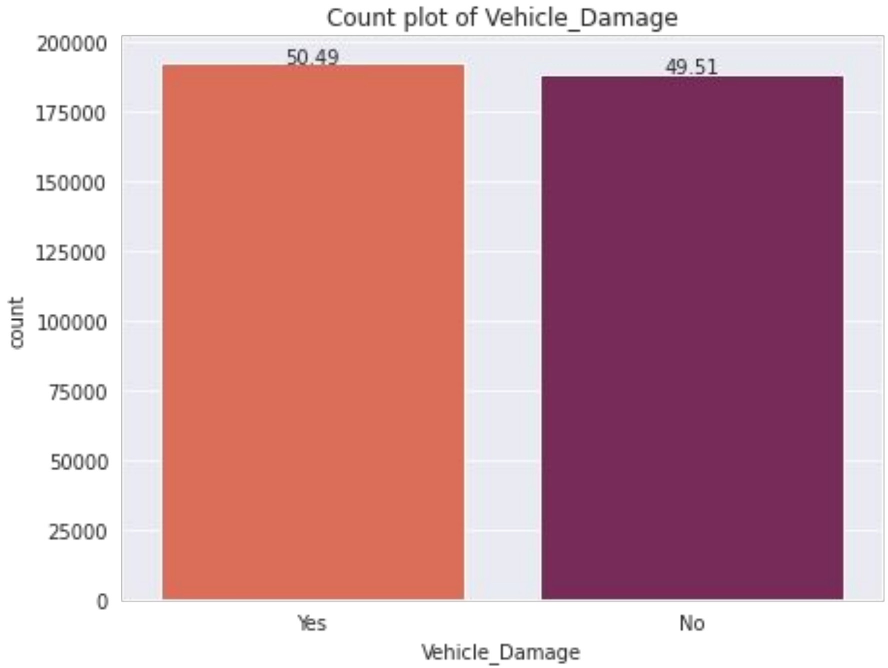
- Customers, with Vehicle age greater than 2 years, are 30% likely of buying Vehicle Insurance.
- Customers with Vehicle age between 1 and 2 years are more likely to interested as compared to the other two categories
- Customers with Vehicle age less than 1 year (new vehicles) have very less chance of buying Insurance.

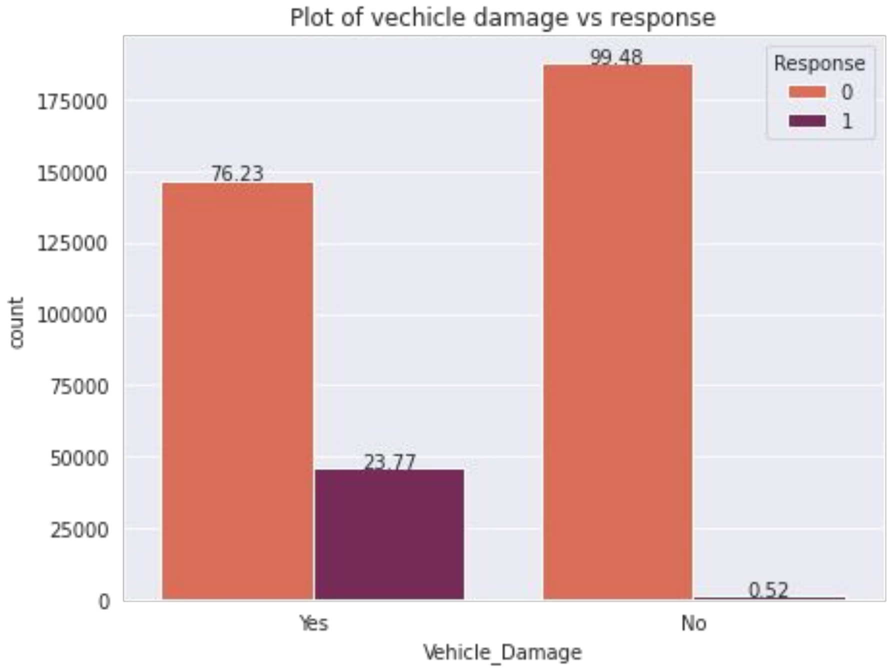




Vehicle Damage Distribution and its effect on Target : Response

- Customers with vehicle damage (Yes and No) are equally distributed with (50.48 % , 49.51 %)
- Customers with no vehicle damage are not interested in Vehicle Insurance





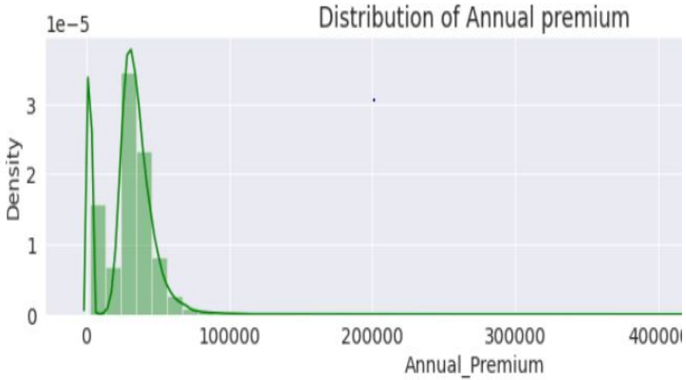
Feature Engineering

Outlier handling for Annual_Premium

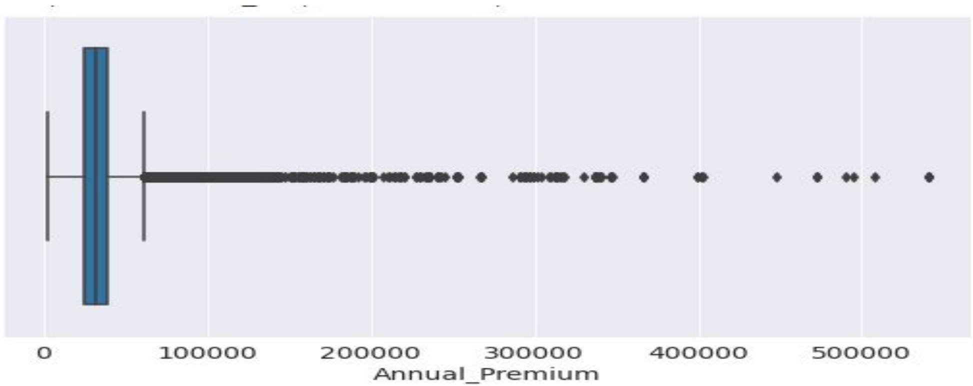
Only Annual_Premium has extreme values :

From the distribution plot, we observed that the annual premium variable is right skewed.

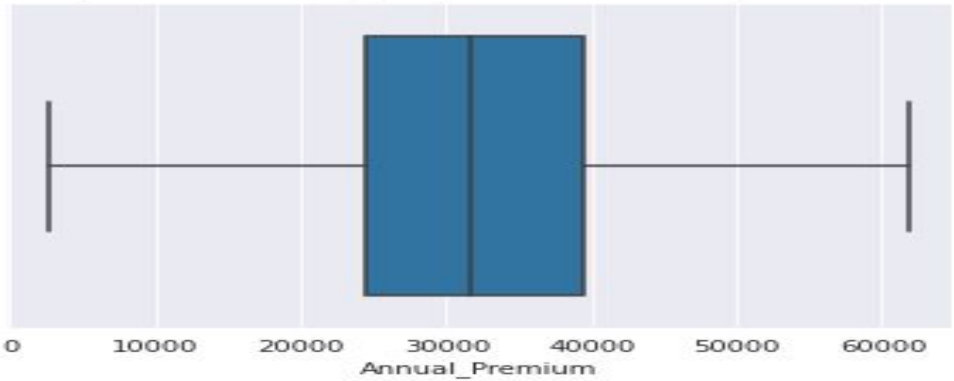
For this, we capped the Extreme Values at Q3 + 1.5xIQR



With Outliers

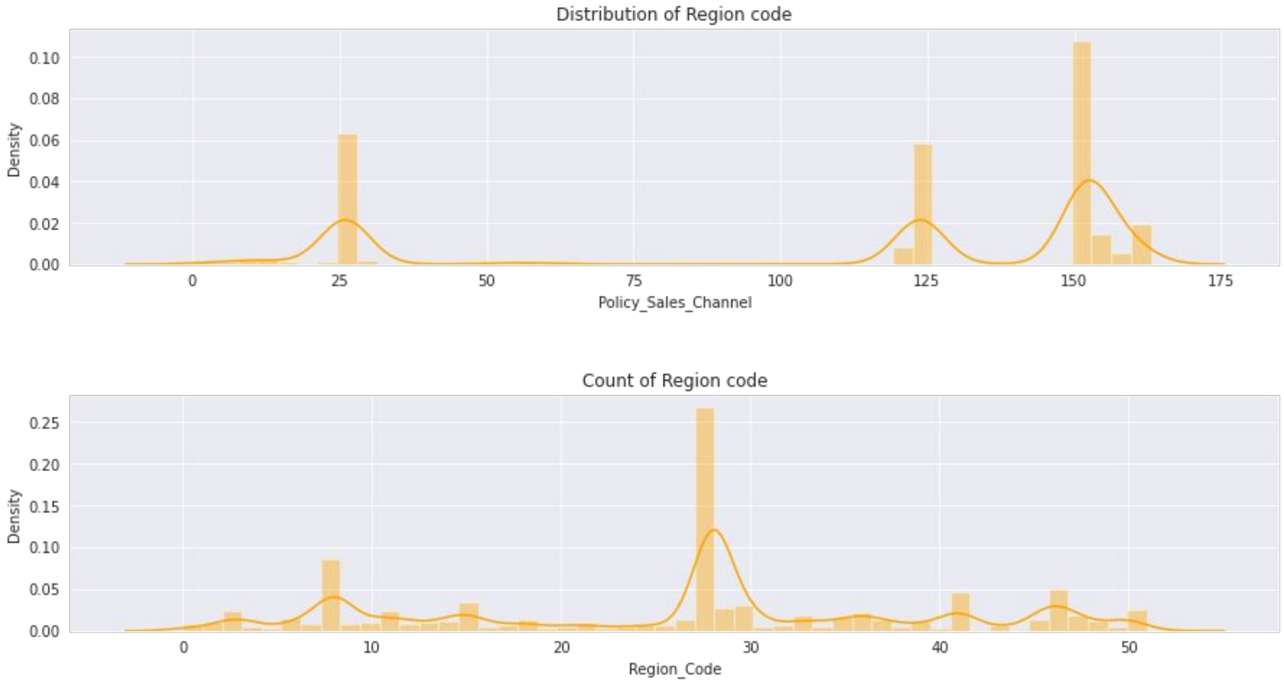


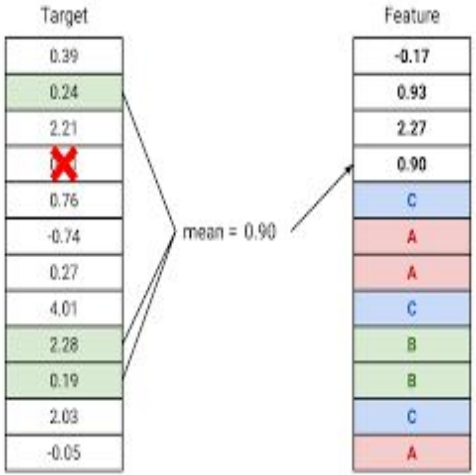
After Outlier Capping



Categorical Variable Target mean encoding:

for the categorical variables : Policy_Sales_Channel and Region_Code.





Machine Learning Modelling

Baseline Algorithms

- KNN
- Logistic Regression

High Performance Algorithms

- Random Forest
- Xgboost
- CatBoost

Metrics Used

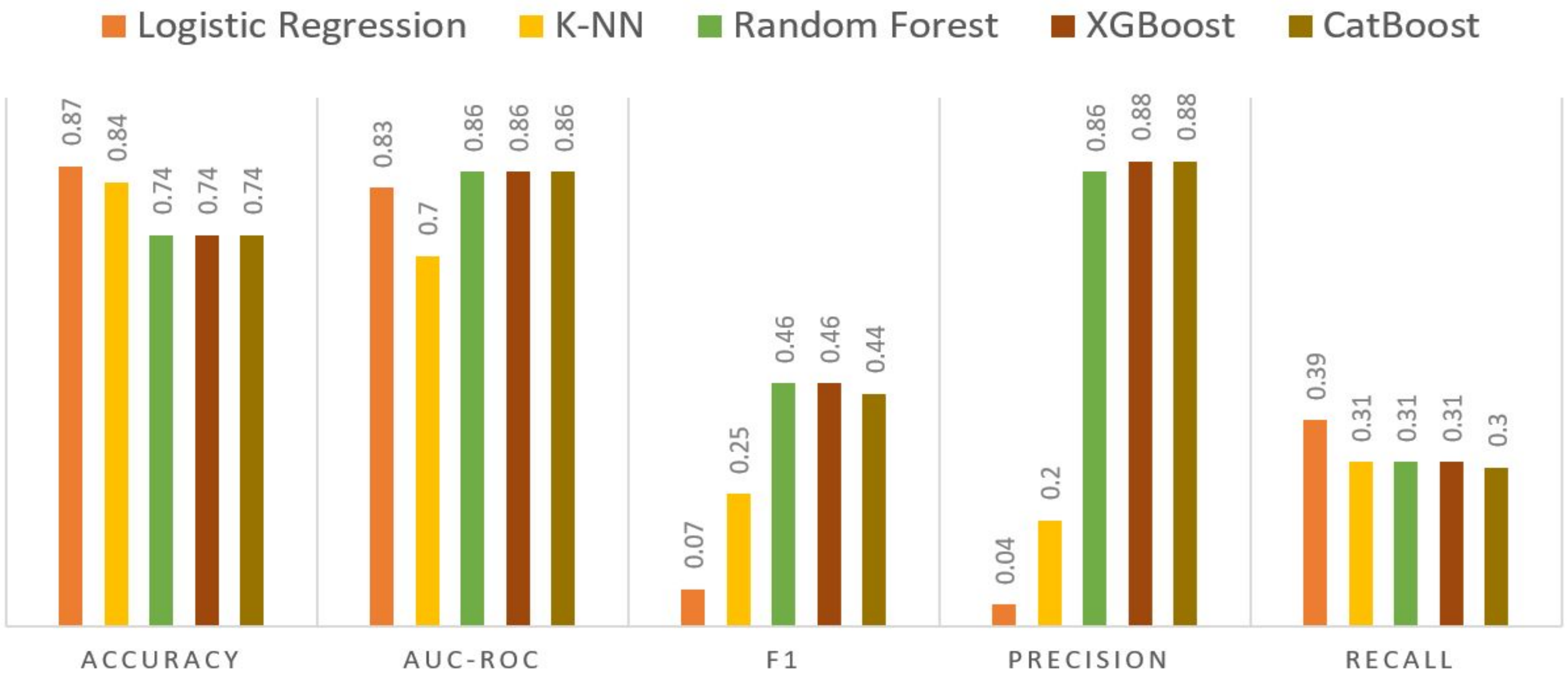
- **F1**-score (Main Metric)
- **Accuracy**
- **Precision**
- **Recall**
- **AUC-ROC** (Area Under Curve - Receiver Operator Characteristics)
- **AUC-PRC** (Area Under Curve - PR Curve / Average Precision)

Hyper-Parameter tuning

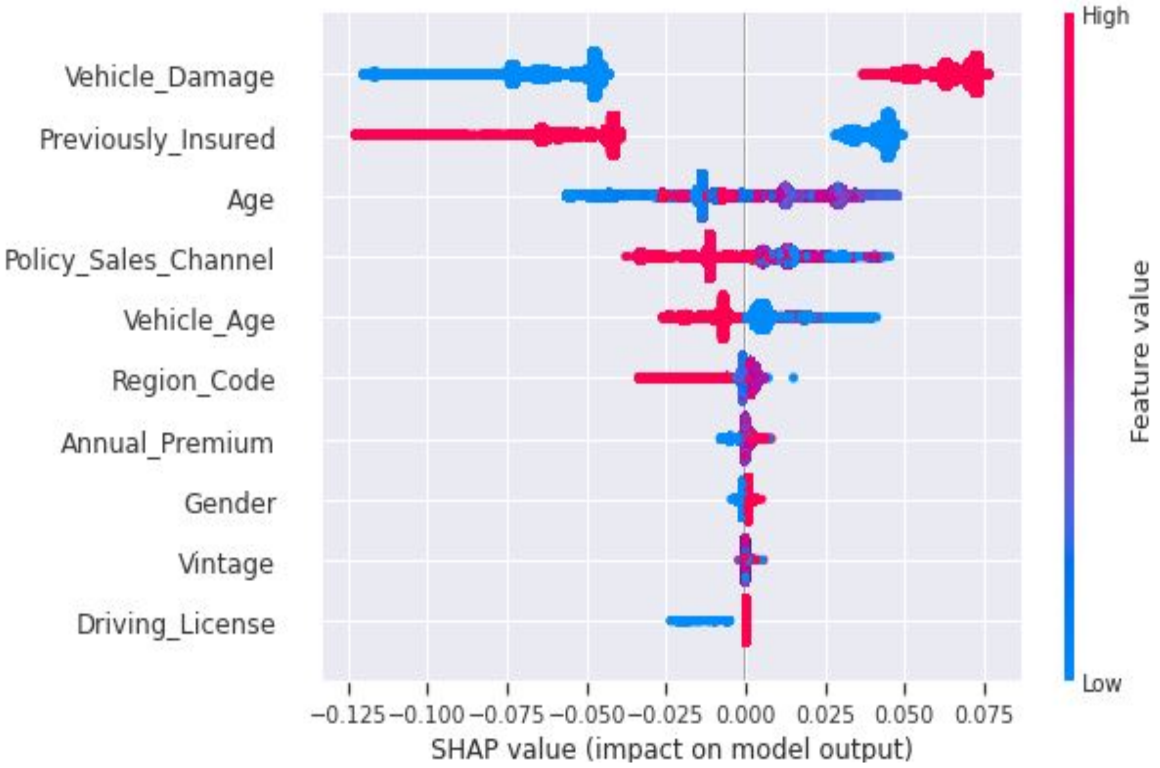
Hyperparameter tuning using **GridSearchCV** and **BayesSearchCV** helped in getting the best out of each algorithm

Final tuning results

which is the best performing model and why



Feature importance



By shap model interpretation, Important features are:

- 'Vehicle Damage',
- ' Vehicle Age',
- 'Previously Insured',
- 'Policy Sales Channel',
- 'Region code'

Inferences

- Customers of **age between 30 and 70** are **more likely to buy** insurance.
- Customers with **Driving Licence** have **higher chance of buying** Insurance.
- Customers with **Vehicle Damage** are **more likely** to buy insurance.
- Age, Previously_insured, Annual_premium are having a large predictive power.
- Comparing ROC Score , we can see that **XGBoost model** performs the best .
- Customers with **Vehicle age between 1 and 2 years** are **more likely** to interested.
- Customer **who are not insured previously** are **more likely** to be interested.

What Worked?

- **Hyperparameter tuning** using **GridSearchCV** and **BayesSearchCV** helped in getting the best out of each algorithm .
- Feature Engineering such as **Target Mean Encoding for Sparse Categorical Values** helped retain useful information in the column , without needing One-Hot encoding which would lead to the Curse of Dimensionality and Severe Overfitting .
- **CatBoost** performed great without extensive Feature Engineering
- **XGBoost** and **RandomForest** have a similar performance of **0.44 F1-Score** and **0.86 AUC** , while they have a good Recall , they suffer from poor Precision .
    - this is tolerable because it is better to make a few extra calls (**False Positives**) , but its more harmful to lose even one potential customer (**False Negatives**)

What didn't Work?

- **Class balancing** via oversampler , undersampler , **SMOTE** was tried in the initial stages but had a detrimental effect on the Model Performance .
- **Logistic Regression ,** which assumes a linear relationship ,

  It did not capture the Variance

  and **severely underfitted** the Dataset with a very **poor Recall**
- **KNN performed poorly** as was expected , in an effort to increase the Recall , the Precision took a hit ,

  and the best F1 Score was at **3 Neighbours** , which implies **severe Bias**

Any Questions