Health Insurance Cross Sell Prediction Model Building

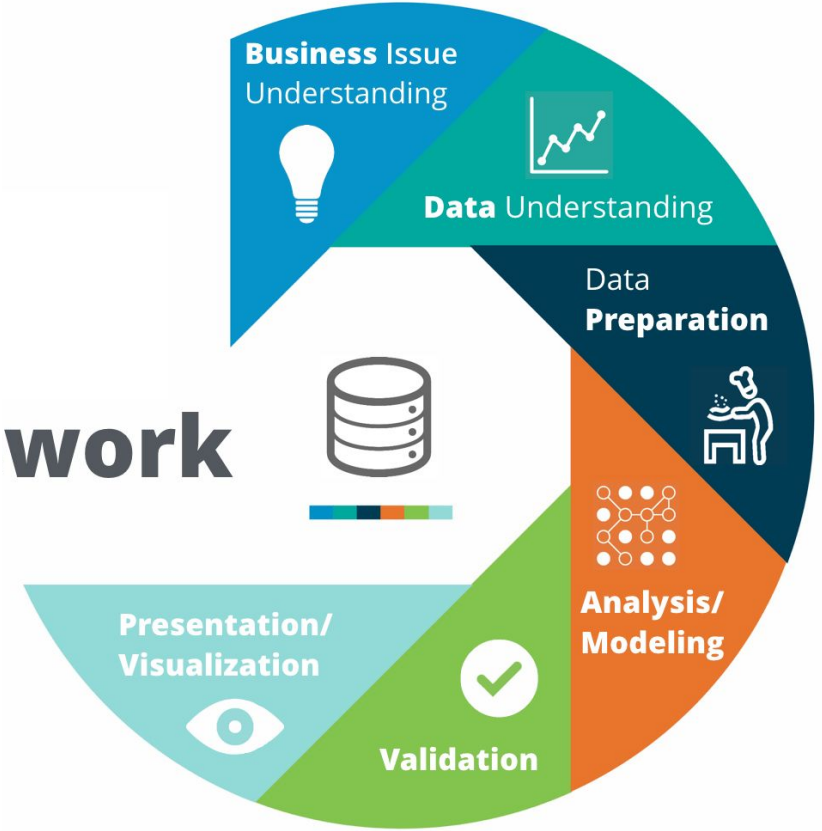
Sarvesh Kumar Yadav
Email- sarveshy995@gmail.com
+91-8826685572

Under the guidance of
Dr. Gautam Panighahi
Department of Mathematics
NIT Durgapur

Problem Statement

Our client is an Insurance company that has provided Health Insurance to its customers,
and now they need a model to predict whether the policyholders (customers) will also be interested in Vehicle Insurance provided by the company.

CRISP-DM Framework :



Exploratory Data Analysis

Understanding the Data

DATASET NAME:

SHAPE:

TARGET VARIABLE:

MISSING DATA CHECK:

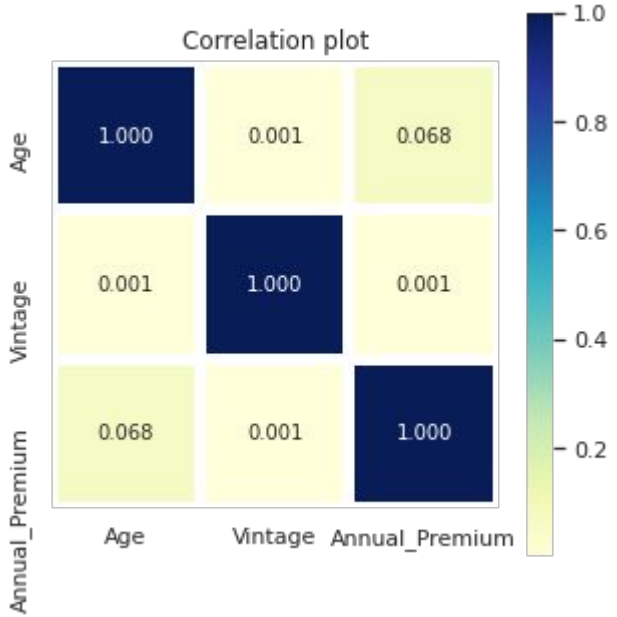
FEATURE:

```
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

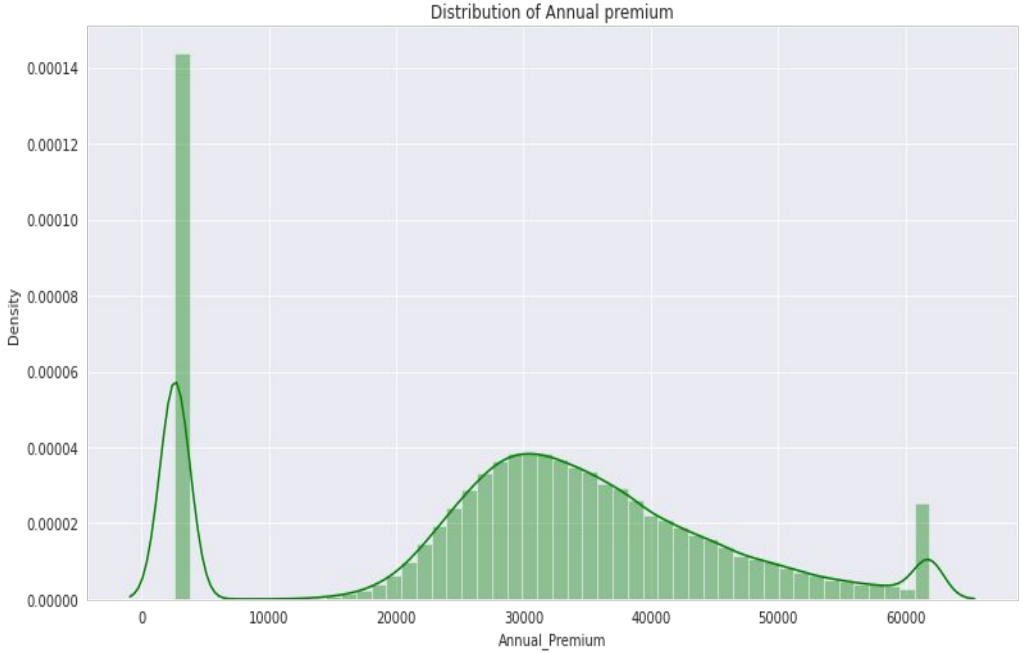


Feature	min	10%	25%	50%	Mean	75%	95%	99%	max
Age	20	22	25	36	38.82	49	69	77	85
Vintage	10	38	82	154	154.35	227	285	297	299
Annual_Premium	2630	2630	24405	31669	30564	39400	55176	72963	540165

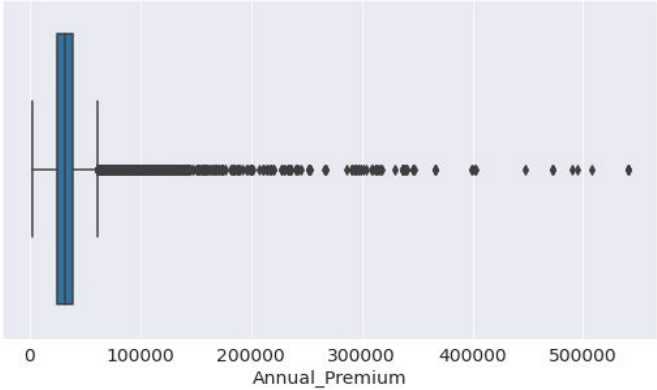
Outlier handling for Annual_Premium

Only Annual_Premium has extreme values :

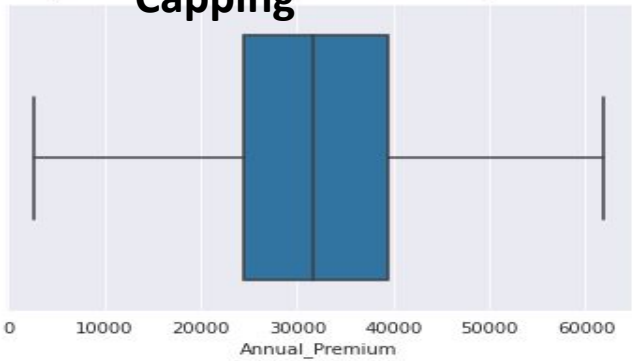
From the distribution plot, I observed that the annual premium variable is right skewed .



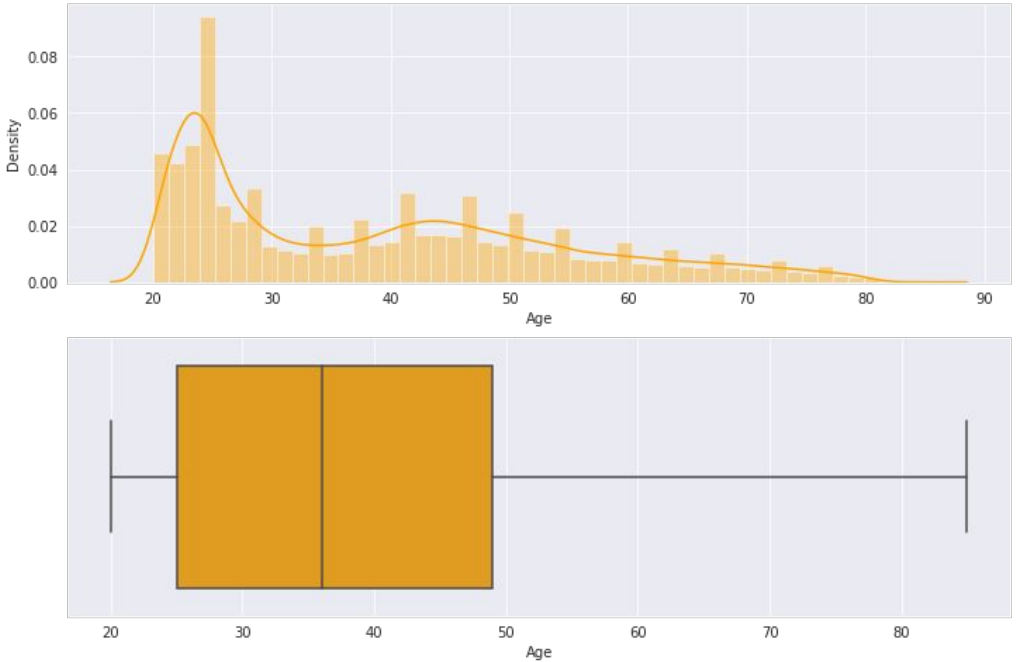
With Outliers

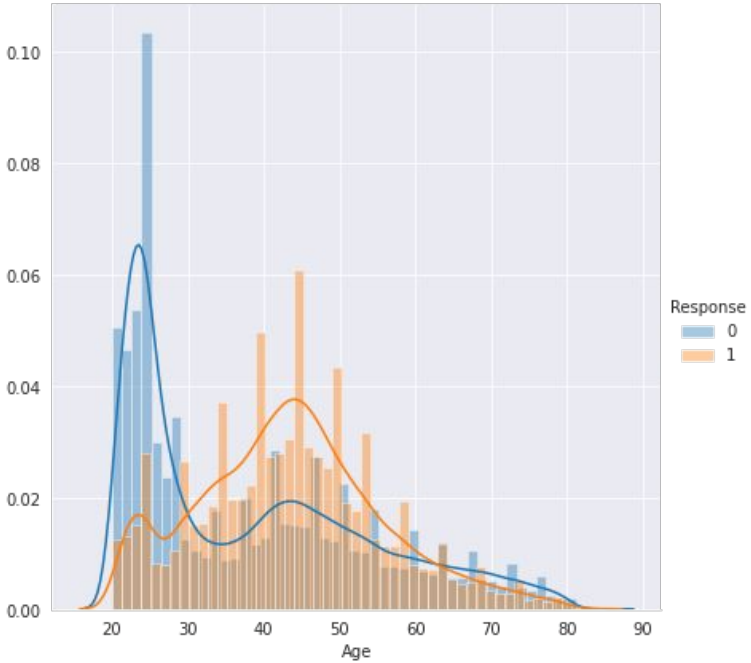


After Outlier Capping



Age Distribution and its effect on Target Variable : Response



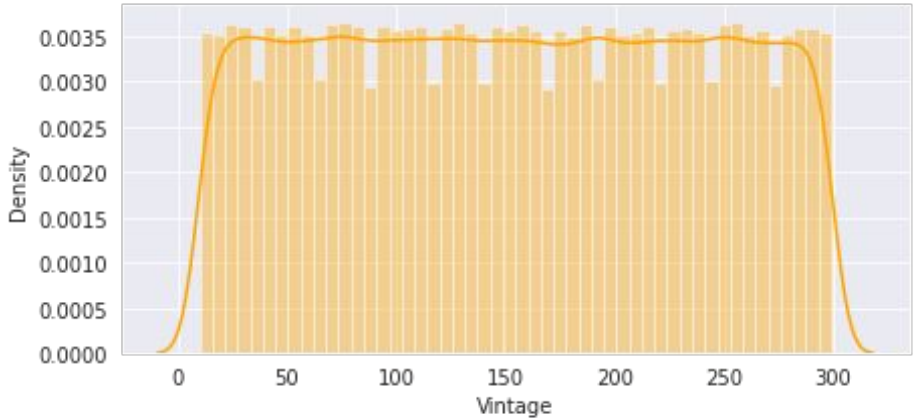


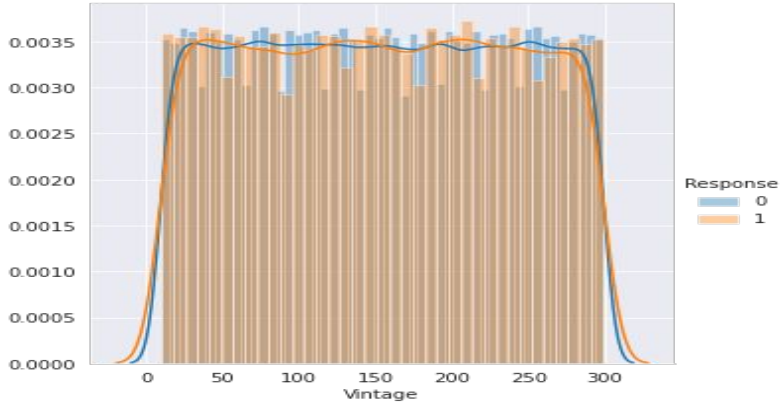
Distribution of Vintage

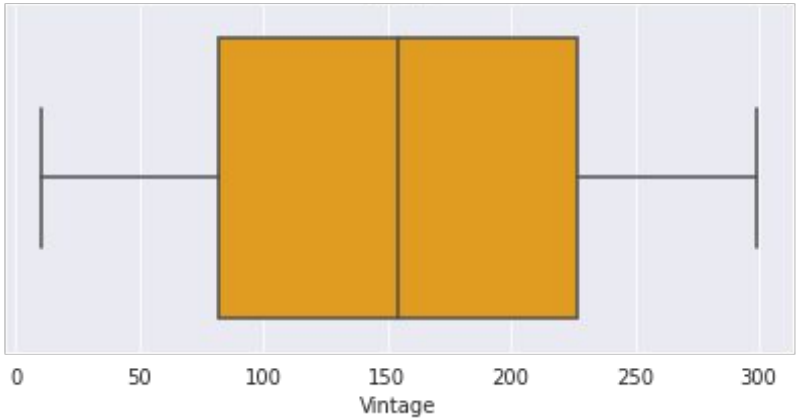
The Feature Vintage has very less information
and is Uniformly Distributed , With no skew .
Also, the Values are uniformly mixed ,
in both the classes of the target variable response .

This was also confirmed with feature importance given by Random Forest and XGBoost

Since this Feature potentially contribute to Over Fitting ,
Or it can also contain hidden information (nonlinear , conditional states) we need to analyze the feature_importances for this feature and decide whether to retain it or not

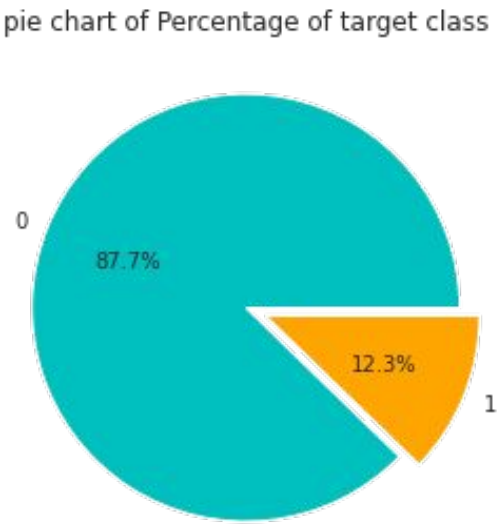






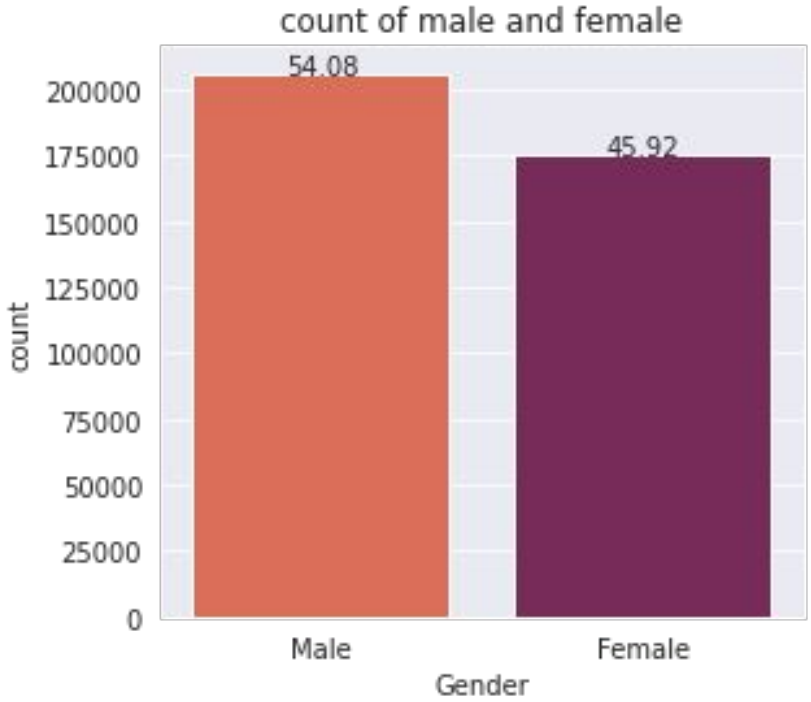
Categorical Features

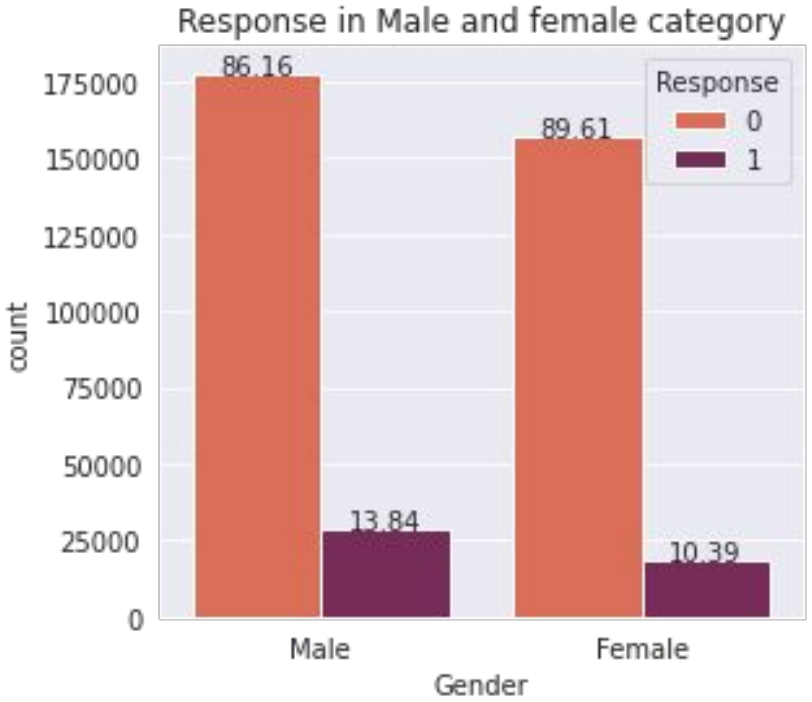
Features	# Categories	Top	% Frequency
Region_Code	53	28	28%
Policy_Sales_Channel	155	152	35%
Vehicle_Age	3	1-2 Year	53%
Gender	2	Male	54%
Driving_License	2	1	100%
Previously_Insured	2	0	54%
Vehicle_Damage	2	Yes	50%



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%.
  This means Males are 30% more likely than females to buy insurance

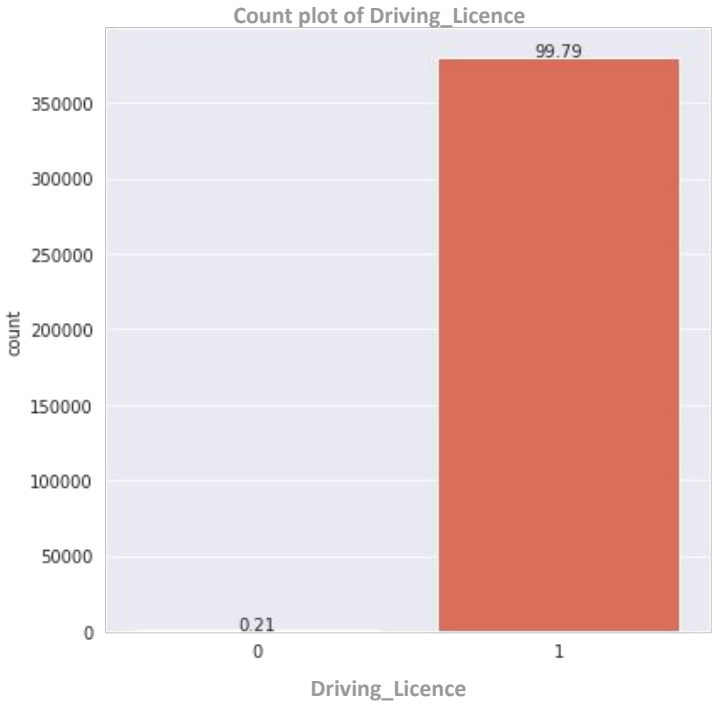


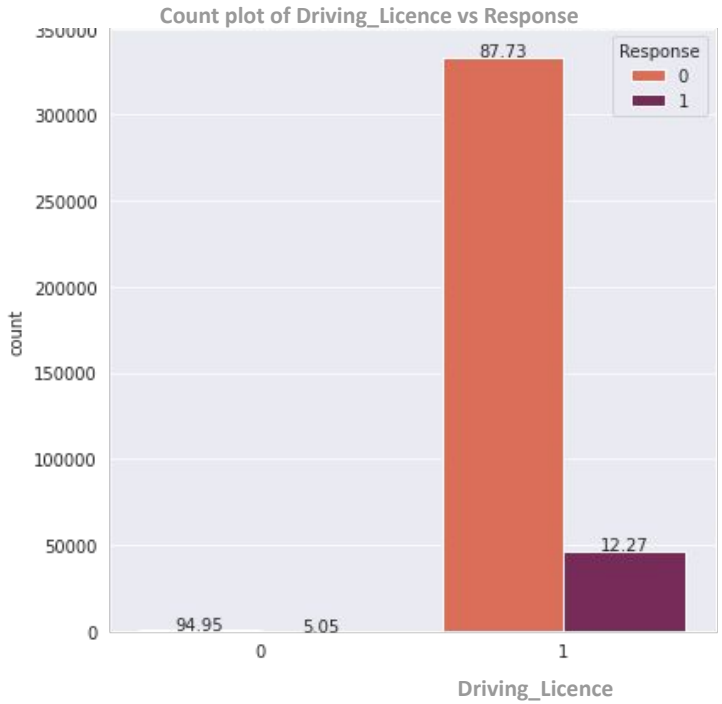


Driving Licence Distribution and its effect on Target Variable : Response

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

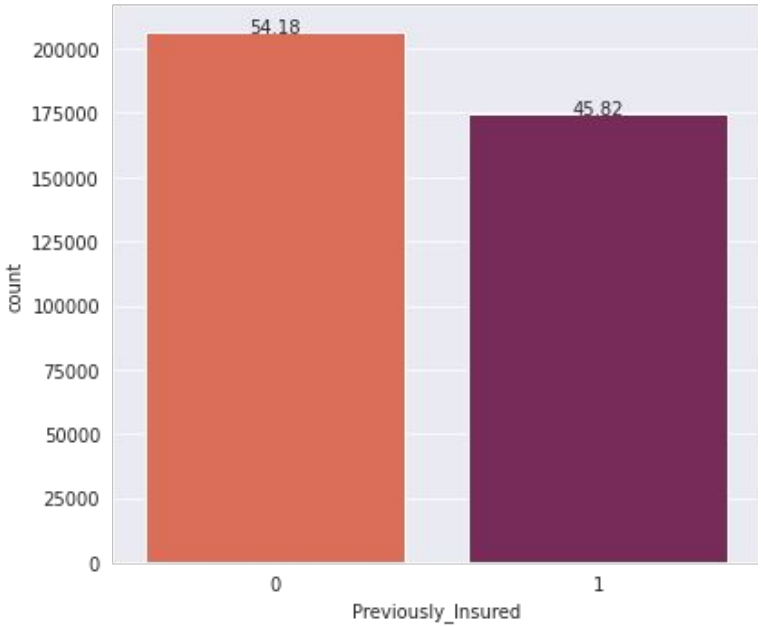


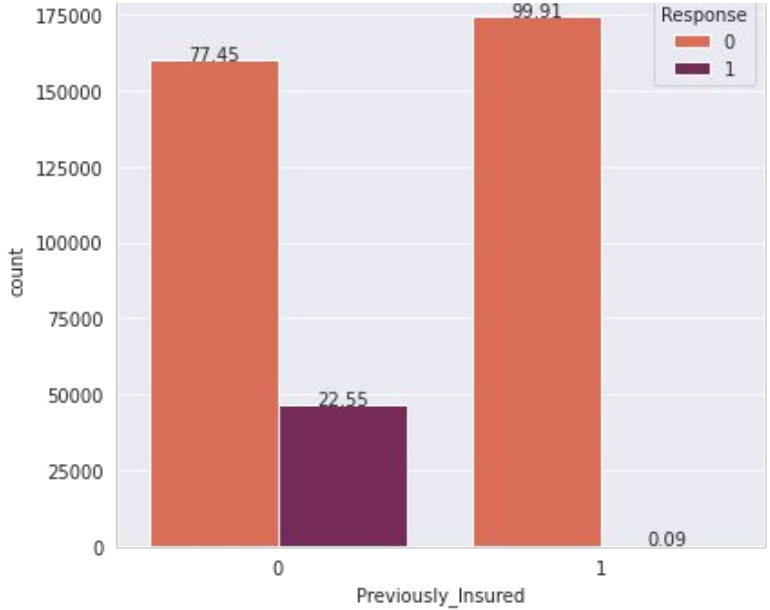


Previously_Insured Distribution and its effect on Target : Response

Customers who were previously insured tend not to be interested.

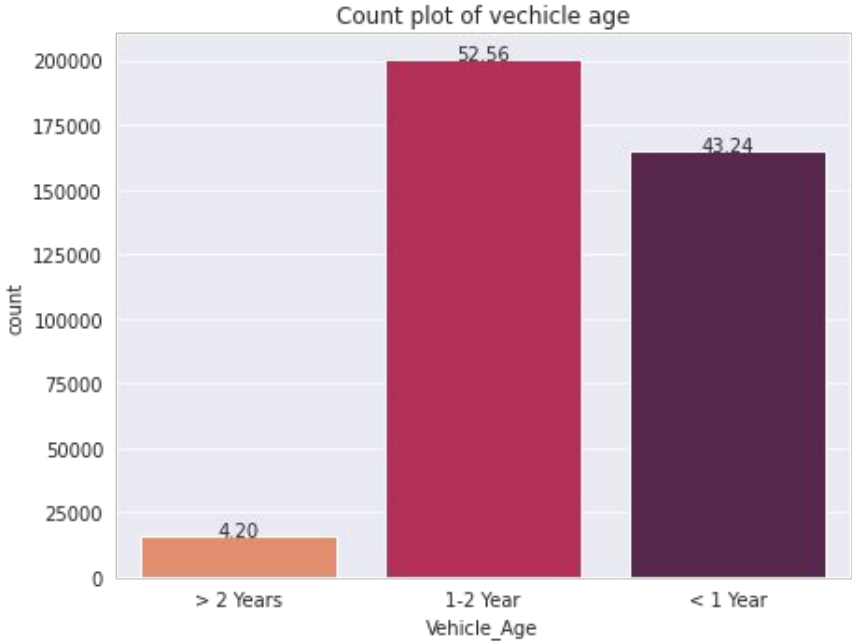
- We can think that the reason for this is that their previous insurance agreement has not expired yet
- Or maybe they are unsatisfied with previously purchased insurance services

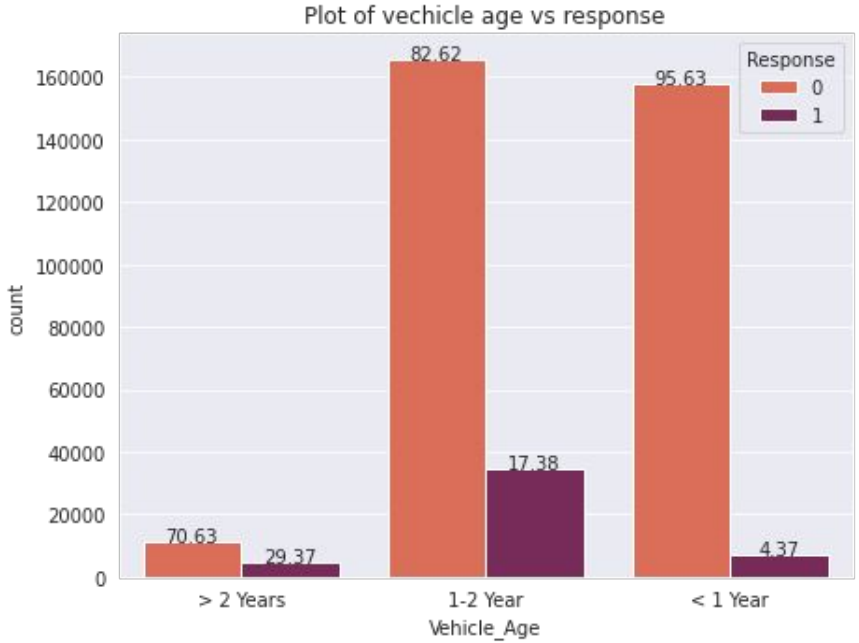




Vehicle Age Distribution and its effect on Target : Response

- 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.

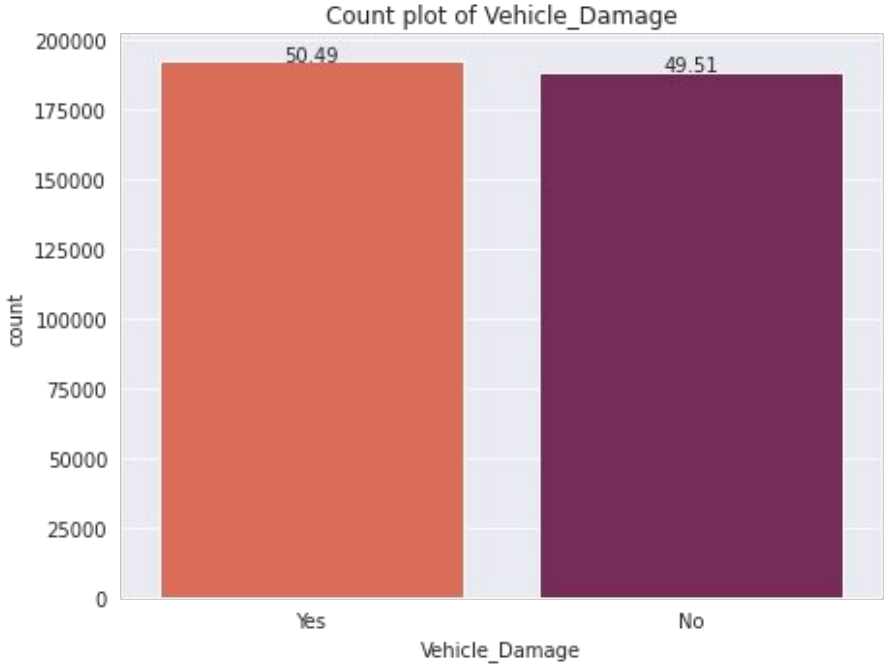


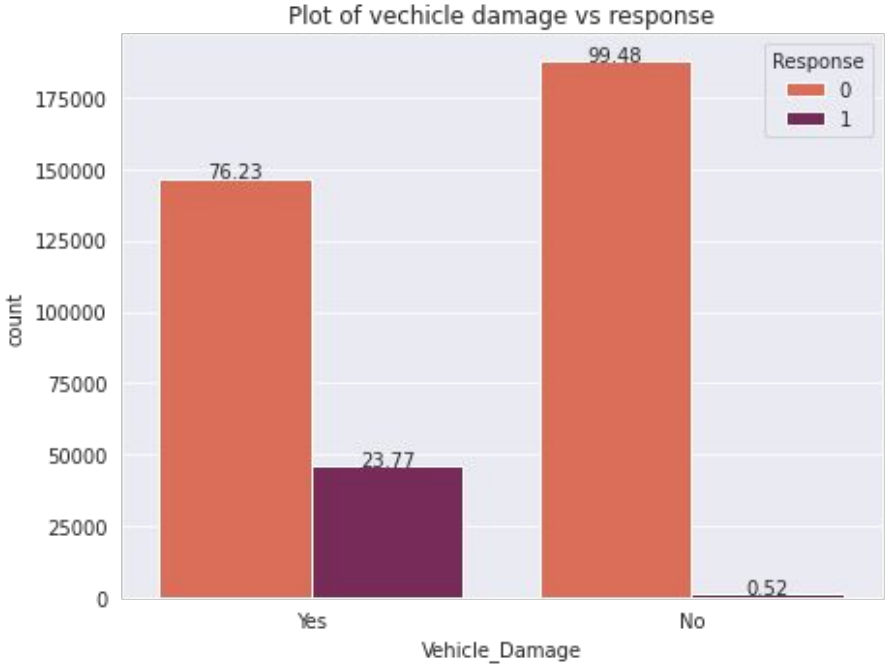


Vehicle Damage Distribution and its effect on Target : Response

Customers 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

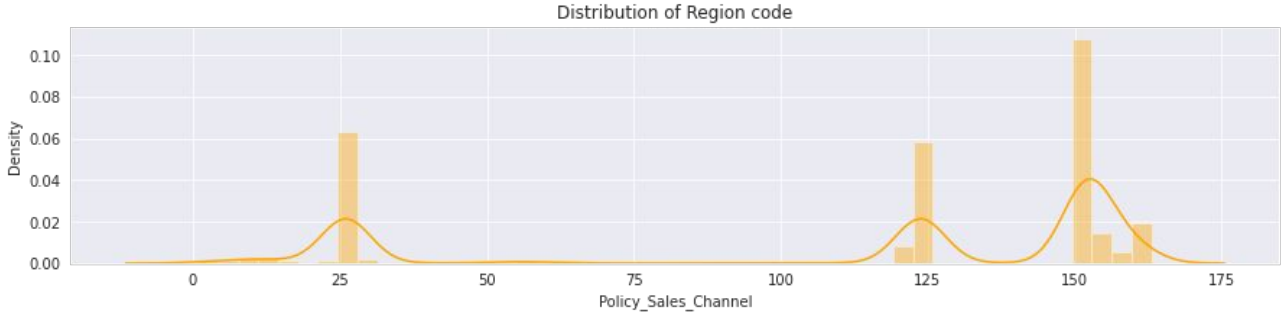


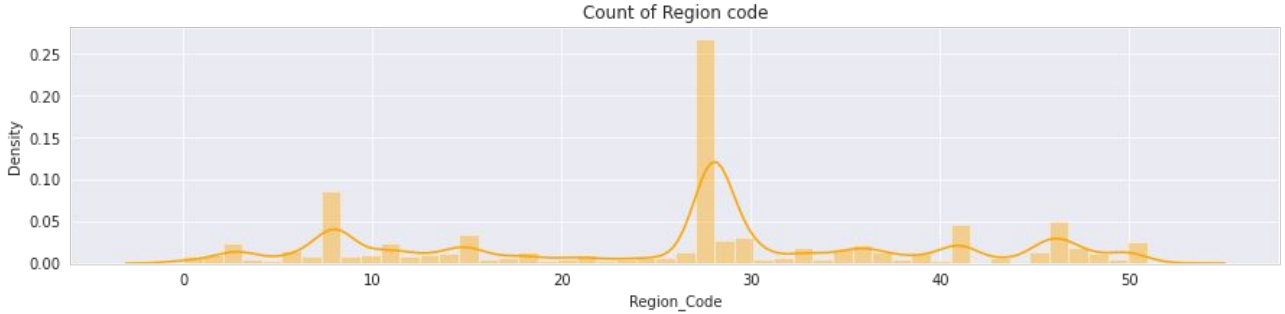


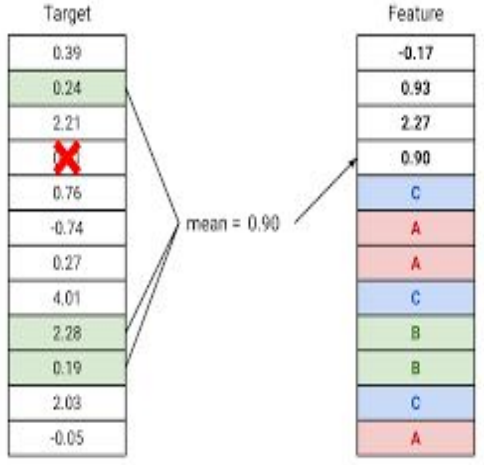
Feature Engineering

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

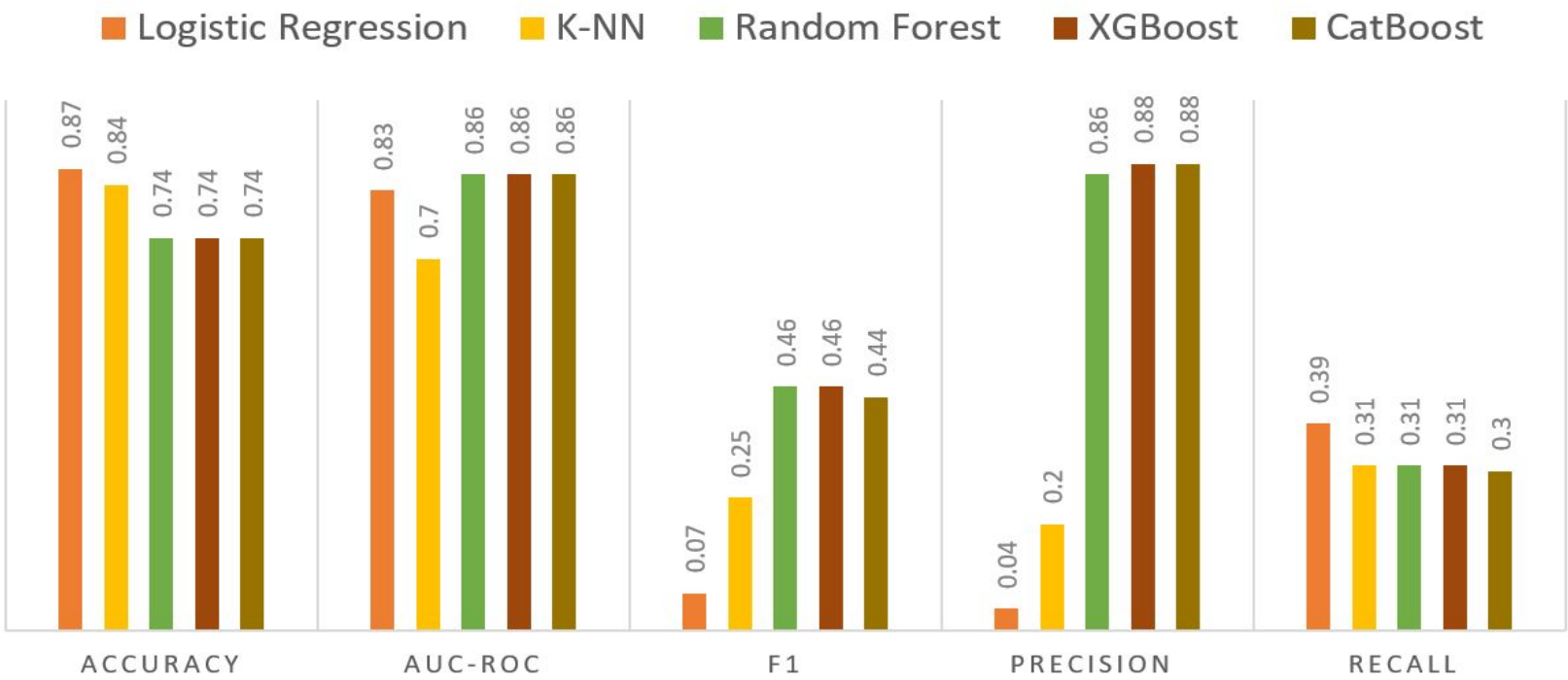
- **F1**-score ( )
- **Accuracy**
- **Precision**
- **Recall**
- **AUC-ROC** (Area Under Curve - Receiver Operator Characteristics)
- **AUC-PRC** (Area Under Curve - PR Curve / Average Precision)

Hyper-Parameter tuning

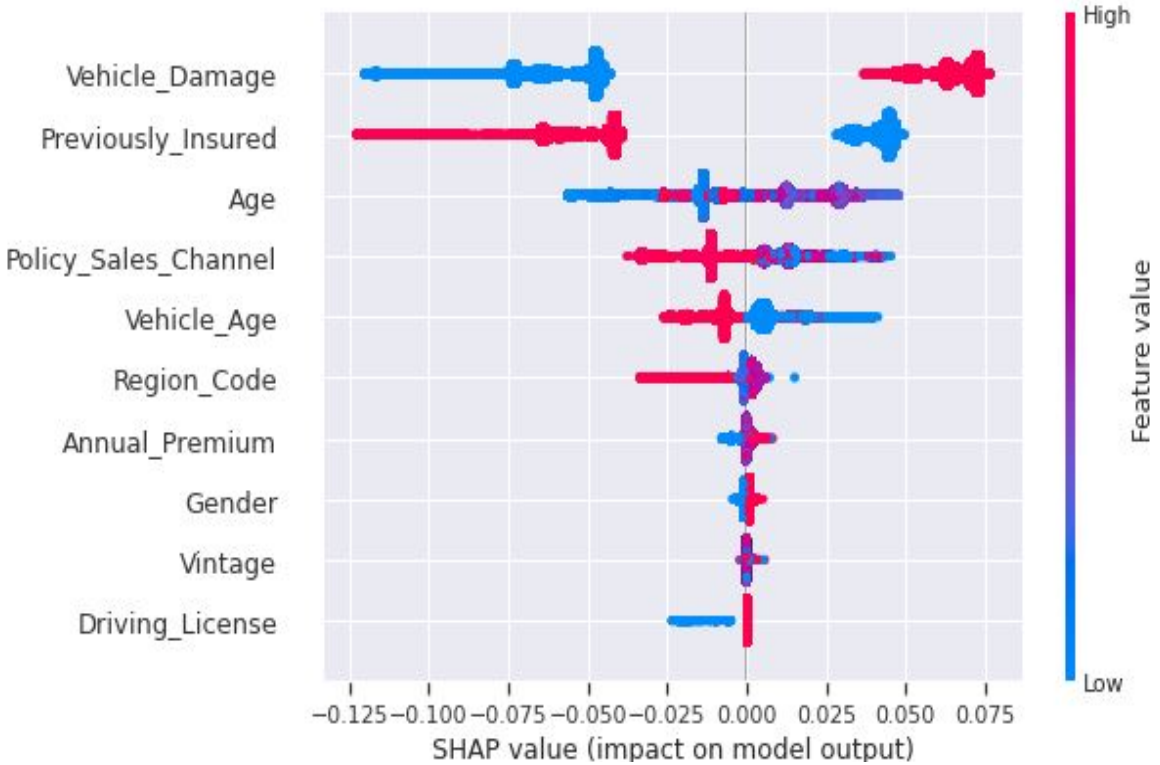
Hyperparameter tuning using **GridSearchCV** and **BayesSearchCV** helped in getting the best out of each algorithm

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’

Conclusion and 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 , **Did not capture the Variance** and **severely underfitted** the Dataset with a very **poor Recall**
- **KNN Failed** 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 means **severe Bias**

Future Scope:

1. It can be easy for marketers to forget about their previous customers while they're putting all of their blood, sweat, and tears into generating new interest and getting more leads to enter the funnel. But sometimes the best opportunities are the ones you've already created.
   Running efficient advertising campaigns to re-convert past customers is often the best use of your working hours. And while it might seem counter-intuitive to re-invest in the same individuals, running paid campaigns for cross-sell/upsell purposes can increase each customer's lifetime value and average order value, increasing your revenue. That's not time wasted!
1. It increases overall customer satisfaction.
2. It builds great relationship with the customers.
3. It is easiest way of marketing, which involves very budget

Research Area:

Tackling imbalanced data:

We may distinguish three main approaches to learning from imbalanced data:

- **Data-level methods that modify the collection of examples to balance distributions and/or remove difficult samples.**
- **Algorithm-level methods that directly modify existing learning algorithms to alleviate the bias towards majority objects and adapt them to mining data with skewed distributions.**
- **Hybrid methods that combine the advantages of two previous groups.**

References:

1. https://www.kaggle.com/search
2. https://docs.microsoft.com/en-us/azure/architecture/data-science-process/overview
3. https://www.lexjansen.com/nesug/nesug07/sa/sa16.pdf
4. https://ijesc.org/upload/c92ccd5cffe580f4fe0935ee520a73c4.Right%20Product%20to%20the%20Right%20Customer%20using%20Cross%20Selling%20in%20Insurance.pdf
5. https://link.springer.com/article/10.1007/s13748-016-0094-0
6. https://www.researchgate.net/publication/271757750_Cross-Selling
7. https://research-repository.uwa.edu.au/files/3218934/Ferguson_Graham_2009.pdf
8. https://www.sciencedirect.com/science/article/abs/pii/S1094996802701631
9. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html
10. https://catboost.ai/docs/concepts/python-reference_catboostclassifier.html
11. https://numpy.org/
12. https://pandas.pydata.org/
13. https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
14. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

Thank You!