CAPSTONE PROJECT

HEALTH INSURANCE CROSS SELL PREDICTION

TEAM MEMBERS

ANKIT KUMAR
FAHAD MEHFOOZ
SANJOG MISHRA
VARUN NAYYAR

CONTENTS

- INTRODUCTION
- PROBLEM STATEMENT
- METHODOLOGY
    1. DATA ACQUISITION
    2. EXPLORATORY DATA ANALYSIS AND DATA CLEANING
    3. TRAIN - TEST SPLIT
    4. DATA PREPROCESSING
    5. FEATURE ENGINEERING
    6. DATA MODELLING
    7. MODEL INTERPRETATION
    8. TRAIN AND TEST INTERPRETATION
- CONCLUSION
- CHALLENGES FACED
- FUTURE SCOPE OF WORK

INTRODUCTION

Insurance Companies are entities which provide insurance which is a form of risk management, primarily used to hedge against the risk of a contingent or uncertain loss. In this arrangement, the insurance company or the insurer provides a guarantee of compensation for specified loss, damage, illness, or death in return of payment of specified premium. A premium is a sum of money that the customer pays regularly to an insurance company for this guarantee.

Vehicle Insurance is one type of insurance in which the customer pays a premium to an insurance company so that in case of unfortunate accident by the vehicle, the insurer will provide a compensation to the customer.

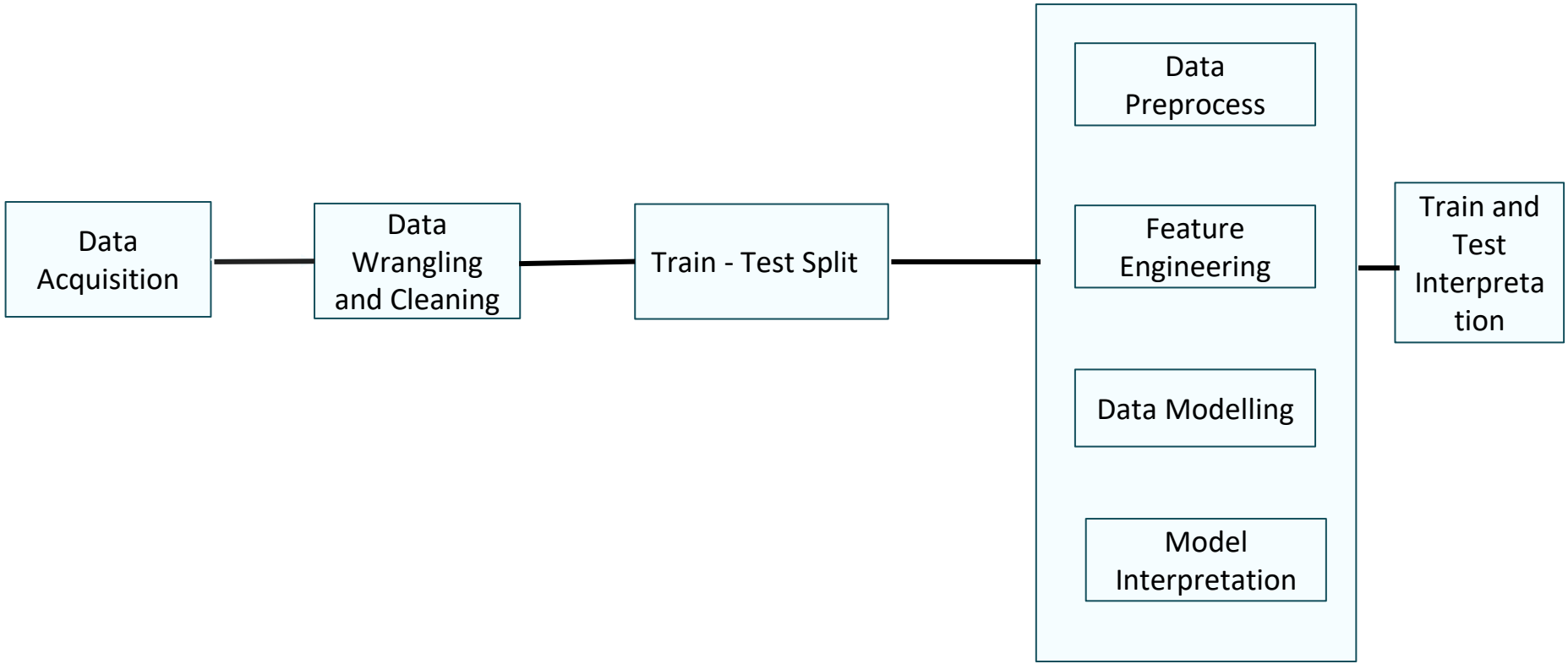
PROBLEM STATEMENT

We have an insurance company as our client that has provided Health Insurance to its customers. Now they want a model to predict whether the policyholders from last year will be interested in Vehicle Insurance provided by the company. Such a model will be extremely helpful to the company because it can then accordingly plan its communication strategy to reach out to those customers and optimise its business model and revenue.

We have been provided with information of customers of the company broadly on following categories:

- **Demographics** - Gender, Age, Region Code Type
- **Vehicles** - Vehicle Age and Damage
- **Policy** - Annual Premium, Previously Insured Status and Sourcing Channel

METHODOLOGY



DATA ACQUISITION

The first step in the pipeline aims at importing our dataset into our environment. In our case; we will import the dataset provided by the insurance company.
The dataset consists of about 390K rows and 12 columns including our target feature.

EXPLORATORY DATA ANALYSIS

First we will have a preliminary look on the information provided by our dataset.

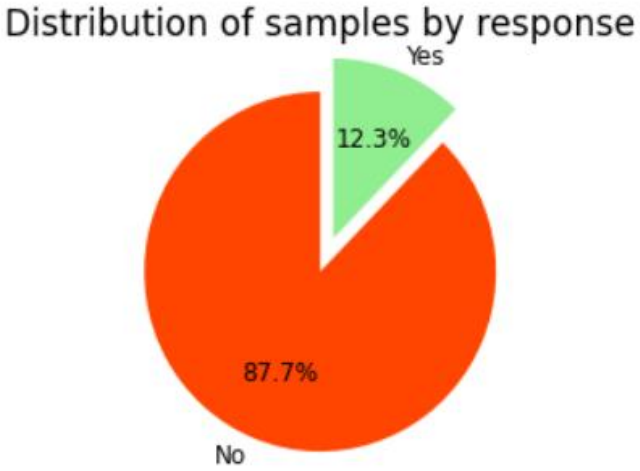
```
id                    381109 non-null  int64
Gender                381109 non-null  object
Age                   381109 non-null  int64
Driving_License       381109 non-null  int64
Region_Code           381109 non-null  float64
Previously_Insured    381109 non-null  int64
Vehicle_Age           381109 non-null  object
Vehicle_Damage        381109 non-null  object
Annual_Premium        381109 non-null  float64
Policy_Sales_Channel  381109 non-null  float64
Vintage               381109 non-null  int64
Response              381109 non-null  int64
```

1. We observe that we have 381109 entries in our dataset and none of the columns contain any null values.
2. Further it is also checked that there are no duplicate rows in our dataset which means that we have data on 381109 unique customers.
3. Each column has all its entries of the same type as described by the column itself which means that there are no miscellaneous data in any column.
4. Now let us take a look at the description of the numerical columns which consist of ‘Age’ , ‘Annual Premium’ and ‘Vintage’.

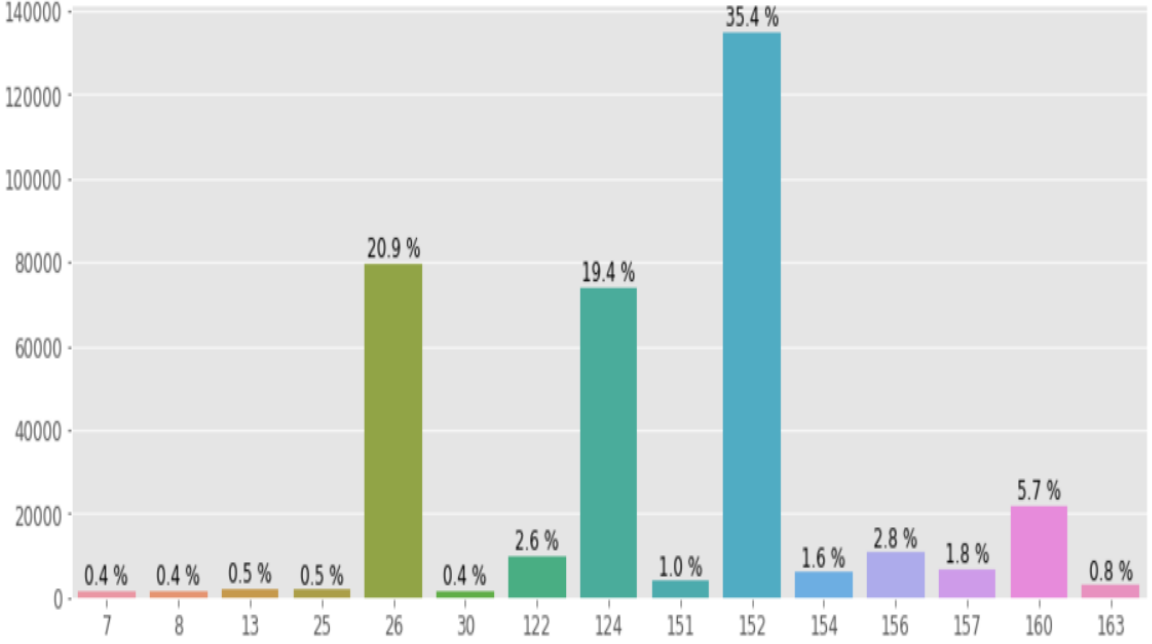
	count	mean	std	min	25%	50%	75%	max
Age	381109.0	38.822584	15.511611	20.0	25.0	36.0	49.0	85.0
Annual_Premium	381109.0	30564.389581	17213.155057	2630.0	24405.0	31669.0	39400.0	540165.0
Vintage	381109.0	154.347397	83.671304	10.0	82.0	154.0	227.0	299.0

UNIVARIATE ANALYSIS

Following are the important observations we obtained from Univariate Analysis



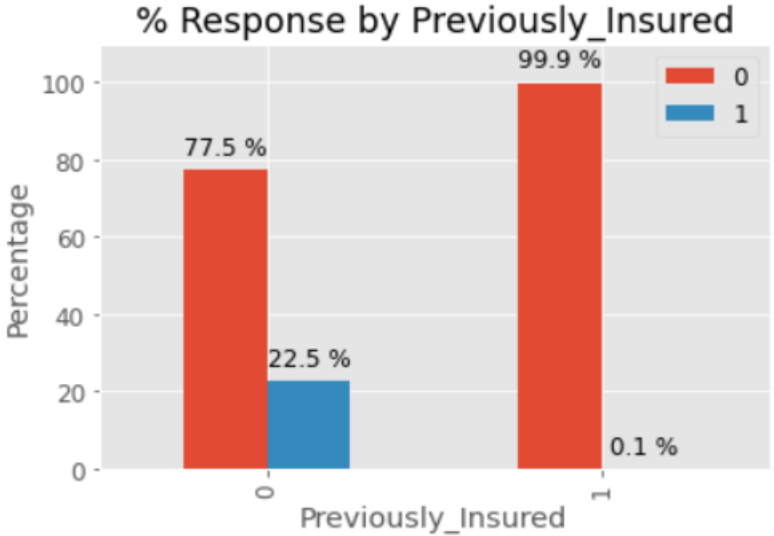
Most of the customers in our dataset are not interested in Vehicle Insurance indicating that customers opting for Vehicle Insurance are a minority in our dataset.

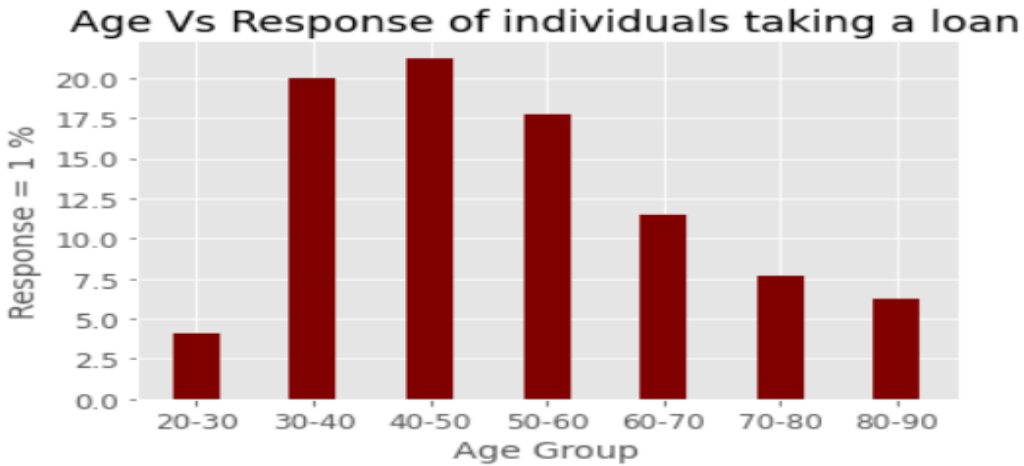


Most of the customers prefer 152, 26, 124 and 160 channels

BIVARIATE ANALYSIS

Following are the important observations from Bivariate Analysis.

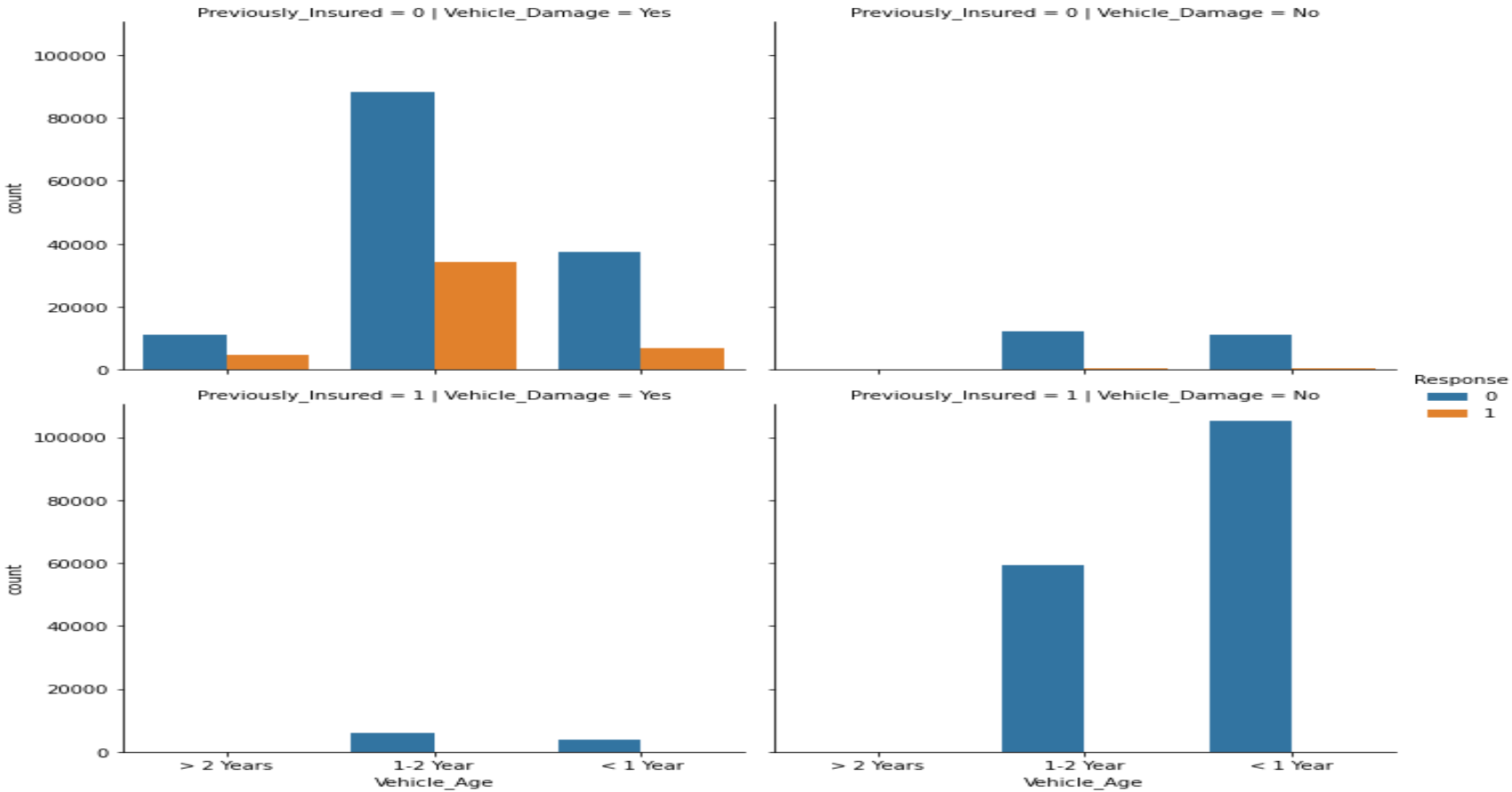




It is inferred that customers who previously are not insured opt for Vehicle Insurance by a far larger margin than customers who are previously insured

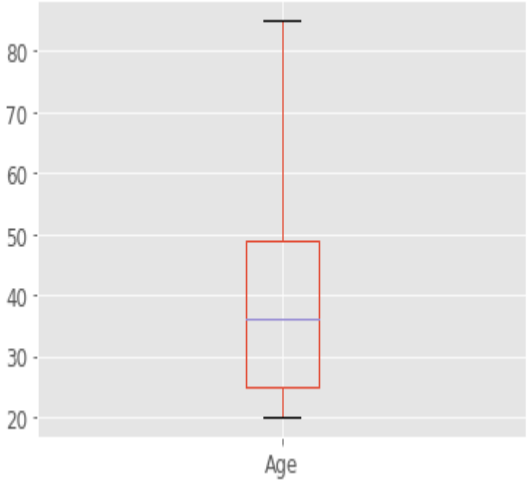
Most of the customers buying Vehicle Insurance are in the age range of 30 - 60 years

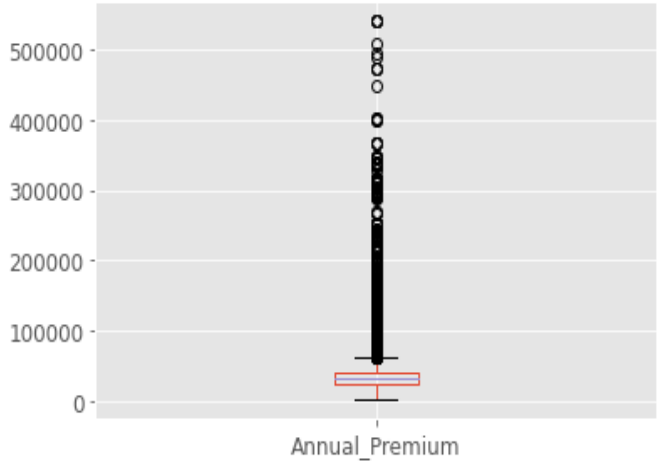
MULTIVARIATE ANALYSIS

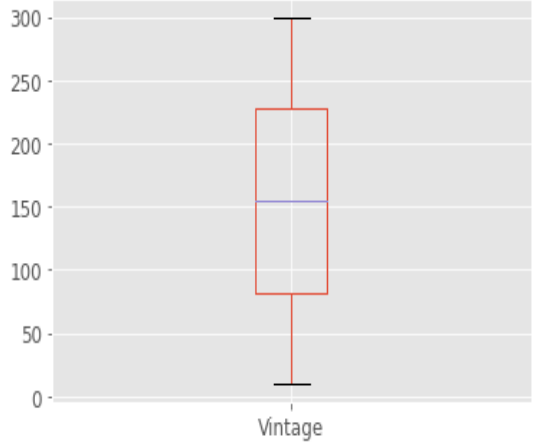


DATA CLEANING

On analyzing the numerical columns in our dataset namely ‘Age’, ‘Annual Premium’ and ‘Vintage’ ; we observe that only Annual Premium consists of outliers.







Applying log transformation to ‘Annual Premium’ column handles the outliers and also in addition gives a normal distribution of the data.

TRAIN TEST SPLIT

After cleaning our data; the dataset is split into Train - Test datasets. This is done to ensure that our test dataset is completely isolated and there is no information leakage during the training process of machine learning models.

DATA PREPROCESSING AND FEATURE ENGINEERING

The following targets are to be achieved by this stage:

- Encoding all categorical columns which cannot be used for modelling
- Removing class imbalance from our dataset where ‘Response’ labelled as 1 is the minority class

The following points describe the encoding process for the categorical columns of the dataset :

1. For ‘Gender’ column; ‘Male’ was labelled as 1 and ‘Female’ was labelled as 0.
2. One-Hot-Encoding was applied to ‘Vehicle Age’ column and three dummy columns were introduced into the dataset corresponding to ‘>2 years’, ‘1-2 years’ and ‘<1 year’.
3. For ‘Vehicle Damage’ column; ‘Yes’ was labelled as 1 and ‘No; with 0.
4. Target Encoding was applied to ‘Region Code’ and ‘Policy Sales Channel’. Target Encoding

To remove class imbalance SMOTE was applied on the training dataset which is an oversampling technique.
Modelling is performed on two datasets; the other being PCA applied; in case we face overfitting.

DATA MODELLING

Taking ROC - AUC as the measure of performance for our classification problem; four models were shortlisted which gave comparable and best results among the classifiers we had adopted.

LGBM Classifier - It is a boosting technique that uses tree based learning algorithm. It grows tree leaf wise rather than level wise.

CatBoost Classifier - It is also a boosting technique that uses tree based learning algorithm.

Stacked Classifier - CatBoost and LGBM are chosen as the base models for the Stacked Classifier which will combine the predictions from the base models.

Voting Classifier - The estimators to be used for voting include CatBoost and LGBM. Also since the number of classifiers is even; the voting is set to ‘soft’.

RESULTS OF DATA MODELLING

MODEL	TRAIN-ROC	TEST-ROC
LGBM	0.869	0.845
CatBoost	0.851	0.845
Stacked Model	0.884	0.839
Voting Model	0.865	0.846

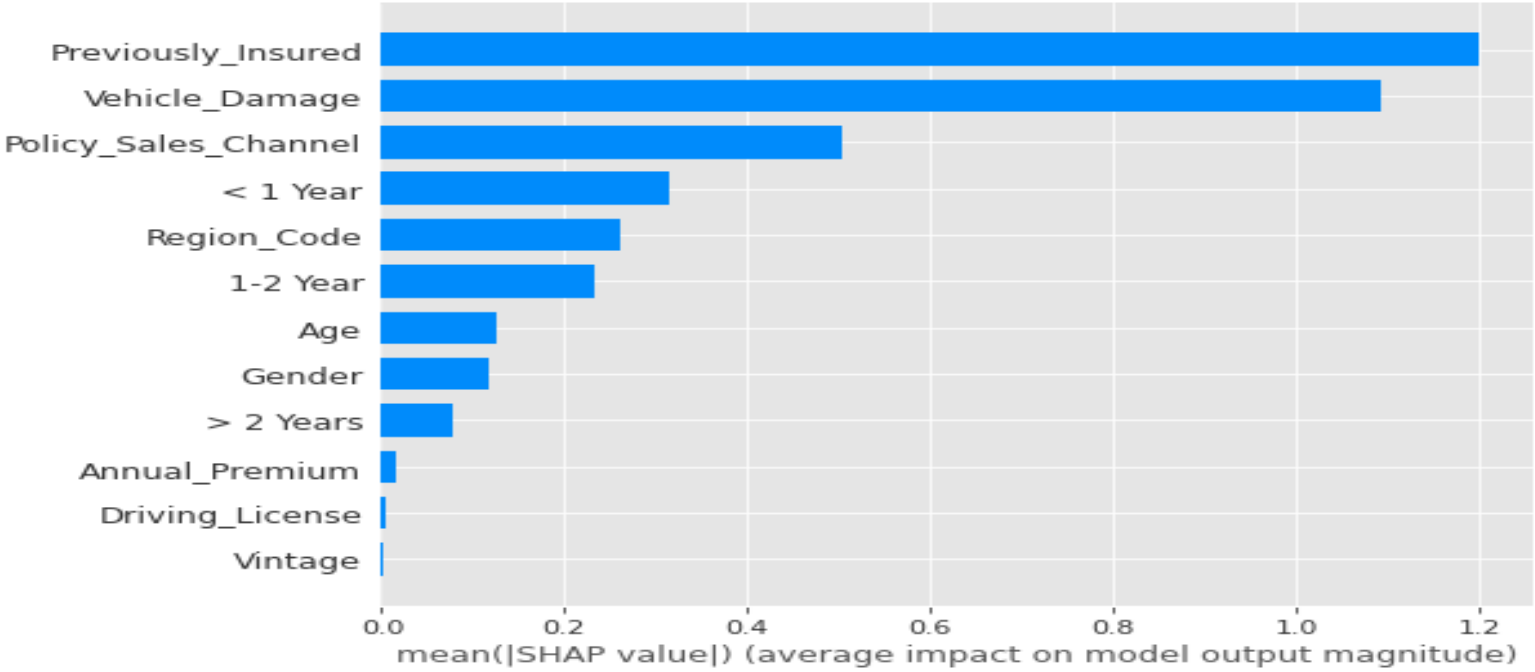
TRAIN AND TEST INTERPRETATION

CatBoost is chosen as the final model for our classification problem owing to best test result and close Train and Test ROC values.

MODEL INTERPRETATION

MODEL INTERPRETATION USING SHAP

SHAP values interpret the impact of having a certain value for a given feature in comparison to the prediction we'd make if that feature took some baseline value.



MODEL INTERPRETATION USING Eli5

Eli5 provides a way to compute feature importances for any estimator by measuring how score decreases when a feature is not available. Permutation Importance is taken as the estimator for calculating the weights.

Weight	Feature
0.0943 ± 0.0013	Vehicle_Damage
0.0847 ± 0.0008	Previously_Insured
0.0541 ± 0.0010	< 1 Year
0.0509 ± 0.0004	Policy_Sales_Channel
0.0045 ± 0.0002	Region_Code
0.0041 ± 0.0005	1-2 Year
0.0017 ± 0.0001	Gender
0.0007 ± 0.0001	Driving_License
0.0006 ± 0.0001	> 2 Years
0.0000 ± 0.0000	Vintage
-0.0000 ± 0.0002	Annual_Premium
-0.0001 ± 0.0002	Age

CONCLUSION

We have established a model for our client's classification problem wherein given a customer's data, the model is able to predict whether he/she will be interested in buying Vehicle Insurance. The model was able to predict with a **ROC** score of **0.845** performance.

Model Interpretation also shows how each feature contributes to the Response. Both SHAP and Eli5 give similar results.

CHALLENGES FACED

- The dataset contained a lot of categorical features out of which certain features had to be encoded in order to be used for modelling purpose.
- We had to deal with Class Imbalance in our dataset where ‘Response’ labelled as ‘Yes’ was a minority class.
- We were facing with some kind of overfitting even in PCA based models.

FUTURE SCOPE OF WORK

- Weights assigned to the encoded columns namely ‘Region Code’ and ‘Policy Sales Channel’ can be better which would lead to better results.
- Creating Application and Model Deployment.
- Various Other Classifiers can be used for the classification problem.