“Cross Sell Prediction: Vehicle insurance recommendation”

Predicting customer interest in vehicle insurance with precision.

Santosh kumar Ravikanti

1st December, 2024

Problem Statement

Challenge:

Predict whether a customer will be interested in buying a vehicle insurance product.

Why It Matters:

- *Vehicle insurance is critical for risk management and compliance.*
- *Optimized targeting reduces costs and improves customer satisfaction.*
- *Informed predictions can enhance sales efficiency and marketing ROI (Return on Investment).*

Our Solution

Overview:
An ML-powered predictive model that forecasts customer interest based on demographics, policy details, and past interactions.

Key Features:

- *High accuracy using advanced ML techniques.*
- *Robust feature engineering to capture insights.*
- *Easy-to-deploy API for real-time predictions.*

Dataset and Exploratory Data Analysis

Dataset

Dataset Overview:

- *Provided by Analytics Vidhya.*
- *Contains demographic details, policy features, and historical data.*
- *Size:*
    - *Train dataset: Over 3,81,109 rows and 12 features*
    - *Test dataset: Over 1,27,037 rows and 11 features*

Dataset Overview

Features Overview:

- *Categorical features:*
  - *Gender, vehicle, age, vehicle damage*
- *Numerical features:*
  - *Region code, annual premium, policy sales channel*
- *Target variable : Response (1 for purchase, 0 otherwise)*

Observations

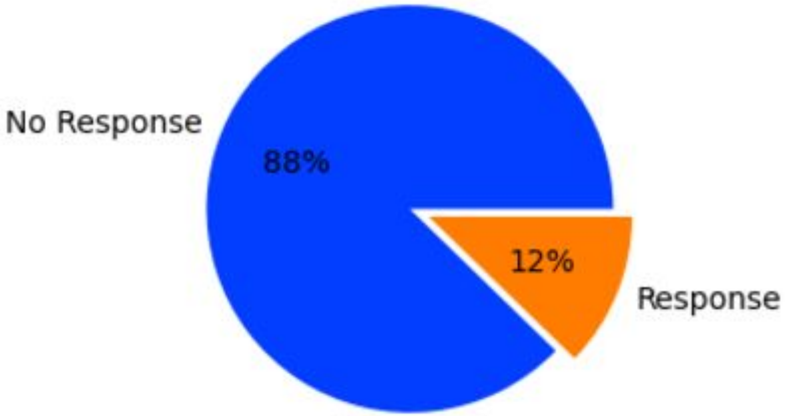
Train dataset Observations:

- *Total columns - 12*
  - *Integer - 6 columns*
  - *Object - 3 columns*
  - *Float - 3 columns*
- *No missing data*
- *No duplicates*
- *No null values*

Observations - Target data

Target data: "Response"

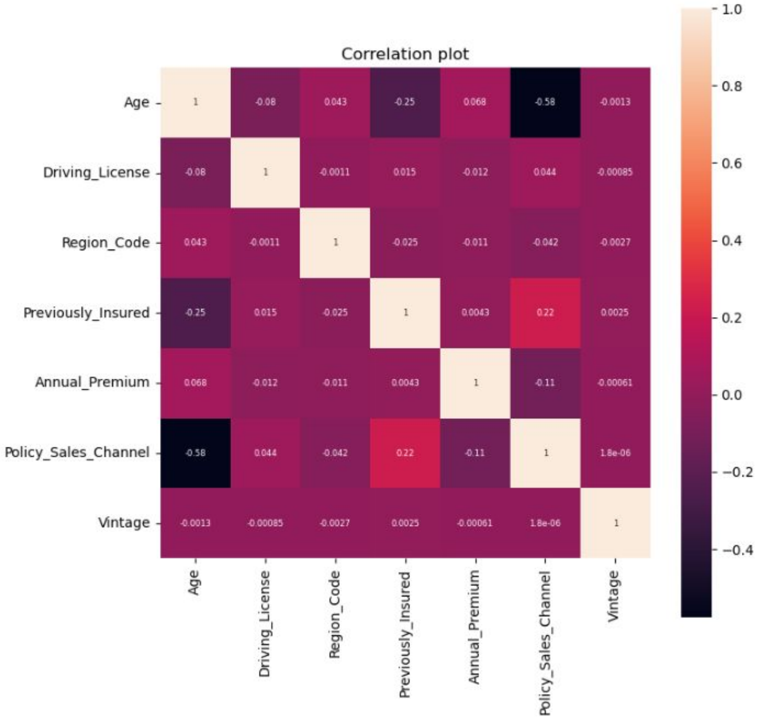
- *Imbalanced target data*
- *Response rate*
  - *0 -* ***87.74*** *%*
  - *1 -* ***12.25*** *%*



Correlation of features

Correlation:

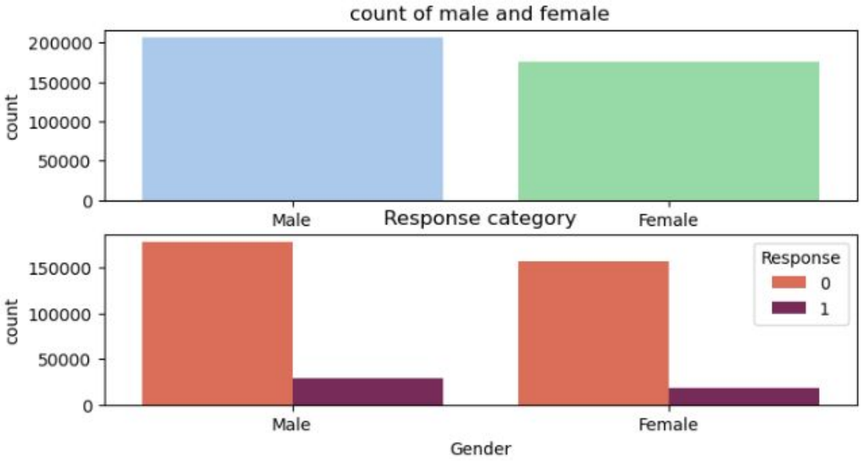
- ***Age*** *and* ***Policy Sales Channel*** *are highly correlated among all the features*
- ***Age*** *and* ***Previously Insured*** *in the second highest correlated among all the features*



Observations - Gender

Gender:

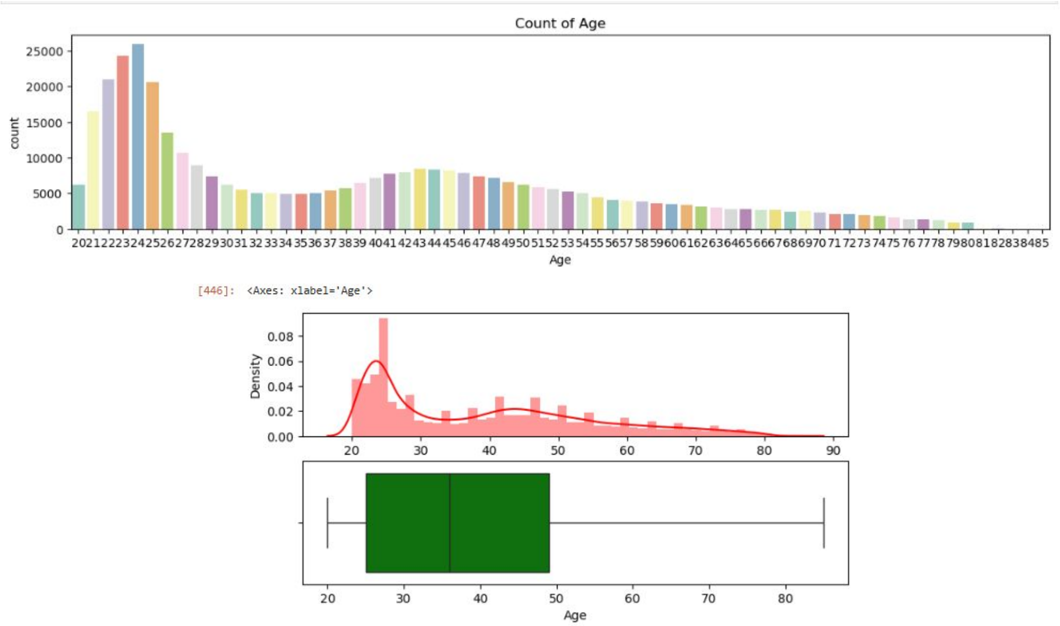
- *Gender is equally distributed in the training population*
- *Male category has slightly high chances of buying an insurance*



Observations - Age

Age:

- *Count of individuals with Age 24 are greater in the dataset*
- *Age data distribution is skewed*
- *No outliers observed in the box plot*

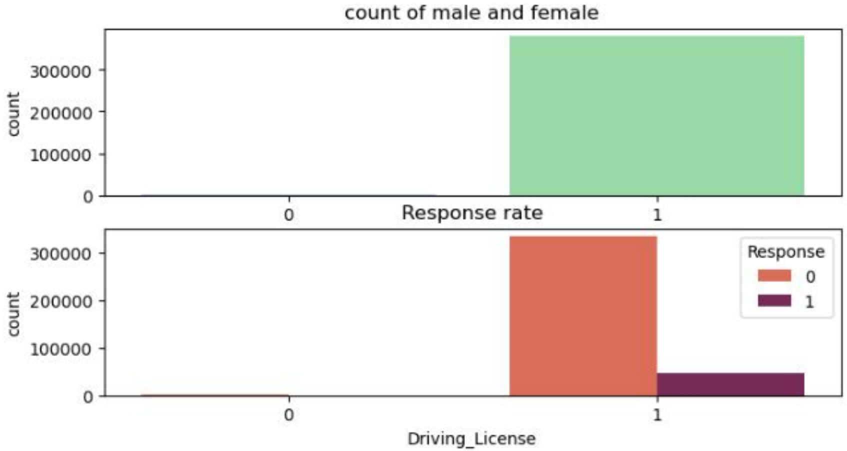


```
[446]:  <Axes: xlabel='Age'>
```

Observations - Driving License

Driving License:

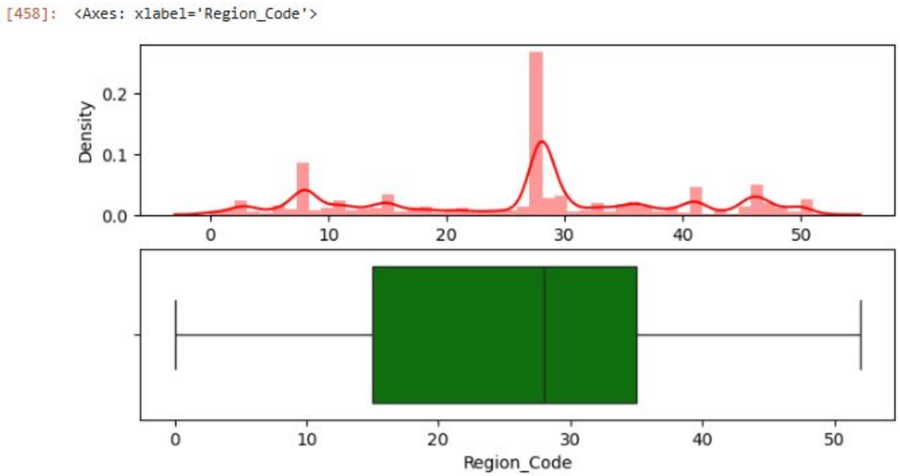
- *People with driving license are more than 99.78 %*
- *People interested in insurance almost have a driving license*



Observations - Region Code

Region Code:

- *People with region code 28 has the highest no of records*
- *No outliers in the box plot*

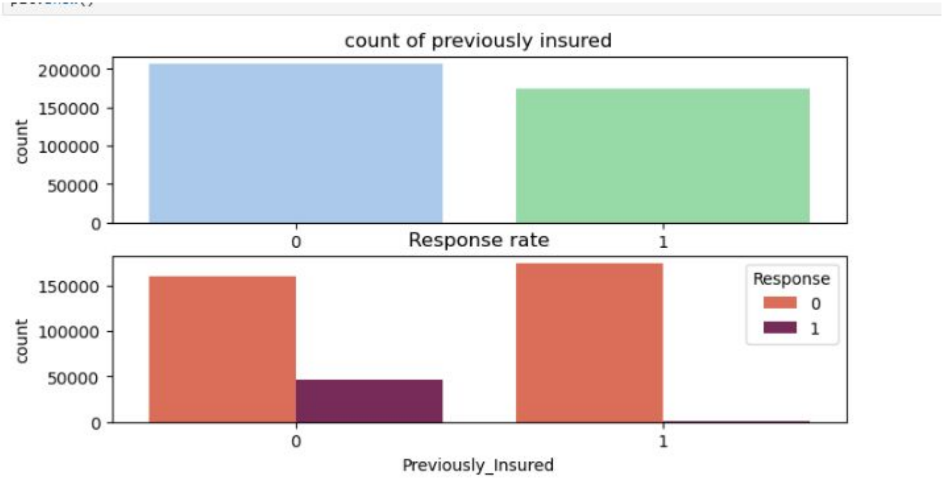


```
[458]:  <Axes: xlabel='Region_Code'>
```

Observations - Previously Insured

Previously Insured:

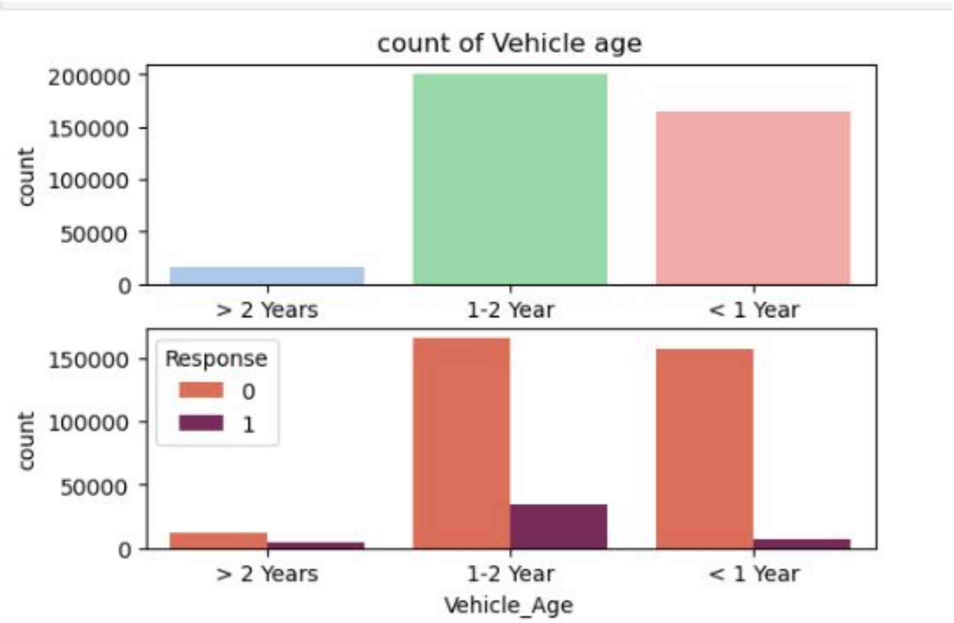
- *People previously insured are almost in equal distribution*
- *Few people who were not previously insured are now interested for insurance*



Observations - Vehicle Age

Vehicle Age:

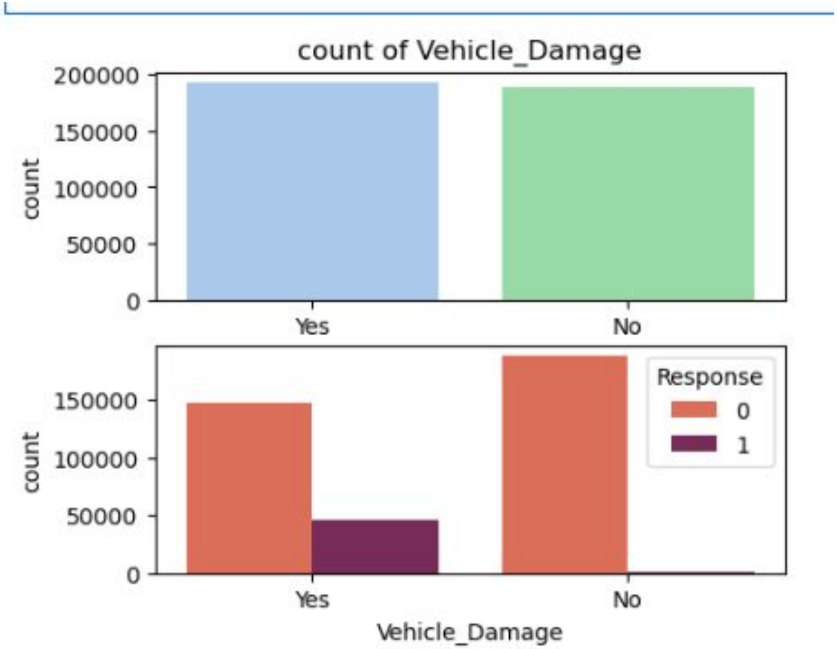
- *Most people are having vehicles less than 2 years*
- *More people with 1-2 years of vehicle age are interested in insurance compared with other categories*



Observations - Vehicle Damage

Vehicle Damage:

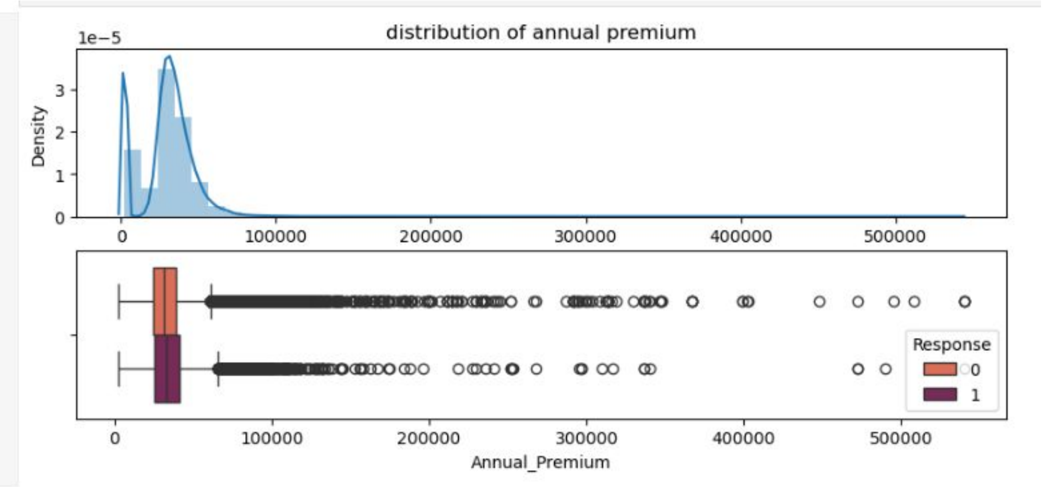
- *Vehicles damaged - Yes and No are equally distributed*
- *People with vehicles damage are most interested in the vehicle insurance*



Observations - Annual premium

Annual premium:

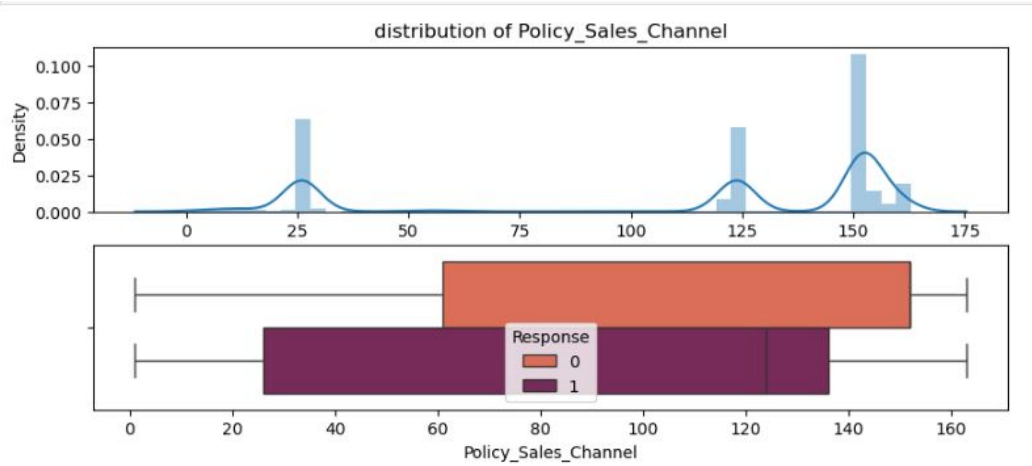
- *Annual premium has got more outliers*
- *It has skewed distribution*



Observations - Policy Sales Channel

Policy Sales Channel:

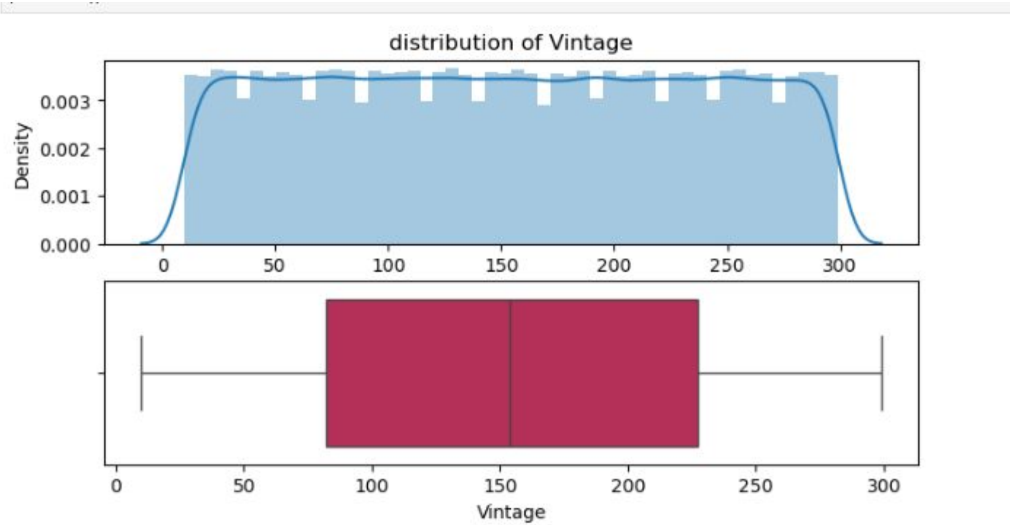
- *Sales channel 150 has got more density than any others*
- *No outliers observed in this data*



Observations - Vintage

Vintage:

- *No outliers observed in this data*
- *Vintage values are mostly equally distributed*

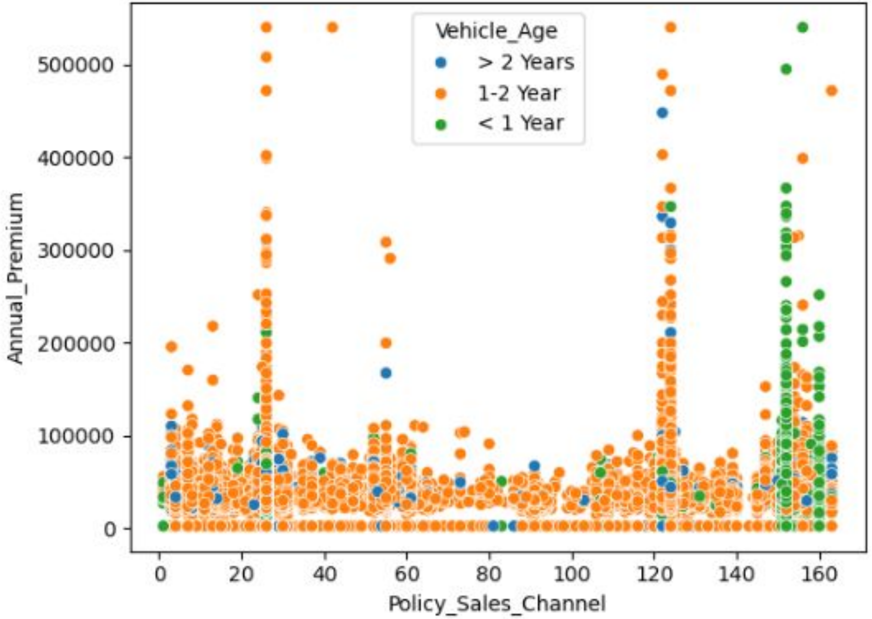


Other Observations

Observation:

- *Sales channel 150 has got more people paying annual premiums with vehicle age <1 year (new vehicles)*

```
[361]:  <Axes: xlabel='Policy_Sales_Channel', ylabel='Annual_Premium'>
```



Model building

Methodology - ML Pipeline

Data Preprocessing:

- Handled outliers.
- Encoded categorical variables using One-Hot Encoding.
- Normalized numerical features for better performance
- Addressed class imbalance using SMOTE.

Feature Engineering:

- Derived meaningful variables like vehicle age bins (<1 year, 1-2 years, > 2 years)

Model Development:

- Tested models: Logistic Regression, Decision Tree classifier, Random Forest, Gradient Boosting, XGBoost, Cat Boost classifier, Light GBM Classifier.
- Hyperparameter optimization using GridSearchCV, RandomizedSearchCV.

Model comparison and evaluation

Model Parameters	Best Model	CV - mean test score	AUC score	Test Solution score
Logistic Regression without penalty	Logistic Regression	0.83	0.5	0.4999
Logistic Regression with l2 penalty, Decision Tree Classifier	Logistic Regression	0.836327	0.782963	0.7947
XGBoost Classifier with eval metric log loss, SMOTE	XGBoost Classifier	0.826957	0.713364	0.7031
Random Forest Classifier with SMOTE	Random Forest Classifier	0.826462	0.796764	0.7463
XGBoost Classifier, LGBM Classifier, CatBoost Classifier	XGBoost Classifier	0.820355	0.808041	0.7664
XGBoost Classifier, LGBM Classifier, CatBoost Classifier, Decsion Tree classifier Logistic regression with penalty	CatBoost Classifier	0.829858	0.807639	0.7664
XGBoost Classifier, LGBM Classifier, CatBoost Classifier, Decsion Tree classifier Logistic regression with penalty, AdaBoost, Random Forest	CatBoost Classifier	0.826287	0.776609	0.77609
Logistic Regression with penalty and Decision Tree with entropy and gini both	Logistic Regression	0.84873	0.502789	0.5005
Logistic Regression with penalty and Decision Tree with entropy, gini, log loss	Logistic Regression	0.848759	0.502789	0.5005
Logistic Regression with penalty and Decision Tree with entropy, gini, log loss - with over sampled data	Decision Tree Classifier	0.943733	0.999923	0.5948
CatBoost Classifier, Light GBM	Light GBM classifier	0.833892	0.833892	0.5018
Logistic Regression, XGBoost classifier with SMOTE and one hot encoding	XGBoost Classifier	0.826957	0.713364	0.7031

Best Performing Model

Logistic Regression

Demo

Visual

- Jupyter Notebook
- Git Hub - link
- Docker Hub - https://hub.docker.com/
- Fast API
  - Local
  - Docker HUB
  - AWS cloud - link
- Streamlit
  - Local
  - Docker HUB
  - AWS cloud - link
- Streamlit io - link

Sample Parameters to use

```
{
 "Gender": "Male",
 "Age": 40,
 "Driving_License": 1,
 "Region_Code": 28,
 "Previously_Insured": 0,
 "Vehicle_Age": "1-2 Year",
 "Vehicle_Damage": "Yes",
 "Annual_Premium": 33762,
 "Policy_Sales_Channel": 7,
 "Vintage": 111
}
```

Methodology - ML Pipeline

Challenges:

- *Managing the class imbalance in data.*
- *Balancing overfitting.*
- *Finding optimal hyperparameters using Grid Search CV and Randomised Search CV*

Learnings:

- *Linear regression with L2 penalty is used and no other model is performing well.*
- *SMOTE combined with XGBoost improved results significantly.*
- *Age and policy customization are key factors influencing interest.*

Model Enhancements

Scope of improvements in the data

- *Incorporate external datasets like customer income or vehicle type*

Model Enhancements:

- *Scope of finding better model hyperparameters*
- *Test deep learning models for further improvement.*

Conclusion

"Our model empowers the insurance company to predict customer interest, reduce costs, and optimize revenue streams effectively."

Thank you