Capstone-Project-HEALTH-INSURANCE-CROSS-SELL-PREDICTION---Classification

This project is a part of the AlmaBetter Premium Program , Banglore/Bengaluru ,Karnataka , India

-- Project Status: [Completed]

Project Summary :

Problem Statement :

The task is to use existing Health and Vehicle Insurance Customer Data to predict whether the any new Customers are open to purchasing Vehicle Insurance from this company .

About the Data :

We have the data of existing Health Insurance Customers , this Data includes 12 relevant data points such as age, gender, sales channel data, vehicle ownership data. And most importantly , the target variable : whether the customer has vehicle insurance or not . The Data is available for 390K existing customers .

Approach taken :

This project is divided into multiple section where each section will have it’s own importance towards our problem statement. the approach that we will be following in this project is given as-

Section 1: Data overview : In this section we will be importing important libraries and simply load our dataset into google colab and will explore the basic information about data.

Section 2: Exploratory Data Analysis : In this section We fill focus on Exploratory data analysis of the dataset using various methods and visualization plot and will be extracting the information from this dataset as much as we can.

Section 3: Outliers handling : In this section we will be dealing with outliers in out dataset and will see how to define our outlier criteria and deal with outliers.

Section 4: Data preprocessing : In this Section we will be splitting the data into training and testing dataset then we will be scaling the data.

Section 6: Model implimentation : In this section we will be applying and evaluting the ML models to predict the results and .

Section 7: Model Explainability : In this section we will use model explainability to explain the results.

Section 8: In this section we will be giving a quick summary of entire notebook.

Technical Details for ML : We trained 3 Different Algorithms

( Logistic Regression , Decision Tree and Random Forest) We used GridSearchCV and BayesSearchCV for HyperParameter Tuning Comparing both F1 and AUC-ROC Score , we can see that Random Forrest and XGBoost model performs the best . Best AUC-ROC = 0.85 , Best F1=0.43

Conclusions : Insights from exploring the Data :

● Men's give higher conversion. ● Target more people having age range 30-55 years. ● Region 29 has given us highest conversions. ● Poor conversion from “< 1year” vehicle age while “>1year” give comparatively good sell counts. ● As per the data we have seen that no insurance conversion is given from no-damage vehicle. Hence organization should make some strategy to attract no-damage vehicle. ● Less records having > 75000 premium and hence less +ve responses. ● We saw that channel no. 26 comes out to be best channel as it gave us 15891 sales count followed by channel 124 which gave us 13996 conversion. ● If a customer is not previously insured then such customers are more likely to buy. ● If we look from the perspective of explainability so we can choose decision tree but Random forest results are slightly better however the model is bit complex. ● Vehicle damage is the most important feature that drive predictions followed by previously insured.

Python Libraries used

Datawrangling :

-Numpy

-Pandas

For Graphing :

-Matplotib

-Seaborn

Machine learning :

-Scikit-Learn

Model explainability :

-Lime

-Shap

Miscellaneous :

-Google colab tools

The Structure of the main IPYNB notebook :

-About this Project

-Problem Statement

-Bussiness Goal

Approach Taken in this Project

Understanding the given Data

Initial Code : Initliaing the Data and Modules

-Installing and Importing Libraries

-Import Dataset and Initial Data Checks

Data Preparation and Cleaning

Exploratory Data Analysis

Initial Macro-Level Data Analysis

Variable wise EDA

-Target Variable (Response)

-Age variable

-Annual\_Premium

-Gender variable

-Driving License

-Previously Insured

-Vehicle Age

-Vechicle damage

-Vintage

-Region Code

-Policy Sales Channel

Correlation Plot for Numeric Features

Data Preprocessing and Feature Engineering

Outlier Treatment in feature : Annual\_Premium

-oneHotencoding

-Standardization

-Cleaned Data Exporting

Building Prediction Systems using ML Models

Import cleaned final data

Classifier Performance Reporting Function

-Overfitting Underfitting Debugging Notes

-Metrics to be used during HyperParameter

-Random Forrest Specific Cutom Defined Metrics

Decision Tree Classifier

-Default Parameters : DecisionTreeClassifier

-Base Estimator Generator Function

-HyperParameter tuning using GridSearchCV

-Final Training Run

Logistic Regression Classifier Algorithm

-LR Classifier Generator Function

-LR Hyper Parameter Tuning : GridSearch

-Final Logistic Regression Training run

Random Forrests of Decsision Trees

-Default Parameters : RandomForestClassifier

-Base Estimator Generator Function

-HyperParameter tuning using GridSearchCV

-Final Training Run

Created by- Abdul Quadir – [abdulec1002@gmail.com](mailto:abdulec1002@gmail.com)