Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:
Name: Shubham Chandrakar Email Id: chandrakar.shubham17@gmail.com Contribution: 1. Data Wrangling • Checking for duplicate row and null/nan values • Checking distribution of dependent variable • Feature engineering • Correlation Analysis • Outlier removal 2. EDA on dataset • Univariate analysis • Bivariate analysis • VIF analysis 3. Categorical Encoding 4. Building machine learning models • ML model with imbalanced data • ML model with oversampling data using SMOTE • ML model with both oversampling and undersampling data using SMOTE TOMEK 5. Identifying best sampling method with best ROC score, f1 score and recall score 6. Feature importance analysis on the best model: Scikit-learn and Shap analysis
Please paste the GitHub Repo link.
Github Link:- https://github.com/chandrakar-shubham/health-insurance-cross-sell-prediction
Please paste the Gdrive Repo link.
Gdrive Link: https://drive.google.com/drive/folders/1d4gz6gjb7eaSm-7syOUSnX-bYvcdEzaV?usp=share_link
Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words) The client is an insurance company involved in the health insurance business. The client has experience in selling health insurance. They need the help of the data science team in building a model to predict whether the policyholders (customers)

from past year will also be interested in Vehicle Insurance provided by the company.An insurance policy is an arrangement by which a company undertakes to provide a guarantee of compensation for specified loss, damage, illness, or death in return for the payment of a specified premium. A premium is a sum of money that the customer needs to pay regularly to an insurance company for this guarantee.
Our aim is to explore, analyze and draw insights from the data as well as predict very efficiently where policyholders (customers) from past year will also be interested or not.

In this project, as a first step data wrangling is performed on data. It is found the dependent variable is highly imbalanced and categorical in nature. In the next step, EDA was performed, In EDA , Univariate analysis reveals male customers are more than female customer,most of the people have driving license, most of the customers are in age group less than 28, most vehicle insured are 1-2 years old, more than 50% people reported vehicle damage, most of the customers have negative response as compared to positive response,Most of the people were not previously insured, major proportion of customers are youth followed by middle aged.

In Bivariate analysis, It is observed that with age of vehicle premium increases, Middle aged people are giving more positive responsive followed by senior, customer who had vehicle damage in the past show more interest than with no vehicle damage, 26 policy sales channel out of 155 have more than generating more than 20% conversion rate in terms of interest, More than 15 region code out of 44 are generating more than 12% conversion rate in terms of customer policy purchase interest.

In the third step we build a model in three ways to attain best model performance are as follows : In the first way, we created various models with imbalanced data, it was observed that only tree based models like decision tree classifiers with hyperparameter tuning are giving best results. In the second way, we built models after performing oversampling of data using SMOTE(**Synthetic Minority Oversampling Technique**). It was observed that the XGBoost Classifier model with hyperparameters tuning is performing best. In the third way, we performed oversampling as well as undersampling of data using SMOTETomek. It was observed that the XGBoost Classifier model with hyperparameters tuning is performing best.

While model building it was observed that SMOTETomek is providing the best representation of data as it does both oversampling as well as undersampling of data as well as it provides best results.

After finding best model, We performed feature importance analysis, It was found that these are top five most important feature

Conclusion : We need to target middle aged, senior people and focus more on converting youth into our possible customers. People who did not have vehicle damage needed to be guided more to purchase policy more as most of the customers showing positive response had vehicle damage. Thirty-six policy channels had generated no response,they needed to be trained robustly. Regions generating less than 8% conversion rate in terms of customer policy purchase interest must focus on increasing brand presence in their respective area. While building ML models, SMOTETonek must be used for re-sampling data after that Random Forest Classification model with hyperparameter tuning can predict whether the policyholders (customers) from past year will also be interested in Vehicle Insurance provided by the company with 94.26 % ROC-AUC score and 88.27% f1 score.