Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:
Contribution Role : **1.** Nikunj Sonule (nikunj.sonule10@gmail.com) 1. Data Preprocessing and data cleaning 2. Exploratory Data Analysis • Univariate Analysis on numeric and categorical features • Bivariate Analysis on numeric and categorical features • Analysis from categorical variables 3. Fitting different models • Logistic Regression • XGBoost Classifier 4. Hyper parameter Tuning with GridSearchCV in Logistic Regression and XGBoost Model 5. Model Evaluation Metrics • Roc_auc score • Precision • recall • F1 score • Test Accuracy 6. Getting detail insights from ROC curve and Confusion Matrix
Please paste the GitHub Repo link.
Github Link:- https://github.com/nikunjsonule/HEALTH-INSURANCE-CROSS-SELL-PREDICTION
Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)

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. We were provided with 'Health Insurance Cross Sell Prediction' dataset to perform machine learning task, were to get insights from the independent and dependent variable of a dataset. We've all record and data with different Age, Driving License, Region Code, Previously Insured, Vehicle Age, Vehicle Damage, Annual Premium, Policy Sales Channel, Vintage, Response (Target variable – 1 and 0). Providing specific data analysis and prediction to done with this data.

As the first step, I performed data preprocessing and EDA part. The Age, Region Code and Policy Sales Channel column was an important feature to gets smaller insights from its. Extracting YoungAge, MiddleAge and OldAge from Age to create a new feature in dataset. Similarly, getting Region A, Region B and Region C from Region Code. For Policy Sales Channel I've got Channel A, Channel B, Channel C and Channel D. The extraction has made very easy to get useful insights. In EDA plotting heatmap for correlation, univariate and bivariate analysis, graphs and other plots to get some insights from the data.

After data preprocessing, our data was ready to fit into the models. Different model used to get optimal roc_auc score, precision, recall, f1 score and test accuracy score, models are – Logistic Regression, XGBoost Classifier model.

The third step is to do some hyperparamter tuning. Doing GridSearchCV on Logistic Regression and XGBoost Classifier part to get best roc_auc score and to shrink the coefficient and get optimal model. As per roc_auc score we find Logistic Regression as optimal model.

As to get more about TPR and FPR point view, the threshold was main important to know. The ROC curve and Confusion matrix has helped a lot to understand how those curve differ. The Confusion matrix is better to know a score for precision, recall, test accuracy and f1 score. The desired score has satisfy the confusion matrix. TPR and TNR is higher which is a good thing in imbalance dataset.

The final section is to observe model evaluation metrics. Logistic regression shows roc_auc score with 98%, test accuracy score with 93.71%, precision with 63%, recall 93% and f1 score with 75%. As for XGBoost Classifier we got roc_auc score with 90.33%, test accuracy with 84.32%, precision with 72%, recall and f1 score with 95% and 82% respectively.

So roc_auc score of the best model is 98% for this dataset. This performance could be due to various reasons: no proper pattern of data, it may be some negligible noise in the data, different model give different accuracy score.

That's how I have accomplished my project work in Health Insurance Cross Sell Prediction.

Drive link :-

https://drive.google.com/drive/folders/1zE7ExRtaciyq4tqw9ooxTXnelk8S4c6T?usp=sharing