Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

1) Ajinkya Dakhale
   Email- Ajinkya.dakhale2408@gmail.com
   - Data inspection
   - Exploratory Data Analysis
     - Checking distribution of features.
     - Checking relation of target feature with independent feature.
     - Used multiple graphs and did analysis on dataset.
   - Feature Engineering
     - Checking null values.
     - Handling outliers.
     - Encoding categorical features.
   - Feature Selection
     - Variance Threshold
   - Model training
     - Logistic Regression
     - Decision Tree
     - Random Forest Model
     - XGB Classifier Model
   - Imbalance Techniques
   - Technical Documentation

2) Harshjot Singh
   Email- hs9158695878@gmail.com
   - Data Cleaning
   - Outlier handling
   - Checked distribution of numerical feature
   - Checked correlation
   - Encoding for categorical features
   - Features Selection
   - Preparation and Model making
   - Model training
     - Logistic Regression
     - Decision Tree
     - Random Forest Model
     - XGB Classifier Model
   - Imbalance Techniques
   - Set up Roc and Precision recall curve for best model
   - Conclusion

3) Suvir kapse
   Email- suvirkapse@gmail.com
   - Exploratory Data Analysis:
     - Analyzing responses based on gender.
     - Age, Previously Insured, Vehicle age, Region code Vs Response
     - Checking distributions.

- Imbalance techniques
- Preparation and Model making
- Model training
  - Logistic Regression
  - Decision Tree
  - Random Forest Model
  - XGB Classifier Model
- Hyperparameter tuning for top 2 models
  1. For Random Forest Classifier
  2. Xgboost classifier

Please paste the GitHub Repo link.

Github Link:- https://github.com/Harshjyot-Singh-Chawla/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)

PROBLEM STATEMENT:

Cross-selling allows insurance companies to grow their bottom line without having to start from scratch. Similar to medical insurance, car insurance entails the customer paying a yearly premium to the insurance company in order for the insurance company to give the customer with compensation in the sad event that an accident involving the vehicle occurs.
The objective is to forecast whether any potential new customers are likely to be interested in obtaining vehicle insurance from this company using the existing health and vehicle insurance customer data.
By creating a model to predict if a customer would be interested in acquiring vehicle insurance, the business can then plan its communication strategy to reach out to those clients and maximise its business model and revenue.
We have a dataset which contains information about demographics (gender, age, region code type), Vehicles (Vehicle Age, Damage), Policy (Premium, sourcing channel) etc. related to a person who is interested in vehicle insurance. We have 381109 data points available.

APPROACH:

- The first step includes loading of dataset and then inspecting the data through which we get to know the summary or description of data, shape and size of data, null value count, and duplicates values in the data and about the data types of column.
- On the basis of univariate, bivariate, and multivariate analysis, we have carried out several visualisations. First, we performed a Univariate analysis since we needed to comprehend each feature or column's individual significance and the insights it would add to our study. Second, we used bivariate analysis to examine how one column or characteristic affects another, as well as the direction these discoveries may take us. Finally, we conducted a multivariate study to determine the effect of various factors on multicollinearity.
- Next step involves visualization of data .In visualization we saw that dependent variable (i.e.response) is highly imbalanced and then used SMOTE technique to balance it. After having a look at the distribution of data we saw that Annual_premium column have outliers**.** We convert Annual_premium column to normal distribution by power transformer.
- Following data visualisation, we utilize onehotencoder and label encoding to perform encoding, which converts categorical data to numerical data. We then performed feature

selection using VIF and removed variable Driving_License because of high VIF value. We then divide the data by 80:20 using train test split. 20% for model testing and 80% for model training.

- Then, various models are applied. We used Logistic Regression, Decision Tree, Random Forest Regression and XGBoost Classifier and then used Bayes search CV for hyperparameter tuning.

CONCLUSION:

- The given dataset is an imbalance dataset since the value of the Response variable, 1, is significantly lower than its value, 0.
- In compared to their female counterparts, male consumers own a little bit more vehicles and have a higher likelihood to get insurance.
- Consumers between the ages of 30 and 60 are the most likely to purchase insurance, whilst consumers under the age of 30 are uninterested in vehicle insurance. Lack of involvement, ignorance of insurance, and perhaps a deficit of expensive vehicles are possible explanations.
- Customers with driving licenses are more likely to purchase insurance
- Compared to consumers with vehicles that are less than one year old, those with 1-2 year old vehicles show greater interest in purchasing insurance.
- Customers who have experienced vehicle damage are more likely to buy insurance since they know firsthand how much it costs to restore a car.
- The variable such as Age, Previously_insured, Annual_premium is more affecting the target variable.
- We used different type of algorithms to train our model like, Logistic Regression, Random Forest model, Decision tree and XGB Classifier. And also we tuned the parameters of XGB Classifier and Random Forest model, comparing the model on the basis of precision, recall, accuracy, F1 score we can see that the XGBClassifier model performs better. Even comparing ROC curve XGB Classifier performed better because curves closer to the top-left corner indicate better performance.