Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

1) Manoj Patil M
   Email- smmanoj208@gmail.com
   - Data inspection
   - Exploratory Data Analysis
     - Checking distribution of features.
     - Pandas profiling of dataset.
     - 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
     - F classification
     - Pearson correlation
   - Imbalance Techniques
   - Model Training
   - Cross validation and Hyper parameter tuning
   - Conclusion
   - Technical Documentation

2) Gulzar
   Email-gulzarkhan9980@gmail.com
   - Data inspection
   - Data Cleaning
   - Outlier handling
   - Checked distribution of numerical features
   - Visualization of all features count with respect to target feature
   - Checked relationship between our features and target
   - Checked correlation
   - Encoding for categorical features
   - Features Selection
   - Preparation and Model making
   - Used SMOTE for balancing data but didn't give promising results
   - Used standard(commented) and minmax both scaler
   - Tried out 10 models and comparing their results with barplot.
   - Written code for comparing of all 10 model's confusion matrix.
   - Hyperparameter tuning for top 2 models
     1. For Random Forest Classifier
     2. For LGBM Classifier
     3. Setting up best parameters after a lot of playing

- Set up Roc and Precision recall curve for best model
- Made feature importance column according to best model
- Conclusion with improvements points
- Made powerpoint presentation.

3) Bindu Kovvada
Email- bindukovvada187@gmail.com
   - Data Inspection
   - Exploratory Data Analysis:
     - Analyzing responses based on gender.
     - Age, Previously Insured, Vehicle age, Region code Vs Response
     - Checking distributions.
   - Feature Engineering:
     - Checking null values
     - Handling outliers
     - Encoding categorical features
   - Feature Selection:
     - F classification
     - Feature importance
     - Pearson correlation
   - Imbalance techniques
   - Model training
   - Cross validation and Hyper parameter tuning
   - Conclusion.

4) Deepak Kumar Gautam
Email- deepakpracheta@gmail.com
   - Data inspection
   - Data Cleaning
   - Exploratory data analysis
     - Checking relation of independent features with target feature.
     - Checked distribution of features.
     - Checked correlation through heatmap.

   - Feature Engineering
     - Checking for null values
     - Handling outliers
     - Encoding categorical features

   - Feature Selection
     - F Classification
     - Variance Threshold
     - Pearson correlation
   - Used Imbalance Techniques like Oversampling and SMOTE
   - Model Training
   - Cross Validation and Hyper Parameter Tuning
   - Conclusion.
   - Project Summary.

5) Saksham Tripathi
Email-saksham757474@gmail.com
   - Data Inspection
   - Exploratory Data Analysis:
     - Analyzing responses based on gender.
     - Checked distribution of features
     - Used multiple graphs and did analysis on dataset
   - Feature Engineering
     - Checking for null values
     - Handling outliers
     - Encoding categorical features
   - Feature Selection:
     - F classification
     - Feature importance
     - Pearson correlation
   - Used Imbalance Techniques SMOTE
   - Used standard and minmax both scaler
   - Model Training
   - Cross Validation and Hyper Parameter Tuning
   - Conclusion.
   - Technical Document.

Please paste the GitHub Repo link.

Github Link:- github 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)

Problem statement:

Our client is an insurance company that has provided Health Insurance to its customers now they need you r help in building a model to predict whether the policyholders (customers) from past year will also be inte rested in Vehicle Insurance provided by the company.

An insurance policy is an arrangement by which a company undertakes to provide a guarantee of compens ation for specified loss, damage, illness, or death in return for the payment of a specified premium. A prem ium is a sum of money that the customer needs to pay regularly to an insurance company for this guarantee .
For example, you may pay a premium of Rs. 5000 each year for a health insurance cover of Rs. 200,000/- s o that if, God forbid, you fall ill and need to be hospitalised in that year, the insurance provider company w ill bear the cost of hospitalisation etc. for upto Rs. 200,000. Now if you are wondering how can company b ear such high hospitalisation cost when it charges a premium of only Rs. 5000/-
, that is where the concept of probabilities comes in picture. For example, like you, there may be 100 custo mers who would be paying a premium of Rs. 5000 every year, but only a few of them (say 2- 3)
would get hospitalised that year and not everyone. This way everyone shares the risk of everyone else.

Just like medical insurance, there is vehicle insurance where every year customer needs to pay a premium of certain amount to insurance provider company so that in case of unfortunate accident by the vehicle, the insurance provider company will provide a compensation (called 'sum assured') to the customer.

Building a model to predict whether a customer would be interested in Vehicle Insurance is extremely help ful for the company because it can then accordingly plan its communication strategy to reach out to those c ustomers and optimise its business model and revenue.

Now, in order to predict, whether the customer would be interested in Vehicle insurance, you have informa tion about demographics (gender, age, region code type), Vehicles (Vehicle Age, Damage), Policy (Premiu m, sourcing channel) etc.

Approach:

- First, we load data set into Panda's frame and initialize all the library which are required for doing EDA.
- Then we did inspection of data on a basic level.
- Then we did data cleaning by removing null values, duplicate values and outliers.
- Then we used the matplotlib and seaborn to do Exploratory Data Analysis on sample data by plotting different graphs like count plot, pie chart, lmplot, bar plot, boxplot, subplot and heat map from this we got useful insights and correlation between target column and other features
- In feature selection use variance threshold and F_Regression to select best features
- In feature engineering check null values, removed outliers in the data set. Use imbalance technique
- use multiple models to predict cross sell did hyper parameters tuning lgbm is best performing model

Conclusion:

- Customers aged between 30 to 60 are more likely to buy insurance.
- Youngsters under 30 are not intrigued by vehicle insurance. Reasons could be the absence of involvement, less awareness about insurance and they may not have costly vehicles yet.
- Consumers with 1-2-year-old vehicles are more interested as compared to others.
- Consumers with less than 1-year-old Vehicles have very less chance of buying Insurance
- Customers with Driving License have a higher chance of buying Insurance.
- Customers with Vehicle Damage are likely to buy insurance.
- The male category is somewhat more noteworthy than that of females and chances of purchasing the insurance are likewise minimally high.

- The variable such as Age, previously insured, Annual premium is more affecting the target variable.
- Comparing the ROC curve, we can see that the Random Forest model performs better. Because curves closer to the top-left corner indicate better performance.
- We can see that **LGBM model perform better** for this dataset.