Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)

Data Visualization:

Data visualization techniques for instance histogram, line graphs, heatmaps, box plots,pie charts etc helps us in understanding the pattern the data follows. Firstly, we used a pie chart plot to look at the target variable - ‘Term Deposit’. Then we used bar plots and boxplots to visualize the other variables, as most of the variables are categorical in nature. Further we went on to look at the age and job type distribution with respect to Term Deposit.

Feature Engineering:

Feature engineering is a method of converting raw data into desirable features before fitting it into a machine learning model which results in better performance. We used feature engineering techniques like One Hot encoding and Label Encoder. Additionally we dropped columns whose most of the values were unknown.

Feature Selection:-

Feature selection helps in reducing the number of input variables to decrease the computational cost of modeling and in some cases it improves the performance of the model.As the number of input variables were not excessive and all of them were significant, we choose all the features for the machine learning algorithm.

Model Evaluation metrics :-

We Evaluated the model on different metrics which helps us to better optimize the performance, fine-tune it, and obtain a better result. And got the results from the best suitable model for our project. Following are the evaluation metrics for our selected model:-

- **Confusion matrix** :- A confusion matrix is defined as the table that is often used to describe the performance of a classification model on a set of the test data for which the true values are known.
- **Accuracy :-** Accuracy simply measures how often the classifier correctly predicts.
- **Precision :-**Precision explains how many of the correctly predicted cases actually turned out to be positive.the precision score calculated for our model as 0.9187
- **Recall (Sensitivity) :-** Recall explains how many of the actual positive cases we were able to predict correctly with our model.The Recall score calculated for our model as 0.9268
- **F1 Score :-** It gives a combined idea about Precision and Recall metrics. The F1 score calculated for our model is 0.9227
- **Receiver Operator Characteristic (ROC)**
- **Area Under the Curve (AUC)**

Model Selection:-

After training the dataset on nine models and evaluating on the five evaluation metrics, the Gradient Boosting Classifier model came out to be the best model with an F1 score of 0.92 and matthews_corrcoef of 0.84.

Gradient boosting does not penalize missed-classified cases but uses a loss function instead. Loss function can be the mean average error for log loss for classification problems. In addition, the gradient boosting algorithm uses the gradient descent method to continuously minimize the loss function to find the optimal point.

Model Interpretation :-

We used Shapley Additive Explanation (SHAP Values) For interpretation of our model.The Shapley value is the average marginal contribution of a feature value across all possible coalitions. The goal of SHAP is to explain the prediction of an instance x by computing the contribution of each feature to the prediction. With SHAP package the calculation is quite simple and straightforward. We only need the model (regressor) and the dataset (X_train) .After calculating the SHAP values we plotted several analyses that will help us to understand the model.

Team Member’s Name, Email and Contribution:

Ajinkya Morade (ajinkya.morade1998@gmail.com)

- **Data Processing**
- **Exploratory Data Visualization**
- **Feature Engineering**
- **Imbalanced Dataset Handling(SMOTE, Stratified Shuffle)**
- **Model Evaluation Metrics**
- **Model Selection**

- **Feature Selection**
- **Feature Scaling**
- **Model Evaluation Metrics**
- **ROC-AUC**
- **Model Selection**
- **SMOTE**
- **Hyperparameter Tuning**
- **Model Interpretation**

Nitesh Gajakosh (niteshgajakosh1998@gmail.com)

- **Data Preprocessing**
- **Feature Engineering(Encoding)**
- **Model Selection**
- **Hyperparameter Tuning**
- **Model Evaluation Metrics**
- **Imbalanced Dataset Handling**
- **Tree Visualization**
- **Model Interpretation**

- **Exploratory Data Visualization**
- **Data Processing**
- **Feature engineering (Dropping unknowns,Label encoding)**
- **Imbalance dataset handling by using SMOT**
- **Model selection**
- **Model evaluation metrics(ROC AUC Curves)**
- **Model Interpretation**

Please paste the GitHub Repo link.

Github Link: https://github.com/ajinkyamorade/Bank-Marketing-Effectiveness-Prediction