Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email, and Contribution:
1) Pruthvi Raj Pruthviraj1698@gmail.com • Data wrangling • Exploratory Data Analysis • Handling Outliers • Data Preprocessing • Encoding • Training and testing models • Oversampling, Training and testing Models

Please paste the GitHub Repo link.

Github Link:- https://github.com/Pruthviraj009/Bank-Marketing-effi
Please write a summary of your Capstone project and its components. Describe the problem statement, your approaches, and your conclusions. (200-400 words)

The data provided is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required in order to access if the product (bank term deposit) would be yes or no (subscribed or not). The goal of the problem statement is to predict whether the client will subscribe a term deposit or not.

The data has total of 45211 rows and 17 columns. Luckily there were no null or duplicated values. We have plotted several plots between each feature. We have taken out categorical features out and plotted several plots. We have plotted bivariate to multivariate plots on the features. We have checked for the outliers and found age feature has outliers and we have handled the outliers. We have used LabelEncoder to convert categorical features to numerical features. All the features were not correlated with each other which was found in correlation heat map. As the data was imbalanced with very less yes in the target feature where it has more no. We have split the data train and test using train test split. We have built logistic regression, SVM classifier, Gaussian Naïve bayes, Random Forest Classifier, and Knn classifier and for the boosting methods we have used ADA boost classifier, Gradient Boost Classifier and XG boosting Classifier, all these models were giving almost same accuracy between 89 to 90.5 and we have used SMOTE (Synthetic Minority Oversampling Technique) which is used to oversample the data. After using Smote we have got same number of yes and no in the target feature. We have split the data for training and testing. We have built the same above models. Out of all this models XG boosting Classifier was giving higher accuracy which is 93%. After closely observing all the performed classification models we can clearly see the oversampled XB boosting Classifier gave the best results and predicted values were on par on the test data.