Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

Asim Siddiqui (siddiquiasim5800@gmail.com)

Contribution- Analyzed some of the attributes of the dataset provided a better visualization in a graphical manner which was very significant in model prediction. Implemented Logistic Regression,Random Forest classifier, KNN and XGBoost classifier.

Suraj Kumar Mishra (suraj170898@gmail.com)

Contribution- Observed some of the key factors and did mean encoding on one of the features which was quite significant and implemented various models such as Logistic Regression,DecisionTree,SGD Classifier , KNN,Random Forest, Gradient Boosting,XGBoost etc.

Sagar Rokad (sagarrokad000@gmail.com)

Contribution- Did some visualization on various features and did a task of feature selection among various features.Performed under sampling and oversampling techniques due to imbalance dataset and implemented some models such as Knn, Logistic Regression, Random Forest, XGBoost etc.

Please paste the GitHub Repo link.

Github Link:- https://github.com/Suraj110597/-Bank-Marketing-Effectiveness-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

The data was related to 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 assess if the product (bank term deposit) would be subscribed(yes) or not. The classification goal was to predict if the client will subscribe to a term deposit (variable y).

The dataset consists of direct marketing campaigns data of a banking institution which consisted of 45211 data points with 17 independent variables out of which 7 were numeric features and 10 were categorical features.

APPROACH

Exploratory Data Analysis:-

We performed exploratory data analysis with to get insights from the data to observe following things:-

- The dataset was imbalanced, where the number of negative classes is close to 8 times the number of positive classes.
- The customers who had a job of admin had the highest rate of subscribing a term deposit, but they were also the highest when it comes to not subscribing. This is simply because we have more customers working as admin than any other profession.
- Majority of the customers were married. Followed by Single, divorced and unknown.
- Majority of the customers had a housing loan.

Models Implementation:-

- **Logistic Regression**
- **Logistic Regression(under sampling)**
- **Random Forest(under sampling)**
- **Random Forest(under sampling)**
- **K-NN(over sampling)**
- **XGBoost(over sampling)**

Model Performance:

Model	Test AUC	Test Accuracy	F1_score	Precision
Logistic Regression	**0.70**	**0.61**	**0.29**	**0.19**
Logistic Regression (Under sampling)	**0.89**	**0.85**	**0.83**	**0.94**
Random Forest (Under sampling)	**0.95**	**0.89**	**0.88**	**0.92**
Random Forest (Over sampling)	**0.93**	**0.88**	**0.81**	**0.82**
KNN (Over sampling)	**0.87**	**0.82**	**0.69**	**0.78**
XGBoost (Over sampling)	**0.91**	**0.86**	**0.77**	**0.82**

Among all models, **Random Forest** and **XGBoost** work the best and provide a reliable prediction.