Capstone Project Submission

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

Introduction- Finance industry is one of the leading industries globally and have the potential to bring huge impact in the growth of nation. Thus, it is important to analyze the data or information that banking sector records about the clients. In this project we trained a model that can predict that **whether the client will opt for a term deposit or not** using given bank-client data, data related with the last contact of the current campaign and some other useful attributes.

Classification Approach- The given dataset was initially loaded for a quick overview. It was observed that our dataset contains **45211 records and 17 features**. Datatypes of features was then checked and it was found that there are **7 numerical (int) and 10 Categorical (object)** datatypes among which no null values and duplicated records were found in our dataset.

Once done with data wrangling, we started with **EDA** and we did all **univariate, bivariate and multivariate analysis** to get important insights for feature selection. We then performed **feature engineering** which is the most crucial part of the project. For this we **removed the outliers** in duration, campaign and previous. Then we **c**onverted some discrete variable into categorical variable and further **reduced 'job' categories** as they were 12 in number.

Thereafter, we plotted **correlation graph** and **build decision tree** to select important features which can outperform our final model to accurately predict output values. From the above Decision tree feature importance, we came to conclusion that only **'marital', 'age', 'day', 'campaign', 'balance', 'new_job_o', 'previous', 'poutcome', 'housing', 'pdays', 'month', 'contact'** are showing significant importance to be considered for model building. But from these features we removed **'previous'** as this feature is showing high collinearity with pdays**.**

After feature selection we started with **model building** by initially balancing the imbalanced data and then normalizing the data. We started with **KNN, Random forest**, but couldn't get good score. So, we opted for **LGBM** and got **recall score of 0.81** which was good enough to be accepted. Therefore, we plotted the feature importance of LGBM and saw that **'duration' is highly correlated with target class**.

Therefore, **to build more realistic model** we again trained LGBM but without 'duration' feature. This time **we got precision score of 0.46.** For our models **with 'duration'** feature included we focused on **recall score** to see the actual conversion rate, but after **removing this feature** we are more focusing on **precision score**, because now we are actually trying to predict the chances of conversion from the list of customers provided.

We also performed **hyperparameter tuning** on LGBM and **ANN model building** but didn't get better scores.

Thus, at last after selecting LGBM as best working model in our case we looked at the **SHAP plot**s for feature importance.

Conclusion- Thus we came to an end of our project and finally concluded that if the clients are contacted twice in first 10 days of the month, then it is more likely that call will be converted to subscription. Also, clients take more subscriptions in a month of January, May or August and if clients are already paying housing loans, then it is more likely that they will opt for term deposit.

Team Member's Name, Email and Contribution:

Cohort Kaimur- SSP

Contributor roles:

1. **Ganeshkumar Patel- (*ganeshkumarpatel452@gmail.com*)**
    a. Dataset Overview
    b. Data wrangling
    c. Exploratory Data Analysis
    d. Feature selection using heatmap and decision tree
    e. Model Building and comparison with different features, making PPT

2. **Yaman Saini- (*ys726507@gmail.com*)**
    a. Dataset Overview
    b. Data wrangling
    c. Exploratory Data Analysis
    d. Feature selection and model building
    e. Hyperparameter tuning and commenting codes in Colab notebook.

3. **Akanksha Agarwal- (*akn.agarwal@gmail.com*)**
    a. Dataset Overview
    b. Data wrangling
    c. Exploratory Data Analysis
    d. Feature selection and model building, feature importance of final model
    e. Writing inferences in Colab notebook, drafting technical document and summary

4. **Saurabh Funde- (*funde.saurabh21@gmail.com*)**
    a. Dataset Overview
    b. Data wrangling
    c. Exploratory Data Analysis
    d. Feature selection and model building
    e. Hyperparameter tuning, drafting technical document and summary

5. **Sanjay Kumar (*sanjay.skm222@gmail.com*)**
    a. Dataset Overview
    b. Data wrangling
    c. Exploratory Data Analysis
    d. Feature selection and model building
    e. Feature importance, making power point presentation and drafting summary

Please paste the GitHub Repo link.

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