**Capstone Project Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| Team Member’s Role:-  Nikhil Machave ([machavenikhil@gmail.com](mailto:machavenikhil@gmail.com))  Contribution:   * 1. Dataset Overview   2. Data wrangling   3. Exploratory Data Analysis   4. Feature selection using heat map and decision tree   5. Model Building and comparison with different features drafting technical document and summary     Aishwarya Methe ([aishwaryamethe252@gmail.com](mailto:aishwaryamethe252@gmail.com))  Contribution:   1. Dataset Overview 2. Data wrangling 3. Exploratory Data Analysis 4. Feature selection and model building, feature importance of final model 5. Writing inferences in Colab notebook   Aditya Tadas ([adityatadas@gmail.com](mailto:adityatadas@gmail.com))  Contribution:   1. Dataset Overview 2. Data wrangling 3. Exploratory Data Analysis 4. Feature selection and model building 5. Feature importance, making power point presentation and drafting summary 6. Shapash Model Explanatory |
| **Please paste the GitHub Repo link.** |
| Github Link:- https://github.com/Methe11/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)** |
| 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.  The given dataset was initially loaded for a quick overview. It was observed that our dataset contains 45211 records and 17 features. Data types of features was then checked and it was found that there are 7 numerical (int) and 10 Categorical (object) data types among which no null values and duplicated records were found in our dataset.  After that we have done some exploratory data analysis technique to find the patterns to predict model and also find which features are most important while predicting our target variable.  After finding the patter and feature importance we have done some feature engineering and categorize some variable according to their attributes i.e. Age variable is categorized into five categories according to age of the customers who are below 25 are grouped as Struggling customers after that the clients in in range of 25- 48 they are grouped as stable customers then when the customer age is in between 48-57 they are grouped as about to retire and all the customers who are above 57 age are grouped as counting a last breadth.  Now we have also done some feature engineering for job column we have categorized join column into 5 categories all the service are job related customer are grouped as cat 1 then retired customers are grouped as cat 2 to after that all the customers which are having there own business have been grouped in cat 3 then all the customer who have lower income such as unemployed , student , housemaid, and unknown are grouped in cat 4.  After than we have done some data pre processing technique to prepare our dataset to fit in our model for that we have applied frequency count encoding for month column because these column has too many unique value then we have applied one hot encoding and create dummy variables of all categorical variable.  We have drop pdays and duration column in our dataset because pdays variable contains 0 value and for the duration column in order to build realistic model we have to drop that column because it is obvious that when the communication time with the customer is zero then it will not agree to Subscribes for term deposit.  Then we have seen that there is huge class imbalance in target variable so we have to overcome these for that we have used SMOTE class imbalance oversampling for handling class imbalance it will increases our number of records from 45211 to 79874 . After that we have divided our data into train and test 75 % of data are use to train our model and 25 % of data are use to test model accuracy of our model.  Now our data is ready to fit into model Firstly we applied K Nearest Neighbour classifier it gives the accuracy of 78% then we have applied Random Forest Classifier then these machine learning model giver the accuracy of 90 % after that finally we have applied XG boost classifier these machine learning algorithm perform best to predict target variable it gives accuracy of 93% to increase the accuracy of XG boost we have Hypertuned that Algorithm. Hypertuning incresses the accurance of model by 1 % now XGB Hypertuning Model gives our accuracy of 94% .  Now to explain model and feature importance we have applied Shapash model explanatory and we visualize feature according to there importance for predict target variable.  From the above model explanatory tool we have seen that poutcome  Unknown is the most important feature while predicting our target variable also from the table we can see that when the poutcome is 0 then it contribute in the negative way and increases the probability of predicting 0.  Marital married is the second most important feature for predicting target variables from the table we can see that when the marital married then it will affect positively and increases the probability of predicting 1.  Also age cat stable variable affect positively on the target variable when the age of clients is stable then it will increases the probability of predicting 1 that means it higher the probability that client will subscribe for term deposit.  Also education secondary affects positively on the target variable when the client education is secondary then it increases the probability that client will agree to subscribe for term deposit.  From these we have seen that poutcome unknown , marital married education secondary ,education territory and marital single are the top 5 features which are predicting target variable.  XGB Classifier Performed Best for predicting target variable  . |