**Capstone Project Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

|  |
| --- |
| **Team Member’s Name, Email, and Contribution:** |
| Contributor Roles:  1. Jayanth V (Email – [arambamvj@gmail.com](mailto:arambamvj@gmail.com)) :   * 1. Data Overview.   2) Null values treatment.  3) Naïve Bayes Model classifier.  4) K nearest Model implementation.  5) Data Manipulation using Correlation.  6) Random Forest Implementation.  7) SVM classifier.  8) XG Boost Classifier.        2. **Gowthaam Kumarasamy (Email -** [**gowthaam02@gmail.com**](file:///C:\Users\Admin\Downloads\gowthaam02@gmail.com)**) :**   1. EDA (Univariate and Bivariate Analysis) 2. Feature Engineering. 3. Outlier Treatment. 4. Data preprocessing for model implementation. 5. SMOTE. 6. Hyper Parameter Tuning. 7. Cross-Validation. 8. Shapash Model Explanatory. |
| **Please paste the GitHub Repo link.** |
| GitHub Link: - <https://github.com/Gowthaam-K/Bank-Marketing-Effectiveness-Prediction>  Google Drive Link: <https://drive.google.com/drive/folders/1u2ahMdP4a9jYy4Q5sUlEfgXRymRJRknK?usp=sharing> |
| **Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches, and your conclusions. (200-400 words)** |
| The Portuguese banking institution was conducting marketing campaigns. The marketing campaigns were based on phone calls. Often, more than one contact with the same client was done, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed. The classification goal is to predict if the client will subscribe to a term deposit.  At first, we performed Data cleaning by creating a data frame and removing unnecessary columns with null values and started getting basic insight from the cleaned dataset.  We also performed Univariate and Bivariate analysis as a part of EDA to get basic insights into individual and columns’ effects on the sales column.  We merged the two datasets after performing basic cleaning and went with feature selection in which the numeric features in our dataset were taken into consideration.  we have also treated the outliers in the data by removing them using Z-score implementation.  After feature selection, we tend to split the data into training and testing pairs and started implementing basic Machine Learning Models to predict the sales  The Models Were:  a) Random Forest  b) Naïve Bayes Model  c) Support Vector Machine (SVM)  d) XG Boost  e) KNN classifier  To handle the class imbalance, we implemented SMOTE for performing oversampling of the data that we have in our dataset in which we achieve equal weightage of data.  So far, we have implemented different machine learning models to predict whether people subscribe to a term deposit. From the evaluation metrics of the different models, we came to the conclusion that the accuracy of the XG boost ensemble model is the highest compared to the remaining models i.e., **Accuracy:** 0.9318026585404298 is achieved.  Thus, the best model is the XG boost ensemble and hence we tend to apply hyperparameter tuning and cross-validation for the same. |