**Capstone Project Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

|  |
| --- |
| **Team Member’s Name, Email, and Contribution:** |
| 1. Name: - Shubham Narendra Kadu 2. Email - [shubhamkadu.ds@gmail.com](mailto:shubhamkadu.ds@gmail.com) 3. Contribution **–**  * Contributed to this project that first I self-done. * Data processing * Data Cleaning * Data Analysis * EDA * Correlation Analysis * SMOTE * One hot encoding * Machine Learning Model * Hyperparametric Tunning * Model Evaluation through different methods like precision and recall, f1 test, roc\_score and Confusion Matrix and ROC AUC curve * Model selection * Conclusion * PPT |
| **Please paste the GitHub Repo link.** |
| GitHub Link:- https://github.com/shubhsk98/Capston-project-Credit-Card-Default-Prediction.git |
| **Please write a summary of your Capstone project and its components. Describe the problem statement, your approaches, and your conclusions. (200-400 words)** |
| **Problem Statement :**  This project is aimed at predicting the case of customers’ default payments in Taiwan. From the perspective of risk management, the result of predictive accuracy of the estimated probability of default will be more valuable than the binary result of classification - credible or not credible clients. We can use the K-S chart to evaluate which customers will default on their credit card payments.  **Objective**: The objective of our project is to predict which customers might default in the next months  **Summary**:  Before we start our ML-classification Project process which is first Data Reading, Data Preparing, Data Cleansing, Data pre-processing, Data Analysis, EDA, and Machine Learning Model.  Our First Step was to import the dataset and check for duplicate and null values if there but luckily, we see that there is no null values or duplicate values.  After that, performed EDA for visualization. And after selecting independent and dependent variables pass it to the next step which is to train models. then applied the machine learning model ( Logistic Regression, Decision Tree, Random forest, XG-Boost, and KNN Classifier) and applied hyperparametric Tunning.  **Conclusion:**  1)first We started with data inspection, viewed the data distribution  2)By Visualization we have checked the distribution of defaulters vs non-defaulters and we see around 78% are non-defaulters and 22% are defaulters.  3)the distribution of sex, Education, and Marriage with respect to the defaulter. and we found in Sex more defaulter is Female, in Education, more number of the defaulters is a university students and in Marriage more number of the defaulters by single.  4)After that we built a model(Logistic Regression, Decision Tree, Random forest, XGBoot classifier, and KNN), and all of them in, the best accuracy has obtained from the Random Forest Classifier.  5)Using a Logistic Regression classifier, we can predict with 86.38% accuracy, whether a customer is likely to default next month. Using the Decision Tree classifier, we can predict with 82% accuracy whether a customer is likely to default next month OR not. Using Random Forest, we can predict with 87% accuracy whether a customer will be a defaulter in the next month. Using XGBoost Classifier, we can predict with 86.64% accuracy whether a customer will be a defaulter in the next month. And By applying KNN Classifier with 85% accuracy whether a customer will be a defaulter in the next month.  6)From the evaluation table Logistic regression model has the highest recall, if the business cares about recall the most, then this model is the best candidate. If the balance of recall and precision is the most important metric, then Random Forest is the ideal model. but Since the Random Forest classifier has also a higher Recall score.so I would recommend Random Forest.  7)From the above evaluation table we can also see that the Random forest Classifier having Recall, F1-score, and ROC Score values equals 90.46%, 87.08%, and 87.65% resp. and XGBoost Classifier having Recall, F1-score, and ROC Score values equals 90.29%, 85.95%, and 86.89% resp.  8)From the models that are applied to the dataset, We can conclude that these two Random Forest and XGBoost are giving the best evaluation metrics(Recall, F1-score, and ROC-AUC score) and with the help of these two models we are the best to predict whether the credit card is the default or not default according to our analysis. |