**Capstone Project Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| **Name: Sayali Kamalapurkar**  **Email:** [kumthekar.sayali19@gmail.com](mailto:kumthekar.sayali19@gmail.com)  **Contribution:**   * Data collection * Understanding the data variables * Importing python libraries * Data Preprocessing * EDA on features * Handling Imbalanced Data using SMOTE oversampling technique * Feature Engineering * One Hot Encoding * Implementing ML Classification models * Hyperparameter tuning using GridsearchCV * Selection of metric and model to use * Conclusions |
| **Please paste the GitHub Repo link.** |
| GitHub Link: [sskamalapurkar/Supervised\_ML\_Classification\_Capstone\_Project (github.com)](https://github.com/sskamalapurkar/Supervised_ML_Classification_Capstone_Project) |
| **Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)** |
| A Taiwan-based credit card issuer wants to better predict the likelihood of default for its customers, as well as identify the key drivers that determine this likelihood. This would inform the issuer’s decisions on who to give a credit card to and what credit limit to provide. It would also help the issuer have a better understanding of their current and potential customers, which would inform their future strategy, including their planning of offering targeted credit products to their customers.  **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.  **APPROACHES:** Firstly, we imported the dataset and python libraries for data exploration, visualization, also from sci-kit learn library imported useful algorithms. Then did some data wrangling to understand more about data and independent variables, get statistical information about the numerical features. Further, renamed some of the columns and replaced values in categorical columns for better understanding. Looked for duplicate values.  Further, did exploratory data analysis – Univariate, Bivariate and Multivariate. Then, handled imbalanced data using oversampling technique SMOTE. Then, created some new columns out of the highly correlated columns. And finally implemented four classification models – Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine. Also performed hyperparameter tuning using GridSearchCV to make models work better.  Understood evaluation metrics to be more precise in classifying defaulters and non-defaulters. At last concluded that Random Forest is the best and robust classification model to identify defaulters with 82% recall score.  **CONCLUSIONS:**   1. According to RandomForest model, features like paid\_total, total\_bill, dues, cred\_lim, age, repay\_status\_sep\_2, repay\_status\_aug\_1 and marriage status as married are found to be most important features to predict future defaulters. 2. If the credit card holder has paid minimum credit amount for past 6 months, also for every month if the dues of that customers are increasing, then it is obvious that the customer will have default for sure! 3. While this age and limit\_bal are the other two important predictors as we discussed in the data preparation part. This also makes sense because as one gets older, one is more likely to accumulate more resource and cares more about his reputations, which makes credit default less likely. Also, if one and one’s family get more given credits, the person is more likely to live in a wealthier environment which also makes credit default less likely. 4. Further, repay status is also playing a vital role in predicting future defaulters. If the creditor is not repaying for the past one or more months, he is more likely to default in upcoming months. |