

## Capstone Project Submission

### Instructions:

- i) Please fill in all the required information.
- ii) Avoid grammatical errors.

### **Team Member's Name, Email and Contribution:**

→ **Aniket Suresh Satpute**

**Email-id:** - aniketsatpute000@gmail.com

→ **Contribution-**

- Checking the unique values for Analyzing the Dataset for further analysis.
- The relationship Between the Variables and Default.
- Is the proportion of defaults correlated with age?
- Is the default proportion affected by credit limit?
- Substituting the values and finding Correlation Between the Variables
- Adding some more features that will be used in the models.
- Feature Selection, Scaling, Splitting and Balancing the Data.

→ **Kaiwalya Dashrath Zankar**

**Email-id:** - [kaiwalyazankar@gmail.com](mailto:kaiwalyazankar@gmail.com)

→ **Contribution -**

- Checking the presence of NAN values by using the tool missing value matrix.
- Is the proportion of defaults correlated with gender?
- Is the default proportion affected by marital status?
- Statistical inference
- Feature Engineering and encoding categorical data.
- VIF and Deviance Reduction for Feature Selection (in Logistic Regression)
- Implementing the model:
  - ◆ Logistic regression
  - ◆ Gradient boosting classifier.

**Please paste the GitHub Repo link-**

**GitHub Link:** - [Aniket-Satpute/Credit-Card-Default-Prediction \(github.com\)](https://github.com/Aniket-Satpute/Credit-Card-Default-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)**

Credit risk has traditionally been the greatest risk among all the risks that the banking and credit card industry are facing, and it is usually the one requiring the most capital. Despite machine learning and big data have been adopted by the banking industry, the current applications used by the banks are mainly focused on credit score predicting. The disadvantage of heavily relying on credit score is banks would miss valuable customers who come from countries that are traditionally underbanked with no credit history or new immigrants who have repaying power but lack credit history. Due to the scope of the project and lack of computational resources, this analysis is not intended to be exhaustive, we only applied 3 classification machine learning models.

The dataset is details of credit card holders of an "important bank in Taiwan" for the period April to September, 2005. The features available include some basic customer demographics (gender, education, marital status and age), available credit line, their history of payment/default for the six months mentioned (Apr--Sep '05), their bill amounts and their payment amounts for that period and a binary target variable indicating default the following month.

I perform some EDA to understand the data and clean the data, engineer relevant features, build predictive models to predict default and perform some statistical analyses to obtain a greater understanding of the features and their interactions. I finish with some business case scenarios where the predictive model could be applied.

Relationship between the variables and default to get more insights like-

- there are more female credit card holders, the default proportion among men is higher.
- The default proportion decreases with higher education level.
- Married people have higher default proportions than single folks.
- The age group of 30s has the high number of counts of defaults.

In feature engineering we have add some of the features like credit\_limit, usage to get required output and also ease further analysis. We also encoded some categorical data like sex, marriage and education.

We used different type of classification algorithms to train our model like, logistic regression, Random Forest Classifier, Gradient Boosting Classifier also found the important features for training the model. Out of them random forest with tuned hyperparameters gave the best result.

**Drive Link:-** [Credit Card Default Prediction - Google Drive](#)