The project is about detecting if the customer will be defaulted or not based on 11 features and the dataset includes 32581 records. As it is a classification problem, I have used 7 machine learning algorithms to find the best one using Python & Jupyter Notebook in Anaconda Environment. The steps are below :

**Step 1 : Exploratory Data Analysis**

* **Finding Missing Values :** No missing values found in any feature
* **Numerical Variables & Their Distribution :** There are total 7 numerical Variables & most of them was not following the assumption of normality as we can see in Histogram.
* **Categorical Variables & Their Cardinality :** There were 4 categorical features. Only 1 of them contained 2 categories and others contained more than 2 categories.
* **Relationship Between Dependent & Independent Variables** : Correlation matrix was used for numerical features and Chi square test was used for Categorical features.
* **Outliers Detection :** Through Box & Whisker plot , outliers were detected.

**Step 2 : Feature Engineering**

* **Dummy Creation :** One Hot Encoding was used to create the dummies.
* **Transformation :** The numerical features were transformed into Log scale in remedy of violating the assumption of normality.
* **Feature Scaling :** Standardization & Min\_Max scaler both were used for feature Scaling.
* **Train\_Test Split**

**Step 3 : Feature Selection**

* Lasso regression was used to select the features. As it assigned weights close to zero to the unimportant features, So we can select the important features. After Creating the dummies, we got 21 features from which only 12 was selected for modelling.

**Step 4 : Data Modelling**

Using K fold cross validation and grid search cv , I tried to find the Optimized Hyperparameter for these models.

* Logistic Regression
* Support Vector Machine (SVM)
* Decision Tree Classifier
* Random Forest Classifier
* ADA Boost
* Gradient Boost
* Naive Bayes

**Step 5 : Model Evaluation**

Classification Matrix and the accuracy score was used to evaluate the best model.

* Logistic Regression ----------------------------- .8586
* Support Vector Machine (SVM)----------------- .9035
* Decision Tree Classifier ------------------------- .9106
* Random Forest Classifier------------------------ .9160
* ADA Boost--------------------------------------- **.9180**
* Gradient Boost----------------------------------- **.9183**
* Naive Bayes-------------------------------------- .8474

Ada Boost & Gradient Boost Model was the best among all and they have

macro avg 0.90 0.84 0.87

weighted avg 0.91 0.91 0.91

\*The outliers were not influencial so I kept them in the model.

\* The interpretation of the model and some visualization will be presented in part 2 IN SHA ALLAH.