**Capstone Project Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| Shajad [shehzadglocal786@gmail.com](mailto:shehzadglocal786@gmail.com) |
| **Please paste the GitHub Repo link.** |
| Github Link:- <https://github.com/Shajad121/Credit-card-fefault-prediction.git> |
| **Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)** |
| This project is about the Credit card default prediction analysis. When we start this project, it is very difficult to understand and look very complex. In first step we upload the data and then we start to find the duplicate values and nulls values and missing values etc. but in this data there is no missing values and duplicate values. After completion of 1st step in 2nd step we plot the data in form of graph, pie plot and bar plot and line plot and also box plot, and in box plot we find some outlier but these outliers are not affect the result because they are very less in number. And we also fine some hidden information by plot the graph like there are 34% male are defaulter but 21% women are defaulter it means that the women are less defaulter as compare to male. And also from graduate school defaulter that is 19% then university comes with 24% defaulter then high school comes with 25% defaulters, according to this chart from age 60-70 the default percent 29% present in this group. from the age group 20-30 There are 7421 customers are not defaulters but 2197 customers are defaulters. From the age group 30-40 there are 9862 customers are not defaulters but 2276 customers are defaulters. from age group 40-50 there are 4979 customers are non defaulter but 1485 customers are defaulters. from age group50-60 there are 1759 customers are non defaulters but 582 customers are defaulter. from the age group of 60-70 there are 225 customers are non default but 89 customers are also defaulters.in SEX plot single customers are least defaulter with 21% then married customers are comes with 23.5% defaulters After plot the data we plot the heatmap of the data and then we find the relationship between the feature and we did some feature engineering in the data. We remove the labels of the data because that labels are not readable and difficult to understand and we make the data’s 1st row as the labels of the data. After the feature engineering we apply some algorithm, 1) Logistic Regression, 2)Random forest,3)XG Boost and we also use the dummy model to describe the our model and make them easy to understand. And we also do some cross validation in every algorithm. After applying these algorithm we select the Random Forest as a best algorithm that gives the best score. It gives the precision 0.668 and recall 0.364 and F1 score 0.471 percent. After selecting the algorithm we do some feature engineering and do some cross validation but there is not much change in the result. Logistic Regression model has the highest recall but the lowest precision, if the business cares 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. Since Random Forest has slightly lower recall but much higher precision than Logistic Regression, I would recommend Random Forest. |