CREDIT CARD DEFAULT

AI FINAL INDIVIDUAL ASSIGNMENT

SUBMITTED BY: GUPTA, DIVYA

MATRIC NO: G2200597K

# **EXECUTIVE SUMMARY**

# As an after effect of the COVID’19 pandemic and the slow mindsight shift towards the cashless economy, there is a trend more and more people are moving towards cashless payment systems. The pandemic in turn has also reduced the propensity to save for people, which force them to rely on credit system. An easy and effective way to get credit readily is the use of credit cards. Consumers get the free hand on spending how much ever they would want. As the demand for money on credit shoots up, to cover their financial cost, banks are offering higher interest rates. Now for each higher rate the bank charges, the consumer has to go ahead with more credit, which ultimately leads to huge number of defaults.

In this scenario, both the consumer’s need for credit and banker’s need for higher interest, banks nowadays are facing huge defaults on credit cards. This phenomenon particularly destroys the financial stability of the bank as well as hinders the borrower’s credibility.

Keeping the above complications in mind, the study on analyzing the credit card default has been focused on three major business issues:

1. **Identifying the highly likely individuals who are to default**
2. **Setting a credit card limit, depending on which individuals are risky in terms of default**
3. **Mitigating the financial loss caused by the defaults, in terms of new policies**

The current study uses machine learn models to predict categorical variable ‘Default or No Default’ using a few independent variables which might impact the default on credit card payments. The models will test the accuracy of prediction using the confusion matrix and look for how many correct predictions we can derive. Since this is a multi-model study, we can effectively judge which model to use in bank’s BAU.

With the implementation of this model, the bank will able to recognize the individuals who are likely to default and accordingly they can reject their credit card application. Banks can work on their credit card management rate as per the default status of a particular individual. This model thus will be a quick evaluation for the bank’s consumers, wherein it can target its policies and pricing with respect to credit cards.

The complete data visualization and the model working have been provided in the subsequent parts of this report.

# TABLE OF CONTENTS

[Introduction 2](#_Toc117435549)

[Business Problem and Key Deliverables 2](#_Toc117435550)

[Data Description and Data cleaning 2](#_Toc117435551)

[Data Visualization 3](#_Toc117435552)

[Modelling and Model Comparison (Pros and Cons of each model) 5](#_Toc117435553)

[Business Implementation 7](#_Toc117435554)

[Future works and improvement 8](#_Toc117435555)

# Introduction

This project deals on mitigating the loss from credit cards defaults to the banks. Default is the failure to make required interest or principal repayments on a debt/credit. Nowadays since the economy is moving towards cashless transactions, people are looking more and to usage of credit cards. However, when using a credit card generally they lose the sense of how much money they are spending, which ultimately is the credit from the bank and need to be paid within a stipulated time. If the credit amount has not been paid after that designated time, it results in a credit card default. This issue has been considered a headache for the banks who need to take complete evaluation before setting the credit limit for the customer’s credit card. They need to be aware about the purchase behavior and credit history of the customer to completely understand if he/she is going to default. On the other side, a default on credit card impact the consumers as well, as that default stays on the consumer profile for 6 years, impacting their credit history and potential to apply for loans and other credits. Post COVID when people have lost their propensity to consumer, and with little cash or savings in hand, they are moving more towards the credit cards. However, because of higher risk of default associated with these groups, higher interest rates are charged to these borrowers. These people in turn end up taking multiple credits to fulfill the interest obligation and end up defaulting many of them. This problem in itself become the vicious circle and impact the financial stability of the banks issuing these credit cards.

# Business Problem and Key Deliverables

With this model, the business problem that deserves the maximum attention is maintaining the financial stability of the of the bank. The key problem can be broken down into two parts given below:

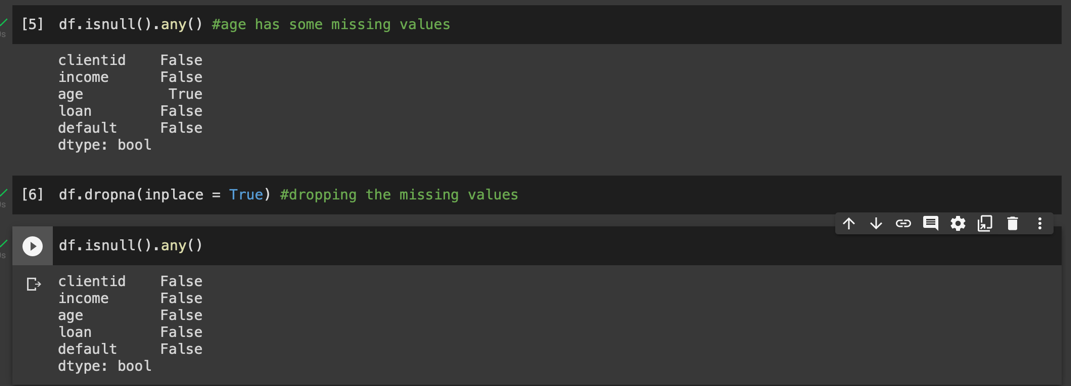
* How to set a credit card limit to minimize the defaults?
* How to mitigate the loss and manage financial risk, if the default has already happened?

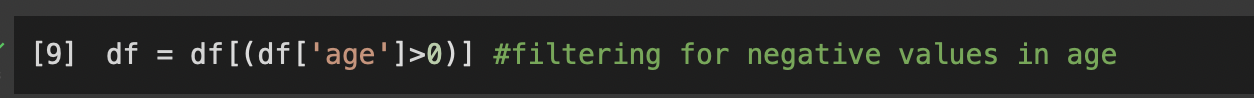
They deliverables of this project will aim to answer how to build comprehensive financial system by collecting relevant data and using machine learning models to predict which borrowers are likely to defaults, helping prompt action to manage financial loss mitigation associated with credit card defaults.

# Data Description and Data cleaning

The dataset targets the 3 major variables namely, **Age, Income, Loan, and Default**. Considering ‘**Default**’ as the dependent Y variable, other 3 will be used as the independent variables to predict Y.

On careful analysis of the independent metrics, **Age** is found to have three missing values which have been removed for better usability of data. Also, there are a few negative values in **Age** column again, which have been removed as well. Additionally, the **Client ID** column has also been removed because it is not required to make our model predictions and finally the new data set 1994 entries has been used for the analysis.

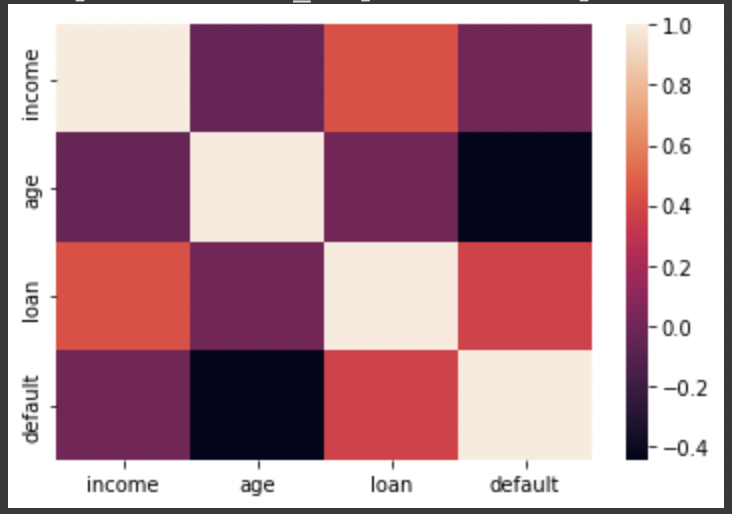






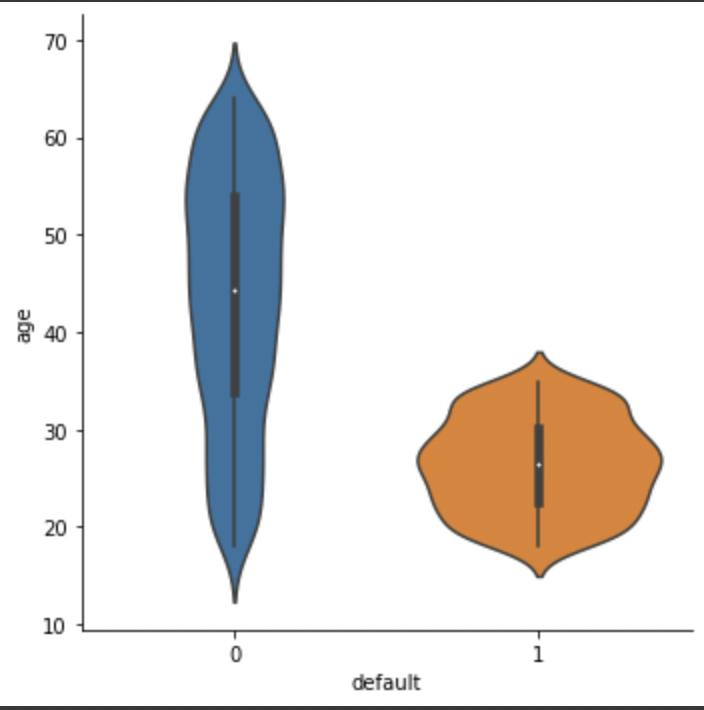
# Data Visualization

As the very first step to understand the data, a correlation analysis has been conducted using seaborn software. Below is the correlation plot to check which independent variables impact the Default rate and how. Also, the correlation insights are mentioned below:

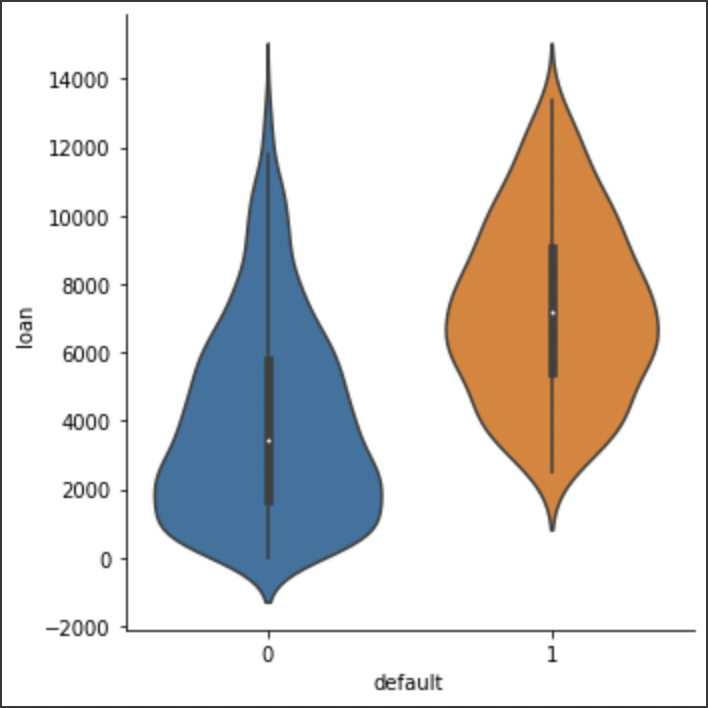


* In the above picture, it is evident loan (amount) is positively correlated with the default rate, which means higher the loan amount, higher will be the default rate.
* Secondly, age is negatively correlated to the default rate i.e. younger population tends to default more.
* Lastly, in terms of income, people with lower income tend to default than those from the higher income level.

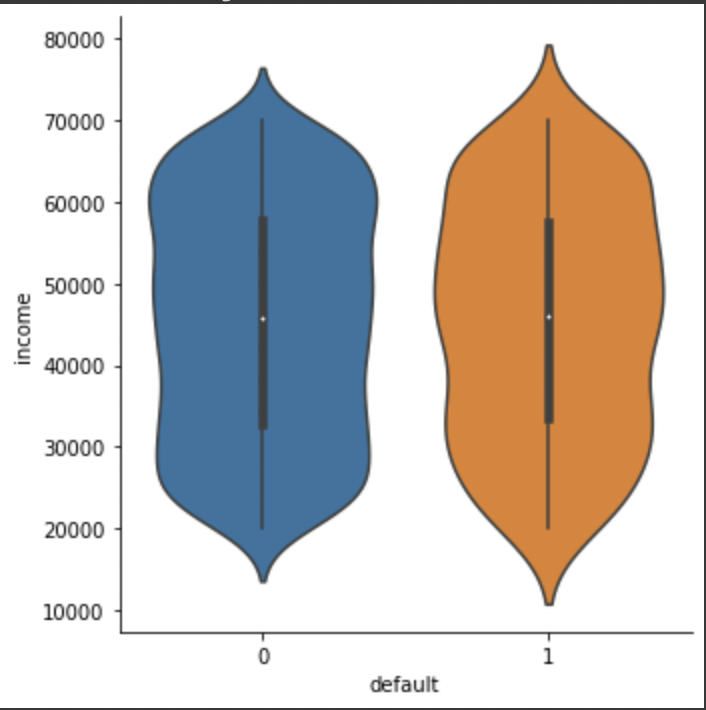
The above relationships can more clearly be studies from the violin charts given below:



*Younger people likely to default more (0- no default, 1-default)*



*People with higher loan, likely to default more (0- no default, 1-default)*



*People default irrespective of the income, however people with lower income are likely to default more (0- no default, 1-default)*

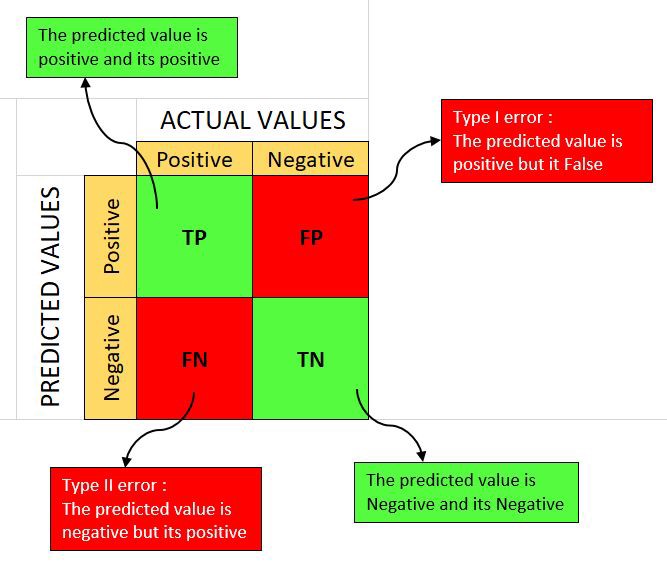
# Modelling and Model Comparison (Pros and Cons of each model)

Before designing the models, the data is first made compatible to use. Since there are no actual values available as of now, the current data has been used for both train and test. The data has been split using the **train\_test\_split** function of Python. Post splitting the data, we Y variables i.e. Default is skewed for 0 (no default) with 1283 values, whereas 1(default) has only 212 values. To make the data uniform, oversampling is done for Y=1, using the **SMOTE** functionality in Python.

After the train test split and oversampling, the independent variables (age, loan, income) are normalized to include all the values of the data between normalized range. This ensures that all the outliers are not impacting the accuracy of any model.

Since, the dependent variable is the categorical variable, the accuracy of the model will be based on the confusion matrix, which shows 4 boxes in this case i.e.:

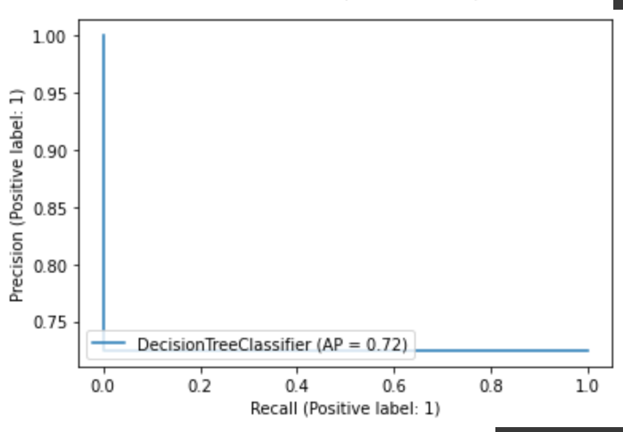
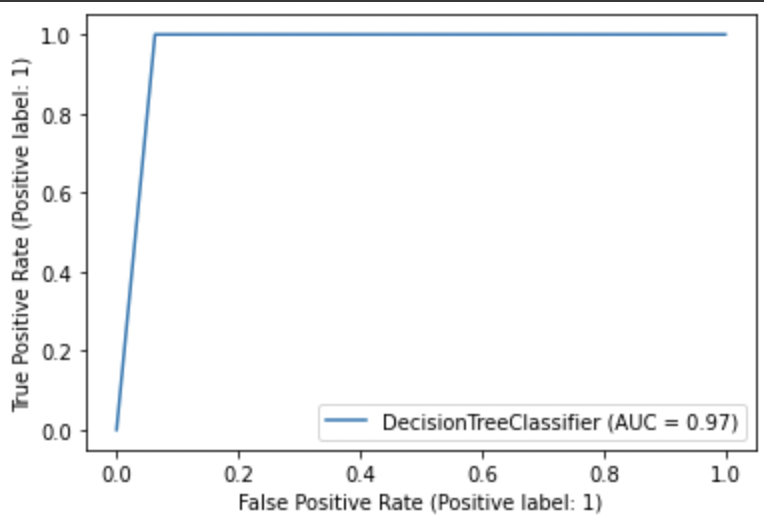
* True Positive – Meaning the predicted value is same as the actual value both default
* True Negative – Meaning the predicted value is same as the actual value, both no default
* False Positive – Meaning the predicted value is predicted as defaulted, however the actual value is no default
* False Negative – Meaning the predicted value is predicted as no defaulted, however, the actual value is default



For modelling, we are going to compare the results using Logistics Regression, Decision Tree, Random Forest, XGBoost, Neural Network. The results in terms of accuracy have been displayed in the below table:

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| Logistic Regression | 87.50% |
| Decision Tree | 94.50% |
| Random Forest | 94.18% |
| Gradient Boosting | 93.78% |
| Neural Network | 90.78% |
| Neural Network (Keras) | 89.50% |

From the above model comparison table, it is evident that Decision tree is performing the best when it comes to predicting the default rate on credit cards. Also, looking into ROC curve (which signifies the relationship between true positive and false positive values with the model), the AUC (area under the curve) is 97% with precision recall at 72%.



The ROC curve and precision recall curves for other models have been added to Appendix.

Finally, analysing the pros and cons of each model:

* **Logistics Regression** - One of the advantages of Logistics regression is that is the easiest model to implement, and always a great a accuracy for the simple data sets. However, Logistics Regression considers a linear relation between the independent and dependent variable and cannot thus measure any non-linear data. It is thus too simplified for complex data such as this, when it becomes tough to trace relationship between the variables under this model.
* **Decision Tree** - One of the major advantage of a decision tree is that it dies not requires scaling or normalisation of data which is bringing the difference accuracy in the current dataset, while requiring less effort to pre-process the same data. However the analysis for the decision tree is based on just tree and it fails to build upon different angles of the model as it's improved version Random forest or gradient boosting does. Also, changing a little value in the whole model can lead the decision tree model to collapse.
* **Random Forest** - Random forest is an improved version of the decision tree model and analyses the data based on multiple trees, following an ensemble model. The multiplicity of trees avoids the issue of overfitting, reducing variance and increasing accuracy. Here as well no scaling is required, which can still lead to higher accuracy. A major drawback is complexity of the model, as it requires the construction and combining a lot of trees and because of this random forest takes much more time to train the dataset than other models.
* **Gradient Boosting** - Gradient boosting generally train faster especially on large datasets. In this model, the missing values are handled proactively. However, this is generally considered the weak model, because the model is quite prone to overfitting and computational and time expensive.
* **Neural Network** - A neural network is designed to learn and improve its results continuously. Once the system is trained, it can produce output without the need for complete inputs, which makes it more user friendly. Neural network enables effective data retrieval as it is not dependent, and all the data is saved over the cloud. However, most neural networks are black-box systems generating results based on experience and not on specified programs, making it difficult for modifications.

# Business Implementation

* **Setting credit limit**

Setting a credit card limit as per the age and income of the borrower can help that they do not get to spend more than their limit. People at a younger age can have a lower credit card limit Same goes for lower income level.

* **Higher interest on likely defaults**

With this model, we would be able to know which people are prone to defaulting the credit card payments, these individuals can be charged high credit card interest varying after every 30 days. Also, the credit card management fee can also be increased after each failure of payment within stipulated period.

* **Auto debit on likely defaults**

An auto debit payment function can be enabled on the bank accounts of highly likely to default individuals. This will trigger an auto payment from their linked bank account. In this, maintaining a minimum bank balance would need to be ensured too.

* **Setting an income to offer free credit cards**

Setting an income limit when the credit card would be offered for free, will discourage people on lower income to sign up for the credit card, hence reducing the default rate.

# Future works and improvement

* Avoid oversampling, by getting more representative data

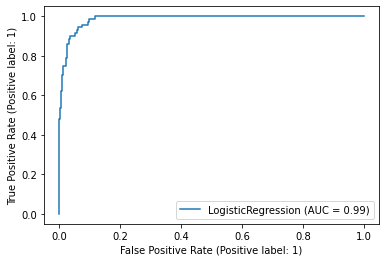
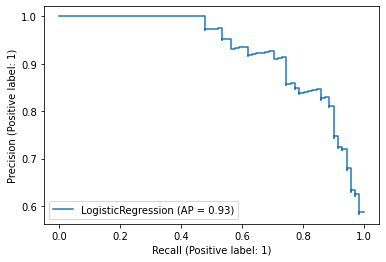
Since our model has to oversample the data on default values, the predictions may not be completely accurate, as the data is only imputed. For better accuracy and reproducibility, a more representative data in terms of both default and no default can be considered.

* Adding more parameters

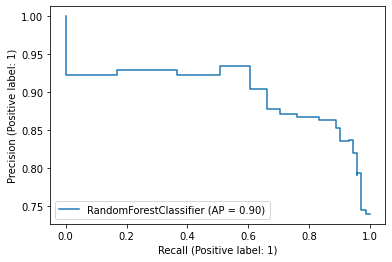
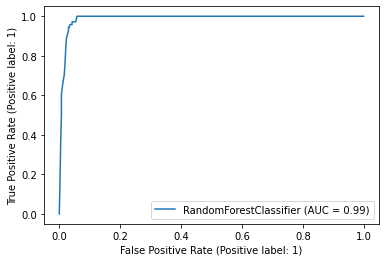
The current model deals with only 3 parameters to predict the credit cart default, more parameter such as interest rates, other loans can be taken into consideration for better operation of this model.

# APPENDIX

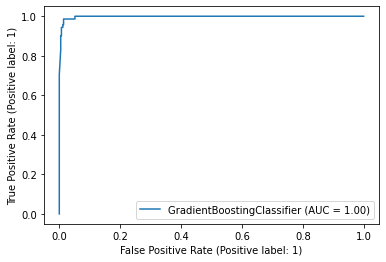
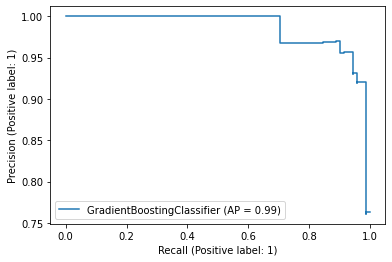
1. ROC and Precision-Recall Curve for Logistic Regression

1. ROC and Precision-Recall Curve for Random Forest



1. ROC and Precision-Recall Curve for Gradient Boosting

1. ROC and Precision-Recall Curve for Neural Network

