**Credit Scoring Model**

A credit scoring model is a statistical model that is used to predict the likelihood that an individual or business will default on a loan or credit obligation. Credit scoring models are commonly used by financial institutions to assess the creditworthiness of loan applicants and to determine the terms and conditions of loans, including the interest rate and the amount of collateral required. Credit scoring models use a variety of inputs, including credit bureau data, financial statements, and other demographic and financial information, to predict the likelihood of default (Hussin Adam Khatir and Bee, 2022).

# **Research Question One**

In this problem, we seek to build a credit scoring model that will aid in the prediction of serious loan defaulters based on some explanatory variables.

# **Data**

The data used in this analysis had 10,000 observations. These observations were borrowers who had taken loans of which some had repaid while others had defaulted seriously. The response variable captures the default status, that is, 1 if the borrower defaulted and 0 if the client did not default on the loan. There were 10 predictor variables which contained information about the 10,000 borrowers.

# **Data Pre-Processing**

Data pre-processing is the process of preparing data for analysis or machine learning. It involves cleaning, formatting, and organizing data in a way that makes it ready for further analysis or modeling. One common step in data pre-processing is data cleaning, which involves identifying and correcting or removing inaccuracies and inconsistencies in the data. This may include dealing with missing values, detecting and fixing errors, and eliminating duplicates. Data formatting involves organizing and structuring the data in a way that is suitable for the intended analysis or modeling. This may include converting data into a specific format, such as transforming categorical data into numerical form, or scaling numerical data to a common range. By taking the time to properly pre-process data, analysts and machine learning practitioners can improve the quality and reliability of their results (Mishra et al., 2020).

## **Dealing with Missing values**

The first step I undertook in this analysis was to check for any missing values in the credit risk dataset. There were no missing values in the 11 variables but monthly income and number of dependents variables. Monthly income variable had 1974 missing values which represented about 19.74% of the total observations while number of dependents had 284 missing values which represented about 2.84% of the total observations. The best way to handle these missing values was to replace these missing observations in the dataset. We cannot drop almost 20% of the observations in the dataset as this may affect the performance of our models. If we drop them, we may end up having biased results and conclusions.

According to Sainani (2015) there are a number of ways of dealing with missing values. These includes:

1. Ignore the missing values: This approach is only suitable if the missing values are few and do not affect the overall analysis. However, ignoring missing values can lead to biased or incomplete results.
2. Remove the rows or columns with missing values: This approach is only suitable if the missing values are few and do not affect the overall analysis. However, removing rows or columns with missing values can result in a loss of important information.
3. Replace the missing values with a statistical measure such as the mean or mode or median of the remaining values: This approach can be suitable if the missing values are few and the distribution of the remaining values is relatively normal. However, replacing missing values with a statistical measure can distort the distribution of the data.
4. Use a machine learning algorithm to predict the missing values based on the other values in the dataset: This approach can be suitable if the missing values are few and there is a strong relationship between the missing values and the other values in the dataset. However, this approach requires additional processing time and may not always produce accurate results.

In my problem case, the first variable (monthly income) is continuous in nature. I thus replaced all the missing values in this variable with the mean of the column. The second variable with missing values, number of dependencies, has most values concentrated around a certain number which is 0. We thus replace all missing values in this column with 0.

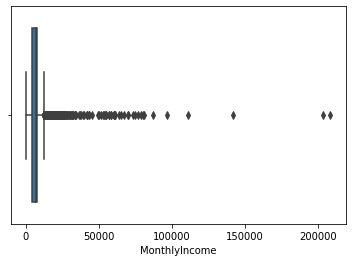
## **B. Outlier Detection**

Outliers are data points that are significantly different from other data points in a dataset. They can be caused by measurement error, or they might represent true variations in the data. Outliers can have a significant effect on statistical analyses, so it’s important to identify and treat them appropriately. Outliers can be detected using various statistical methods such as the boxplot method, the Z-score method among other methods. Contacting summary statistics can help us note observations which are far apart (Wang et al., 2019).

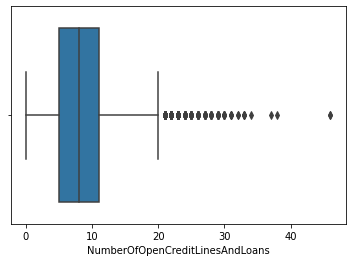
In this analysis, I used both the descriptive statistics of the dataset and the boxplots to identify the outliers in the dataset.

From the descriptive statistics, we clearly note that the variables have significant outliers. For instance, looking at the debt to ratio variable, the values for this variable cannot logically exceed 1 as debt ratio is typically expressed as a percentage, so the range of values is 0% to 100% (0.0 to 1.0). But from the descriptive statistics, we clearly see that the maximum value for this variable is 168835 which is impossible. Any value for this variable that is above 1 was dropped. The variable “**RevolvingUtilizationOfUnsecuredLines”** also has significant outliers. Considering the boxplot, we shall drop the observations whose values are greater than 1. The boxplots below of some of the variables clearly indicating the outliers in the dataset.

***Plot1: Boxplot of monthly income***



***Plot2: Boxplot of number of open credit lines and loans***



***Plot3: Boxplot of number of times (90 days late)***



***Plot4: Boxplot of total balance on credit cards***



Generally, there are outliers in the credit score dataset. Closely evaluating the min, median, 3rd quantile and the max values of the variables, I dropped the values which were extreme from the rest of the observations. I also dropped the impossible values from the dataset.

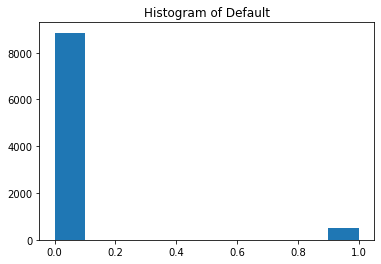
Having dropped the outliers from my dataset, I remained with 9378 observations from 10,000 observations.

# **C. Exploratory Data Analysis**

This type of analysis typically involves visualizing the data in various ways, such as plotting graphs, to gain an understanding of the underlying trends and patterns in the data. EDA is often used to identify outliers or anomalies in the data and to verify relationships between variables. It is also used to identify potential explanatory or predictive variables that may be used in further analyses.

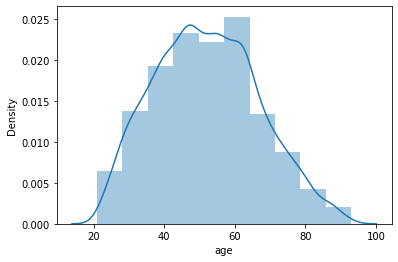
In this data exploration, we have already dwelt with outliers and anomalies in the dataset.

***Figure1: Histogram of Serious default***

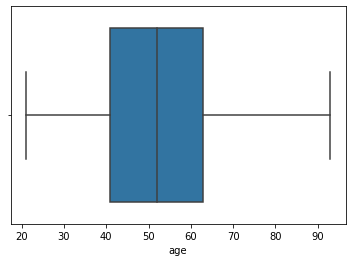


From the plot, we can see that majority of the borrowers repaid their loans while only a few defaulted the loans.

***Figure2: Histogram of Age distribution of the borrowers***

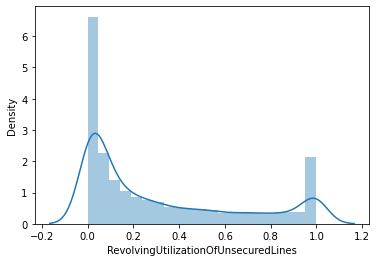


***Figure3: Boxplot of Age***

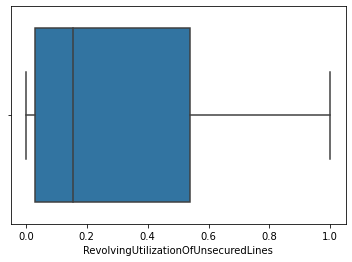


From the visualizations of the age variable, we clearly see that there are no outliers. Majority of the borrowers are of the age of 40 to 65 years.

***Figure4: Histogram of total balance on credit cards***



***Figure5: Boxplot of total balance on credit cards***

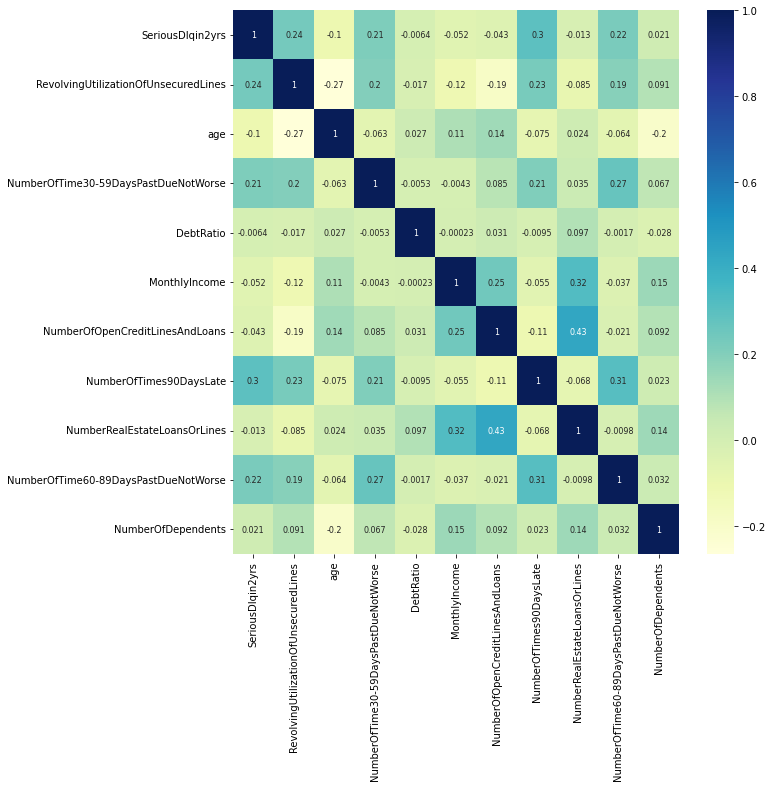


From the histogram and box plot of total balance on credit cards, it is clear that the credit balances are varied across the borrowers. There are no outliers as they had been removed before.

***Figure6: Comparison of the boxplots of total balances on credit card with outliers and after outliers were removed***

|  |  |
| --- | --- |
| Boxplot before outliers were removed | After outliers were removed |
|  |  |

***Figure7: Heatmap of the correlations between the variables***

******

The heatmap above shows the associations that exists between different variables. For instance, we see a positive association between the default status and total balances on one’s credit card. There is a strong positive correlation between default and the number of times one has delayed to repay the loan.

# **D. Splitting the Dataset**

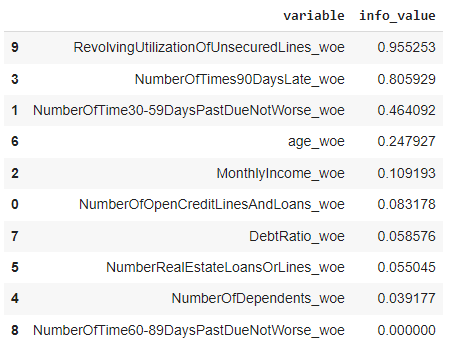
The dataset was split in the ratio 7:3, that is, 70% train and 30% test datasets.

# **E. Weights of Evidence Transformation**

Weights of Evidence is a technique used to quantify the impact of a particular independent variable on the dependent variable. It works by assigning a numerical value to each value of an independent variable. I performed variable binning after which coded I the variables with the weights of evidence.

Variable binning is the process of dividing a set of continuous or ordinal variables into a set of bins or categories. This is often done to transform a continuous or ordinal variable into a categorical variable, which can be useful for certain types of data analysis and modeling.

Having performed variable binning and added weights of evidence, I did variable filtering to remove the variables which were not predictive. A variable with information score less than 0.1 was removed from the dataset. Below are the variable names and the information value of importance.



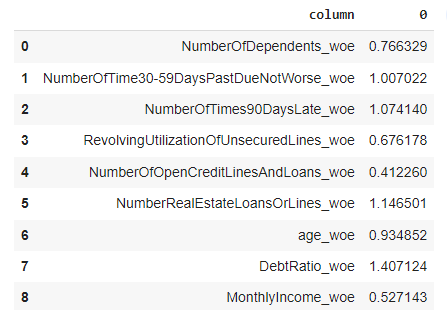
From the table, NUMBEROFTIMES60-89DAYSPASTDUENOTWORSE\_WOE variable was removed as it was not predictive according to the table results. The variables are in order of their predictive power from the most to the least predictive one.

# **Logistic Regression classifier**

**Variable impact on the response**

The coefficients of the logistic regression classifier are presented in the table below in which we clearly see that all the coefficients have a positive sign implying a positive effect on the response variable with non-defaulters as the base variable.

*Table of Logistic Regression Coefficients*



**Model Performance**

To evaluate the model performance, I considered various metrics which can are computed from the confusion matrix.

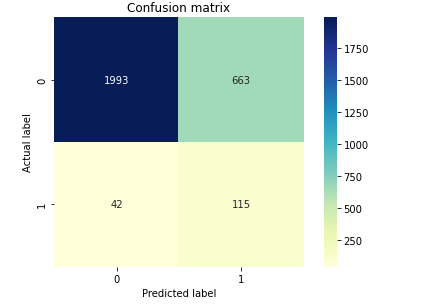
Precision: Precision is a measure of how well a model is able to identify the relevant data points for a given problem. It is the ratio of true positives to the sum of true positives and false positives.

Recall: Recall is a measure of how well a model is able to identify all relevant data points for a given problem. It is the ratio of true positives to the sum of true positives and false negatives (Novakovic et al., 2017).

In our problem case, True Positive refer to the borrowers who actually defaulted, True Negative are those borrowers who were actually repaid their loans, False positive are the borrowers who repaid their loans but they were classified as defaulters whereas False Negatives are the borrowers who defaulted their loans but they were classified as those who repaid their loans.

The figure below shows the classification report.

***Figure 8: Confusion Matrix for Logistic Regression Classifier***



From the classification report, we clearly see that 1993 out of 2656 borrowers who repaid their loans were correctly classified. It is also clear that 115 out of 157 defaulters were correctly predicted by the model.

The precision of the model was 73.24% whereas the recall was 14.78%.

Coming to the Random Forest classifier, the performance report is also summarized in the confusion matrix as shown below.

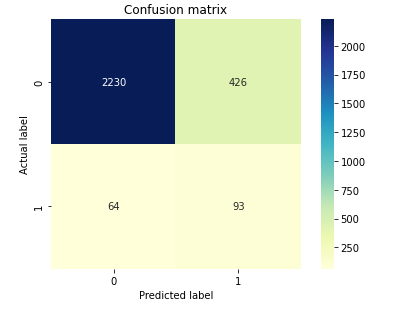
**ROC Curve**

An ROC curve (receiver operating characteristic curve) is a graphical representation of the performance of a binary classification model. It plots the true positive rate (TPR) on the y-axis and the false positive rate (FPR) on the x-axis, for different classification thresholds.

In general, A model is better if its ROC curve is higher up and to the left of another model’s ROC curve. Area under the curve (AUC) is a metric that quantifies the overall model performance. The higher the AUC, the better the model.

The logistic regression had an AUC of 80.9%

*Confusion Matrix for Random Forest Classifier*



For the Random Forest model, we clearly see that 2230 out of 2656 borrowers who repaid their loans were correctly classified. It is also clear that 93 out of 157 defaulters were correctly predicted by the model.

The precision of the model was 59.23% whereas the recall was 17.91%.

**Comparison of scorecard for the logistic classifier and the random forest classifier**

|  |  |  |
| --- | --- | --- |
| **Metric** | **Logistic Classifier** | **Random Forest** |
| **Value** | **Value** |
| **Precision** | 73.24% | 59.23% |
| **Recall** | 14.78% | 17.91% |
| **AUC for ROC curve** | 80.9% | 79.5% |

Comparing the two models, we see that Logistic Regression model performed the better. This is because it has a greater AUC value than the random classifier. It is also good noting that according to precision, logistic performed better but random forest has a better recall value.

**Why banks typically use Logistic Regression as their base classifier? What do banks win and**

**lose by doing this?**

Banks typically use Logistic Regression as their base classifier because it is a reliable and efficient way to predict binary outcomes. It is a simple but powerful model that can easily be trained on limited data and can still produce results with high accuracy. Logistic Regression is also easy to interpret and explain, which is important when dealing with complex data sets. The main advantage of using Logistic Regression is its ability to give banks precise results. It is a reliable and relatively simple method of predicting outcomes and can yield accurate results with limited training data. The main disadvantage of using Logistic Regression is that it can be prone to overfitting. If the data set is too large or complex, the model can become too specific to the data and may not generalize well. Additionally, Logistic Regression may not be the best choice for certain types of data or problems, such as image recognition or natural language processing.

# **Question Two**

Title: "Predicting Credit Risk of Companies Using XGBoost and Deep Neural Networks"

Authors: Ting-Wei Chen, Ching-Hsien Hsu, and Chien-Lun Shih

Citation: INFORMS Journal on Computing, Vol. 32, No. 3, pp. 623-634, 2020

**Data Mining Problem**

The data mining problem considered in the paper was credit risk assessment for companies. The authors aimed to to develop machine learning models that can accurately predict the credit risk of companies, which is vital for financial institutions that need to make informed lending decisions.

**Data Mining Methodology**

The data mining methodology used in the paper was supervised learning approach. XGBoost and deep neural network (DNN) algorithms were adopted in the analysis where the authors preprocessed the data by selecting relevant features and standardizing the values. The models were then trained on a labeled dataset of historical credit risk data for companies.

**Results**

The results reported in the paper show that the XGBoost model outperforms the DNN model in terms of prediction accuracy, with an overall accuracy of 93.7% and a precision of 91.3%. The DNN model has an overall accuracy of 92.1% and a precision of 87.2%.

**A critical discussion of the model and results:**

The authors made the assumption that the historical credit risk data used to train the models is representative of the current credit risk environment. This assumption may not always hold true, as economic conditions and company characteristics can change over time. Additionally, the performance of the models may be affected by the choice of evaluation metrics and the specific hyperparameter values used. It would be useful to further evaluate the robustness and generalizability of the models to different datasets and conditions.

In terms of limitations, the authors note that their model is only applicable to predicting credit risk for Taiwanese companies, as the data used in the study consists of financial information from the Taiwan Stock Exchange. Additionally, the model does not consider company-specific information or macroeconomic factors, which could potentially affect credit risk.

Overall, the "Predicting Credit Risk of Companies Using XGBoost and Deep Neural Networks" paper presents a promising approach for credit risk prediction using machine learning. However, it is important to consider the limitations of the model and the specific context in which it is applied. Further research could potentially explore the incorporation of additional factors or the application of the model to different types of credit risk.

# **Question 3**

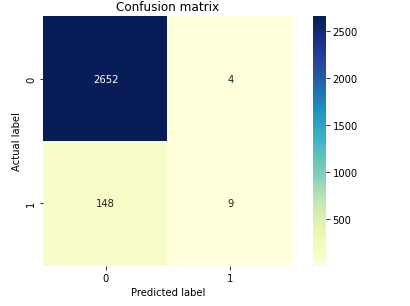
The paper I reviewed in question two had two models, the XGBoost model and the Deep Neural Network model. I choose to apply the XGBoost classifier to the credit score dataset. XGBoost is a popular and efficient implementation of the gradient boosting algorithm for decision tree ensembles.

**Analytic Steps of applying XGBoost Classifier to credit dataset**

1. Data cleaning to remove all the anomalies and dealing with the missing observations appropriately
2. Splitting the dataset into train and test set. This was in the ratio 7:3 with train set being 70% of the original dataset. The splitting was done randomly (data shuffled).
3. Having split the dataset, I converted the data into an XGBoost-compatible dataset, that is, in the form of an ‘xgb.DMatrix’ which is a special data structure that efficiently stores the data in a wat compatible with XGBoost’s implementation.
4. I then specified the hyperparameters of the XGBoost model then fitted the classifier on the dataset.

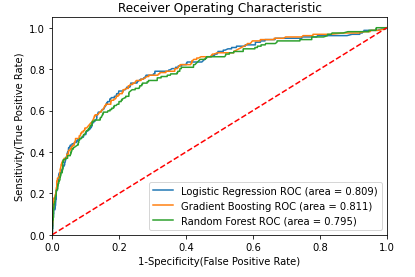
**Performance of the Model**

***Figure9: Confusion Matrix of XGBoost Classifier***



The XGBoost model was able to correctly classify 2652 non defaulters out of 2656 total defaulters. 4 non-defaulters were misclassified. Coming to total defaulters, only 9 were correctly classified out of 157 defaulters. The XGBoost model had very small values for precision and recall. It did not perform well when it comes to the prediction of loan defaulters.

**Comparison of the three Models using ROC curve**



According to the Area Under Curve (AUC), XGBoost model had the highes AUC followed by the logistic regression then the random classifier.

We can say from this metric that XGBoost model performed the best.

But when it comes to predicting defaulters, logistic regression is found to have the highest precision and thus the better model.

**Business Implication of using XGBoost model on credit data**

1. Improved risk assessment: By training an XGBoost model on credit data, you can improve the accuracy and efficiency of risk assessment, which can help you make more informed decisions about lending and credit risk management.
2. Enhanced customer segmentation: You can use an XGBoost model to identify patterns and trends in credit data that can help you segment customers into different risk categories, which can inform marketing and credit strategies.
3. Increased loan approval rate: By using an XGBoost model to predict the likelihood of loan repayment, you can potentially increase the approval rate for loan applications, which can help you expand your customer base and increase revenues.
4. Reduced default rates: By accurately predicting the risk of default for each loan, you can reduce the default rate and improve the overall performance of your loan portfolio.

It is however good to note that the specific business implications of using an XGBoost model on credit data will depend on the quality and characteristics of the data, as well as the specific goals and objectives of the analysis.

# **References**

Hussin Adam Khatir, A.A., Bee, M., 2022. Machine Learning Models and Data-Balancing Techniques for Credit Scoring: What Is the Best Combination? Risks 10, 169. https://doi.org/10.3390/risks10090169

Mishra, P., Biancolillo, A., Roger, J.M., Marini, F., Rutledge, D.N., 2020. New data preprocessing trends based on ensemble of multiple preprocessing techniques. TrAC Trends Anal. Chem. 132, 116045. https://doi.org/10.1016/j.trac.2020.116045

Novakovic, J.D., Veljovic, A., Ilic, S.S., 2017. Evaluation of Classiﬁcation Models in Machine Learning.

Sainani, K.L., 2015. Dealing With Missing Data. PM&R 7, 990–994. https://doi.org/10.1016/j.pmrj.2015.07.011

Wang, H., Bah, M.J., Hammad, M., 2019. Progress in Outlier Detection Techniques: A Survey. IEEE Access 7, 107964–108000. https://doi.org/10.1109/ACCESS.2019.2932769