DATA ANALYSIS TO PREDICT CREDIT CARD DEFAULTERS

Name – Ashutosh Nandkeolyar

Student ID – 501140245

Supervisor – Ashok Bhowmick

Date of Submission – December 05, 2022

Ryerson University

Table of Contents

1. **Introduction** 3
2. **Literature Review** 4
3. **Data** 11
4. **Data Preparation** 19
5. **Model Implementation** 19
6. **Conclusion** 233
7. **REFERENCES** 7-34

Table of Figures

Figure 1: Resulting Self-Organizing Map (With Markers) 9
Figure 2: features overview 13
Figure 3: Null values count 14
Figure 4: Data distribution of defaulter group 155
Figure 5: Distribution of target with gender 16
Figure 6: Education and default staus 16
Figure 7: Marriage and default status 17
Figure 8: Age and default status 18
Figure 9: Correlation matrix 18
Figure 10: Performance evaluation with imbalance data 200
Figure 11: Comparison of imbalanced and balanced data 21
Figure 12: Performance evaluation with balanced dataset 22

1. Introduction

Making the appropriate decisions is crucial when deciding who should be approved and who shouldn't be when offering credit cards or loans to customers. To assist bank managers in deciding whether to approve loans for potential consumers, a data analytics-based strategy must be created. Classification has been used as a primary theme for the analysis. A list of the issues (research questions) this project has attempted to solve is as follows: Will granting a loan to a borrower with bad credit risk (a defaulter) cost the bank money? How can this risk be reduced to ensure the bank makes the most money possible? What characteristics of the dataset can be used to predict defaulters as the main determinant? This project used 'Default of Credit Card Clients' Dataset having 25 attributes in which data was collected for 30,000 Taiwanese clients during a span of six months (April to September 2005). The data contains a class attribute (which is qualitative) that displays the overall consumers' good or bad credit risk (default payment). In addition to this, there are ten qualitative attributes (Client ID, Sex, Education, Marriage status, six payment status for six months) and 14 numeric attributes (Credit Limit, Age, Amount of bill statement for six months, Amount of previous payment for six months). This dataset was taken from Google Dataset Search engine [1].

First, data preparation and preprocessing were done in order to comprehend the data and lay the groundwork for answering the study questions. Finding missing values, analyzing if numerical attributes have any bearing on the class attribute, determining which attribute is most closely related to the class attribute, etc. are some examples of stages that were included. Second, the primary analysis was conducted using the chosen theme (Classification). This entailed actions like segmenting the dataset and selecting attributes with the aid of 5 classification models based on both unbalanced and balanced datasets, in order to better comprehend algorithm results, establishing performance metrics, contrasting classification schemes, etc. The analysis' key conclusions and suggested solutions were included in the conclusion and recommendations section, which contain steps that the banks can take to address their issues.

2. Literature Review

The financial problem of creditworthiness has been extensively studied, both in terms of issuing credit cards and providing various types of credit. Literature shows that cardholders, especially those with high balances, are still very sensitive to interest rates.

As a result, they almost always seek lower credit card interest rates to reduce cash costs. Research shows that relatively few consumers choose their banking services as an alternative, despite their dissatisfaction with their banks. Additionally, various methodologies from the areas of statistical and decision-making technology have been successfully implemented.

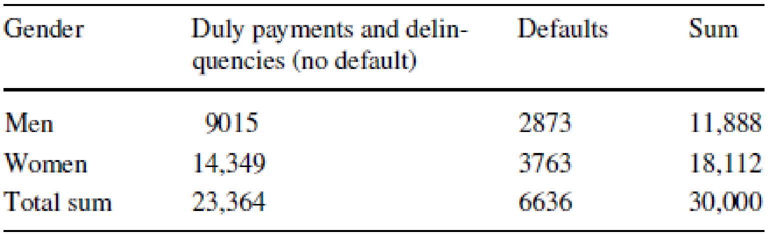
This section first looks at a few papers that have used the same dataset that this project has taken. Then there will be some insights about a few more papers that have done similar work using different datasets.

One paper has used accounting, demographic, and credit criteria to look for credit card defaults and obtain credit positions from customers. As the credit card industry has grown significantly over the past few years, analysts and financial institutions need to predict the creditworthiness of their customers [2].

The prediction accuracy of 7 machine learning classifiers has been used, which include - H. ANN, Logistic Regression, Naive Bayes, Decision Trees, Random Forest, SVC, and Linear SVC. Classification methods allow you to predict which category your financial situation belongs to using key characteristics. Loan default prediction is a good example of the application of machine learning techniques.

This study finds that the most important predictor for all seven methods used is PAY_0, representing cardholder repayment performance in September 2005.

Table 1: Gender-wise results

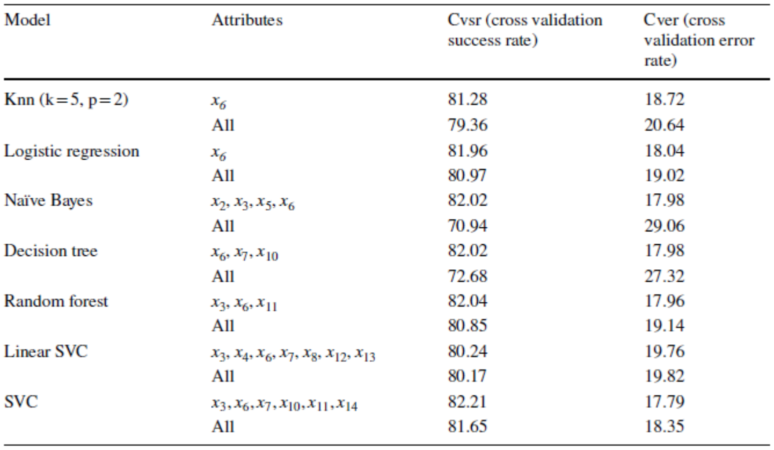


Gender	Duly payments and delinquencies (no default)	Defaults	Sum
Men	9015	2873	11,888
Women	14,349	3763	18,112
Total sum	23,364	6636	30,000

Analyzing the samples in Table 1, it can be seen that the relative proportion of the highest creditworthy borrowers has graduate degrees. This indicates that skill development through education is generally interpreted as indicating that borrowers will have more influence and will need the credit facilities they seek to meet their needs. Those with graduate degrees in education appear to be more conservative about lending than those with college degrees. On the other hand, among the loan portfolios studied, high school seniors had proportionally the highest default rates, suggesting that this category of borrowers may require better credit evaluation monitoring. [2]

The accuracy of the resulting models ranges from approximately 70% to 82%. Their accuracy can therefore be considered satisfactory and can therefore be used by financial institutions or credit card companies to classify potential customers with less information according to their solvency conditions during the approval process. [2]

Table 2: Comparative results

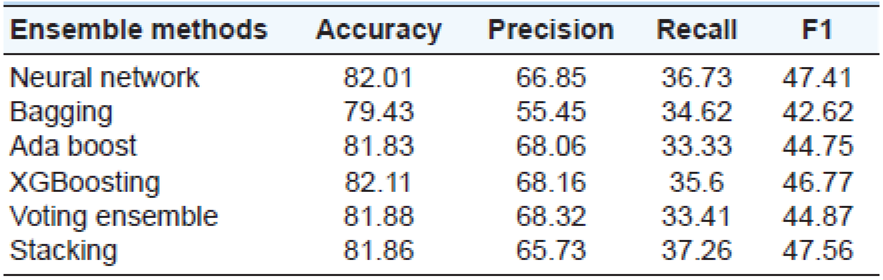


Model	Attributes	Cvsr (cross validation success rate)	Cver (cross validation error rate)
Knn (k=5, p=2)	x_6	81.28	18.72
	All	79.36	20.64
Logistic regression	x_6	81.96	18.04
	All	80.97	19.02
Naïve Bayes	x_2, x_3, x_5, x_6	82.02	17.98
	All	70.94	29.06
Decision tree	x_6, x_7, x_{10}	82.02	17.98
	All	72.68	27.32
Random forest	x_3, x_6, x_{11}	82.04	17.96
	All	80.85	19.14
Linear SVC	$x_3, x_4, x_6, x_7, x_8, x_{12}, x_{13}$	80.24	19.76
	All	80.17	19.82
SVC	$x_3, x_6, x_7, x_{10}, x_{11}, x_{14}$	82.21	17.79
	All	81.65	18.35

Another article has applied various ensemble techniques and compared the performances in order to detect credit card defaulters using the same dataset from Taiwan. The ensemble method aims to improve the predictability of the model by combining multiple models to create a stable model. By training multiple models to train a meta-estimator, ensemble learning aims to improve prediction efficiency. [3]

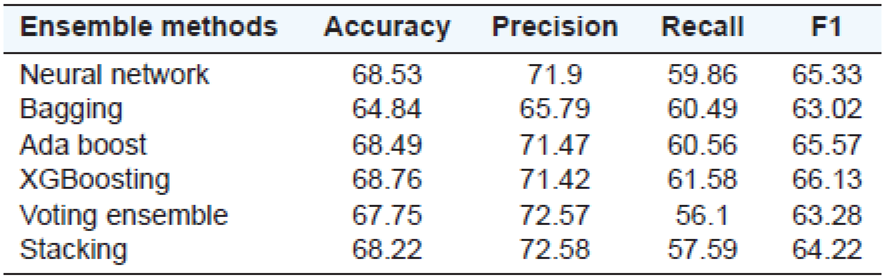
As with all types of risk assessment datasets, the ratio of positive to negative samples creates large imbalances in the dataset. Only 22% of customers failed in this dataset (defaulters). There are no missing values in the dataset. Ensemble learning results were recorded in two separate experiments. The first time with the original imbalanced data set and the second time after the imbalanced aspects were eliminated. [3]

Table 3: Results of imbalanced dataset



Ensemble methods	Accuracy	Precision	Recall	F1
Neural network	82.01	66.85	36.73	47.41
Bagging	79.43	55.45	34.62	42.62
Ada boost	81.83	68.06	33.33	44.75
XGBoosting	82.11	68.16	35.6	46.77
Voting ensemble	81.88	68.32	33.41	44.87
Stacking	81.86	65.73	37.26	47.56

Table 4: Results of balanced dataset



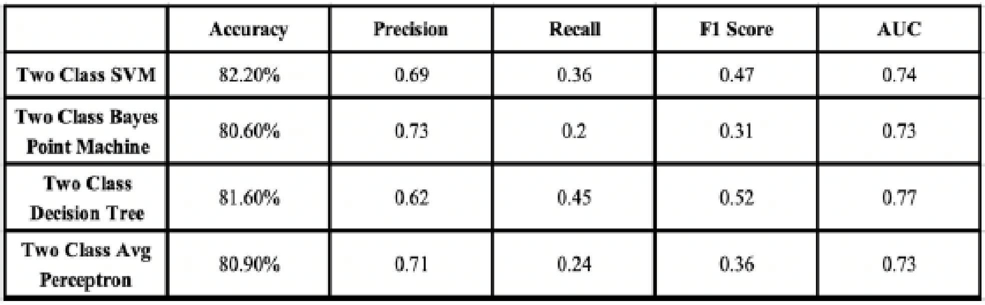
Ensemble methods	Accuracy	Precision	Recall	F1
Neural network	68.53	71.9	59.86	65.33
Bagging	64.84	65.79	60.49	63.02
Ada boost	68.49	71.47	60.56	65.57
XGBoosting	68.76	71.42	61.58	66.13
Voting ensemble	67.75	72.57	56.1	63.28
Stacking	68.22	72.58	57.59	64.22

While considering the imbalanced dataset, better results were given by Stacking & XGBoosting. And while considering the balanced dataset, XGBoosting performed slightly better than all other methods. The concept of boosting methods is to improve the accuracy of poor classification methods by incorporating multiple instances into a more reliable estimate. [3]

A. Shivanna and D. P. Agrawal (used the same Taiwanese dataset) have emphasized the fact that data mining and machine learning techniques are being widely used by various financial institutions in order to predict credit defaulters. In this work, different algorithms such as Deep Support Vector Machine (DSVM), Boosted Decision Tree (BDT), Average Perceptron (AP), and Bayes Point Machine (BPM) are used to build different models and analyze the payouts and better-predicted defaults. [4]

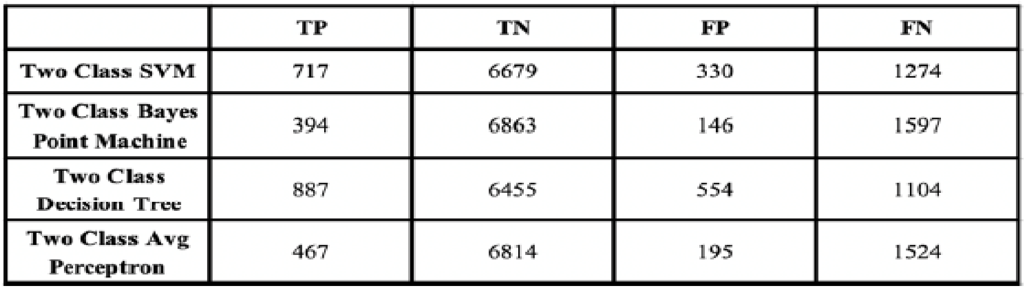
Some of the new observations made in this paper are – Female customers are in majority, older customers are less likely to be defaulters, and 'payment' and 'bill amount' features are positively correlated. [4]

Table 5: Confusion Matrix



	Accuracy	Precision	Recall	F1 Score	AUC
Two Class SVM	82.20%	0.69	0.36	0.47	0.74
Two Class Bayes Point Machine	80.60%	0.73	0.2	0.31	0.73
Two Class Decision Tree	81.60%	0.62	0.45	0.52	0.77
Two Class Avg Perceptron	80.90%	0.71	0.24	0.36	0.73

Table 6: Classification Results

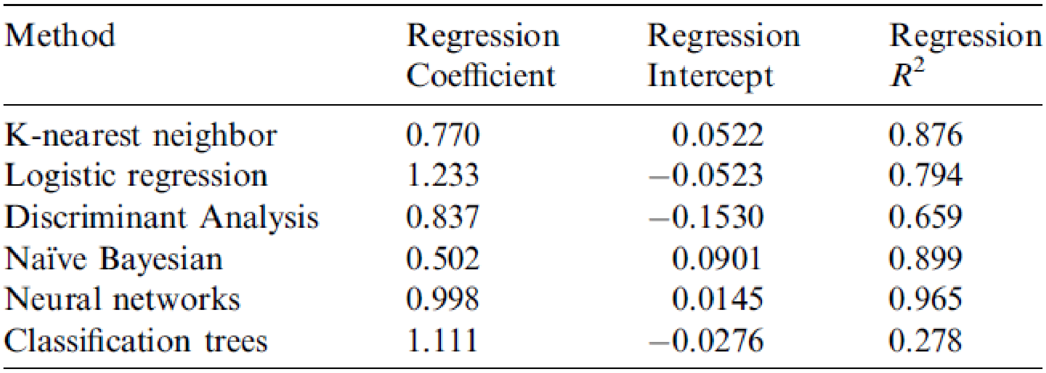


	TP	TN	FP	FN
Two Class SVM	717	6679	330	1274
Two Class Bayes Point Machine	394	6863	146	1597
Two Class Decision Tree	887	6455	554	1104
Two Class Avg Perceptron	467	6814	195	1524

The paper accurately predicted credit card defaults by training ML models with BDT, BPM, DSVM, and AP algorithms on the Azure Machine Learning platform. Accuracy and other metrics are evident from the table above. The model with DSVM performs better than all other models. The overall accuracy is 82.2% and the AUC is 0.74. [4]

One paper suggests that error rate has often been used as a measure of a model's classification accuracy. However, most records in the credit card customer dataset (Taiwanese) are non-defaulters (87.88%). Therefore, the error rate is not affected by the model's classification accuracy. This study examined the classification accuracy of six data mining methods using area ratio instead of error rate. [5]

Table 7: Summary of linear regression between real probability and predictive probability of default



Method	Regression Coefficient	Regression Intercept	Regression R^2
K-nearest neighbor	0.770	0.0522	0.876
Logistic regression	1.233	−0.0523	0.794
Discriminant Analysis	0.837	−0.1530	0.659
Naïve Bayesian	0.502	0.0901	0.899
Neural networks	0.998	0.0145	0.965
Classification trees	1.111	−0.0276	0.278

Regarding the classification accuracy among the six data mining techniques, the results show that there is little difference in the error rate among the six methods. However, there is a relatively large difference in the area ratios of the six techniques. The area ratio is more sensitive and a good criterion for measuring the classification accuracy of the model. Artificial neural networks (ANN) perform classification more accurately than the other five methods. This paper has concluded that the real default probability is represented most accurately by ANN, and suggested that it should be preferred over other techniques of data mining to score clients. [5]

U. K. Panchal and S. Verma have taken the help of Self Organizing Maps (SOM) to predict future defaulters from a list of current non-defaulters. SOM is an unsupervised artificial neural network that generates a low-dimensional discrete representation of the input space of data samples, called a map, and retains the original topology. Self-organizing maps are especially useful when the relationships between different attributes in the data set are non-linear. A slightly similar dataset (to the Taiwanese dataset) was used in this paper (source of dataset not provided). [6]

In figure 1, defaults are represented by green squares, and non-defaulters are represented by red circles. A count of defaulters and non-defaulters is calculated for every 625 nodes (25x25). The brightness of each square indicates the normalized sum of the distances between the neuron and its neighbors. Clusters containing more than a certain percentage of the default majority are now selected. This percentage of the selected majority is called the limit percentage value (LPV). The number of non-defaults (identified as potential future defaults) in these groups, compared to this threshold's percentage of the majority of defaults in this group, is shown in table 8. The higher the percentage of non-payers, the more data points in these clusters that belonged to the non-payer class are identified as potential future non-payers. [6]

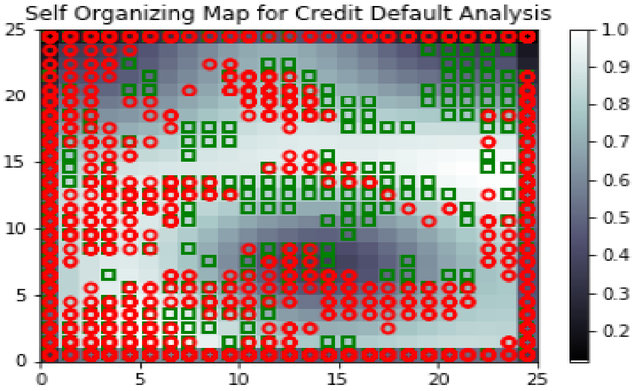
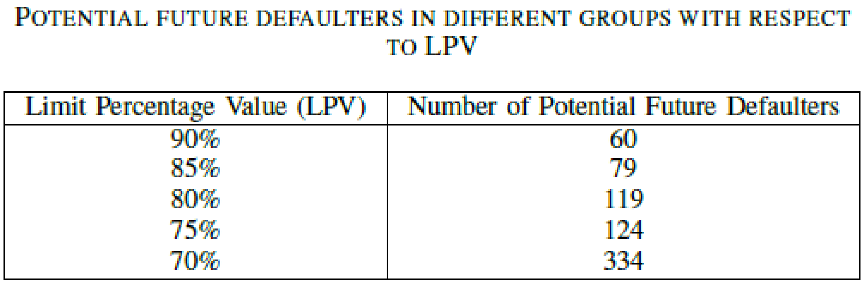


Figure 1: Resulting Self-Organizing Map (With Markers)

Table 8: Potential feature defaulters

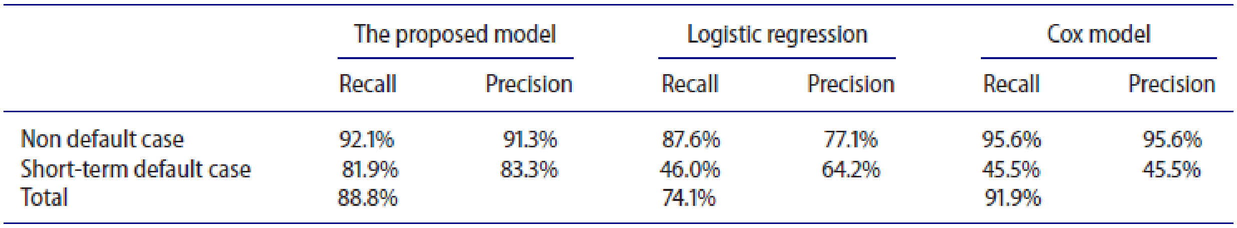


POTENTIAL FUTURE DEFAULTERS IN DIFFERENT GROUPS WITH RESPECT TO LPV

Limit Percentage Value (LPV)	Number of Potential Future Defaulters
90%	60
85%	79
80%	119
75%	124
70%	334

Loan managers predict when loans will default in the short term, as short-term defaults are costly for financial institutions, as explained in another research article from China. A loan manager is more interested in identifying potential default applications that may default in a short time period. This paper proposes a decision tree-based short-term credit risk assessment model for assessing credit risk. The goal is to create a highly accurate model that can use decision trees to filter short-term defaults and distinguish between loans at default. This paper has integrated bootstrap aggregation (bagging) using the synthetic minority oversampling (SMOTE) technique into credit risk models to improve decision tree stability and performance on imbalanced data. [7]

Table 9: Discriminant recall rate of risk assessment model



	The proposed model		Logistic regression		Cox model	
	Recall	Precision	Recall	Precision	Recall	Precision
Non default case	92.1%	91.3%	87.6%	77.1%	95.6%	95.6%
Short-term default case	81.9%	83.3%	46.0%	64.2%	45.5%	45.5%
Total	88.8%		74.1%		91.9%	

This paper used a dataset provided by another Taiwanese bank that gave short-term credit loans to SMEs in the period from Nov 2001 to Dec 2010. The proposed model can filter short-term default events. This helps financial institutions estimate potential financial losses and adjust lending policies. Compared to traditional risk assessment models that estimate whether a case will default, the proposed model focuses on avoiding losses by identifying cases that are likely to default in the short term. Hence, financial institutions will continue to be profitable even with bad debt. On the other hand, compared to other credit risk assessment models that estimate the timing of default using the survival analysis approach, the proposed method succeeds in identifying short-term default cases that require a higher level of response through binary classification. Experimental results show that the recall of the proposed model is better than those of logistic regression and Cox model. [7]

3. Data

3.1. Data description

The dataset extracted for this study uses information on Taiwanese credit card portfolio customers from April 2005 to September 2005. This includes the customer's default payments in the next month's credit card repayments, as well as the cardholder's billing statements, plus their accounting, demographic, and credit factor data.

This dataset was taken from Kaggle, a Google dataset search engine. It is part of the UCI Machine Learning Repository which is available from Irvine University College [8]. The website provides information about data such as the number of columns and rows, year created, type of data, and jobs they are suitable for. Out of the 25 attributes, a binary variable is taken, the default payment (Yes=1, No=0) which is used as a response variable. The dataset has a total of 30,000 entries taken from clients in Taiwan. The description of the attributes is as follows -

ID - Client’s Identification number.

LIMIT_BAL - Amount of given credit, NT dollars (individual & family/supplementary credit included). Here NT means New Taiwan Dollars, which is the official currency.

SEX - Gender (1=male, 2=female) which is binary variable

EDUCATION - (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown). The level of education is an important aspect for analyzing default payments.

MARRIAGE - Marital status (1=married, 2=single, 3=others)

AGE - Age in years. The values for age are taken as full integers.

PAY_0 - Repayment status in September, 2005 (-1=pay duly, 1=payment delay for one month, 2=payment delay for two months, …up until 8=payment delay for eight months, 9=payment delay for nine months and above). It is clear that the longer the clients wait to pay credit card or revert to their current credit balance, the more likely they're to default next month.

PAY_2 - Repayment status in August, 2005 (scale same as above)

PAY_3: Repayment status in July, 2005 (scale same as above)

PAY_4: Repayment status in June, 2005 (scale same as above)

PAY_5: Repayment status in May, 2005 (scale same as above)

PAY_6: Repayment status in April, 2005 (scale same as above)

BILL_AMT1: Amount of bill statement in September, 2005 (NT dollar). The amounts of bill payments are taken for the same time period.

BILL_AMT2: Amount of bill statement in August, 2005 (NT dollar)

BILL_AMT3: Amount of bill statement in July, 2005 (NT dollar)

BILL_AMT4: Amount of bill statement in June, 2005 (NT dollar)

BILL_AMT5: Amount of bill statement in May, 2005 (NT dollar)

BILL_AMT6: Amount of bill statement in April, 2005 (NT dollar)

PAY_AMT1: Amount of previous payment in September, 2005 (NT dollar)

PAY_AMT2: Amount of previous payment in August, 2005 (NT dollar)

PAY_AMT3: Amount of previous payment in July, 2005 (NT dollar)

PAY_AMT4: Amount of previous payment in June, 2005 (NT dollar)

PAY_AMT5: Amount of previous payment in May, 2005 (NT dollar)

PAY_AMT6: Amount of previous payment in April, 2005 (NT dollar)

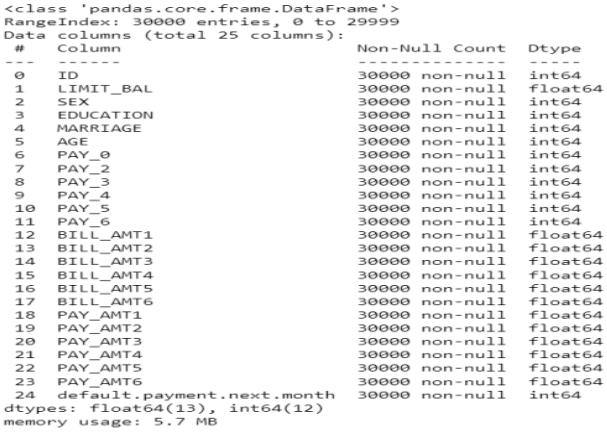
default.payment.next.month: Default payment (1=yes, 0=no). This is the only class attribute in the dataset.

3.2. Data Preprocessing and data analysis

Data preprocessing is the process of checking data for null values or any other ambiguous or incorrectly entered data values. Data analysis is done to find out the deep insights of the dataset. It gives us a better understanding of the data.

3.2.1. Data Overview

The below figure is a quick overview of the features involved in this dataset.



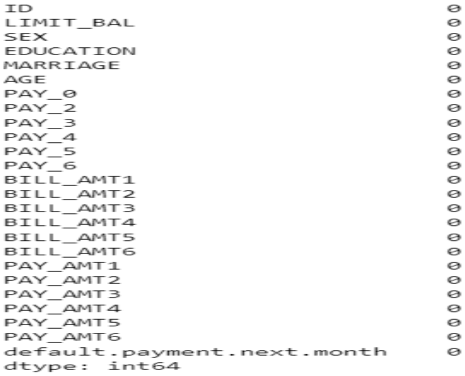
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30000 entries, 0 to 29999
Data columns (total 25 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   ID                          30000 non-null  int64
 1   LIMIT_BAL                   30000 non-null  float64
 2   SEX                         30000 non-null  int64
 3   EDUCATION                   30000 non-null  int64
 4   MARRIAGE                    30000 non-null  int64
 5   AGE                         30000 non-null  int64
 6   PAY_0                       30000 non-null  int64
 7   PAY_2                       30000 non-null  int64
 8   PAY_3                       30000 non-null  int64
 9   PAY_4                       30000 non-null  int64
 10  PAY_5                       30000 non-null  int64
 11  PAY_6                       30000 non-null  int64
 12  BILL_AMT1                   30000 non-null  float64
 13  BILL_AMT2                   30000 non-null  float64
 14  BILL_AMT3                   30000 non-null  float64
 15  BILL_AMT4                   30000 non-null  float64
 16  BILL_AMT5                   30000 non-null  float64
 17  BILL_AMT6                   30000 non-null  float64
 18  PAY_AMT1                    30000 non-null  float64
 19  PAY_AMT2                    30000 non-null  float64
 20  PAY_AMT3                    30000 non-null  float64
 21  PAY_AMT4                    30000 non-null  float64
 22  PAY_AMT5                    30000 non-null  float64
 23  PAY_AMT6                    30000 non-null  float64
 24  default.payment.next.month  30000 non-null  int64
dtypes: float64(13), int64(12)
memory usage: 5.7 MB
```

Figure 2: features overview

There are a total of 24 independent features with one dependent or target feature. The total number of records is 30000. All are integer or float-type features.

3.2.2. Checking for null values

The empty cells in a column that have been left blank because of improper data handling or entering are referred to as having null values.



```
ID                            0
LIMIT_BAL                     0
SEX                           0
EDUCATION                     0
MARRIAGE                      0
AGE                           0
PAY_0                         0
PAY_2                         0
PAY_3                         0
PAY_4                         0
PAY_5                         0
PAY_6                         0
BILL_AMT1                     0
BILL_AMT2                     0
BILL_AMT3                     0
BILL_AMT4                     0
BILL_AMT5                     0
BILL_AMT6                     0
PAY_AMT1                      0
PAY_AMT2                      0
PAY_AMT3                      0
PAY_AMT4                      0
PAY_AMT5                      0
PAY_AMT6                      0
default.payment.next.month    0
dtype: int64
```

Figure 3: Null values count

We made use of the isnull() function, which iterates through all of the cells and returns the cells that are empty. The sum() function will then take the count of all of those cells and provide a single count value that represents the cells that are missing. However, in our case, there is no missing value in our dataset as shown in the above figure.

3.2.3. Changing feature and label names

For better understanding, we changed the name of the ‘default.payment.next.month’ to ‘default’, which contains two distinct labels on whether a particular customer will default for payment or not. We also changed labels in Sex, Education, and marriage from numerical labels to categorical ones. It will make our analysis understandable and unambiguous.

3.2.4. Distribution of target feature

Default is the target feature of our dataset. It contains two distinct labels whether a particular customer will default or not next month.

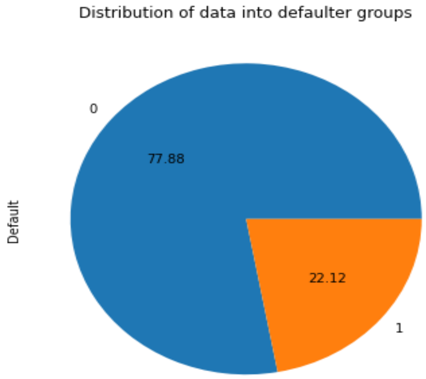


Figure 4: Data distribution of defaulter group

In the above figure, the distribution of defaulter and non-defaulter groups are displayed in a pie chart. Most of the customers fall under the non-defaulter group. Only 22% of customers are prone to default.

3.2.5. Gender wise default analysis

The credit dataset contains information about both males and females. There are 18112 female customers in this data and 11888 males.

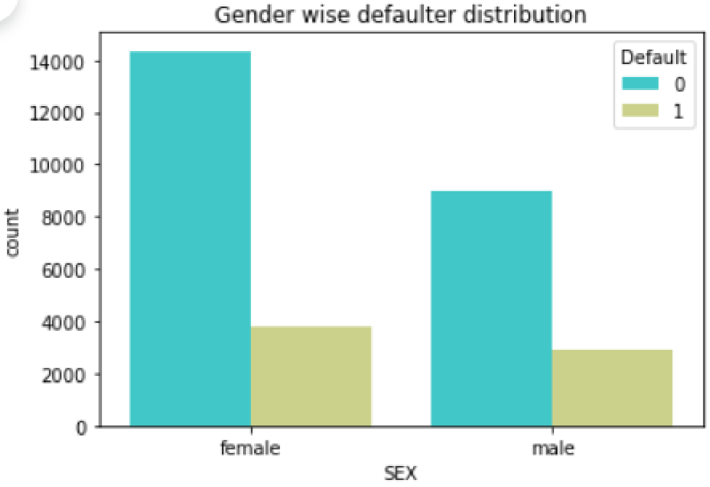


Figure 5: Distribution of target with gender

The default ratio for female is more than male as shown in above figure. It is also due to the larger number of females than males in this credit dataset.

3.2.6. Education type and default ration

Education has also an important role in the defaulter position of the customers. There are 3 groups of education - university, graduate, and high school and one group is unknown and put under the other.

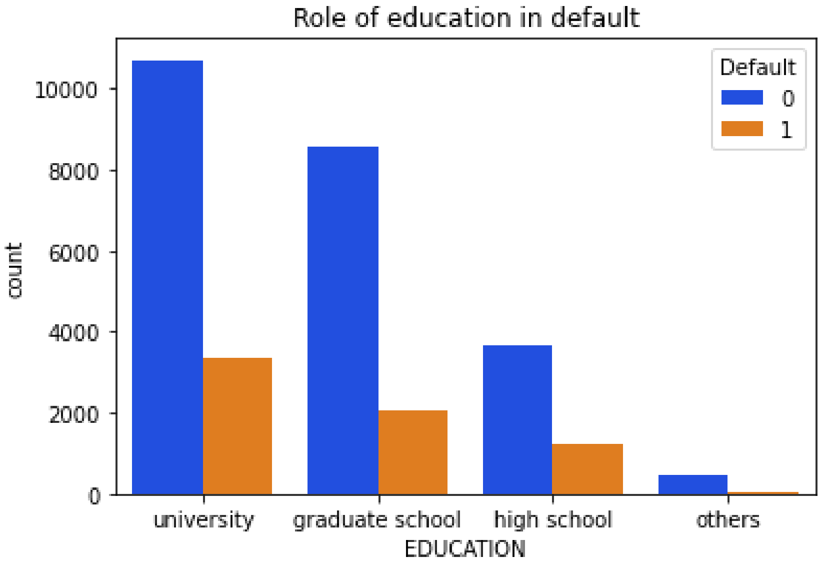


Figure 6: Education and default status

As compared to graduate and high school, customers with university education have more chances to default next month. The reason may also be the higher number of customers with a university education than other educational groups.

3.2.7. Marital Status and default position

The customers included in this credit dataset are with different marital statuses. There are 15964 single,13659 married and 377 with unidentified marital status.

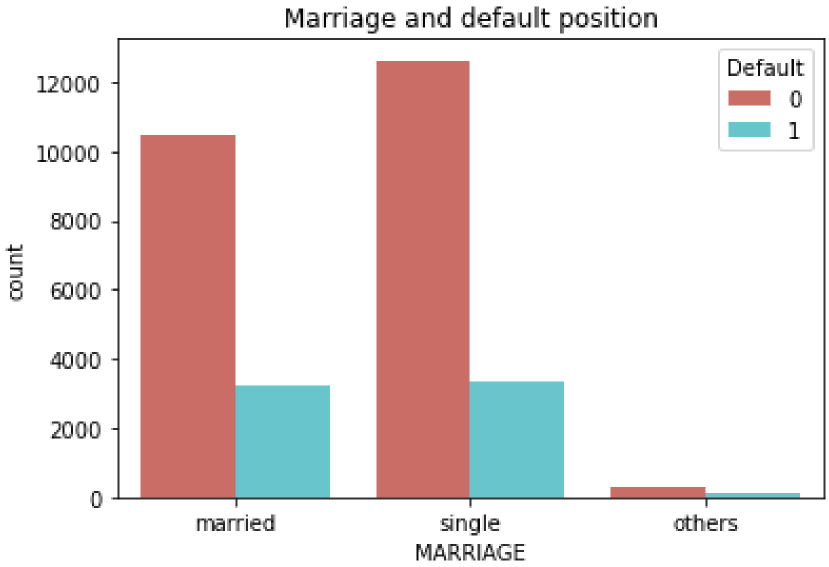


Figure 7: Marriage and default status

There is not much difference in the number of defaulters with single and married marital status. However, singles have a greater number than married in the non-defaulter group.

3.2.8. Age distribution and default position

There are different age groups from the adult group in this dataset. The customers on the credit list range from 21 years to 79 years.

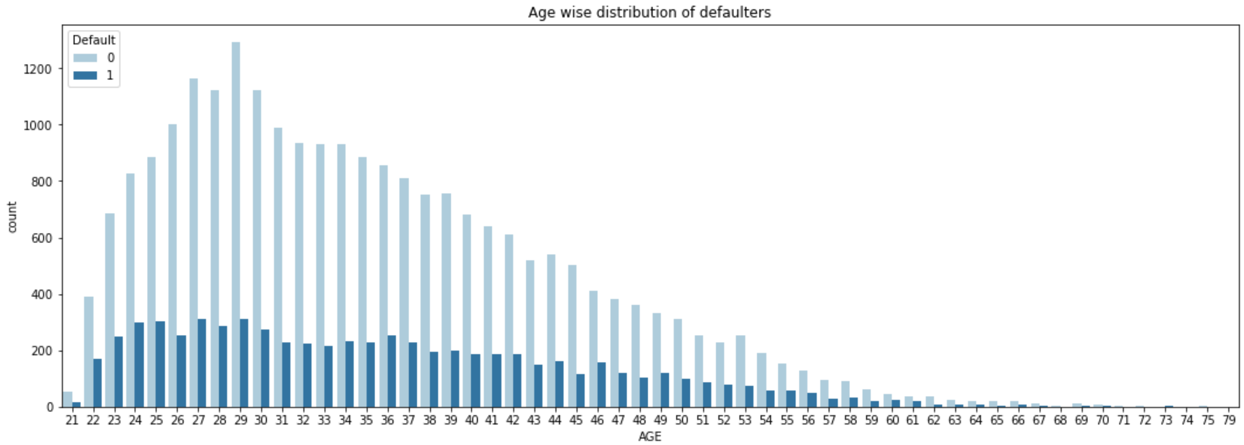


Figure 8: Age and default status

From the above figure age group from 22 to 46 have a significant number of people with default position. Whereas age group from 57 to onward have the least existence in both default and non-default category.

3.2.9. Correlation

Correlation shows the close relationship between two features in a dataset. A correlation number close to 1 between the two features shows their strong relationship. The below figure is the correlation graph of the features included in this dataset.

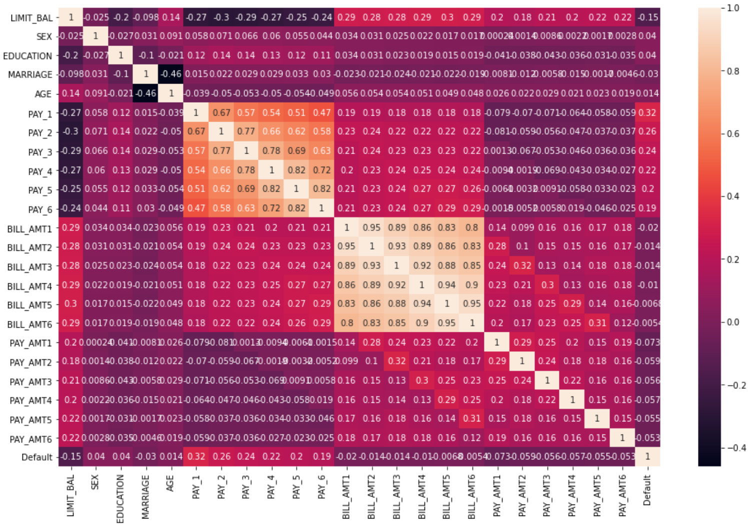


Figure 9: Correlation matrix

The correlation numbers in the lighter colors show a strong correlation between the two features. For example, bill amount 1 and bill amount 2 are strongly correlated with a higher correlation number of (0.95). The scale on the right side is the scale according to correlation numbers.

4. Data Preparation

Data preparation is the step to prepare the dataset for training of the machine learning model. This usually involves the conversion of categorical data features into numerical ones because machine learning algorithms cannot work with categorical data types. The label encoder from sklearn was used to change the categorical data into numerical values. After preprocessing, the dataset was divided into training and testing sets with 80 and 20% ratios.

5. Model Implementation

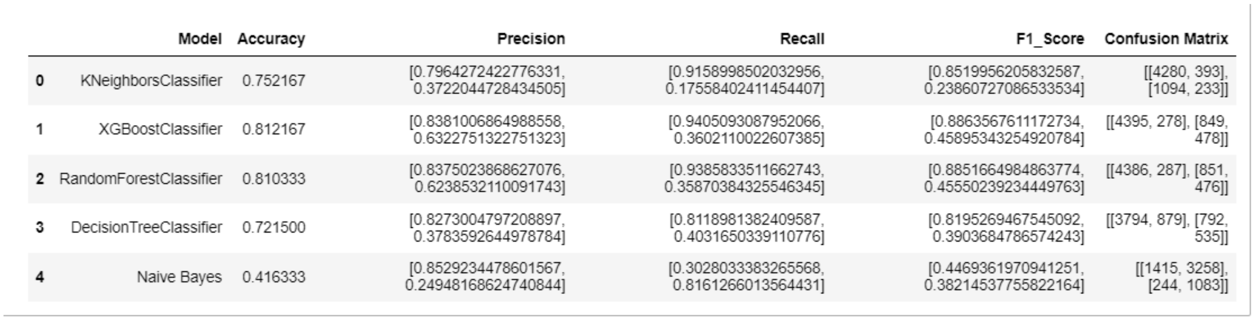
Machine learning models are trained to predict whether a customer is going to default in the next month or not. We will implement each model two times for imbalanced and balanced data. The machine learning models used in this project are - Decision tree classifier, random forest classifier, XGBoost classifier, kneighbors classifier and naïve bayes classifier. First, the machine learning model is trained on a training set. Then, the trained model is tested and evaluated using an evaluation matrix.

5.1. Balanced and imbalanced dataset

A dataset is balanced when the proportion of records for each class is the same or when the difference in proportion is negligible. A balanced dataset would mean 50% records for each class in the case of 2 classes. Proportion is negligible in the case of 45% and 55% for 2 classes. A dataset is considered imbalanced when the proportions of records for the classes are significantly different. In the case of 2 classes, the dataset is imbalanced when 80% of the records are for class A and 20% for class B. A machine learning model performs best on balanced data. In our case, the dataset is imbalanced as the almost 78% data is for class 0 and 22% for class 1 for the target class.

5.2. Results with imbalanced dataset

Decision Tree Classifier, naïve bayes, Random Forest, KNearest Neighbors, and XGBoost Classifier are fitted on the training subset of imbalanced data. Once fitted, the machine learning models make predictions on the held-out test set, using the predicted values for the test dataset, each machine learning model is evaluated. Below are the results from these models:



	Model	Accuracy	Precision	Recall	F1_Score	Confusion Matrix
0	KNeighborsClassifier	0.752167	[0.7964272422776331, 0.3722044728434505]	[0.9158998502032956, 0.17558402411454407]	[0.8519956205832587, 0.23860727086533534]	[[4280, 393], [1094, 233]]
1	XGBoostClassifier	0.812167	[0.8381006864988558, 0.6322751322751323]	[0.9405093087952066, 0.3602110022607385]	[0.8863567611172734, 0.45895343254920784]	[[4395, 278], [849, 478]]
2	RandomForestClassifier	0.810333	[0.8375023868627076, 0.6238532110091743]	[0.9385833511662743, 0.35870384325546345]	[0.8851664984863774, 0.45550239234449763]	[[4386, 287], [851, 476]]
3	DecisionTreeClassifier	0.721500	[0.8273004797208897, 0.3783592644978784]	[0.8118981382409587, 0.4031650339110776]	[0.8195269467545092, 0.3903684786574243]	[[3794, 879], [792, 535]]
4	Naive Bayes	0.416333	[0.8529234478601567, 0.24948168624740844]	[0.3028033383265568, 0.8161266013564431]	[0.4469361970941251, 0.38214537755822164]	[[1415, 3258], [244, 1083]]

Figure 10: Performance evaluation with imbalanced data

This table figure shows the performance results from all models with imbalanced data. The accuracy is very average for all models. Only random forest and xgboost are giving 81% accuracy score which is just satisfactory, not a very good score. Moreover, the recall has a very high difference for both classes such as for kneighbours, it is giving 92% for 0 labels and 18% for 1. This performance matrix needs to improve and that can be done by balancing the imbalanced data.

5.3. Model implementation on balanced dataset

Because the dataset was originally imbalanced, and this may have an effect on the performance of some machine learning algorithms, the random oversampling method was used to make the dataset balanced in order to find out how using a balanced dataset would affect the performance of the algorithms.

5.3.1. Random oversampling

A dataset can be balanced using a technique called random oversampling, which involves randomly duplication of data from a specific(minority) class. This process continues until both classes monitory and majority have same number of records. The application of Random Oversampling results in a balanced dataset that has a 55.76 percentage increase in data points as compared to the initial data. Below figure depicts the distribution of values before and after balancing dataset.

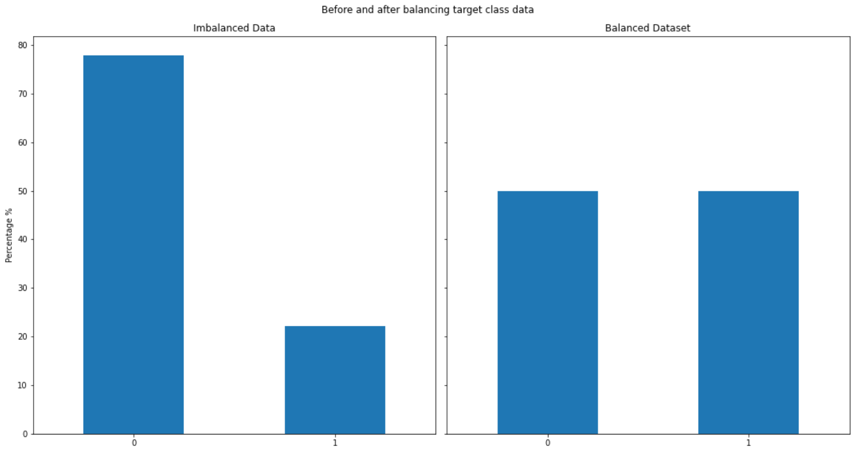
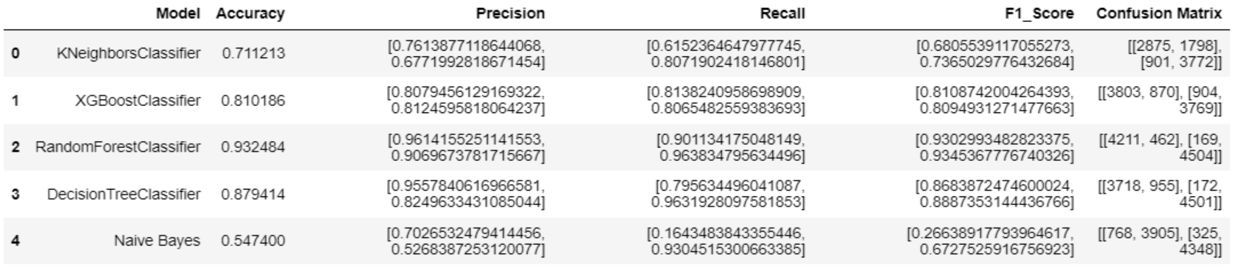


Figure 11: Comparison of imbalanced and balanced data

5.3.2. Results with balanced dataset

After balancing the data, the machine learning models were again trained on the training set and tested with the test set. Below figure shows the new results after balancing the dataset.



	Model	Accuracy	Precision	Recall	F1_Score	Confusion Matrix
0	KNeighborsClassifier	0.711213	[0.7613877118644068, 0.6771992818671454]	[0.6152364647977745, 0.8071902418146801]	[0.6805539117055273, 0.7365029776432684]	[[2875, 1798], [901, 3772]]
1	XGBoostClassifier	0.810186	[0.8079456129169322, 0.8124595818064237]	[0.8138240958698909, 0.8065482559383693]	[0.8108742004264393, 0.8094931271477663]	[[3803, 870], [904, 3769]]
2	RandomForestClassifier	0.932484	[0.9614155251141553, 0.9069673781715667]	[0.901134175048149, 0.963834795634496]	[0.9302993482823375, 0.9345367776740326]	[[4211, 462], [169, 4504]]
3	DecisionTreeClassifier	0.879414	[0.9557840616966581, 0.8249633431085044]	[0.795634496041087, 0.9631928097581853]	[0.8683872474600024, 0.8887353144436766]	[[3718, 955], [172, 4501]]
4	Naive Bayes	0.547400	[0.7026532479414456, 0.5268387253120077]	[0.1643483843355446, 0.9304515300663385]	[0.26638917793964617, 0.6727525916756923]	[[768, 3905], [325, 4348]]

Figure 12: Performance evaluation with balanced dataset

The results show a significant improvement in the models' performance. Although not all models still give the best accuracy, but we gained the maximum accuracy score with random forest. It is giving 93% accuracy score. Decision tree has also improved the accuracy with the balanced dataset from 72 percent to 87% which is a remarkable change. Recall scores have also been improved, for the random forest it is giving 90 and 96%. And F1-score for the random forest on test data is also above 90% which makes it best the model among others. So, in conclusion, we can say that random forest is the best model for the prediction of the default position for this credit data.

6. Conclusion

In this Project, five different machine learning algorithms were trained for the prediction of the customers to default or not in the next month. After training on the imbalanced data, the models did not produce higher accuracy rates as expected from imbalanced data. Therefore, the data was balanced using over-sampling which was retrained and re-evaluated. This time with the balanced dataset, models produced significant performance in their resulting scores. The **random forest** was the best performing with the highest accuracy of **93%** (higher than other models used in previous research papers). The outcomes of this experiment can be utilized as a benchmark to assist the bank administration to take effective measures to control the default groups through efficient policy measures. This will help them make more informed decisions.

7. REFERENCES

[1] Google Dataset Search. [online] Available: https://www.kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset

[2] Default avoidance on credit card portfolios using accounting, demographical and exploratory factors: decision making based on machine learning (ML) techniques [online] Available: https://link-springer-com.ezproxy.lib.ryerson.ca/article/10.1007/s10479-019-03188-0

[3] Comparison of Different Ensemble Methods in Credit Card Default Prediction [online] Available: https://journals.uhd.edu.iq/index.php/uhdjst/article/view/806/640

[4] Prediction of Defaulters using Machine Learning on Azure ML [online] Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9284884

[5] The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients [online] Available: https://www-sciencedirect-com.ezproxy.lib.ryerson.ca/science/article/pii/S0957417407006719

[6] Identification of Potential Future Credit Card Defaulters from Non Defaulters using Self Organizing Maps [online] Available: https://ieeexplore.ieee.org/document/8944605

[7] Establishing decision tree-based short-term default credit risk assessment models [online] Available: https://www-tandfonline-com.ezproxy.lib.ryerson.ca/doi/pdf/10.1080/03610926.2014.968730?needAccess=true

[8] UCI Machine Learning; default of credit card clients Data Set [online] Available: https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients