FRAUD DETECTION IN FINANCIAL SYSTEMS USING STATISTICAL LEARNING

Project report submitted in partial fulfillment of the requirement for the degree of

Bachelor of Technology

By
Ojas Srivastava
(1710110238)

And

Yatharth Jain
(1710110400)

Under supervision

Of

Prof. Madan Gopal

DEPARTMENT OF ELECTRICAL ENGINEERING

SCHOOL OF ENGINEERING

SHIV NADAR UNIVERSITY

(November 2020)

Abstract

This project will discover methods to detect fraud in Financial Systems. Our aim is to create a robust system that would detect patterns of default and predict any potential default which could be encountered by the system. We desire to make a model that would not only give true negatives but would also produce fewer false positives (which would be one of our parameters for model performance). This project would fall under the broad category of Anomaly Detection Methods under the Data Science field.

The idea is to use the dataset UCI_Credit_Card [1] for this purpose. This dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005. Our proposed model will be trained and tested on this dataset.

Table of Contents

List of Figures

List of Tables

Chapter-1
Introduction

Problem Description and Aim

1. This project will discover methods to detect credit default in Financial Systems. Our aim is to create a robust system that would detect patterns of credit card payments by card holders and predict any potential defaulter which could be encountered by the system.

2. We desire to make a model that would not only give true negatives but would also produce fewer false positives (which would be one of our parameters for model performance). This project would fall under the broad category of Anomaly Detection Methods under the Data Science field.

Background

1. **Credit Cards (Debt Instrument):** Credit Cards are a physical card which are issued by various financial institutions allowing the holder to purchase goods or services on credit. Credit Cards are popularly combined with benefits like loyalty points, discounts etc. These cards are a great innovation in financial technology offering its holders many facilities **[2]** like easy overdraft & credit, security features, staggered repayment etc.
    a. Credit cards are one of the many debt-based products (also called Debt Instruments) offered by financial institutions. Debt instruments **[3]** often allow a fixed income to the financial institution and can be utilized to raise capital. Financial Institutions, thus, push towards wider adoption of such instruments like the credit card to acquire greater capital. It has been observed that institutions, in greed of capital, have lent out credit to non-deserving customers or even fraudsters.

2. **Credit Risk Aversion:** Whenever a financial institution lends out to a customer, it bears an inherent risk of not being paid back. In the case of credit cards, this risk is manifested by default in payment by the holder for the credit borrowed for transactions over a period of time.

Any institution cannot afford to have a huge risk on its Debt Instruments as it will threaten to burden them with huge credit losses.

a. The problem is so grave that risky credits by certain influential financial firms eyeing for big money ended up defaulting and thus led to a big pile-up of debt in the financial system which eventually resulted in the great stock market crash in 2008 [4] .

3. **Best Practices for Credit Risk Management:** According to SAS, following techniques should be adopted for a successful credit risk management at an institution **[5]** :
    a. Better model management that spans the entire modelling life cycle.
    b. Real-time scoring and limits monitoring.
    c. Robust stress-testing capabilities.
    d. Data visualization capabilities and business intelligence tools that get important information into the hands of those who need it, when they need it.
    e. For our project's implementation, we will try and accomplish all the suggested methods by SAS.

4. **Taiwan Credit Crisis:** Beginning in 1990, the Taiwanese government allowed the formation of new banks. These new banks lent large sums of money to real estate companies with the goal of expanding their businesses and increasing profits. However, after a couple of years of expansion, the real estate market became saturated and profits from the sector stopped growing.
    a. The new banks turned to other new businesses – credit cards and cash cards. In expanding this area of business, banks lavished money on commercials encouraging people to apply for credit cards to consume, apparently without consequences. These banks lowered the requirements for credit card approvals to get more customers. In time, young people became target customers. Although young people tend not to have enough income, banks still issued credits cards to them. **[6]**

The Problem & Need

- **The Problem**
    - In Taiwan, in February 2006, debt from credit cards and cash cards reached *$268 billion USD*. More than half a million people were not able to repay their loans. They became "credit card slaves", a term coined in Taiwan to refer to people who could only pay the minimum balance on their credit card debt every month. This issue resulted in significant societal problems. Some debtors and their families committed suicide because of the debt, some became homeless due to repossession of their homes, and others could not afford to pay their children's tuition.
    - The suicide rate in Taiwan is the second highest in the world. The suicide rate increased *22.9%* compared to the rate in 2005, and the main reason is unemployment and credit card debt. [6]
- **The Need & Motivation**
    - We have taken up this problem because we feel that such a grave mistake could have been avoided with the help of analytics. Many people could have had been saved if better monitoring and due-diligence of customers had been taken up by the bank. By this project we also aim to enhance our knowledge in the field of *Business Analytics*.

Problem Formulation & Target of the Project

The default detection problem in credit cards can be formalized based on the data provided for paid amount, billed amount, and other demographic data like gender, age, education background, etc. The task is to predict the customer who could become a defaulter in coming month.

Ideally, any bank would like to detect as many defaulters as possible while not wrongly labelling any of its loyal & honest customers. However, experience tells us that all Machine Learning models will have a trade-off between detecting higher number of defaulter & higher number of honest customers.

The goal of the project, therefore, is to predict more true positives (customers who will become defaulters) and less false positives (honest customers wrongly labelled as defaulters).

Metrics to evaluate robustness

We will use Precision & Recall Metrics. Formula for both have been provided below:

$$recall = \frac{true\ positives}{true\ positives\ + false\ negatives} \qquad precision = \frac{true\ positives}{true\ positives + false\ positives}$$

A few points to keep in mind:

- According to the formula of the metric it is evident to see that we need to have higher recall for class - 0. Which means lesser number of honest customers being wrongly labelled
- At the same time we need high recall for class -1 which will imply higher detection of default and lower number of wrongly labelled dishonest customers
- For this purpose, the metrics used are recall for class '0' or majority class or non-defaulter class and recall for class '1' or minority class or defaulter class. Both of these values should give a high value for a better model.

Brief Overview of Work Done

First the dataset was searched for. After a deep search along with tuning of keywords we landed up with this dataset which contains information of real customers.

Once the dataset was finalized, we preprocessed the dataset visualizing the features and finding some discrepancies in the data which are discussed in detail in coming chapters.

After this the categorical features were binary encoded and then the dataset was split into training and test sets (75% for training and 25% for testing). After all of this different machine learning models along with different methods were implied to find the best model for this kind of problem. All of this is discussed in detail in coming chapters.

---------End of Chapter 1-------

Chapter 2

Literature Survey

After going through various research works done in this field, we found that the researchers have mainly used two approaches. One being based on probability while the other on classifiers. Hence, we would be using both of these approaches in our model. Given are a few research papers which we would be taking reference from.

1. **Prediction of default payment of credit card clients using Data Mining Technique**s [7]

    1.1. This research works on customers' default payments for the accurate prediction of the probability of default payment. This paper suggests Synthetic Minority Over-Sampling Technique (SMOTE) to deal with the imbalanced dataset. By utilizing SMOTE method, the predicting model produced by Random Forest has the best accuracy and performance with low error rate. Consequently, among the seven data mining algorithms, Random Forest is a good alternative to precisely predict the default payment. The proposed Random Forest model classifies Default of Credit.

2. **Deep Neural Network a Step by Step Approach to Classify Credit Card Default Customer** [8]

    2.1. This study and research aimed is to classify and predict the credit card default customers' payment by means of contemporary approach of artificial neural network (ANN) known as deep neural network. This paper explains the dataset which signifies Taiwan credit card defaults in 2005 and their previous payment histories taken from popular machine learning dataset resource known as UCI. The paper enlightens each and every concept and step require to build, train, validate and test a deep neural network model for classification task that has never been discussed before. Moreover, we tried to elaborate the relevant and important concepts associated with deep neural network model that must be kept in mind during model building. This paper mainly tries to classify the default payment customer with more than 82% accuracy. For this purpose, various deep neural network techniques with different libraries are used to attain maximum accuracy and we have tried to build a best possible model which can be used for future prediction. This study proves deep neural network is the only one that can accurately estimate the real probability of default. So, by using this network model, which is more complex, sophisticated and most widely used than a simple neural

network and logistic regression model, the classification simulation shall have a better performance and accuracy.

3. **The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients** [9]**.**

3.1. This research aimed at the case of customers' default payments in Taiwan and compares the predictive accuracy of probability of default among six data mining methods. From the perspective of risk management, the result of predictive accuracy of the estimated probability of default will be more valuable than the binary result of classification - credible or not credible clients. Because the real probability of default is unknown, this study presented the novel ''Sorting Smoothing Method" to estimate the real probability of default. With the real probability of default as the response variable (Y), and the predictive probability of default as the independent variable (X), the simple linear regression result (Y = A + BX) shows that the forecasting model produced by artificial neural network has the highest coefficient of determination; its regression intercept (A) is close to zero, and regression coefficient (B) to one. Therefore, among the six data mining techniques, artificial neural networks are the only one that can accurately estimate the real probability of default.

4. Detecting Default Payment Fraud in Credit Cards [10]

4.1. A credit card account is said to be in default if the payment into the credit card account is not done in due time because of credit card fraud. In many cases the default account credit card holders are found to be fraudsters. It is necessary for financial Institutions to determine which accounts may go into default so that they can take necessary preventive steps to curb any fraudulent transactions. The transactions of the clients can be observed and using this information if any account resembles a fraud, the financial institutions can take necessary actions like blocking the account. We have considered some standard machine learning algorithms for the classification of the default accounts. Later, we have proposed a model by combining the classification algorithms and critically examined it with other machine learning and deep learning models

---------End of Chapter 2-------

Chapter 3
Work Done

Dataset

1. Our dataset was taken from *UCI Machine Learning Repository* [11] . The dataset for *"Default of Credit Card Clients"* [1] was chosen.
2. This dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005.
3. A csv file of the original dataset is provided in the *Resources (see 1)* section of this document.
4. The dataset has a total of ***23 features or independent variables*** & ***1 dependant variable***
5. The dataset has a total of ***30,000 rows & 24 columns.***
6. The dataset contains a mix of qualitative as well quantitative features.
7. **Categorical Features Handling:**
    a. The dataset has 4 categorical variables.
    b. All Categorical variables were already in *Nominal Numeric Encoding* (for example in Gender column: 1=male & 2=female).
    c. Detailed description of preparation of categorical variables to feed to our models is discussed later. *(see here)*
8. Detailed description of the dataset with all the features and description is presented in the table below:

Variables	Variable Names	Variable Type	Description
$s^{(i)}$	ID	Numeric (Unique)	Id of each client
$x_{(1)}$	LIMIT_BAL	Numeric	Amount of given credit in NT dollars (includes individual and family/supplementary credit

$x_{(2)}$	SEX	Categorical	Gender (1=male, 2=female)
$x_{(3)}$	EDUCATION	Categorical	1:graduate school, 2:university, 3:high school, 4:others
$x_{(4)}$	MARRIAGE	Categorical	Marital status (1=married, 2=single, 3=others)
$x_{(5)}$	AGE	Categorical	Age in years
$x_{(6)}$	PAY_1	Numeric	Repayment status in September, 2005 (0=pay duly, 1=payment delay for one month, 2=payment delay for two months, ... 9=payment delay for nine months and above)
$x_{(7)}$	PAY_2	Numeric	Repayment status in August, 2005 (scale same as above)
$x_{(8)}$	PAY_3	Numeric	Repayment status in ,July 2005 (scale same as above)
$x_{(9)}$	PAY_4	Numeric	Repayment status in June, 2005 (scale same as above)
$x_{(10)}$	PAY_5	Numeric	Repayment status in May, 2005 (scale same as above)
$x_{(11)}$	PAY_6	Numeric	Repayment status in April, 2005 (scale same as above)
$x_{(12)}$	BILL_AMT1	Numeric	Amount of bill statement in September, 2005 (NT dollar)
$x_{(13)}$	BILL_AMT2	Numeric	Amount of bill statement in August, 2005 (NT dollar)
$x_{(14)}$	BILL_AMT3	Numeric	Amount of bill statement in July, 2005 (NT dollar)
$x_{(15)}$	BILL_AMT4	Numeric	Amount of bill statement in June, 2005 (NT dollar)
$x_{(16)}$	BILL_AMT5	Numeric	Amount of bill statement in May, 2005 (NT dollar)
$x_{(17)}$	BILL_AMT6	Numeric	Amount of bill statement in May, 2005 (NT dollar)
$x_{(18)}$	PAY_AMT1	Numeric	Amount of previous payment in September, 2005 (NT dollar)
$x_{(19)}$	PAY_AMT2	Numeric	Amount of previous payment in August, 2005 (NT dollar)
$x_{(20)}$	PAY_AMT3	Numeric	Amount of previous payment in July, 2005 (NT dollar)

$x_{(21)}$	PAY_AMT4	Numeric	Amount of previous payment in June, 2005 (NT dollar)
$x_{(22)}$	PAY_AMT5	Numeric	Amount of previous payment in May, 2005 (NT dollar)
$x_{(23)}$	PAY_AMT6	Numeric	Amount of previous payment in April, 2005 (NT dollar)
y	default.payment.next.month	Categorical	Default payment (1=yes, 0=no)

Table 3.1: Description of Dataset

Analysis & Cleaning of Dataset

A few inconsistencies were spotted in the dataset. To maintain logical consistency throughout the project we shall analyse & clean all the inconsistencies present. Help from a Kaggle Notebook [12] was taken for cleaning. All of them are discussed in this section.

Before proceeding further, we would like to reiterate that description of variables are given in Table 3.1 *(see here)*.

Cleaned dataset is provided in csv form in Resources point 2 *(see here)*

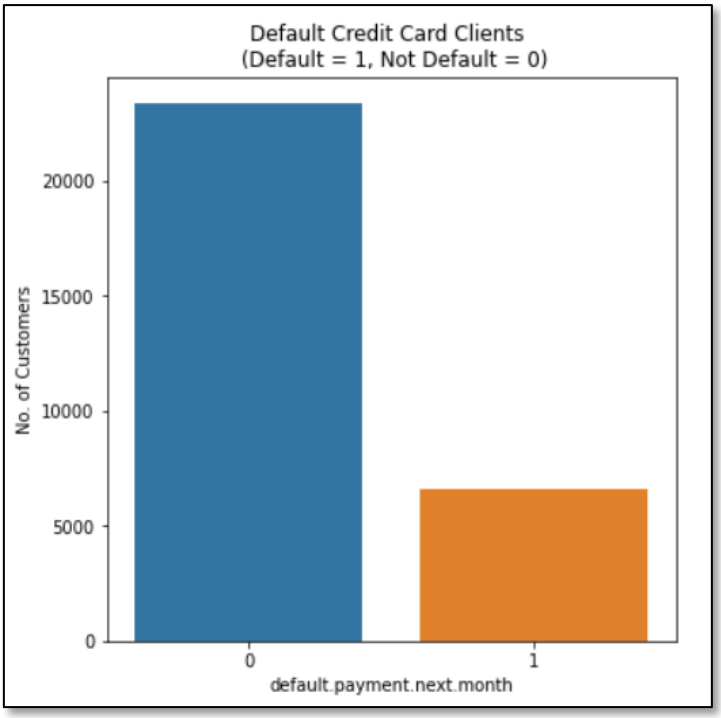
The following table lists all the inconsistencies and cleaning methods used:

Inconsistency in Dataset	Remedy Used
EDUCATION column has categories 5 & 6 which are undocumented	Entries of EDUCATION with 5 & 6 are clubbed with category 4:unknown
MARRIAGE column has a category 0 in the dataset which is undocumented	Entries of MARRIAGE with 0 are merged with category 3: unknown
PAY_n* has entries of -2 & 0 which are not documented *(n ={1,2,3,4,5,6})	Entries of PAY_n with -2 & 0 are merged with category -1: paid duly . Further -1 category was changed to label 0.

Table 3.2 Inconsistency in dataset & Remedy used

Exploratory Analysis

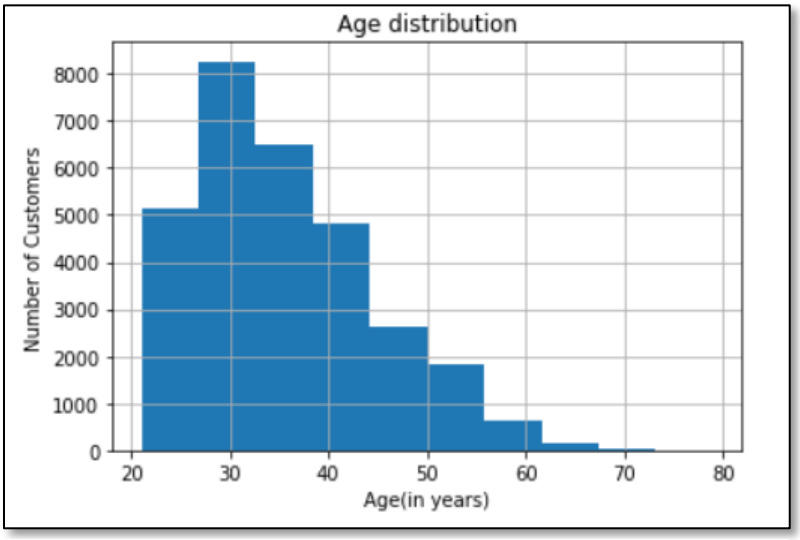
Bar Graph of Defaulters & Non- Defaulters



- The graph shows the large ***class imbalance in the dataset***
- About ***77% for Non- defaulters*** & ***22% for Defaulters.***
- We will need to take in account the class – imbalance while training & evaluating our model.

Figure 1 Chapter 3 Imbalance in Dataset

Histogram of Age of Customers



- We see that the graph is skewed towards the younger age group.
- Maybe the reason for high default in Taiwan could be lending excessively to younger population.

Figure 2 Chapter 3 Age v/s No of Customers

Bar Graph of Marital Status of Customers

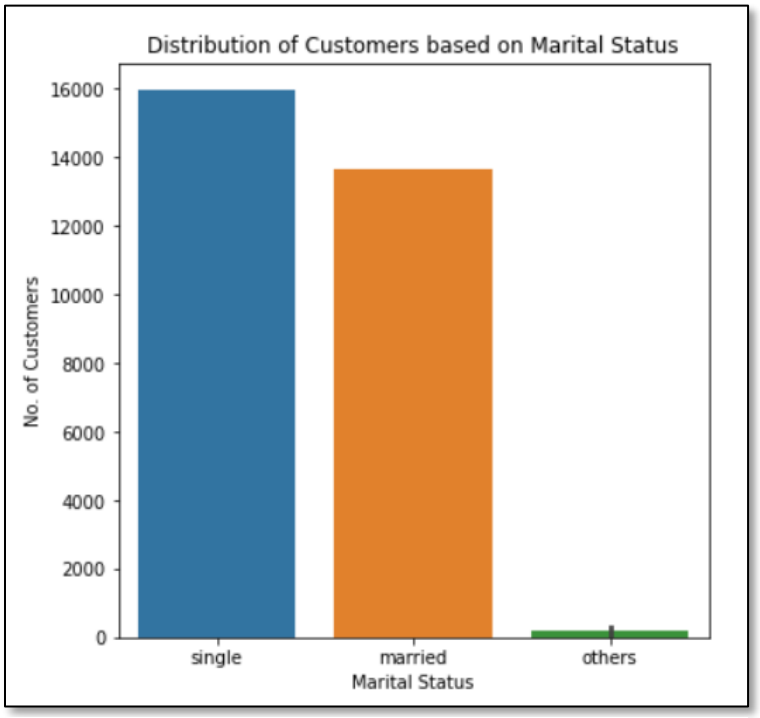


Figure 3 Chapter 3 Marital Status v/s Customers

- Highest number of credit cards were given to Single people.
- It seems that younger and single population might be the key driving force behind the relatively high defaulting rate.

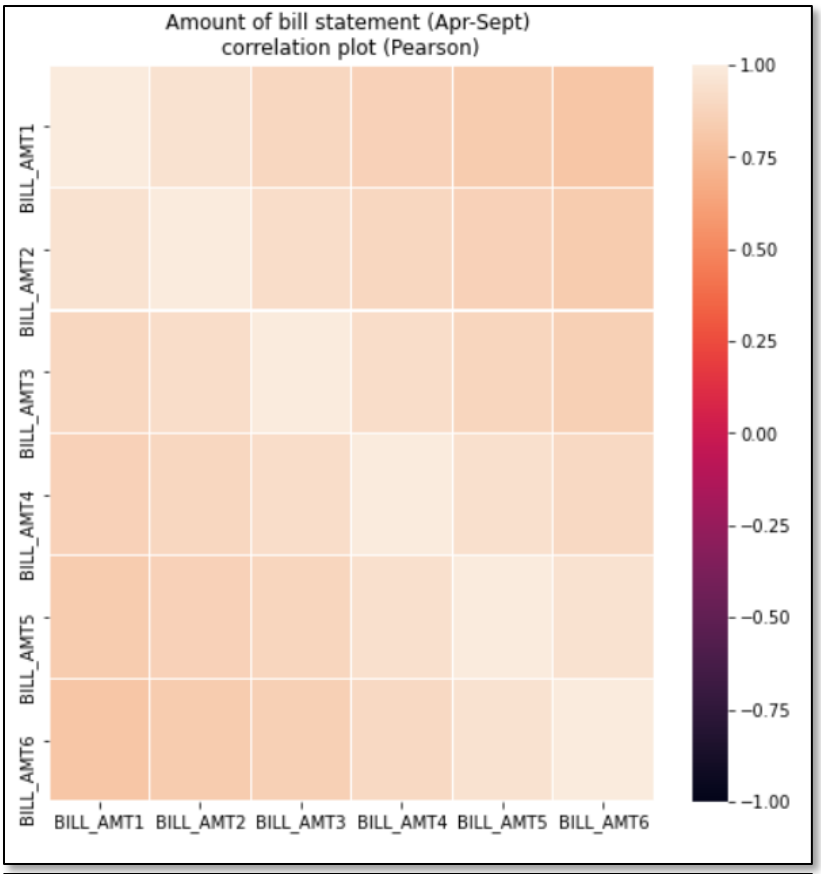
Matrix of Default v/s Marital Status

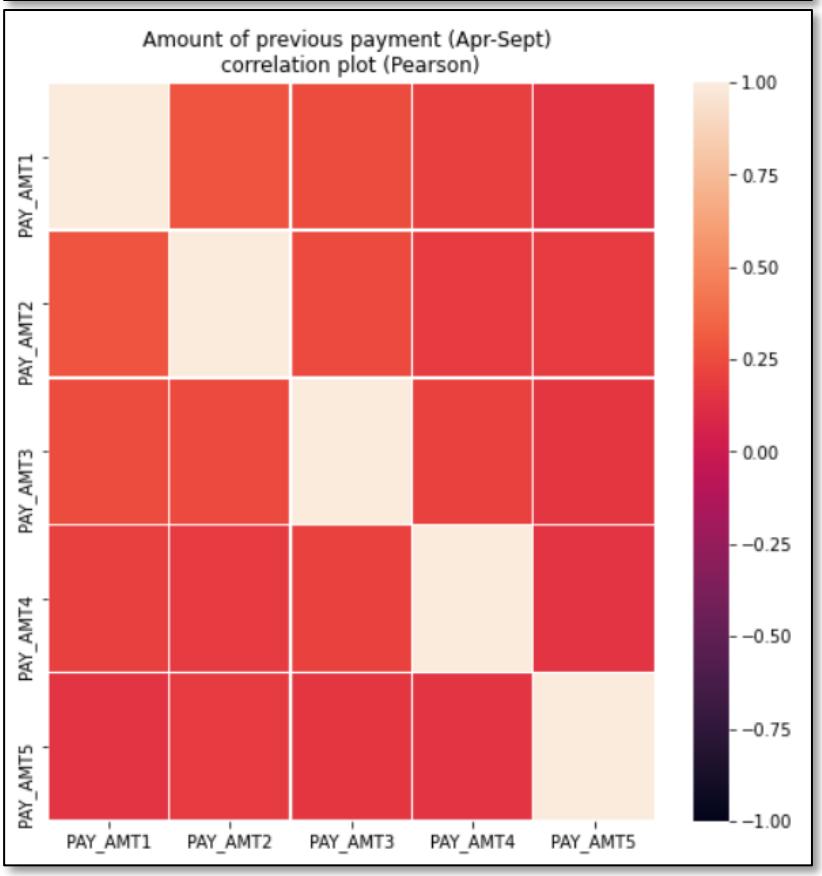
def_pay	0	1	perc
MARRIAGE			
1	10453	3206	0.234717
2	12623	3341	0.209283
3	288	89	0.236074

Figure 4 Chapter 3 Default v/s Marital Status

- This table has brought a new perspective.
- Although the customers are majorly young but they young & single are equally likely to default as the elderly.
- We say category 3 as elderly because our analysis shows that category 3 has customers majorly above 50 years of age

Correlation Plots





- Correlation plots between all the features were plotted, however in the interest of space only two important portions are presented.
- No definitive correlation was found between dependent variable and any of the independent variables.
- We observe that for `BILL_AMT_n` there is a high correlation which indicates towards constant credit borrowing of the customer.
- However, payment `PAY_AMT_n` which should also show a similar trend doesn't manifest such trend. It shows that the Payment done by the customers is fairly irregular over the months.
- The correlation plots gave us confidence to not implement Feature Reduction methods on our dataset

Figure 5 Chapter 3 Correlation Plots

Data Pre-Processing (Categorical & Numeric Variables)

1. **Categorical Variables Pre-Processing**
    a. **Dummy Variables (Binary Encoding)**
        i. All the categorical variables were binary encoded. In this type of encoding, the categorical feature is first converted into numerical using an ordinal encoder. Then the numbers are transformed in the binary number. After that binary value is split into different columns. Since, we already had our categorical features in numerical form we only transformed those to binary numbers and then we made a different column for each digit of this binary equivalent.
        ii. A csv file of encoded data is provided in Resources point 3 *(see here)*
2. **Data Pre-Processing**
    a. **Data Normalization**
        i. Data normalisation is the step which is used to remove the dominance of a particular feature by bringing all the features on same scale.
        ii. Normalized data to zero mean and unit variance
        iii. The formula used is: $x_j^{(i)\prime} = \left(x_j^{(i)} - \mu\right)/\sigma$

Train – Test Split

Throughout the project we have employed a 75-25 train-test split. Separate data frames for the following were made:

1. X_train: Data frame of all features in the training set. *(csv file : see here)*
2. Y_train: Dependant variable data frame of the training set. *(csv file : see here)*
3. X_ test: Data frame of all features in the testing set. *(csv file : see here)*
4. Y_test: Dependant variable data frame of the testing set. *(csv file : see here)*

Data Manipulation

We employed a few data manipulation techniques to overcome the class imbalance. All of the below-mentioned techniques are only applied on training dataset, details about the methods used are given:

1. **SMOTE (Synthetic Minority Over Sampling Technique):**
    a. SMOTE is an over-sampling method. It creates synthetic (not duplicate) samples of the minority class. Hence making the minority class equal to the majority class. [13]
    b. A csv file of SMOTE training dataset is provided *(see here)*

2. **Over-sampling**
    a. It creates duplicate samples of the minority class. Hence making the minority class equal to the majority class It is synthetic
    b. A csv file of Over-Sampling training dataset is provided *(see here)*

3. **Under-sampling**
    a. Under sampling consists of reducing the data by eliminating examples belonging to the majority class with the objective of equalizing the number of examples of each class. [14]
    b. A csv file of Under-Sampling training dataset is provided *(see here)*

Performance Metrics Used

The performance metrics employed by us are listed below.

Here, `TP` means True Positive, `TN` means True Negative, `FP` means False Positive and `FN` means False Negative.

1. **Recall:** Also called Sensitivity and Recall, `TP` rate is the ratio of correctly classified positives to total positives. This tells us how sensitive our classification technique is in the detection of abnormal (positive) events. A classification method with high sensitivity will rarely miss the positive event when it occurs.

$$recall = \frac{true\ positives}{true\ positives\ + false\ negatives}$$

2. **Precision:** It is the ratio of the correctly classified positives by the total number of predicted positives. High Precision indicates samples labelled as positives are indeed positive. While recall expresses the ability to find all positive sample in the dataset, precision expresses the proportion of the samples our method says were positive to those that actually were positive.

$$\text{Precision} = \frac{TP}{TP + FP}$$

3. **F_1-Score:** In statistical analysis of binary classification, the F-score or F-measure is a measure of a test's accuracy. It is calculated from the precision and recall of the test.

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

   The F1 score is the harmonic mean of the precision and recall. The highest possible value of an F-score is 1, indicating perfect precision and recall, and the lowest possible value is 0, if either the precision or the recall is zero.

4. **Precision-Recall Curves:** The precision-recall curve as the name suggests plots precision on y-axis and recall on x-axis. High scores for both precision and recall indicate

that the classifier is returning accurate results as well as returning a majority of all positive results.

a. Optimum Point: The optimum threshold point in a Precision-Recall Curve is found out by locating the point having maximum F_1 score in the curve and the threshold corresponding to it is taken as the decision threshold of the model.

b. Break-Even Point: The Break-even point on the precision-recall curve is sought when the business need of the model is not fully clear. In this method we locate the point where precision equals recall and use the decision threshold in our model.

PTO --->

Model Fitting

During the whole course of the project, we applied a lot of classification algorithms with different input formats, hyper parameters and new features. However, in the interest of this report we are reporting a few of the best models. Our overall work is divided into three major parts.

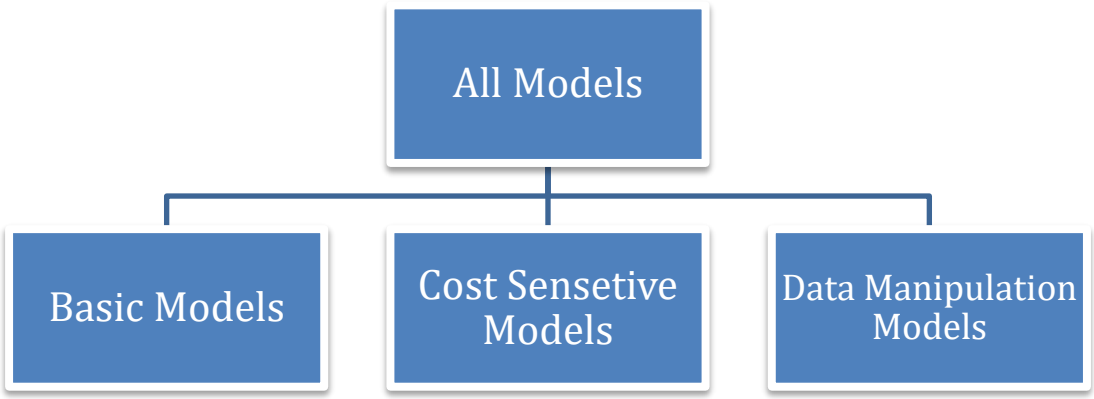


Figure 6 Chapter 3 Models Description

Overall, we have used two ANNs.structures of both the ANNs used is given below

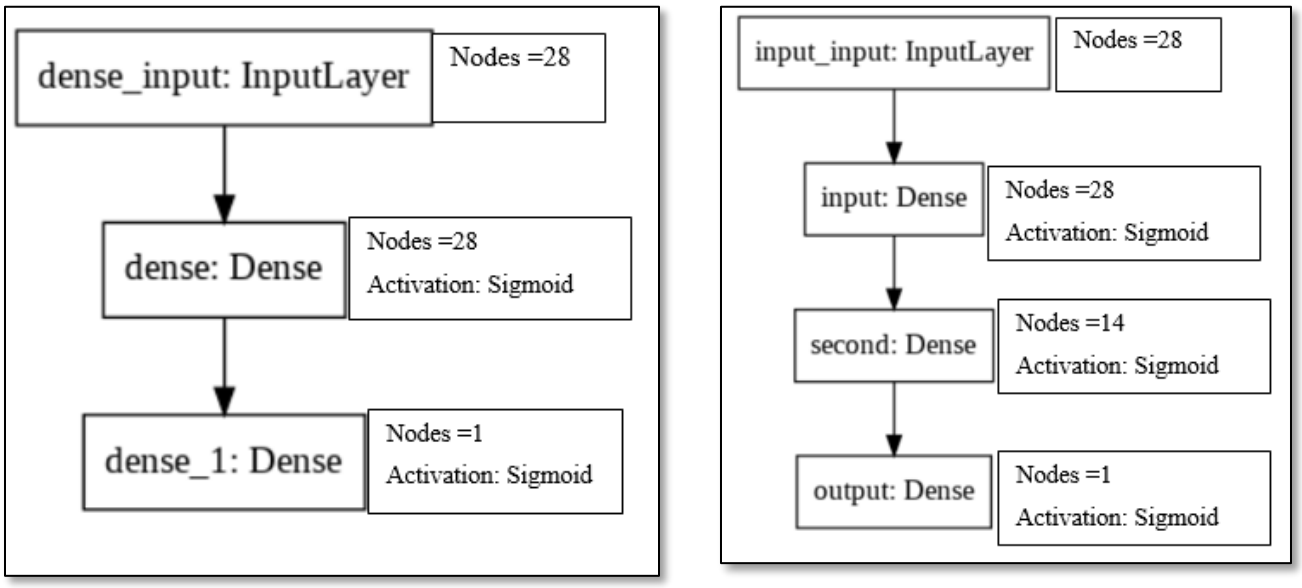


Figure 7 Chapter 3 ANN-1 & ANN-2 Structures respectively

Basic Models

- Basic Models are models where the dummy (binary encoded) dataset was fed.
- No threshold shifting or data manipulation methods were employed.
- Default Threshold = 0.5

Algorithm	***Class 0 Precision & Recall***	***Class 1 Precision & Recall***	***Accuracy Rate***	***Misclassification Rate***	***F_1-score*** ***Class -0*** ***Class-1***
Logistic Regression	0.84 0.95	0.69 0.36	82.04%	17.96%	0.89 0.47
SVM	0.83 0.96	0.69 0.33	81.64%	18.36%	0.89 0.44
KNN	0.83 0.96	0.66 0.30	80.97%	19.03%	0.89 0.42
Random Forest	**0.84** **0.94**	**0.65** **0.39**	**81.54%**	**18.46%**	**0.89** **0.48**
ANN-1	0.84 0.95	0.64 0.35	81.64%	18.36%	0.89 0.45
ANN-2	0.84 0.95	0.68 0.36	82.41%	17.59%	0.89 0.47

Table 3.3 Results of Basic Models

Precision-Recall Curve of the Best Model (in yellow)

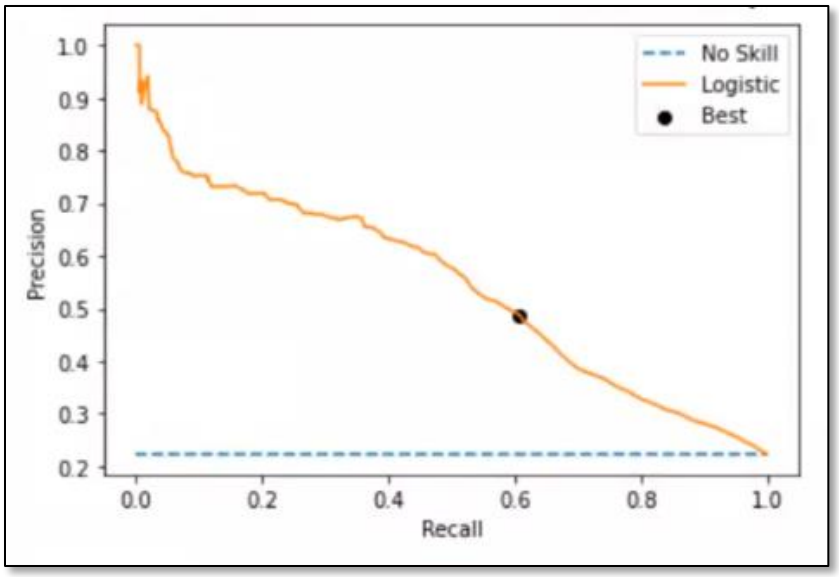


Figure 8 Chapter 3 Precision-Recall Curve of Best Baseline Model

Confusion Matrix

	Actual	
Predicted	**0**	**1**
0	**5470**	**354**
1	**1030**	**646**

Cost Sensitive Models (Threshold Shifting)

- Cost Sensitive Models are improvements of the models in the previous section. We employed threshold-shifting on models based on precision-recall curve.
- The optimal point on the PR curve is found by locating the point with the highest F_1 score. The threshold given by this point is then used as decision threshold.

Algorithm	***Class 0 Precision & Recall***	***Class 1 Precision & Recall***	***Accuracy Rate***	***Misclassification Rate***	***Threshold***	***F_1-score Class -0 Class-1***
Logistic Regression	0.87 0.84	0.50 0.57	78.20%	21.80%	0.23	0.86 0.53

SVM	0.86 0.89	0.56 0.49	80.00%	20.00%	0.17	0.87 0.52
KNN	0.86 0.88	0.55 0.50	79.58%	20.42%	0.30	0.87 0.52
Random Forest	0.88 0.83	0.50 0.59	77.44%	22.16%	0.29	0.85 0.54
ANN-1	0.87 0.82	0.47 0.58	76.80%	23.20%	0.25	0.85 0.52
ANN-2	**0.87** **0.86**	**0.52** **0.56**	**79.32%**	**20.68%**	**0.27**	**0.87** **0.54**

Table 3.4 Results of Cost-Sensitive Learning (Threshold Shifting)

Precision-Recall Curve of the Best Model (in yellow)

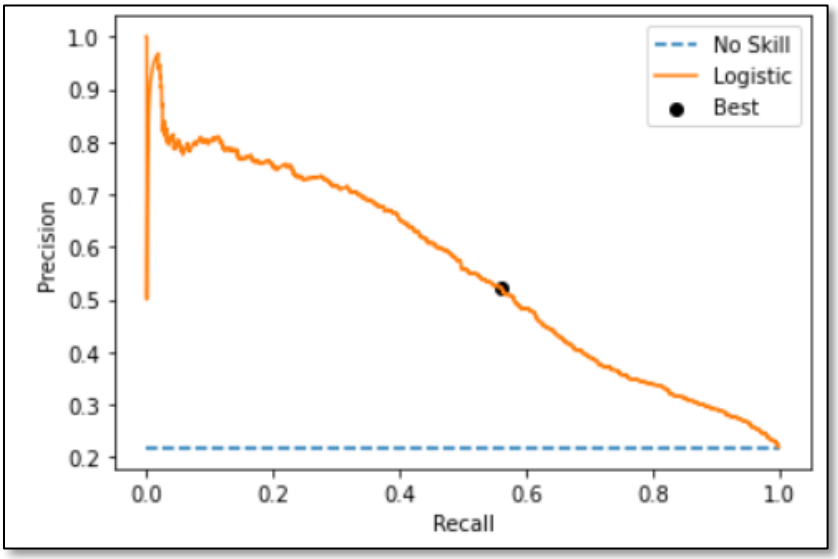


Figure 9 Chapter 3 Precision Recall Curve of Best Model in Cost Sensitive (Threshold) Model

Confusion Matrix

	Actual	
Predicted	**0**	**1**
0	**1083**	**353**
1	**317**	**99**

Cost Sensitive Models (Break-Even Point)

- Cost Sensitive Models in this section are further improvements of the models in the previous section. We have employed threshold-shifting on models based on precision-recall curve.
- The break-even point on the PR curve is found by locating the point where precision and recall are equal. The threshold given by this point is then used as decision threshold.

Algorithm	***Class 0 Precision & Recall***	***Class 1 Precision & Recall***	***Accuracy Rate***	***Misclassification Rate***	***Threshold***	***F_1-score*** ***Class -0*** ***Class-1***
Logistic Regression	0.86 0.86	0.53 0.53	79.00%	21.00%	0.27	0.86 0.53
SVM	0.86 0.86	0.52 0.52	78.34%	21.66%	0.16	0.86 0.52
KNN	**0.87** **0.87**	**0.52** **0.52**	**79.59%**	**20.41%**	**0.30**	**0.87** **0.52**
Random Forest	0.85 0.85	0.54 0.54	77.89%	22.11%	0.30	0.85 0.54
ANN-1	0.86 0.86	0.53 0.53	78.26%	21.74%	0.25	0.86 0.53
ANN-2	0.86 0.86	0.52 0.52	78.26%	21.74%	0.26	0.86 0.53

Table 3.5 Results of Cost-Sensitive Learning (Break-Even Point)

Precision-Recall Curve of the Best Model (in yellow)

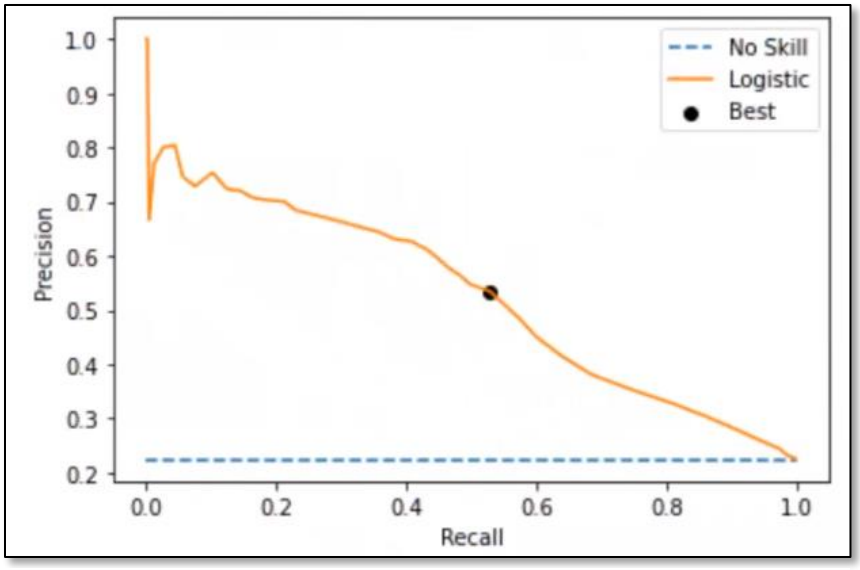


Figure 10 Chapter 3 Precision Recall Curve of Best Model in Cost Sensitive (Break Even) Model

Confusion Matrix

	Actual	
Predicted	**0**	**1**
0	**5134**	**690**
1	**841**	**835**

Data-Manipulated Models

Using Under-sampled data

Algorithm	*Class 0 Precision & Recall*	*Class 1 Precision & Recall*	*Accuracy Rate*	*Misclassification Rate*	*F_1-score* *Class -0* *Class-1*
Logistic Regression	0.87 0.83	0.49 0.58	77.36%	22.64%	0.85 0.54
SVM	0.88 0.77	0.45 0.63	74.2%	25.8%	0.82 0.52
KNN	0.87 0.84	0.50 0.56	77.48%	22.52%	0.85 0.53
Random Forest	0.88	0.45	74.36%	25.64%	0.82

	0.77	0.65			0.53
ANN-1	**0.87** **0.88**	**0.54** **0.52**	**79.94%**	**20.06%**	**0.87** **0.53**
ANN-2	0.86 0.86	0.50 0.49	78.20%	21.80%	0.86 0.50

Table 3.6 Results of Under Sampling Dataset

Using Over-sampled data

Algorithm	***Class 0 Precision & Recall***	***Class 1 Precision & Recall***	***Accuracy Rate***	***Misclassification Rate***	***F_1-score*** ***Class -0*** ***Class-1***
Logistic Regression	0.87 0.83	0.50 0.58	77.47%	22.53%	0.85 0.53
SVM	0.88 0.79	0.45 0.62	74.87%	25.13%	0.83 0.52
KNN	0.88 0.76	0.43 0.63	72.91%	27.09%	0.81 0.51
Random Forest	**0.85** **0.91**	**0.59** **0.45**	**80.8%**	**19.2%**	**0.88** **0.51**
ANN-1	0.86 0.88	0.52 0.49	79.14%	20.86%	0.87 0.50
ANN-2	0.86 0.87	0.52 0.50	79.05%	20.95%	0.87 0.51

Table 3.7 Results of Over Sampling Dataset

Using SMOTE data

Algorithm	*Class 0 Precision & Recall*	*Class 1 Precision & Recall*	*Accuracy Rate*	*Misclassification Rate*	*F_1-score* *Class -0* *Class-1*
Logistic Regression	0.87 0.83	0.50 0.58	77.4%	22.6%	0.85 0.53
SVM	0.88 0.79	0.45 0.61	74.76%	25.24%	0.83 0.52
KNN	0.88 0.71	0.40 0.66	69.96%	30.04%	0.79 0.50
Random Forest	**0.86** **0.87**	**0.53** **0.50**	**78.93%**	**21.07%**	**0.87** **0.52**
ANN-1	0.86 0.86	0.49 0.48	77.76%	22.24%	0.86 0.49
ANN-2	0.85 0.85	0.46 0.48	76.64%	23.36%	0.85 0.85

Table 3.8 Results of SMOTE Dataset

Precision-Recall Curve of the Best Model (in yellow)

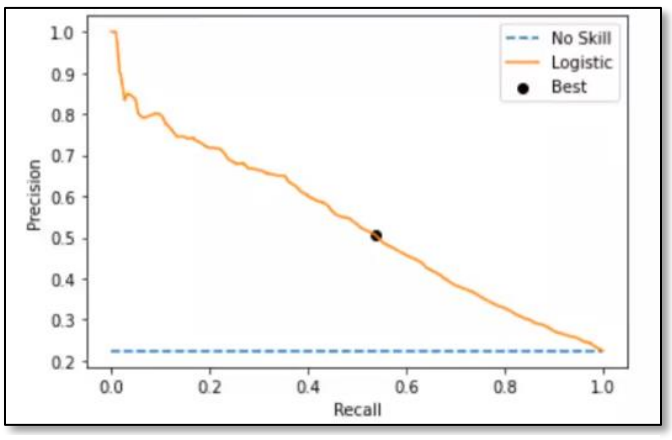


Figure 11 Chapter 3 Precision-Recall Curve of the best model in Data Manipulation

Confusion Matrix

	Actual	
Predicted	**0**	**1**
0	**4998**	**826**
1	**800**	**876**

Chapter 4
Conclusion

Comparison with Literature

Final Remarks

Throughout the project we have learned and observed a lot of things. Following are our concluding remarks:

1. There is a visible trade-off between identifying higher number of defaulters at the cost of genuine customers being labelled wrongly. Also, Identification of good customers at a higher level comes with a lower level of detecting defaulters.

2. This trend is supported by the fact that if *recall* for class 0 goes higher in one model then the *recall* for class 1 gets lower and vice-versa.

3. There was a general improvement observed in all our models with respect to basic models when we applied cost-sensitive (threshold shifting) models.

4. At this point, one may be presented with the dilemma of selecting the best model for deployment. This problem is a managerial issue and hence should only be decided by the concerned banks. In this project we did not aspire to put any benchmark on precision and recall values. As a result, we have chosen to report al our models at the *break-even point* and declare the model with highest F_1 score as the best. With that being said we still feel a thorough examination of the business problem at the deploying institution shall be done.

Chapter 5
Future Prospects

Anomaly Detection in banking systems has always been a challenge for financial bodies around the world. Each and every money lending body around the world loses some of its capital because of various types of frauds that are committed either by the customers or with the customers. Credit Card default is one of such problems which we are trying to cope up with.

We believe that if we are able to detect the default customers before lending them money or credit card in general then we can to some extent reduce the burden on our money lending bodies and hence we will also be able to reduce one type of frauds that prevailing in our financial systems.

The future prospects of this project are by suggesting the best machine learning Model which we can apply for solving this type of problem. We are trying to help the governing bodies to reduce the number of credit-card defaulters and curb the crime in this sector of financial services. We would also encourage app developers to use our findings for making a user-friendly app for bank officials specially to ease the process of identifying whether a particular customer is going to be a defaulter or not in the coming months.

References

[1] I.-C. Yeh, “UCI Machine Learning, Credit Card Default Dataset,” [Online]. Available: https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients.

[2] L. Irby, "thebalance.com," 04 June 2020. [Online]. Available: https://www.thebalance.com/pros-and-cons-of-credit-cards-960222.

[3] J. Chen, "Investopedia.com," [Online]. Available: https://www.investopedia.com/terms/d/debtinstrument.asp.

[4] P. Kosakowski, "Investopedia," [Online]. Available: https://www.investopedia.com/articles/economics/09/subprime-market-2008.asp.

[5] "SAS.com," [Online]. Available: https://www.sas.com/en_us/insights/risk-management/credit-risk-management.html.

[6] sevenpillarsinstitute.org, [Online]. Available: https://sevenpillarsinstitute.org/case-studies/taiwans-credit-card-crisis/.

[7] A. S. a. S. Cankurt, ""Prediction of default payment of credit card clients using Data Mining Techniques,"," in *International Engineering Conference (IEC)*, , Erbil, Iraq, 2019.

[8] S. M. A. Waseem Ahmad Chishti, "Deep Neural Network a Step by Step Approach to Classify Credit Card Default Customer," in *2019 International Conference on Innovative Computing (ICIC)*, 2019.

[C.-h. L. I-Cheng Yeh a, "The comparisons of data mining techniques for the
9 predictive accuracy of probability of default of credit card clients," *Expert*
] *Systems with Applications*, vol. 36 (2009) 2473–2480, no. Issue 2, Part 1, pp. Pages 2473-2480,, 2009.

[A. S. K. D. P. a. S. K. R. S. S. H. Padmanabhuni, ""Detecting Default Payment
1 Fraud in Credit Cards,"," in *2019 IEEE International Conference on Intelligent*
0 *Systems and Green Technology (ICISGT)*, Visakhapatnam, India,, 2019.
]

[UCI, "UCI MACHINE LEARNING," Center for Machine Learning and
1 Intelligent Systems, [Online]. Available:
1 https://archive.ics.uci.edu/ml/index.php.
]

[L. Basanisi, "Kaggle.com," [Online]. Available:
1 https://www.kaggle.com/lucabasa/credit-card-default-a-very-pedagogical-
2 notebook.
]

[S. A. Rahim, "SMOTE AND NEAR MISS IN PYTHON: MACHINE
1 LEARNING IN IMBALANCED DATASETS," [Online]. Available:
3 https://medium.com/@saeedAR/smote-and-near-miss-in-python-machine-
] learning-in-imbalanced-datasets-b7976d9a7a79#:~:text=SMOTE%20(Synthetic%20Minority%20Over%2Dsampling%20Technique)&text=SMOTE%20does%20this%20by%20selecting,to%20see%20how%20this%20works..

[J. Brownlee. [Online]. Available:
1 https://machinelearningmastery.com/undersampling-algorithms-for-imbalanced-classification/#:~:text=Undersampling%20refers%20to%20a%20group,has%20

[14] a%20skewed%20class%20distribution.&text=Undersampling%20methods%20can%20be%20used,fit%20a%20machine%20learning.

[15] S.-T. L.-P. W.-S. H. C.-L. Te-Cheng Hsu, *Enhanced Recurrent Neural Network for Combining Static and Dynamic Features for Credit Card Default Prediction*, National Tsing Hua University, Hsinchu 30013, Taiwan: IEEE.

[16] F. A. Armin Lawi, *Classification of Credit Card Default Clients Using LS-SVM Ensemble*, Makassar, Indonesia: IEEE.

[17] A. S. K. D. P. S. k. R. S. S. Harshini Padmanabhuni, "Detecting Default Payment Fraud in Credit Cards," in *2019 IEEE International Conference on Intelligent Systems and Green Technology (ICISGT)*, 2019.

[18] [Online]. Available: https://machinelearningmastery.com/imbalanced-classification-with-python-7-day-mini-course/.

[19] [Online]. Available: https://machinelearningmastery.com/cost-sensitive-decision-trees-for-imbalanced-classification.

Appendix A

Resources

[1] UCI Machine Learning: Credit Card Original Dataset [Link]

[2] Cleaned Dataset [Link]

[3] Dummy Variable (Binary Encoded) Dataset [Link]

[4] Training set of Features (X_train) [Link]

[5] Training set of Dependant Variable (Y_train) [Link]

[6] Testing set of Features (X_test) [Link]

[7] Testing set of Dependant Variable (Y_test) [Link]

[8] Training Set SMOTE Dataset [Link]

[9] Training Set Oversampling Dataset [Link]

[10] Training Set Undersampling Dataset [Link]

---------End of Document-------

Thank You