RISK MODELLING – CREDIT SCORING USING LOGISTIC REGRESSION

A Major Project Report

Submitted to

CHHATTISGARH SWAMI VIVEKANAND TECHNICAL UNIVERSITY

BHILAI (C.G.), INDIA

In partial fulfillment of requirement for the award of the Degree

of

Bachelor of Technology

in

Electronics & Telecommunication Engineering

by

Riya Deshmukh (BG1840)
Shiwani Mishra (BG1850)
Vanshika Agrawal (BG1869)

Under the Guidance of
Prof. Reynolds Duddu
Assisstant Professor

DEPARTMENT OF ELECTRONICS & TELECOMMUNICATION ENGINEERING

BHILAI INSTITUTE OF TECHNOLOGY, BHILAI HOUSE, DURG (C.G.) -491001, INDIA

ASPIRE TO EXCEL

SESSION 2021-2022

RISK MODELLING – CREDIT SCORING USING LOGISTIC REGRESSION

A Major Project Report

Submitted to

CHHATTISGARH SWAMI VIVEKANAND TECHNICAL UNIVERSITY

BHILAI (C.G.), INDIA

In partial fulfillment of requirement for the award of the Degree

of

Bachelor of Technology

in

Electronics & Telecommunication Engineering

by

Riya Deshmukh (BG1840)
Shiwani Mishra (BG1850)
Vanshika Agrawal (BG1869)

Under the Guidance of
Prof. Reynolds Duddu
Assisstant Professor

DEPARTMENT OF ELECTRONICS & TELECOMMUNICATION ENGINEERING

BHILAÍ INSTITUTE OF TECHNOLOGY, BHILAI HOUSE, DURG (C.G.) -491001, INDIA

SESSION 2021-2022

DECLARATION

I the undersigned solemnly declare that the report of the Project work entitled ***"RISK MODELLING – CREDIT SCORING USING LOGISTIC REGRESSION"***, is based on my own work carried out during the course of my study under the supervision of Prof. Reynolds Duddu, Department of Electronics and Telecommunication Engineering, Bhilai Institute of Technology, Durg, Chhattisgarh.

I assert that the statements made and conclusions drawn are an outcome of the project work. I further declare that to the best of my knowledge and belief that the report does not contain any part of any work which has been submitted for the award of any other degree/diploma/certificate in this University/ deemed University of India or any other country. All help received and citations used for the preparation of the Report have been duly acknowledged.

Riya Deshmukh
Roll No.: 300102818042
Enrollment No.: BG1840
Department of Electronics
& Telecommunication Engg.
BIT, Durg

Shiwani Mishra
Roll No.: 300102818032
Enrollment No.: BG1850
Department of Electronics
& Telecommunication Engg.
BIT, Durg

Vanshika Agrawal
Roll No.: 300102818034
Enrollment No.: BG1869
Department of Electronics
& Telecommunication Engg.
BIT, Durg

CERTIFICATE BY THE SUPERVISOR

This is to certify that the report of the Project submitted is an outcome of the project work entitled ***"Risk Modelling – Credit Scoring using Logistic Regression"***, carried out by

Riya Deshmukh bearing ***Roll No.: 300102818042 & Enrollment No. : BG1840***
Shiwani Mishra bearing ***Roll No.: 300102818032 & Enrollment No. : BG1850***
Vanshika Agrawal bearing ***Roll No.: 300102818034 & Enrollment No. : BG1869***

carried out under my guidance and supervision for the award of Degree, ***Bachelor of Technology in Electronics and Telecommunication Engineering*** of Chhattisgarh Swami Vivekanand Technical University, Bhilai (C.G.), India.

To the best of my knowledge and the Report

i. Embodies the work of the candidate him/herself,
ii. Has duly been completed,
iii. Fulfills the requirement of the ordinance relating to the B.E. Degree of the University and
iv. Is up to the desired standard for the purpose of which is submitted.

Signature of the Project Incharge
Prof. Kiran Dewangan
Assisstant Professor
Department of Electronics
& Telecommunication Engg.
BIT, Durg

Signature of the Supervisor
Prof. Reynolds Duddu
Assisstant Professor
Department of Electronics
& Telecommunication Engg.
BIT, Durg

The project work as mentioned above is hereby being recommended and forwarded for examination and evaluation.

Signature of
Head of the Department
With Seal

Head of Electronic &
Telecommunication Engineering Department
Bhilai Institute of Technology
BHILAI HOUSE
durg (C.G.) 491001

CERTIFICATE BY THE EXAMINERS

This is certify that the project work entitled

"Risk Modelling – Credit Scoring using Logistic Regression"

Submitted by

Riya Deshmukh , Roll No. 300102818042 , Enrollment No. BG1840
Shiwani Mishra , Roll No. 300102818032 , Enrollment No. BG1850
Vanshika Agrawal , Roll No. 300102818034 , Enrollment No. BG1869

has been examined by the undersigned as a part of the examination for the award of the ***Bachelor of Technology Degree in Electronics and Telecommunication Engineering of*** Chhattisgarh Swami Vivekanand Technical University, Bhilai, (C.G.).

Internal Examiner
Date: 23/06/2022

External Examiner
Date: 23/06/22

ACKNOWLEDGEMENT

With deep regards and profound respect, I avail this opportunity to express my deep sense of gratitude and indebtedness to **Prof. Reynolds Duddu,** Department of Electronics and Telecommunication Engineering, BIT Durg for his valuable guidance and support. I am deeply indebted for the valuable discussions at each phase of the project. I consider it my good fortune to have got an opportunity to work with such a wonderful person.

I express my sincere gratitude to **Dr. Arun Arora,** Director, **Dr. Mohan Kumar Gupta,** Principal and **Dr. Manisha Sharma,** Vice Principal, Bhilai Institute of Technology, Durg for providing adequate infrastructure to carry out present investigations and also motivating for research work, which has been a constant source of inspiration in completing this work.

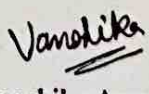
I take immense pleasure to thank **Dr. Naveen Kumar Dewangan,** HOD (ETC), Bhilai Institute of Technology, Durg, for motivating to work in research direction and providing opportunities to connect with global research.

I whole heartedly extend my gratitude to **Prof. Kiran Dewangan,** Department of Electronics and Communication Engineering, BIT Durg for constant feedbacks and encouragements and endless support and help throughout this project work.

Lastly, I feel immensely moved in expressing my indebtedness to my revered parents whose sacrifice, guidance and blessings helped me to complete my work.

Riya Deshmukh

Shiwani Mishra



Vanshika Agrawal

ABSTRACT

In the present day, most of the consumers, purchasers and business ventures are highly dependent on credit from banks. All the financial institutions gives credit on the basis of credit score of the consumer to evaluate the probability that the consumer will repay loan in a timely manner.

Borrower's repayment history, length of credit history, number of credit enquiries in the past and number of active credit cards and loans, all contribute to computation of a statistical three number, called credit score. A statistical analysis is performed by banks and financial institutions to determine a borrower's credit score which is called credit scoring.

Credit risk refers to the chance that a burrower will be unable to make their payments on time and default on their debt. Credit risk modelling refers to the process of using data models to find out two important things. The first is the probability of the burrower defaulting on the loan. The second is the impact on the financials of the lender if this default occurs.

In this project, we have developed a credit risk model, which evaluates the probability of a loan being a good loan and helps the banks and financial institutes to decide whether to give loan to a particular consumer or not. We have built an in – house risk model which gives data driven lending decisions.

The given model is evaluated using logistic regression and decile methodology is used to assess and arrange the data of consumers in a significant way and make the management of the consumer's data easy for the financial institutions. We have also generated ROC curve which helps the financial institutes to maximise their profit.

To sum up, we've analyzed and pre-processed our data, trained and evaluated our model, namely logistic regression, for their ability to predict loan defaults and their probability. This will help the banks to make their processes easier, faster and with minimum errors.

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
AIRB	Advanced Internal Rating Based
AUC	Area Under the ROC Curve
B2B	Business – To – Business
B2C	Business – To – Consumer
CART	Classification And Regression Tree
EAD	Exposur At Default
FICO	Fair Isaac Corporation
LGD	Loss Given Default
NNs	Neural Networks
PD	Probability of Default
ROC	Receiver Operator Characteristics
SVM	Support Vector Machine Models

LIST OF FIGURES

LIST OF TABLES

Table of Contents

CHAPTER – 1

INTRODUCTION

1.1 ABOUT THE PROJECT

In this project, we have developed a credit risk model using machine learning, which will evaluate if the consumer is worthy of giving the credit, based on his credit history, repayment history, number of active credit cards and loans and various other factors.

This model uses the consumer's data to determine whether the banks and financial institutes will get a profit or a loss in giving the consumer the credit he wants. We have applied decile methodology to determine if the loan will be good or bad for the financial institutes.

We have used logistic regression for making ROC curve that helps in studying the model easily. In simple words, this project is a risk model which uses the concept of credit scoring and with the help of logistic regression tells whether the banks and financial institutes can take the risk of giving credit to consumers.

1.2 CREDIT SCORE

Borrower's repayments history, length of credit history, number of credit enquiries in the past, and number of active credit cards and loans; all contribute to computation of a statistical three number, called credit score. This number acts as a proxy for banks to gauge credit worthiness of any loan applicant.

In simple words, banks use this number to make their lending decisions. A credit score is a number between 300 – 850 that depicts a consumer's creditworthiness. The higher the score, the better a borrower looks to potential lenders. A credit score is based on credit history, number of open accounts, total levels of debt, and repayment history, and other factors.

Lenders use credit scores to evaluate the probability that an individual will repay loans in a timely manner. The concept of credit score can be illustrated using the following figure: [8]

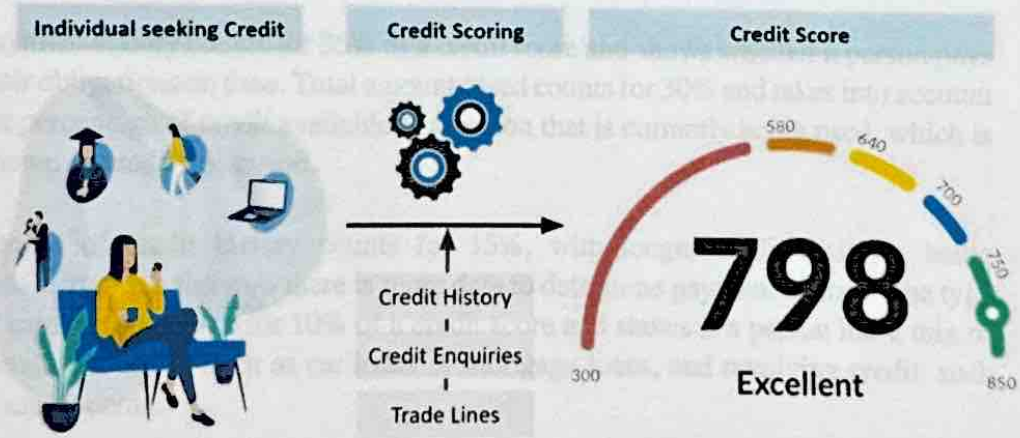


Fig.1.2 (a) Credit History

1.3 HOW CREDIT SCORE WORK?

A credit score can significantly affect your financial life. It plays a key role in a lender's decision to offer you credit. People with credit scores below 640, for example, are generally considered to be subprime burrowers. Lending institutions often charge interest on subprime mortgages at a rate higher than a conventional mortgage in order to compensate themselves for carrying more risk.

They may also require a shorter repayment term or a co-signer for borrowers with a low credit score. Conversely, a credit score of 700 or above is generally considered good and may result in a borrower receiving a lower interest rate, which results in their paying less money in interest over the life of the loan.

Scores greater than 800 are considered excellent. While every creditor defines its own ranges for credit scores, the average FICO score range is often used. A person's credit score may also determine the size of an initial deposit required to obtain a smartphone, cable service or utilities, or to rent an apartment.

And lenders frequently review borrowers' scores, especially when deciding whether to change an interest rate or credit limit on a credit card.

1.4 CREDIT SCORE FACTORS

There are three major credit reporting agencies in the United States which report, update, and store consumers' credit histories. While there can be differences in the information collected by the three credit bureaus, there are five main factors evaluated when calculating a credit score:

- Payment history
- Total amount owed
- Length of credit history
- Types of credit
- New credit

Payment history counts for 35% of a credit score and shows whether a person pays their obligations on time. Total amount owed counts for 30% and takes into account the percentage of credit available to a person that is currently being used, which is known as credit utilization.

Length of credit history counts for 15%, with longer credit histories being considered less risky, as there is more data to determine payment history. The type of credit used counts for 10% of a credit score and shows if a person has a mix of instalment credit, such as car loans or mortgage loans, and revolving credit, such as credit cards.

New credit also counts for 10%, and it factors in how many new accounts a person has, how many new accounts they have applied for recently, which result in credit inquiries, and when the most recent account was opened.

1.5 CHALLENGES WITH CREDIT SCORE

1. **Not all borrowers would have it:**
   All borrowers do not have a significant credit history to have a good credit score, or in fact any credit score

2. **Who are you, as business:**
   Credit score has different significance for different kinds of businesses. It depends how big or small the bank is. A borrower having a good credit score might not visit a small lender. Making this small lender's decision process not so straight forward

1.6 CREDIT SCORING

Credit scoring is basically a statistical analysis performed by banks and financial institutions to determine a borrower's credit score. It is used to determine the creditworthiness of a person or a small, owner-operated business. Credit scoring is used by lenders to help decide whether to extend or deny credit.

Lenders use credit scoring in risk-based pricing in which the terms of a loan, including the interest rate, offered to borrowers are based on the probability of repayment.

1.7 LIMITATIONS OF CREDIT SCORING

Although credit scoring ranks a borrower's credit riskiness, it does not provide an estimate of a borrower's default probability. It merely assesses a borrower's riskiness from highest to lowest. As such, credit scoring suffers from its inability to determine whether Borrower A is twice as risky as Borrower B.

Another interesting limit to credit scoring is its inability to explicitly factor in current economic conditions. If Borrower A has a credit score of 800, for instance, and the economy enters a recession, then Borrower A's credit score would not adjust unless Borrower A's behaviour or financial position changed.

More-advanced methods of credit risk modelling, including structural models and reduced-form models, are used to assess default probability. Advances in technology, such as machine learning and other analytics-friendly computer languages, continue to scientifically refine the accuracy of credit risk modelling.

1.8 CREDIT RISK

Credit risk refers to the chance that a burrower will be unable to make their payments on time and default on their debt. It refers to the risk that a lender may not receive their interest due or the principal lent on time. This results in an interruption of cash flows for the lender and increase the cost of collection.

In extreme cases, some part of the loan or even the entire loan may have to be written off resulting in a loss for the lender. It is extremely difficult and complex to pinpoint exactly how likely a person is to default on their loan. At the same time, properly assessing credit risk can reduce the likelihood of losses from default and delayed repayment.

Interest payments from the borrower are the lender's reward for bearing credit risk. If the credit risk is higher, the lender or investor will either charge a higher interest or forego the lending opportunity altogether. For example, a loan applicant with a superior credit history and steady income will be charged a lower interest rate for the same loan than an applicant with a poor credit history.

1.9 DIFFERENT TYPES OF CREDIT RISK

There are a number of different types of credit risk which arise based on the type of loan and the situation. Obviously, different credit risk models work better for different kind of credit and credit risk model validation differs accordingly. Here are some credit risks that lenders undertake.

- There is a risk that an individual borrower may fail to make a payment due on a credit card, a mortgage loan, line of credit, or any other personal loan.
- A business or individual fails to pay a trade invoice on the due date. This is a common risk that both B2B and B2C businesses that work on credit carry.
- A company that burrows money is unable to repay fixed or floating charge debt.
- An insurance company that is insolvent does make a claim payment which is due.
- A company or a government may have issued a bond that it does not pay the interest or principal amount on.
- A business does not pay an employee's salary or wages when they become due.
- A bank that is now bankrupt does not return money that has been deposited.

1.10 RISK MODELLING

There are many different factors that affect a person's credit risk. This makes assessing a burrower's credit risk a highly complex task. With so much money

riding on our ability to accurately estimate the credit risk of a borrower, credit risk modelling has come into the picture.

Credit risk modelling refers to the process of using data models to find out two important things. The first is the probability of the burrower defaulting on the loan. The second is the impact on the financials of the lender if this default occurs.

Financial institutions rely on credit risk models to determine the credit risk of potential borrowers. They make decisions on whether or not to sanction a loan as well as on the interest rate of the loan based on the credit risk model validation. As technology has progressed, new ways of modelling credit risk have emerged including credit risk modelling using R and Python.

These include using the latest analytics and big data tools to model credit risk. Other factors like the evolution of economies and the subsequent emergence of different types of credit risk have also impacted how credit risk modelling is done.

1.11 FACTORS AFFECTING RISK MODELLING

The risk for the lender is of several kinds ranging from disruption to cash flows, and increased collection costs to loss of interest and principal. That's why it is important to be able to forecast credit risk as accurately as possible. Credit risk modelling depends on a variety of complex factors.

That's why it is important to have sophisticated credit risk rating models. There are several major factors to consider while determining credit risk. From the financial health of the burrower and the creditor to a variety of macroeconomic considerations.

Here are three major factors affecting the credit risk of a borrower:

- **The Probability of Default (PD)**

  This refers to the likelihood that a borrower will default on their loans and is obviously the most important part of a credit risk model. For individuals, this score is based in their debt – income ratio and existing credit score. For institutions, that issue bonds, this probability is determined by rating agencies like Moody's and Standard & Poor's.

  The PD generally determines the interest rate and amount of down payment needed.

- **Loss Given Default**

  This refers to the total loss that the lender will suffer if the debt is not repaid. This is a critical component in credit risk modelling. For instance, two borrowers with the same credit score and a similar debt – income ratio will present two very different credit risk profiles if one is borrowing a much large amount.

  That's because the loss to the lender in case of default is much higher when the amount is larger. This again plays a big role in determining interest rates and down payments. If the borrower is willing to offer collateral, then that has a big impact on the interest rate offered.

- **Exposure at Default**

  This is a measure of the total exposure that a lender is exposed to at any given point of time. This also has an impact on the credit risk because it is an indicator of the risk appetite of the lender. It is calculated by multiplying each loan by a certain percentage depending on the particulars of the loan.

1.12 LOGISTIC REGRESSION

Logistic regression is a classification algorithm. It is used to predict a binary outcome based on a set of independent variables. A binary outcome is one where there are only two possible scenarios—either the event happens (1) or it does not happen (0).

Independent variables are those variables or factors which may influence the outcome (or dependent variable).
So: Logistic regression is the correct type of analysis to use when you're working with binary data. You know you're dealing with binary data when the output or dependent variable is dichotomous or categorical in nature; in other words, if it fits into one of two categories (such as "yes" or "no", "pass" or "fail", and so on).

1.13 WHAT IS LOGISTIC REGRESSION USED FOR?

Logistic regression is used to calculate the probability of a binary event occurring, and to deal with issues of classification. Logistic regression is used to predict the likelihood of all kinds of "yes" or "no" outcomes. By predicting such outcomes, logistic regression helps data analysts (and the companies they work for) to make informed decisions.

In the grand scheme of things, this helps to both minimize the risk of loss and to optimize spending in order to maximize profits. For example, it wouldn't make good business sense for a credit card company to issue a credit card to every single person who applies for one.

They need some kind of method or model to work out, or predict, whether or not a given customer will default on their payments. The two possible outcomes, "will default" or "will not default", comprise binary data—making this an ideal use-case for logistic regression.

Based on what category the customer falls into, the credit card company can quickly assess who might be a good candidate for a credit card and who might not be.

1.14 TYPES OF LOGISTIC REGRESSION

The three types of logistic regression are:

- **Binary logistic regression:**

  It is the statistical technique used to predict the relationship between the dependent variable (Y) and the independent variable (X), where the dependent variable is binary in nature. For example, the output can be Success/Failure, 0/1, True/False, or Yes/No.

- **Multinomial logistic regression**

  It is used when you have one categorical dependent variable with two or more unordered levels (i.e., two or more discrete outcomes). It is very similar to logistic regression except that here you can have more than two possible outcomes. For example, let's imagine that you want to predict what will be the most-used transportation type in the year 2030. The transport type will be the dependent variable, with possible outputs of train, bus, tram, and bike (for example).

- **Ordinal logistic regression**

  It is used when the dependent variable (Y) is ordered (i.e., ordinal). The dependent variable has a meaningful order and more than two categories or levels. Examples of such variables might be t-shirt size (XS/S/M/L/XL), answers on an opinion poll (Agree/Disagree/Neutral), or scores on a test (Poor/Average/Good).

1.15 ADVANTAGES OF LOGISTIC REGRESSION

- Logistic regression is much easier to implement than other methods, especially in the context of machine learning. A machine learning model can be described as a mathematical depiction of a real-world process. The process of setting up a machine learning model requires training and testing the model.

Training is the process of finding patterns in the input data, so that the model can map a particular input (say, an image) to some kind of output, like a label. Logistic regression is easier to train and implement as compared to other methods.

- Logistic regression works well for cases where the dataset is linearly separable. A dataset is said to be linearly separable if it is possible to draw a straight line that can separate the two classes of data from each other. Logistic regression is used when your Y variable can take only two values, and if the data is linearly separable, it is more efficient to classify it into two separate classes.

- Logistic regression provides useful insights. Logistic regression not only gives a measure of how relevant an independent variable is (i.e., the coefficient size), but also tells us about the direction of the relationship (positive or negative). Two variables are said to have a positive association when an increase in the value of one variable also increases the value of the other variable.

1.16 DISADVANTAGES OF LOGISTIC REGRESSION

- Logistic regression fails to predict a continuous outcome. Logistic regression only works when the dependent or outcome variable is dichotomous.
- Logistic regression assumes linearity between the predicted (dependent) variable and the predictor (independent) variables. While linearly separable data is the assumption for logistic regression, in reality, it's not always truly possible.
- Logistic regression may not be accurate if the sample size is too small. If the sample size is on the small side, the model produced by logistic regression is based on a smaller number of actual observations. This can result in overfitting. In statistics, overfitting is a modelling error which occurs when the model is too closely fit to a limited set of data because of a lack of training data. Or, in other words, there is not enough input data available for the model to find patterns in it. In this case, the model is not able to accurately predict the outcomes of a new or future dataset.

1.17 RISK SPECTRUM

Based on the scale of operation, banks and financial institutions are placed differently on the risk spectrum and have different business strategies, as we can see here.

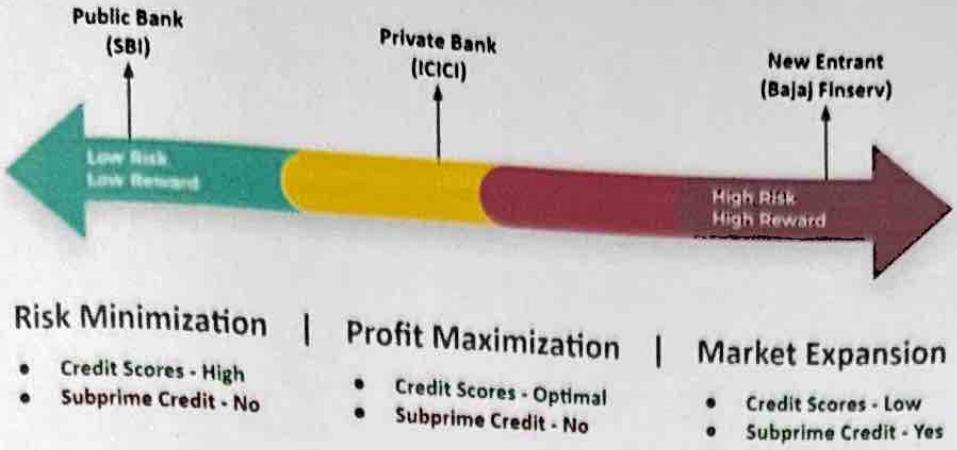


Fig. 1.17 Risk Spectrum

- **Public bank:**
  A public bank, which comes to the left of the spectrum, would always want to minimise their business risk, by considering only high credit score loan applicants for lending

- **Private Bank:**
  A private bank would want to maximise their profit, which requires them to identify their optimal credit score tolerances

- **New Entrant:**
  A new entrant that comes to the right of the risk spectrum, however, would have no other choice to but to consider every applicant that comes to their door, be it the one having low credit score or the one having no credit score at all.

Thus, every business in this risk spectrum would have a unique approach to assess loan applications and make lending decision based on their business strategy

CHAPTER – 2

LITERATURE SURVEY

LITERATURE SURVEY

Bensic M. et al. in 2005 have extracted important features for credit scoring in small – business lending on a dataset with specific transitional economic conditions using a relatively small dataset. They compared the accuracy of the best models extracted by different methodologies, such as logistic regression, neural networks (NNs), and CART decision trees. Four different NN algorithms are tested, including backpropagation, radial basis function network, probabilistic and learning vector quantization, by using the forward nonlinear variable selection strategy. Although the test of differences in proportion and McNemar's test do not show a statistically significant difference in the models tested, the probabilistic NN model produces the highest hit rate and the lowest type I error. According to the measures of association, the best NN model also shows the highest degree of association with the data, and it yields the lowest total relative cost of misclassification for all scenarios examined. The best model extracts a set of important features for small-business credit scoring for the observed sample, emphasizing credit programme characteristics, as well as entrepreneur's personal and business characteristics as the most important ones. [1]

Adams N. et al. in 2019 demonstrated the class imbalance problem and suggested a relabelling solution based on clustering the minority class. The class imbalance problem arises in two-class classification problems, when the less frequent (minority) class is observed much less than the majority class. This characteristic is endemic in many problems such as modelling default or fraud detection. Recent work by Owen has shown that, in a theoretical context related to infinite imbalance, logistic regression behaves in such a way that all data in the rare class can be replaced by their mean vector to achieve the same coefficient estimates. They have built on Owen's results to show the phenomenon remains true for both weighted and penalized likelihood methods. Such results suggest that problems may occur if there is structure within the rare class that is not captured by the mean vector. In a simulation and a real mortgage dataset, they have shown that logistic regression is not able to provide the best out-of-sample predictive performance and that an approach that is able to model underlying structure in the minority class is often superior. [2]

Sukono et al. in 2014 have done the analysis of credit scoring using logistic regression model, which is estimated using genetic algorithms. One of the Cooperative of Financial Services is disbursed loans to debtors (members and prospective members). In lending (provision of credit) is likely to arise the problem, namely the possibility of debt default by the debtor. To anticipate the risk of default (credit risk), to prospective debtors applying for credit risk analysis was performed using credit scoring. As a numerical illustration, the method used to analyse the credit scoring on a cooperative of financial services in Indonesia. Of the eight factors were analysed, it was only six factors that significantly influence to the risk of default. Six of these factors include:

number of dependents, the amount of savings, the value of collateral, monthly income, credit limit is realized, and the loan repayment period. [3]

Soureshjani M. et al. in 2012, have used two high-usage methods used for credit scoring (Credit scoring is a method used to estimate the probability of default or becoming delinquent of a loan applicant or existing borrower.), such as traditional statistics models like probit and logistic regression, data mining approaches and also artificial intelligence algorithms, on real data of legal customers of a commercial and also compared their performance. They found that logistic regression as a statistic model can estimate a good econometrics model which is able to calculate the probability of defaulting, and also neural networks is a very high-performance black box method which can be used in credit scoring problems. Also, the best cut off point in both logistic regression and neural network is calculated by these methods which have minimum errors on the available data. [4]

Beasens B. et al. in 2005 have applied a gradual approach that balances the interpretability and predictability requirements to rate banks. The Basel II capital accord encourages banks to develop internal rating models that are financially intuitive, easily interpretable and optimally predictive for default. Standard linear logistic models are very easily readable but have limited model flexibility. Advanced neural network and support vector machine models (SVMs) are less straightforward to interpret but can capture more complex multivariate non-linear relations. First, a linear model is estimated; it is then improved by identifying univariate non-linear ratio transformations that emphasize distressed conditions; and finally, SVMs are added to capture remaining multivariate non-linear relations. [5]

Wang H. et al. in 2015 have considered the plausibility of ensemble learning using regularized logistic regression as the base classifier to deal with credit scoring problems. Various ensemble learning methods with different base classifiers have been proposed for credit scoring problems. However, for various reasons, there has been little research using logistic regression as the base classifier. In this research, the data is first balanced and diversified by clustering and bagging algorithms. Then a Lasso-logistic regression learning ensemble is applied to evaluate the credit risks. They have shown that the proposed algorithm outperforms popular credit scoring models such as decision tree, Lasso-logistic regression and random forests in terms of AUC and F-measure. They have also provided two importance measures for the proposed model to identify important variables in the data. [6]

CHAPTER – 3
PROBLEM IDENTIFICATION

PROBLEM IDENTIFICATION

The problems identified in the above research papers are as follows:

- **No- Hits**
  Credit scores are based upon the historical data of the consumer or business venture. If any new consumer comes to apply for a loan who did not have any historical data then the evaluation of the credit score for that particular person is not possible. Hence the consumer remains and unscorable and the financial institutes might have problem lending him the money.

- **Short Term Outlook:**
  The financial institutes use a bureau record to handle the scorecards of the consumers while evaluating the credit score. These bureau scores have a limited time period. So the banks can not use them or refer to them always.

- **Score Boosting:**
  Another issue around credit scores revolves around credit bureaus trying to be more transparent by disclosing how they calculate their scores. If the bureaus reveal how they evaluate the credit scores then the consumer may use it in a bad way for their benefit which will be a loss for the banks and financial institutes.

CHAPTER – 4
METHODOLOGY

METHODOLOGY

The workflow of the processes involved in this project is demonstrated below:

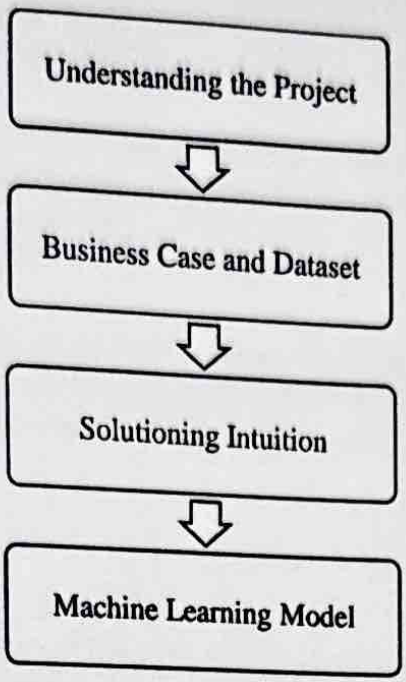


Fig. 4 Flowchart of methodology

4.1 UNDERSTANDING THE PROJECT

In this project we have assumed an ABC Bank Limited, which is a private bank. This bank intends to build an in – house risk model to make data driven lending decisions for subprime mortgages. Subprime mortgages are a type of loan granted to individuals with poor credit scores.

ABC Bank Limited wants to maximize profit, with an eye on market expansion as well. This implies that this ABC bank limited lies between the private bank and the new entrant banks in the risk spectrum. In order to achieve its objective of maximum profit and market expansion, the bank should provide loans to the consumers having optimal to low credit scores.

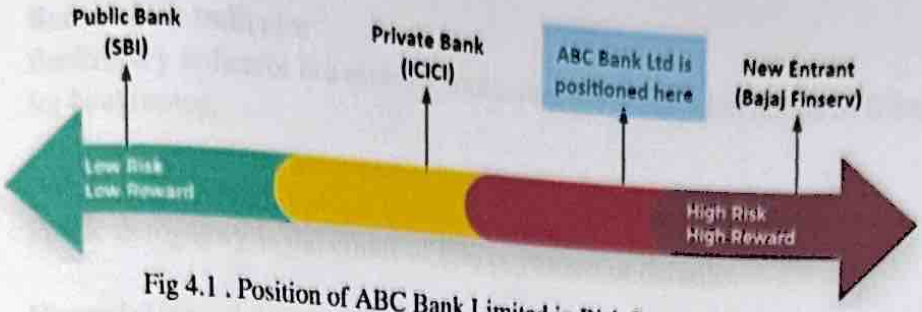


Fig 4.1 . Position of ABC Bank Limited in Risk Spectrum

We also have details on profit business books on a good loan and loss incurred on a bad loan i.e., which loan can be considered as the good or bad. The profit from a good customer can be considered as ₹100. Similarly, a loss from a bad customer can be considered as ₹500.

Our client has also shared historical customer data with us, that has the details on borrower's credit bureau records, captured at the time of loan application and final outcomes on these loans, viz., loan turned good or bad. Vaguely this is how our dataset is structures.

4.2 DATASET

There are 30 variables and 3000 observations in our dataset. The descriptions on these 30 variables are given in the following table:[9]

Table. 4.2 Variables of the Dataset

Variable Name	Label	Role
Target	Target = 1 (Defaulters), Target = 0 (Good Loans)	Target
BanruptcyInd	Bankruptcy Indicator	Input
TLBadDerogCnt	Bad Dept plus Public Derogatories	Input
CollectCnt	Collections	Input
InqFinanceCnt24	Finance Inquires 24 Months	Input
InqCnt06	Inquiries 6 Months	Input
DerogCnt	Number Public Derogatories	Input
TLDel3060Cnt24	Number Trade Lines 30 or 60 Days 24 Months	Input
TL50Utilent	Number Trade Lines 50 pct Utilized	Input
TDel60Cnt24	Number Trade Lines 60 Days or Worse 24 Months	Input
TDe160CntAll	Number Trade Lines 60 Days or Worse Ever	Input
T75UtilCnt	Number Trade Unes 75 pct Utilized	Input
Tel90Cnt24	Number Trade Unes 90+ 24 Months	Input
TLBadCnt24	Number Trade Unes Bad Debt 24 Months	Input
TLDel60Cnt	Number Trade Unes Currently 60 Days or Worse	Input

Variable Name	Label	Role
TLSatCnt	Number Trade Lines Currently Satisfactory	Input
TLCnt12	Number Trade Lines Opened 12 Months	Input
TLCnt24	Number Trade Unes Opened 24 Months	Input
TLCnt03	Number Trade Unes Opened 3 Months	Input
TSatPct	Percent Satisfactory to Total Trade Lines	Input
TLBalHCPct	Percent Trade Line Balance to High Credit	Input
TLOpenPct	Percent Trade Lines Open	Input
TLOpen24Pct	Percent Trade Unes Open 24 Months	Input
TLTimeFirst	Time Since First Trade Une	Input
InqTimeLast	Time Since Last Inquiry	Input
TLTimeLast	Time Since Last Trade Line	Input
TLSum	Total Balance All Trade Lines	Input
TLMaxSum	Total High Credit All Trade Lines	Input
TL_Cnt	Total Open Trade Unes	Input
ID	Customer ID	ID

Variable "Target" is our dependent variable. It is a binary variable and has 0's for good loans and 1's for bad loans. So, this is the variable we want our credit scoring model to predict.

We have the independent variable captured by the credit bureau at these historic loan application filings. The details on some of the independent variable which are highlighted are mentioned below:

1. **Bankruptcy indicator**
   Bankruptcy Indicator is a number indicating borrower's likelihood of filing for bankruptcy.

2. **Public derogatory**
   Public derogatory is the count of late payments or defaults.

3. **Financial Enquiries**
   Financial enquiries are the count of credit enquiries made in last few months

4. **Trade Line**
   Trade line is a number indicating the number of credits accounts a borrower has, be it: loan, credit card, or other debt obligations. So, basically if you have a credit card and a home loan at a certain point in time, you have 2 trade lines open

5. **Customer ID**
   Customer ID, is a unique identifier for every applicant in database. It would not have any impact on loan turning good or bad. So, we would not include this variable into our analysis.

4.3 SOLUTIONING INTUITIONS

At a high level, our solution architecture can be explained using the following diagram.

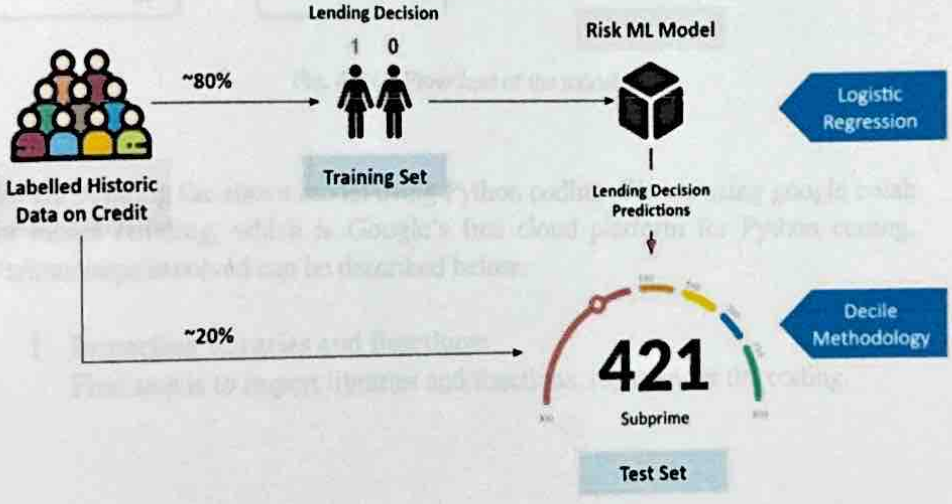


Fig. 4.3 Solution Architecture

Using 80% of our labelled dataset of historic loan applications, we would train a logistic regression risk model. And then we would use this Machine Learning

model to predict the likelihood of loan repayment for the remaining 20% applications.

Towards the end, we would use decile methodology to identify the business rules for accepting or rejecting any new application, targeting business profitability, as well as, penetration.

4.4 MACHINE LEARNING MODEL

The flow of the processes involved in this credit risk model can be described using the following flowchart.

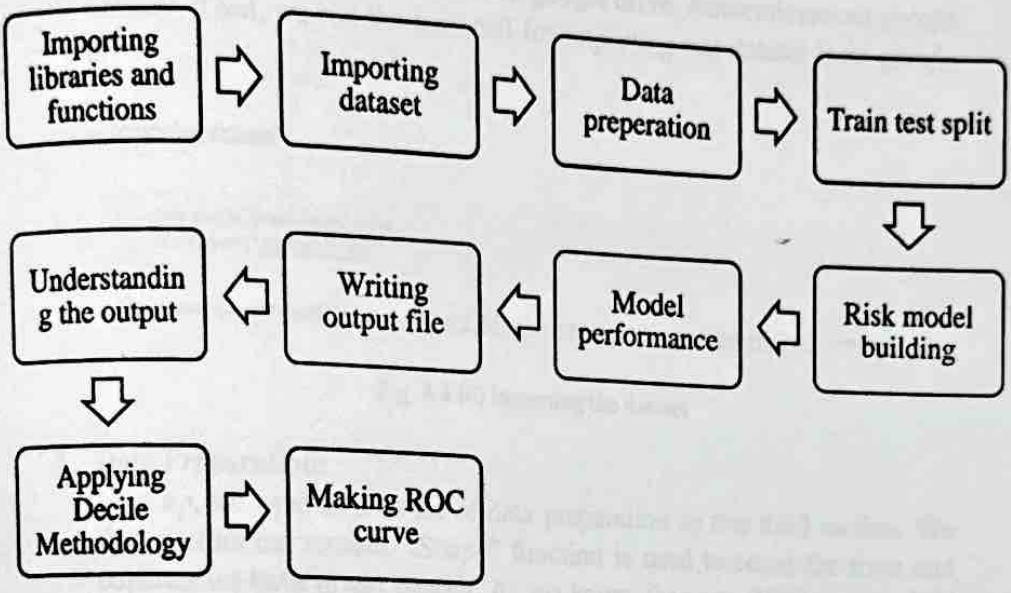


Fig. 4.4 (a) Flowchart of the model

We are building the above model using Python coding. We are using google colab for model building, which is Google's free cloud platform for Python coding. Various steps involved can be described below:

1. **Importing libraries and functions:**
   First step is to import libraries and functions, required for the coding.

Importing libraries & functions

```
[ ] import pandas as pd
    import numpy as np

    from sklearn.model_selection import train_test_split
    from sklearn.preprocessing import StandardScaler
    from sklearn.metrics import classification_report, confusion_matrix, accuracy_score
    from sklearn.linear_model import LogisticRegression
```

Fig. 4.4 (b) Importing library and functions

2. **Importing Dataset:**

Next up, we are importing our dataset. The dataset has been saved in our google drive. First, we need to mount google drive. Authenticate our google account. Then, we run the next cell for importing our dataset from google drive.

Importing dataset

```
[ ] from google.colab import drive
    drive.mount('/content/drive')

[ ] dataset=pd.read_excel("/content/drive/My Drive/Project1_Credit_Scoring/a_Dataset_Credit Scoring.xlsx")
```

Fig. 4.4 (c) Importing the dataset

3. **Data Preparation:**

Next up, we need to do a bit of data preparation in this third section. We first explore our dataset. "Shape" function is used to count the rows and columns we have in our dataset. As we know, there are 3000 rows and 30 columns.

```
[4] # shows count of rows and columns
    dataset.shape

    (3000, 30)
```

```
    #shows first few rows of the code
    dataset.head()
```

	TARGET	ID	BorogCnt	CollectCnt	BanruptcyInd	InqCnt06	InqTimeLast	InqFinanceCnt24	TLTimeFirst	TLTimeLast	TLCnt03	TLCnt12
0	0	85	1	1	0	7	1.0	4	125	3	1	3
1	0	116	1	1	0	2	1.0	0	262	18	0	0
2	0	124	0	0	0	1	1.0	4	254	12	0	1
3	0	128	0	0	0	6	3.0	6	154	3	1	0
4	0	143	0	0	0	1	0.0	1	311	17	0	0

Fig. 4.4 (d) Exploring the dataset

Next, we have this "head" function, which would show us the first few rows of our dataset. First column here is our target variable. Then we have customer ID as our column 2 and then have our 28 independent variables.

As we discussed previously, we would drop customer ID from our analysis. We will do this using "drop" function. We may validate the revised count of columns using the "shape" function, again. Columns have reduced to 29 now.

```
#dropping customer ID column from the dataset
dataset=dataset.drop('ID',axis=1)
dataset.shape
```

```
(3000, 29)
```

Fig. 4.4 (e) Dropping Customer ID from Dataset

Next up we check for missing values using "isna" and "sum" function, which return a count of na or not available fields, for all variables. We can see that we have missing values in our dataset. We would fill these missing values using "fillna" function with the mean value of each variable.

```
# explore missing values
dataset.isna().sum()
```

```
TARGET              0
DerogCnt            0
CollectCnt          0
BanruptcyInd        0
InqCnt06            0
InqTimeLast       188
InqFinanceCnt24     0
TLTimeFirst         0
TLTimeLast          0
TLCnt03             0
TLCnt12             0
TLCnt24             0
TLCnt               3
TLSum              40
TLMaxSum           40
TLSatCnt            4
```

✓ 0s completed at 23:19

Fig. 4.4 (f) Exploring missing values

Again, we check for the missing values jus to be sure.

```
[8]  # filling missing values with mean
     dataset=dataset.fillna(dataset.mean())

     # explore missing values post missing value fix
     dataset.isna().sum()

TARGET              0
DerogCnt            0
CollectCnt          0
BanruptcyInd        0
InqCnt06            0
InqTimeLast         0
InqFinanceCnt24     0
TLTimeFirst         0
TLTimeLast          0
TLCnt03             0
TLCnt12             0
TLCnt24             0
TLCnt               0
TLSum               0
TLMaxSum            0
TLSatCnt            0
TLDel60Cnt          0
TLBadCnt24          0
TL75UtilCnt         0
TL50UtilCnt         0

✓ 0s  completed at 23:20
```

Fig. 4.4 (g) Filling missing values with mean

4. **Train Test Split:**
   Moving on, in this next step, we are splitting our dataset into 80% training and 20% test. Thus, we would have 80% of 3000, viz, 2400 observations going into training the model and the rest 20%, viz, 600, going into testing model performance.

   Next up, we are doing data normalization. This is a data science practise where we scale all our independent variables between 0 and 1.

```
y = dataset.iloc[:, 0].values
X = dataset.iloc[:, 1:28].values

# splitting dataset into training and test (in ratio 80:20)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)

sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)
```

Fig. 4.4 (h) Splitting dataset

5. **Risk Model Building:**
Now, we come to the model building part. We are using logistic regression as our classifier.

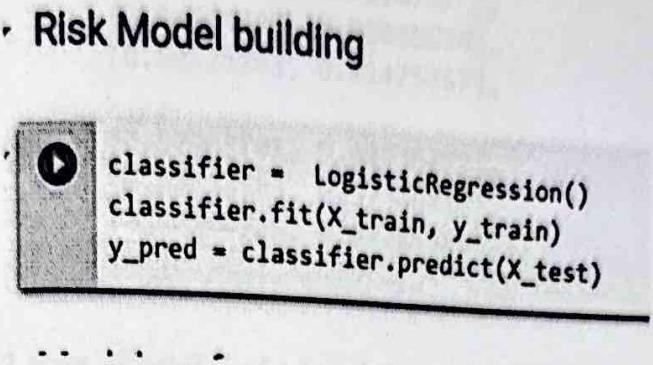


Fig. 4.4 (i) Risk model building

6. **Model Performance:**
In this next step, we analyse our model performance and results. First, we have the confusion matrix. Over here, we check how efficient our model is in classifying good loans and bad loans as bad. Moving on, accuracy score gives us the percentage of correct classifications.

Higher this accuracy value, better the model.

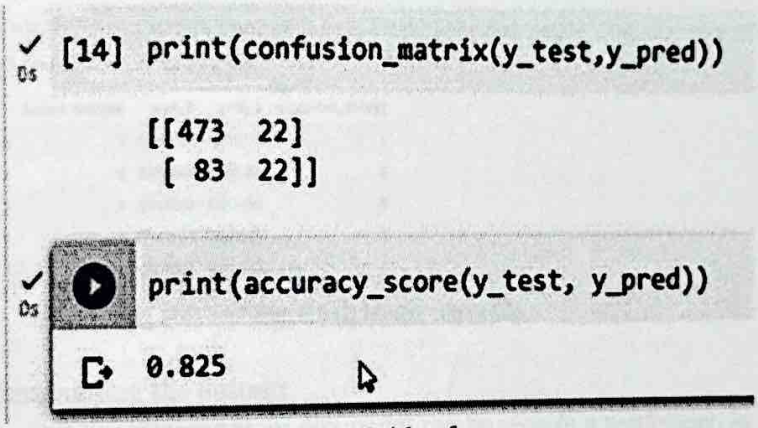


Fig. 4.4 (j) Model performance

7. **Writing Output File:**
At last, we are capturing probabilities for our predicted Y's using "predict_proba" function. It returns an array. The first column provides the probability corresponding to Target = 0, viz, probability of a loan being good.

```
predictions = classifier.predict_proba(X_test)
predictions
```

```
array([[0.2879243 , 0.7120757 ],
       [0.96954426, 0.03045574],
       [0.98524243, 0.01475757],
       ...,
       [0.57963701, 0.42036299],
       [0.62633583, 0.37366417],
       [0.92284615, 0.07715385]])
```

Fig. 4.4 (k) Probability of predicted Y

Column 2 gives probability of a loan being bad, and is nothing but (1 – probability of a loan being good). Next, we write our output file back to google drive, which should have model probabilities, along with predicted Y's and actual Y's.

Using "head" function, we can have a look at the output file.

```
e writing model output file

df_prediction_prob = pd.DataFrame(predictions, columns = ['prob_0', 'prob_1'])
df_prediction_target = pd.DataFrame(classifier.predict(X_test), columns = ['predicted_TARGET'])
df_test_dataset = pd.DataFrame(y_test,columns= ['Actual Outcome'])

dfx=pd.concat([df_test_dataset, df_prediction_prob, df_prediction_target], axis=1)

dfx.to_csv("/content/drive/My Drive/Project1 Credit Scoring/c1 Model Prediction.xlsx", sep=',', encoding='UTF-8')

dfx.head()
```

	Actual Outcome	prob_0	prob_1	predicted_TARGET
0	1	0.287924	0.712076	1
1	0	0.969544	0.030456	0
2	0	0.985242	0.014758	0
3	0	0.990957	0.009043	0
4	0	0.678374	0.321626	0

Fig. 4.4 (l) Model output file

8. **Understanding the output:**
The result we obtained from our model can be seen in a worksheet. In this worksheet, there are 6000 observations. As we have split our dataset into 80% training and 20% test, so, 20% of 3000 is 600 observations and that's exactly what we get.

Moving on to columns, we have five columns in this file. First one is serial number, starting from 0, thanks to python indexing. Next up, we have actual Y's from our test set. Here, zeros mean historical loans that turned out to be good.

Table 4.4.1 Result obtained after the coding

S No	Actual Outcome	Probability_Good	Probability_Bad	predicted_TARGET
0	1	28.63%	71.37%	1
1	0	96.90%	3.10%	0
2	0	98.53%	1.47%	0
3	0	99.11%	0.89%	0
4	0	67.66%	32.34%	0
5	0	86.42%	13.58%	0
6	0	93.02%	6.98%	0
7	0	92.20%	7.80%	0
8	0	54.48%	45.52%	0
9	0	92.67%	7.33%	0
10	0	62.88%	37.12%	0
11	0	94.34%	5.66%	0
12	0	65.50%	34.50%	0
13	0	94.45%	5.55%	0
14	0	42.42%	57.58%	0
15	0	74.59%	25.41%	1
16	1	90.10%	9.90%	0
17	1	58.70%	41.30%	0
18	0	96.64%	3.36%	0
19	0	92.62%	7.38%	0
20	0	80.92%	19.08%	0
21	1	17.80%	82.20%	1
22	0	93.84%	6.16%	0

And 1's mean historical loans that turned out to be bad. Columns C and D are model predicted probabilities for good and bad loans. Column E is model predicted Y. logistic regression classifier consider 50% as the default probability for marking predicted Y as 0 and 1.

Thus, predicted Y is 1 for probability of good less than 50% and is 0, when probability of good is over 50%. This predicted Y would not help us much, because we do not want to keep the probability threshold at 50%. Rather we want to find a custom threshold value for the bank on the basis of their business needs.

So to do that we are sorting our data in descending order of probability of good loans. By doing this we have these loans sequenced in descending order of their loan repayment likelihood.

Table 4.4.2 Result in Descending order

Probability_Good

A	B	C	D	E
S No	Actual Outcome	Probability_Good	Probability_Bad	predicted_TARGET
251	0	99.14%	0.86%	0
3	0	99.11%	0.89%	0
101	0	99.06%	0.94%	0
105	0	98.92%	1.08%	0
314	0	98.92%	1.08%	0
180	0	98.71%	1.29%	0
468	0	98.70%	1.30%	0
522	0	98.61%	1.39%	0
410	0	98.60%	1.40%	0
339	0	98.57%	1.43%	0
2	0	98.53%	1.47%	0
455	0	98.52%	1.48%	0
286	0	98.50%	1.50%	0
269	0	98.41%	1.59%	0
94	0	98.37%	1.63%	0
591	0	98.36%	1.64%	0
170	0	98.32%	1.68%	0
274	0	98.32%	1.68%	0
452	0	98.31%	1.69%	0
368	0	98.25%	1.75%	0
260	0	98.23%	1.77%	0
83	0	98.21%	1.79%	0
386	0	98.15%	1.85%	0

9. **Applying Decile Methodology:**

So, definitely as a bank, we want to approve loans for these initial few customers. But to figure out where to stop approving loans, we need to aggregate our data set first, using a statistical technique called decile. Using decile, we would divide this data, already sequenced in descending order of probability of good, into ten equal parts.

Table 4.4.3 Applying decile

S No

A	B	C	D	E	F
S No	Actual Outcome	Probability_Good	Probability_Bad	Predicted Outcome	Decile
251	0	99.14%	0.86%	0	1
3	0	99.11%	0.89%	0	1
101	0	99.06%	0.94%	0	1
105	0	98.92%	1.08%	0	1
314	0	98.92%	1.08%	0	1
180	0	98.71%	1.29%	0	1
468	0	98.70%	1.30%	0	1
522	0	98.61%	1.39%	0	1
410	0	98.60%	1.40%	0	1
339	0	98.57%	1.43%	0	1
2	0	98.53%	1.47%	0	1
455	0	98.52%	1.48%	0	1
286	0	98.50%	1.50%	0	1
269	0	98.41%	1.59%	0	1
94	0	98.37%	1.63%	0	1
591	0	98.36%	1.64%	0	1
170	0	98.32%	1.68%	0	1
274	0	98.32%	1.68%	0	1
452	0	98.31%	1.69%	0	1
368	0	98.25%	1.75%	0	1
260	0	98.23%	1.77%	0	1
83	0	98.21%	1.79%	0	1
386	0	98.15%	1.85%	0	1

So, essentially , we need to put 1 against the first 60 observations here as we have a total of 600. And then we use a simple formula to apply the decile methodology. As a result, we have a new file with 10 deciles in 10 rows.

Table 4.4.4 Deciled file

Decile	Count of Decile	Sum of Actual Outcome	MIN of Probability_Good	Good	Cumm. Good	Cumm. Bad	Cumm. Good %	Cumm. Bad %	Cumm. Bad Avoided %
1	60	1	97.15%	59	59	1	12%	1%	99%
2	60	1	95.85%	59	118	2	24%	2%	98%
3	60	5	93.84%	55	173	7	35%	7%	93%
4	60	6	92.20%	54	227	13	46%	12%	88%
5	60	6	89.05%	54	281	19	57%	18%	82%
6	60	8	86.47%	52	333	27	67%	26%	74%
7	60	13	79.73%	47	380	40	77%	38%	62%
8	60	17	72.45%	43	423	57	85%	54%	46%
9	60	18	55.98%	42	465	75	94%	71%	29%
10	60	30	9.73%	30	495	105	100%	100%	0%
Grand Total	600	105	9.73%						

Against this we have count of observations in each decile, which is 60 for all sets. And then we have sum of actual outcomes. As you know the actual outcome is zero for a good loan and 1 for a bad loan. So, the summation of actual outcome essentially gives us the count of bad loans.

Quite naturally, for initial few deciles sets, as the probability of good is almost approaching 100%, actual Y's would mostly be 0, so, there are no bad loans. And as we go down the decile set, we have more bad loans i.e., 1's.

Column D here has minimum of probability good for each decile set. So, for decile 1, if we look at our data, we have minimum of probability good as 97.15%. similarly, we have the cut off probabilities for another decile sets as well.

Next, we have count of good loans, which is nothing but column B minus column C. cumulative good and cumulative bad have cumulation of column E and column C, respectively. Then, the column H and column I, have percentages of these cumulative good or bad values.

And then we have a percentage of cumulative bad avoided at each level, which is 1 minus column I, basically. For decile 1, practically speaking, we have our best customers here. And if a business wants to be extremely conservative with their loan approvals, they may take decile 1 minimum probability or the cut – off probability as for their approval probability threshold for loans.

This way, business is funding loans for 12 % of total good customers that come to them and avoiding 99% bad customers, as well. Of course, we are avoiding 88% good customers as well. As we go down the decile sets, business is getting more good customers, but at the cost of getting exposure to more and more bad customers.

10. Making ROC Curve:

By the way, in statistics, cumulative good percentage is nothing but model sensitivity. And percentage of bad avoided is called specificity. When these two are plotted on a 2 - dimensional chart, we get an ROC curve. Area under this Roc curve gives us model performance which we already know is over 80%.

Table 4.4.5 Sensitivity/Specificity

Decile	Count of Decile	Sum of Actual Outcome	Min of Probability_Good	Good	Comm. Good	Comm. Bad	Sensitivity	1 - Specificity	Specificity
							Comm. Good %	Comm. Bad %	Comm. Bad Avoided %
1	80	1	97.15%	59	59	1	12%	1%	99%
2	80	1	95.85%	59	118	2	24%	2%	98%
3	80	6	93.84%	55	173	7	35%	7%	93%
4	80	6	92.20%	54	227	13	46%	12%	88%
5	80	6	89.05%	54	281	19	57%	18%	82%
6	80	8	85.47%	52	333	27	67%	26%	74%
7	80	13	79.73%	47	380	40	77%	38%	62%
8	80	17	72.45%	43	423	57	85%	54%	46%
9	80	18	65.86%	42	465	75	94%	71%	29%
10	80	30	9.73%	30	495	105	100%	100%	0%
Grand Total	800	105	8.73%						

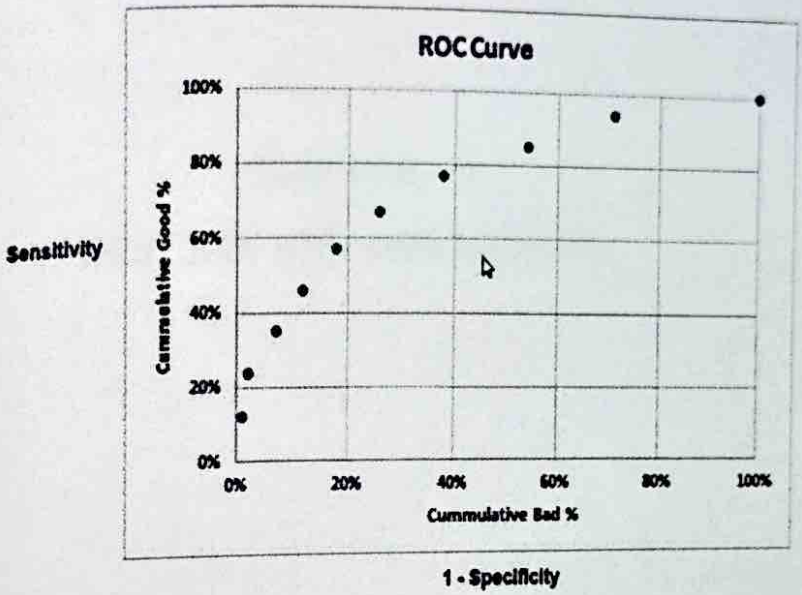


Fig. 4.4 (m) ROC Curve

CHAPTER – 5

RESULTS AND DISCUSSIONS

5.1 RESULTS

We have developed this risk model all this while for ABC Bank Limited. They wish to maximize their profits, while keeping an eye on market expansion, as part of their business strategy. We also know that they incur a loss of 500 ₹ for each bad loan and book a 100$ profit for each good loan.

We have created this model using the concept of logistic regression and decile methodology. The output we obtained at different stages have different applications. The first output which we get is after we apply the logistics regression through coding on the dataset we had.

The output obtained was in a tabular format with different columns and with all the 600 observation we had for analysis. The first column showed the serial number which starts from 0 because of the python coding. Then we have the actual outcome generated due to the code.

Table 5.1.1 Result showing predicted target

fx | S No

	A	B	C	D	E
	S No	Actual Outcome	Probability_Good	Probability_Bad	predicted_TARGET
0	0	1	28.63%	71.37%	1
1	1	0	96.90%	3.10%	0
2	2	0	98.53%	1.47%	0
3	3	0	99.11%	0.89%	0
4	4	0	67.66%	32.34%	0
5	5	0	86.42%	13.58%	0
6	6	0	93.02%	6.98%	0
7	7	0	92.20%	7.80%	0
8	8	0	54.48%	45.52%	0
9	9	0	92.67%	7.33%	0
10	10	0	62.88%	37.12%	0
11	11	0	94.34%	5.66%	0
12	12	0	65.50%	34.50%	0
13	13	0	94.45%	5.55%	0
14	14	0	42.42%	57.58%	1
15	15	0	74.59%	25.41%	0
16	16	1	90.10%	9.90%	0
17	17	1	58.70%	41.30%	0
18	18	0	96.64%	3.36%	0
19	19	0	92.62%	7.38%	0
20	20	0	80.92%	19.08%	0
21	21	1	17.80%	82.20%	1
22	22	0	93.84%	6.16%	0

The actual outcome is based upon the probability of good loans and the probability of bad loans. If the probability of good loan is greater than 50% then it is considered

as Target 0 i.e., the consumer can be trusted for repayment and can be given loan to earn profit.

The second objective we had, was to provide the ABC Bank Limited an assessed and arranged dataset which will make the data – driven lending decision easy and faster. To apply such arrangement, we used decile methodology to group the 600 observations into 10 groups with 60 each. These 10 groups have their cumulative good and bad percentages, sum of actual outcomes, minimum probability of good in the group and also profit to business which all help in making the data driven decisions easily and saves time.

Table 5.1.2 Result after delice methodology

Decile	Count of Decile	Sum of Actual Outcome	MIN of Probability_Good	Good	Cumm. Good	Cumm. Bad	Cumm. Good %	Cumm. Bad %	Cumm. Bad Avoided %	Profit to Business
1	60	1	97.15%	59	59	1	12%	1%	99%	5400
2	60	1	96.86%	59	118	2	24%	2%	98%	10800
3	60	5	93.84%	55	173	7	36%	7%	93%	13800
4	60	6	92.20%	54	227	13	46%	13%	88%	16200
5	60	6	88.06%	54	281	19	57%	19%	82%	18600
6	60	8	85.47%	52	333	27	67%	26%	74%	19800
7	60	13	79.73%	47	380	40	77%	39%	62%	18000
8	60	17	72.45%	43	423	57	86%	54%	46%	13800
9	60	18	55.98%	42	465	75	94%	71%	29%	9000
10	60	30	9.73%	30	495	105	100%	100%	0%	-3000
Grand Total	600	105	9.73%							

5.2 DISCUSSIONS

Using this information, we can compute profitability across our deciles. At decile 6, business is maximizing on their profit. So, this is the place our bank would want to be if profitability is their only objective. The cut – off probability here is 85.47% at this level, which means business may accept only those loans applications that have good loan probability more than this.

However, as we are categorically told to keep an eye on market expansion as well, business may consider decile 7 as well, where they gain more market share at the cost of some of their potential profits. And the cut – off probability at this level reduces to 79.73%.

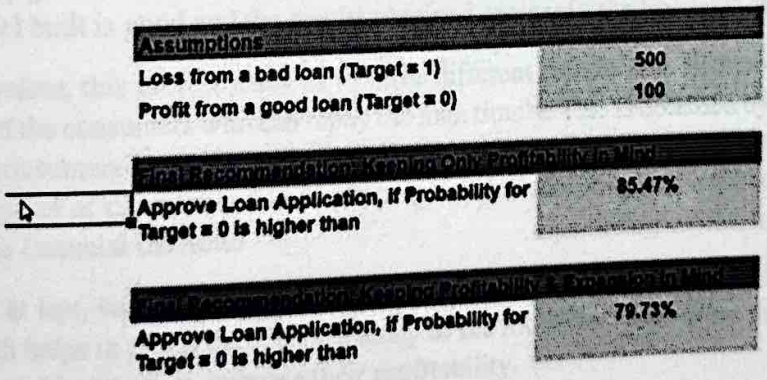


Fig. 5 (a) Final Recommendations

Our job is to give data driven insights to business for making better decisions. So, this is what we shall go back to our client with. Wherein we give them both these options and allow them to make the final call. The final call will be according to the financial institutes on the basis of their respective business strategies.

Along with this we also evaluated and drew the ROC curve using the above information. ROC curve is very useful in helping the banks and financial institutes to maximise their profit. The ROC curve obtained from the above result can be drawn as follows:

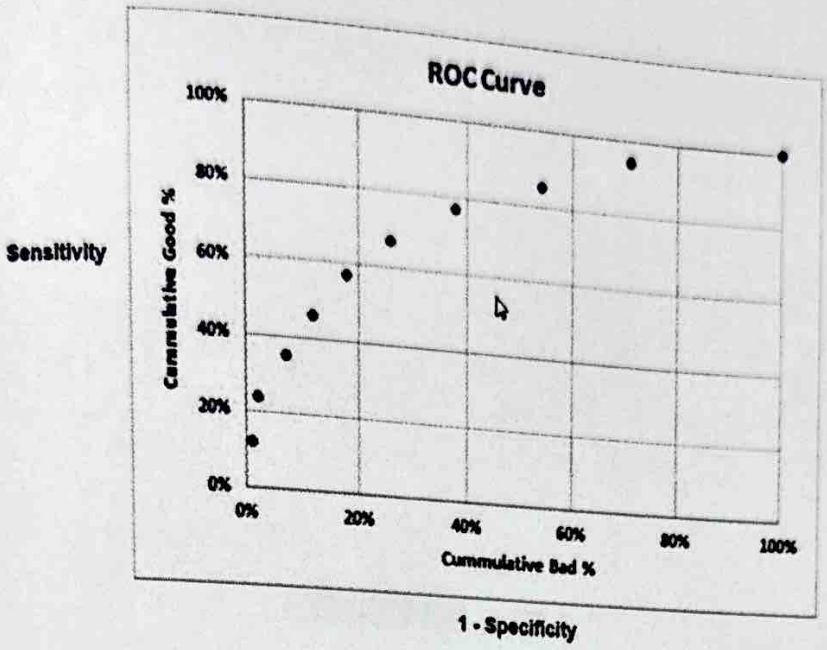


Fig. 5 (b) Output ROC curve obtained

The area under the ROC curve is very useful in evaluating predictive performance of the model. We already predicted the model performance above 80% which is pretty good. After the results, by observing the ROC curve, it is evident that the model built is good and the results obtained can profit the financial institutes.

Therefore, this project leads us to three different results. The first one tells us the list of the consumers who can repay the loan timely. This is obtained by considering the consumers having Target = 0. Next, we arranged the previous result to obtain the group of consumers which can be given loan to achieve the business strategy of the financial institutes.

And at last, we obtained the ROC curve between the sensitivity and specificity, which helps in predicting the efficiency of the model and also helps the banks and financial institutes to increase their profitability.

CHAPTER – 6

CONCLUSION AND FUTURE SCOPE

6.1 CONCLUSION

Our scope was to build a machine learning based in – house risk model, allowing business to make lending decisions for subprime mortgages. Based on the understanding, we knew that ABC Bank Limited is placed towards the right in the risk spectrum and their business objectives shall entail both: profitability and market expansion.

Now we tell them about the assumptions we made while building our model, including: missing value imputation with mean. And we used logistic regression as our classifier. And finally, as our deliverable we tell them that we have built an in – house risk model that gives 83% accuracy that is built with google colab's free online resources.

Fig. 6.1 (a) Details of built model

As our next deliverable, we give them these lending strategy options where they focus only on profitability or prioritise market expansion while compromising a bit on their profitability.

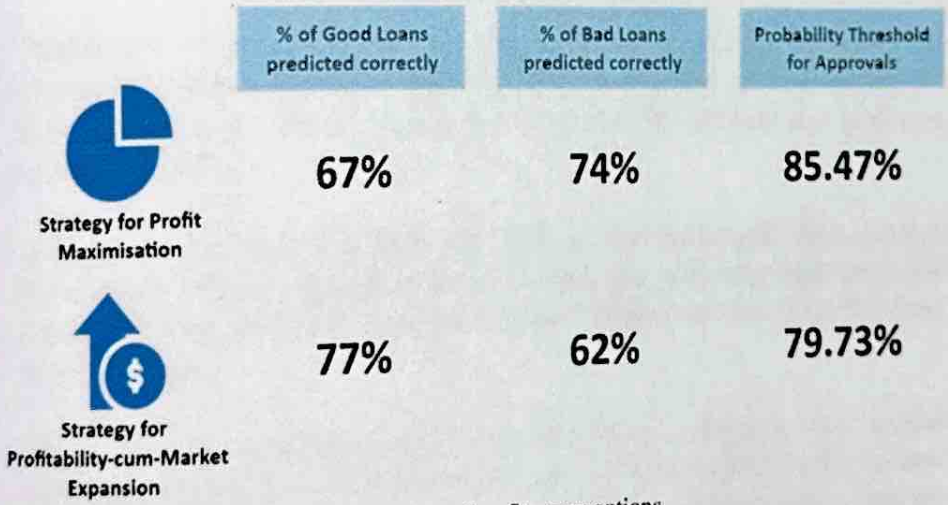


Fig. 6.1 (b) Lending Strategy options

In conclusion, we can conclude that credit risk model built will assess and arrange the data of consumers in a significant way and make the management of the consumer's data easy for the financial institutions. It successfully evaluates the probability of good or bad loans on the basis of which the banks will decide whether to give loan or not.

The demonstrated work creates an in – house risk model which make lending decisions on the basis of consumer's dataset. The ROC curve based on logistic regression and decile methodology on the details of the output will help the banks and financial institutes to maximize their profit.

It is a data – driven lending strategy which will make the decisions of the banks and financial institutes easier and faster and will help them to save a lot of time. It leads to significant reduction in misclassification costs to the benchmark logistic regression.

To sum up, we've analyzed and pre-processed our data, trained and evaluated our model, namely logistic regression, for their ability to predict loan defaults and their probability. We evaluated the models' performance at predicting probability of good loans.

We also used ROC curve that can help the financial institutions in great ways. We demonstrated how machine learning can be applied to the world of credit risk assessment. Machine learning is often seen as difficult to apply in banking due to the sheer amount of regulation the industry faces.

ML is successfully used in numerous, heavily regulated industries. Because of this innovative approach it is possible to increase the sustainability of the loans sector and make loans even more affordable to bank customers.

6.2 FUTURE SCOPE

Machine learning algorithms can detect anomalies in data. The algorithms can find potentially strange data entries that need more investigation. AI technologies can also be used for the automation of report generation, for example data quality or monitoring reports.

In this way, employees can spend more time on other meaningful tasks. Another development that we are seeing is that the data and modelling departments are working, and probably will continue to keep working more closely together than they did before.

Previously the modelling department prepared the data-sets they needed themselves, but now that the data quality plays a more important role, the data department has taken over. The data department has more knowledge about the data and can prepare a data-set more thoroughly and deliver a higher quality data-set.

Since this is a new field, there are not yet any established best practices regarding the use of AI in models. Machine learning algorithms can become part of PD (Probability of Default), LGD (Loss Given Default) or EAD (Exposure at Default) models, because they can find relationships that traditional methods cannot.

Initially, banks may start with small projects that investigate how AI can be used to improve credit risk models, independent of regular modelling process. Banks can use the same modelling steps, control framework, and role of model validation and hence treat AI as normal AIRB models.

This will also show supervisors that AI models are not just an experiment.
Another option is when banks develop AI models parallel to the current "normal" AIRB models and use them for insights instead of reporting purposes.

This way, banks will be able to show supervisors the value additions of their AI models over 'normal' AIRB models. Also, both banks and supervisors will gain more knowledge about AI models and warm up to their use.

This seems to be the most likely way for AI credit risk models to be used in practice. However, we know that the probability that AI models will not be taken into use any time soon, because banks simply do not have the time and resources to start developing them and or are busy enough with redeveloping their 'normal' AIRB models, as well as implementing new regulations like Basel IV.

The credit risk model landscape is heavily regulated. Supervisors are possibly not keen on allowing unknown modelling methods since they are apprehensive of the so-called 'black box' algorithms. I think that supervisors will start allowing more complicated AI models gradually, once both banks and supervisors have more experience with them.

REFERENCES

REFERENCES

[1] Agresti, A. (2003). Categorical data analysis. John Wiley & Sons.

[2] Davison, A.C. (2003). Statistical models. Cambridge University Press.

[3] Dobson, A. J. and A. Bearnett (2008). An introduction to generalized linear models. CRC press Hosmer, D. W. T.

[4] Hosmer, S. Le Cessie, S. Lemeshow, et al. (1997). A comparison of goodness-of-fit tests for the logistic regression model. Statistics in medicine 16 (9), 965-980

[5] E.L. Altman and G. Sabato, Modelling credit risk for smes: Evidence from the US market, *Abacus*, **43** (2007), 332-357.

[6] G. E. Batista, R. C. Prati and M. C. Monard, A study of the behavior of several methods for balancing machine learning training data, *ACM SIGKDD Explorations Newsletter*, **6** (2004), 20-29.

[7] C. Bravo, L. C. Thomas and R. Weber, improving credit scoring by differentiating defaulter behaviour, *Journal of the Operational Research Society*, **66** (2015), 771-781.

[8] Fig. 1.2 (a) https://images.app.goo.gl/URBUnurMVx9xHR8y5

[9] Table-4.2
https://docs.google.com/file/d/1jFI0hWZdBwwF5lS8dU60gqS73Nz_n0e3/edit?usp=docslist_api&filetype=msexcel

APPENDIX

CODE

```
### Importing libraries & functions
"""

import pandas as pd
import numpy as np

from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score
from sklearn.linear_model import LogisticRegression

"""### Importing dataset"""

from google.colab import drive
drive.mount('/content/drive')

dataset=pd.read_excel("/content/drive/My
Drive/Project1_Credit_Scoring/a_Dataset_CreditScoring.xlsx")

"""### Data preparation"""

# shows count of rows and columns
dataset.shape

#shows first few rows of the code
dataset.head()

#dropping customer ID column from the dataset
dataset=dataset.drop('ID',axis=1)
```

```
dataset.shape

# explore missing values
dataset.isna().sum()

# filling missing values with mean
dataset=dataset.fillna(dataset.mean())

# explore missing values post missing value fix
dataset.isna().sum()

# # count of good loans (0) and bad loans (1)
# dataset['TARGET'].value_counts()

# # data summary across 0 & 1
# dataset.groupby('TARGET').mean()

"""### Train Test Split"""

y = dataset.iloc[:, 0].values
X = dataset.iloc[:, 1:29].values

# splitting dataset into training and test (in ratio 80:20)

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)

sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)
```

```
# Exporting Normalisation Coefficients for later use in prediction

import joblib
joblib.dump(sc, '/content/drive/My
Drive/Project1_Credit_Scoring/f2_Normalisation_CreditScoring')

"""### Risk Model building"""


classifier =  LogisticRegression()
classifier.fit(X_train, y_train)
y_pred = classifier.predict(X_test)


# Exporting Logistic Regression Classifier for later use in prediction


# import joblib
joblib.dump(classifier, '/content/drive/My
Drive/Project1_Credit_Scoring/f1_Classifier_CreditScoring')


"""### Model *performance*"""


print(confusion_matrix(y_test,y_pred))


print(accuracy_score(y_test, y_pred))


"""### Writing output file"""


predictions = classifier.predict_proba(X_test)
predictions


# writing model output file
```