




Document Information

Analyzed document	Credit Risk Analysis Using Machine Learning.pdf (D144128715)
Submitted	9/15/2022 12:38:00 PM
Submitted by	Mr. P.Venkateswarulu
Submitter email	venkat.it@jntukucev.ac.in
Similarity	2%
Analysis address	venkat.it.jntuk@analysis.urkund.com

Sources included in the report

W	URL: https://www.slideshare.net/nikhileshMane/beproject-pressure-reducing-and-desuperheater-station Fetched: 4/25/2022 6:25:29 PM	 2
W	URL: https://su-plus.strathmore.edu/bitstream/handle/11071/6789/Consumer%20credit%20risk%20modelling%20using%20machine%20learning%20algorithms.pdf?sequence=3&isAllowed=y Fetched: 6/15/2021 8:15:58 AM	 2
W	URL: https://link.springer.com/article/10.1007/s44230-022-00004-0 Fetched: 5/12/2022 6:49:23 AM	 1

Entire Document

Credit Risk Analysis Using Machine Learning A
project report submitted

80%	MATCHING BLOCK 1/5	W
-----	--------------------	----------

in partial fulfilment of the requirements for the award of the degree of MASTER OF

COMPUTER APPLICATIONS

By
Jakka Kanaka Naveena
19VV1F0010

Under the Esteemed Guidance of Dr.B.Tirimula Rao, M.Tech,Ph.D Assistant Professor & HOD
Department of Information Technology
DEPARTMENT OF INFORMATION TECHNOLOGY JAWAHARLAL NEHRU TECHNOLOGICAL
UNIVERSITY GURAJADA
UNIVERSITY COLLEGE OF ENGINEERING VIZIANAGARAM - 535003, A.P. 2019-
2022
DEPARTMENT OF INFORMATION TECHNOLOGY
JNTUG - UNIVERSITY COLLEGE OF ENGINEERING VIZIANAGARAM CERTIFICATE

This is to certify that the dissertation report entitled "Credit Risk Analysis Using Machine Learning" submitted by Jakka Kanaka Naveena bearing registration number 19VV1F0010 in partial fulfillment for the award of the degree of Master of Computer Applications (MCA) from Jawaharlal Nehru Technological University Gurajada - University College of Engineering Vizianagaram. This bonafide work was carried out by him under my guidance and supervision during year 2019- 2022. The results embodied in this dissertation have not been submitted to any other University or Institute for the award of any degree or diploma.

Signature of Project Guide Signature Head of the Department B.TIRIMULA RAO B.TIRIMULA RAO
Assistant Professor&HOD Assistant Professor&HOD Dept.of Information Technology Dept.of Information Technology JNTUG - VIZIANAGARAM JNTUG - VIZIANAGARAM
(Signature of External Examiner)

DECLARATION
I, Jakka Kanaka Naveena (Reg. No: 19VV1F0010)

declare that this submission represents my ideas in my own words, and where others' ideas or words have been included, I have adequately cited and referenced the source. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal

Signature)

Jakka Kanaka Naveena

(19VV1F0010)

Date:

Place:

ACKNOWLEDGEMENT

This acknowledgment transcends the reality of formality when I express deep gratitude and respect to all those people behind the screen who inspired and helped us in the completion of this project work.

I take the privilege to express my heartfelt gratitude to my guide Dr. B.Tirimula Rao, Assistant Professor & HOD, Department of Information Technology, JNTUG University College of Engineering Vizianagaram for her valuable suggestions and constant motivation that greatly helped me in the successful completion of the dissertation. His wholehearted cooperation and the keen interest shown to him at all stages are beyond words of gratitude.

With great pleasure and privilege, I wish to express my heartfelt sense of gratitude and indebtedness to Dr.B.Tirimula Rao, Assistant Professor, Head of Department of Information Technology, JNTUG -UCEV Vizianagaram, for her supervision.

I extend heartfelt thanks to our principal Prof. R.Rajeswara Rao for providing intensive support throughout my dissertation.

I am also thankful to all the Teaching and Non-Teaching staff of the Information Technology Department, JNTUG University College of Engineering Vizianagaram, for their direct and indirect help provided to me in completing the dissertation.

I extend my thanks to my parents and friends for their help and encouragement in the success of my dissertation.

J.K Naveena

(19VV1F0019)

ABSTRACT

ABSTRACT

Credit scoring technology is often utilised in the risk evaluation of loan applications as a type of statistical model. Based on the data provided by the borrowers, particularly their past data and data from the financial system, it can estimate the credit risk of applicants. This study focuses on credit risk analysis and evaluation based on machine learning techniques such as the Logistic Regression algorithm, Decision Tree algorithm, and Random Forest algorithm, using data provided by a financial institution. AUC, KS, and F1 score, which show that the Random Forest algorithm, Decision Tree method, and Logistic Regression algorithm can all apply to financial risk analysis, were also found in the study's training and testing samples.

TABLE OF CONTENTS

TABLE OF CONTENTS Page.No

ABSTRACT (I)	LIST OF FIGURES (II).....	1.INTRODUCTION .(1-7)
.....	1.1.Introduction to Credit Risk Analysis (3-4)	1.1 1.Problem statement 3
.....	1.1.2.Objective 4	1.1.3.Scope Of the Project 4
1.1.4.		
Applications 4	1.2.Introduction to Machine Learning (5-8)	2.LITERATURE SURVEY (9-11)
.....	3.SOFTWARE AND HARDWARE REQUIREMENTS (12-13)	4.DATA SET (14-15)
.....	5.ARCHITECTURE (16-17)	6.DATA VISUALISATION (17-18)
.....	7.METHODOLOGY (18-36)	7.1.Data Preprocessing
(24-30)		
.....	7.1.1.Filling Data Using Imputation Methods 20	7.1 2.Removing Outliers 20
.....	7.1.3.One Hot Encoder 20	7.1.4.Label Encoder
21		
7.1.5.Standardization 21	7.2.Algorithms	
(24-30)	7.2.1..Logistic Regression (24-28)	
.....	7.2.2.KNN Classifier (29-30)	7.2.3.Random Forest (34-33)
.....	7.2.4.Naive Bayes (30-33)	7.2.5.XGBoost
(34-33)		
.....	8.IMPLEMENTATION (37-45)	9.RESULTS AND ANALYSIS (46-49)
.....		
10. CONCLUSION. (50-51)	11. BIBLIOGRAPHY	
(52-54)		

LIST OF FIGURES

S.NO Figure Name Page No
 1 Credit Risks and Types 3
 2 Work Flow Of Machine Learning 5
 3 Dataset and features 16
 4 Features of dataset 17
 5 Model Architecture 19
 6 Data Insights Dashboard 23
 7 Insights of Grade A Dashboard 24
 8 Insights of Grade B Dashboard 25
 9 Logistic Regression 32
 10 shows the before and after the KNN 35
 11 Random Forest Working 37
 12 GNB Normal distribution graph 40
 13 Attribute to Attribute Correlation 45
 14 Classification Report 46
 15 Confusion Matrix 49
 16 Output 52
 17 TPR and FPR Graph 53
 18 Resultant ROC Curve 54

| Page 1 CHAPTER-1 INTRODUCTION

| Page 2

INTRODUCTION

1.1.Introduction to credit risk Analysis:

When a borrower fails to meet their contractual loan commitments for any reason, the lender may suffer a loss. This is determined by performing a credit risk analysis. Interest for credit-risk assumption forms the revenues and benefits from such debt commitments and risks.

When the principal and interest due are not paid, it affects the financier's cash flow. Additionally, collections are becoming more expensive. The technique of intelligent credit analysis can help lessen the severity of the total loss of the borrowings and their recovery, even though it is difficult to predict who and when will default on loans. In the past, typical commercial banking systems were limited to determining whether a loan's credit risk was high or low based on a specified weighting of the borrowers' capital value, income, liabilities, use of the loan, and credit history. This outdated method of assessing credit risk would overestimate the credit risk of borrowers who take significant financial risks in the future and undervalue the credit risk of borrowers who have proven but unproven sources of income, lowering economic efficiency and security. Additionally, guarantee institutions may be able to assist these borrowers in their attempts to conceal high credit risk in order to increase their credit, something that conventional risk evaluation used by commercial banks is unable to reveal. Data mining and analysis are now possible thanks to the Internet's rapid expansion.

What is credit Risk:

The process of determining a company's or organisation's creditworthiness is called credit analysis. In other words, it is the assessment of a company's capacity to meet

| Page 3 its financial commitments. When a significant corporation offers or has issued bonds, the audited financial accounts may be examined.

If the borrower poses an acceptable level of default risk, the analyst may recommend approving the credit application at the agreed-upon terms. The results of the credit risk analysis affect the borrower's risk rating and their ability to obtain credit.

Fig1: credit Risks and types

1.1.1.Problem Statement:

typical commercial banking systems were limited to determining whether a loan's credit risk was high or low based on a specified weighting of the borrowers' capital value, income, liabilities, use of the loan, and credit history. This outdated method of assessing credit risk would overestimate the credit risk of borrowers who take significant financial risks in the future and undervalue the credit risk of borrowers who have proven but unproven sources of income, lowering economic efficiency and security. Additionally, guarantee institutions may be able to assist these borrowers in their attempts to conceal high credit risk in order to increase their credit, something that conventional risk evaluation used by commercial banks is unable to reveal. Data mining and analysis are now possible thanks to the Internet's rapid expansion 1.1.2.Objective:

The objective of the credit risk analysis are:

1.To analyse the credit risk of the borrower's.

| Page 4 1.1.3.Scope Of the Project

It considers a complete message about its organization rather than single words. It can be referred to as the intelligent approach due to its message examining criteria. It provides sensitivity to the client and adapts well to future spam techniques. Even if the spam word is slightly modified, this algorithm still succeeds and notices the spam content. Even though this framework is accessible for the URL groped embedded in an email, it lacks action taking for IT security groups. The degree of business in real-time is one more major drawback that is faced by this framework. It does not provide any clarity about how well it is performed in real-time for the spam campaigns. Spammers can easily develop techniques to meet the preventive measures of their to framework like making legitimate domains fall into the list of illegitimate emails the results that are obtained do not provide any aggregate view of the large groups of emails. Moreover, it does not let the network administrator have an online monitoring system across the network.

1.1.4.Applications:

- A technique used by credit analysts to assess a borrower's capacity to repay debts is called credit risk analysis.
- Credit analysis measures the lender's exposure to loss in order to assess the creditworthiness of potential borrowers.
- The likelihood of default, the loss in the event of default, and the exposure to default are the three variables that lenders use to calculate credit risk.

1.2.Introduction to Machine Learning

Machine Learning is a subset of artificial intelligence that has primarily been used to create algorithms that allow the computer to benefit from data and prior knowledge on its own. Arthur Samuel first demonstrated terminological machine learning in 1959. It can be described as follows.

Working of Machine Learning:

"The Machine Learning Model must learn from real data, establish prediction models, and predict outputs while we provide new data," says the author. The

| Page 5 amount of data influences the accuracy of the output because the majority of the data is used to build the best model for forecasting the new data point. Fig: 2

Work Flow Of Machine Learning

Classification Of Machine Learning:

Machine Learning can be categorized into three types:

- 1.Supervised learning
- 2.Unsupervised learning
- 3.Reinforcement learning

Supervised Learning:

Supervised machine learning is a process in which we receive labeled sample data for the machine learning scheme to train it and predict the outcome. The machine generates a model with fewer details so that datasets and studies each dataset. Once the preparation and processing are completed, we will have sample data to validate the model and see if it predicts the same performance. The goal of supervised learning is to convert input data into output data. Simple supervised learning is dependent on supervision and is analogous to a student learning something under the supervision of an instructor. The goal of supervised learning is to connect input and output data. Supervised learning is dependent on supervision.

Supervised machine learning is depicted graphically. Original preprocessed data sets containing known variables and targets are divided into training data and

| Page 6 test data in supervised learning. (Above) The training data are used to train a learning algorithm in an attempt to develop an accurate predictive model during the training phase. (Center) The test data are then applied to the model to validate it, and the predictive accuracy is evaluated. (Below) After the model has been validated, new data is fed into it in an attempt to make new predictions.

Advantages of Supervised learning:

- This type of learning is simple to grasp. It is the most widely used type of learning method. People should begin learning ML by practicing supervised learning.
- The training data is only required to train the model. Because of its size, it takes up a lot of room. However, it is erased from memory once training is completed because it is no longer relevant.
- We'd already know how many classes are in the data.
- After training, the model will determine which specific data needs to be predicted, as all of the data in the collection is unimportant.
- It is extremely useful in solving real-world computational problems.

Disadvantages of supervised learning:

- Its performance is limited by the fact that it cannot handle complex ML problems.
- It is unable to generate its labels. This means that, unlike unsupervised learning, it cannot discover data on its own.
- Any new data we enter must come from one of the given classes. If you enter watermelon data into a collection of apples and oranges, it may classify watermelon as one of these, which is incorrect.
- To train a supervised learning-based model, a good computer with quality processors is required. It necessitates a high level of computation power, which not all PCs may have.

Unsupervised Learning:

| Page 7 Unsupervised learning is a mechanism that allows a computer to learn without supervision. Because training data is presented to a computer with data that has not been labeled or classified the algorithm must function without any control over that data. Unsupervised learning attempts to reconstruct input data into new features or a collection of items with matching patterns.

Reinforcement Learning:

Strengthening learning is an interpretive-based learning process in which the learning assistant is rewarded for each correct action and fined for each incorrect action. The assistant learns from the feedback and improves its performance as a result. The assistant deals with and examines the world to strengthen learning. The goal of the assistant is to earn more reward points.

Features Of Machine Learning:

- Machine learning employs data to detect various patterns in a given dataset, and it can learn from previous data and improve itself automatically.
- It is a technology that is driven by data.
- Machine learning is similar to data mining in that it deals with massive amounts of data.

| Page 8 CHAPTER-2

LITERATURE SURVEY

| Page 9 LITERATURE SURVEY

Credit risk management was chosen as the independent variable for the study since it was thought that the robustness of such a system might change or be affected by different financial institutions. The research mainly uses bank performance as the main component to examine how changes to the credit risk management systems have affected it.

A deficient credit risk management system that is unable to effectively manage the default risk of a bank or lending business is indicated by the number of non-performing loans. In order to understand how external economic factors affect the link between credit and the economy, it is consequently adopted in the study as an intervening variable, while macroeconomic factors are included in the review as a moderating variable.

Danjuman, Ibrahim, Kola, Ibrahim Abdullateef, Magaji, Badiya Yusuf, Kumshe & Hauwamodu (2016) explained the credit risk management and customer satisfaction. It shows the positive relationship between credit risk management and customer satisfaction and there is no need for banks management to pay attention to other factors that contributes towards the customer satisfaction other than granting of credits. Bank needs to focus on its credit policy in order to make more profits.

Hameeda Abu, Hussain, Al Ajmi & Jasim (2012) examined the administration of risk practices followed by the ordinary banks and found that the risk levels confronted by banks are higher in case of traditional banks. Hence, nationwide, residual and settlement, operational, risks are seen to be higher if there must be an event to occur in traditional banks.

Abmed, Sufi Fizan, Malik & Qaisarali (2015) assessed the credit risk administration and advance execution of micro scale banks. The consequences of

| Page 10 the examination are demonstrating that there is a positive connection between the credit term and execution of advance. While, there is a positive connection between gathering approach and control of Credit risk however they are insignificantly affecting the advance.

Parsley & Mark (1996) found that credit and market risks alone cannot explain the earnings volatility they experience and against which they want to allocate capital.

Measuring operational risk will provide banks a way to price a new and lucrative source of business. Hence bank needs to concentrate more on controlling its operational risk in order to increase its source of business.

Meighs & Frank E (1995) Analysed by utilizing conventional credit instruments with regards to interest rate swaps which offers the credit officers to sufficiently deal with another source of credit risk. End clients of financing interest rate swaps can fundamentally decrease their credit chance by taking insurance. It acts as new instrument to deal with the risk required in loaning credit to the clients.

Gupta V K (1991) Reviewed that asset and liability administration has extended to incorporate into financing interest rate risk, cash chance, liquidity risk and operational risk. Modelling utilizing all the risk elements empowers investors to get ready for any instabilities. It is required for every banks to set their fixation towards the credit displaying which permits the broker to decrease and face any eventuality.

Weber, Olaf, Fenchel, Mareus, Scholz & Roland W (2008) examined the reconciliation of natural risks into credit risks administration methods and techniques of banks and finding the huge contrasts in incorporating environmental risks between banks that are signatories of UNEP proclamation by the banks on the earth and sustainable advancement in coordinating environmental risks and banks that had not consented to this arrangement so far could be found.

Jobs, Norbert J. Zenios & Stavros A (2005) found that spread risk and interest rate risk are Essential variables which won't broaden away in a substantial portfolio

| Page 11 setting and particularly when top notch instruments are considered Bank should focus on limiting such risks keeping in mind the goal to accomplish long run development in the business level.

Sensarma, Rudra & Jaydev M (2009) found a novel method for taking a glimpse at banks financial related aspects that is from the risk management perspective. This review helps in creating outline scores of risk management capacities of banks. As risk management is appeared to be an imperative determinant of stock return of banks, bank ought to adjust all around prepared instrument to control credit risk and push the stock comes back to another level.

Ljaz & Maha (2015) found that examination in credit risk management has considerably moved from estimation of credit risk to the evaluating of credit risk which is more pivotal process for the bank. There is reliable increment in the zone of interest rate risk. In any case, different parts of the region are not yet mindful of its fullest potential. Ghosh S K & Maji S G (2000) analysed and included a risk management based process review into credit review function. It also emphasises interest towards the risk and finds that risk is not a threat but is instead an opportunity to outperform the competition. And also made an explanation of the evolution of the credit review function.

Hussain A & Hassan Al Tamimi (2007) examined the level of risk management techniques and tools used by banks when it considers the different kinds of risks faced by banks are compared by taking look into the questionnaire prepared for the two sets of bank. It can be concluded that commercial banks have to face more foreign exchange risk, operating risks and credit risks. It shows that significant relationship between the foreign and national banks.

Sensarma R & Jayadev M (1998) analysed down the credit risk estimation work over the past 20 years. It thought about the credit chance estimation of individual

| Page 12 advances and portfolio advances. It additionally centered around the new methodology around the mortality hazard to quantify the return on advances and bonds. It examined the hazard return structures of arrangement of credit chance uncovered obligation instruments.

Chahal H, Kaur Sahi G & Rani A (2008) analyzed a calculated model to be utilized further in understanding credit hazard the board arrangement of business bank in an economy by assessing less created money related area. It found that the part of credit hazard the executives framework vary in business banks working in a less created economy from those in a created economy. This gives a suggestion to the earth where bank conveys its task.

Khare A, Khare A & Singh S (2005) analyzed the observational connection between the total credit security recuperation rates and default rates, Found a negative connection between the default rate and bond recovery rates, Once the default rate is taken into account, the supply of outstanding distressed debt remains variable whereas, such macro-economic variables as stock market return do not. Found that systematic macroeconomic risk factor is less periodic than expected and the relative supply and demand of defaulted bonds has a role in ascertaining the average recovery rates.

| Page 13 CHAPTER-3 SOFTWARE

AND HARDWARE REQUIREMENTS

| Page 14 Software Requirements:

Environment : Python

Tool : Jupyter notebook Operating System : Mac OS Hardware

Requirements:

Processor : I5 RAM : 8GB

Storage : 1024 GB

| Page 15 CHAPTER-4

DATA SET

Content:

CREDIT RISK DATASET : The credit risk dataset consists of 32 thousand samples of data that consists of customers financial statements that is used to train the model which predicts the credit risk of the customers. below figure shows the features of the dataset. The current state of various variables' data exploration will be discussed later. Can the borrowers' ages, for instance, be used as an indicator of whether they are in default? Reviewing the age distribution for a good and a terrible person will help (the left figure is the age distribution for a good person, while the right figure is the age distribution for a bad person). Fig:3Dataset and features

Features Description: Our dataset consists of 12 features they are

| Page 17

Fig:4 features of dataset

| Page 18

CHAPTER-5

ARCHITECTURE

| Page 19 Architecture

A dataset from the "Kaggle" website is used as the training dataset in this suggested system. To improve machine performance, the inserted dataset is first examined for duplicates and null values. The dataset is then divided into two smaller datasets, say the "train dataset" and "test dataset," in a ratio of 70:30. The "train" and "test" datasets are then provided as text-processing input parameters.

Fig:5 Model Architecture

| Page 20 Punctuation marks and terms on the stop words list are taken out during text processing and replaced with clean words. The term "Feature Transform" is then passed using these clear phrases. The clean words that the text-processing returned in feature transform are then used in "fit" and "transform" to build a vocabulary for the machine.

Using this data, Classification models are built using classification techniques. Based on these performances predictions are made on the new data.

| Page 21 CHAPTER-6

DATA VISUALIZATION

| Page 22 Data Visualization:

The depiction of data through the use of typical graphics, such as infographics, charts, and even animations, is known as data visualization. These informational visual representations make complex data relationships and data-driven insights simple to comprehend. Data visualization is frequently employed to encourage collaborative brainstorming. They are typically used to promote the collection of various viewpoints and to draw attention to the shared problems of the group during brainstorming or design thinking sessions at the beginning of a project. Even though these visualizations are typically rough around the edges, they assist in laying the groundwork for the project and guarantee that the team is on the same page regarding the issue that they're trying to solve for important stakeholders.

Tableau:

Tableau is a fantastic business intelligence and data visualisation application for reporting and analysing huge amounts of data. It was founded in America in 2003, and in June 2019, Salesforce bought Tableau. It assists users in producing a variety of graphs, maps, dashboards, and stories for the purpose of visualising and analysing data to aid in business decision-making. Tableau is one of the most well-liked business intelligence tools since it has so many interesting, distinctive features (BI). Let's explore some of the key Tableau Desktop features in more detail. Now that we are clear on what exactly the tableau is, let's examine some of its key characteristics.

Features of Tableau

- Users of Tableau may quickly find answers to crucial queries thanks to the platform's robust data search and exploration capabilities.
- Users without relevant experience need not have any prior programming knowledge in order to begin building visualisations with Tableau. I
- t can link to a number of data sources that are not supported by other BI products. Users of Tableau can combine and produce reports from many datasets.
- A centralised location to manage all published data sources inside an organisation is supported by Tableau Server.

| Page 23

Fig:6 Data Insights Dashboard

| Page 24 Fig:7 Insights of Grade A Dashboard

| Page 25 Fig:8 Insights of Grade B Dashboard

| Page 26 CHAPTER-7

METHODOLOGY

| Page 27 7.1. Data Preprocessing :

Data preprocessing is an important step in developing a Machine Learning model, and the quality of the data depends on how well it has been preprocessed. The changes made to our data prior to feeding it to the algorithm are referred to as preprocessing. Data preprocessing is a method for transforming unclean data into clean data sets. In other words, anytime data is collected from various sources, it is done so in a raw manner that makes analysis impossible.

Need of Data Preprocessing:

The format of the data in machine learning projects must be correct in order to get better results from the applied model. For example, the Random Forest algorithm does not accept null values, so null values must be handled from the original raw data set in order to execute the Random Forest algorithm.

Some specific machine learning models require data in a specific format. Another consideration is that the data set should be formatted so that many machine learning and deep learning algorithms can run in parallel and the best one is selected.

1. Filling Data Using Imputation:

Imputation in statistics refers to the process of substituting values for missing data. It is known as "unit imputation" when replacing a data point and as "item imputation" when covering a component of a data point. Missing data can introduce a significant degree of bias, make the process of analyzing the data more complex, and reduce efficiency, which are the three main issues it causes. Imputation is viewed as a technique to avoid the problems inherent with the list-wise deletion of cases that have missing values since missing data might pose problems for data analysis.

fillna() method:

The data frame's fillna() method can be used to impute missing values using the mean, median, mode, or constant value. You might also want to look at the linked

| Page 28 page sklearn.impute: Imputing Missing Data Using Sklearn SimpleImputer. Using the mean, median, mode, or constant value, SimpleImputer is used to impute missing data. Simple methods for imputing missing values are provided by the SimpleImputer class. The statistics (mean, median, or most frequent) of each column in which the missing values are present can be used to impute missing values, or they can be done using a constant value that is given.

- Replacing or filling values using mean value of the features.
- Replacing or filling values using median value of the features.
- Replacing or filling values using mode value of the features.
- Replacing or filling values using constant value of the features.

2. Removing Outliers:

Preparing the data sample before analysis is crucial to ensure that the observations accurately reflect the issue. A dataset may occasionally contain extreme values that are dissimilar from the other data and beyond the expected range. These are known as outliers, and by comprehending and even getting rid of these outlier values, machine learning modelling and model quality, in general, may often be improved.

Outliers are often thought of as samples that deviate greatly from the median of the data. Outliers can be challenging to identify, even with a detailed understanding of the data. The removal or modification of values should be done with extreme caution, especially if the sample size is tiny.

One of the most challenging aspects of data cleanup is identifying outliers and flawed data in your dataset, and it takes time to get it right. It's always a topic to examine cautiously, even if you have a thorough understanding of statistics and how outliers could affect your data.

| Page 29 • Identifying the outliers and dropping the values

- Ignoring the outliers.
- Or replacing outliers with mean value.
- Or replacing outliers with median value.
- Or replacing outliers with mode value.
- Or replacing outliers with constant value.

3. One Hot Encoder:

A one-hot encoding is the representation of binary vectors for category variables. The categorical values must first be converted to integer values in order to do this. Then, each integer value is represented as a binary vector, where all the values are zero except for the integer's index, which is denoted by a 1. Categorical data can be represented more expressively with a one-hot encoding.

Categorical data cannot be directly used by many machine learning techniques. It is necessary to translate the categories into numbers. For categorical input and output variables, this is necessary.

Direct usage of an integer encoding with appropriate scaling is possible. This might be effective for issues where there is a clear ordinal link between the categories and the integer values, like labels for temperature 'cold', 'warm', and 'hot'.

4. Label Encoder:

Label encoding is the process of converting labels into a numeric form so that they may be read by machines. The operation of those labels can then be better determined by machine learning methods. It is a significant supervised learning preprocessing step for the structured dataset. While label encoding converts the data into machine-readable form, it also gives each class of data a distinct number (beginning at 0). This could cause priority concerns to emerge during the training of data sets. A label with a high value could be given more priority than one with a

| Page 30 low value.

5. Standardisation:

Feature Scaling: One of the most critical information preprocessing steps in machine learning is feature scaling. If the data is not scaled, algorithms that estimate the distance between the features are biased towards numerically larger values.

The scale of the features has very little effect on tree-based algorithms. Furthermore, feature scaling promotes faster training and divergence of deep learning and machine learning algorithms. The most widely used and, simultaneously, most perplexing feature scaling methods are normalisation and standardisation.

| Page 31 7.2. Algorithms

7.2.1. Logistic Regression:

- Logistic regression is a popular Machine Learning algorithm that is part of the Supervised Learning technique. It is used to forecast the categorical dependent variable from a set of independent variables.
- Logistic regression is used to forecast the outcome of a categorical dependent variable. As a result, the outcome must be categorical or discrete. It can be Yes or No, 0 or 1, true or False, and so on, but instead of giving the exact values as 0 and 1, it gives the probabilistic values that fall between 0 and 1.
- Logistic Regression is very similar to Linear Regression, with the exception of how they are used. Logistic regression is used to solve classification problems, whereas linear regression is used to solve regression problems.
- Instead of fitting a regression line, we fit an "S" shaped logistic function that predicts two maximum values in logistic regression (0 or 1).
- The logistic function curve indicates the likelihood of something like whether the cells are cancerous or not, whether a mouse is obese or not based on its weight, and so on.
- Logistic Regression is an important machine learning algorithm because it can provide probabilities and classify new data using both continuous and discrete datasets.
- Logistic Regression can be used to classify observations using various types of data and can quickly determine which variables are most effective for classification. The logistic function is depicted in the image below.

| Page 32

Fig:9 Logistic Regression

Logistic Function(Sigmoid Function):

- The sigmoid function is a mathematical function that is used to convert predicted values into probabilities.
- It transforms any real value between 0 and 1 into another.
- The logistic regression value must be between 0 and 1, and it cannot exceed this limit, forming a curve similar to the "S" form. The Sigmoid function or logistic function is another name for the S-form curve.
- The concept of the threshold value is used in logistic regression to define the probability of either 0 or 1. For example, values above the threshold value tend to be 1, while values below the threshold value tend to be 0.

Type of Logistic Regression:

Logistic Regression can be classified into three types:

• Binomial • Multinomial • Ordinal

| Page 33 Logistic Regression Assumptions :

- Assumptions of Logistic regression are given below
- The target variables in binary logistic regression must always be binary, and the desired outcome is represented by factor level 1.
- The model should not have any multi-collinearity, which means that the independent variables must be independent of one another.
- Our model must include meaningful variables.
- For logistic regression, we should use a large sample size.

Pros:

- Logistic regression is easier to use, interpret, and train.
- It makes no assumptions about class distributions in feature space.
- It is easily scalable to multiple classes (multinomial regression) and provides a natural probabilistic view of class predictions.
- It not only indicates the appropriateness of a predictor (coefficient size), but also the direction of association (positive or negative).
- It classifies unknown records very quickly.
- It performs well when the dataset is linearly separable and has a high accuracy for many simple data sets.
- Model coefficients can be interpreted as indicators of feature importance.
- Logistic regression is less inclined to over-fitting but it can overfit in high-dimensional datasets. To avoid over-fitting in these scenarios, regularisation (L1 and L2) techniques may be considered.

Cons:

- Logistic Regression should not be used if the number of observations is less than the number of features; otherwise, it may result in overfitting.
- It creates linear boundaries.
- The assumption of linearity between the dependent and independent variables is the major limitation of Logistic Regression.
- It is only capable of predicting discrete functions. As a result, the Logistic Regression dependent variable is restricted to the discrete number set.

| Page 34 • Because logistic regression has a linear decision surface, it cannot solve non-linear problems. In real-world scenarios, linearly separable data is uncommon.

- Logistic Regression necessitates average or no multicollinearity among independent variables.
- Complex relationships are difficult to obtain using logistic regression. This algorithm is easily outperformed by more powerful and compact algorithms, such as Neural Networks.
- Linear Regression involves the linear relationship of independent and dependent variables.

2. KNN(K- Nearest Neighbour)

- One of the simplest machine learning algorithms, based on the supervised learning method, is K-Nearest Neighbour.
- The K-NN algorithm places the new case in the category that is most comparable to the available categories, assuming that the new case/data and the existing cases are similar.

- The K-NN algorithm saves all available data and categorises new data based on similarity. This means that utilising the K-NN method, fresh data can be quickly and accurately sorted into a suitable category.
- The K-NN algorithm can be used for both classification and regression problems, but it is most frequently utilised for classification issues.
- Because it is non-parametric, it makes no assumptions about the basic data. Other name for it is a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
- During the training phase, the KNN algorithm simply stores the dataset, and when new data is received, it classifies it into a category that is similar to the new data.

| Page 35 Why do we need a K-NN Algorithm?

If there are two categories, Category A and Category B, and we have a new data point, x_1 , which category does this data point belong in? We require a K-NN algorithm to address this kind of issue. K-NN makes it simple to determine the category or class of a given dataset. Consider the following diagram:

Fig 10 shows the before and after the KNN

Pros:

The KNN method has the following benefits:

- No training period: Because the data already contains a model that will serve as the foundation for future predictions, KNN modelling does not require training. As a result, it is very time-effective when improvising for random modelling on the given data.
- Simple implementation: The only thing that needs to be calculated for KNN is the distance between various points using information from various features, and this distance can be calculated with ease using distance formulas like Euclidean or Manhattan.
- Since there is no training period, fresh data can be uploaded whenever you want because it won't change the model.

| Page 36 Cons:

- It does not function well with large datasets because it would be highly expensive to compute the distances between each data instance.
- It also does not work well with high dimensionality since it makes it more difficult to calculate the distances for each dimension.
- Sensitive to erratic and incomplete data.

3. Random Forest Classifier:

• Random Forest is a well-known machine learning algorithm from the supervised learning technique. It can be applied to both classification and regression problems in machine learning. It is based on the concept of ensemble learning, which is a process that involves combining multiple classifiers to solve a complex problem and improve the model's performance.

• "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset," as the name implies. Instead of relying on a single decision tree, the random forest takes the predictions from each tree and predicts the final output based on the majority vote of predictions.

• The greater the number of trees in the forest, the higher the accuracy and the lower the risk of overfitting.

| Page 37

Fig:11 Random Forest Working

Assumptions for Random Forest:

The random forest combines multiple trees to predict the class of the dataset, some decision trees may correctly predict the output while others may not. However, when all of the trees are combined, they correctly predict the outcome. As a result, two assumptions for a better Random forest classifier are as follows:

- There should be some actual values in the dataset's feature variable so that the classifier can predict accurate results rather than guesses.
- The predictions from each tree must have very low correlations.

Why use Random Forest:

The following are some reasons why we should use the Random Forest algorithm:

- It requires less training time than other algorithms.
- It predicts output with high accuracy, and it runs efficiently even on large datasets.

| Page 38 • It can also maintain accuracy when a significant amount of data is missing. It can also maintain accuracy when a significant amount of data is missing. Features Of Random Forest:

- It is the most accurate algorithm available today.
- It performs well on large databases.
- It can handle thousands of input variables without deleting any of them.
- It estimates which variables are significant in the classification.
- As the forest is built, it generates an internal unbiased estimate of the generalization error.
- It has methods for balancing errors in unbalanced class population data sets.
- Forests created can be saved for later use on other data.
- Prototypes are created to provide information about the relationship between variables and classification.
- It computes proximities between pairs of cases, which can be used for clustering, locating outliers, or providing interesting views of the data (via scaling).

How does Random Forest Works:

Assume our dataset contains "m" features:

- "k" features were chosen at random to satisfy the condition $k \leq m$.
- Calculate the root node among the k features by selecting the node with the highest Information gain.
- Dividing the node into child nodes
- Rep the preceding steps n times.
- You end up with a forest made up of n trees.
- Bootstrapping is the process of combining the results of all Decision Trees.

Pros:

- Outliers are not a problem.
- It performs well with non-linear data.
- Reduced risk of overfitting.

| Page 39 • Runs quickly on large datasets. • superior to other classification algorithms in terms of accuracy.

Cons:

- When dealing with categorical variables, random forests are found to be biased.
- Slowly train.
- Linear methods with a large number of sparse features are not suitable.

4. Naïve Bayes: • Naïve Bayes algorithm is a supervised learning algorithm, which is based on the Bayes theorem and used for solving classification problems. • It is mainly used in text classification that includes a high-dimensional training dataset. • Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building fast machine learning models that can make quick predictions. • It is a probabilistic classifier, which means it predicts on the basis of the probability of an object. • Some popular examples of the Naïve Bayes Algorithm are spam filtration, Sentimental

Why It is called Naïve Bayes: • The Naïve Bayes algorithm is comprised of two words Naïve and Bayes, Which can be described as: • Naïve: It is called Naïve because it assumes that the occurrence of a certain feature is independent of the occurrence of other features. Such as if the fruit is identified on the bases of color, shape, and taste, then red, spherical, and sweet fruit is recognized as an apple. Hence each feature individually contributes to identifying that it is an apple without depending on each other.

Gaussian Naïve Bayes: Bayes: It is called Bayes because it depends on the principle of the Bayes Theorem. given data and based on that model's performance, further predictions are done.

| Page 40 • In Gaussian Naïve Bayes, continuous values associated with each feature are assumed

to follow a Gaussian distribution. • This distribution is also called Normal Distribution. • When plotted, it produces a bell-shaped curve that is symmetric about the feature values' mean. Fig:12 GNB Normal distribution graph • Because the features' likelihood is assumed to be Gaussian, the conditional probability is given by:

| Page 41 Prons of GNB: • It is the most significant classifier for binary classification.

- As we get the features as continuous, it handles these features.
- It scales well in terms of predictors and data points.
- It is quick and can make real-time predictions.
- It is not affected by irrelevant characteristics.

Cons of GNB: • It assumes that all predictors (or features) are independent, which is rarely the case in practice. This limits the algorithm's applicability in real-world use cases.

- The 'zero-frequency problem' is encountered by this algorithm, which assigns zero probability to a categorical variable whose category in the test data set was not available in the training dataset.

7.2.5.XGBoost

- XGBoost stands for “Extreme Gradient Boosting”, in which the term “Gradient Boosting” originates from the paper Greedy Function Approximation: A Gradient Boosting Machine, with the aid of using Friedman.

- In gradient boosting, it trains many versions sequentially. Each new version step by step minimizes the loss function ($y = ax + b + e$, e desires unique interest as it's miles an mistakes term) of the entire machine the use of Gradient Descent method. The mastering manner consecutively in shape new fashions to offer a greater correct estimate of the reaction variable.

- The precept concept in the back of this set of rules is to assemble new base newcomers which may be maximally correlated with poor gradient of the loss function, related to the entire ensemble.

| P a g e 42 • The gradient boosted timber has been round for a while, and there are a number of substances at the topic. This educational will provide an explanation for boosted timber in a self- contained and principled manner the use of the factors of supervised mastering. We suppose this rationalization is cleaner, greater formal, and motivates the version system utilized in XGBoost.

- XGBoost is an optimized allotted gradient boosting library designed to be surprisingly efficient, bendy and portable. It implements system mastering algorithms beneath Neath the Gradient Boosting framework. XGBoost affords a parallel tree boosting (additionally called GBDT, GBM) that resolve many information technological know- how issues in a quick and correct.

Pros:

- Less feature engineering required (No need for scaling, normalizing data, can also handle missing values well)
- Feature importance can be found out (it output importance of each feature, can be used for feature selection)
- Fast to interpret
- Outliers have minimal impact.
- Handles large sized datasets well.
- Good Execution speed
- Good model performance (wins most of the Kaggle competitions)
- Less prone to overfitting

Cons:

- Difficult interpretation, visualization tough.
- Overfitting possible if parameters not tuned properly.
- Harder to tune as there are too many hyperparameters.

| P a g e 43 CHAPTER-7

IMPLEMENTATION

| P a g e 44 IMPLEMENTATION

Pre-processing

The Correlation Heatmap is made for simultaneously examining the correlations between numerous parameters. It can assist you in identifying how changes to input fields may affect output fields and in finding unanticipated correlations that may lead to underlying causes. Early in the exploratory process is when the Correlation Heatmap is most helpful. Following the discovery of pairs of interest, you can further explore these pairs using some of the additional tools, such as the Curve Fit Analysis. The analysis of missing data is the first stage in the Preprocessing Target,num_characters, num_words as well num_sentences all have a value of zero, suggesting that the dataset is missing values. To increase the model's efficiency, the missing values should

be addressed. As a result, we substituted missing data with the average values of each column. The dataset is next subjected to the Spearman technique, which is used to determine the degree of correlation between the values.

| P a g e 45

Fig: 13 Attribute to Attribute Correlation

| P a g e 46

Fig:14 Classification Report

Exploratory Data Analysis (EDA): Exploratory Data Analysis is a critical step in data science. It assists the data scientist in comprehending the data at hand and relating it to the business context. Word Cloud is an open-source tool that I will use to visualize and analyze my data. Word Cloud is a text-representation data visualization tool. The image's text sizes represent the frequency or importance of the words in the training data.

Framework For Classifiers:

Using the feature extraction as input, the data is trained using classification algorithms such as Support vector machines, Gaussian Naive Bayes, Logistic Regression, and Random Forest. This data fits into these models with the appropriate parameters for the type of classifiers. After analyzing their performance, hyperparameter optimization is required to improve Accuracy and other evaluation metrics such as Precision, Recall, F1-score, and so on. The results and predictions are provided based on the hyperparameters provided. According to the data, the ensemble learning method Random Forest outperforms all other classification methods in all evaluation metrics. The new data is fit into the model using this as the classifier model, and then the actual predictions are made.

| P a g e 47 CHAPTER-8

RESULTS AND ANALYSIS

| P a g e 48 RESULTS AND ANALYSIS

These terms are also of extreme importance in Machine Learning. We need them for evaluating ML algorithms or better their results. We will present in this Python Machine Learning four important metrics. These metrics are used to evaluate the results of classifications. The metrics are:

- Accuracy • Precision • Recall • F1-Score We will present each of these measures and go over the advantages and disadvantages of each. Each metric evaluates a classifier's performance in a unique way. The most crucial factor in machine learning will be the metrics.

Fig: 15 Confusion Matrix

Accuracy:

Since machine learning is of importance to us, accuracy is also utilized as a statistical metric. A classifier's accuracy is measured statistically as the ratio of right predictions (both TruePositives (TP) and TrueNegatives (TN)) to the total of all predictions (including False Positives (FP) and False Negatives (FN)) made by the classifier (FN). Consequently, the equation for calculating binary accuracy.

| P a g e 49

Accuracy = $(TP + TN) / (TP + TN + FP + FN)$

Recall:

It is the number of correct positive results divided by the number relevant samples (all samples that should have been identified as positive).

Recall = $(TP) / (TP + FN)$

Precision:

Precision is the ratio of the correctly identified positive cases to all the predicted positive cases, the correct and the incorrect case predicted as positive. Precision is the fraction of retrieved documents that are relevant to the query. The formula is:

Precision = $(TP) / (TP + FP)$

J.A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, 109-1311

The formula for the F1 score is :

Precision and recall both contribute equally to the F1 score, which can be viewed as a weighted average of these two metrics.

$$F1 = (2 * \text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$$

A measurement tool for binary classification issues is the Receiver Operator Characteristic (ROC) curve. In essence, it separates the "signal" from the "noise" by plotting the TPR against the FPR at different threshold values. The capacity of a classifier to differentiate between classes is measured by the Area Under the Curve (AUC), which is used as a summary of the ROC curve.

| Page 50 The model performs better at differentiating between the positive and negative classes. the higher the AUC. AUC = 1 indicates that the classifier can accurately distinguish between all positive and negative class points.

However, if the AUC was 0, the classifier would be predicting all negatives as positives and all positives as negatives. There is a good possibility that the classifier will be able to tell the difference between the positive class values and the negative class values when 0.5 AUC. This is the case because more True positives and True negatives can be detected by the classifier than False positives and False negatives.

The classifier cannot discriminate between Positive and Negative class points when AUC=0.5. This indicates that the classifier is either predicting a random class or a fixed class for each data point.

Therefore, a classifier's capacity to discriminate between positive and negative classes is improved by a higher AUC value.

| Page 51 How Does the AUC-ROC Curve Work?

A higher X-axis value on a ROC curve denotes more false positives than true negatives. A larger Y-axis value, however, denotes a greater proportion of true positives compared to false negatives. The capacity to strike a balance between false positives and false negatives will therefore influence the threshold selection. Let's look more closely at how our ROC curve would appear for various threshold settings and how specificity and sensitivity would change. By creating a confusion matrix for each point on the graph that corresponds to a threshold, we can attempt to comprehend it and discuss the effectiveness of our classifier:

Fig:16 TPR and FPR Graph

| Page 52 Accuracy:

• By using four algorithms i.e., Support vector machines, Random Forest classifier, Logistic Regression and Gaussian Naïve Bayes algorithms, we find the higher accuracy. • Support Vector Machines have the highest accuracy when compared to other algorithms. Fig: 17 Output

Fig:18 Resultant ROC Curves

| Page 53 CHAPTER-9 CONCLUSION

AND FUTURE SCOPE

| Page 54 CONCLUSION:

The experimental results exhibit machine learning algorithms are feasible approaches to evaluating credit. Besides, within the study's training and testing samples, such objective evaluation parameters as precision, recall, AUC, KS, and F1 score, indicate Random Forest algorithm, Decision Tree algorithm, and Logistic Regression algorithm can all apply to financial risk analysis.

To sum up, the AUC of the four algorithms is satisfactory in this experiment. Among that, AUC of Logistic regression and naive bayes forest is slightly low than that of the decision tree and random forest indicating can have a stronger ability to distinguish between good and bad samples.

FUTURE SCOPE:

First and foremost, understanding the importance of the role played by the risk management department as represented by the risk administration office and put into practise when the bank extends cash loans to its clients comes first and foremost. To ensure that the loan given by the bank won't be marked as default, it is essential to understand the risk involved in doing so. Furthermore, to identify the essential steps that the bank must take to reduce the risk associated with its lending. The different factors that the bank must consider when making a loan decision in order to lower its loan and preserve profitability are listed below.

| Page 55 CHAPTER-10

BIBLIOGRAPHY

| Page 56 BIBLIOGRAPHY

[1] Science-Management Science: Investigators from North West University Report New Data on Management Science (An Optimised Credit Scorecard To Enhance Cut-off Score Determination)

[2] Cuizhu Meng, Bisong Liu and Li Zhou (2019) The Application Study of Consumer Credit risk model in Auto Financial Institution Based on Logistic Regression. *Advances in Computer Science Research, International Conference on Modelling, Simulation and Big Data Analysis (MSBDA 2019)*

[3] P. Burns and

C. Ody, Validation of consumer credit risk models, Conference Summary, Federal Reserve Bank of Philadelphia & Wharton School's Financial Institutions Center, 2004.

[4] Bagga, S., Goyal, A., Gupta, N., & Goyal, A. (2020). Credit Card Fraud Detection using Pipeling and Ensemble Learning. *Procedia Computer Science*, 173, 104-112.

[5] Caruso, G., Gattone, S. A., Fortuna, F., & Di Battista, T. (2020). Cluster Analysis for mixed data: An application to credit risk evaluation. *Socio- Economic Planning Sciences*, 100850.

[6]

Rtayli, N., & Enneya, N. (2020). Selection Features and Support Vector Machine for Credit Card Risk Identification. *Procedia*

Manufacturing, 46, 941-948.

[8]

TianZ, Xiao, J., Feng, H., & Wei, Y. (2020). Credit Risk Assessment based on Gradient Boosting Decision Tree. *Procedia Computer Science*, 174, 150-160.

[9] Wang, Y., Zhang, Y., Lu, Y., & Yu, X. (2020). A Comparative Assessment of Credit Risk Model Based on Machine Learning—a case study of bank loan data. *Procedia Computer Science*, 174, 141-149.

| Page 57 [10] Zhang, J. L., & Härdle, W. K. (2010). The Bayesian additive classification tree applied to credit risk modelling. Computational Statistics & Data Analysis, 54(5), 1197-1205.

[11] Martin,D(1977).Early warning of bank failure:A logit regression approach. Journal of Banking&Finance, 1(3),249-276

[12] Altman,E.I Haldeman,R. G.,&Narayanan,P.(1977).Zeta tm analysis a new model to identify bankruptcy risk of corporations. Journal of Banking&Finance

[13] Ohlson,

Ki Mun Jung;Lyn C Thomas;;Mee chi So(2015),When to build or when to adjust scorecards The Journal of The Operational Research Society
[16] Machine Learning - Intelligent Systems; Recent Studies from Sun Yat-sen University Add New Data to Intelligent Systems (Domain Adaptation Learning Based On Structural Similarity Weighted Mean Discrepancy for Credit Risk Classification), Journal of Robotics & Machine Learning

16. José Rômulo de Castro Vieira; Flavio Barboza; Vinicius Amorim Sobreiro; Herbert Kimura; (2019) Machine learning models for credit.
17. Danjuma, I., Kola, I. A., Magaji, B. Y., & Kumshe, H. M. (2016). Credit Risk Management and Customer Satisfaction in Tier-one Deposits Money Banks: Evidence from Nigeria. International Journal of Economics and Financial Issues, 6(3S), 225- 230.
18. Duffie, D., & Singleton, K. J. (2012). Credit risk: pricing, measurement, and management. Princeton University Press.
19. Kalra, R. (2012). Credit appraisal system in Allahabad bank. International Journal of Management IT and Engineering, 2(5), 537-559.
20. Bhattacharya, H. (2011). Banking Strategy, Credit Appraisal, and Lending Decisions: A Risk-Return Framework. Oxford University Press.
21. Bodla, B. S., & Verma, R. (2009). Credit risk management framework at banks in India. The IUP Journal of Bank Management, 8(1), 47-72.

Ht and source - focused comparison, Side by Side

Submitted text As student entered the text in the submitted document.
Matching text As the text appears in the source.

SUBMITTED TEXT

15 WORDS

80% MATCHING TEXT

15 WORDS

in partial fulfilment of the requirements for the award of the degree of MASTER OF

in partial fulfillment of the requirement for the award of the degree of Bachelor of

W
<https://www.slideshare.net/nikhileshMane/beproject-pressure-reducing-and-desuperheater-station>

SUBMITTED TEXT

193 WORDS

96% MATCHING TEXT

193 WORDS

declare that this submission represents my ideas in my own words, and where others' ideas or words have been included, I have adequately cited and referenced the source. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed. (

declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

W
<https://www.slideshare.net/nikhileshMane/beproject-pressure-reducing-and-desuperheater-station>

SUBMITTED TEXT

31 WORDS

100% MATCHING TEXT

31 WORDS

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit- risk models via machine-learning algorithms. Journal of Banking & Finance, 34(11), 2767-2787.
[7]

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine- learning algorithms. Journal of Banking & Finance, 34(11), 2767{2787.

<https://su-plus.strathmore.edu/bitstream/handle/11071/6789/Consumer%20credit%20risk%20modelling%20...>

SUBMITTED TEXT

14 WORDS

100% MATCHING TEXT

14 WORDS

J.A.(1980).Financial ratios and the probabilistic prediction of bankruptcy. Journal of Accounting Research,109-1311
[14]

J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. Journal of accounting research, 109{131.

<https://su-plus.strathmore.edu/bitstream/handle/11071/6789/Consumer%20credit%20risk%20modelling%20...>

SUBMITTED TEXT

20 WORDS

Rtayli, N., & Enneya, N. (2020). Selection Features and Support V ector Machine for Credit Card Risk Identification. Procedia

82% **MATCHING TEXT**

20 WORDS

Rtayli N, Enneya N. selection features and support vector machine for credit card risk identification. Procedia

<https://link.springer.com/article/10.1007/s44230-022-00004-0>