International Journal of Applied Mathematics
Volume 38 No. 4s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

DEBENTURE BANKRUPT PREDICTION USING MACHINE LEARNING TECHNIQUES

Dr.M.P.Vani

Associate Professor Sr.

Vellore Institute Of Technology, Vellore, India

mpvani@vit.ac.in

Abstract

In Debenture bankrupt prediction using machine learning techniques many people take loan from different bank and companies like India Bulls but the authority of that company decide whether to sanction the loan for the loan seeker or not based on certain criterion. But nowadays technology become so powerful that using machine leaning techniques we can decide whether it would be good decision to sanction the loan or not based on the individual's transaction history. And here four techniques CNN, Naives-Bayes Classifier, CVM(Support Vector Machine) is used to determine the loan seeker is a good returner or not and also it see's which method's among the used four methods is best on the basis of efficiency.

A strong ability to predict the factors affecting repayment is of great importance to mitigate information asymmetry among borrowers, lending platforms and lenders should be displayed on the borrowing webpage.

Keywords: CNN, Naives-Bayagees Classifier, CVM(Support Vector Machine), asymmetry, webpage, screening,loan defaulters, webpage

Introduction:

Lending companies in and around the world have undergone extensive reorganization. In the past Few years, classic statistical analysis and mathematical programming have led to the development Of several data mining techniques in the domain of credit rating and assessment of borrowers' default risk. While human experts are limited and may ignore important signals, interest in machine learning to determine the driving factors of repayment failure has arisen in the field of lending due to multiple factors and to improve the accuracy and efficiency of decision making with respect to borrower screening by platforms and risk supervision by regulators. Increasingly complicated machine learning techniques have helped decision makers to analyze a large amount of data in a timely manner. Considering the importance and urgency of evaluating the growing risk of borrowers, we endeavor to improve the predictive power by taking advantage of the latest Machine learning methods. Standard Logistic Regression attempts to maximize the log likelihood of the estimates. Since log(h(x(i))) and log(1-h(x(i))) are always zero or negative, misclassification results in a large negative number. As our goal is to minimize default risk by focusing on increasing precision, we modify the log likelihood

Received: August 11, 2025

International Journal of Applied Mathematics
Volume 38 No. 4s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

estimate by multiplying the y(i)=0 term by a penalty factor beta (β). $l(\theta)=\Sigma y(i)\log(h(x(i)))+\beta(1-y(i))\log(1-h(x(i)))$ (1)$mi=0$ If the classifier incorrectly classifies a defaulted loan (y(i)=0) as a good loan (h(x(i))~1), the log likelihood of this same will be multiplied by the factor β. A high β will cause the classifier to avoid incorrectly classifying defaulted loans (y(i)=0) as paid off loans (y(i)=1), even if that means increasing the misclassification of a paid off loans as defaulted loans. This effectively gives preference for precision at the cost of recall and overall accuracy. Introducing β into the log likelihood, it follows that the first and second derivatives are: $\partial\partial\theta jl(\theta)=(y-(y+\beta+\beta y)h(x))\ xj$ (2) $\partial 2(\theta)\partial\theta j\partial\theta k=-(y+\beta+\beta y)h(x)(1-h(x))\ xjxkT$ Note that when β=1, all three equations above revert to the original Logistic Regression equations. With the first and second derivatives, we used Newton-Raphson as the optimization method: $\theta:=\theta-H-1\ \nabla(\theta)$ (4)Financial enterprises operating on the internet are developing rapidly with the advent of the Web 3.0 era. Peer-to-peer () lending is one online financial platform that satisfies the need for providing small and micro loans among strangers. The first lending company, Zola, was created in London in 2005 and was followed by high-speed growth of lending companies around the world. As a developing country, China has experienced a major boom in lending over the past five years in terms of both the number of borrowers and the total value of loans. The first Chinese lending platform was established in 2006 (Huang, 2017). Five years later 50 platforms were in operation, and the number jumped to 1,931 as of December 2017. However, Chinese internet-based financial institutions have accumulated substantial financial risks due to the prosperity of the lending industry. By the end of June 2018, the number of on-going lending companies had dropped to 1,836, 95 less than at the end of 2017. In the past 5 years, lending companies in China have failed due to variety problems, such as unreal investments, inability to repay loans, platform revocation, and juridical person escape, making lending a potential disastrous venture full of financial risks

Project Description

The existing models of loan defaulter prediction are potentially faced with problems such as scale

Limitation and variety. The conventional machine learning is difficult to train and test large data And tree-based classification methods with high performance require many features. Also, Statistical methods and hand-crafted methods are limited in extracting features by capturing the Relationship between complex variables inherent in various data. Most of the companies review the borrower's information and give each borrower a score according to a forecasting method. Understanding the factors that lead to repayment failure is crucial and necessary for lending platforms to perform borrower screening and for regulatory agencies to conduct risk management. A strong ability to predict the factors affecting repayment is of great importance to mitigate information asymmetry among borrowers, lending platforms and lenders. Given that the existing assessments for borrowers are inadequate and subject to unknown factors, a wide variety of personal information of the borrowers should be verified, and whether the information was verified should be displayed on the borrowing webpage. If a model is able to identify credit-worthy customers that were not recognized by traditional credit scores while

Received: August 11, 2025

International Journal of Applied Mathematics
Volume 38 No. 4s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

minimizing their risk of default on the loans, this can be a lucrative niche market or micro-market. Pushing higher the profit margin of the financial institution or investor. Although, the prospect of more customers seems positive, we need to be careful as to not lend to people that will default on the loan, this would cause a drop in the earlier stated objective. Thus a conservative approach and rigorous evaluation metrics will be kept in mind throughout the project. The loan default prediction is a problem of binary classification (do we lend or not). The existing models of loan defaulter prediction are potentially faced with problems such as scale limitation and variety. The conventional machine learning is difficult to train and test large data and tree-based classification methods with high performance require many features. Also, statistical methods and hand-crafted methods are limited in extracting features by capturing the complex variables inherent in various data Goal. Peer to Peer allows the users with a special loan approval service system which allows all courier transactions through a smart mobile phone. It allows the user to select how much they want to charge them for sending the parcel. It provides a map for training datasets which will enable them to easily navigate to their picking and delivery point. Product Perspective The system would allow Users to make instant delivery orders allowing them to select locations for pick-up and drop-off of the parcel. Loan seekers can request with an option to accept or reject requests. Once a company accepts a loan approval , the app enables the history checker and includes his loan transactions in the datasets. Meanwhile, the server maintains and tracks the transactions among entities, thus enabling connections to favorable customers.

Paper Goal

This paper intends to provide a platform which transforms data sets into an easy, instantaneous and dynamic sector, thus eliminating the limitations in customer responsiveness faced currently. The platform widens the scope of training data sets by ensuring real-time training services for products of any shapes and sizes over any distance metrics without the customers needing to step out of their houses. The ultimate goal is that the end user does not have to perform those complex set of operations like training and deciding which methodology is best to train the datasets so that they can decide whether to sanction the loan for the loan seeker. This series of steps can be performed just by simple clicks of button on your mobile. This can save user lots of useful things.

like, Time Reduction: user can save the time to reach the decision. Rate: No charges required. Training datasets: The programmer and developers can train the datasets. Less chance of wear and tear of products, and much more.

Assumptions

It is assumed that the company authority must have a working system and have good knowledge About using it. Also, Company labs should have good internet connection to use this real time application. The application is easy to learn and use so it will not be challenging task to get used to As per the initial concept for Peer to Peer, people will have to register oneself by signing in the application. This will create their account which can then be used from anywhere and also with any other android mobile phone as well. The sign up requires a

Received: August 11, 2025

International Journal of Applied Mathematics
Volume 38 No. 4s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Valid E-mail with a Phone number which has to be provided at the time of on boarding. As this are the default channels of communication in this paper.;

User Characteristics Company – they can use these methods to train various datasets, Loan seeker-They can seek loan through this platform. Coders and Developers- Coders and developers will train the datasets.

Domain Requirements - The system should work with all the windows system and Linux systems Also, the system should have internet connectivity. User Requirements and Product Specific System Requirements - A different way to train, Real time information sharing, Eco-friendly, Ease in Delivering, Time Saving Digitalization in Exchange sector.

Business Strategy and Infrastructure Challenges: The loan is one of the most important products of the banking. All the banks are trying to figure out effective business strategies to persuade customers to apply their loans. However, there are some customers behave negatively after their application are approved. To prevent this situation, banks have to find some methods to predict customers' behaviors. Machine learning algorithms have a pretty good performance on this purpose, which are widely-used by the banking. Here, it will work on loan behaviors prediction using machine learning models.

1.6 Infrastructure Challenges the modus operandi of credit rating by banks and rating agencies. The impact of economic downturn on the behaviors of borrowers as well as lenders. 3. Mode of calculation of default probability. One of the first steps engaged was to outline the sequence of steps that we will be following steps of elaboration as mentioned below

Phase I

Data Extraction and Cleaning:

Missing Value Analysis and Treatment. In the dataset the target shows that 94% have not defaulted and 6% are defaulters or charged off.

So this is clearly an unbalanced dataset. The first issue was to know if the columns were filled with

Useful information or were mostly empty. Data exploration uncovered many empty or almost empty columns which were removed from the dataset because it would prove a difficult task to go back and try to answer for each data point that did not seem necessary at the time of the loan application. Our dataset has 855969 rows × 73 features including the target out of which 32 have missing values or NAN. Below we will look at a plot and get some insights. Insights: So, we can see from the above plot that there are 20+ columns in the dataset where all

the values are NA.As we can see there are 855969 observations & 73 columns in the dataset, it will

be very difficult to look at each column one by one & find the NA or missing values. So find out all columns missing values which are more than certain percentage, say 40%.remove those columns as it is not feasible to input missing values for those columns. Out of 73 features only kept 53 was kept. So removed about 21 features that had more than40% missing values

Received: August 11, 2025

International Journal of Applied Mathematics
Volume 38 No. 4s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

since it will not make any sense in further exploration. Pay close attention to data leakage, which can cause the model to overfit. This is because the model would be also learning from features that wouldn't be avail-able when using it makes predictions on future loans By checking the variance remove low variance variables (excluding out target).However, since this is a sensitive problem remove based on once own discretion and business knowledge. Some irrelevant columns Unique ID's such as "id","member_id" because they did not provide any useful information about the customer. As last 2 digits f zip code is masked 'xx', remove that as well. still it had five more features with null values (emp_length,revol_util,tot_coll_amt,tot_cur_bal,total_rev_hi_lim) which have inputted using fill method the appropriate statistic .Feature Extraction Decide On A Target Column Now, let's decide on the appropriate column to use as a target column for modeling keeping in mind the main goal is predict who will pay off a loan and who will default. By We earning from the description of columns in the preview Data Frame that default is the only field in the main dataset describe a loan status, so use this column as the target column.94% have not defaulted(fully paid) and 6% are defaulters or charged off.

Transformation

It has transformed term, emp-length and grade to integer values using transformation technique

so that they provide some information to the model.

Purpose of loan

Drop records where values are less than 0.75%. Analyze only those categories which contain more than 0.75% of records. Also, there is no awareness about what comes under 'Other' remove this category as well. Drop the columns that have more than 7 levels (categories) since it is not feasible to one hot encode them. The variables removed are 'earliest_cr_line','addr_state'.

Mapping and Transformation

The columns "status" in table "loan" is the target variable, which stands for the customers' loan

Behaviors. Then select useful features for the model. The standard is based on relevant to the target and some business sense. For example, the columns "bank to" and "account to" have nothing to do with loan behaviors, so drop them.

Next step is the data preprocessing. First, convert categorical variable into numeric.

Usually it is part of feature engineering in Python. Use two methods.

Analyze only those categories which contain more than 0.75% of records. Also, there is no

Awareness what comes under 'Other' remove this category as well.

Drop columns that has more than 7 levels (categories) since it is not feasible to one hot

encode them. The variable removed were 'earliest_cr_line','addr_state'.do not remove issued because the variable need to be used for splitting feature Engineering Casting continuous

Received: August 11, 2025

International Journal of Applied Mathematics
Volume 38 No. 4s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

variables to numeric:Cast all continues variables that are necessary for analysis to numeric so that correlation between them can be found .

Mapping: Mapping the issue date from "Jun-2015" to "Dec-2015" as Test for the ease of splitting the

data to test set using Dictionary.

Correlation: Finding the correlation between variables. look at the correlation structure between the variables that we selected above. This will tell us about any dependencies between different variables and help us reduce the dimensionality a little bit more The variables checked for correlation are:[loan_amnt','funded_amnt','funded_amnt_inv','installment','int_rate','annual_inc','dti','to-tal_pymnt','total_pymnt_inv','total_rec_prncp`]Insights

It is clear from the Heatmap that how 'loan_amnt','funded_amnt' &'funded_amnt_inv' are closely

interrelated . So we can take any one column out of themfor our analysis. Also

'total_pymnt','total_pymnt_inv' are highly correlated.

Transformation:

Box-cox

Before training the data, first transform the data to account for any scenes in the variable distribution. Various transformation techniques ranging from log trans-form to power transformation are available. For the analysis, use Box-co transformation. It is used to modify the distributional shape of a set of data to be more normally distributed so that tests and confidence limits that require normality can be ppropriately used.

Feature Scaling:

Scaling the data so that each column has a mean of zero and unit standard deviation. Scaled the training set and test set as well so as to reproduce the same results.

One Hot Encoding

Since there are some categorical variables for the analysis and the machine learning algorithms doesn't take categorical and string variables directly, we have to create dummy variables for them. We can either encode them using label encoder available for python but it would be wrong in the analysis since a lot of these variables have multiple categories. Just using weights can cause discrepancies in the algorithm. Instead, use one hot encode so that one can have a 1 wherever that category turns up and 0 otherwise. This will also create separate columns for each level of category. Also, we'll be dropping one of the categories so that we have N-1 columns instead of N.;

Business Insights and Recommendations

Successfully machine learning algorithm is built to predict the people who might default on their loans. Also, might want to look on other techniques or variables to improve the prediction

Received: August 11, 2025

International Journal of Applied Mathematics
Volume 38 No. 4s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

power of the algorithm. One of the drawbacks is just the limited number of people who defaulted on their loan in the 8 years of data (2011-2019) present on the dataset. use an updated data frame which consist next 3 years values (2016-2019) and see how many of the current loans were paid off or defaulted or even charged off. Then these new data points can be used for predicting them or even used to train the model again to improve its accuracy. Since there are lot of categorical data, one cannot apply PCA for dimensionality reduction. Because of this, try some different type of variable selection method like 'MULTIPLE CORRESPONDENCE ANALYSIS' to reduce the dimensionality and select the most important variables from the columns. Since the algorithm puts around 40% of non-defaulters in the default class, one might want to look further into this issue to make the model more robust. Business Insights and Recommendations:

The facts from our analysis shows that Applicants who has taken the Loan for 'small business' has

The highest probability of charge off of 14%. Hence, bank should take extra caution like take some

Asset or guarantee while approving the loan for purpose of 'small business' Banks should consider;

Grade" as a major variable while providing loans. Also, pas the annual income is decreasing the

Probability that person will default is in-creasing with highest of 7% at (0 to 25000) salary bracket.

The banks should either start with less principal loan

Amount and check the credibility. Finally, As

The interest rate is increasing the probability that person will default with highest of 9% at 15% &

Above bracket. Banks should consider minimizing their inter strange for Applicants who are self-

Employed & less than 1 year of experience as they are more probable of charged off

Project Schedule and Milestones

develop a generic framework that could be used to deploy any use case quickly. After working with

some popular application development tools, decided to go with understanding Firebase Database since

the whole environment is integration – ready and robust. After the following above task, the Milestones

for the project are stated as follows:

Received: August 11, 2025

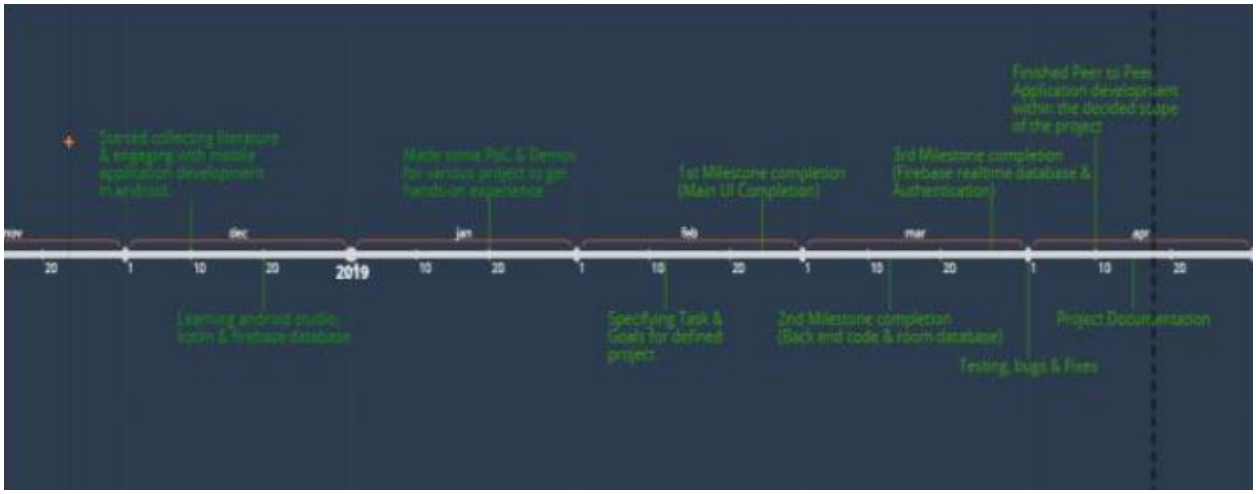
International Journal of Applied Mathematics
Volume 38 No. 4s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

• Start to collect datasets

• Decide on what platform, you want to train

• Then try to list out the methodologies you want to use.

•Then try to train them in that methodologies;



;

Methodologies used:

1. Supporting vector machine(SVM):

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating

hyper plane. In other words, given labeled training data (supervised learning), the algorithm outputs

an optimal hyper plane which categorizes new examples. In two dimensional space this hyperplane

is a line dividing a plane in two parts where in each class lay in either side.

Algorithm

1. Define an optimal hyper plane: maximize margin

2. Extend the above definition for non-linearly separable problems: have a penalty term for

misclassifications.

3. Map data to high dimensional space where it is easier to classify with linear decision surfaces:

reformulate problem so that data is mapped implicitly to this space.

Received: August 11, 2025

International Journal of Applied Mathematics
Volume 38 No. 4s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

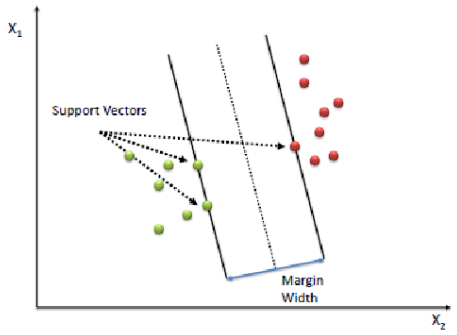
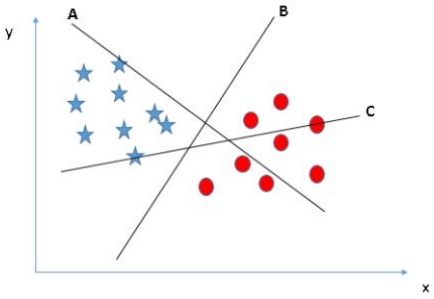


Fig 5.1: An SVM graph



;

Above, got accustomed to the process of segregating the two classes with a hyper-plane. Now

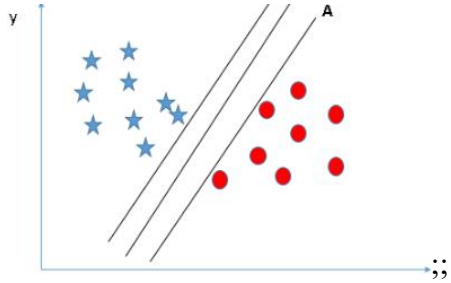
The burning question is “How can we identify the right hyper-plane?”. Don’t worry, it’s not as hard

As you think!

Understand:

Identify the right hyper-plane (Scenario-1): Here, three hyper-planes (A, B and C). Now,

Identify the right hyper-plane to classify star and circle.



;;

one way of plotting the points in the graph;

Here, maximizing the distances between nearest data point (either class) and hyper-plane will help

to decide the right hyper-plane. This distance is called as Margin. look at the below

snapshot:

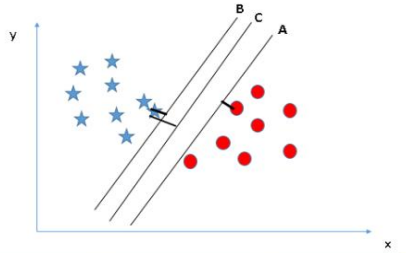


Fig 5.4: hyper-plane with higher margin is robustness

;

2. K-Nearest neighbors Classifiers (KNN):

Received: August 11, 2025

International Journal of Applied Mathematics
Volume 38 No. 4s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

K-nearest neighbors (KNN) algorithm uses 'feature similarity' to predict the values of new Data points which further means that the new data point will be assigned a value based on how Closely it matches the points in the training set. We can understand its working with the help of

Following steps:

Step 1 − For implementing any algorithm, we need dataset. So during the first step of KNN, must load the training as well as test data.

Step 2 − Next, choose the value of K i.e. the nearest data points. K can be any integer.

Step 3 − For each point in the test data do the following −

3.1 − Calculate the distance between test data and each row of training data with the help of any

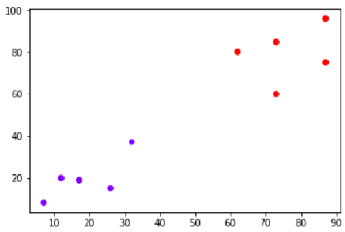
of the method namely: Euclidean, Manhattan or Hamming distance. The most commonly used method to calculate distance is Euclidean.

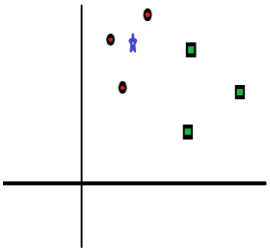
3.2 − Now, based on the distance value, sort them in ascending order.

3.3 − Next, it will choose the top K rows from the sorted array.

3.4 − Now, it will assign a class to the test point based on most frequent class of these rows.

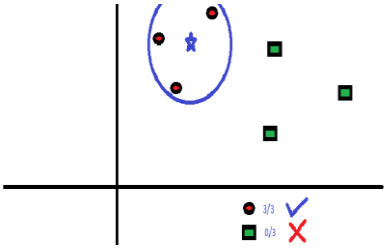
Step 4 – End





ig 5.10: SVM plotted in a box

How does the KNN algorithm work? Let's take a simple case to understand this algorithm. Following is a spread of red circles (RC) and You inptend to find out the class of the blue star (BS). BS can either be RC or GS and nothing else. The "K" is KNN algorithm is the nearest neighbor to take the vote from. Let's say K = 3. Hence, now make a circle with BS as the center just as big as to enclose only three datapoints on the plane. Refer to the following diagram for more details:

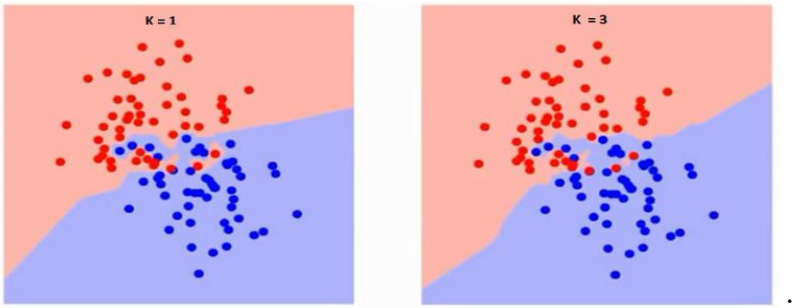


Received: August 11, 2025

International Journal of Applied Mathematics
Volume 38 No. 4s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

The three closest points to BS is all RC. Hence, with a good confidence level, say that the BS should belong to the class RC. Here, the choice became very obvious as all three votes from the closest neighbor went to RC. The choice of the parameter K is very crucial in this algorithm. Next, we understand what are the factors to be considered to conclude the best K

How to choose the factor K? First try to understand what exactly does K influence in the algorithm. If the last example is seen, given are the all the 6 training observation which remain constant, with a given K value it can effect of value "K" on the class boundaries

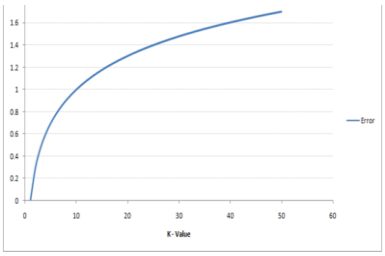


;

: KNN at k=1 and k=3;

if watched carefully, one can see that the boundary becomes smoother with increasing value of K. With K increasing to infinity it finally becomes all blue or all red depending on the total majority. The training error rate and the validation error rate are two parameters we need to access

different K-value. Following is the curve for the training error rate with a varying value of K



As it is seen, the error rate at K=1 is always zero for the training sample. This is because the closest

point to any training data point is itself. Hence the prediction is always accurate with K=1. If Validation error curve would have been similar, choice of K would have been 1. Following is the validation error curve with varying value of K

Using Bayes theorem, one can find the probability of A happening, given that B has occurred. Here,

B is the evidence and A is the hypothesis. The assumption made here is that the predictors features

Received: August 11, 2025

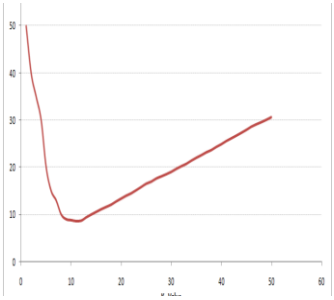
International Journal of Applied Mathematics
Volume 38 No. 4s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

are independent. That is presence of one particular feature does not affect the other. Hence it is called naive.

Suppose in a data set X is input matrix and Y is output matrix. These are specified as

Here x1,x2….xn represent the features. Y is a column;

$X = (x_1, x_2, x_3, ..., x_n)$;



This makes the story more clear. At K=1, overfitting the boundaries were done. Hence, error rate

initially decreases and reaches a minima. After the minima point, it then increase with increasing

K. To get the optimal value of K, now it can segregate the training and validation from the initial;

dataset. Now plot the validation error curve to get the optimal value of K. This value of K should

be used for all predictions.

Breaking it Down – Pseudo Code ofp KNN

implement a KNN model by following the below steps:

1. Load the data

2. Initialise the value of k

3. For getting the predicted class, iterate from 1 to total number of training data points

1. Calculate the distance between test data and each row of training data. Here we will

use Euclidean distance as our distance metric since it’s the most popular method.

The other metrics that can be used are Chebyshev, cosine, etc.

2. Sort the calculated distances in ascending order based on distance values

3. Get top k rows from the sorted array

4. Get the most frequent class of these rows

Received: August 11, 2025

International Journal of Applied Mathematics
Volume 38 No. 4s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

5. Return the predicted class

Naive Bayes Classifier:

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features,

all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'. Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

Bayes theorem provides a way of calculating posterior probability **P(c|x)** from P**(c), P(x)** and

P(x|c).

Look at the equation below:

A Naive Bayes classifier is a probabilistic machine learning model that's used for classification

task. The crux of the classifier is based on the Bayes theorem.

Using Bayes theorem, one can find the probability of A happening, given that B has occurred. Here,

B is the evidence and A is the hypothesis. The assumption made here is that the predictor's features are independent. That is presence of one particular feature does not affect the other. Hence it is

called naive.

Suppose in a data set X is input matrix and Y is output matrix. These are specified as

Here x1,x2….xn represent the features. Y is a column matrix which contains two or more classes.

So according to bayes theorem, / $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

Using Bayes theorem, find the probability of A happening, given that B has occurred. Here,

B is the evidence and A is the hypothesis. The assumption made here is that the predictors features

are independent. That is presence of one particular feature does not affect the other. Hence it is

called naive.

Suppose in a data set X is input matrix and Y is output matrix. These are specified as

Here x1,x2….xn represent the features. Y is a column matrix which

$X = (x_1, x_2, x_3, ..., x_n)$

Received: August 11, 2025

International Journal of Applied Mathematics
Volume 38 No. 4s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

here x1,x2….xn represent the features. Y is a column matrix which contains two or more classes.

So according to bayes theorem

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)}$$

$$P(y|x_1, ..., x_n) = \frac{P(x_1|y)P(x_2|y)...P(x_n|y)P(y)}{P(x_1)P(x_2)...P(x_n)}$$

Now, one can obtain the values for each by looking at the dataset and substitute them into the equation. For all entries in the dataset, the denominator does not change, it remain static. Therefore,

the denominator can be removed and a proportionality can be introduced.

$$P(y|x_1, ..., x_n) \propto P(y) \prod_{i=1}^{n} P(x_i|y)$$

Proposed approach :

1. Data collection Phase : This phase UCI credit data is taken as it is having higher event rate and low class imbalance

2. Data Normalization Phase: In this phase the values are normalized between 0 and 1 to avoid over fitting.

3. Train-Test Split Phase: This phase splits data into training and testing data subsets. For Example, data are divided into two parts per a ratio of 80% training data and 20% test data.

4. Data-Preprocessing Phase: Before the data is fed to the model all the null and redundant Values are removed.

5. Model-Building Phase: In this phase we are using sclera package of python which contains Many packages for classification and regression task. Here we are using four mostly used Classifiers SVM, KNN and naïve bias, Artificial Neural Network and compare their accuracy To choose the best model. It is observed that SVM performs better than the other two models.

6. Prediction Phase: In this phase we test our model with the test input data and make the Prediction. Then that output is compared with testing data to calculate loss and accuracy. The Accuracies for KNN and Naïve bias are 100%

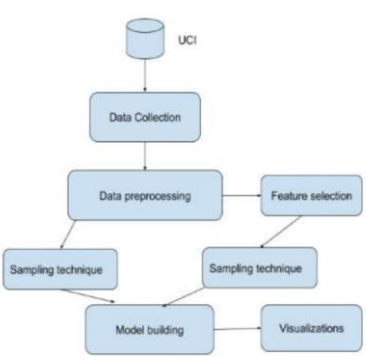
How Naive Bayes algorithm works?
Let’s understand it using an example. Below I have a training data set of weather and corresponding target variable ‘Play’ (suggesting possibilities of playing). Now, we need to classify
whether players will play or not based on weather condition. Let’s follow the below steps to

Received: August 11, 2025

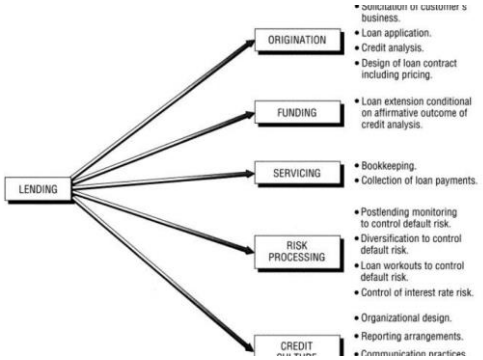
International Journal of Applied Mathematics
Volume 38 No. 4s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

perform it.
Step 1: Convert the data set into a frequency table
Step 2: Create Likelihood table by finding the probabilities like Overcast probability = 0.29 and
probability of playing is 0.64.
Step 3: Now, use Naive Bayesian equation to calculate the posterior probability for each class. The
class with the highest posterior probability is the outcome of prediction.
What are the Pros and Cons of Naive Bayes?
Pros: It is easy and fast to predict class of test data set. It also perform well in multi class , prediction When assumption of independence holds, a Naive Bayes classifier performs better compare to other models like logistic regression and you need less training data. It perform well in case of categorical input variables compared to numerical variable(s). For numerical variable, normal distribution is assumed (bell curve, which is a strong assumption)

Proposed Architecture:



;



	C0	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15
0	b	30.83	0.000	u	g	w	v	1.25	t	t	1	f	g	202	0	+
1	a	58.67	4.460	u	g	q	h	3.04	t	t	6	f	g	43	560	+
2	a	24.5	0.500	u	g	q	h	1.50	t	f	0	f	g	280	824	+
3	b	27.83	1.540	u	g	w	v	3.75	t	t	5	t	g	100	3	+
4	b	20.17	5.625	u	g	w	v	1.71	t	f	0	f	s	120	0	+

Fig 7.1: Dataset 1

1st dataset plotting of good credit and bad credit

;

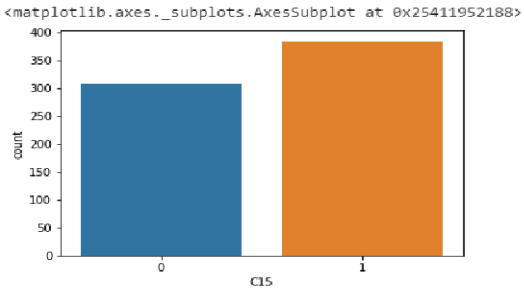


Fig 7.2: plotting of good credit and bad credit

;

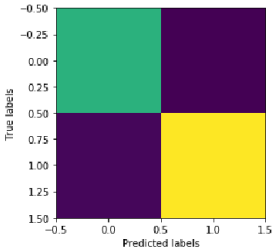
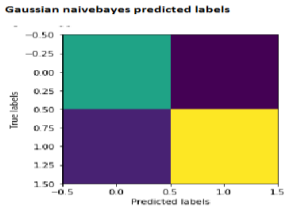


Fig 7.5: KNN graph

International Journal of Applied Mathematics
Volume 38 No. 4s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)



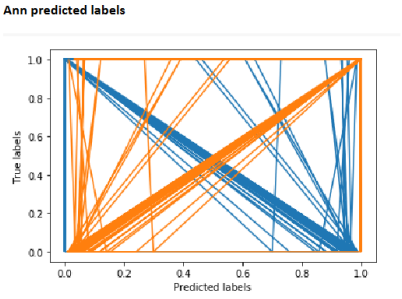


Fig 7.4: Naïve Bayes graph on dataset 1 p ;

2nd dataset

Performance matrix

	svm	knn	Nb
0	0.578125	0.640625	0.625000
1	0.578125	0.656250	0.781250
2	0.578125	0.750000	0.953125
3	0.562500	0.593750	0.875000
4	0.578125	0.718750	0.843750
5	0.578125	0.656250	0.828125
6	0.571429	0.666667	0.809524
7	0.571429	0.619048	0.746032
8	0.571429	0.650794	0.952381
9	0.571429	0.603175	0.825397

	checking_balance	months_loan_duration	credit_history	purpose	amount	savings_balance	employment_length	installment_rate	personal_status	c
0	< 0 DM	6	critical	radio/tv	1169	unknown	> 7 yrs	4	single male	
1	1 - 200 DM	48	repaid	radio/tv	5951	< 100 DM	1 - 4 yrs	2	female	
2	unknown	12	critical	education	2096	< 100 DM	4 - 7 yrs	2	single male	
3	< 0 DM	42	repaid	furniture	7882	< 100 DM	4 - 7 yrs	2	single male	
4	< 0 DM	24	delayed	car (new)	4870	< 100 DM	1 - 4 yrs	3	single male	

; ;;

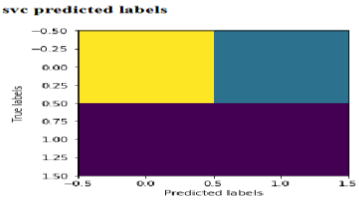
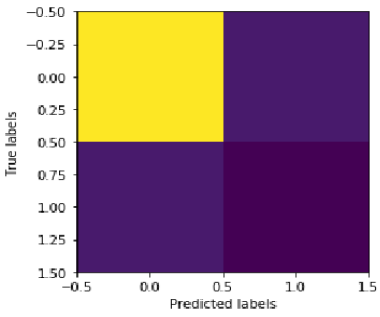
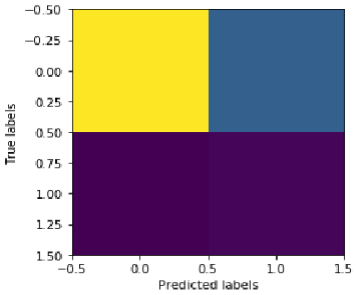


Fig 7.7: svm graph on dataset 2

;

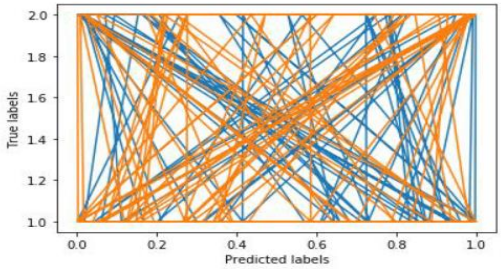
Gaussioan naïve bayes predicted labels





;**ig 7.9: a graph on KNN**

Ann predicted labels



;

International Journal of Applied Mathematics
Volume 38 No. 4s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Cost Analysis:

The proposed Loan Defaulters' system is built by using open sources, there isn't a direct cost associated with it. Further if we consider the cost to the user – Peer to Peer service will be a free application so the price user needs to consider is only the one for the approval of the loan.

Conclusion

The performance of lending platforms can be improved with fewer features. The processing procedure using GPU leads to reduced runtime. Therefore, the workload of the staff credit assessment may be educed, because they do not have to take into database a large number of features during the valuation procedure. The experimental results show that the method is effective in predicting loan defaulters afferent classification algorithms or even different machine learning architecture such as neural networks could provide better results. In at least one known case, it has be used by a Fitch company such as Underwrite.ai (Underwrite, 2016). Deep learning can be another way to discover and push the boundaries of what is possible in predicting the outcome of loans. Along with different algorithms, we could create more features. We could as well ensemble models to make better predictions.

References

1. Guo, Y. Zhou, W. Luo, C. Liu, C. Xiong, H. Instance-based credit risk assessment for investment decisions in P2P lending. Eur. J. Oper. Res. 2016, 249, 417–426.
2. Xu, J.; Chen, D.; Chau, M. Identifying features for detecting fraudulent loan requests on P2P platforms. In Proceedings of the IEEE Conference on Intelligence and Security Informatics, Tucson, AZ, USA, 28–30 September 2016; pp. 79–84.
3. Serrano-Cinca, C.; Gutiérrez-Nieto, B.; López-Palacios, L. Determinants of default in P2P lending. PLoS ONE 2015, 10, e0139427. [CrossRef] [PubMed]
4. Yan, J.; Yu, W.; Zhao, J.L. How signaling and search costs affect information asymmetry in P2P lending: The economics of big data. Financ. Innov. 2015, 1, 19.
5. Serrano-Cinca, C.; Gutiérrez-Nieto, B. The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (P2P) lending. Decis. Support Syst. 2016, 89, 113–122.
6. Everett, C.R. Group membership, relationship banking and loan default risk: The case of online social lending. Bank. Financ. Rev. 2015, 7.
7. Lin, X.; Li, X.; Zheng, Z. Evaluating borrower's default risk in peer-to-peer lending: Evidence from a lending platform in China. Appl. Econ. 2017, 49, 3538–3545.
8. Malekipirbazari, M.; Aksakalli, V. Risk assessment ipn social lending via random forests. Expert Syst. Appl. 2015, 42, 4621–4631.
9. Jiao, Z.; Gao, X.;Wang, Y.; Li, J.; Xu, H. Deep convolutional neural networks for mental load classification based on EEG dataPattern Recognit. 2018, 76, 582–595

Received: August 11, 2025