# CREDIT ANALYSIS USING K-NEAREST NEIGHBOURS MODEL

[1]Prajwal Pitlehra, [2]Dr. R. C. Jaiswal
[1]Student, [2]Associate Professor,
[1] E&TC Department,
[1]Pune Institute of Computer Technology, Pune, India.

***Abstract :*** Many banks are revamping their business models because of the technology related to computational power and big data. Credit risk predictions and monitoring, model reliability and loan processing are detrimental to decision-making. There are several methods to analyse credit risk. In this paper, we have explained one of the algorithms that run behind the curtains- K Nearest Neighbours (KNN). We start our discussion by discussing the various methodologies prevalent in the industry and then focus our attention to KNN. We end with a discussion on the table of our results thus obtained from the methodology.

***IndexTerms*** **- Credit analysis, Machine learning, Confusion matrix.**

## I. INTRODUCTION

Providing credits to people is one of the many activities performed by banks all over the world. Credit scoring requires a lot of empirical analysis of the borrower's past history. Customer's loyalty and ability to neutralize the credit along with interests is thoroughly tested so that the banks can prevent themselves from fraud. This involves collecting a lot of data and creating a big picture out of it. Since it involves a form of "Prediction" of whether the client is worthy of the credit or not, machine learning models can be used to determine the same. There are a lot of models that are being used in the industry for credit scoring like logistic regression, neural networks, decision trees, etc. The following sections focus on one of the many machine learning models used for credit scoring- K - Nearest Neighbours (KNN).

## II. DATASETS

The dataset we have used in the paper to test is from the UC Irivine's machine learning library (UCI-ML)[1]. We use the following databases:

1. Australian credit dataset- 14 attributes and 690 cases.
2. German credit dataset- 20 characters and 1000 states.

Both the datasets have some fields in common such as-loan purpose, salary, credit, age, etc.We wish to determine whether a client is trustworthy or not.

## III. LITERATURE SURVEY

In this section, we discuss the various methodologies that are prevalent in the credit scoring industry. Nowadays, Machine learning algorithms are extensively used for Internet traffic recognition and so many other applications [9-14].The scoring models can be either subjective or statistical [2].

Subjective scoring: This scoring ideology gives a qualitative output on the basis of inputs from the loan officer and the executives involved.

Statistical scoring: Self explanatory, the quality of the worthiness is a function of quantitative metrics that are stored in a database of the organisation providing credit.

There has been a shift of scoring methods prevalent from subjective to statistical over the years, though the former is not completely obsolete. We will now iterate some of the traditional models.

1. Linear Regression

This analysis is practically simple and accurate in explaining various parameters such as ability to repay the credit. Linear regression models are generally fitted using least squares approach. This model involves modelling of a relationship between dependent and independent variables (positive or negative)[3].

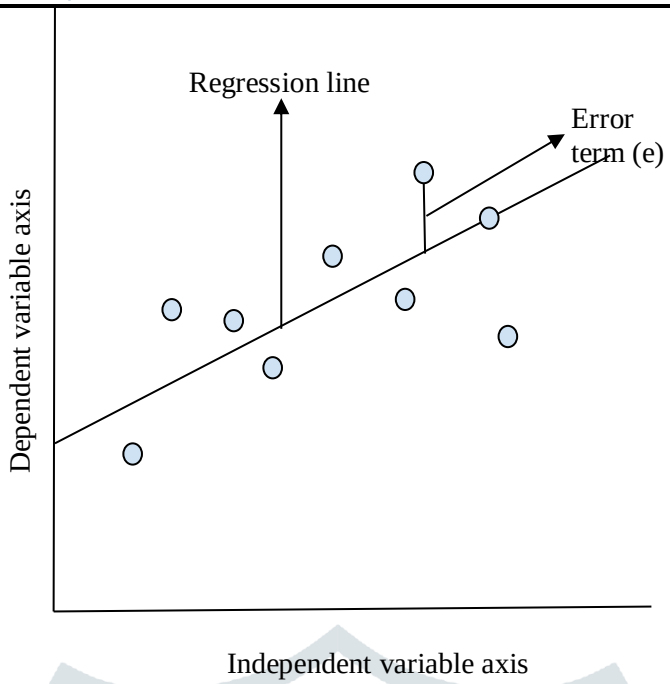


Figure 3.1 Linear regression

Figure 3.1 shows a typical linear regression model. The goal is to reduce the error term(e).

2. Discriminant Analysis

This model is a variant of regression analysis built on categorical data. One of the variations is a label having 2 categories. A label having 2 categories is a variation of this model(in our case- "default" and "non-default")[4]. Altman's model is still prevalent in practical applications. The main aim of this model is to help find a linear combination of features that differentiates multiple classes of objects and was developed by Sir Ronald Fisher

The original Z-score model was [4]:

$Z = 1.2P + 1.4Q + 3.3R + 0.6S + 1.0T$

Where,

P = working capital / total assets ratio

Q = retained earnings / total assets ratio

R = earnings before interest and taxes / total assets ratio

S = market value of equity / total liabilities ratio

T = sales / total assets ratio

3. Judgment-Based Models

These come into play when we have exceptions or when cases are underrepresented in the data, when situations require human intervention and capabilities. Decisions are made on the basis of information available. The information is then arranged in a hierarchical manner and relationships between the data points are made to come to a conclusion. This is the analytic hierarchy process, abbreviated as AHP. This falls under the category of qualitative credit scoring methods. An example of this model is the one by Bana e Costa, Barroso, and Soares[5].

These were some of the traditional methods used in the industry. We now shift our discussion to more recent developments which are gaining momentum. Machine learning and artificial intelligence enables the replacement of human interference and sophistication with computational tools and algorithms.

On a broader picture, the following are the steps involved in any machine learning algorithm:

• Access raw data.

• Collect and manipulate input data accordingly.

• Add features (manual or automatic).

• Selection of useful features according to requirements.

• Application of machine learning algorithms to the training data set.

• Inferences from the results thus obtained.

• Feedback loop enabling the algorithm to learn.

Following are some of the Machine learning paradigms (classified as supervised and unsupervised learning techniques) on which organisations have become heavily dependent.

A. Supervised Learning Techniques

This technique is predictive rather than descriptive, meaning, a value is predicted by this paradigm. It uses dependent and independent variables to predict a value. Some of the techniques are:

1. Decision trees

Decision trees effectively classify on the basis of responses to a specific condition. It is sometimes compared to an if-else loop. The classification can be categories or numeric.

Decision trees are used for:

1. Classification (i.e. default and non default).
2. Prediction of a quantitative measure.

Decision tree is a decision support tool which has a representation of decisions that resembles a tree.

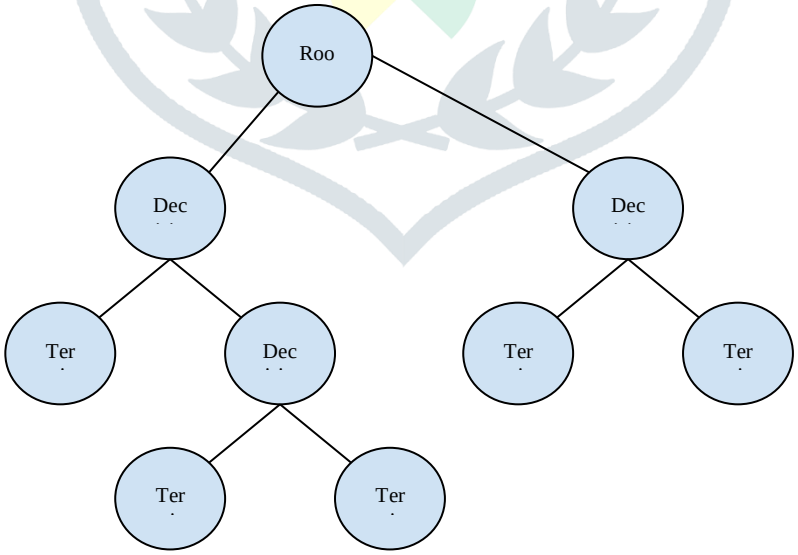


Figure 3.2 A decision tree

Figure 3.2 shows a typical decision tree. Every decision has an output and those outputs are used to make decisions further. This creates a tree-like hierarchy.

2. Random forests

Random forests combine the simplicity of decision trees and flexibility resulting in an improvement in accuracy. Decision trees form the building blocks of random forest. In this method, we train a bunch of decision trees, hence the name forest, and then take a vote amongst the trees. Every tree imparts a single vote. In case of classification, each tree spits out a class prediction. The class getting the most number of votes, becomes the output of the model. In case of regression, a simple average of each individual tree's prediction becomes the output. The key idea harnessed here is- there is wisdom in the crowd. Insights drawn from a group of models have greater accuracy as compared to a single model. Figure 3.3 depicts a typical random forest.

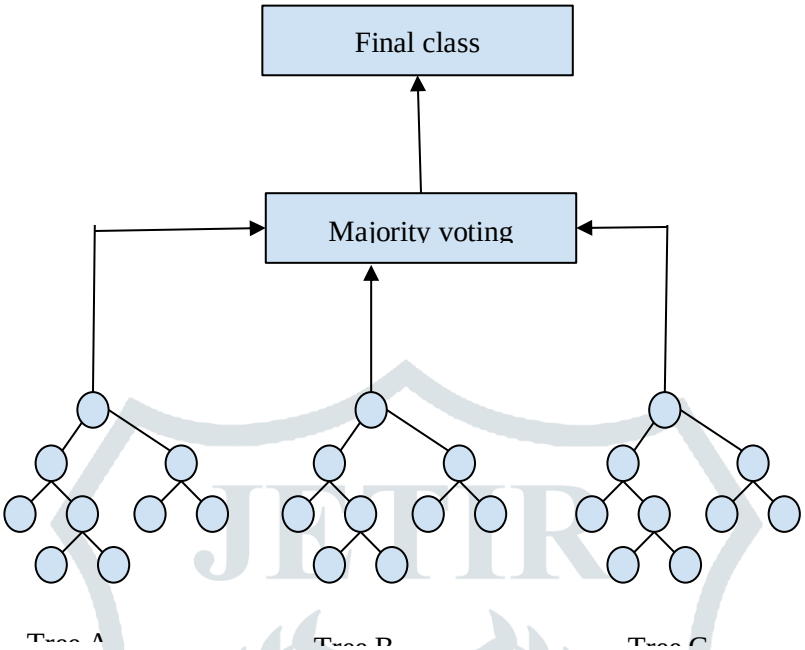


Figure 3.3 Random forests

3. Deep Neural Networks

Neural networks train from the input data by recognizing patterns. They then predict output values on new data feeds. They are made up of neurons arranged in the form of layers. Refer to figure 3.4.

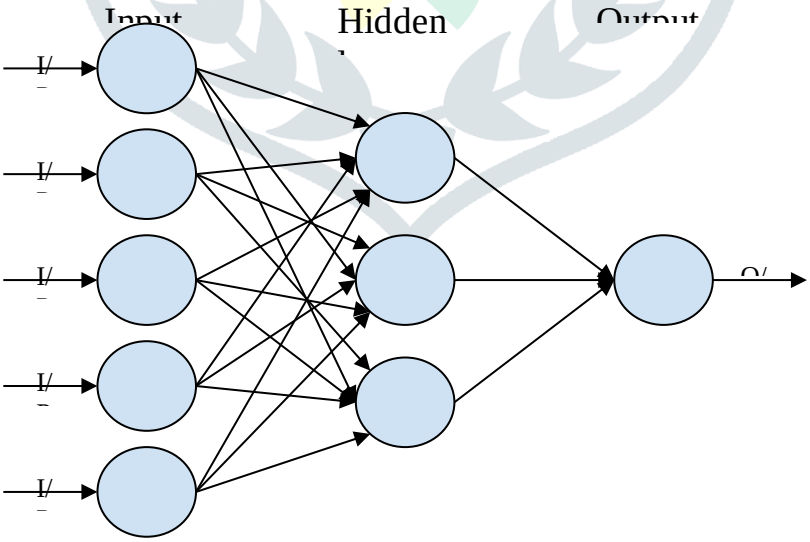


Figure 3.4 A neural network

The input layer receives the input data. The calculations are performed by the hidden layer(s) and the output is processed by the output layer. The neurons are connected by channels which have their own numerical values called weights. These weights are multiplied to the inputs. Bias is a numerical value associated with neurons and is added to the input values. Activation of a neuron is determined by the activation function. The data is passed on to further layers depending on the activation of neurons. This data flow in a neural network is called forward propagation. The output value is then compared to the actual values in order to "learn".

Changes to the network are made on the basis of deviations of predicted value from the actual value. This "learned network" is then used for a new data set, in our case credit analysis.

B. Unsupervised Learning Techniques

These learning techniques, rather than predicting a probability or an unknown value, explore the characteristics and patterns in the input data. We discuss a few prevalent unsupervised learning techniques.

1. Clustering

Unlike supervised learning techniques, clustering doesn't analyze class labeled data. It instead creates groups that have the most similar characteristics and patterns. These algorithms are descriptive. Suppose we wish to find borrowers that are easy to assess. Clustering algorithms can be implemented to find groups that have the same set of desirable characteristics. We now differentiate between K- clustering and hierarchical clustering:

K Clustering: This model aims to divide the data in K different clusters. Centroids are then placed in the Euclidean space. Data points are then assigned to the cluster closest to them in terms of Euclidean distance.

Hierarchical clustering: This clustering process begins by the assignment of data points as their own cluster. Then 2 nearest data points are combined to form a single cluster based on Euclidean distance.

Statistical scoring has evolved from predicting models based on empirical data to systems using machine learning algorithms. In this report, we explore K- nearest neighbours.

## IV. METHODOLOGY

This algorithm (KNN) does not have any assumptions in the distribution of data and is thus a type of non parametric lazy learning algorithm. This is very helpful, as in the real world, many inactive details do not comply with the general existing theories[6]. Non-parametric algorithms like KNN are helpful here. Outcomes are engendered on the basis of set training data in KNN. The primary idea of the model is- whenever a new predictive point (k) is discovered, its neighbors are selected from the training data set. Post which, the prediction of a new point can be a measure of the value of its nearest neighbors. Minkowski distance is the measure for nearest distance where q = 2 is often used to measure the Euclidean distance, or q = 1 for Manhattan distance measurement.

$$d(x,y) = \left(\sum_{i=1}^{k} (|x_i - y_i|)^q\right)^{\frac{1}{q}}$$

The data is divided into training and testing data sections using an 80% - 20% proportion post scaling and normalizing the data (80%- training and 20%-testing)[7]. Consider K = 5 and the Euclidean distance for the data sets. The accuracy (calculated using k-fold cross-validation) is given in table 3.1. The results, after tuning use:

1. q = {1, 2, 3, 4, 5, 6, 7}
2. K = {1, 3, 5, 7, 9, 11, 13, 15}

For German data:

1. K = 7
2. q = 2 (Euclidean distance)

For Australian data we have:

1. K = 13
2. q = 1 (Manhattan distance)

## V. PERFORMANCE EVALUATION PARAMETERS FOR DIFFERENT ML TECHNIQUES

Sensitivity = TP / (TP + FN)

Specificity = TN / (FP + TN)

Precision = TP / (TP + FP)

Accuracy = (TP + TN) / (P + N)

F1 Score = 2TP / (2TP + FP + FN)

Where:

TP- True positive

TN- True negative

FP- False positive

FN- False negative

## VI. EXPERIMENTAL RESULTS

Confusion matrix of both the data sets are given by Table 5.1 and Table 5.2.

Table 5.1: German data confusion matrix.

| | **True Positive** | **True Negative** |
|---|---|---|
| **Predicted Positive** | TP=123 | FP=28 |
| **Predicted Negative** | FN=17 | TN=32 |

Table 5.2: Australian data confusion matrix

| | **True Positive** | **True Negative** |
|---|---|---|
| **Predicted Positive** | TP=48 | FP=8 |
| **Predicted Negative** | FN=12 | TN=70 |

From confusion matrices, it is observed that KNN engenders less FP for the Australian dataset. False-negatives suggest a customer is non-credible instead of credible. FPs are undesirable[8] and, thus, the model fits better on our Australian dataset.

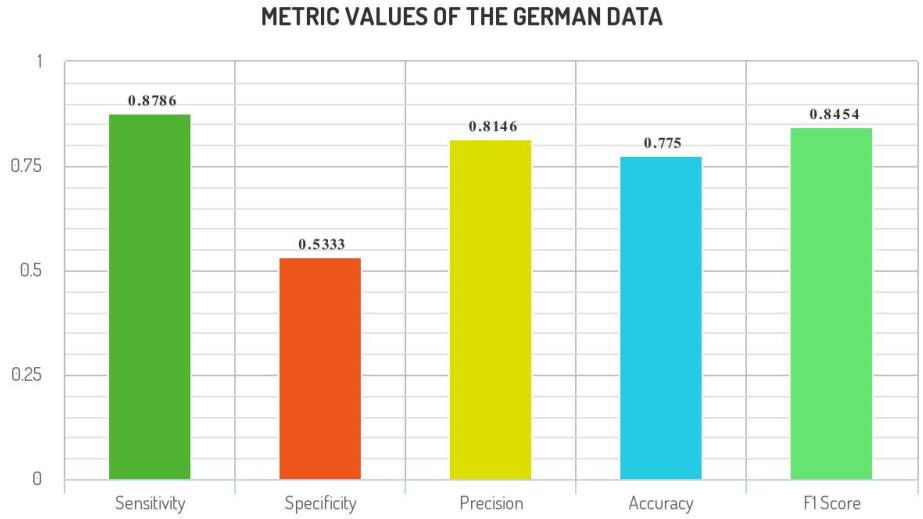


Figure 5.1 Metric values of the German data

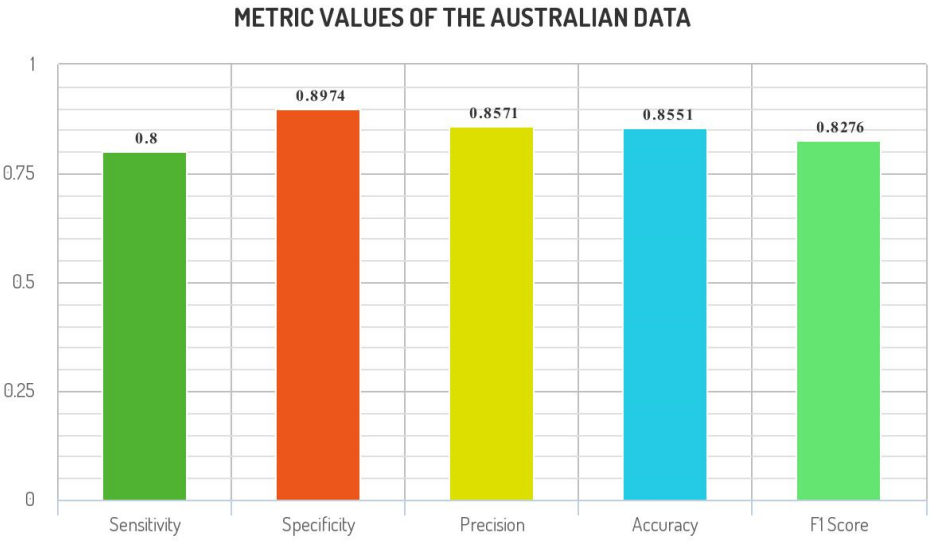


Figure 5.2 Metric values of the Australian data

Figure 5.1 and 5.2 demarcate the various metrics for the German and Australian dataset respectively. The KNN model has a greater accuracy and precision for the Australian dataset.

## VII. Conclusion

In conclusion, machine learning proves to be a powerful toolbox for financial analysts to make predictions and discover patterns in the data with confidence. Many different models and validation techniques exist to augment data mining and decision-making processes. KNN is an efficient model which can be used by banks for credit scoring reliably. It is not easy to boil down to a single machine learning technique. In fact, as a data scientist and/or financial expert, it is more beneficial to harness the strengths of different methods and combine them to make better business decisions.

## VIII. ACKNOWLEDGMENT

I would like to express my gratitude towards my mentor, Dr. R. C. Jaiswal for guiding me through the research. His contribution in providing the research guidance required for this was immense and his insight in the topic gave the paper what was needed.

## REFERENCES

[1] UC Irvine Machine Learning Repository Australian Credit Data, UC Irvine Machine Learning Repository German Credit Data.

[2] Schreiner, Mark. 2003. "Scoring: The Next Breakthrough in Microcredit?" Occasional Paper 7. Washington, D.C.: CGAP. https://www.cgap.org/research/publication/scoring-next-breakthrough-microcredit

[3] "Linear regression", *Wikipedia*, Wikimedia Foundation,06 May 2021, https://en.wikipedia.org/w/index.php?title=Linear_regression&oldid=1021750493

[4] Altman, Edward. 1968 "Financial Ratios, Discriminant Analysis, and the Prediction of Corporate Bankruptcy." Journal of Finance 23 (4): 589–609.

[5] Bana e Costa, Carlos A. E., Luís A. Barroso, João O. Soares.2002. "Qualitative Modelling of Credit Scoring: A Case Study in Banking." European Research Studies 5 (1–2): 37–51.

[6] M A Mukid et al 2018 J. Phys.: Conf. Ser. 1025 012114.

[7] Thanawala, Dhruv Dhanesh, "CREDIT RISK ANALYSIS USING MACHINE LEARNING AND NEURAL NETWORKS", Open Access Master's Report, Michigan Technological University, 2019.

[8] Wynants, L., van Smeden, M., McLernon, D.J. *et al.* Three myths about risk thresholds for prediction models. *BMC Med* 17, 192 (2019). https://doi.org/10.1186/s12916-019-1425-3.

[9] Jaiswal R.C. and Lokhande S.D, "A Novel Approach for Real Time Internet Traffic Classification", ICTACT Journal on Communication Technology, September 2015, volume: 06, issue: 03, pp. 1160-1166.(Print: ISSN: 0976-0091, Online ISSN:2229-6948 (Impact Factor: 0.789 in 2015).

[10] Jaiswal R.C. and Lokhande S.D "Measurement, Modeling and Analysis of HTTP Web Traffic", IMCIET-International Multi Conference on Innovations in Engineering and Technology-ICCC-International Conference on Communication and Computing -2014, PP-242-258, ISBN:9789351072690, VVIT, Bangalore.

[11] Jaiswal R.C. and Lokhande S.D, "Comparative Analysis using Bagging, LogitBoost and Rotation Forest Machine Learning Algorithms for Real Time Internet Traffic Classification", IMCIP-International Multi Conference on Information Processing –ICDMW- International Conference on Data Mining and Warehousing-2014, PP113-124, ISBN: 9789351072539, University Visvesvaraya College of Engg. Department of Computer Science and Engineering Bangalore University, Bangalore.

[12] Jaiswal R.C. and Lokhande S.D, "Statistical Features Processing Based Real Time Internet Traffic Recognition and Comparative Study of Six Machine Learning Techniques", IMCIP- International Multi Conference on Information Processing-(ICCN- International Conference on Communication Networks-2014, PP-120-129, ISBN: 9789351072515, University Visvesvaraya College of Engg. Department of Computer Science and Engineering Bangalore University, Bangalore.

[13] Jaiswal R.C. and Lokhande S.D, "Analysis of Early Traffic Processing and Comparison of Machine Learning Algorithms for Real Time Internet Traffic Identification Using Statistical Approach ", ICACNI-2014-International Conference on Advanced Computing, Networking, and Informatics),Kolkata, India,DOI: 10.1007/978-3-319-07350-7_64, Volume 28 of the book series Smart Innovation, Systems and Technologies (SIST),Page:577-587.

[14] Jaiswal R.C. and Lokhande S.D, "Machine Learning Based Internet Traffic Recognition with Statistical Approach", INDICON-2013-IIT BOMBAY IEEE CONFERENCE. INSPEC Accession Number: 14062512, DOI: 10.1109/INDCON.2013.6726074.