Smart Intelligent Computing and Communication Technology
V.D. Ambeth Kumar et al. (Eds.)

doi:10.3233/APC210004

Performance Analysis of K-Nearest Neighbor Classification Algorithms for Bank Loan Sectors

Hemachandran K[a,1], Preetha Mary George[b], Raul V. Rodriguez[c], Raviraj M. Kulkarni[d] and Sourav Roy[e]

[a,1] *Professor, Woxsen School of Business, Woxsen University, Hyderabad, India*
[b] *Professor, Dept of Physics, Dr. MGR Educational & Research Institute, India*
[c] *Dean, Woxsen School of Business, Woxsen University, Hyderabad, India*
[d] *Professor, Dept of ECE, KLS Gogte Institute of Technology, Karnataka, India*
[e] *Postdoctoral Research Fellow, Xi'an Jiaotong University, China*

Abstract. An attempt has been made to develop an algorithm for banks to check the credibility of borrowers to avoid nonperformance assets. People move towards different banks for loan purpose to fulfil their financial needs. Approaching bank for loan is increasing day by day mainly for child marriage, education, agriculture, business, home loan etc. Some people take the loan and they won't pay back in time or some will move out of the country without any intimation, so that bank will go in loss. Even now in covid-19 pandemic many industries were closed but they need to give salary to the employees, need to pay rent and electricity bills too for that they will approach bank for loan. For all these cases bank first need to analyse their Credit Information Bureau India Limited score and check whether they had done loan repayments in appropriate time or not. In the present work the effectiveness of K nearest neighbor algorithm were analysed. This research were carried out using python. The accuracy of this classifier is analysed using following metrics such as Jaccard index, F1-score and LogLoss. This helps to find the potential of the customer which is much higher than the data mining classification algorithm and thus it helps in sanctioning loans.

Keywords. F1-score, Jaccard Index, K-Nearest Neighbor Algorithm, Log loss

1. Introduction

For the past two decades, there was a rapid increase in demand of sanctioning loans. In the present times, decision to approve loan depended on human decision to gauge the default risk. The increase demand of credit, directed to a leap in the use of official and objective methods of credit-scoring. Credit scoring provides to decide whether to credit bank loan to the applicant or not [1,2]. The intention of credit-scoring is to aid credit suppliers to manage and quantify the monetary risk. By analyzing through the

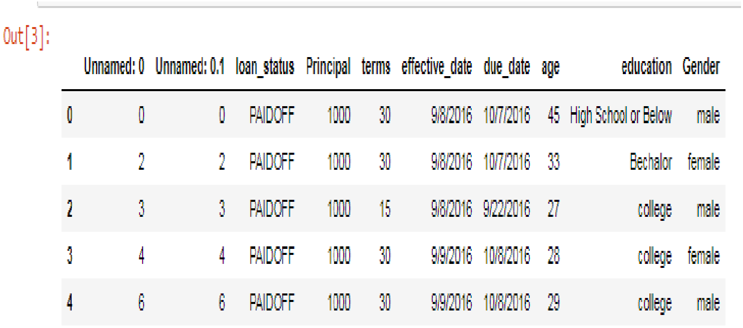
[1] Hemachandran K, Professor, Woxsen School of Business, Woxsen University, Hyderabad, India. E-mail:hemachandran.k@woxsen.edu.in

method of credit scoring the applicants of loan can take better lending decisions quickly. It provides a system for the creditors to assign loan to the applicants who are either a "good canditure" or a "bad canditure".

The implementation of K-Nearest Neighbor, is a criterion in recognizing patterns and in non-parametric statistics to the problem of credit scoring [4-6]. The K-NN strategy helps in making a decision about positive or negative danger prospect of the candidate. Also, once more, it sorts great and terrible extent among the most closest comparative k focuses in the preparation tests. An acceptable distance can access the point metric similarity [3].

2. Build Data Model

To test this algorithm, the historical data was retrieved from the website of UCI. The KNN algorithm to find on client loans has a definite effect, to help and guide the bank to take a sure decision to reduce the financial risk [7-9]. In Fig.1 it shows 346 records with 10 attributes and to represents the client active-index score which is worthy or ruthless we used 1 or 0. If it gives yes, it tells the customer is healthier and can consider for the sanction of the client's loan, and on the other hand, if it gives no, then that customer is not eligible [16-19].



Out[3]:

	Unnamed: 0	Unnamed: 0.1	loan_status	Principal	terms	effective_date	due_date	age	education	Gender
0	0	0	PAIDOFF	1000	30	9/8/2016	10/7/2016	45	High School or Below	male
1	2	2	PAIDOFF	1000	30	9/8/2016	10/7/2016	33	Bechalor	female
2	3	3	PAIDOFF	1000	15	9/8/2016	9/22/2016	27	college	male
3	4	4	PAIDOFF	1000	30	9/9/2016	10/8/2016	28	college	female
4	6	6	PAIDOFF	1000	30	9/9/2016	10/8/2016	29	college	male

Figure 1. Dataset

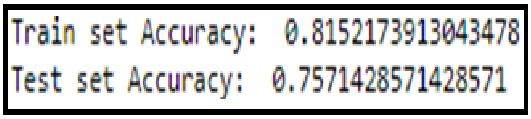
After analyzing the data, how many paid off the previous loan and whoever who made default were also analyzed. Two attributes gender and age have been selected based on that data visualization and pre-processing have been carried out and then convert the categories into numerical values and finally data has been normalized. If the data is not huge, we can do the data cleaning or delete attributes which are not necessary[20-26].

3. K-NN Algorithms

In this method, a set of training sets are taken. Euclidean distance metric estimates the closeness between each preparation model and another model. In this example, highlights of k-NN calculation [10], won't be weighted and they are dealt with identically. The procedure finds the nearest k-value to the next example. The new example allots the class wherein most of the k neighbor models have a place. Neighbors will be weighed by the reverse of their distance during casting a ballot[11, 13]. The number of neighbors is a deciding element. A reasonable k is to be exactly decided to smoothen the noise. The Selection of suitable nearest-distant points has a significant influence in K-NN strategy. The Euclidean distance is given by

$$D(X,Y) =\{(X-Y)^T(X-Y)\}^{1/2} \tag{1}$$

Based on the train set and test set. The data has been trained and it has been tested using test set. So that the accuracy evaluation has been made and shown in figure 2.



```
Train set Accuracy:  0.8152173913043478
Test set Accuracy:  0.7571428571428571
```

Figure 2. Accuracy Evaluation

4. Performance Metrics

Assessment metrics clarify the performance of a model. On the off chance that we have a recorded dataset of client agitates, in the wake of preparing the model we need to ascertain the exactness utilizing the test set. We finish the assessment set to the model to discover anticipated labels [14]. There are three distinctive model evaluation metrics they are Jaccard, F1 score and logloss.

4.1. Jaccard Similarity Score

Jaccard Index additionally called a Jaccard comparability coefficient. For instance, if y shows the genuine names of the agitate dataset and $\hat{y}$shows the anticipated qualities by our classifier [12, 22]. Precision = True '+'ve / (True '+'ve + False '+'ve). Furthermore, Recall is the genuine positive rate. It is characterized as: Recall = True '+'ve / (True '+'ve + False '-'ve). In this way, we can compute the exactness and review of each class

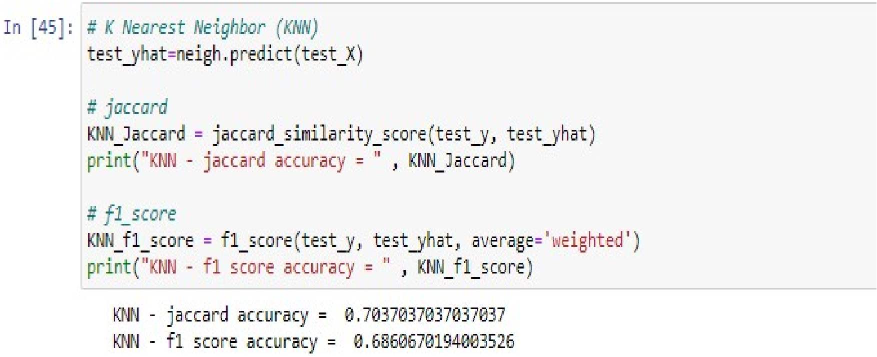
4.2. F1 Score

We can compute the F1 scores for each label, in view of the precision and recall of that label. The F1 score is determined dependent on the precision and recall of each class. The F1-score arrives at its ideal incentive at the very least at 0. It is a generally excellent approach to show that grouping has a decent review and accuracy esteems.

4.3. Log Loss

We can utilize the Log loss in situations where the result of the classifier is a class likelihood and not a class name like in instances of logistic regression models. Log loss quantifies the exhibition of a model where the anticipated result is a likelihood esteem somewhere in the range of 0 and 1.

Based on the K- Nearest Neighbor accuracy evaluation the three parameter metrics has been evaluated. Whereas log loss value for this model is not applicable.



```
In [45]: # K Nearest Neighbor (KNN)
         test_yhat=neigh.predict(test_X)

         # jaccard
         KNN_Jaccard = jaccard_similarity_score(test_y, test_yhat)
         print("KNN - jaccard accuracy = " , KNN_Jaccard)

         # f1_score
         KNN_f1_score = f1_score(test_y, test_yhat, average='weighted')
         print("KNN - f1 score accuracy = " , KNN_f1_score)

           KNN - jaccard accuracy =  0.7037037037037037
           KNN - f1 score accuracy =  0.6860670194003526
```

Figure 3. Performance metrics accuracy

5. Result & Conclusion

In this paper, the performance of K Nearest Neighbor classification algorithms were analyses based on three performance metrics Jaccard, F1 score and logloss. These proposed algorithms are used to predict the loan repayment capability behavior of a customer in a cost effective way. The bank officers need to determine whether to approve loan for the applicant or not. This proposed methodology will protect the bank from further misuse, fraud applications etc by identifying the customers whose repayment capability status is risky especially in the banking sector. This research gives that the classification accuracy of KNN is 70%.

Table 1. Accuracy Result

Algorithm	Jaccard	F1 Score	Log Loss
K-Nearest Neighbor	0.703704	0.686067	N/A

References

[1] Arun, K.,Ishan, G., Sanmeet, K., Loan approval prediction based on machine learning approch. IOSR Journal of Computer Engineering, NCRTCSIT. 2016, pp. 18-21.

[2] Aafer Y., Du W., Yin H. DroidAPIMiner. Mining API-Level Features for Robust Malware Detection in Android. Security and Privacy in Communication Networks.2013.pp 86-103.

[3] Hand, D. J., &Vinciotti, V.,Choosing k for two-class nearest neighbor classifiers with unbalanced classes. PatternRecognition Letters.2003. 24, (9-10), pp.1555-1562.

[4] HanumanthaRao, K., Srinivas, G., Damodhar, A., &Vikar Krishna, M., Implementation of Anomaly Detection Technique Using Machine Learning Algorithms. International Journal of Computer Science and Telecommunications. 2011. 2(3), pp.25-30.

[5] Haodong Zhu., &Zhong Yong., Based on improved ID3 information gain feature selection method.Computer engineering.2010. 36(8).

[6] He Y., Han J., Zeng S., Classification Algorithm based on Improved ID3 in Bank Loan Application. Information Engineering and Applications.2012. Pp.1124-1130.

[7] Houxing You., A Knowledge Management Approach for Real-time Business Intelligence. 2nd International Workshop on Intelligent Systems and Applications, DOI:10.1109/IWISA.2010.5473385. 2010.

[8] Huang, L., Zhou, CG., Zhou, Yu-qin., & Wang, Zhe., Research on DataMining Algorithms for Automotive Customers' Behavior Prediction Problem. In:2008 Seventh International Conference on Machine Learning and Applications.2008. DOI: 10.1109/ICMLA.2008.23.

[9] J.M. Chambers., Computational methods for data analysis. Applied Statistics, Wiley.1977. 1(2), pp.1–10.

[10] Li Xia, ID3 classification algorithm application in bank customer's erosion. Journal Computer technology and development,.2009.19(3).

[11] M. J. Islam., Q. M. J. Wu., M. Ahmadi., & M. A. Sid-Ahmed, Investigating the Performance of Naive-Bayes Classifiers and K- Nearest Neighbor Classifiers. In: International Conference on Convergence Information Technology (ICCIT 2007).2007. pp. 1541-1546.

[12] Marinakis, Y., Marinaki, M., Doumpos, M. et al., Optimization of nearest neighbor classifiers via metaheuristic algorithms for credit risk assessment. Journal of Global Optimization.2008. 42, pp.279–293.

[13] Ram, B., & Rama Satish, A., Improved of K-Nearest Neighbor Techniques in Credit Scoring. International Journal For Development of Computer Science & Techno logy.2013. 1(2).

[14] Sahay, B,S., &Ranjan, J., Real time business intelligence in supply chain analytics. Information Management & Computer Security.2008. 16(1), pp. 28-48.

[15] Ambeth Kumar.V.D, S. Sharmila, Abhishek Kumar, A. K. Bashir, Mamoon Rashid, Sachin Kumar Gupta &Waleed S. Alnumay .A novel solution for finding postpartum haemorrhage using fuzzy neural techniques. Neural Computing and Applications (2021) (https://doi.org/10.1007/s00521-020-05683-z)

[16] AnkitKumar,VijayakumarVaradarajan,AbhishekKumar, PankajDadheech, SurendraSinghChoudhary, V.D. AmbethKumar, B.K.Panigrahi, KalyanaC.Veluvolug . Black hole attack detection in vehicular ad-hoc network using secure AODV routing algorithm .Microprocessors and Microsystems, In Press,(https://doi.org/10.1016/j.micpro.2020.103352)

[17] Ambeth Kumar.V.D . A Cognitive Model for Adopting ITIL Framework to Improve IT Services in Indian IT Industries.Journal of Intelligent Fuzzy Systems. (DOI: 10.3233/JIFS-189131)(Accepted - Inpress)

[18] Ambeth Kumar.V.D .Efficient Data Transfer in Edge Envisioned Environment using Artificial Intelligence based Edge Node Algorithm. Transactions on Emerging Telecommunications Technologies (Accepted - Inpress)(DOI: 10.1002/ett.4110)

[19] AmbethKumar.V.D , Malathi.S ,AbhishekKumar, Prakash M and Kalyana C. Veluvolu .Active Volume Control in Smart Phones Based on User Activity and Ambient Noise .Sensors 2020, 20(15), 4117; https://doi.org/10.3390/s20154117

[20] Ambeth Kumar S. Malathi R. Venkatesan K Ramalakshmi, Weiping Ding, Abhishek Kumar .Exploration of an innovative geometric parameter based on performance enhancement for foot print recognition. Journal of Intelligent and Fuzzy System , vol. 38, no. 2, pp. 2181-2196, 2020.

[21] V.D.Ambeth Kumar, Dr.S.Malathi, V.D.Ashok Kumar (2015).Performance Improvement Using an Automation System for Segmentation of Multiple Parametric Features Based on Human Footprint. for the Journal of Electrical Engineering & Technology (JEET) , vol. 10, no. 4, pp.1815-1821 , 2015. [http://dx.doi.org/10.5370/JEET.2015.10.4.1815]

[22] V.D.Ambeth Kumar and M.Ramakrishan (2013).Temple and Maternity Ward Security using FPRS. in the month of May for the Journal of Electrical Engineering & Technology (JEET) ,Vol. 8, No. 3, PP: 633-637. [http://dx.doi.org/10.5370/JEET.2013.8.3.633]

[23] V.D.Ambeth Kumar and M.Ramakrishan (2013) .A Comparative Study of Fuzzy Evolutionary Techniques for Footprint Recognition and Performance Improvement using Wavelet based Fuzzy Neural Network. for the International Journal of Computer Applications in Technology (IJCAT-Inderscience), Vol.48, No.2,pp.95 – 105, [DOI: ttp://dx.doi.org/10.1504/IJCAT.2013.056016]

[24] Navadeepika.K.M.R, Dr.V.D.Ambeth Kumar.Online Conversion of Tamil Language into Tactile for Visual Impaired People.International Journal of Advanced Science and Technology Vol. 29, No. 5, pp. 3401 - 3407, (2020).

[25] Rajendiran.M.,Subhashini.P. Energy efficient cluster based leader selection protocol for wireless ad hoc network. International Journal of Computer Science and Engineering Communications (IJCSEC), 3(3), 2015.

[26] Vimala.S, Thilagavathi.S, Valarmathi.K, Priya.R, Sathya.S .Massive Data Processing Using Map Reduce Aggregation to make Digitized India. Advances in Engineering Research, vol. 142, Pp.88-92, 2019