LITERATURE SURVEY

DOMAIN: **Applied Data Science**

TOPIC: **Smart Lender - Applicant Credibility Prediction For Loan Approval**

TEAM ID: **PNT2022TMID00272**

TEAM MEMBERS:

JAGADEESH. A (Team Leader)

ARUL SAMUEL. I

BADRI NARAYAN. J

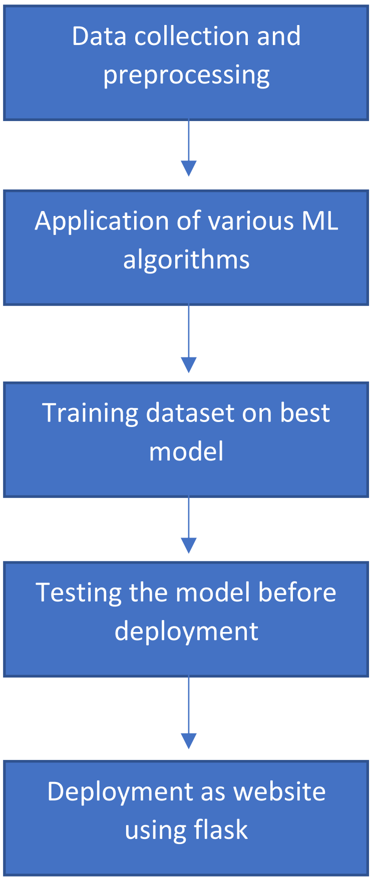
BALAKUMARAN. M

ABSTRACT

The enhancement in the banking sector lots of people are applying for bank loans but the bank has its limited assets which it has to grant to limited people only, so finding out to whom the loan can be granted which will be a safer option for the bank is a typical process. So in this paper we try to reduce this risk factor behind selecting the safe person so as to save lots of bank efforts and assets. This is done by mining the Big Data of the previous records of the people to whom the loan was granted before and on the basis of these records/experiences the machine was trained using the machine learning model which give the most accurate result. The main objective of this paper is to predict whether assigning the loan to particular person will be safe or not.

This paper is divided into five sections:

1) Data Collection
2) Comparison of machine learning models on collected data.
3) Training of system on most promising model
4) Testing
5) Deployment using Flask


Data collection and preprocessing
Application of various ML algorithms
Training dataset on best model
Testing the model before deployment
Deployment as website using flask

1. AN APPROACH FOR PREDICTION OF LOAN APPROVAL USING MACHINE LEARNING ALGORITHM

AUTHOR: Mohammad Ahmad Sheikh, Amit Kumar Goel, Tapas Kumar

ABSTRACT:

In our banking system, banks have many products to sell but main source of income of any banks is on its credit line. So, they can earn from interest of those loans which they credit. A bank’s profit or a loss depends to a large extent on loans i.e., whether the customers are paying back the loan or defaulting. By predicting the loan defaulters, the bank can reduce its Non-Performing Assets. This makes the study of this phenomenon very important. Previous research in this era has shown that there are so many methods to study the problem of controlling loan default. But as the right predictions are very important for the maximization of profits, it is essential to study the nature of the different methods and their comparison. A very important approach in predictive analytic is used to study the problem of predicting loan defaulters: The Logistic regression model. The data is collected from the Kaggle for studying and prediction. Logistic Regression models have been performed and the different measures of performances are computed. The models are compared on the basis of the performance measures such as sensitivity and specificity. The final results have shown that the model produce different results. Model is marginally better because it includes variables (personal attributes of customer like age, purpose, credit history, credit amount, credit duration, etc.) other than checking account information (which shows wealth of a customer) that should be taken into account to calculate the probability of default on loan correctly. Therefore, by using a logistic regression approach, the right customers to be targeted for granting loan can be easily detected by evaluating their likelihood of default on loan. The model concludes that a bank should not only target the rich customers for granting loan but it should assess the other attributes of a customer as well which play a very important part in credit granting decisions and predicting the loan defaulters.

Keywords: loan, outlier, Prediction, component, Over-fitting, Transform.

2. LOAN DEFAULT FORECASTING USING DATA MINING

AUTHOR: Bhoomi Patel , Harshal Patil , Jovita Hembram , Shree Jaswal

ABSTRACT:

Estimation or assessment of default on a debt is a crucial process that should be carried out by banks to help them to assess if a loan applicant can be a defaulter at a later phase so that they process the application and decide whether to approve the loan or not. The conclusion derived from such assessments helps banks and other financial institutions to lessen their losses and eventually increase the number of credits. Hence, it becomes vital to construct a model that will take into account the different aspects of an applicant and derive a result regarding the concerned applicant. All available means to loan the money from their illicit activities are used for criminal activities in today's technology-based realm. The increasing number of bad debts resulting from commercial banks' loans reflects the growing problem of distraught banks within the economic system. We have used data mining algorithms to predict the likely defaulters from a datasets that contains information about home loan applications, thereby helping the banks for making better decisions in the future.

Keywords: loan, credit, prediction, data mining

3. PREDICTION OF LOAN STATUS IN COMMERCIAL BANK USING MACHINE LEARNING CLASSIFIER

AUTHOR: G. Arutjothi,Dr. C. Senthamarai

ABSTRACT:

Banking Industry always needs a more accurate predictive modeling system for many issues. Predicting credit defaulters is a difficult task for the banking industry. The loan status is one of the quality indicators of the loan. It doesn't show everything immediately, but it is a first step of the loan lending process. The loan status is used for creating a credit scoring model. The credit scoring model is used for accurate analysis of credit data to find defaulters and valid customers. The objective of this paper is to create a credit scoring model for credit data. Various machine learning techniques are used to develop the financial credit scoring model. In this paper, we propose a machine learning classifier-based analysis model for credit data. We use the combination of Min-Max normalization and K

Nearest Neighbor (K-NN) classifier. The objective is implemented using the software package R tool. This proposed model provides the important information with the highest accuracy. It is used to predict the loan status in commercial banks using machine learning classifier.

Keywords: Credit Scoring; K-NN; Loan status; Loan Lending Process; Min-Max Normalization

4. OVERDUE PREDICTION OF BANK LOANS BASED ON LSTM-SVM

AUTHOR: Xin Li, Xianzhong Long, Guozi Sun, Geng Yang, and Huakang Li

ABSTRACT:

In the aspect of bank loans, the accuracy of traditional user loan risk prediction models, such as KNN, Bayesian,DNN, are not benefit from the data growth. This article is based on the work of Overdue Prediction of Bank Loans Based on Deep Neural Network. And we propose to analyze the dynamic behavior of users by LSTM algorithm, and use the SVM algorithm to analyze the user's static data to solve the current prediction problems. This article uses users basic information, bank records, user browsing behavior, credit card billing records, and loan time information to evaluate whether users are delinquent. These static data are the basic input for SVM. For LSTM model, we extract user's recent transaction type from browsing behavior as input to LSTM, to predict the probability of users' overdue behavior. Finally, we calculate the average of the two algorithms as the final result. From the experimental results, this LSTM-SVM model shows a great improvement than traditional algorithms.

Keywords: Overdue Prediction, Bank Loans, LSTM, SVM

5. PREDICTION DEFAULTS FOR NETWORKED-GUARANTEE LOANS

AUTHOR: Dawei Cheng, Zhibin Niu†, Yi Tu and Liqing Zhang

ABSTRACT: Networked-guarantee loans may cause the systemic risk related concern of the government and banks in China. The prediction of default of enterprise loans is a typical extremely imbalanced prediction problem, and the networked-guarantee make this problem more difficult to

solve. Since the guaranteed loan is a debt obligation promise, if one enterprise in the guarantee network falls into a financial crisis, the debt risk may spread like a virus across the guarantee network, even lead to a systemic financial crisis. In this paper, we propose an imbalanced network risk diffusion model to forecast the enterprise default risk in a short future. Positive weighted k-nearest neighbors (pwkNN) algorithm is developed for the stand-alone case – when there is no default contagious; then a data-driven default diffusion model is integrated to further improve the prediction accuracy. We perform the empirical study on a real-world three years loan record from a major commercial bank. The results show that our proposed method outperforms conventional credit risk methods in terms of AUC. In summary, our quantitative risk evaluation model shows promising prediction performance on real-world data, which could be useful to both regulators and stakeholders.

6. PERSONAL CREDIT RATING USING ARTIFICIAL INTELLIGENCE TECHNOLOGY FOR THE NATIONAL STUDENT LOANS

AUTHOR: Jian HU, Zibo, China

ABSTRACT: National student loans have the general features of commercial loans, and are a financial credit services provided by commercial banks. But the general personal credit rating assessment system of commercial bank can not make the correct credit rating because the lender, college students, have no credit history. To avoid the credit risk, a rational credit assessment system must to be established for college Students. With the self-learning, self-organizing, adaptive and nonlinear dynamic handling characteristics of Artificial Neural Network, a Back Propagation neural network was developed to evaluate the credit rating about a college student. Several samples, which were provided by a bank, were used for network training and testing by MATLAB. The maximum value of the error between the prediction value of the network and actual value is only 2.92that the algorithm developed is fairly efficient for the assessment about the college student’s personal credit situation

Keywords: National Student Loans, credit rating, Artificial Intelligence, Back Propagation neural network.

7. DYNAMIC LOAN SERVICE MONITORING USING SEGMENTED HIDDEN MARKOV MODELS

AUTHOR: Haengju Lee, Nathan Gnanasambandam, Raj Minhas, and Shi Zha

ABSTRACT: We describe how to apply Hidden Markov Model (HMM) to automate the loan service monitoring process. To predict the probability of defaulting in the near future, we build a statistical model of HMM from borrowers' historical payment data. The predicted probability is dynamic in a sense that the probability keeps changing as new realized data is added to the current historical data. The time series sequence data is obtained from the composite information of the loan status and days delinquent on each month. In the training stage, various HMMs are trained: one is paid HMM and the others are defaulted HMMs. We show that more accurate monitoring can be achieved by segmenting the defaulted data and training them separately (i.e., segmented HMM method) than by training a single defaulted HMM (i.e., simple HMM method). In the prediction stage, for each active loan, we apply the following two steps:

1) Classification of the loan
2) Calculation of the default probability over a prospective time period.

Finally, the monitoring system sends a signal if the probability is greater than a per specified threshold. We also explore how to select the optimal threshold level using precision and recall analysis.

Keywords: loan default prediction, sequences and sequential data analysis, Hidden Markov Model, default monitoring system.

8. AZURE ML BASED ANALYSIS AND PREDICTION LOAN BORROWERS CREDITWORTHY

AUTHOR: Khaldoon Alshouiliy, Ali Al Ghamdi, Dharma P Agrawal

ABSTRACT: In the era of big data, it would be beneficial for lenders to use modern technology such as machine learning this research, our aim is to analyze Lending Club datasets to make it well understood datasets features. Then, we upload our clean datasets to Microsoft Azure machine learning (Azure ML) platform to use for building our model. Which aims to predict whether the customers are going to pay back their loans or not. This model predicts the loan status going to be default or fully paid.

Moreover, the Lending Club datasets we used in this work is gathered from 2007 to 2018 used accept loans. We used Azure ML platform with Two Jungle algorithm and the Two Decision tree. Thereafter, we assess their performance (algorithms) in terms of Accuracy, Precision, Recall, F1 and AUC. Finally, we compare our work with other researchers and our work shows a good result compared to others.

Keywords: machine learning algorithms, P2P lending; credit scoring, big data, data analytic, Azure ML, jungle algorithm, two decision tree

9. ANALYSIS OF FEATURE SELECTION AND EXTRACTION ALGORITHM FOR LOAN DATA: A BIG DATA APPROACH

AUTHOR: Girija Attigeri, Manohara Pai M M*, Radhika M Pai

ABSTRACT:

Fraudulent activities in financial institutes can break the economic system of the country. These activities can be identified using clustering and classification algorithms. Effectiveness of these algorithms depend on quality of the input data. Moreover, financial data comes from various sources and forms such as financial statements, stakeholders activities and others. This data from various sources is very vast and unstructured big data. Hence, parallel distributed pre-processing is very significant to improve the quality of the data. Objective of this work is dimensional reduction considering feature selection and extraction algorithm for large volume of financial data. In this paper an attempt is made to understand the implications of feature extraction and transformation algorithm using Principal Feature Analysis on the financial data. Effect of reduced dimension is studied on various classification algorithms for financial loan data. Parallel and distributed implementation is carried out on IBM Bluemix cloud platform with spark notebook. The results show that reduction of features has significantly improved execution time without compromising the accuracy.

Keywords: Classification; Financial big data; Feature selection and extraction; Support Vector

10. CREDIT COLLECTIBILITY PREDICTION OF DEBTOR CANDIDATE USING DYNAMIC K-NEAREST NEIGHBOR ALGORITHM AND DISTANCE AND ATTRIBUTE WEIGHTED

AUTHOR: Tiara Fajrin, Ragil Saputra, Indra Waspada

ABSTRACT:

BPR Bank Jepara Artha is one of the banks that provide loan for activist of MSME (Micro, Small and Medium Enterprises). The activity of loaning in BPR Bank Jepara Artha has bad loan issue that often occured especially on loan MSME activist, therefore it needs an application to predict the loan collectibility of debtor applicant to minimize the issue. This research applied one of Data Mining classification algorithms in the application that produces output that can serve as information sources or second opinion for the consideration in decision making to accept or reject the loan applicant. The algorithm that be used was Dynamic K-Nearest Neighbor and Distance and Attribute Weighted algorithm which is a dynamic selection of k, addition of attribute and distance weight on k-Nearest Neighbor algorithm. The attributes that be used to determine the prediction result are 5C (Character, Capacity, Capital, Collateral, Condition of Economic), monthly income, debt status elsewhere, number of dependents, age, type of commodity and business status. The results of Dynamic K Nearest Neighbor and Distance and Attribute Weighted algorithm performance measurement use historical data of 240 old customer, the order of importance of the attribute specified by domain expert and 10-fold Cross Validation yield the highest accuracy of 65.83value of 56.10attribute in this algorithm performs higher accuracy, precision and recall than the one which does not use it. The change in the order of importance of the attributes determined by Correlation Attribute Evaluation yield in a higher recall value of 54.35 attributes determined by the domain expert.

Keywords: Loan prediction, Data Mining, Classification, Dynamic K-Nearest Neighbor and Distance and Attribute Weighted, Cross Validation

REFERENCES

[1] PhilHyo Jin Do ,Ho-Jin Choi, “Sentiment analysis of reallife situations using loca- tion, people and time ascontextual features,” International Conference on Big Data and Smart Computing (BIGCOMP), pp. 39–42. IEEE, 2015.

[2] Bing Liu, “Sentiment Analysis and Opinion Mining,” Morgan Claypool Publishers, May 2012.

[3] Bing Liu, “Sentiment Analysis: Mining Opinions, Sentiments, and Emotions,” Cambridge University Press, ISBN:978-1-107-01789-4.

[4] Shiyang Liao, Junbo Wang, Ruiyun Yu, Koichi Sato, and Zixue Cheng, “CNN for situations understanding based on sentiment analysis of twitter data,” Procedia computer science, 111:376–381, 2017.CrossRef.

[5] K I Rahmani, M.A. Ansari, Amit Kumar Goel, “An Efficient Indexing Algorithm for CBIR,” IEEE- International Conference on Computational Intelligence Communication Technology, 13-14 Feb 2015.

[6] Gurlove Singh, Amit Kumar Goel ,”Face Detection and Recognition System using Digital Image Processing” , 2nd International conference on Innovative Mechanism for Industry Application ICMIA 2020, 5-7 March 2020, IEEE Publisher.