IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 3, March 2021

DOI 10.17148/IJARCCE.2021.10337

Survey Paper on Credit Risk Management using Machine Learning and User Personalization

Hrutvik Desai[1], Heet Savla[2], Rahil Vithalani[3], Prachi Pradhan[4]

K J Somaiya college of Engg. Vidyavihar, Mumbai, India[1,2,3,4]

Abstract: The banking industry is changing and new trends of using technology in banking are emerging. Credit risk assessment has an important role in the banking sector. Various researches have been done to automate the process of predicting the loan default probability to speed up the process and reduce human errors. In this paper, a comparative study of various methods to identify the rightful customers for the bank. The paper also demonstrates how social media networks can be useful for gathering data which can be used for finding the loan eligibility of the user. The outcome of this paper is to find which method or algorithm gives the most accurate result while prediction loan eligibility of a user.

Keywords: Machine Learning, Credit Risk, Personalization, Social Networks

I.INTRODUCTION

According to experts machine learning can benefit humans in various ways by identifying the different trends and patterns, by continuous improvement, used in a wide range of applications, no human intervention is needed, and handles multivariate and multidimensional data. The idea is to automate all the possible human interventions by using machine learning algorithms in fields such as banking, agriculture, business, etc. Businesses and banks in the monetary industry use machine learning algorithms for two crucial purposes 1. To prevent fraud 2. To observe important insights in data. The understanding can help investors know when to trade and also recognize investment opportunities. The initial motive of user personalization is to render content and purpose that matches specific user interest or needs, with no effort from the intended users. The user personalized system creates the profiles of the user and adjusts the interface according to that profile.

II.LITERATURE SURVEY

In 2019, S. Z. H. Shoumo, M. I. M. Dhruba, S. Hossain, N. H. Ghani, H. Arif and S. Islam devised a machine learning model which helps in credit risk assessment system by showing great promise in identifying loan defaulters. The authors have also stated a comparative analysis of various machine learning algorithms and concluded that among all the models, the combination of tuned Support Vector Machine and Recursive Feature Elimination with Cross- Validation has shown more accurate results. The paper also introduces the five step proposed model by the authors. The support vector machines can outperform other tree-based models or regression models. Furthermore, in the debate over which dimensionality reduction technique to use, their model has shown us that recursive feature elimination with cross-validation can outperform other models [1].

In 2018, Addo, Peter & Guegan, Dominique & Hassani, Bertrrs demonstrated an approach for credit risk analysis using machine and deep learning models. The authors worked on building binary classifiers using these models to predict loan default probability. According to the authors the final decision depends on the choice of algorithms, parameters, relevant features and the role of evaluation criteria. The paper gives the proper explanation of the algorithms used, norms used to classify the outputs, dealing with imbalanced datasets, the results and the corresponding parameters and criteria used for each model are given, and how the machine learning and deep learning algorithms are used to provide a loan to an individual or a company. The authors have retained six approaches: two machine learning models (a random forest model, a gradient boosting machine) and four deep learning models and observed that tree-based models are more stable than models based on artificial neural networks [2].

In 2008 G. Bartolomeo, F. Petersen and M. Pluke have cited the importance of Personalization of user needs according to their preferences, context of use in various situations. The authors also have analyzed how user profiles of different users will be managed according to their needs. This paper describes the profile management and personalization activities at the European Telecommunications Standards Institute Technical Committee Human Factors. Personalization basically identifies how different users can be benefitted based on their different needs along with the ease of operation. The users interact with the profiles by checking, adding to, modifying, or deleting information in their profile and get notified when the other people access their profile, and understand how their profile affects the service or capabilities that the users experience. There are two agents mentioned in the system as Profile Storage agent who is responsible for storing the information about the user data and the locations of data repositories of data related to users and Profile Processing agent who is responsible to ensure all operations required by the profile rules are carried out. A better user experience will be provided by individualized devices and services. The reuse of existing users' knowledge will help to manage new terminal services and devices and lead to easier and faster take up of upcoming technologies. Harmonization and Synchronization of user profiles across devices and services leads to easier and faster use of devices and services. The user profile that will suit a particular situation and that even handles many areas will only be needed to be defined once and also the end-users will not have to enter their preferences again and again as each time the users acquire new devices and services. The main consequences were that during the user management, all the information of the user is already stored and hence the privacy of the user is not maintained. Data of the user profile should be all-time available to the profile storage and processing agent.[3]

In 2020, M. A. Sheikh, A. K. Goel and T. Kumar s demonstrated a machine learning model which predicts the customers for granting loans as users will be fulfilling most of the criteria. Authors have used a logistic regression model which uses a dataset from kaggle to predict the right customers. The authors have also stated a method to create a machine learning model starting with data collection (Kaggle dataset), preprocessing (handling noisy data and filling in the missing values), and feature engineering (techniques used to prepare a proper dataset which is compatible with machine learning model). This approach saves a lot of time and also reduces human error, however such models should be built very properly or else wrong predictions will result in huge losses for banks [4].

In 2011, Lops, Pasquale & de Gemmis, Marco & Semeraro, Giovanni & Narducci, Fedelucio & Musto, Cataldo proposed how social networks like LinkedIn can be used to gather information about the users. The paper proposes an idea where user profiles can be made and based on their interests and specialties research papers can be recommended to the users. Various methods for recommendation of research papers are given and one of such methods is that researchers from the same area of field/interest tend to read the same research articles. The authors have used the LinkedIn extractor system of the users and their connections and found that this group of people tend to have similar interests. This same thing can be used in loan prediction where people in the connection can help to find whether the customer will default or not [5].

In 2019(Kumar, R., Jain, V., Sharma, P. S., Awasthi, S., & Jha) and others talked about the safety of the loan using three machine learning models, Logistic Regression (LR), Decision Tree (DT) and Random Forest (RF) and concluded that the decision tree is more accurate as compared to other two algorithms. Loan Recovery is a very risky task so loan approval has to be done with great care. So, this paper suggests to automate the loan approval process using ML algorithms to make the process less risky and also to deal with less loss. The paper also mentions various other techniques to predict the loan status and also to calculate the credit score in its literature survey using machine learning, web services, business process execution language, fuzzy model, etc. Three ml algorithms are applied to a dataset one-by-one where 70 percent of the data is used as training data and the remaining 30 percent is used as the testing data. The columns which are considered to build this prediction model are: Gender, Married, Dependents, Education, Self-employed, Applicant Income, ComplicantIncome, Loan Amount, Loan_Amount_Term, Credit_History and Property_Area. After running all three models, the accuracy of these models are calculated, with LR having 93.04% accuracy, DT having 95% and RF having 92.53% accuracy. So after this it was concluded that the Decision tree is more accurate in comparison to Logistic Regression and Random Forest. [6]

In 2020, Zhenya Tian, Jialiang Xiao, Haonan Feng, Yutian Wei talked about the comparative analysis of various algorithms with its advantages and disadvantages. The support vector machines are a traditional method used to calculate the binary classification problems. SVM is a general linear classifier, uses supervised learning to classify data. The decision boundary is the margin of the hyper plane for solving the samples which are linear in nature. To calculate the risk SVM uses a loss function and adds a regulator term to the solution system to optimize the structure of the risk. SVM is a linear classifier with stability and sparsity.The advantages is that stability of SVM is high as it works only with linear data with binary outputs, can work with sparse data, and is easy to implement when the data is efficient. The disadvantage is that SVM was developed in 1964, hence due to the age of the algorithm the method is time inefficient and also performance inefficient when compared to newer methods used to calculate risks. Logistic regression method works with regression analysis when the data variable is binary. Logistic regression like other regression techniques, is predictive analysis of data. This method is used to identify the relation between differential binary data. The only disadvantage is that Logistic regression is sensitive to multi variable collinearity of independent data points (variables) in the model and will largely affect other relationships of variables. Decision tree method is the most commonly used reasoning approach in machine learning. Decision tree analyzes training samples to summarize concepts and knowledge, it is a method to approximate discrete objective functions which are represented in a decision tree. The disadvantages of decision trees are that any change in data may lead to the creation of a completely different decision tree hence making this approach highly unstable and when the data is unbalanced with inconsistent sample size, information gain may be biased towards samples with more values, making it more challenging. Multilayer perceptron (MLP) method falls under the category of artificial neural networks that tends to its structure, mapping inputs to outputs in vectors. MLP can be considered as a directed graph with an interconnected structure with each and every node. Since this technique is ANN based it can also accept non-linear inputs in forms of activation functions. The disadvantage is that MLP has strict requirements for feature selection and data normalization. The AdaBoost method uses and selects the weak classifiers for weighted combination step after step. The disadvantage of AdaBoost is sensitive to the outlier data, which has a great influence on the accuracy of prediction results. Random Forest selects the features and the data to generate many decision trees and then summarize the results of the decision trees. The advantage is that Random Forest improves the prediction without increase of the calculated amount. The Disadvantage is that it may generate many similar decision trees that may cover up the real results. Gradient Boosting Decision Tree algorithm ensembles several weak classifiers (decision tree) together to form strong and effective classifiers. It's a process of emerging all weak classifiers together to have a model with better performance. [7]

In 2018 A. Mittal, A. Srivastava, A. Saxena and M. Manoria discussed that in the finance sector, credit risk has become one of the most crucial tasks and the paper cites that due to that, there is a high level of competition within various banking institutions. The complexity of designing such a framework for credit risk modelling is quite complex and therefore the paper provides a brief study of the risk assessment by the risk based approach as well as reducing the dimensionality of the data and also improving the classification of the data in comparison to other existing methods. The paper states that over the years the concept of the banking sector is that banking institutions lend the surplus community funds that banks have, to the people in the manner of loans. This lending should be done very carefully with accurate customer assessment which is very crucial in the banking sector. Accurate risk assessment is not only important for minimizing credit risk but is also important to avoid errors by rejecting a valid customer. Today credit risk assessment is performed for credibility of the customer, but the size of the dataset is huge, and therefore, there is the need of automation by applying the soft computing approach. Since, the Indian economy is in the period of liberalization, risk management and risk analysis are of paramount importance. The paper states that for the management of business risks, risks such as market risk, credit risk and operational risk need to be translated into a composite measure. As regards the Basel Committee Agreements and the RBI Guidelines, the study of risk analysis and risk management in cooperative banks is of fundamental importance. Credit risk modelling at the transactional level can be done by two methods, the analytical approach based on the accounting and forecasts and statistics. The paper states that Business Intelligence (BI) provides all the necessary functionalities for the identification, integration and the analysis of data and thereby provides decision support for strategic management. Data Mining along with the framework of Business Intelligence (BI) systems is very popular and helps in solving banking as well as financial problems by finding correlations, casualties and attributes that are not immediately visible to the institution managers as the data that has been generated is too fast or too large on the screen. The authors in this paper have performed a comparative analysis of the data mining techniques to find the default probability of the credit card holders and the best statistical forecasting methods to analyze this indicator. Methods such as Support Vector Machines (SVM), Logistic Regression and Naïve Bayes algorithm have been implemented by the authors. The authors' study in this paper is mainly aimed at achieving objectives such as Identifying significant risks, Quantifying the risk, a comparative analysis of the risk assessment models and Designing a risk assessment model for understanding of the significant risks facing in the banks and enhancing the accuracy of the system [8]

In 2017 T. N. Pandey, A. K. Jagadev, S. K. Mohapatra and S. Dehuri and others discussed that the major activity of the banking industry is to lend money to those that are in need of money. In this paper, the authors have surveyed various techniques for the credit risk analysis. The authors have discussed various techniques for credit risk management in this paper such as Bayesian Classifier, Decision Tree, K-Nearest Neighbor, K-Means, Multilayer Perceptron, Extreme Learning Machine and Support Vector Machine. The authors have also analyzed and compared their accuracies using various types' classifiers. This paper opens the doors for further research in the credit risk management using the machine learning classifier in the financial field.[9]

III.CONCLUSION

This paper describes the different machine learning technologies and how these technologies can help in calculating the risk in banks while approving the loans. A comparative study of the different machine learning algorithms was done with proper advantages and disadvantages. How new banking trends are emerging in the banking sectors were studied. The importance of user personalization was also studied from the various papers. Nowadays satisfaction of the user is considered as crucial in systems. Users need the systems to be customized according to their needs and interests. This paper also cites how risk can occur while providing loans to the users and how the risks in the banking sector can be managed by giving different approaches.

DOI 10.17148/IJARCCE.2021.10337

REFERENCES

[1] S. Z. H. Shoumo, M. I. M. Dhruba, S. Hossain, N. H. Ghani, H. Arif and S. Islam, "Application of Machine Learning in Credit Risk Assessment: A Prelude to Smart Banking," TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON), Kochi, India, 2019, pp. 2023-2028, doi: 10.1109/TENCON.2019.8929527.

[2] Addo, Peter & Guegan, Dominique & Hassani, Bertrand. (2018). Credit Risk Analysis Using Machine and Deep Learning Models. Risks. 6. 38. 10.3390/risks6020038.

[3] Petersen, Françoise & Bartolomeo, Giovanni & Pluke, Mike. (2008). Personalization and User Profile Management. International Journal of Interactive Mobile Technologies (iJIM). 2. 10.3991/ijim.v2i4.666.

[4] M. A. Sheikh, A. K. Goel and T. Kumar, "An Approach for Prediction of Loan Approval using Machine Learning Algorithm," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2020, pp. 490-494, doi: 10.1109/ICESC48915.2020.9155614.

[5] Lops, Pasquale & de Gemmis, Marco & Semeraro, Giovanni & Narducci, Fedelucio & Musto, Cataldo. (2011). Leveraging the LinkedIn social network data for extracting content-based user profiles. RecSys'11 - Proceedings of the 5th ACM Conference on Recommender Systems. 293-296. 10.1145/2043932.2043986.

[6] Kumar, R., Jain, V., Sharma, P. S., Awasthi, S., & Jha, G. (2019). Prediction of Loan Approval using Machine Learning. *International Journal of Advanced Science and Technology*, *28*(7), 455 - 460. Retrieved from http://sersc.org/journals/index.php/IJAST/article/view/460

[7] Zhenya Tian, Jialiang Xiao, Haonan Feng, Yutian Wei,
Credit Risk Assessment based on Gradient Boosting Decision Tree, Procedia Computer Science, Volume 174,
2020, Pages 150-160, ISSN 1877-0509,
https://doi.org/10.1016/j.procs.2020.06.070.

[8] A. Mittal, A. Shrivastava, A. Saxena and M. Manoria, "A Study on Credit Risk Assessment in Banking Sector using Data Mining Techniques," 2018 International Conference on Advanced Computation and Telecommunication (ICACAT), Bhopal, India, 2018, pp. 1-5, doi: 10.1109/ICACAT.2018.8933604.

[9] T. N. Pandey, A. K. Jagadev, S. K. Mohapatra and S. Dehuri, "Credit risk analysis using machine learning classifiers," 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), Chennai, India, 2017, pp. 1850-1854, doi: 10.1109/ICECDS.2017.8389769.