PROJECT AND DESIGN PHASE
LITERATURE SURVEY

DATE	19 September 2022
TEAM ID	PNT2022TMID28895
PROJECT TITLE	Smart Lender - Applicant Credibility Prediction for Loan Approval
TEAM MEMBERS	KOTHAI S 411719104029 SHARMILA K 411719104046 KALAIVANI L 411719104021 ASHMITHA R 411719104004

Smart Lender - Applicant Credibility Prediction for Loan Approval

LITERATURE SURVEY

[1] Ms. Kathe Rutika Pramod uses the decision tree for the loan prediction. In Decision tree each node represents a feature (attribute), each link (branch) represents a decision (rule) and each leaf represents an outcome (categorical or continues value). Using different data analytics tools loan prediction and there severity can be forecasted. In this process it is required to train the data using different algorithms and then compare user data with trained data to predict the nature of loan. Several R functions and packages were used to prepare the data and to build the classification model. The work proves that the R package is an efficient visualizing tool that applies data mining techniques. Using R Package, customer's data analysis can be done and depends on that bank can sanction or reject the loan. In real time customers data sets may have many missing and imputed data which needs to be replaced with valid data generated by making use of the available completed data. The dataset has many attributes that define the credibility of the customers seeking for several types of loan. The values for these attributes can have outliers that do not fit into the regular range of data. DT is a supervised learning algorithm used to solve classification and regression problems too. Here, DT uses tree representation to solve the prediction problem, i.e., external node and leaf node in a tree represents attribute and class labels respectively. The analytical process started from data cleaning and processing, Missing value imputation with mice package, then exploratory analysis and finally model building and evaluation. The best accuracy on

public test set is 0.811. This brings some of the following insights about approval. Applicants with Credit history not passing fails to get approved, Probably because that they have a probability of a not paying back. Most of the Time, Applicants with high income sanctioning low amount is to more likely get approved which make sense, more likely to pay back their loans. Some basic characteristic gender and marital status seems not to be taken into consideration by the company.

[2] Shubham Nalawade, Suraj Andhe, Siddhesh Parab, Prof. Amruta Sankhe proposed system includes a web application with a model trained by using machine learning algorithms deployed in it. There are a total 11 fields in the form which the user needs to fill. The dataset that we have used for training the model also includes 11 attributes. This dataset is pre-processed before using it for training the model. The pre-processing is done by replacing the null values in the dataset with mean and mode method and replacing the string values with 1 and 0 using label encoder. Then the dataset was divided into two parts: train and test. 90% of the dataset is used for training purposes and 10% is used for testing the accuracy that the model will give for different algorithms. After splitting the dataset different algorithms were applied and each of them gave different accuracy. The best we got was from Logistic Regression i.e., 88%. Once the model is trained a pickle file is created of the model. When the client wants to predict his/her loan approval the client has to first fill a form by visiting our web application. After filling the form, the user has to just click on the MAKE PREDICTION button and depending on the pickle file or the model that we have trained it will give the result as whether the loan of the customer will be approved or not. As we have also done the comparison of different machine learning algorithms in terms of their accuracy. The web application also includes a bar plot graph of the comparison of algorithms, insights of the dataset that we have used for training the model. This system will make it easier for the banks or organizations to do the job of loan approval prediction. Here author compared different machine learning algorithms for the Property Loan dataset; they are Random Forest, Naive Bayes, Logistic Regression and K Nearest Neighbors. The Logistic Regression algorithm gave the best accuracy (88.70%). Following this approach, we found that apart from the logistic regression, the rest of the algorithms performed satisfactory in terms of giving out the accuracy. The accuracy range of the rest of the algorithms were from 75% to 85%. Whereas the logistic regression gave us the best possible accuracy (88.70%) after the comparative study of all the algorithms.

[3] Soni P M, Varghese Paul introduces a new hybrid feature selection algorithm using wrapper method and fisher score method. The new algorithm is termed as wrapper-fisher feature selection algorithm. In this work, LCPS uses a wrapperfisher feature selection algorithm to select the most significant features which will improve the accuracy of Random Forest (RF) classification. After studying various past data from the bank it is possible to identify several attributes that can influence the customer behaviour. The most influencing attribute can be considered while a new customer approaches the bank for loan and thus we can identify the potential of customer. Here

by enabling the bank officers to identify fraud applicants by using the final application of this research work. The accuracy level considerably increased after feature selection methods were applied to the classifier. The proposed algorithm had produced better accuracy than existing methods. Experiments on standard data sets proved that the proposed algorithm for loan credibility prediction system outperforms many other feature selection methods. , a novel hybrid feature selection approach is proposed to predict the loan repayment capability behavior of a customer in a cost effective way. Complex set of decision making are need to be taken by bank officers to determine whether to approve loan applicants or not. Normally classification technique solved the problem up to an extent. Now the experiment proved that a model that use feature selection before classification can help the bank officers to take proper decision more accurately. 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 co-operative banking sector. The experiment proved that the classification accuracy have considerably increased after feature selection. The proposed algorithm had produced better accuracy than existing methods. Experiments on standard data sets proved that the proposed algorithm for loan credibility prediction system outperforms many other feature selection methods

[4] In Dr.AMIT KUMAR GOEL proposed model for loan prediction, Dataset is split into training and testing data. After then training datasets are trained using the decision tree algorithm and a prediction model is developed using the algorithm. Testing datasets are then given to model for the prediction of loan. The motive of this paper is to predict the defaults who will repay the loan or not. Various libraries like pandas, numpy have been used. After the loading of datasets, Data preprocessing like missing value treatment of numerical and categorical is done by checking the values. Numerical and categorical values are segregated. Outliers and frequency analysis are done. developed a prediction model for Loan sanctioning which will predict whether the person applying for loan will get loan or not. The major objective of this project is to derive patterns from the datasets which are used for the loan sanctioning process and create a model based on the patterns derived in the previous step. This model is developed by using the one of the machine learning algorithms. Here the author used decision tree algorithm for development. Based on the segregated value the decision tree able to work and predict the loan approval. Here author is able to conclude that Decision tree version is extraordinary efficient and gives a higher end result. Developed a model which can easily predict that the person will repay its loan or not. we can see our model has reduced the efforts of bankers. Machine learning has helped a lot in developing this model which gives precise results.

[5] Mehul Madaan used two machine learning algorithms, the Random Forest and Decision Trees to work out a model for loan prediction and credit risk assessment. The results of both the model are shown below with their classification report and confusion

matrix to get a better understanding of the accuracy and other scores of the two models. This paper aimed to explore, analyse, and build a machine learning algorithm to correctly identify whether a person, given certain attributes, has a high probability to default on a loan. This type of model could be used by Lending Club to identify certain financial traits of future borrowers that could have the potential to default and not pay back their loan by the designated time. The Random Forest Classifier provided us with an accuracy of 80% while the Decision Tree method provided us with an accuracy of 73%. Hence, the Random Forest model appears to be a better option for such kind of data. Lending Club must be careful when identifying potential borrowers who fit certain criteria. For example, borrowers who do not own a home and are applying for a small business or wedding loan, this could be a negative combination that results in the borrower defaulting on a loan. One of the drawbacks is simply the limited number of people who defaulted on their loan in the 8 years of data (2007-2015). We could use an updated data frame that consists of the next 3 years' values (2015-2018) and see how many of the current loans were paid off, defaulted, or even charged off. Then, these new data points can be used for prediction or and training new models for better and more accurate results. Since the algorithm puts some of the non-defaulters in the default class, we might want to look further into this issue to help the model accurately predict capable borrowers.

[6] In the paper presentation of AFRAH KHAN, EAKANSH BHADOLA, ABHISHEK KUMAR and NIDHI SINGH, It will be comparing different prediction models and deduce their limitations as well as advantages. Since all the research papers used different sets of data to infer the accuracy and for cross validation of data, the authors have used the same data for all the models which will give a clearer view on their performance and lead to a better comparison of the same. On the basis of the results, a modified prediction model will be created to ensure maximum accuracy and performance. The predictive models based on Logistic Regression, Decision Tree and Random Forest, give the accuracy as 80.945%, 93.648% and 83.388% whereas the cross-validation is found to be 80.945%, 72.213% and 80.130% respectively. This shows that for the given dataset, the accuracy of model based on decision tree is highest but random forest is better at generalization even though it’s cross validation is not much higher than logistic regression.

REFERENCES

[1] Ms. Kathe Rutika Pramod Information Technology Engineering SVIT, Nashik Maharashtra, India a An Approach For Prediction Of Loan Approval Using Machine Learning Algorithm-2021 IJCRT | Volume 9, Issue 6 June 2021

[2] Shubham Nalawade, Suraj Andhe, Siddhesh Parab, Prof. Amruta Sankhe- Loan Approval Prediction-Loan Approval Prediction -2021 IJCRT | Volume: 09 Issue: 04 | Apr 2022

[3] Soni P M, Varghese Paul- Algorithm For the Loan Credibility Prediction System-International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-1S4, June 2019

[4] Dr.AMIT KUMAR GOEL, M.Tech., Ph.D - LOAN PREDICTION SYSTEM - APRIL / MAY- 2020

[5] Loan default prediction using decision trees and random forest: A comparative study-IOP Conference Series Mehul Madaan et al 2021 IOP Conf. Ser.: Mater. Sci. Eng. 1022 012042

[6] AFRAH KHAN, EAKANSH BHADOLA, ABHISHEK KUMAR and NIDHI SINGH - LOAN APPROVAL PREDICTION MODEL A COMPARATIVE ANALYSIS | Advances and Applications in Mathematical Sciences Volume 20, Issue 3, January 2021