Literature Survey

Smart Lender - Applicant Credibility Prediction For Loan Approval

Applied Data Science: A Literature Review Paperwork

Applied Data Science:

Applied Data Science students will learn to extract relevant information from a compilation of large datasets from different sources. The ability to create, manage and exploit data has become one of the most important challenges for practitioners in most disciplines, sectors, and industries. Students with expertise in Applied Data Science will be highly sought after in the future labor market, where they will contribute to the development of smart solutions and digitization.

Introduction:

The two most pressing issues in the banking sector are: 1) How risky is the borrower? 2) Should we lend to the borrower given the risk? The response to the first question dictates the borrower's interest rate. Interest rate, among other things (such as time value of money), tests the riskiness of the borrower, i.e. the higher the interest rate, the riskier the borrower. We will then decide whether the applicant is suitable for the loan based on the interest rate. Lenders (investors) make loans to creditors in return for the guarantee of interest-bearing repayment. That is, the lender only makes a return (interest) if the borrower repays the loan. However, whether he or she does not repay the loan, the lender loses money. Banks make loans to customers in exchange for the guarantee of repayment. Some would default on their debts, unable to repay them for a number of reasons. The bank retains insurance to minimize the possibility of failure in the case of a default. The insured sum can cover the whole loan amount or just a portion of it. Banking processes use manual procedures to determine whether or not a borrower is suitable for a loan based on results. Manual procedures were mostly effective, but they were insufficient when there were a large number of loan applications. At that time, making a decision would take a long time. As a result, the loan prediction machine learning model can be used to assess a customer's loan status and build strategies. This model extracts and introduces the essential features of a borrower that influence the customer's loan status. Finally, it produces the planned performance (loan status). These reports make a bank manager's job simpler and quicker.

Description:

One of the most important factors which affect our country’s economy and financial condition is the credit system governed by the banks. The process of bank credit risk evaluation is recognized at banks across the globe. “As we know credit risk evaluation is very crucial, there is a variety of techniques are used for risk level calculation. In addition, credit risk is one of the main functions of the banking community

Literature review:

We start our literature review with more general systematic literature reviews that focus on the application of machine learning in the general field of Banking Risk Management. Since the global financial crisis, risk management in banks has to take a major role in shaping decision-making for banks. A major portion of risk management is the approval of loans to promising candidates. But the black-box nature of Machine learning algorithms makes many loan providers vary the result. Martin Leo, Suneel Sharma and k. Maddulety's [1] extensive report has explored where Machine Learning is being used in the fields of credit risk, market risk, operational risk, and liquidity risk only to conclude that the research falls short of extensive research is required in the field.

We could not find any literature review for loan prediction for specific Machine learning algorithms to use which would be a possible starting point for our paper. Instead, since loan prediction is a classification problem, we went with popular classification algorithms used for a similar problem. Ashlesha Vaidya [2] used logistic regression as a probabilistic and predictive approach to loan approval prediction. The author pointed out how Artificial neural networks and Logistic regression are most used for loan prediction as they are easier comparatively develop and provide the most accurate predictive analysis. One of the reasoning behind this that that other Algorithms are generally bad at predicting from non-normalized data. But the nonlinear effect and power terms are easily handled by Logistic regression as there is no need for the independent variables on which the prediction takes place to be normally distributed.

Logistic regression still has its limitations, and it requires a large sample of data for parameter estimation. Logistic regression also requires that the variables be independent of each other otherwise the model tends to overweigh the importance of the dependent variables.

A solution to this multicollinearity problem among the categorical explanatory variables is the use of a categorical principal component analysis which can be seen used by Guilder and Ozlem [3] on a case study for housing Loan approval data. The goal of Principal component analysis is to reduce the number of m variables where many of them would be highly correlated with each other, to a smaller set of n uncorrelated variables called principal components which account for the variances between the previous m variables. Methods such as PCA are known as dimension reduction of the data. It may be suitable for scaled continuous variables but it isn’t quite an appropriate method of dimension reduction for categorical variables. Thus, the authors here used a tweaked version of PCA for categorical data called CATPCA or categorical (nonlinear) principal components analysis which is specifically developed for where the dependent variables are a mix of nominal, ordinal, or numeric data which may not have linear relationships with each other. CATPCA works by using a scaling process optimized to convert the categorical variables into numeric variables.

Similar to PCA, Zaghdoudi, Djebali &amp; Mezni [4] compared the use of Linear Discriminant Analysis versus Logistic Regression for Credit Scoring and Default Risk

Prediction for foreseeing default risk o small and medium enterprises. Linear Discriminant Analysis (LDA) is like PCA for dimensionality reduction but instead of looking for the most variation, LDA focuses on maximizing the separability among the know categories. This subspace that well separates the classes is usually in which a linear classifier can be learned. The classification of those enterprises correctly in their original groups through both these methods was inconsequential with Logistic regression having a 0.3% better accuracy score than LDA.

Another novel approach for T.Sunitha and colleagues [5] was to predict loan Status using Logistic Regression and a Binary Tree.&nbsp;Decision Tree is an algorithm for a predictive type machine learning model.

Classification and Regression Trees are referred to as CART (in short) introduced by Leo Breiman. It best suits both predictive and decision modeling problems. This Binary Tree methodology is the greedy method is used for the selection of the best splitting. Although Decision trees gave us a similar accuracy. The benefits of Decision Trees, in this case, were due to the latter giving equal importance to both accuracy and prediction. This model became successful in making a lower number of False Predictions to reduce the risk factor.

Rajiv Kumar and Vinod Jain [6] proposed a model using machine learning algorithms to predict the loan approval of customers. They applied three machine learning algorithms, Logistic Regression (LR), Decision Tree (DT), and Random Forest (RF) using Python on a test data set. From the results, they concluded that the Decision Tree machine learning algorithm performs better than Logistic Regression and Random Forest machine learning approaches. It also opens other areas on which the Decision Tree algorithm is applicable.

Some machine learning models give different weights to each factor but in practice sometimes loans can be sanctioned based on a single strong factor only. To eliminate this problem J. Tejaswini and T. Mohana Kavya [7] in their research paper have built a loan prediction system that automatically calculates the weight of each feature taking part in loan processing and on new test data the same features are processed concerning their associated weight. They have implemented six machine learning classification models using R for choosing the deserving loan applicants. The models include Decision Trees, Random Forest, Support Vector Machine, Linear Models, Neural Network and Adaboost. The authors concluded that the accuracy of the Decision Tree is highest among all models and performs better on the loan prediction system.

Predicting loan defaulters is an important process of the banking system as it directly affects profitability. However, loan default data sets available are highly imbalanced which results in poor performance of the algorithms. Lifeng Zhou and Hong Wang [8] in their call for paper made loan default prediction on imbalanced data sets using an improved random forests approach. In this approach, the authors have employed weights in decision tree aggregation. The weights are calculated and assigned to each tree in the forest during the forest construction process using Out-of-bag (OOB) errors. The experimental results conclude that the improved algorithm performs better and has better accuracy than the original random forest and other popular classification algorithms such as SVM, KNN, and C4.5. The research opens improvements in terms of efficiency of the algorithm if parallel random forests can be used for further work.

Anchal Goyal and Ranpreet Kaur [9] discuss various ensemble algorithms. Ensemble algorithm is a supervised machine learning algorithm that is a combination of two or more algorithms to get better predictive performance. They carried out a systematic literature review to compare ensemble models with various stand-alone models such as neural network, SVM, regression, etc. The authors after reviewing different literature reviews concluded that the Ensemble Model performs better than the stand-alone models. Finally, they concluded that the concept of combined algorithms also improves the accuracy of the model.

Data Mining is also becoming popular in the field banking sector as it extracts information from a tremendous amount of accumulated data sets.&nbsp;Aboobyda Jafar Hamid and Tarig Mohammed Ahmed [10] focused on implementing data mining techniques using three models j48, bayesNet, and naiveBayesdel for classifying loan risk in the banking sector. The author implemented and tested models using the Weka application. In their work, they made a comparison between these algorithms in terms of accuracy in classifying the data correctly. The operation of sprinting happened in a manner that 80% represented the training dataset and 20% represented the testing dataset. After analyzing the results the author came up with the results that the best algorithm among the three is the J48w algorithm in terms of high accuracy and low mean absolute error.

Reference:

1.Kumar Arun, GargIshan, Kaur Sanmeet, May-Jun. 2016. Loan Approval Prediction based on Machine Learning Approach, IOSR Journal of Computer Engineering (IOSR-JCE)

2. Wei Li, Shuai Ding, Yi Chen, and Shanlin Yang, Heterogeneous Ensemble for Default Prediction of Peer-to-Peer Lending in China, Key Laboratory of Process Optimization and Intelligent Decision-Making, Ministry of Education, Hefei University of Technology, Hefei 2009, China