International Journal of Communication and Information Technology

E-ISSN: 2707-6628
P-ISSN: 2707-661X
IJCIT 2020; 1(2): 41-45
Received: 15-05-2020
Accepted: 21-06-2020

Reddi Babusamikeri
GATE College, Tirupati, Andhra Pradesh, India

Using machine learning algorithm for loan providing to house

Reddi Babusamikeri

DOI: https://doi.org/10.33545/2707661X.2020.v1.i2a.19

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: K-Nearest Neighbor (K-NN) classifier, Commercial loan, Borrowers', Credit

Introduction
In commercial loan lending, scoring of borrowers' creditworthiness is one of the most important problems to be addressed in the Banking Industry. Credit risk is defined as the risk that borrowers will fail to meet their loan obligations [1]. The Credit scoring system is used to predict the credit risk and to reduce the illegal activities. This credit scoring systems are used to make decisions under information about the borrowers [3]. In order to make loan decisions, lenders want to minimize the risk of default of each lending decision, and realize the return that compensates for the risk. In general, Banking Industry success and failure is based on their credit risk. The credit amount couldn't collect properly, then the bank will be loss. So, bank profit is correlated to their credit risk. Credit risk is a crucial challenge and a complex task to manage and evaluate. Credit scoring tasks can be divided into two groups such as, application scoring and behavioral scoring. Application Scoring is to classify the credit applicant into ' good' and ' bad' risk groups. Behavioral scoring task is to classify the existing customers based on their payment history and personal information.

Related Works
Bank credit risk analysis with K Nearest-neighbor classifier: Case of Tunisian banks
Credit risk is defined as the risk that borrowers fail to pay its debt obligations. In recent years, a large number of banks have developed sophisticated systems and models to assist bankers in measuring, integrating and managing risk. The outputs of these models also play important roles in banks risk management and performance measurement processes. In this study we seek to address the question of short-term debt default assessment for Tunisian commercial bank. We use a database of 924 credit records of Tunisian companies granted by the Tunisian Commercial Bank from 2003 to 2006. The k-nearest neighbor classification algorithm is carried out and the results show that the best data set is related to the compilation and cash flow and that the good classification rate is 88.63% (for k = 3). A curved ROC is plotted to evaluate the performance of the model. The result shows that the AUC (Area under Curve) criterion is in the order of 87.4% (for the first model), 95% (for the third model) and 95.6% for the best model with cash flow information.

Comparison of Feature Selection Methods for Credit Risk Assessment
Credits' granting is a fundamental question for which every credit institution is confronted and one of the most complex tasks that it has to deal with. This task is based on analyzing

Corresponding Author:
Reddi Babusamikeri
GATE College, Tirupati, Andhra Pradesh, India

and judging a large amount of receipts credits' requests. Typically, credit scoring databases are often large and characterized by redundant and irrelevant features. With so many features, classification methods become more computational demanding. This difficulty can be solved by using feature selection methods. Many feature selection methods are proposed in literature such as filter and wrapper methods. Filter methods select the best features by evaluating the fundamental properties of data, making them fast and simple to implement. Wrapper methods select the best features according to the classifier accuracy, making results well-matched to the predetermined classification algorithm. However, they typically lack generality since the resulting subset of features is tied to the bias of the used classifier. The purpose of this thesis is to build simple and robust credit scoring models based on selecting the most relevant features. Three feature selection methods are proposed. First, we propose a new filter rank aggregation method based on optimization using genetic algorithms and similarity. Second, we introduce an ensemble wrapper feature selection method based on an improved exhaustive search. Combining both methods seems a natural choice to benefit from their advantages and avoid their shortcomings. Thus, a three-stage feature selection using quadratic programming is considered. Based on different performance criteria and on four real credit datasets our three methods are evaluated. Results show that feature subsets selected by the proposed methods are either superior or at least as adequate as those selected by their competitor

Credit Risk Evaluation using Hybrid Feature Selection Method

As a novel financing method, peer-to-peer (P2P) lending has drawn extensive attention as it provides those financers who cannot participate in the traditional financial market with funds. In P2P lending marketplaces, one of the crucial challenges that P2P online lending platforms are facing is to accurately predict the default risk of each loan by tapping into default prediction models, thus effectively helping P2P lending companies avoid credit risks. That traditional credit risk prediction models fail to meet the demand of P2P lending companies for default risk prediction, which is because of the uneven distribution of credit data samples in the P2P lending marketplaces (i.e., the default sampled data are scarce). In this paper, we designed a multi-round ensemble learning model based on heterogeneous ensemble frameworks to predict default risk. In this model, an extreme gradient boosting (XGBoost) is initially used for ensemble learning, and the XGBoost, deep neural network, and logistic regression are then regarded as heterogeneous individual learners to undergo a linear weighted fusion. To verify the designed default risk prediction model, real credit data from a famous P2P online lending marketplace in China were used in a test. The results of the experiment indicate that this model can effectively increase the predictive accuracy compared with traditional machine learning models and ensemble learning models

The Problem of Normalization and a Normalized Similarity Measure by Online Data

Case-based reasoning, image or data retrieval is based on the similarity determination between the actual case and the cases in the database. It is advisable to generalize the similarity values between 0 and 1 to compare different similarity values based on the scale. In this way the analogy is given by semantic meaning. The main problem arises when the case base is not yet complete and there are only a small number of cases, other cases increase as soon as they come into the system. In this case the upper and lower bounds of the feature values cannot be approximated to the true values. This paper deals with possible methods for estimating the upper and lower bounds of feature value, and the problems that arise when these values are not accurately estimated due to a limited number of models or one-priori unavailable parameter distributions. Its aim is to develop a method for learning the upper and lower bounds of attribute value and to deal with the change in the semantic meaning of similarity.

Improved of K-Nearest Neighbor Techniques in Credit Scoring

Credit scoring is gaining more attention in the academic world and business community today. Several modeling techniques have been developed to solve credit scoring tasks. Financial institutions are increasingly using credit scoring models to determine whether credit customers belong to a good applicant group or a bad applicant group. Benefits of using credit scoring models K-nearest neighbors can be described as reducing the cost of credit analysis, initiating faster credit decisions, and insuring credit collections. This model is compared to other models. Each of the models is based on credit scoring

Proposed System

We propose a machine learning classifier-based analysis model for credit data. We use the combination of Min-Max normalization and KNearest 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.

Algorithm

KNN algorithm

The KNN algorithm is a parametric and lazy learning algorithm. Non-parametric means no underlying data distribution. In other words, the model structure is determined from the dataset. This is very helpful in practice where real-world datasets do not follow mathematical theoretical umphs. Lazy algorithm means that it does not require training data points for model generation. All training data used during the test phase. This makes the training faster and the testing phase slower and more expensive. An expensive test phase means time and memory. In the worst case, KNN needs more time to scan all data points and more memory to store training data to scan all data points.

Logistic regression is the appropriate regression analysis to perform when the dependent variable is dichotomous (binary). As with all regression analyses, logistic regression is a predictive analysis. Logistic regression is used to describe the data and the relationship between one dependent binary variable and one or more nominal, ordinal, interval, or ratio-level independent variables. Sometimes logistic regressions are difficult to understand; The Intellect Statistics tool allows you to easily perform the analysis, and then explain the output in plain English.

4. Results and Discussions

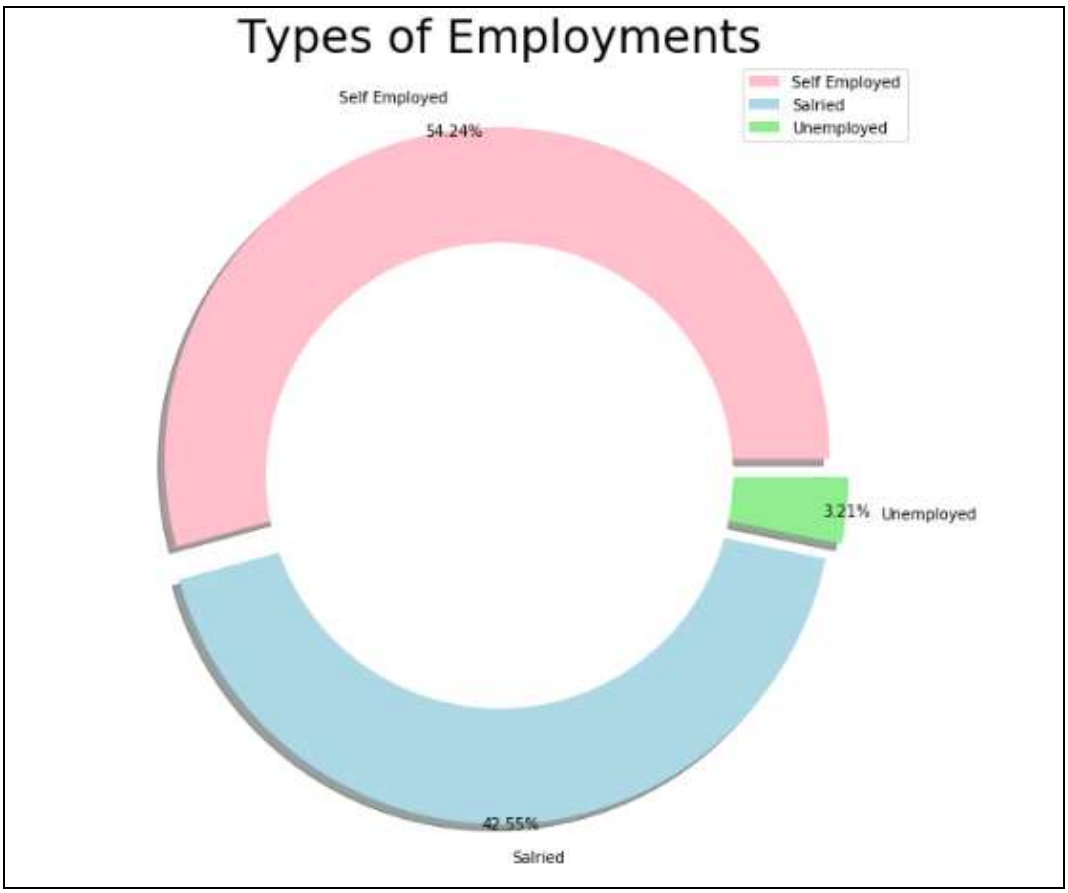

Fig 1: Types of employments

Here the above-mentioned Fig 1 represents the kinds of employments for Employees of any firm or Organization.

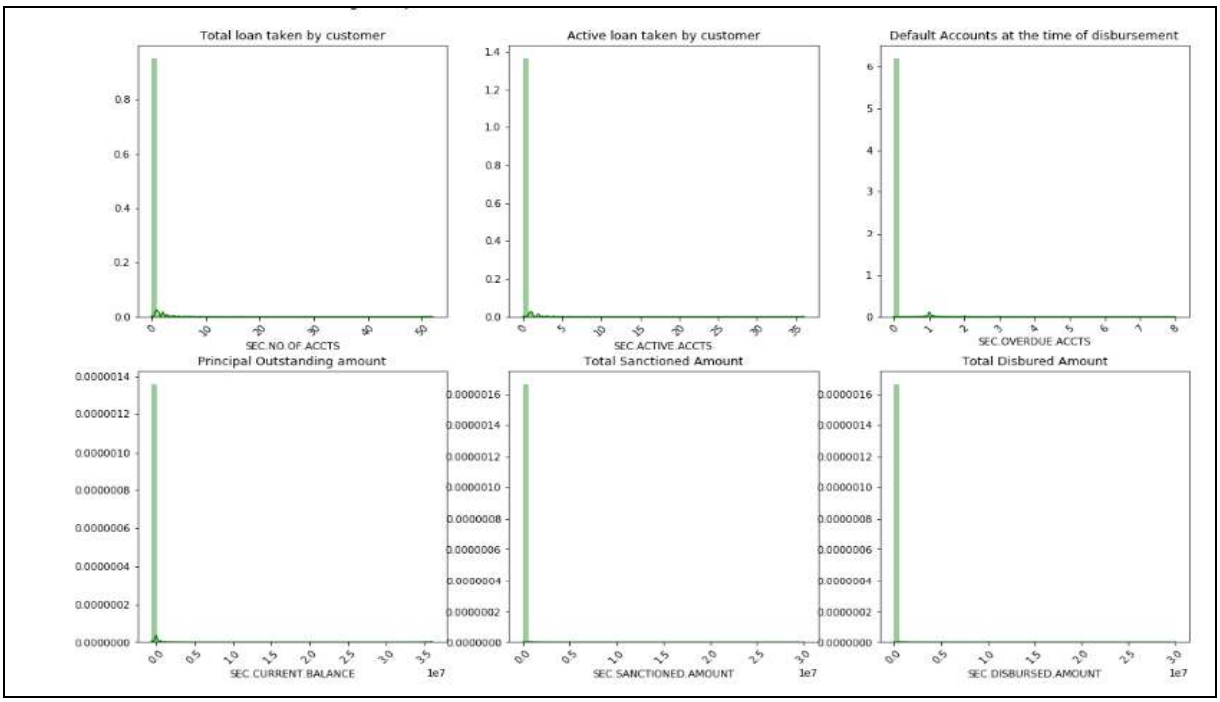

Fig 2: Loan Status

The above Fig 2 explains the customer's loan status of the loans like total principal amount disbursed, outstanding balance and other details.

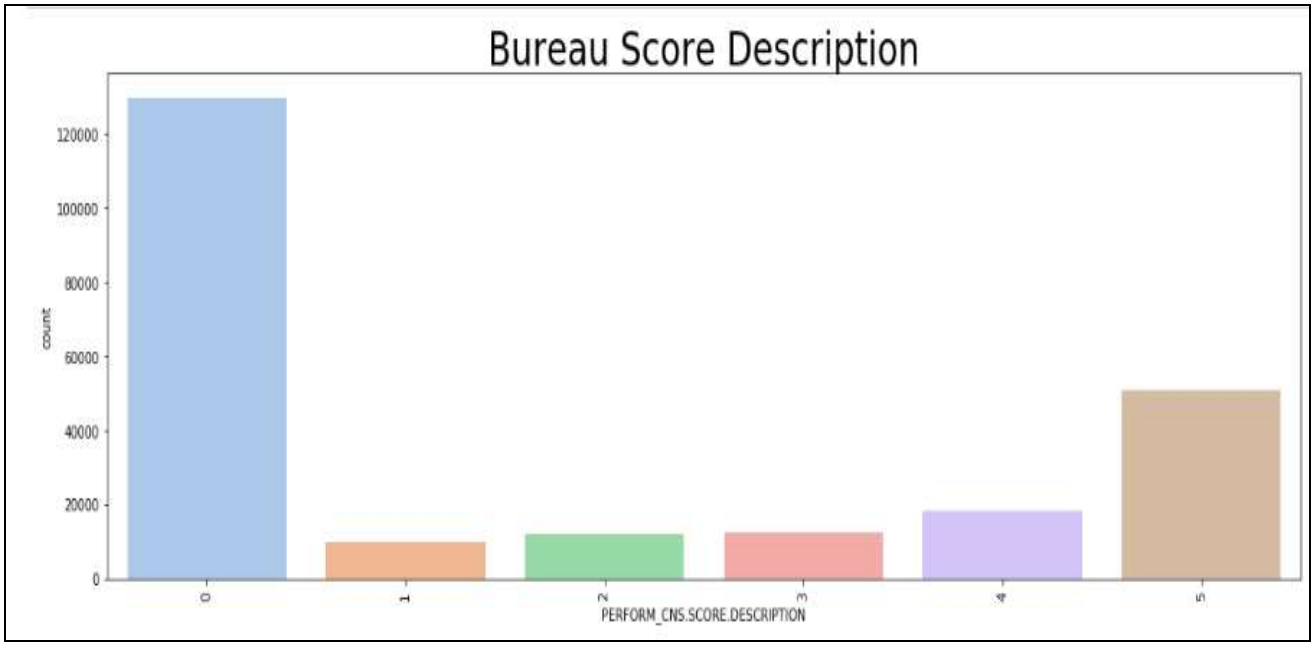

Fig 3: Bureau Score Description

The above Fig 3 explains the Bureau Score description for employees who have availed the Loans.

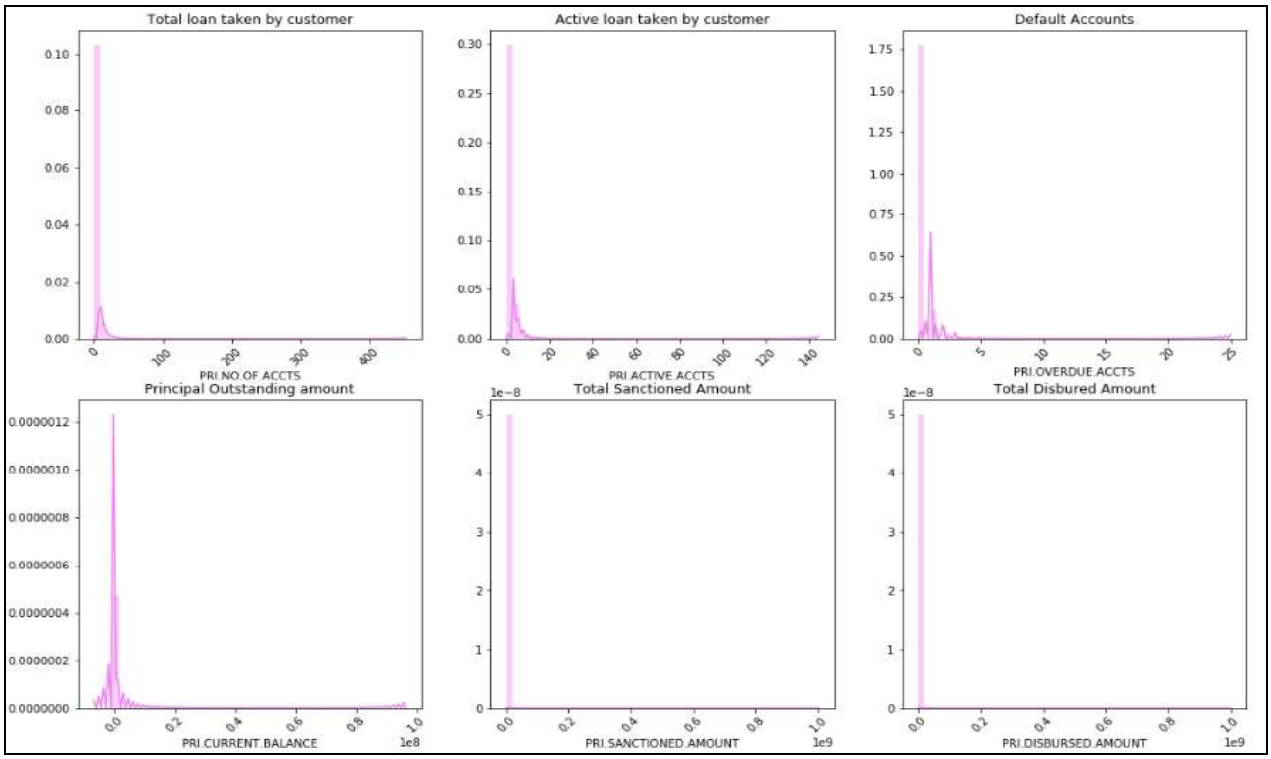

Fig 4: Loan results

In Fig 4 can explains the customer's loan Results.

```
accuracy_score(y_test,y_pred)
```

```
0.7305817261640957
```

```
knn.predict([[50578,58400,89,67,22807,45,1441,1,6,1998,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0]])
```

```
array(['yes'], dtype=object)
```

Fig 5: Accuracy and Prediction

In Fig 5 we are checking the accuracy and predicting the loan will be provided or not. Here the accuracy that we have attained is approximately 73% using KNN Classifier.

Conclusion

We have a loan status model to predict the loan applicant as a valid customer r or default customer. The proposed model shows 75.08% accuracy result in classifying credit applicant using R package. The credit customers based on their payment history and personal information Data mining is the process to discover useful information from large dataset. It consists of classification, clustering and association rule mining.

References

1. Abdelmoula, Aida Krichene. "Bank credit risk analysis with KNearest-neighbor classifier: Case of Tunisian banks." Accounting and Management Information Systems. 2015; 14(1): 79.
2. Arutjothi G, Dr. C Senthamarai. Comparison of Feature Selection Methods for Credit Risk Assessment, International Journal of Computer Science. 2017; 5(1):5.
3. Arutjothi G, Dr. C Senth Amari. Credit Risk Evaluation using Hybrid Feature Selection Method Software Engineering and Technology. 2017; 9(2):23-26.
4. Attig Anja, Petra Perner. The Problem of Normalization and a Normalized Simil arity Measure by Online Data. Tran. CBR. 2011; 4(1):3-17.
5. Babu Ram, A Rama Satish. Impro ved of K-Nearest Neighbour Techniques in Credit Scoring. International al Journal for Development of Computer Science & Technology, 2013, 1.