Intelligent defaulter Prediction using Data Science Process

Chilagani Ravitej¹ and D Teja Santosh²

¹Application Software Associate, Software Development Department, Accenture, Financial District, Nanakramguda, Hyderabad, T.S., India.
ravitejakrish7@gmail.com
²Associate Professor, I.T Department, Sreenidhi Institute of Science and Technology, Yammampet, Ghatkesar, T.S., India.
tejasantoshd@sreenidhi.edu.in

Abstract. The machine learning classifiers employed are not good enough to clearly classify the loan defaulters. In order to alleviate this problem, the data science process is adopted. During the machine learning phase, the random forest classifier with tuned hyperparameters is trained on the loan lending dataset for obtaining the bank loan defaulter model. This classifier is learned by using a smaller number of features as predictor variables obtained after the Exploratory Data Analysis phase. The ability of the model to correctly classify the unseen loan applicant’s data is evaluated in terms of the classifier accuracy and attending the false positives during the model diagnosis from the confusion matrix which unattended will prove fatal for the loan lending banks. The conclusion drawn from this analysis is that the performance metric of the classifier namely the classifier accuracy for Random Forest has outperformed the state-of-art statistical classifiers.

Keywords: Classification, defaulter prediction, Random Forest, statistical learning classifiers

1. Introduction

Loan lending by the financial institutions to the applicant is one among many services as provided by the bank. The individuals due to the increase in the financial constraints, the activity of borrowing money in the form of loan have become inevitable [1]. The small to large firms lend the loans for managing their affairs and to make the system function in the smooth manner [2].

Although this is beneficial to both individual loan applicants and the financial organizations, it does carry great risk [3]. In some situations, the applicant may default the loan (loan that cannot be paid back) and becomes a loan defaulter. This situation when not tackled will lead to liquidity of the bank and finally the bankruptcy [4]. It is understood from a study that the probability of risk in the loan lending process by the financial institutions is high [5]. Throughout the loan lending community, the lenders are in the situation of constant risk in order to minimize the risk of a loan borrower defaulting upon their loan [6]. This is by thoroughly evaluating and verifying the ability of the borrower to satisfy the obligation of fully repaying the loan.

In the machine learning world, this problem is formulated as supervised machine learning task in which the bank loan defaulter is predicted automatically by training the classifiers on the historical loan provided data by using the machine learning algorithms.
The purpose of this study is to adopt the data science process and model the Random Forest classifier, interpret the performance of the model for possible diagnosis and subsequent insights generation and improve the loan defaulter model predictions.

The paper is organized as follows. The related works for carrying out loan defaulter prediction is described in section 2, learning of Random Forest classifier by using Machine Learning workflow and classifier diagnosis is described in section 3, evaluating the accuracy of the learned random forest classifier against the accuracies of the related works is performed in section 4 and conclusion and future scope is specified in section 5.

2. Related Works
Loan defaulter prediction is an essential research area in the field of financial research. There has been a multitude of techniques and classifiers used throughout on this topic. The machine learning algorithms namely Support Vector Machine (SVM), Naive Bayes, Logistic Regression, Decision Trees, Neural Networks and recently the Adaptive Boosting based Random Forest were used for developing the classifiers.

Sudhakar and Reddy used [7] decision trees in order to predict the likelihood of repayment of the loans by the applicants. Vimala and Sharmili proposed [8] a SVM+Naive Bayes hybrid machine learning algorithm in order to predict the loan defaulters. Agbemava et al., implemented [9] Logistic Regression classifier to predict the loan defaulters of non-traditional banks in Ghana. Byanjankar et al., predicted [10] the loan defaulters in peer-to-peer lending by using Neural Network classifier. Recently, Joseph et al., developed [11] Adaptive Boosting based Random Forest classifier to decrease the risk of loan default. The comparison of the classifiers and the exhibited limitations by them are tabulated in Table 1 below.

| Related Work          | Classifier                   | Limitations                      |
|-----------------------|------------------------------|----------------------------------|
| Sudhakar & Reddy [7]  | Decision Tree                | Unstable Decision Tree           |
| Vimala & Sharmili [8] | SVM + Naive Bayes            | Large number of predictors       |
| Agbemava et al., [9]  | Logistic Regression          | Large number of predictors       |
| Byanjankar et al., [10]| Artificial Neural Network    | Need of large amount of training data |
| Joseph et al., [11]   | Adaptive Boosting based Random Forest | Increased complexity of the classifier and large training time |

The authors of this paper to the best of their knowledge strongly present their statement of working with Random Forest classifier in order to better predict the bank loan defaulters.

3. Random Forest Classifier for Predicting the Bank Loan Defaulters Using Data Science Process
The learning of Random Forest classifier starts with asking the Quantitative Question (QQ) on the bank loan defaulter prediction problem. The QQ is as follows: How banks discriminate fully paid bank loan consumers with charged-off bank loan consumers, so that they lend loans in a better manner? The
QQ formulation is one of the steps in the data science process as a part of business understanding. The workflow is illustrated in figure 1 as shown below:

![Data Science process](image)

**Figure 1.** Data Science process.

The QQ drives forward to acquire the data relevant to the business understanding. The dataset acquired is Lending Club Loan Data. The number of features in the dataset is 79 and is considered as potential predictors for the loan defaulters.

The pre-machine learning insights are drawn from the dataset by first cleaning the dataset. This is carried out by first identifying and removing the thin rows and the sparse columns in the dataset by using a statistical threshold. Then replacing the ‘n/a’ cells of certain columns with the relevant universal string. Further, drop certain rows by performing statistical analysis for only having the needed number of records. Finally, the number of features available after cleaning the dataset is 48. The metrics which are of imminent need are derived from the existing columns by constructing them in the form of transformations. These new metrics are useful in uncovering the insights from the dataset. Exploratory Data Analysis (EDA) [12] is performed in the following manner. The univariate, bivariate and multivariate analyses were carried out to obtain the crucial insights. These insights are as follows.

The insights obtained from Univariate Analysis are as follows: Most of the loan amounts are distributed between 8000 to 20000 USD. The visualization of this insight is presented in Figure 2 below.

![Distribution of loan amounts](image)

**Figure 2.** Distribution of loan amounts.

Next, approximately 60% of the loan applicants have applied loan for paying their other loans. The visualization of this insight is presented in Figure 3 below.
Figure 3. Purpose of loan across the loan applicants.

Also, 70% of applicants have 36 months loan term period. The visualization of this insight is presented in Figure 4 below.

Figure 4. Loan Term periods across loan applicants.

The insights obtained from Bivariate/Multivariate Analysis are as follows: The Last payment amount, interest rate and loan amount are highly correlated. The visualization of this insight is presented as a correlation matrix in Figure 5 below.
Next, the probability of a loan applicant being charged-off (derived metric) is high when the applicant employee grade moves towards G from A. The visualization of this insight is presented in Figure 6 below.

Finally, the probability of a loan applicant being charged-off is increasing when the interest rate on the loan is increasing. The visualization of this insight is presented in Figure 7 below.
Figure 7. Probability of the loan applicant being charged off across various interest rates.

The feature selection is done in the following manner after performing EDA as above. The important features useful as predictors from the obtained insights are learned from the decision tree classifier. The final features obtained are last payment amount, loan term, interest rate and the funded amount.

The machine learning algorithm selected for training the classifier is Random Forest algorithm [12]. The Random Forest algorithm is selected because this algorithm reduces the variance in the classifier which in turn improves the classifier accuracy. Also, this algorithm is stable as it is the ensemble of many decision trees. The Random Forest algorithm takes the input predictors and uses the rules from the generated multiple decision trees to predict the outcome. The highest voted predicted target is shown as the final prediction by the algorithm.

The formula for modeling the random forest is given below.

Random forest = base Decision Tree + Bagging + Column sampling + Majority vote ... (1)

The hyper-parameters and the parameters used to train the classifier are tabulated in Table 2 below:

Table 2. Hyper-parameters and Parameters for training Random Forest Classifier.

| Machine Learning input                              | Value considered               |
|------------------------------------------------------|--------------------------------|
| Classifier learned                                  | Random Forest                  |
| Number of predictors                                | 4                              |
| Bootstrapping                                       | True                           |
| Splitting Criterion                                 | Entropy                        |
| Threshold (to prevent misclassification of incoming test data) | 0.03                           |
| Number of Estimators (Decision Tree as Estimator)   | 100                            |
The excerpt of one of the many decision trees from the Random Forest classifier is shown in the figure 8 below.

![Decision Tree Diagram]

**Figure 8.** Excerpt of one of the many decision trees of the Random Forest classifier.

The accuracy of the classifier during the initial run on the test portion of the data is obtained as 89.3%. It is observed from the confusion matrix that the count of false positives is good in number which is of prime concern. In other words, the 6788 loan consumers who are to be predicted as charged-off are wrongly predicted as fully paid consumers.

This is reduced by diagnosing the classifier in order to detect the source which has caused this count. This is called as “Root cause analysis” or “Model diagnosis” [16]. The detected problem is the imbalance in the considered dataset in between the classes. This detected problem is called as postmachine learning insight.

This is addressed by implementing the variation of the traditional re-sampling process with Synthetic Minority Oversampling Technique (SMOTE). The working process of SMOTE technique is given below.

- Initially, the total number of oversampling observations (N) is set up. /* Generally, it is selected such that the binary class distribution is 1:1. But that could be tuned down based on need. Then the iteration starts by first selecting a positive class instance at random.*/
- Next, the K-Nearest Neighbors for that instance is obtained.
- Finally, N of these K instances are chosen to interpolate new synthetic instances. /*To do that, using any distance metric the difference in distance between the feature vector and its neighbors is calculated. Now, this difference is multiplied by any random value in (0,1] and is added to the previous feature vector.*/

It has been observed that there is a significant decrease of 33% test examples being correctly predicted as charged-off. The accuracy of the classifier has increased by 4.1% reaching to 93.4%. The accuracy of the Random Forest classifier when compared with the classifiers from the related works has increased in the most significant manner with a maximum increase in accuracy of 14.4% with [8] and a good increase of 4.4% with [7].

4. Evaluation of the Learned Random Forest Classifier

The comparison of the performance of Random Forest classifier with the classifiers mentioned in the literature section is carried out below.

First, the feature selection techniques incorporated in the related works and in the proposed work for getting predictors are discussed here. These selected features will have the impact on the downstream task, the machine learning and their classifier performance. In other words, the subset of features whose instances contribute to clear distinction at the time of patterns classification are to be identified.

The feature selection technique used in [7], [8] is the ranking technique. The researchers used one of the six statistical and entropy-based feature ranking techniques named Gain Ratio feature selection. The disadvantage of this technique is that there is no guarantee that the selected feature subset will be
used in the learning process. This will have impact on the overall learning process which misclassifies
the some of the test records by the classifier.

In [9], the feature subset is selected by employing the greedy selection algorithms which internally
searches using Depth First Search (DFS) technique. The disadvantage with these algorithms is that the
obtained feature subset is not to be treated as the desired subset as DFS does not examine the alternate
subsets to learn for a better or in some cases the optimal subset. This results in the reduced
performance of the classifier.

In [10], the feature selection technique employed is a measure of proportion contribution by the
Artificial Neural Network which was proposed by Garson in [14]. This method will not give the true
interpretation on the connection weights when there is a combination of positive and negative weights.
These weights will cancel each other when there are similar inputs during the neural computation. This
will remove certain inputs which could become the potential predictor variables.

The feature selection carried out in [11] used Gini index based random forest classifier. The
drawback with Gini index-based feature selection is that different estimators depend on the sample
sizes of the training data. This impacts the selected features which are having less importance in the
patterns learning process by the classifier.

In the proposed work, the subset of features is selected by implementing the random forest
classifier as feature subset selection technique. The features that have an importance of more than 0.15
are selected as predictors. This upper threshold is provided on the basis of EDA provided insights.
This eliminates all kinds of limitations exhibited by the feature selection techniques. Next, the number
of features that were given as input to the machine learning algorithms and the number of features are
actually found as useful predictors of the target class label for bank loan defaulter prediction
is tabulated in Table 3 below.

Table 3. Input features and Predictors.

| Works                  | Number of input features | Number of predictors |
|------------------------|--------------------------|----------------------|
| Sudhakar & Reddy [7]   | 24                       | 6                    |
| Vimala & Sharmili [8]  | 21                       | 6                    |
| Agbemava et al., [9]   | 16                       | 6                    |
| Byanjankar et al., [10]| 14                       | 5                    |
| Joseph et al., [11]    | 20                       | 15                   |
| Proposed Work          | 48                       | 4                    |

It is observed from the above table is that the number of input features that were considered for the
learning process by the researchers from the related works when compared with the proposed dataset
are at a minimum of 41.66% less with [11] and the maximum of 71% less with [10]. On an average
56% of the additional features were missing in their analysis.

The proposed work considers 48 input features after a thorough cleansing process on the dataset by
implementing both the general knowledge and the banking services domain knowledge. This size
specifies that the critical decision of predicting the bank loan defaulter requires tall and wide dataset
which is called as Big Data today.

It is also observed that only 4 features are useful as predictors out of 48 in discriminating fully
paid loan consumers with the charged-off loan consumers. This value is less when compared with the
predictors from the remaining works which helps in providing loan defaulter prediction results in the faster manner. The comparison of the accuracies of the classifiers from the related works with the Random Forest classifier is tabulated in Table 4 below:

| Works                        | Classifier                  | Classifier Accuracy (%) |
|------------------------------|-----------------------------|-------------------------|
| Sudhakar & Reddy [7]         | Decision Tree               | 89%                     |
| Vimala & Sharmili [8]        | SVM + NB                    | 79%                     |
| Agbemava et al., [9]         | Logistic Regression         | 86%                     |
| Byanjankar et al., [10]      | Artificial Neural Network   | 74%                     |
| Joseph et al., [11]          | Adaptive Boosting based Random Forest | 66.6%                 |
| Proposed Work                | Random Forest               | 93.4%                   |

It is observed from the above table that the accuracy of the decision tree in [7] is 89% which is the outcome on the basis of classification on the test portion of the dataset and is not cross-validated. The limitation in the performance of the decision tree classifier is the instability of the learned decision tree when applied on other samples. The accuracies of SVM + NB based hybrid and Logistic Regression classifiers is less in [8], [9] as learning the linear decision boundary has become a challenge as more number of input features are used as input for these algorithms. The accuracy of the Artificial Neural Network classifier in [10] is less than Random Forest classifier accuracy because the Artificial Neural Network classifier outcome varies for different values of the input features. The accuracy of the Adaptive Boosting based Random Forest classifier in [11] is less than the proposed Random Forest classifier as the complexity of the classifier has been increased owing to increased misclassifications. Furthermore, it is learned from [15] that the standard deep learning architecture based classifiers also performed in the poor manner when it comes to credit risk evaluation of the bank loan consumers.

It is understood from the above discussion that the accuracy of the Random Forest classifier in predicting the loan defaulter outperformed the accuracies of all the classifiers specified in the related works section. Also, this is the answer to the QQ which was asked at the start of this work. This is because Random Forest is an ensemble based classifier that becomes strong by combining all the individual performances of the decision trees.

5. Conclusion and Future Scope

The bank loan defaulter prediction using Random Forest classifier was carried out successfully. The objective is to correctly predict the bank loan defaulters in maximum number thereby reducing the false positive rate. The experiment carried out in this work has achieved 93.4% accuracy with the Random Forest classifier on the test portion of the dataset and with 4 predictors. This has outperformed the state-of-art classifiers accuracies.

In future, the deep learning neural network architectures-based classifiers with model diagnosis are investigated in order to improve the accuracy of the classifier and raise the confidence in deploying the deep learning classifiers for better loan defaulter prediction.
References

[1] Perera, H A P L, and Premaratne S C 2016 An Artificial Neural Network Approach for the Predictive Accuracy of Payments of Leasing Customers in Sri Lanka.

[2] Marqués, Ana I, Vicente García, and José Salvador Sánchez 2012 Exploring the behaviour of base classifiers in credit scoring ensembles. Expert Systems with Applications 39.11 10244-10250.

[3] A. Haldorai and A. Ramu, Security and channel noise management in cognitive radio networks, Computers & Electrical Engineering, vol. 87, p. 106784, Oct. 2020. doi:10.1016/j.compeleceng.2020.106784.

[4] A. Haldorai and A. Ramu, Canonical Correlation Analysis Based Hyper Basis Feedforward Neural Network Classification for Urban Sustainability, Neural Processing Letters, Aug. 2020. doi:10.1007/s11063-020-10327-3.

[5] Li H, McCarthy J and Pantalone C, 2014 High-yield versus investment-grade bonds: Less risk and greater returns? Applied Financial Economics 24(20) pp1303–1312.

[6] Pandit A 2016. Data Mining on Loan Approved Dataset for Predicting Defaulters (Doctoral dissertation, Rochester Institute of Technology).

[7] Sudhakar M and ReddyCVK 2016. Two step credit risk assessment model for retail bank loan applications using decision tree data mining technique. International Journal ofAdvanced Research in Computer Engineering & Technology (IJARCET) 5(3) pp705–718.

[8] Vimala S and SharmiliK C 2018 Prediction of loan risk using naive bayes and support vector machine. Int Conf Adv Comput Technol (ICACT) 4 No 2.

[9] Agbemava, Edinam, et al. 2016 Logistic regression analysis of predictors of loan defaults by customers of non-traditional banks in Ghana. European Scientific Journal 12.1.

[10] Byanjankar A, Heikkilä M and Mezei J 2015, Predicting Credit Risk in Peer-to-Peer Lending: A Neural Network Approach. 2015 IEEE Symposium Series on Computational Intelligence, December; IEEE. Pp719–725.

[11] Sanjaya J, Renata E, Budiman V, Anderson F, and Ayub M 2020 Prediksi Kelalaian Pinjaman Bank Menggunakan Random Forest dan Adaptive Boosting. Jurnal Teknik Informatika Dan Sistem Informasi, 6(1). https://doi.org/10.28932/jutisi.v6i1.2313.

[12] Tukey, John W 1977 Exploratory data analysis. Biometrical Journal, 2.https://doi.org/10.1002/bimj.4710230408.

[13] Breiman, Leo 2001 Random forests. Machine learning 45.1: 5-32.

[14] Garson, David G 1991 Interpreting neural network connection weights. AI Expert, 47-51.

[15] Addo, Peter Martey, Dominique Guegan, and Bertrand Hassani 2018 Credit risk analysis using machine and deep learning models. Risks 6.2: 38.

[16] Cielen, Davy, Arno Meysman, and Mohamed Ali 2016 Introducing data science: big data, machine learning, and more, using Python tools. Manning Publications Co.,