Advances and Applications in Mathematical Sciences
Volume 20, Issue 3, January 2021, Pages 427-435
© 2021 Mili Publications

LOAN APPROVAL PREDICTION MODEL A COMPARATIVE ANALYSIS

AFRAH KHAN, EAKANSH BHADOLA, ABHISHEK KUMAR and NIDHI SINGH

ABES Engineering College

Ghaziabad

Abstract

The cost of assets is increasing day by day and the capital required to purchase an entire asset is very high. So purchasing it out of your savings is not possible. The easiest way to get the required funds is to apply for a loan. But taking a loan is a very time consuming process. The application has to go through a lot of stages and it's still not necessary that it will be approved. To decrease the approval time and to decrease the risk associated with the loan many loan prediction models were introduced. The aim of this project was to compare the various Loan Prediction Models and show which is the best one with the least amount of error and could be used by banks in real world to predict if the loan should be approved or not taking the risk factor in mind. After comparing and analysing the models, it was found that the prediction model based on Random Forest proved to be the most accurate and fitting of them all. This can be useful in reducing the time and manpower required to approve loans and filter out the perfect candidates for providing loans.

1. Introduction

A Prediction Model uses data mining, statistics and probability to forecast an outcome. Every model has some variables known as predictors that are likely to influence future results. The data that was collected from various resources then a statistical model is made. It can use a simple linear equation or a sophisticated neural network mapped using a complex software. As more data becomes available the model becomes more refined and the error decreases meaning then it'll be able to predict with the least risk and consuming as less time as it can. The Prediction Model helps the

2010 Mathematics Subject Classification: 60G25, 62M45, 97D10.

Keywords: Lending Decision, Loan Portfolio, Decision Tree, Random Forest, Logistic Regression, Prediction Model, Machine Learning.

Received May 20, 2020; Accepted July 31, 2020

banks by minimizing the risk associated with the loan approval system and helps the applicant by decreasing the time taken in the process.

The main objective of the Project is to compare the Loan Prediction Models made implemented using various algorithms and choose the best one out of them that can shorten the loan approval time and decrease the risk associated with it. It is done by predicting if the loan can be given to that person on the basis of various parameters like credit score, income, age, marital status, gender, etc. The prediction model not only helps the applicant but also helps the bank by minimizing the risk and reducing the number of defaulters.

In the present scenario, a loan needs to be approved manually by a representative of the bank which means that person will be responsible for whether the person is eligible for the loan or not and also calculating the risk associated with it. As it is done by a human it is a time consuming process and is susceptible to errors. If the loan is not repaid, then it accounts as a loss to the bank and banks earn most of their profits by the interest paid to them. If the banks lose too much money, then it will result in a banking crisis. These banking crisis affects the economy of the country. So it is very important that the loan should be approved with the least amount of error in risk calculation while taking up as the least time possible. So a loan prediction model is required that can predict quickly whether the loan can be passed or not with the least amount of risk possible.

2. Literature Survey

The author, Vaidya, Ashlesha [1] uses logistic regression as a machine learning tool in paper and shows how predictive approaches can be used in real world loan approval problems. His paper uses a statistical model (Logistic Regression) to predict whether the loan should be approved or not for a set of records of an applicant. Logistic regression can even work with power terms and nonlinear effect. Some limitations of this model are that it requires independent variables for estimation and a large sample is required for parameter estimation.

A work by Amin, Rafik Khairul and Yuliant Sibaroni [2] was referenced which used Decision tree algorithm called C4.5 to implement a predictive model. This algorithm creates a decision tree that generally gives a high accuracy in decision making problems. Dataset of 1000 cases is used in which 70% is approved and rest is rejected. This paper shows C4.5 algorithm performance in recognizing the eligibility of the applicant to repay his/her loan. From the conducted tests, it is found that the highest precision value is 78.08% which was found using data partition of 90:10. The greatest recall value is 96.4% and was reached with data partition of 80:20. Partition of 80:20 is considered to be best since it has a high recall and the highest accuracy.

The research and work done by Arora, Nisha and Pankaj Deep Kaur [3] aimed at forecasting whether an applicant can be a loan defaulter or not. It uses Bolasso to select most relevant attributes based on their robustness and then applied to classification algorithms like Random Forest, SVM, Naive Bayes and KNearest Neighbours (KNN) to test how accurately they can predict the results. It is concluded that Bolasso enabled Random Forest algorithm (BS-RF) provides the best results in credit risk evaluation and gives better accuracy by using optimised feature selection methods.

In paper authored by Yang, Baoan, et al. [4], the use of artificial neural networks in an early warning system for predicting loan risk is discussed wherein it covers the early warning signals for deteriorating financial situations. The ability of an applicant to repay the loan is determined to be the most relevant aspect in the financial analysis. The early warning system in this paper uses artificial neural network that is utilizing the traditional early warning concepts. This system based on ANN proves to be a very effective decision tool and early warning system for banks and other commercial lending organizations.

The scope of using Genetic Algorithms in building prediction models was also discussed in the paper by Metawa Noura, M. Kabir Hassan and Mohamed Elhoseny [5]. This paper discusses a prediction model made using Genetic Algorithm which can facilitate banks in making lending decisions in case of decrease in lending supply. The main focus of the GA model is two-fold: maximising profit and minimising errors in loan approval in case of dynamic lending decisions. Several factors like type of loan, rating of creditor and expected loan loss are integrated to GA chromosomes and then validation is done. The result shows that GAMCC increases the profits of the bank by 3.9% to 8.1%.

Yet another approach was used by Hassan, Amira Kamil Ibrahim and Ajith Abraham[6] wherein they used German dataset and built a prediction model working basically on backpropagation and implemented with three different back propagation algorithms. They also used two different methods for two filtering functions for the attributes which resulted in DS2 giving highest accuracy using PLsFi filtering function.

3. Proposed Methodology

The paper 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.

4. Implementation

4.1. Flow Chart

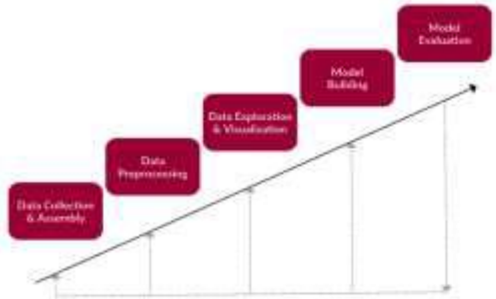


Figure 1. Flow chart.

4.2. Data collection and importing

The training set was imported in csv format and a simple function is applied to check whether it's working or not.

Figure 2. Importing and checking training data.

4.3. Study of distribution of attribute

Box plot and histogram are used for study of distribution factors. In the snapshot below one such factor (applicant income) has been used as an example.

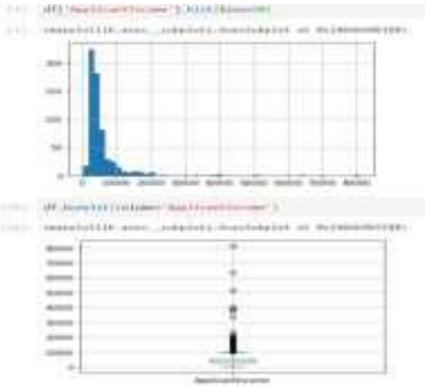


Figure 3. Study of distribution of data.

There are many extreme values due to income gap and difference in education levels.

4.4. Categorical Variable analysis

The following snapshot shows the method to calculate the chances of getting a loan on the basis of credit history.

Figure 4. Checking importance of data.

4.5. Plotting of graph to infer more results

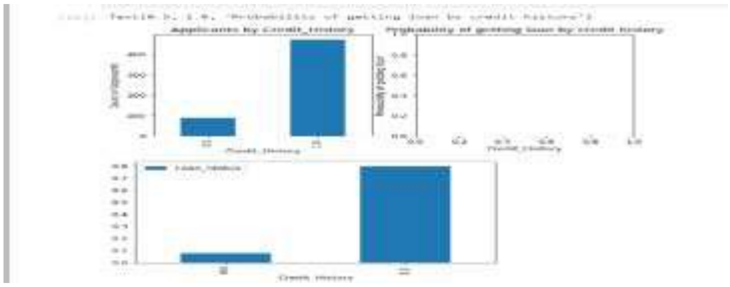
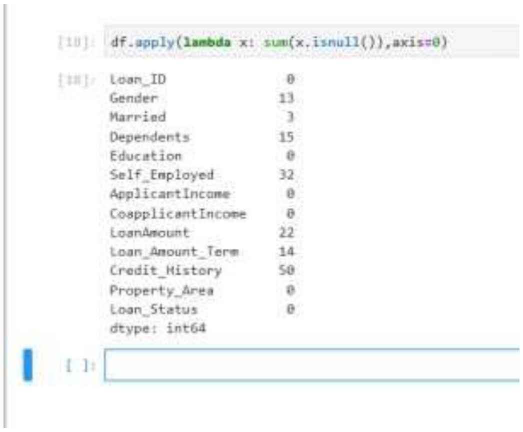


Figure 5. Plotting relation between factor and result to recognise patterns.

The graph depicts that it's eight times easier to get a loan if an applicant has a valid credit history.

4.6. Checking missing values

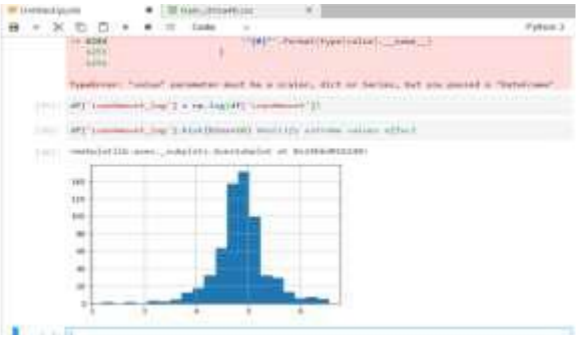
Data is being processed so that it can be determined how many values are missing from each column. The count of missing values present in the non-numerical attributes are processed and computed using the statistics.



```
[10]: df.apply(lambda x: sum(x.isnull()),axis=0)

[10]: Loan_ID              0
      Gender              13
      Married              3
      Dependents          15
      Education            0
      Self_Employed       32
      ApplicantIncome      0
      CoapplicantIncome    0
      LoanAmount          22
      Loan_Amount_Term    14
      Credit_History      50
      Property_Area        0
      Loan_Status          0
      dtype: int64

[ ]:
```

Figure 6. Checking missing values.

4.7. Finding extreme values and nullifying them

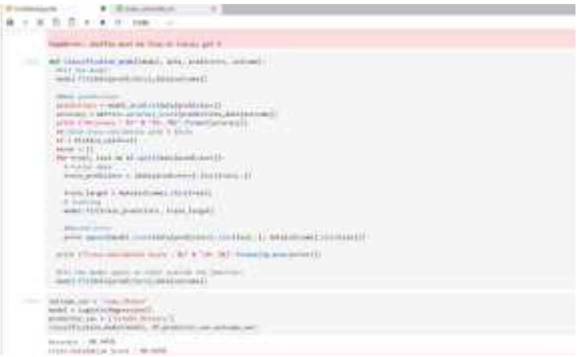


```
-> [illegible]            "{0}".format(type(value).__name__)
   [illegible]
   [illegible]

TypeError: "value" parameter must be a scalar, dict or Series, but you passed a "DataFrame"

[ ]: df['LoanAmount_log'] = np.log(df['LoanAmount'])

[ ]: df['LoanAmount_log'].hist(bins=20) #nullify extreme values effect

[ ]: <matplotlib.axes._subplots.AxesSubplot at [illegible]>
```

Figure 7. Nullifying extreme values.

4.8. Building predictive model using Logistic Regression

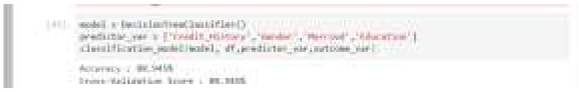
A generic classification function is defined whose input is a model that helps determine the crossvalidation scores and accuracy using *K*-fold method.



```
[illegible]

def classification_model(model, data, predictors, outcome):
  #Fit the model:
  model.fit(data[predictors],data[outcome])

  #Make predictions on training set:
  predictions = model.predict(data[predictors])
  accuracy = metrics.accuracy_score(predictions,data[outcome])
  print ("Accuracy : %s" % "{0:.3%}".format(accuracy))
  #Perform k-fold cross-validation with 5 folds
  kf = KFold(n_splits=5)
  error = []
  for train, test in kf.split(data[predictors]):
    # Filter training data
    train_predictors = (data[predictors].iloc[train,:])

    train_target = data[outcome].iloc[train]
    # Training
    model.fit(train_predictors, train_target)

    #Record error
    error.append(model.score(data[predictors].iloc[test,:], data[outcome].iloc[test]))

  print ("Cross-Validation Score : %s" % "{0:.3%}".format(np.mean(error)))

  #Fit the model again so that it can be refered outside the function:
  model.fit(data[predictors],data[outcome])

outcome_var = 'Loan_Status'
model = LogisticRegression()
predictor_var = ['Credit_History']
classification_model(model, df,predictor_var,outcome_var)

Accuracy : 80.945%
Cross-Validation Score : 80.946%
```

Figure 8. Predictive model using Logistic Regression using Credit History.

4.9. Building predictive model using Decision Tree

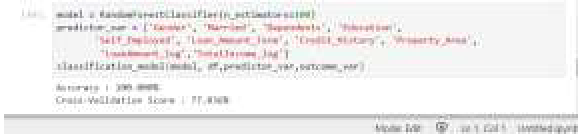


```
[44]: model = DecisionTreeClassifier()
      predictor_var = ['Credit_History','Gender','Married','Education']
      classification_model(model, df,predictor_var,outcome_var)

      Accuracy : 80.945%
      Cross-Validation Score : 80.946%
```

Figure 9. Predictive model using Decision Tree.

4.10. Building predictive model using Random Forest

Model based on Random forest has an advantage that it can find the most important features that greatly affect the accuracy of the result among all the features using feature importance matrix.

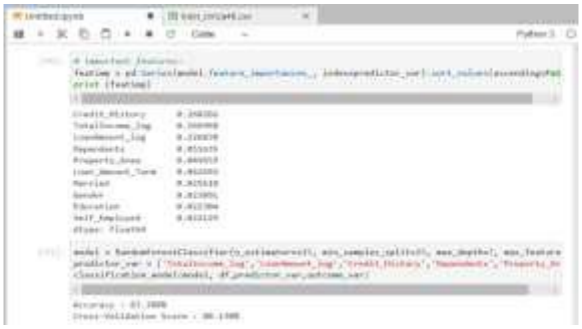


```
[ ]: model = RandomForestClassifier(n_estimators=100)
     predictor_var = ['Gender', 'Married', 'Dependents', 'Education',
            'Self_Employed', 'Loan_Amount_Term', 'Credit_History', 'Property_Area',
            'LoanAmount_log','TotalIncome_log']
     classification_model(model, df,predictor_var,outcome_var)

     Accuracy : 100.000%
     Cross-Validation Score : 77.036%
```

Figure 10. Predictive model using Random Forest.

4.11. Improving random forest predictive model

Since 100% accuracy was observed in the previous results, the most important features are taken from the feature importance matrix now to avoid overfitting.



```
# important features
featimp = pd.Series(model.feature_importances_, index=predictor_var).sort_values(ascending=False)
print (featimp)

Credit_History       [illegible]
TotalIncome_log      [illegible]
LoanAmount_log       [illegible]
Dependents           [illegible]
Property_Area        [illegible]
Loan_Amount_Term     [illegible]
Married              [illegible]
Gender               [illegible]
Education            [illegible]
Self_Employed        [illegible]
dtype: float64

model = RandomForestClassifier(n_estimators=[illegible], min_samples_split=[illegible], max_depth=[illegible], max_feature
predictor_var = ['TotalIncome_log','LoanAmount_log','Credit_History','Dependents','Property_Ar
classification_model(model, df,predictor_var,outcome_var)

Accuracy : 83.388%
Cross-Validation Score : 80.130%
```

Figure 11. Improvised predictive model using Random Forest.

5. Conclusion

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] Vaidya and Ashlesha, Predictive and probabilistic approach using logistic regression: Application to prediction of loan approval, 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT). IEEE, 2017.

[2] Amin, Rafik Khairul and Yuliant Sibaroni, Implementation of decision tree using C4. 5 algorithm in decision making of loan application by debtor (Case study: Bank pasar of Yogyakarta Special Region), 2015 3rd International Conference on Information and Communication Technology (ICoICT). IEEE, 2015.

[3] Arora, Nisha and Pankaj Deep Kaur, A Bolasso based consistent feature selection enabled random forest classification algorithm: An application to credit risk assessment, Applied Soft Computing 86 (2020), 105936.

[4] Yang, Baoan, et al, An early warning system for loan risk assessment using artificial neural networks, Knowledge-Based Systems 14.5-6 (2001), 303-306.

[5] Metawa, Noura, M. Kabir Hassan and Mohamed Elhoseny, Genetic algorithm based model for optimizing bank lending decisions, Expert Systems with Applications 80 (2017), 75-82.

[6] Hassan, Amira Kamil Ibrahim and Ajith Abraham. "Modeling consumer loan default prediction using ensemble neural networks, 2013 International Conference On Computing, Electrical And Electronic Engineering (ICCEEE). IEEE, 2013.