International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 8 Aug 2021, pp: 741-745 www.ijaem.net ISSN: 2395-5252

Loan Prediction using Machine Learning

N.Jayalakshmi* S.Subhash** K.Mallesh** G.Hanesha** Rahul**
*Assistant Professor
**IV CSE - Student
Department of Computer Science and Engineering Raghu Institute of Technology (Autonomous) , Visakhapatnam, A.P,india

Submitted: 01-08-2021 Revised: 10-08-2021 Accepted: 15-08-2021

ABSTRACT:
In the banking sector lots of people are applying for bank loans but which it has to grant to limited people only. To find out to whom the loan can be granted which will be a safer option for the bank. This is done by comparing data of the previous records of the people to which the loan was granted before and on the basis of these records the machine was trained using the machine learning model which gives the most accurate result. We predict the loan data by using decision tree algorithm,linear regression and random forest.Both the algorithms have been used on the same dataset and the conclusions have been made with results showing that the Random Forest algorithm outperformed the Decision Tree algorithm with much higher accuracy.

Keywords: Data Analysis, Loan Prediction, Machine Learning,Random Forest,Decision Tree

I. INTRODUCTION:

The main problem in banking system is to invest their assets in safe hands where it is. Banks or financial companies approves loan after a regress process of verification and validation but still there is no surety whether the chosen applicant is right applicant or not.Loan Prediction is very helpful for banks as well as for the applicant also. The main objective is to predict whether assigning the loan to a particular person will be safe or not We implemented this loan prediction problem using Decision tree algorithm and random forest. once we find the missing values we replace those missing values with mean and mode of the relevant attributes.The main aim of this project is to compare different applications of machine learning models and then chose which one is more accurate.Classifiers that we used to build the model are decision tree and random forest. They were used separately to analyze the same dataset and identify the patterns in the dataset and learn from those. Based on that analysis, predict whether a new applicant is likely to default on a loan or not.

II. LITERATURE REVIEW:

Loan prediction is a much-talked-about subject in the sectors of banking and finance. Credit scoring has become a key tool for the same in this competitive financial world. Furthermore, following the recent improvements in data science and several notable developments in the field of artificial intelligence, this topic has gained more attention and research interest. In recent years, it has attracted more focus towards research on loan prediction and credit risk assessment. Due to the high demands of loan now, demand for further improvements in the models for credit scoring and loan prediction is increasing significantly. A multitude of techniques have been used to assign individuals a credit score and much research has been done over the years on the topic. Unlike previously, where experts were hired and the models depended on professional opinions were used for assessing the individual's creditworthiness, the focus has shifted to an automated way of doing the same job. In recent years, the researchers

III. MACHINE LEARNING:

It is a concept that enables machines to learn from real-world interactions and observations and behave like human beings and improve their ability to learn and perform using data given as input. In the recent years, ML has gained a huge focus and interest of researchers and technologists that they are trying to implement various machine learning models and algorithms in fields which will make various important tasks and lives of common man. Two popular examples are the banking sector and finance. With the help of various ML models, banking authorities and financial firms are observing patterns and making conclusions in areas like credit card frauds, loan default prediction. It has made the process much easier now and more accurate. The models mentioned above are based on various machine

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 8 Aug 2021, pp: 741-745 www.ijaem.net ISSN: 2395-5252

learning methods. It is almost impossible to compile and provide a list of all the ML methods. Usually, the name given to a model is a combination of data structure, design, estimator, ensemble mechanism, and more [6]. Regarding thispaper, the two algorithms in the domain of machine learning used are Random Forest and Decision Trees.

3.1 Decision Trees:

There are many versatile algorithms used to perform the tasks of classification and regression. One of the most popular algorithms used for classification are Decision trees, which comprise several branches, leaf nodes, and root nodes. This algorithm generates a structure like a tree by classifying the instances and utilizing a Recursive Portioning Algorithm (RPA). A class label is represented by a leaf node and the branches represent test results. These tests are represented by internal nodes for an attribute.

3.2 Random Forest:

Random Forest belongs to the supervised learning algorithm. Like decision trees, they are also used for classification and regression. A predictor ensemble is built with several decision trees that expand in randomly selected data subspaces.

IV. METHODOLOGY:

Since the prediction of loan is an important research field, there were many different algorithms and techniques.we compare different machine learning models on the data set in order to find which algorithm is best.Machine Learning techniques are very useful in predicting outcomes for large amount of data.

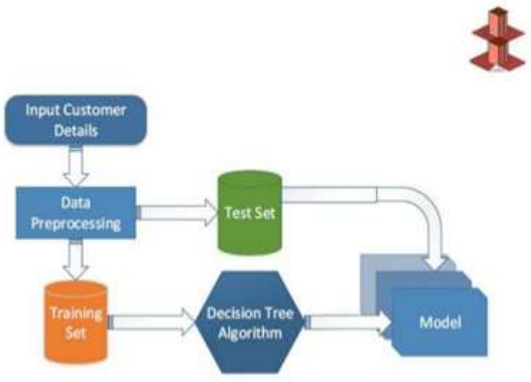
The three machine learning algorithms, Logistic Regression, Decision Tree and Random Forest are applied to predict the loan approval of customers. The experimental results conclude that the accuracy of Decision Tree machine learning algorithm is better as compared to Logistic Regression and Random Forest machine learning approaches.

The methodology adopted for predicting loan using Decision tree technique is shown below. The steps involved in Building the data model is depicted below

The below architecture was defined for decision tree same type of modes was also used for random forest classifier.

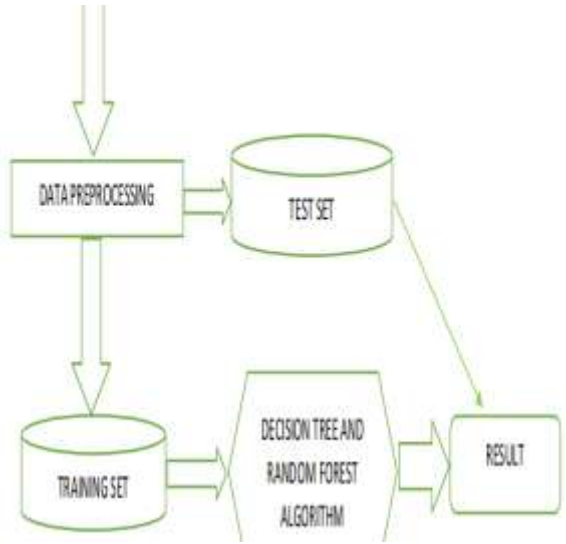
This was the data we used for this is

Download link: https://www.kaggle.com/altruistdelhite04/loan-prediction-problem-dataset



Architecture of Proposed Model

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 8 Aug 2021, pp: 741-745 www.ijaem.net ISSN: 2395-5252



Testing and Training:
In this work we used 75 percent for training set and 25 percent for testing the data.
We also used cross validation to improve the accuracy. if we use cv = 3 the two parts for training set and 1 part for testing set. if we used cv = 5 then four parts for training set and 1 part for testing. We used cv = 5.

V. IMPLEMENTATION:

Data cleaning: There are many columns with null values in the dataset. It is necessary to identify the percentage of null values in each column to drop certain columns that don't meet a percentage threshold. Data cleaning needs to be done before performing the Exploratory Data Analysis.

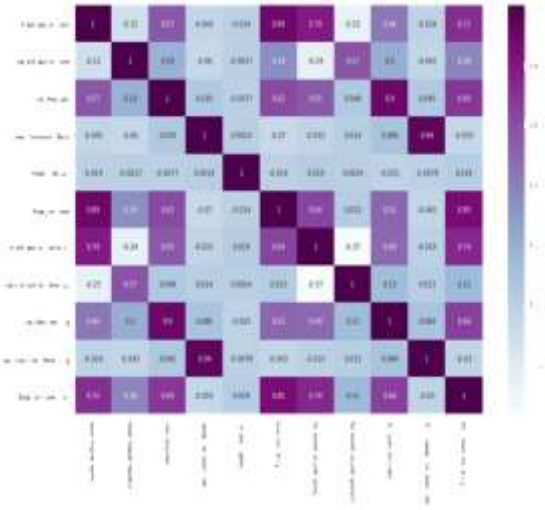
Data columns (total 13 columns)
: # Column Non-Null Count Dtype
0 Loan_ID non-null object
1 Gender non-null object
2 Married non-null object
3 Dependents non-null object
4 Education non-null object
5 Self_Employed non-null object
6 ApplicantIncome non-null int64
7CoapplicantIncome non-null float64
8 LoanAmountnon-null float64
9Loan_Amount_Term non-null float64
10 Credit_Historynon-null float64
11 Property_Areanon-null object
12 Loan_Statusnon-nullobject

Count of null values before cleaning:
Loan_ID0
Gender13
Married 3
Dependents 15
Education 0
Self_Employed32
ApplicantIncome0
CoapplicantIncome 0
Loan_Amount22
Loan_Amount_Term 14
Credit_History 50
Property_Area0
Loan_Status0
D_type: int64

After finding the missing values we fill values for numerical terms with the mean and the categorical values with mode.

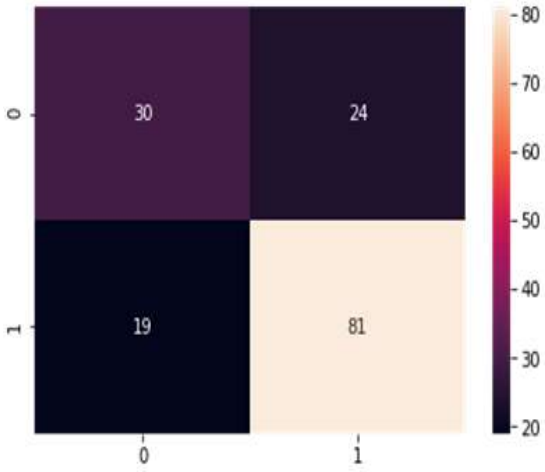
Correlationheap between all features:

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 8 Aug 2021, pp: 741-745 www.ijaem.net ISSN: 2395-5252



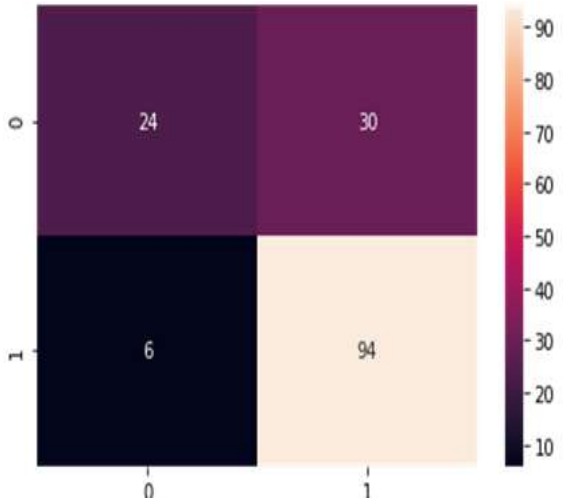
In this work, we used two machine learning algorithms, the Random Forest and Decision Trees to work out a model for loan prediction.
The results of both the model are shown below with their classification report and confusion matrix to get a better understanding of the accuracy and other scores of the two models.

Decision tree:
The decision tree classifier gives an accuracy of 72.727.By doing Cross validation it gives 71.686
Confusion matrix for decision tree:



Random forest:
The random forest classifier gave us an accuracy of 77.272Cross validation is 78.341
Confusion matrix for random forest:

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 8 Aug 2021, pp: 741-745 www.ijaem.net ISSN: 2395-5252



After doing Hyper parameter tuning to random classifier the accuracy
Becomes 76.623
Cross validation is 79.986

VI. CONCLUSION:

This paper aimed to explore, analyze, 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 72%. Hence, the Random Forest model appears to be a better option for such kind of data.

REFERENCES:

[1]. Short-term prediction of Mortgage default using ensembled machine learning models, Jesse C. Sealand on july 20, 2018.
[2]. Research on bank credit default prediction based on data mining algorithm, The International Journal of Social Sciences and Humanities Invention 5(06): 4820-4823, 2018.
[3]. R. O. Duda, P. E. Hart, and D. G. Stork, Pattern classification. John Wiley & Sons, 2012.
[4]. I. H. Witten and E. Frank, Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2005.
[5]. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Van- derplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
[6]. J. Li, H. Wei, and W. Hao, "Weight-selected attribute bagging for credit scoring," Mathematical Problems in Engineering, vol. 2013,
[7]. R. K. Mahapatra, "Business data mininga machine learning perspective," Information &management, vol. 39, no. 3, pp. 211–225, 2001.
[8]. Alshouiliy K, Alghamdi A and Agrawal D P 2020 AzureML based analysis and prediction loan borrowers creditworthy The 3rd Int. Conf. on Information and Computer Technologies (ICICT) 1 pp 302–6
[9]. Li M, Mickel A and Taylor S 2018, "Should this loan be approved or denied?": a large dataset with class assignment guidelines Journal of Statistics Education 26 pp 55–66
[10]. Murphy K P 2012 Machine learning: a probabilistic approach