**Loan Approval prediction using Predictive Analytics**

**Team Members:**

***19BDS0103 YASHASVA SHARMA***

***19BDS0158 TUSHAR KRISHNAN***

**Report submitted for the**

**First Project Review of Loan Approval Prediction**

**Course Code: CSE3045  
Predictive Analysis**

**Slot: A2 Slot**

**Professor: Dr. Ilanthenral Kandasamy**

**Introduction**

Loans have made our life easier, providing us with the financial leverage that extends beyond our earnings. Loans are the credit extended to us by lenders on fulfilling certain key parameters. However, getting a loan can often be a tedious process for the uninitiated, but not for individuals with a good credit score. Whenever you apply for a loan, banks check your CIBIL Score and Report to evaluate your credit history and credit worthiness. The higher your score the better are the chances of your loan application getting approved. Apart from your CIBIL Score, loan eligibility criteria differs from lender to lender and across loan types. ​​Predicting credit defaulters is a difficult task for the banking industry. The system approves or rejects the loan applications. Recovery of loans is a major contributing parameter in the financial statements of a bank. It is very difficult to predict the possibility of payment of loan by the customer. With the help if predictive analytics we are able to see the relation in data which usual

**2. Literature Review Summary Table**

Kindly go through projects, review papers related to your project and study them. **Minimum at least five** projects/papers should be reviewed so that you have a considerable understanding of what is achieved in your project area. (After 2018)

| *Authors and Year (Reference)* | *Title (Study)* | *Concept / Theoretical model/ Framework* | *Methodology used/ Implementation* | *Dataset details/ Analysis* | *Relevant Finding* | *Limitations/ Future Research/ Gaps identified* |
| --- | --- | --- | --- | --- | --- | --- |
| *J. Tejaswini1,*  *T. Mohana, Kavya2,*  *R. Devi Naga Ramya3,*  *P. Sai Triveni4*  *Venkata Rao Maddumala5*  *Vol 11, Issue 4, April/ 2020*  *ISSN NO: 0377-9254* | *ACCURATE LOAN APPROVAL PREDICTION BASED ON*  *MACHINE LEARNING APPROACH* | *Logistic*  *Regression (LR), Decision Tree (DT), Random*  *Forest (RF)*  *Support Vector Machine (SVM), Adaboost (ADB) and Neural Network (Net), are applied to predict the loan approval*  *of customers.* | *Six machine learning classification models have been*  *used for the prediction of android applications. The*  *models are available in R open-source software.*  *After the operation of testing, the model*  *predicts whether the new applicant is a fit case for*  *approval of the loan or not based upon the inference*  *it concludes on the basis of the training data sets.* | *-* | *It can be*  *concluded that the product is a highly efficient*  *component. This application is working properly and*  *meeting to all Banker requirements. This component*  *can be easily plugged in many other systems.* |  |
| *L. Udaya Bhanu1, Dr S. Narayana2*  *International Journal of Scientific and Research Publications, Volume 11, Issue 6, June 2021 403*  *ISSN 2250-3153* | *Customer Loan Prediction Using Supervised Learning Technique* | ***Model***  *Scikit learn*  *Random forest Model*  *Logistic Regression*  *K-nearest neighbour (KNN)*  *Support Vector Machine*  ***Preprocessing***  *MinMax scalar* | *Machine Learning, using semi-automated extraction of knowledge of data for identifying whether a loan would be approved or not.*  *Classification could be supervised learning within which the response is categorical; that is, its values are a unit in a finite unordered set.*  *MinMaxScalar, for each value in every feature MinMaxScalar cipher the minimum value within the feature then divided by the vary. The range is the distinction between the first most and original minimum. It preserves the shapes of the first original distribution.* | *-* | *Optimum results were obtained using Logistic Regression, Random Forest, KNN, and SVM, decision Tree Classifier. Compared to these five algorithms, random forest is more accurate. From a correct analysis of positive points and constraints on the part, it can be safely concluded that the merchandise could be an extremely efficient part. This application is functioning properly and meeting all or any Banker necessities. This part is often simply obstructed in several different systems.* |  |
| *Ms Kathe Rutika Pramod,*  *Ms Panhale Sakshi Dattatray,*  *Ms Avhad Pooja Prakash, Ms Dapse Punam Laxman,*  *Mr Ghorpade Dinesh B.*  *Volume 9, Issue 6 June 2021 | ISSN: 2320-2882* | *An Approach For Prediction Of Loan Approval Using Machine Learning Algorithm* | *Decision tree*  *logistic artificial neural network, logistic regression and support vector machine.*  *Missing value imputation with mice packages* | *Uses a prediction model which is constructed using three different training algorithms to train a supervised two-layer feed-forward network. The results show that the training algorithm improves the design of the loan default prediction model.*  *Decision tree algorithms in machine learning methods efficiently perform both classification and regression tasks. It creates decision trees. Decision trees are widely used in the banking industry due to their high accuracy and ability to formulate a statistical model in plain language.* | *-* | *Applicants with Credit history not passing fails to get approved, Probably because that they have a probability of not paying back. Most of the Time, Applicants with high income sanctioning low amounts are more likely to get approved which make sense, more likely to pay back their loans.* |  |
| *Amruta S. Aphale,*  *Prof. Dr Sandeep. R. Shinde*  *International Journal of Engineering Research & Technology (IJERT)*  *Vol. 9 Issue 8, August-2020* | *Predict Loan Approval in Banking*  *System Machine Learning Approach*  *for Cooperative Banks Loan Approval* | *Neural Networks, Discriminant Analysis,*  *Naive Bayes, K-Nearest Neighbor,*  *Linear Regression,*  *Ensemble Learning/method* | *Classification algorithms is good for*  *grouping data that are never seen before into their various*  *groupings and are therefore extensively used in machine*  *learning tasks.*  *The proposed model focuses on predicting the credibility*  *of customers for loan repayment by analyzing their behaviour.*  *The input to the model is the customer behaviour collected. On*  *the output from the classifier, the decision on whether to approve*  *or reject the customer request can be made.* | *-* | *The experiment revealed that,*  *apart from the Nearest Centroid and Gaussian Naive Bayes,*  *the rest of the algorithms perform credibly well in term of*  *their accuracy and other performance evaluation metrics.*  *Each of these algorithms achieved an accuracy rate between*  *76% to over 80%. We also determined the most important*  *features that influence the credit worthiness of customers.* |  |
| *Dr C K Gomathy, Ms Charulatha, Mr AAkash, Ms Sowjanya*  *International Research Journal of Engineering and technology me: 08 Issue: 10 | Oct 2021* | *THE LOAN PREDICTION USING MACHINE LEARNING* | *Decision Tree* | *Decision tree algorithm efficiently performs both family and retrogression tasks. It creates decision trees. Decision trees are universally used in banking assiduousness due to their high exactitude and capability to formulate a statistical model in plain language. In the Decision tree each knot represents a criterion (diagnostic), each link (branch) represents a decision (rule) and each chip represents an ounce (categorical or continuous value).* | *-* | *From a proper analysis of positive points and constraints on the member, it can be safely concluded that the product is a considerably productive member. This use is working duly and meeting all Banker requisites.* |  |

**3. Objective of the project:**

As mentioned, loans, once a niche service for the rich have become increasingly popular for many people these days. As such, the process to approve a loan can be very tedious, both in terms of work hours for bank staff, and also be especially stressful for the applicants as they wait for their applications to be accepted or rejected.

Thus, we have decided to implement a model to completely automate this aspect of the application process. Based on the applicant's gender, marital status, education, dependents, employment status, income and loan duration, the model predicts whether the loan should be approved or rejected, thus expediting the process for everyone involved.

**4. Innovation component in the project:**

On the face value its hard to discover what new can be done with a simple application that predicts whether a loan is to be sanctioned or not, but once we brainstrom over the application of this technology, the innovative part of this project is ease of integration of this technology into other aspects of technology, with the rise of unregulated currency it will be hard to give out loans for the banks as banks do not have control on the flow of currency, we can integrate this technology in peer to peer loan providing service in which one can ask for a loan and depending on his/her record the algorithm will tell us whether to give that individual a loan or its a bad deal and it would increase the flow of currency without any help of banks.

**5. Work done and implementation**

**a. Methodology adapted:**

First, we need to perform data exploration and preparation. There are 7 steps for this:

1. Variable Identification - identify the predictor variable and the target variable
2. Univariate Analysis - depends on whether the variable type is continuous or categorical
3. Bi-variate Analysis - this finds the relationship between two variables
4. Missing values treatment - Missing values can reduce the fit of a model or lead to a biased model. Thus, we must find a way to deal with the missing values. There are a few ways to do this, such as deletion, imputation with mean, median or mode, or prediction. All of these methods have their own advantages and disadvantages.
5. Outlier treatment - simply put, an outlier is an observation that appears far away from the rest of the data, and diverges from the overall pattern significantly. These outliers can be of two types: Univariate or Multivariate, and can significantly alter the results of the modelling. Outliers can be detected using various visualization techniques such as box plots, histograms, or scatter plots. As a general rule of thumb, any value, which is beyond the range of -1.5 x IQR to 1.5 x IQR (IQR being the interquartile range) is considered to be an outlier, although simpler rules can also be used. Some methods to deal with outliers are Deletion, Transformation, Imputation, or treating them separately.
6. Variable transformation - Now we can proceed to feature engineering, which is essentially extracting more information from existing data. The first step is variable transformation, which refers to the replacement of a variable by a function. In other words, transformation is a process that changes the distribution or relationship of a variable with others. Some common methods for this are Logarithm, square/cube root, or binning.
7. Variable creation - It is a process to generate a new variable based on existing variables. This process is especially useful to highlight hidden relationships in a variable. We can do this by creating derived variables, or creating dummy variables.

The Methodology used should be discussed in detail. Hardware and software requirements must also be mentioned.

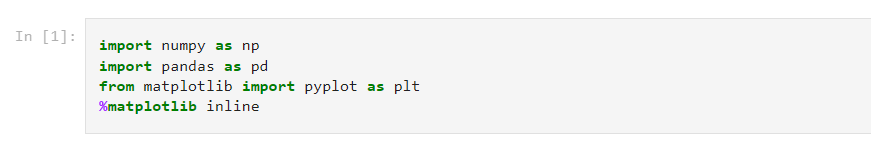
**b. Dataset used:**

The dataset for this problem has been taken from Kaggle.

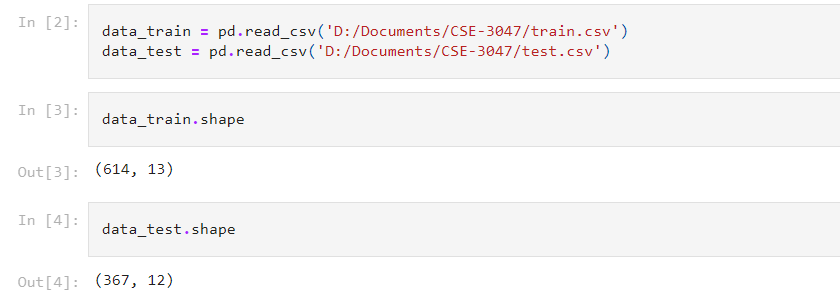
There are 12 columns, and 981 unique entries. There are missing values in some entries, which will need to be dealt with in pre-processing and cleaning.

**c. Tools to be used:** We will be using Jupyter Notebook for this project since it is a very feature-rich, robust and user-friendly environment for data-related projects. We will be using libraries like numpy and pandas.

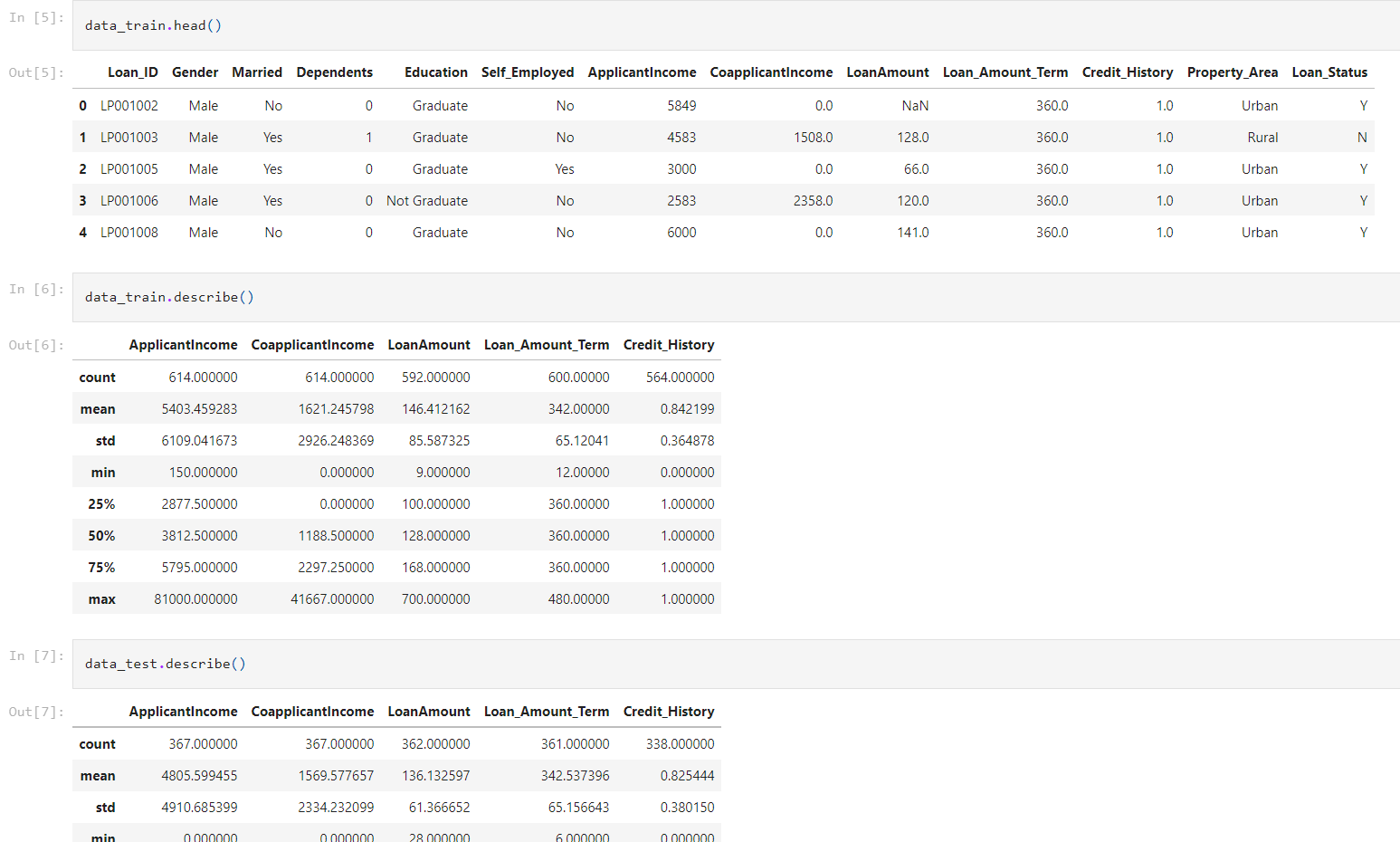
**d. Screenshot and Demo along with Visualization:** Each result and necessary coding part should be substantiated with a related screenshot.

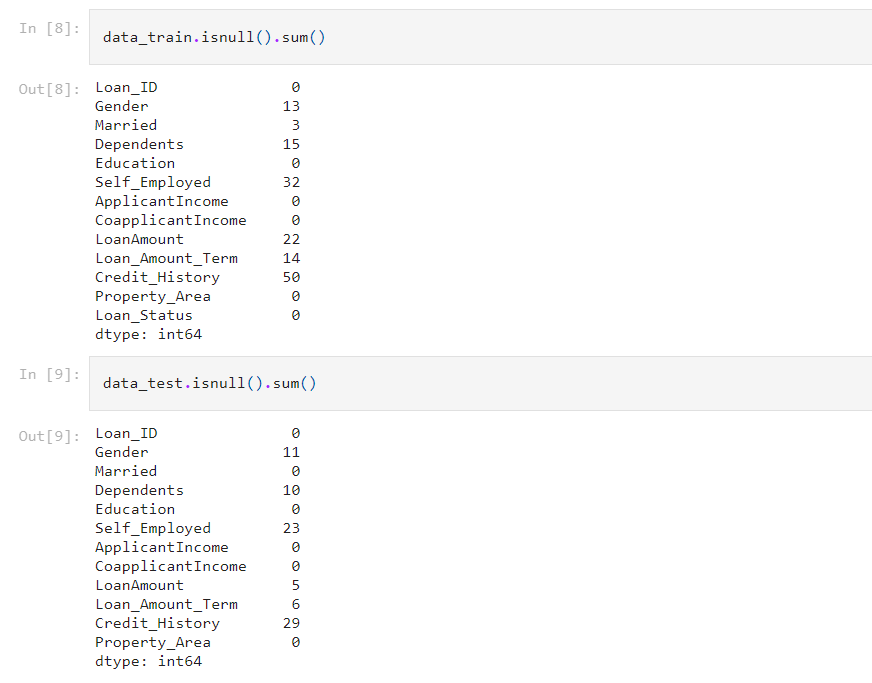


* Importing the necessary libraries to perform data preprocessing

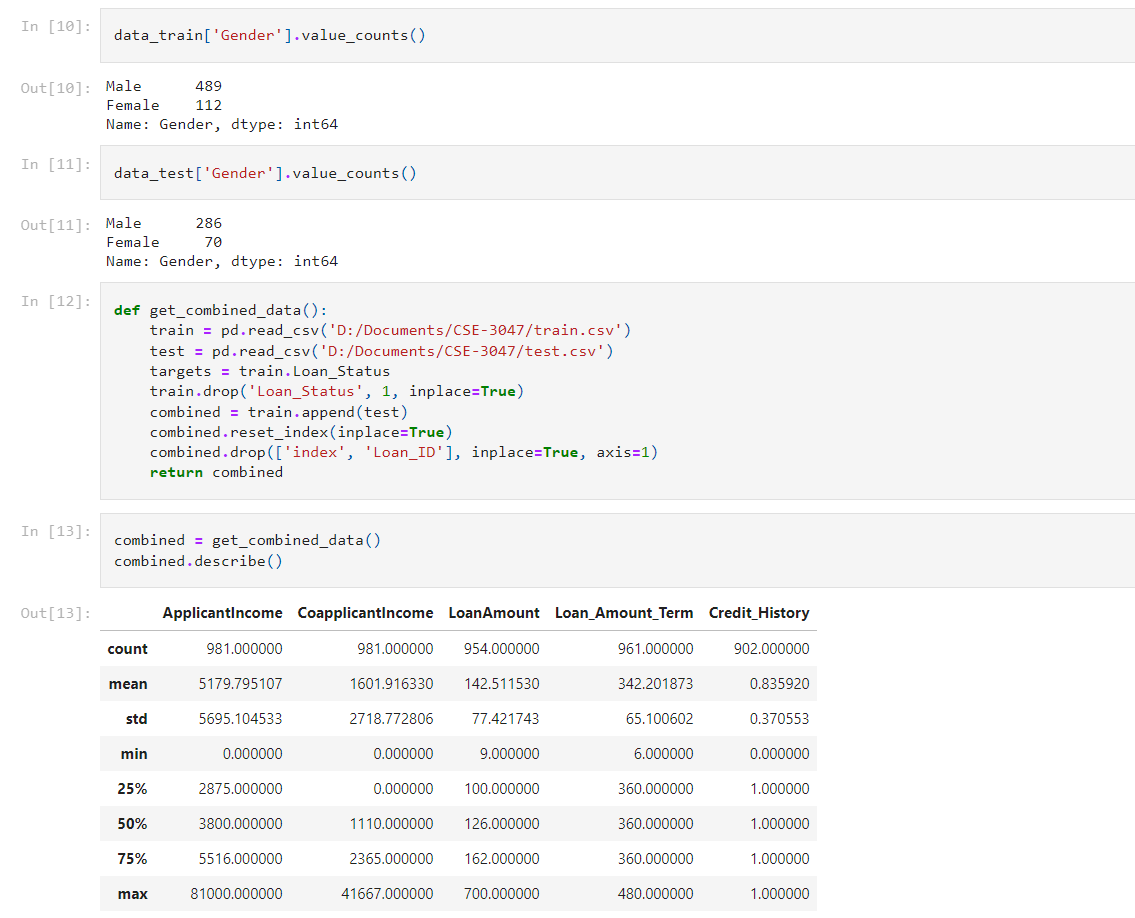


* Reading the training and test data and checking their rows and columns.



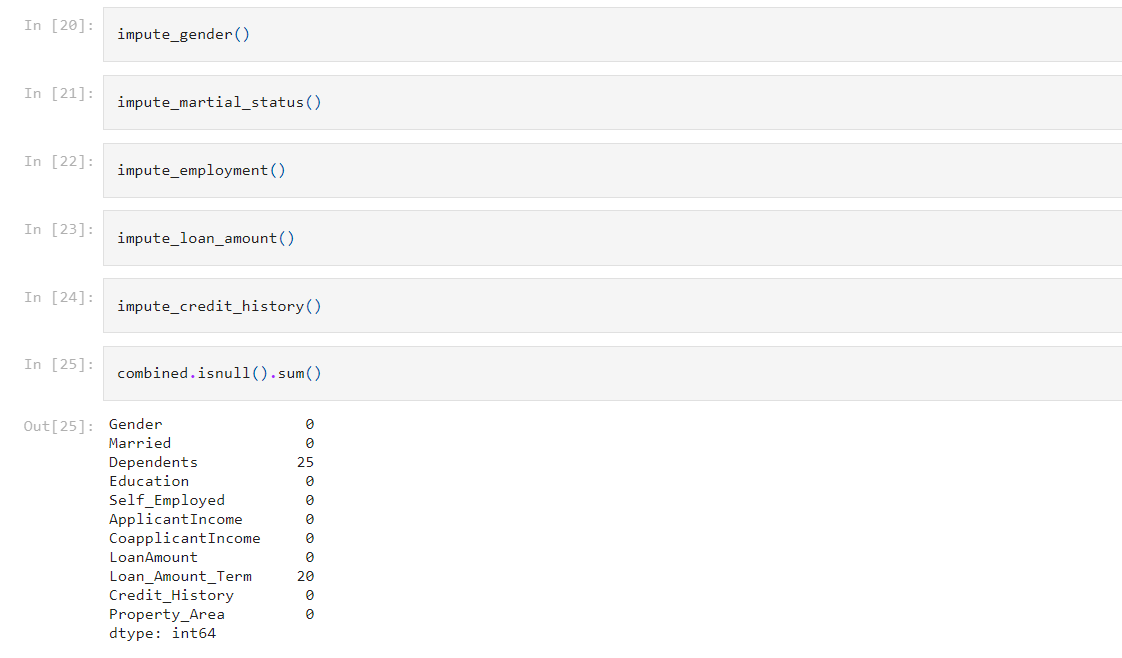


* Checking the number of null (missing) elements in the training and test data.



There are many ways to deal with missing data, as discussed. We shall be imputing the missing entries.





**6. Expected Results**

The program will take the input data for all the loan applicants, and in turn, return a yes or no in a separate column as to whether the loan should be approved or not for each applicant.

**7. References - IEEE std.**

* **Yu Jin and Yudan Zhu, “A data-driven approach to predict default risk of loan for online Peer-to-Peer (P2P) lending,” School of Information, Zhejiang University of Finance and Economics, 310018 Hangzhou, China.**
* **Bhoomi Patel, Harshal Patil, Jovita Hembram, Shree Jaswal “Loan default forecasting using data mining” Department of Information Technology, St. Francis Institute of Technology, Mumbai, India (2020)**
* **Venkata Rao Maddumala, R. Arunkumar, and S. Arivalagan (2018)“An Empirical Review on Data Feature Selection and Big Data Clustering” Asian Journal of Computer Science and Technology Vol.7 No.S1, pp. 96-100.**
* **Clustering Loan Applicants based on Risk Percentage using K-Means Clustering Techniques, Dr. K. Kavitha, International Journal of Advanced Research in Computer Science and Software Engineering.**
* **Amruta S. Aphale , Dr. Sandeep R. Shinde, 2020, Predict Loan Approval in Banking System Machine Learning Approach for Cooperative Banks Loan Approval, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 09, Issue 08 (August 2020),**

**Links for papers used in this project :**

[**https://www.researchgate.net/publication/357449126\_THE\_LOAN\_PREDICTION\_USING\_MACHINE\_LEARNING**](https://www.researchgate.net/publication/357449126_THE_LOAN_PREDICTION_USING_MACHINE_LEARNING)

[**https://jespublication.com/upload/2020-110471.pdf**](https://jespublication.com/upload/2020-110471.pdf)

[**https://www.researchgate.net/publication/353007922\_Customer\_Loan\_Prediction\_Using\_Supervised\_Learning\_Technique**](https://www.researchgate.net/publication/353007922_Customer_Loan_Prediction_Using_Supervised_Learning_Technique)