ISSN No:-2456-2165

# Education Loan Prediction Analysis

Sanskruti Naik
Department of Information Technology and Engineering
Goa College of Engineering
Goa, India

Ganesh Manerkar
Department of Information Technology and Engineering
Goa College of Engineering
Goa, India

**Abstract:- Education loans help students to cover the cost of tuition, books and supplies, and living expenses while in the process of pursuing a degree. Education loans are granted by private banks and by government organizations. This paper isan analysis on student loan data for the interest free education loans granted to students as per the standards and rules of Goa Education Development Corporation(GEDC). The dataset is prepared complying the standards of criteria mentionedby organization. The accuracy of prediction is compared using models like Support Vector Machine(SVM), Random forest(RF), Logistic regression(LR), Decision tree classifier and XG-boost.**

***Keywords:-*** *Loan, Prediction, Support Vector Machine, Ran-Dom Forest, Logistic Regression, Decision Tree Classifier, XG- Boost.*

## I. INTRODUCTION

It is education that uplifts the society at the macro leveland the individuals at the micro level from their all round backwardness, whether social or economical , cultural or political. Education loan helps promote pursuit of higher and technical education by younger population to ensure that economic and financial difficulties do not come in the way of such pursuit. The Government of Goa launched the Interest free Education Loan Scheme under which eligible candidates can undertake approved degree and diploma courses at undergraduate and post graduate levels in India or Abroad. The Rules and Standards are maintained by the Goa Education Development Corporation(GEDC) for a Candidate to Apply for such Loan. The Various parameters like Residence, familyincome, percentage of marks obtained in (10th ,12th, Diploma), Whether any sibling of the applicant has taken a education loan earlier are involved in the processing of Loan underGEDC. The following are the primary Criteria followed by GEDC in granting student Loan.

- The Applicant necessarily should be the resident of Goa for not less than 15 years.
- Maximum 5 years of study period/course duration is coveredin India and maximum 2 years of Study is covered for Abroad under this scheme.
- Any person below age of 30 years, shall be entitled to applyfor and receive loans under this Scheme.
- Applicant who wishes to pursue higher education in India must have obtained 55% or more marks in the qualifying examination(for ST/SC/OBC marks will be relaxed by 10 percent).
- Applicant who wishes to pursue higher education Outside India must have obtained 60% or more marks in the qualifying examination(for ST/SC/OBC marks will be relaxed by 10 percent).
- Family Income should not exceed 7LPA for applicants taking courses within India. In the event that brother or sister of the applicant is also pursuing studies at Higher/Technical education level(whether or not such sibling has applied for, or availed , loan under this scheme), the eligibility limit for family will be raised to 8LPA.
- Family Income should not exceed 12LPA for applicants taking courses Outside India. In the event that brother or sister of the applicant is also pursuing studies at Higher/Technical education level(whether or not such sibling has applied for , oravailed loan under this scheme), the eligibility limit for familywill be raised to 14LPA.

This paper aims to provide loan to a deserving applicant adhering to all the above criterias. The loan approval historyof past applicant forms is considered for training the model and an efficient, non-biased system is formulated to reduce the institutions time employed in checking every application for granting loan on a priority basis . The analysis of parameters such as residence, Category, Education etc. which are linkedto each other is Visualized in this paper. Section II shows literature survey of systems and approaches for granting of loans in various domains . We discuss the Framework of the proposed system in section III. Results obtained, Comparison of models is carried out in Section IV. Finally we conclude in section V.

## II. RELATED WORKS

Predictive analytics is a branch of advanced analytics that uses many techniques from data mining, statistics, modeling, machine learning, and artificial intelligence to analyze current data to make predictions. “Adyan Nur Alfiyatin, Hilman Taufiq and their friends have worked on the house price predic- tion. They have used regression analysis and Particle Swarm Optimization (PSO) to predict house price”. “Mohamed El Mohadab, Belaid Bouikhalene [3] and Safi have put a work to predict the rank for scientific research paper using supervised learning”. “Kumar Arun, Garg Ishan and Kaur Sanmeet [1] have worked on bank loan prediction on how to approve a loan and proposed a model with the help of SVM and Neural networks like machine learning algorithms”. ”Anshika Gupta, Vinay Pant , Sudhanshu Kumar and Pravesh Kumar Bansal[4] worked on Bank loan prediction has implemented algorithms like random forest(RF), logistic regression(LR) to make predictions”. These literature reviews helped us to carry out this work and

propose a reliable Education loan predictionmodel.

## III. PROPOSED SYSTEM

### A. System Model – Using Flow Graph

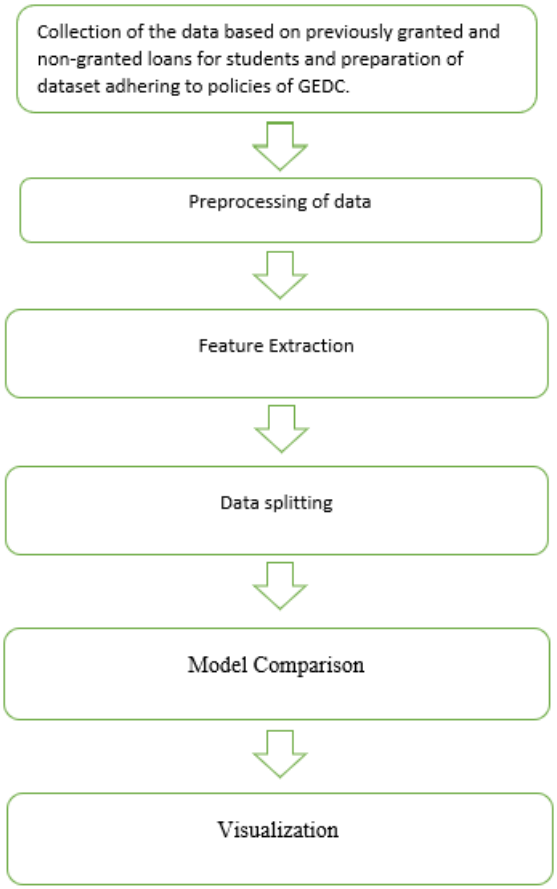


Fig. 1. System architecture

The data is initially prepared complying with the policy document viewed on GEDC portal. The loans of students are approved only when they meet the criteria required by GEDC.The data then is treated for preprocessing where the outliers like missing values are handled and unnecessary attributes are dropped. Label encoding is done to convert the categorical data to a numeric form wherever needed. The feature extraction step begins by thus selecting only the attributes which are essential for predicting grant of loan to a student and thereby dropping unnecessary columns. The data is then split into test and train. The system is trained using various Machine Learn- ing models such as support vector Machine(SVM), Logistic regression(LR), Random Forest(RF), Decision tree classifier and XG-boost. The model accuracy is thus compared of allthe above models. Data Visualization using the pandas library Seaborn is also achieved which helps to get an analysis of various parameters that are involved in granting of loan.

1) *Dataset Used:* Dataset is prepared by complying the policy document for granting education loan to a student available on GEDC portal. It is prepared following the criteria's mentioned in GEDC brochure. This rules are also listed in the Introduction section of this paper. Dataset in total comprises of 290 rows and 11 columns. The column attributesare named such as Loan id, Gender, Highest Degree, Board, Sibling loan, Annual income, caste, Loan amount, Residence, Status and Age.

2) *Preprocessing:* The data obtained from the dataset Preparation Step (step 1) is then preprocessed by removing the unwanted data. Pre-processing data transformation operations, are used to transform the dataset into a structure suitable for machine learning. This step basically helps incleaning the dataset by removing irrelevant or corrupted data that can affect the accuracy of the dataset, which makesit more efficient. In this process removal of missing data, duplicate entries and normalization is carried out. Removal of missing data is the process, where null values such as missing values and Nan values are replaced by 0. Label encoding is done in order to convert the categorical data to a form thatthe machine understands i.e., numerical data. The columns having answers of Yes/No are converted to 1/0 respectively. Similarly, the Gender column attributes Male/Female are converted to 0/1 respectively and so on.

3) *Feature Extraction:* Feature Extraction reduces the number of features in a dataset by creating new features from the existing dataset (and then discarding the original features). These new reduced set of features should then be able to give most of the information contained in the original set of features. In this case, unnecessary columns are dropped whose removal do not affect the loan procedure. Ex. Loan idin our case.

4) *Data Splitting:* Data splitting splits the data into a train, test, or validation set. The train set would contain the data which will be fed into the model or in other words model would learn from this data. The validation set is usedto validate the trained model. The test set contains the dataon which we test the trained and validated model. It tellsus the efficiency/performance using evaluation metrics (like precision, recall, accuracy, etc.

5) *Model Comparison:* The models used for evaluation of accuracy scores are - Support vector Machine (SVM), Logisticregression (LR), Random Forest (RF), Decision tree classifier, and XG-boost. A comparison of the models help us in using the best algorithm for prediction purpose yielding the bestoutput on system under consideration.

### B. Visualization

Visualization is carried out in order to graphically representvarious attributes and their link to granting of loan procedure. It will also help in the survey process. Similarly graphical count of candidates belonging to Categories such as General, OBC, SC, ST can be observed who have got the loan approvaletc.

*C. Algorithms Implemented*

The various supervised Machine learning algorithms used for prediction of availing education loans to students are as follows:

1) Support Vector Machine: The support vector machine (SVM) is used primarily for classification problems in Machine learning. SVM create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors. and hence the algorithm is termedas Support Vector Machine.
2) Logistic regression: Logistic regression is used for predicting the categorical dependent variable using a given set of independent variables. So, it predicts the output of a categorical dependent variable. Therefore the outcome mustbe a categorical or discrete value. It can be either Yes orNo, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1. In this analysis, it will specify whether the education loan to the applicant is granted or not.
3) Decision Tree: It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a decision tree, for predicting the class of the given dataset, the algorithm starts from the root node of the tree. This algorithm compares the values of the root attribute with the record attribute or real dataset attribute and, based on the comparison, follows the branch and jumps to the next node. For the next node, the algorithm again compares the attribute value with the other sub-nodes and moves further. It continues the process until it reaches the leaf node of the tree. The complete process can be better understood using the below algorithm
4) Random Forest: Random Forest (RF) is a popular machine learning algorithm that belongs to the supervised learning technique. Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accu- racy of that dataset. The greater number of trees in the forest leads to higher accuracy.
5) XG-Boost: XG-Boost is an algorithm that has been widely known for prediction of results faster. In this model each tree is built only after the previous one using all cores. This makes XG-Boost a very fast algorithm.

## IV. RESULTS

1) *Model Comparison:* The accuracy score on test data obtained is highest by XG-boost model. However the lowest accuracy is obtained by Support Vector Machine. It can also beobserved that Decision tree too performs well for the dataset under consideration. The comparison of the models helps us employ the best algorithm for prediction purpose on the systemunder consideration.

| ML Models | Accuracy on train samples |
|---|---|
| Support Vector Regression | 46% |
| Logistic regression | 48% |
| Decission tree classifier | 92% |
| Random forest | 89% |
| XG-boost | 94% |

Fig. 2. Model Comparison

2) *Confusion matrix:* The Confusion matrix is a summary of prediction results on a given classification problem. Here heatmap describes that '0'(Loan Not Granted ) samples classi-fied correctly were 26 and incorrectly classified samples were
3) Similarly '1' (Loan Granted) samples classified correctly were 29 and incorrectly classified were 0.

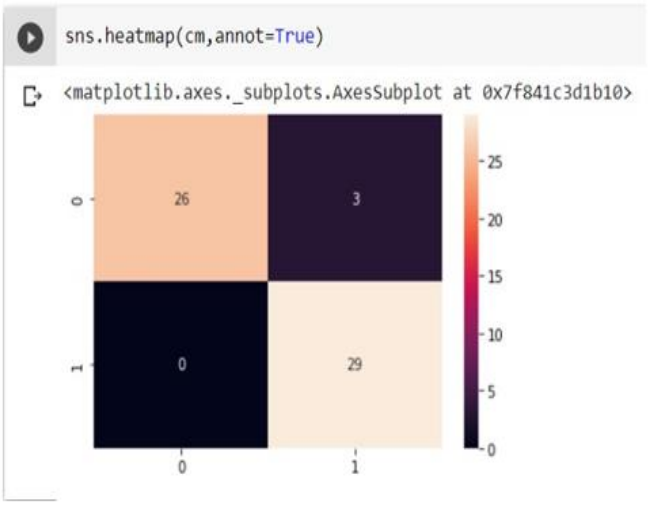


Fig. 3. Confusion matrix

4) *Data Analysis based on Visualization:* Data Visualiza-tion using the panda's library Seaborn is also achieved which helps get an analysis of various parameters that are involved in granting of the loan. The Data Visualization may help the organization to keep a track of various parameters suchas family income, category, highest education, etc. that help students access the loan.

The graph below depicts the volume of male and female candidates who either are been granted the loan(Orange) and not been granted the loan(blue). The 0 on the x-axis is themale candidates and the 1, is for female candidates.

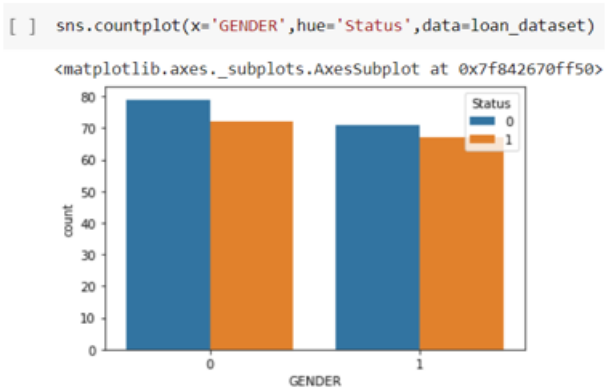


Fig. 4. Represents volume of male and female who either are been granted the loan(Orange) and not granted the loan(blue)

The below graph shows the clear result that only candidatesabove 15 years of residence are been granted the loan.

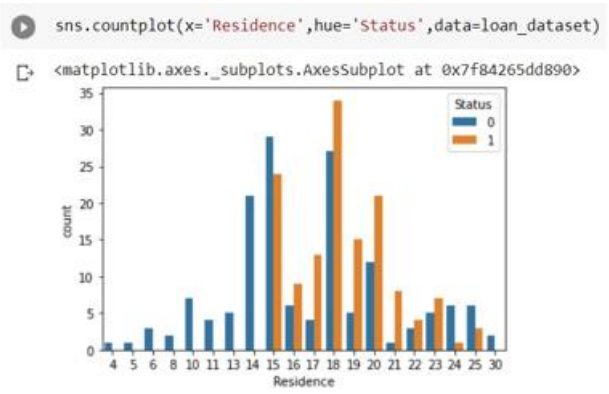


Fig. 5. Representation that only candidates above 15 years of residance are been granted the loan.

The graph underneath gives a virtual representation of can- didates of each category been granted and not granted with the loans. The 0 on the x-axis(caste-ST) shows candidates of that category been availed the loan facility(orange) and not been availed the loan facility(blue).same goes for 1(SC),2(OBC) and 3(General) Categories.

## V. CONCLUSION

The Education Loan Prediction System is trained using various ML models such as SVM, Logistic regression, RF, Decision tree classifier and XG-boost. The accuracy score on testdata obtained is highest by XG-boost model. However the low- est accuracy is obtained by Support Vector Machine(SVM). This suggest that the boosting algorithm can be used for most of the prediction based environment as it best yields the output.Confusion met- rics was visualized using a heatmap. Data Visualization using the pandas library Seaborn is also achieved which helps get an analysis of various parameters that are involved in granting of loan. The dataset can be increasedwith more entries of rows to get even better accuracy scores and train the model in a more better manner.

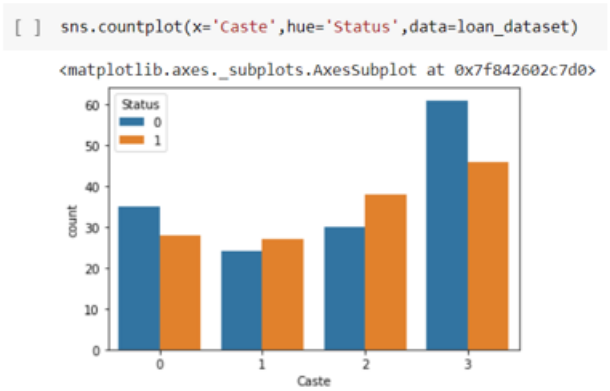


Fig. 6. Representation of candidates of each category been granted and not granted with the loans

## ACKNOWLEDGMENT

I thank my Project Guide, Mr. Ganesh Manerkar for mo- tivating and guiding me to carry out this research seminar. I express my gratitude and earnest thanks to Dr. Nilesh Fal Dessai, Head of Information Technology Department, Goa College of Engineering, in providing me with all the facilities throughout the research seminar work. My sincere and kind thanks to the Principal of our college, Dr. Rajesh Basant Lohani for providing all the facilities and resources to me. I also heatly thank the personell's of GEDC Goa for providing me the necessary inputs. I am indebted to my Parents and my Husband for motivating me in partial fulfillment of this research seminar work.

## REFERENCES

[1]. K. Arun, G. Ishan, and K. Sanmeet, "Loan Approval Prediction based on Machine Learning Approach", IOSR Journal of Computer Engineering, pp. 18-21, 2009.

[2]. Adyan Nur Alfiyatin, Hilman Taufiq, Ruth Ema Febrita, Wayan Firdaus Mahmudy, 'Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization: International Journal of Advanced Computer Science and Applications (Vol. 8, No. 10, 2017).

[3]. Mohamed El Mohadab, Belaid Bouikhalene, Said Safi, 'Predicting rank for scientific research papers using supervised learning applied Computing and Informatics 15 (2019) 182–190.

[4]. Anshika Gupta, Vinay Pant, Sudhanshu Kumar, Pravesh Kumar Bansal, Bank Loan Prediction System using Machine Learning: 9th International Conference on System Modeling and Advancement in Research Trends 4th–5th December 2020

[5]. Vishal Singh, Ayushman Yadav, Rajat Awasthi, N.Partheeban, ' Predic- tion of Modernized Loan Approval system based on Machine Learning Approach '2021 International Conference on Intelligent Technologies (CONIT)

[6]. Mohamed Alaradi, Sawsan Hilal, 'Tree-Based Methods for Loan Ap- proval,2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI)

[7]. L. Al-Blooshi and H. Nobanee, “Applications of Artificial Intelligence in Financial Management Decisions: A Mini-Review”,
[8]. R. Kumar, V. Jain, P.S. Sharma, S. Awasthi, and G. Jha. “Prediction of Loan Approval using Machine Learning”, International Journal of Advanced Science and Technology, vol. 28, pp. 455-460, 2019. SSRN Electronic Journal, 2020.
[9]. Rising Odegua,” Predicting Bank Loan Default with Extreme Gradient Boosting
[10]. ”Goa Education and Development Cooperation”, https://gedc-goa.org/.