Ilkogretim Online - Elementary Education Online, 2021; Vol 20 (Issue 3): pp. 3442-3449
http://ilkogretim-online.org
doi: 10.17051/ilkonline.2021.03.359

Home Loan Prediction Using Machine Learning Models

[1]Vipashi Kansal, [2]Dr.Upma Jain, [3]Ashutosh Kumar Gupta, [4] Ms Manisha Aeri

[123] Graphic Era Deemed to be University.

[4] Assistant Professor, Department of Computer Science and Engineering, Graphic Era Hill University, Dehradun.

ABSTRACT

Buying a home of their choice is a dream of every person. To fulfill their dreams banks provide different loan schemes for requesting customers. Although, with assistance from the banks many people get their loan but there are also many people who face difficulties in getting a loan. The cancellation of loan's request is due to various reasons such as poor civil score, less income, etc.

A huge amount of customer's time is wasted if his/her loan is rejected so with the help of this loan prediction project a customer can get initial inference whether he/she will get a loan or not, also this project will help bank employees in checking the eligibility of a customer. We have used six classification algorithms in this project and for balancing the unbalanced class I have also used SMOTE. After training and predicting the results we have also compared all the Models.

Keywords: KNN, SVM, SMOTE, LR, Decision Tree, Machine Learning.

1. INTRODUCTION:

Machine learning is a subset of Artificial Intelligence in which we try to train our machine with the help of historic data and then according to their training we will predict the results. Machine learning is used in many industrial sectors such as banking, medical field, business analysis, etc. Machine learning allows us to get insight of a huge amount of data in less time and in some cases it is more accurate than the human labor. In short, it is a branch of AI (Artificial Intelligence) that is primarily focused in developing algorithms that permit a computer to learn from data and past experiences, just as humans do. In banking sector machine learning plays a very important role such as predicting whether a person is eligible for getting a loan or not, fraud detection, getting the price of house, etc. Banks can reliably identify the extremely subtle and frequently concealed events and correlations in user behavior that may signify fraud because of anti-money laundering in ML. Financial organizations may analyze considerably more data much quicker than human rule-based

systems by automating the difficult anomaly detection process. The majority of a bank's assets are directly tied to the revenue generated by the loans. In a banking system, the main goal is to place their assets in trustworthy hands. There are many banks and companies which gives the loan after a lengthy process of verification and validation, but still no guarantee that the chosen applicant is the most worthy candidate among all applicants. Through this machine learning methods like Support Vector Machine, Decision Tree, KNN and many more which helps to determine whether a specific application is secure or not, and the entire feature validation process is automated using machine learning. In this paper, we have chosen to work on "home loan prediction, the objective of this paper is to classify whether a person is eligible for getting home loan from bank or not based on his data available. This problem can be solved by using classification algorithms which are a part of supervised machine learning algorithms. The following is how this paper is organized , IInd section describes the work has been done in the field of machine learning in Loan prediction, IIIrd section displays the methodology we have used in this paper, IVth section displays the results and discussion, Vth section provides the conclusion and VIth section shows the references.

2. RELATED WORK:

A technique for assessing credit credibility that aids businesses in forming the proper judgments about whether to approve or deny customer credit requests. This facilitates the development of efficient distribution channels by the banking diligence. This indicates that their method can reduce future risks if the consumer has a minimal repayment capability. It has been built and evaluated for domains with various methods that are superior to the basic data mining model[1].The author proposes a credit status model for classifying loan applications as legitimate or regular clients. When categorizing loan applicants using R-Package, the suggested model yields a score of 75.08. This interpretation can be used by lenders to decide which mortgages to provide for mortgage operations. Additionally, comparative research was carried out at several iteration levels. A 30-grounded ANN model that gives a higher level of delicacy than previous scenarios is used to represent the replication position. In marketable banks, this technique can be utilized to prevent significant losses [2]. To predict Android applications, author has used six different machine learning classification models which are accessible in R (open-source software). Their application was functioned effectively and complies with all bank standards. The only disadvantage was that it was assigned varied weights to each aspect, but in fact, it could be conceivable to authorize a loan just on the basis of one strong element, which is not achievable with this method. Many additional systems may be simply linked to this component. The most significant weights of content mistakes and characteristics are corrected by the automated prediction system in circumstances of computer failure, and soon, so-called software may be safer, more consistent, and more [3]. As we all know,

many people are applying for bank loans but due to have a certain amount of assets to lend to, they can only give to certain number of applicants. In order to save a lot of bank resources and effort, we want to minimize the risk component involved in choosing the safe person in their paper. The Big Data of the individuals to whom the loan was previously issued is mined for this information, and the machine was trained using the machine learning model that produces the most accurate result based on these records and experiences [4].

3. METHODOLOGY USED:

I. **SMOTE:** Stands for synthetic minority oversampling technique. It is used to balance the imbalance label class. SMOTE crates new data points for minority class using Euclidean distance. It chooses any two random minority class data point and creates a new data point in the line joining them.

II. **SVM:** The concepts of classification and regression are the core topics covered by support vector machines, which is the application of supervised learning. SVM's first versions could only categorize data linearly by creating hyperplanes. Vapnik, Boser, and Guyon later developed a method of creating a nonlinear classifier by utilising the Kernel function in 1992. Since then, SVM has emerged as one of the most popular supervised learning—that is, learning from datasets with features and class labels—classification algorithms. Later, SVM clustering was developed by Vapnik and Siegelmann to execute unsupervised learning, or datasets lacking class labels and output features [5].

III. **Random Forest (RF):** It is a part of supervised learning methodology. It can be applied to Machine learning issues involving both classification and regression. It is built on the idea of ensemble learning, which is a method of integrating many classifiers to address difficult issues and enhance model performance[6].Each tree in a random forest depends on the values of a random vector that was sampled randomly and with the same distribution for all the trees in the forest.

IV. **ADABOOST:** The boosting algorithm is well-known in the machine learning field[7].In boosting family of algorithms' Adaboost is the most common one. Is is used to combine several weak classifiers into one strong classifier, improving classification accuracy[8].

V. **KNN:** A simple yet effective machine learning approach is the k-Nearest Neighbor (KNN) algorithm. Both classification and regression can be done successfully using it. However, classification prediction is where it is commonly utilized. The KNN identifies newly entered data based on its similarity to previously trained data and organizes the data into coherent clusters or subsets. The class to which the input belongs is determined by its closest neighbors[9].

VI. **Decision Tree:** It is a supervised learning algorithm is a. It is a visual depiction of every potential answer[10]. Every choice was based on a set of circumstances. With

the help of this technique, a population is divided into segments that resemble branches and form an inverted tree with a root node, internal nodes, and leaf nodes[11].

VII. **Linear Regression(LR):** The process of examining the connection between an independent variable and a dependent variable is known as regression analysis[12].Finding the best-fitting linear line and the ideal intercept and coefficient values such that the error is minimized is the major goal of a linear regression model. Linear Regression are of two types LR and multiple LR[13].

4. RESULTS AND DISCUSSION:

We have done training and testing on different classification models dataset which containing 614 rows and 14 columns. The target column in this dataset is "**Loan Status**" which has 422 values of '1' type and 192 of '0' type. '1' represents loan approved and '0' represent rejected. After applying SMOTE it has 422 values in each type.

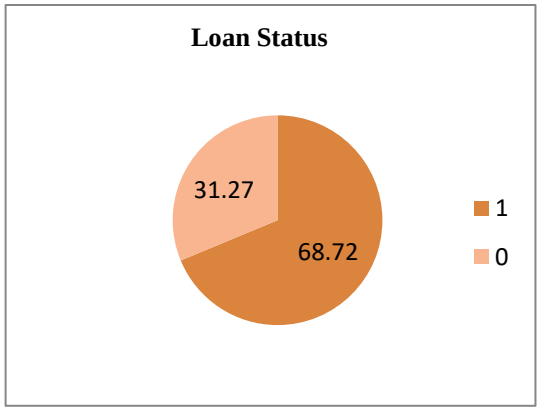

Fig 1: Loan Status

1) **Data Processing:** In this step we are checking the data types of each column in the dataset .

```
Loan_ID              object
Gender               object
Married              object
Dependents           object
Education            object
Self_Employed        object
ApplicantIncome       int64
CoapplicantIncome   float64
LoanAmount          float64
Loan_Amount_Term    float64
Credit_History      float64
Property_Area        object
Loan_Status          object
dtype: object
```

Fig 2: Dataset Columns

Here as we can see there are several columns which have object data types, but for building a machine learning model all the object type columns must be encoded into numerical type because machine learning only works on numerical data. We can do the encoding with the help of LabelEncoder() which is present in Sklearn library. After encoding we have to check for missing values in the dataset and if we have any missing values then we have to either fill them or remove them.

```
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
```

Fig 3 : Dataset

2) **Data Visualization:** After filling the null values we have to get the insight of the data, first of all we will find correlation between coloumns. We use graphs such as bar graphs , pie charts , count plots, etc. for data visualization. Data Visualization helps in getting the inside knowledge of the data, with proper data knowledge we can improve our performance.

```
Loan_Status          1.000000
Credit_History       0.540556
Married              0.096657
Property_Area        0.032112
Self_Employed        0.010880
Gender               0.008690
Dependents          -0.007318
Loan_Amount_Term    -0.022549
Total_Income        -0.031271
LoanAmount          -0.033214
Education           -0.085884
Name: Loan_Status, dtype: float64
```

Fig 4: Data Visualization

Here we can see that loan status is highly correlated with Credit_History. After getting the insight we have to split our data set first into two variables i.e. for dependent and independent variables. After splitting the dataset into we have to split our variables into testing and training data. Training data is used to train the model, whereas testing data is used to test the accuracy of the model. After splitting the dataset now its time to build our models and train them.

3) **Without Using SMOTE:**

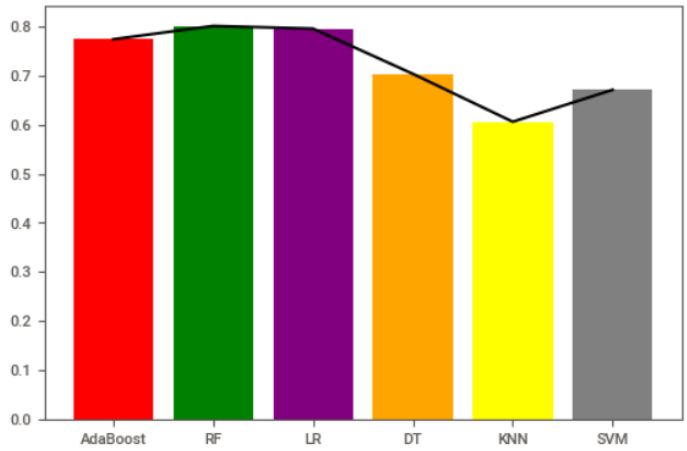

Fig 5 : Comparison of Models based on their Accuracy Without SMOTE

Here we can see that Random forest classifier and logistic regression algorithms are giving the best result on dataset without using SMOTE. The performance of KNN model is least among all 6 algorithms.

4) **With Using SMOTE:**

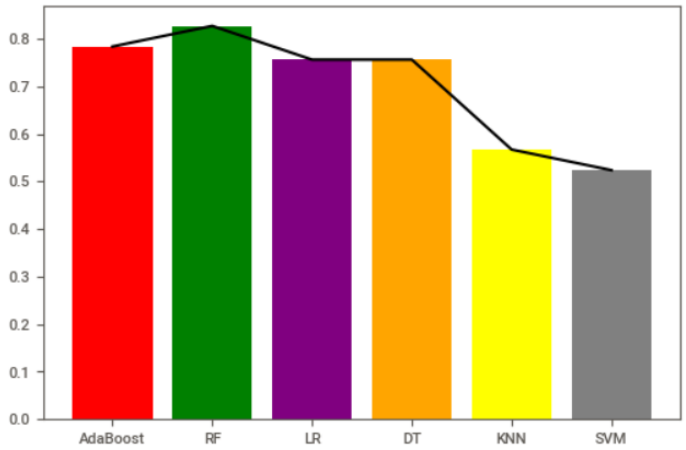

Fig 6 : Comparison of Models based on their Accuracy
With SMOTE

Here we can see that the accuracy of Random Forest Classifier increases slightly and the accuracies of Logistic Regression and SVM were decreased.

5. CONCLUSION:

We have used various concept of machine Learning like data preprocessing, splitting the dataset, balancing the dataset , training and testing of different classification models and comparing their accuracies in different circumstances. These method works well in predicting the eligibility of customer getting a loan. Our work is able to help to understand several machine learning algorithms, different approaches to follow. Also shows that the accuracy of Random Forest classifier is not changed with the use of SMOTE and it is maximum in both the cases, therefore the model to be chosen for prediction can be Random Forest Classifier either by balancing the dataset or not.

REFERENCES :

1. Kumar, A., Garg, I., Kaur, S.,(2016) . Loan Approval Prediction based on Machine Learning Approach .In: National Conference on Recent Trends in Computer Science and Information Technology e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 3, Ver. I, PP 79-81.
2. S. M S, R. Sunny T, "Loan Credibility Prediction System Based on Decision Tree Algorithm," International Journal of Engineering Research & Technology (IJERT) Vol. 4 Issue 09, (2015).
3. A. Kumar, I. Garg and S. Kaur, "Loan Approval Prediction based on Machine Learning Approach," IOSR Journal of Computer Engineering, (2016).

4. Dr K. Kavitha, “Clustering Loan Applicants based on Risk Percentage using K-Means Clustering Techniques,” IJARCSSE - Volume 6, Issue 2, (2016).
5. S. Ghosh, A. Dasgupta and A. Swetapadma, "A Study on Support Vector Machine based Linear and Non-Linear Pattern Classification," 2019 International Conference on Intelligent Sustainable Systems (ICISS), 2019, pp. 24-28, doi: 10.1109/ISS1.2019.8908018.
6. Breiman, Leo. “Random Forests.” Machine Learning, vol. 45, no. 1, 2001, pp. 5–32, doi:10.1023/a:1010933404324.
7. Chengsheng, Tu, et al. “AdaBoost Typical Algorithm and Its Application Research.” MATEC Web of Conferences, vol. 139, 2017, p. 00222, doi:10.1051/matecconf/2017139002
8. Zhang, Yanqiu, et al. “Research and Application of AdaBoost Algorithm Based on SVM.” 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), IEEE, 2019, pp. 662–666.
9. K. Taunk, S. De, S. Verma and A. Swetapadma, "A Brief Review of Nearest Neighbor Algorithm for Learning and Classification," 2019 International Conference on Intelligent Computing and Control Systems (ICCS), 2019, pp. 1255-1260, doi: 10.1109/ICCS45141.2019.9065747.
10. Amrutha, K. “Decision Tree Machine Learning Algorithm.” Analytics Vidhya, 31 Jan. 2022, https://www.analyticsvidhya.com/blog/2022/01/decision-tree-machine-learning-algorithm/.
11. Song, Yan-Yan, and Ying Lu. “Decision Tree Methods: Applications for Classification and Prediction.” Shanghai Archives of Psychiatry, vol. 27, no. 2, 2015, pp. 130–135, doi:10.11919/j.issn.1002-0829.215044.
12. Rong, Shen, and Zhang Bao-wen. “The Research of Regression Model in Machine Learning Field.” MATEC Web of Conferences, vol. 176, 2018, p. 01033, doi:10.1051/matecconf/201817601033.
13. Deepanshi. “All You Need to Know about Your First Machine Learning Model – Linear Regression.” Analytics Vidhya, 25 May 2021, https://www.analyticsvidhya.com/blog/2021/05/all-you-need-to-know-about-your-first-machine-learning-model-linear-regression/.