An approach for prediction of loan approval using Supervised Algorithm in Machine learning algorithm.

[1] **Kondreddy Umesh kumar reddy,** [2] **K. Tulasi Krishna kumar,**
[1] MCA 2nd Year, [2] Associative professor,
[1] Master of Computer Applications
[1] Sanketika Vidya Parishad Engineering College, Visakhapatnam, India

***Abstract*— In the banking industry, loans are the primary source of income for banks, as they earn from the interest charged on these loans. The profitability of a bank largely depends on the timely repayment of loans by customers. Therefore, it is crucial for banks to predict loan defaulters in order to minimize their Non-Performing Assets (NPA). Several methods have been proposed to address this problem, but it is essential to compare these methods to determine their effectiveness in predicting loan defaulters accurately. One popular approach for predictive analytics is the logistic regression model. To study loan default, data was collected from Kaggle, and logistic regression models were used to predict the likelihood of loan default. Performance measures such as sensitivity and specificity were computed to compare the models. The results indicated that the model that included personal attributes such as age, purpose, credit history, credit amount, and credit duration, in addition to checking account information (which reflects a customer's wealth), produced marginally better results. This suggests that a customer's other attributes should also be taken into account while calculating the probability of default on loans to accurately predict loan defaulters. By using the logistic regression approach, banks can identify the right customers to grant loans to by evaluating their likelihood of default on loans. This approach highlights the importance of assessing all aspects of a customer's profile, rather than just their wealth, when making credit decisions and predicting loan defaulters.**

***Index Terms*— Prediction, component, Overfitting, banking system, credit line, interest, loan defaulters, Non-Performing Assets, predictive analytics, logistic regression model, Kaggle.**

I.INTRODUCTION

The purpose of this study is to predict the safety of loans using a machine learning model trained on data from previous customers of various banks who were approved for loans based on a set of parameters. The logistic regression algorithm is utilized to make this prediction. The data set used in this study consists of 1500 cases and includes 10 numerical and 8 categorical attributes. Before training the model, the data is cleaned to eliminate any missing values. Several parameters are taken into account when deciding whether to grant a loan [1][3] to a customer, including their CIBIL score (credit history), business value, and assets.

Qualification	Categorical
In Service / Business Owner	Categorical
Individual income of Applicant	Qualitative
Individual income of Co- Applicant (if Any)	Qualitative
Amount of Loan required	Qualitative
Term for which loanRequired	Qualitative
Credit History ofApplicant	Qualitative
Area of Property	Categorical

II.LITERATURE SURVEY

Logistic regression is a useful machine learning algorithm that is commonly used for classification problems. One of its advantages is that it can be used for predictive analysis. It is used to describe data and to explain the relationship between a single binary variable and one or more independent variables that are nominal, ordinal, or ratio level. To develop a prediction model, logistic regression utilizes the sigmoid function, which is appropriate for binary outcomes of 0 or 1[11][15]. The dataset used in this study consists of bank customer data that has been divided into training and test sets. The training dataset contains approximately 600+ rows and 13+ columns, while the test dataset contains 300+ rows and 12+ columns and does not include the target variable [13]14]. Both datasets have some missing values, which are filled using mean, median, or mode values. Feature engineering techniques are then applied, followed by exploratory data analysis using statistical concepts such as normal distribution and probability density function to study the dependent and independent variables through univariate, bivariate, and multivariate analysis. The focus of the model is to target eligible customers for loans, and logistic regression is enabled using the sigmoid function to divide the probability into binary output. This allows the prediction model to be developed.

III.PROBLEM STATEMENT

Banks, housing finance companies, and NBFCs offer various types of loans in rural, semi-urban, and urban areas. Before approving a loan application, these companies must evaluate a customer's eligibility. This paper proposes an automated solution using a machine learning algorithm to streamline this process. Customers fill out an online loan application form that includes

details such as sex, marital status, education level, dependents, annual income, loan amount, and credit history. The algorithm identifies eligible customer segments, enabling the bank to focus on those customers [4][7]. Automating the loan approval process can save time and improve customer service, resulting in increased customer satisfaction and reduced operational costs are significant[9]. However, a robust model is necessary to accurately predict which loans to approve and which to reject, thus minimizing the risk of loan default.

IV.EXISTING SYSTEM

The success or failure of a bank heavily relies on its loan portfolio, which includes the ability to accurately predict whether a customer will repay their loan or default. However, the current method of predicting loan defaulters relies on human effort and CIBIL score, and it can be time-consuming due to manual verification by officers. In order to streamline the loan eligibility process, factors such as gender, marital status, dependents, education, applicant income, loan amount, and loan amount term are taken into consideration. Nonetheless, accurately predicting a customer's ability to repay their loan remains challenging with the current system.

V. PROPOSED MODEL

The proposed model aims to predict whether a bank should grant a loan to a customer or not, using classification as the target. To achieve this, Logistic Regression with a sigmoid function is utilized. Preprocessing is the time-consuming step, followed by Exploratory Data Analysis, Feature Engineering, and Model Selection. The model is then trained on two separate datasets before being applied to new data. Logistic Regression is a statistical machine learning algorithm that creates a logarithmic line to differentiate between extreme outcomes, allowing for accurate predictions.

V.A) Data Collection

Data has been collected from the Kaggle one of the most data source providers for the learning purpose and hence the data is collected from the Kaggle, which had two data sets one for the training and another testing [12]. The training dataset is used to train the model in which datasets is further divided into two parts such as 80:20 or 70:30 the major datasets is used for the train the model and the minor dataset is used for the test the model and hence the accuracy of our developed model is calculated.

V.B) Pre-Processing

Data mining technique has been used in Pre-Processing for transforming raw data which is collect using online form into useful and efficient formats. There is a need to convert it in useful format because it may have some irrelevant, missing information and noisy data. To deal with this problem data cleaning technique has been used. Before data mining the data, reduction techniques are used to deal with huge volume of data. So, data analysis will become easier, and it intends to get accurate results. So, data storage capacity increase and cost to analysis of data reduces.

Id	0
Sex	13
Married	3
No_Dependents	15
Qualification	0
In Service / Self_Employed	32
Annual_Income_Applicant	0
Annual_income_Coapplicant	0
Amount_Loan	22
Term	14
Credit_History _ Applicant	50
Assets	0
Status_Loan	0

V.C) Feature Engineering

The size of data can be reduced by encoding mechanisms. So, it may be lossy or lossless. If the original data is obtained after reconstruction from compressed data, such reductions are called lossless reduction else it is called lossy reduction. Wavelet transforms and PCA (Principal Component Analysis) methods are effective for reduction. In feature engineering a proper input dataset which is compatible as per machine learning algorithm requirements is prepared. In our model **Pandas** and **NumPy** library has been imported to run. So, the performance of machine learning model improves.

import pandas as pd import NumPy as np.

V.D) List of Techniques

D. i) Imputation

Imputation[17] is the process of replacing missing data with substituted values. There is one more measure problem i.e., missing values when data is prepared for our machine learning model. There may be many reasons of missing values like human errors, interruptions in flow of data, security concerns, and so on. The performance of machine learning model severely affected by missing values.

Train['Gender'].fillna(tra in['Gender'].mode()[0],inplace =Tr ue)
train['Married'].fillna(train['Married'].mode()[0],inplace=True)
train['Dependents'].fillna(train['Dependents'].mode()[0],in place=True)

Id	*0*
Sex	0
Married	0
No Dependents	0
Qualification	0
In Service / Self_Employed	0
Annual_Income_Applicant	0
Annual_income_Coapplicant	0
Amount_Loan	22
Term	14
Credit_History _ Applicant	0
Assets	0
Status_Loan	0

D. ii) Handling Outliers

In order to identify outliers in data, one common approach is to visually inspect the data and then make decisions about how to handle any outliers that are found. This method can be effective when the decisions made based on the visualizations are precise and accurate. Another approach is to use percentiles as a mathematical method for identifying outliers. With this method, a certain percentage of values are assumed to be outliers, either from the top or bottom of the distribution. However, setting the percentage value requires careful consideration of the underlying distribution of the data, as this will determine what is considered to be an outlier**.**

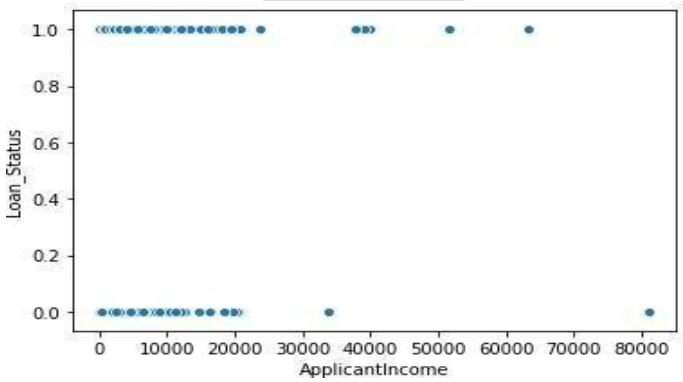


Fig. 1. Application income vs Loan Status

D. iii) Binning

Binning is often the critical factor in balancing performance and overfitting of a model. However, in the case of numerical columns, binning may not always be necessary except for certain scenarios where overfitting is a concern, as it can negatively impact the model's performance. On the other hand, for categorical columns, low-frequency labels can affect the model's robustness in a negative way. Assigning a common category to these infrequent values can help maintain the model's robustness and improve its performance.

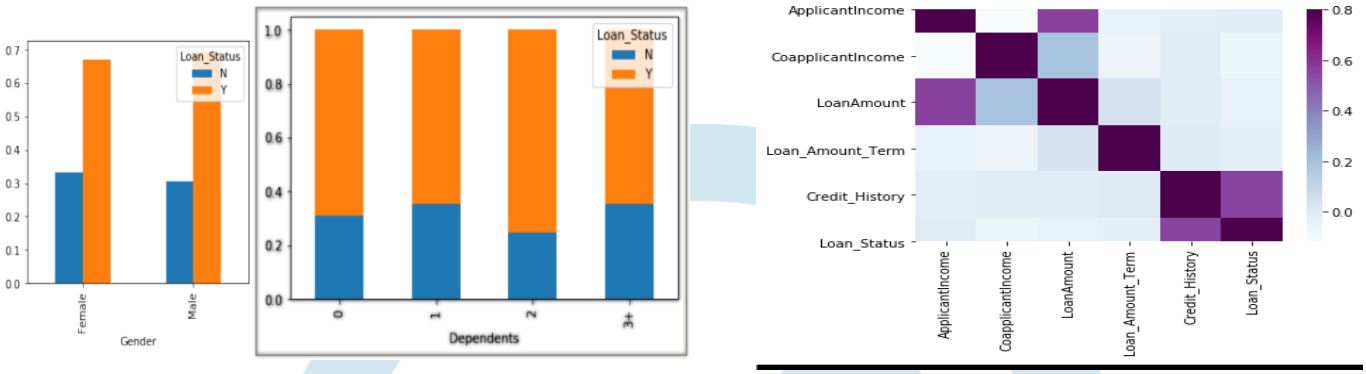
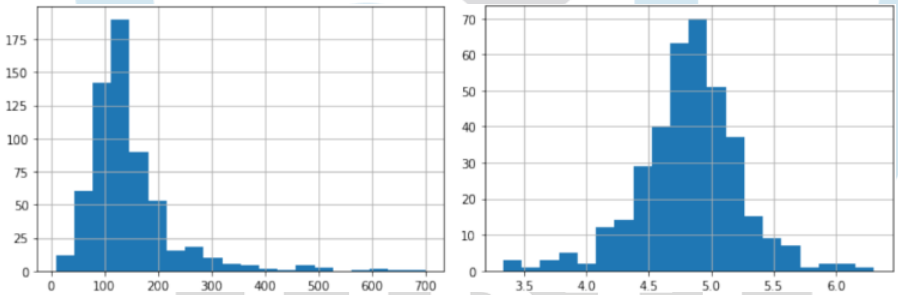


Fig. 2. Heap Map

D. iv) Log Transform

Logarithm transformation, also known as log transform, is a widely used mathematical technique in feature engineering. Its primary advantage is its ability to handle skewed data, making the distribution of the data more closely resemble a normal distribution. Additionally, log transformation can reduce the impact of outliers by normalizing the magnitude differences, which can increase the robustness of the machine learning model.



D.iv)Before Log Transform D.iv) After Log Transform

D. v) One Hot Encoding

One hot encoding [18] is a widely used method of encoding categorical data in machine learning. This technique involves transforming categorical data into a format where each value is represented by a single column with values of 0 or 1. The resulting columns indicate the relationship between the encoded and grouped columns. By converting categorical data into numerical format through one hot encoding, the resulting data can be easily processed by machine learning algorithms. This method also enables grouping of categorical data without any loss of information

from sklearn.preprocessing import LabelEncodernumber=LabelEncoder()

G	M	D	E	SE	AI	CAI	LA	CH	LAT		PA	LAL
1	0	0	0	0	5849	0	128	360	1		2	4.85203
1	1	1	0	0	4583	1508	128	360	1		0	4.85203

G	Sex
M	Married
D	No_Dependents
E	ualification
SE	In Service / Self_Employed
AI	Annual_Income_Applicant

CAI	Annual_income_Coapplicant
LA	Amount_Loan
CH	Credit_History _ Applicant
LAT	Loan Amount Transfer
PA	Assets
LAL	loan Amount log

VI.MODEL SELECTION

Model selection is the process of choosing the best machine learning model from a group of candidate models for a specific loan customer training dataset. There are various types of models such as logistic regression, SVM, KNN, etc. Each model has its own advantages and disadvantages, such as predictive error caused by statistical noise in the data, incompleteness of sample data, and limitations of different model types. The chosen model should meet the requirements and constraints of the stakeholders involved in the project, including both the bank and the customers. The selected model should be skillful in comparison to naive models, other tested models, and the state-of-the-art in loan approval classification problems. Therefore, the loan approval prediction problem can be considered a classification problem, and a suitable model can be chosen to address it. From sklearn.linear_model import LogisticRegression model =LogisticRegression()model.fit(x_train, y_train)

VII.MODEL EVALUATE

VII.A) Confusion Metrics

$Confsuion\ metrics$[20]are the tables that is used in classification problems to assess where errors in the model were made.

Fig. 3. Confusion Matrix

VII.B) Accuracy

The accuracy of a model is typically evaluated using predefined metrics. In a balanced class model, high accuracy is often observed. However, in the case of an unbalanced class, the accuracy can be significantly lower.

$$\frac{(TP+TN)}{(TP+FP+TN+FN)}$$

VII.C) Precision

The precision value of a model is calculated by taking the percentage ratio of positive instances and the total predicted positive instances. The denominator in the precision equation represents the total number of positive predictions made by the model for the entire dataset. A high precision value indicates the model's ability to make accurate positive predictions. Fortunately, our dataset has yielded a good precision value, indicating the high degree of accuracy of our model.

$$\frac{TP}{(TP+FP)}$$

VII.D) Recall

Recall is a metric that measures the percentage ratio of correctly identified positive instances to the total number of positive instances in the dataset. The denominator of the recall formula, which is the sum of true positives (TP) and false negatives (FN), represents the actual total number of positive instances in the dataset. Therefore, recall indicates how many correct positive instances the model can identify and how many it may miss. A high recall value means the model is able to correctly identify most of the positive instances in the dataset, while a low recall value indicates that the model is missing a significant number of positive instances.

$$\frac{TP}{(TP+FN)}$$

VII.E) F1 Score

The F1 Score is calculated as the harmonic mean [19] of precision and recall values. The model that achieves the highest F1 Score is considered the best performer. The numerator of the F1 Score is the product of precision and recall, and if either precision or recall is low, the final F1 Score is significantly affected. Therefore, a model performs well in terms of F1 Score if it has a high positive predicted value (precision) and does not miss positive instances (recall) by predicting them as negative.

$$\frac{2}{\frac{1}{\text{precision}}+\frac{1}{\text{recall}}}=\frac{2*\text{precision}*\text{recall}}{\text{precision}+\text{recall}}$$

VIII.CONCLUSION

The prediction process begins with cleaning and processing the data, imputing missing values, performing experimental analysis on the dataset, building a model, evaluating the model, and testing it on a separate test dataset. On the original dataset, the best-case accuracy achieved is 0.811.From the analysis, it is concluded that applicants with the worst credit scores are more likely to be denied loan approval due to a higher probability of defaulting on the loan. Additionally, applicants with high income and lower loan amounts are more likely to be approved, which is reasonable as they are more likely to repay their loans. However, factors such as gender and marital status do not seem to be taken into consideration by the company.

IX.FUTURE SCOPE

There are several potential avenues for future improvement and expansion when it comes to using supervised machine learning algorithms to predict loan approvals. Here are a few possibilities: Integration with other financial systems: Machine learning models can be integrated with other financial systems such as credit scoring, fraud detection, and loan servicing systems to create a more comprehensive financial ecosystem. This would help improve the accuracy of the model and provide a more seamless experience for customers.

REFERENCES

[1] Toby Segaran, "Programming Collective Intelligence: Building Smart Web 2.0 Applications." O'Reilly Media.

[2] Drew Conway and John Myles White," Machine Learning for Hackers: Case Studies and Algorithms to Get you Started," O'Reilly Media.

[3] Trevor Hastie, Robert Tibshirani, and Jerome Friedman ,"The Elements of Statistical Learning: Data Mining, Inference, and Prediction," Springer ,Kindle

[4] PhilHyo Jin Do ,Ho-Jin Choi, " Sentiment analysis of real-life situations using loca- tion, people and time as contextual features," International Conference on Big Data and Smart Computing (BIGCOMP), pp. 39–42. IEEE, 2015.

[5] Bing Liu, "Sentiment Analysis and Opinion Mining," Morgan & Claypool Publishers, May 2012.

[6] Bing Liu, " Sentiment Analysis: Mining Opinions, Sentiments, and Emotions," Cambridge University Press, ISBN:978-1-107-01789-4.

[7] Shiyang Liao, Junbo Wang, Ruiyun Yu, Koichi Sato, and Zixue Cheng, " CNN for situations understanding based on sentiment analysis of twitter data," Procedia computer science, 111:376–381, 2017.CrossRef.

[8] K I Rahmani, M.A. Ansari, Amit Kumar Goel, " An Efficient Indexing Algorithm for CBIR,"IEEE- International Conference on Computational Intelligence & Communication Technology ,13-14 Feb 2015.

[9] Gurlove Singh, Amit Kumar Goel ,"Face Detection and Recognition System using Digital Image Processing" , 2nd International conference on Innovative Mechanism for Industry Application ICMIA 2020, 5 -7 March 2020, IEEE Publisher.

[10] Amit Kumar Goel, Kalpana Batra, Poonam Phogat," Manage big data using optical networks", Journal of Statistics and Management Systems "Volume 23, 2020, Issue 2, Taylors & Francis.

[11] Raj, J. S., & Ananthi, J. V., "Recurrent neural networks and nonlinear prediction in support vector machine" Journal of Soft Computing Paradigm (JSCP), 1(01), 33-40, 2019.

[12] Aakanksha Saha, Tamara Denning, VivekSrikumar, Sneha Kumar Kasera. "Secrets inSource Code: Reducing False Positives usingMachine Learning", 2020 InternationalConference on Communication Systems &Networks (COMSNETS), 2020.

[13] X.Frencis Jensy, V.P.Sumathi,Janani Shiva Shri, "An exploratory Data Analysis for Loan Prediction based on nature of clients", International Journal of Recent Technology and Engineering (IJRT E),Volume-7 Issue-4S, November 2018.

[14] Pidikiti Supriya, Myneedi Pavani, Nagarapu Saisushma,Namburi Vimala Kumari, k Vikash,"Loan Prediction by using Machine Learning Models", International Journal of Engineering and Techniques.Volume 5 Issue 2, Mar-Apr 2019

[15] Nikhil Madane, Siddharth Nanda,"Loan Prediction using Decision tree", Journal of the Gujrat Research History,Volume 21 Issue 14s, December 2019.

[16] An article reference from https://brill.com/view/journals/lega/83/1-2/article-p179_9.xml

[17] A book reference of https://www.routledge.com/Flexible-Imputation-of-Missing-Data-Second-Edition/Buuren/p/book/9781032178639#

[18] An article reference from https://drbeane.github.io/python_dsci/pages/one_hot_encoding.html

[19] Website Reference from

https://www.google.com/search?q=harmonic+mean+book+reference&rlz=1C1RXQR_enIN1013IN1013&oq=harmonic+mean+book+reference&aqs=chrome.69i57j0i546l5.26430j0j4&sourceid=chrome&ie=UTF-8

[20] Reference taken from IEEE platform. https://ieeexplore.ieee.org/abstract/document/6270271

BIBILIOGRAPHY

Kondreddy Umesh Kumar Reddy is studying his 2nd year Master of Computer Applications in Sanketika Vidya Parishad Engineering College, Visakhapatnam, A.P

With his interest in Python, machine and as a part of academic project he chose Image processing using Python. The article has been evolved from an idea to understand the flaws in conventional reporting and keeping time consistency, quality report generation in pulmonary emphysema. A full-fledged project along with code has been submitted for Andhra University as an Academic Project.

Kandhati Tulasi Krishna Kumar: Project Guide & Training & Placement Officer with decade plus experience in training & placing the students into IT, ITES & Core profiles & trained more than9,000 UG, PG candidates & trained more than 350 faculty through FDPs. Authored various books for the benefit of the diploma, pharmacy, engineering & pure science graduating students. He is a Certified Campus Recruitment Trainer from JNTUA, did his Master of Technology degree in CSE from VTA and in process of his Doctoral research. He is a professional in Pro-E, CNC certified by CITD He is recognized as an editorial member of IJIT (International Journal for Information Technology & member in IAAC, IEEE, MISTE, IAENG, ISOC, ISQEM, and SDIWC. He published articles in various international journals on Databases, Software Engineering, Human Resource Management and Campus Recruitment & Training.