CASE STUDY REPORT
DATA MINING
HOME LOAN APPROVAL(2023)

Anakha Ajkumar , Department of Computer Science
Rajagiri college of social sciences,Msccs2206@rajagiri.edu

ABSTRACT

The Home Loan Approval dataset is a group of information about housing loan applications. The dataset has 13 variables, such as the applicant's loan amount, income, employment status, and credit history.
The goal of this project is to use Orange to make a model that can accurately predict whether or not a loan application will be approved. We can predict the most accurate model based on how well different models work. This will help lenders make better decisions and lower the chance of loans not being paid back.

1. INTRODUCTION

Data mining is the process of using techniques like machine learning, statistical analysis, and visualization to find patterns and insights in large sets of data. The purpose of data mining is to discover useful information in a lot of data.This information can then be used for many things, like making business decisions, improving customer relationships, and finding fraud.Data mining usually involves several steps, such as cleaning and preparing the data, analyzing and modeling the data, interpreting and evaluating the results, and so on. Data mining can use a number of different methods, such as classification, clustering, association rule mining, and finding outliers.

The banking industry plays a big role in the present, especially in developing countries where money is usually needed by everyone. By making money, banks can increase the market value of their capital and grow their market share. Customers can save a lot of money in their own accounts at banks. So, banks let business owners and other people borrow money, which they can use to grow their capital and meet their business needs. The money has to be paid back to the bank within a certain amount of time, plus interest. So, interest is the money that banks make when they lend money to people who need it. But banks are worried about whether or not the person who got a loan will be able to pay it back on time. So that they can predict it, they mostly look at things like the applicant's credit score and income. Score on credit. Here, the customer's credit score is the most important piece of information. Most of the time, you need a good credit score to get a loan. If the applicant didn't pay back the loan amount, his credit score would eventually go down on its own. Giving loans to people is one of the most common ways that banks make money. Most of the money made from the loan will go to the bank in the form of interest. The main goal of bank officials is to give loans to people they can trust to pay them back when the time comes. In recent years, banks have been giving loans to customers after going through a step-by-step process, but there's still no way to know for sure if the loan was approved or not. Before approving a loan, banks will try to figure out how risky the application is. This is important for them because they can't give money to people who can't pay it back on time, which could hurt their finances in this very competitive market. So, we've gathered a dataset from Kaggle that has information about loan applicants. It has fields like Gender, Income, Credit History, etc

3. IMPLEMENTATION

a) **Tool Used:-**Orange

Users can engage with and study data in a visual and natural manner with Orange, an open-source application for data visualization and analysis. It supports a variety of data mining and machine learning techniques and is created for both new and experienced users. Orange allows users to quickly input data from numerous sources and display it using a variety of graphs and charts. Moreover, they are capable of operations like filtering, sorting, and merging data tables. Orange's capability to construct machine learning models without requiring any coding is one of its primary characteristics. Users have access to a large selection of algorithms, a graphical interface for configuring their parameters, and a number of model evaluation tools for assessing the performance of their models.Data preparation, visualization, clustering, classification, regression, and text mining are just a few of the many sorts of analysis that Orange may be used for. It is a strong and easy-to-use tool.

b) Data Description

On Kaggle, you may find the publicly accessible dataset known as the Home Loan Approval dataset. Information on house loan applications and their approval status may be found in the Home Loan Approval dataset on Kaggle. There are 614 rows and 13 columns of data in it. The columns include details about the borrower's financial status, the loan's terms, and whether the loan was accepted or not. Each row represents a single loan application.The dataset is intended to assist analysts and data scientists in creating prediction models that can be used to evaluate if a loan application is more likely to be accepted or declined. This is a typical use case in the banking and finance sector, where lenders must determine the risk involved in providing financial assistance to applicants. The attributes in the data set are:

- LoanID
- Gender
- Married
- Education
- Dependents
- Self_Employed
- Property Area
- ApplicantIncome
- CoApplicantIncome
- LoanAmount
- Loan Term
- Credit_History
- Loan Status

Si. No	Attributes	Description	Type
1	LoanID	Unique Identification for each loan application	text
2	Gender	Gender of applicant Female-1/ male-0	Categorical
3	Married	Whether the applicant is married or not	categorical
4	Education	Level of education of applicant Graduate/ Nongraduate	categorical
5	Dependent	Number of dependent applicant has	text
6	Self_Employed	Whether the applicant is self employed or not	categorical
7	Property Area	Area where the property is located Rural/Semi Urban/urban	categorical
8	ApplicantIncome	Income of applicant	Numeric
9	CoApplicantIncome	Income of co applicant	Numerical
10	Loan amount	Amount of loan requested	numeric

11	Loan Term	Term of loan in months	numeric
12	Credit History	Binary variable indicating whether the applicant has a credit history or not	categorical
13	Loan Status	Binary variable indicating whether the loan has approved or not	categorical

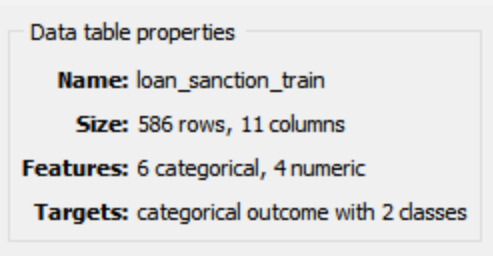
In this set of data, the loan status is the target variable. By looking at the data and making predictive models, lenders can figure out which loan applications are more likely to be approved or denied based on income, credit history, employment status, and other things.

c) Data Preprocessing

Data preprocessing in data mining is the process of cleaning, changing, and reducing data before it is analyzed with machine learning algorithms or other data mining techniques. The goal of data preprocessing is to put the data into a format that can be used for analysis. This can help data mining models be more accurate and work better.

In home loan approval data set,data preprocessing involve several steps:

- Set loan status as target variable.On the basis of that evaluation and model building happens.
- Then, we have to get the statistics for each feature using feature statistics. Then, we have to find the missing values for each attribute and also check if the data is balanced.
- Eliminate columns with meta data and find outliers with a box plot. Because of that, some rows were also taken out.Model-based imputer and most frequent or average can help reduce missing values.
- Change the attribute gender from a categorical to a binary variable in Edit Domain. By making the accuracy better.



d) Data Visualization

Data visualization is the use of pictures to show information and data. It is the process of showing data in a visual form, like charts, graphs, and maps, to help people understand and interpret the data more easily. Data visualization is an important part of data analysis because it can help find patterns, trends, and relationships that might not be obvious in a raw data set. By putting data in a visual format, we can easily find outliers, compare different data sets, and draw attention to interesting areas.

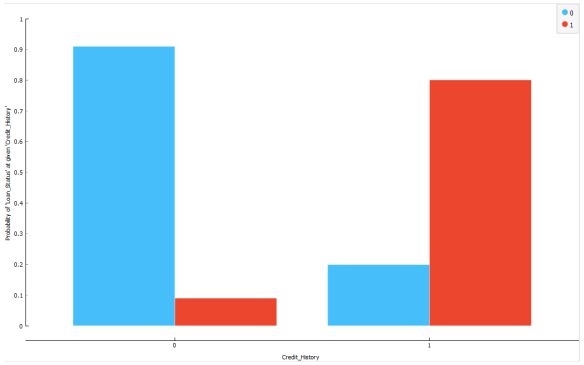


Fig:Credit History v/s Loan Status

From the above plot, we can see that banks can decide whether or not to give a loan to someone based on their credit history. Here, 0 means that the history is not satisfactory, so the loan is not approved, and 1 means that the history is satisfactory, so the loan is approved.

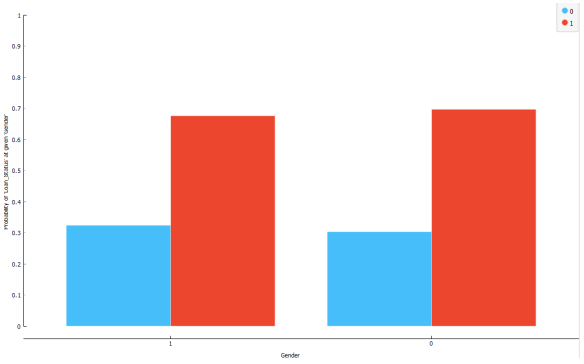


Fig:Gender v/s Loan Status

From the above graph, we can see that loan approval has nothing to do with gender. We changed the categorical variable into a binary variable to make it more accurate.
Then, from the plot, we can see that both men and women have the same chances of getting a loan.

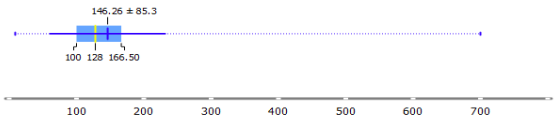


Fig:Loan Amount-Boxplot

We can use the box plot to find the outliers in the above plot. So here, we see the loan amount as a box plot, and from that, we can see that we need less than 250. so we get rid of the rest of the rows.

e) Data Modeling

Data prediction is done using machine learning algorithms,which consist of a target variable that is to be predicted from a given set of predictors We create a function that converts input data into the desired output using this set of variables.The training procedure is carried out repeatedly until the models accuracy on the training set reaches the target level.
Logistic Regression, Decision Tree,Random Forest,SVM,Naive Bayesian are the algorithms used in the study.The Orange connection diagram:

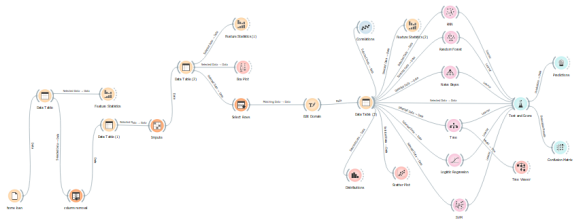
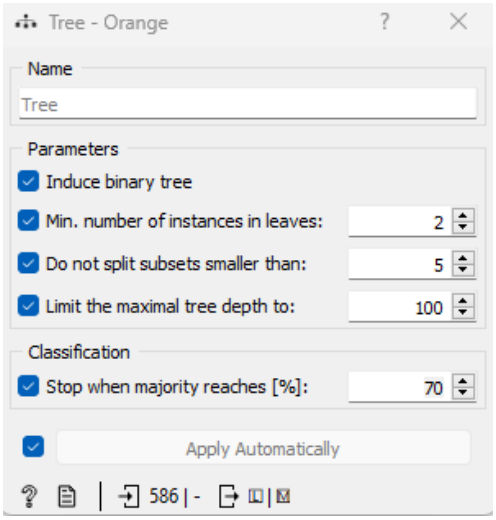


Fig:Model building for home loan approval

The model parameters are given below:

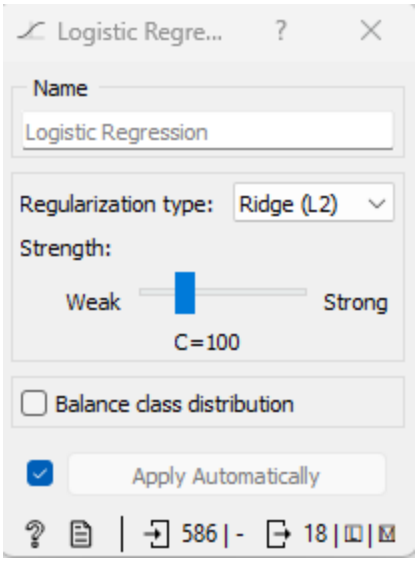
1) Decision Tree

In machine learning and data mining, a decision tree is used as a tool for making predictions. It is a type of algorithm for supervised learning that can be used for both classification and regression.
A decision tree is a model in the shape of a tree that shows a series of decisions and the possible results of each one. The tree structure is made up of nodes and branches. Each node represents a decision or test on a specific feature, and each branch represents a possible result of the decision or test. The tree's leaf nodes show what the end result will be.



2) Logical Regression

Logistic regression is a statistical method used to predict one of two possible outcomes, such as "yes" or "no," "true" or "false," or "0" or "1." It is a type of generalized linear model in which a logistic function is used to model the relationship between the independent variables and the dependent variable.
The dependent variable in logistic regression is a binary, and the independent variables can be either continuous or categorical. The goal is to find the best fit between the independent variables and the probability that the dependent variable is true.

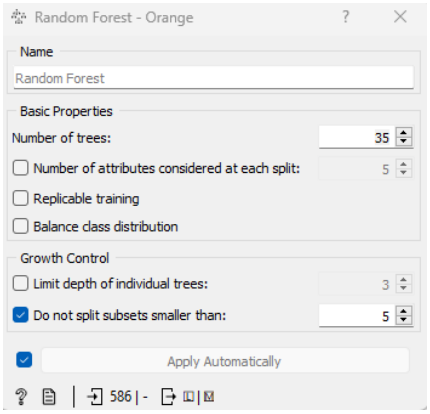


3) Random Forest

It is made up of several trees that work together as a whole. Each tree in the random forest makes a prediction, and the model prediction will be the one that gets the most votes.

The random forest is a classifier that takes the average of a number of decision trees on different parts of a given dataset to improve the accuracy of that dataset's predictions.

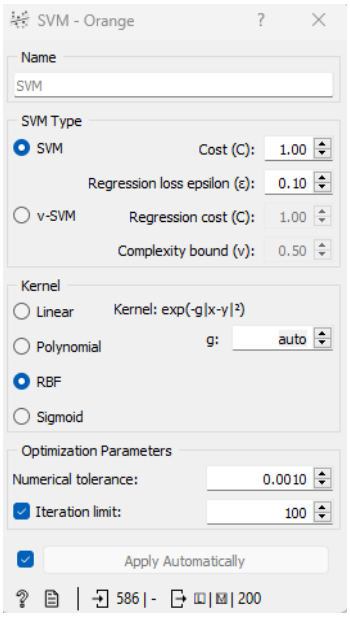
When there are more trees in the forest, the accuracy is better and the problem of overfitting doesn't happen.



4) SVM

Support Vector Machines (SVM) is a flexible and effective machine learning algorithm which is suitable for both classification and regression.. SVM works by finding the hyperplane that best divides the data into two groups (for classification tasks) or that best fits the data (for regression tasks) (in the case of regression tasks).

The main idea behind SVM is to get the distance between the hyperplane and the closest data points from each class to be as big as possible. This is done by solving a constrained optimization problem, which involves finding the hyperplane that maximizes the margin while also making sure that all data points are correctly classified.



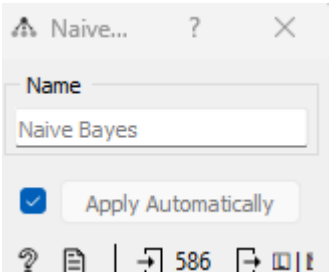
5) Naive Bayesian

It is a supervised classification method based on the Bayes Theorem, which says that predictors are independent of each other. It's based on the idea that the features in the dataset don't have any effect on each other. This is the Naive Bayes formula:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

where A and B are events and $P(B) \neq 0$.

- $P(A \mid B)$ is a conditional probability: the likelihood of event A occurring given that B is true.
- $P(B \mid A)$ is also a conditional probability: the likelihood of event B occurring given that A is true.
- $P(A)$ and $P(B)$ are the probabilities of observing A and B independently of each other; this is known as the marginal probability.



f) RESULT AND DISCUSSION

The performance value obtained for each model:

Model	AUC	CA	F1	Precision	Recall
kNN	0.559	0.643	0.610	0.598	0.643
Tree	0.711	0.817	0.796	0.834	0.817
SVM	0.747	0.817	0.796	0.832	0.817
Random Forest	0.767	0.816	0.794	0.831	0.816
Naive Bayes	0.773	0.816	0.796	0.826	0.816
Logistic Regression	0.770	0.817	0.796	0.832	0.817

From the data, it's clear that these models can be used to predict whether or not a home loan will be approved.

Logistic regression is the most accurate(81.7%) of the classification models that are used.The confusion matrix for the model is:

		Predicted		
		0	1	Σ
Actual	0	82	98	180
	1	9	397	406
	Σ	91	495	586

3 CONCLUSION

In the end, using Orange data mining software to build a model for home loan approval can be a useful way to predict loan approvals. Using techniques like data preprocessing, feature selection, and model evaluation, we can build an accurate and reliable model that can predict the likelihood of getting a loan based on a variety of factors like income, credit score, employment status, etc.

By using Orange's visualization tools, we can learn more about the data and how the variables relate to each other. This can help us find important predictors and get rid of ones that aren't useful or are redundant. Once the model is built, we can use metrics like accuracy, precision, and recall to measure how well it works. We can also use techniques like cross-validation to make sure the model doesn't overfit the data.

Overall, using Orange to build a home loan approval model can give lenders a powerful tool for making decisions and lowering the risk of default or loan delinquency.

By looking at the Home loan approval dataset with different models, we found that logistic regression is the most accurate, with an accuracy of 81.7%.

From what I can see, the thing that most affects the loan status is my credit history.Credit history of 1 means it's good enough to get a loan, and credit history of 0 means it's not good enough to get a loan.But gender doesn't change the dataset much more.Since men and women are almost as likely to get a loan.Even though it is clear from the dataset that the loan can be approved or denied based on the applicant's credit history, income, etc.

4 References

- https://www.kaggle.com/datasets/rishikeshkonapure/home-loan-approval
- https://www.analyticsvidhya.com/blog/2022/02/loan-approval-prediction-machine-learning/?utm_source=related_WP&utm_medium=https://www.analyticsvidhya.com/blog/2022/05/loan-prediction-problem-from-scratch-to-end/
- https://medium.com/@meetpatel12121995/naive-bayes-machine-learning-algorithm-aaf57bdc8d87
- https://www.analyticsvidhya.com/blog/2022/02/loan-approval-prediction-machine-learning/?utm_source=related_WP&utm_medium=https://www.analyticsvidhya.com/blog/2022/05/loan-prediction-problem-from-scratch-to-end/