Loan Approval Prediction Using Machine Learning

Alexandra L. Dakhniuk, Nishant Patel, Siddhant Pillai, Vinu Ratnayake
University of Maryland – Robert Smith School of Business

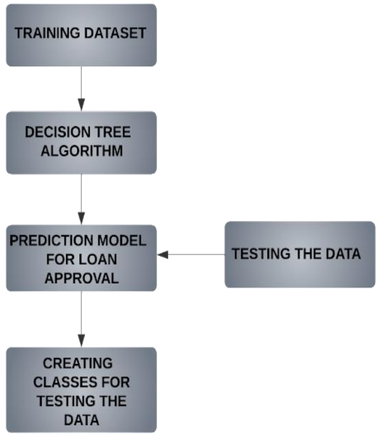
INTRODUCTION

In our project, we explored various machine learning techniques to build a model that could efficiently analyze data and make decisions about loan approvals. We cleaned and prepared a dataset for analysis, and then trained and tested it to improve the model's accuracy. Our research showed that this model could potentially be integrated into an automated loan processing system in the future. To ensure the safety of loan disbursements, we emphasized the importance of collecting and preserving detailed financial information, such as credit history and any delinquencies, for use in exploratory data analysis. We also suggested using a decision tree induction algorithm and the map function to generate loan predictions based on the available data.

LETERATURE REVIEW

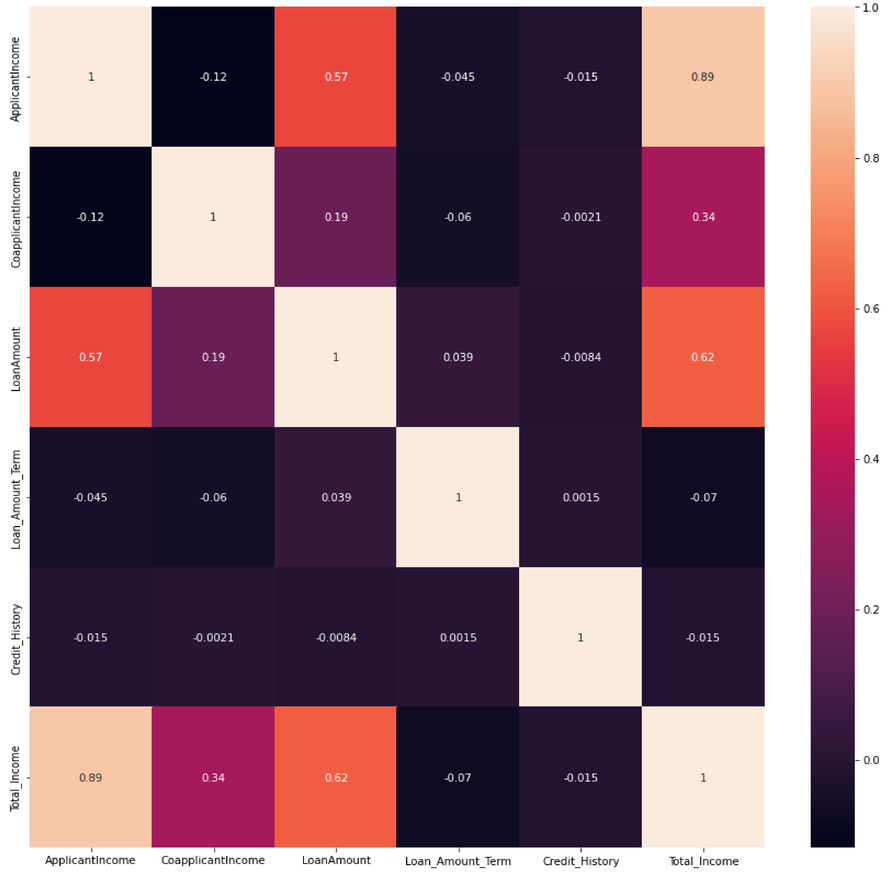
Obtaining loans is a crucial aspect of a bank's operations, as the interest earned on these loans is a significant source of revenue. However, loan companies must carefully evaluate and verify an applicant's ability to repay a loan before granting it. Despite these efforts, there is still a risk that the borrower may experience difficulty repaying the loan. Therefore, it is important for loan companies to have thorough processes in place to assess an applicant's creditworthiness and ensure that they are able to make timely payments.

Figure 1. Architecture of Proposed System



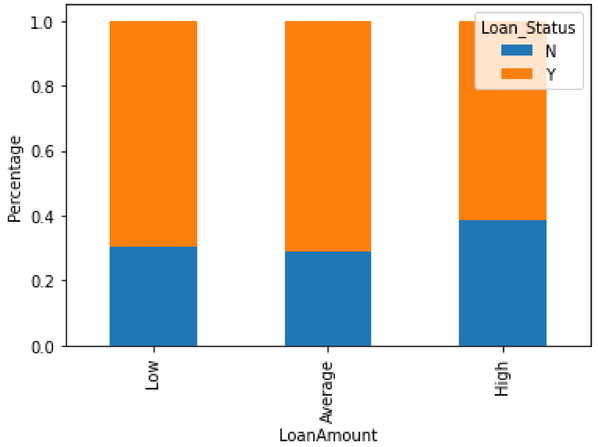
PROJECT APPROACH

Our team conducted research and developed a model to predict loan repayment and improve the efficiency of the loan approval process. The project involved four main steps: selecting and cleaning a dataset, analyzing multiple machine learning methods, training and testing the dataset, and integrating the loan approval system with an automated processing system. To ensure accuracy, we emphasized the importance of collecting and preserving detailed financial information, including credit history and any delinquencies, for use in exploratory data analysis. We suggested using the map function and a decision tree induction algorithm to generate loan predictions based on the available data. We tested several machine learning algorithms, including KNN classification, Naive Bayes, Logistic Regression, Random Forest, and Decision Tree, and found that Random Forest produced the most accurate results. Our classifier technique, which combined KNN and normalization, had a 75.08% accuracy rate. The goal of the model was to screen out applicants. We used a dataset from Kaggle.com and applied data mining techniques, including data cleaning, classification, and data adjustment, to develop the prediction model. min-max.



To understand the relationship between numerical variables, we will use a heat map to visualize the correlations between them. Heat maps use variations in color to represent data, with darker colors indicating a stronger correlation between variables. This visual representation will allow us to easily see which variables are most strongly related to each other. According to the heat map, the variables that show the strongest correlation are ApplicantIncome and LoanAmount, as well as Credit_History and Loan_Status. Additionally, we see that there is a correlation between LoanAmount and CoapplicantIncome.

According to the chart, it appears that the probability of loan approval is higher for low loan amounts, compared to average and high loan amounts. This supports our initial hypothesis. In order to use logistic regression, we will need to convert the categorical variables in our data to numeric values: "N" will be converted to 0 and "Yes" to 1, and the "3+" dependent variable will be changed to a numeric value of 3. We will also convert the target variable categories to 0 and 1 in order to analyze their correlation with the numerical variables in our data.



CONCLUSION

In conclusion, this model compares the logistic regression and decision tree algorithms to determine which one produces the most accurate predictions. Our analysis showed that the logistic regression algorithm is more effective. This model can be used to predict whether an applicant will be able to repay a loan, which will reduce the workload of loan officers and improve the efficiency of the bank. Our findings suggest that machine learning can be applied to other areas of the banking industry to enhance efficiency.