Predicting Approval of Home Loans

Shiying Cai [1] **Xixiang Liu** [1] **Paru Meyyappan** [1] **Willy Lan** [1]

Abstract

In order to make it easier for customers to take out loans, they would find it important to figure out what are the biggest factors that would influence the loan eligibility process. Just looking at the data based on what customers filled out in application forms, there are several factors that can influence loan eligibility. In this paper, we use various methods such as linear regression, LASSO, ridge regression,logistic regression, decision trees, random forest, support vector machines, and neural networks, to explore the relationship between some factors and whether the loan was approved. We found the logistic regression to predict loan approval best with an accuracy of 0.8699.

1. Introduction

1.1. About the Dataset

Dream home financing is dedicated to providing consumers with an education in the various mortgage programs and lending alternatives. They also have an extensive network of lenders which offer a wide variety of programs and we are considered to be authority in this area. At Dream Home Financing, they have decades of lender relationships and we are constantly monitoring those lender's guidelines to fully understand the various mortgage options. This enables them to pair clients with the best lender based upon your personal scenario.

Dream home financing has a presence across all urban, semi-urban and rural areas. The customer first applies for a home loan after that company validates the customer's eligibility for a loan. The company wants to automate the loan eligibility process (real-time) based on customer detail provided while filling out the online application form. These details are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and others. To automate this process, they have given a problem identifying the customer segments eligible for loan amounts to target these customers specifically.

1.2. Objectives and Plan of Analysis

We wanted to create a model that could predict whether a client would be approved for a loan or not with high accuracy. We started with data pre processing, outlier detection, feature transformation, and missing value imputation. We then implemented various supervised and unsupervised classification models to best fit our data including: linear regression, lasso, ridge, k nearest neighbors clustering, logistic regression, decision trees, random forest, support vector machines, linear discriminant analysis and a simple neural network. We decided to find the model of best fit by choosing the model with highest accuracy. Lastly, we discuss the conclusions, dataset limitations, and potential future work.

2. Data

2.1. Data Pre-processing

Based on the nature of our dataset, utilizing packages and local methods to preprocess it is necessary. Pandas is used to analyze and manipulate the tabular dataset. Numpy, which provides the much more efficient ndarray and its supporting functions, is used to reduce time and space complexity. To implement machine learning, we transform non-numerical features into reasonable numerical values. For example, the home loan organization takes the applicant's education status into consideration, and (Graduate, Not Graduate) is converted to (1, -1).

Numerical features are also transformed to improve our algorithms. For example, income is a vital factor, while it may vary much more than other features, since both one with a high salary and one unemployed may apply for the loan. This can hinder our models from making correct predictions. We take a log transformation to make the data as normal as possible to reduce the skewness, so that our analysis results can be more valid. In fact, such transformations do bring significant improvements to our models. For example, the accuracy of our logistic regression model is 65% before the log transformation of income, while it reaches 87% after the transformation.

Due to the dataset's relatively small size, each data is valuable. Data of applicants with missing features are also preprocessed so that they become usable. Missing entries are assigned a reasonable value, such as a new number meaning

uncategorized, or the median of this feature's other data.

2.2. Outliers, Imputations, and Transformations

Despite the fairly clean dataset we decided to do outlier detection. We decided to remove points that are 3 standard deviations away from the mean. This is a commonly used method of outlier removal. Less than 5 points were removed in this stage. Further we decided to impute missing values to keep the size of our dataset, as it is already not super large. We imputed numerical features with the median and imputed categorical variable with the mode, or most commonly chosen level. Lastly, we made some simple transformations. We decided to take the log of the 2 features that are measured in dollars: applicant income and co-applicant income. A log transformation helps reign in outliers in terms of income so that the range of values is more reasonable. Thus, super large or super small incomes don't have a lot of influence or leverage over the regression models.

2.3. Exploratory Data Analysis

Before pre-processing data, exploratory data analysis is performed to check the data quality and characteristics. The dataset under consideration comprises 614 observations, of which 68.73% were approved and 31.27% were declined. Moreover, out of all the applicants, 81.36% were male and only 18.64% were female. We sought to investigate the basic relationship between client features, such as marriage status, gender, and education level, and loan approval. To do so, we generated bar plots that compared loan status grouped by different characteristics. The bar plots led us to a few initial observations, including that married applicants were more likely to receive approval and that having a credit history was also advantageous for applicants. One important aspect of this analysis is to check for multicollinearity among numeric features. Correlation matrix is used to visualize pairwise correlations between features, and the resulting heatmap provides a quick way to identify any strong correlations. In our case, we observed a moderate positive correlation of 0.57 between the applicant income and loan amount variables, but we decided to keep both variables in our dataset after careful consideration of their relevance and potential insights. It is important to note that high correlation does not necessarily imply causation, and in our case, it may simply reflect the fact that applicants with higher income tend to apply for larger loans.

Figure 1. The heatmap displays the correlation matrix of the numeric features in the dataset, with ApplicantIncome and LoanAmount having the strongest positive correlation.

3. Methodologies

3.1. Multiple Linear Regression

Although we approached this problem as a classification task, we also wanted to explore linear methods to gain insights into the relationship between the independent variables and the dependent variable. Multiple linear regression is utlilized for us to understand how these independent variables affect the dependent variable. To achieve consistency with other methods, we set a threshold such that predicted y values greater than 0.5 are considered as approved (1) in our case. We started with a simple multiple linear regression, fitting all 11 variables in the dataset. While the resulting model had a low test MSE of 0.13 and an accuracy of about 0.85, the R-squared value was only 0.296. This indicates that the model can explain only 29.6% of the variation in the dependent variable. The fact that many independent variables in the model are categorical variables can impact the model's performance and interpretation, even though we encoded them as dummy variables. Interpreting the coefficients of these variables can be challenging, as unlike continuous variables, the coefficients for categorical variables do not represent the change in the dependent variable for a one-unit increase in the predictor variable.

3.2. LASSO Regression

LASSO regression is a regression analysis method that applies a penalty term to the regression coefficients to shrink them towards zero, which is often used when we want to select important predictors and reduce the model complexity. We attempted to improve the performance of the model by adding penalty terms through LASSO regression. However, the accuracy decreased to 0.683, and the negative R2 score of -0.47 indicated the presence of overfitting, and the model is performing worse than a model that simply predicts the

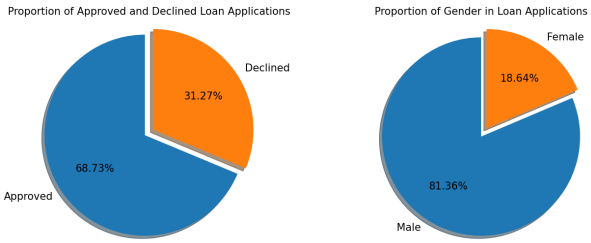

Figure 2. The pie charts show the distribution of two variables in loan applications. The chart on the left displays the proportion of approved and declined loan applications, with 68.7% of applications being approved. The chart on the right displays the gender distribution of loan applicants, with 81.36% of applicants being male.

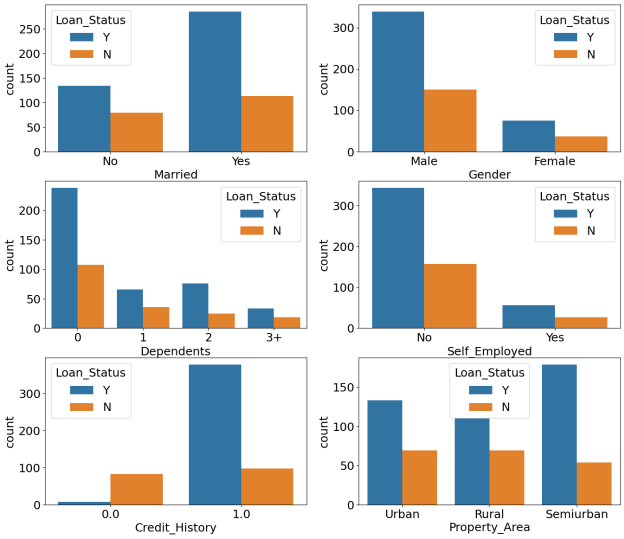

Figure 3. The plots show how loan approval varies with different demographic and financial factors. Each plot displays the count of loan approvals and rejections for different categories of a factor. The factors include marital status, gender, dependents, self-employment, credit history, and property area. These plots provide valuable insights into the factors that influence loan approval.

mean of the dependent variable. One possible reason for the decrease in performance is that LASSO regression is better suited for continuous predictors than categorical predictors, and the categorical variables in our dataset may not be well-handled by LASSO regression.

3.3. Logistic Regression

Logistic Regression is a supervised classification algorithm used to assign observations into a discrete set of classes. In our case, there are exactly 2 classes, so we are doing a binary logistic regression. Logistic Regression uses a sigmoid function to map predicted values to probabilities between 0 and 1. This algorithm typically performs well when the dataset is simple and linearly separable. The accuracy of our Logistic Regression model is 0.8699. This means that around 87% of the time this model correctly predicts whether a client's loan application was approved or not.

3.4. K-nearest Neighbors

The K-nearest Neighbors algorithm is a non-parametric, supervised learning classifier. This algorithm utilizes the concept of proximity to predict the classification of each individual data point. We train various KNN models with different values of parameters, like definition of proximity and leaf size, and then choose the optimal one. The best performing KNN model has an accuracy of 0.7073. This means that around 71% of the time this model correctly predicts whether a client's loan application was approved or not.

3.5. Ridge Regression

Ridge regression is a regularization technique used in linear regression to prevent overfitting by adding a penalty term to the cost function. This penalty term is proportional to the square of the magnitude of the coefficients, forcing them to be smaller. The strength of the penalty is controlled by a hyperparameter called the regularization parameter. The accuracy of our ridge regression was 0.78862. This means that around 79% of the time this model correctly predicts whether a client's loan application was approved or not.

3.6. Decision Tree

A decision tree is a type of supervised machine learning algorithm used for classification and regression tasks. It works by recursively splitting the data into smaller subsets based on the values of the input features, using a set of decision rules. At each split, the algorithm chooses the feature that best separates the data into the target classes or produces the best regression fit. The accuracy of our decision tree was 0.7496. This means that around 75% of the time this model correctly predicts whether a client's loan application was approved or not. The pruned random forest of max depth 3 layers has an accuracy of 0.7804 which is a 78% accuracy in predicting approval of loan applications.

3.7. Random Forest

Random forest is a popular ensemble learning technique used for classification and regression tasks. It works by building multiple decision trees using randomly selected subsets of the input data and features, and then combining

their predictions through a voting or averaging process. This helps to reduce overfitting and improve the accuracy and robustness of the model.The accuracy of our random forest was 0.7804: around 78% of the time this model correctly predicts whether a client's loan application was approved.

3.8. Support Vector Machine

Support vector machines (SVMs) are a type of supervised machine learning algorithm used for classification and regression tasks. They work by finding the hyperplane that best separates the data into the target classes or produces the best regression fit. The hyperplane is chosen such that the distance between the hyperplane and the nearest data point from each class (called the margin) is maximized. The accuracy of our SVM was 0.78862. This means that around 79% of the time this model correctly predicts whether a client's loan application was approved or not.

3.9. Linear Discriminant Analysis

Linear discriminant analysis (LDA) is a classification technique used in supervised machine learning. It works by projecting the input data onto a lower-dimensional space while maximizing the separation between the classes. The projection is chosen such that the ratio of the between-class variance to the within-class variance is maximized. The accuracy of our LDA was 0.78862. This means that around 79% of the time this model correctly predicts whether a client's loan application was approved or not.

3.10. Simple Neural Network

A simple neural network is a type of machine learning algorithm inspired by the structure and function of the human brain. It consists of an input layer, one or more hidden layers, and an output layer of interconnected nodes (neurons) that pass signals (activations) from one layer to the next. The network learns to make predictions by adjusting the strengths of the connections (weights) between the neurons during training using a mathematical optimization algorithm. The accuracy of our NN was 0.63414. This means that around 63% of the time this model correctly predicts whether a client's loan application was approved.

4. Discussion

Based on the accuracy, the best model for predicting whether or not a clients' loan application has been approved is the logistic regression with an accuracy of 0.8699. This follows as logistic regression is a well known classification method. For the logistic regression we specified 500 as the maximum number of iterations and 0.1 as the C-value. Figure 5 shows the confusion matrix for the logistic regression that a small part of the test dataset was misclassified: 2 people who were actually approved for the loan were predicted to not be approved and 14 people were predicted to be approved and in actuality weren't.

Table 1. Prediction accuracies on test data for different models

MODEL	ACCURACY
MLR	0.8470
LASSO	0.6831
RIDGE	0.7886
LOGISTIC	0.8699
KNN	0.7739
D.TREE	0.7804
RANDOM FOREST	0.7804
SVM	0.7886
LDA	0.7862
SNN	0.7886

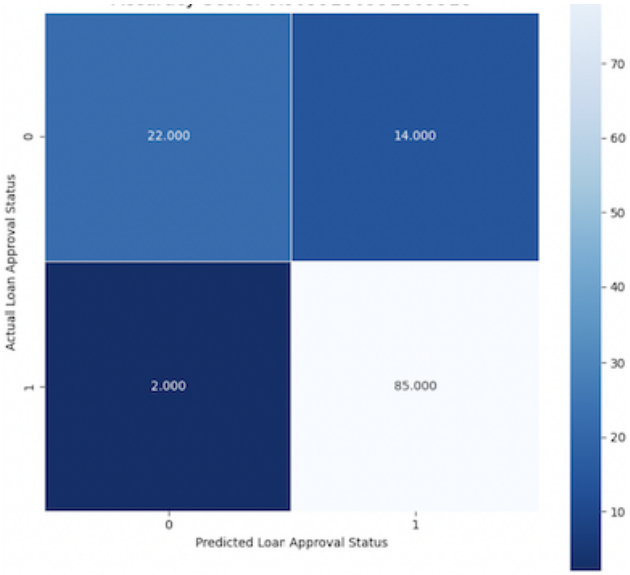

Figure 4. Confusion matrix for logistic regression.

Despite the fairly high accuracy score and well fit model our dataset had some limitations mainly the size. Classification models improve their predictive ability and become more robust as they train on more and more data. Since this dataset was relatively small, the accuracies and metrics aren't as high as they could be. Further certain variables had many missing values or weren't very important. Perhaps the analysis could be improved with more features about the clients to help aid in the process. There's a lot of potential work in this domain: banks and financial institutions are slowly using more and more technology and data to aid in decision making. Given the sheer amount of transaction and financial data this is a large untapped market that could bring about really useful innovation.

5. References

[1] About Us. Dream Home Financing. (2023, February 10). Retrieved May 7, 2023

[2] Koehrsen, W. (2018, January 17). Random Forest in python. Medium. Retrieved May 7, 2023

[3] Konapure, R. (2023, January 12). Home loan approval. Kaggle. Retrieved May 7, 2023

[4] Navlani, A. (2018, December 28). Python decision tree classification tutorial: Scikit-Learn Decisiontreeclassifier. DataCamp. Retrieved May 7, 2023, from https://www.datacamp.com/tutorial/decision-tree-classification-python