# Determining Loan Eligibility Via Machine Learning Methods

Paulie Lee, Peter Kelleher
EE 600: Machine Learning
San Diego State University
16 December 2021

## Introduction

Many Americans, since the 2008 market crash, are experiencing increasing financial stress and reliance on loans. Federal minimum wage has not increased since 2009, while inflation, cost of living, and home prices have increased tremendously. Total outstanding personal loan debt has doubled from $72B in 2015 to $143B in Q1 2019 [1]. Over half of borrowers intend to use personal loans for refinancing or consolidating debt [4].

Loans carry great risk and reward for both borrower and lender. Thus, it is essential to accurately and quickly assess a borrower's eligibility for approval. Machine learning techniques can be employed to analyze applicant information and determine approval status.

To explore the feasibility and performance, a dataset of home loan information was subject to supervised learning algorithms and compared. The techniques utilized are logistic regression, neural network, and Bayes' naive classification.

## Related Work

[2] In this paper, the problem was defined as trying to find a way to determine credibility of the borrower without sorting through variables or factors for each individual applicant. They note the weaknesses of the existing CIBIL score system.

Their data analysis included looking for missing values, as well as outliers. They handled missing values by the same method that we intend to, which is replacing categorical values with the most common occurring value in that feature (i.e., mode imputation) and replacing continuous values with the mean of that feature (i.e., mean imputation). Their method for dealing with the outliers that they identified is to normalize those features by applying a log transformation. We will also normalize our variables, but not by a log method.

They have chosen to only consider variables from the dataset that they consider to have "a direct impact on loan eligibility" (credit history, education, self-employment, property area). They did this to avoid overfitting due to too many variables. This method risks leaving out variables that could end up being influential in determining whether someone is approved for a loan. Our analysis will include all variables.

They perform analysis only using logistic regression and do not compare it to other methods. The authors do not go into detail about their results, but describe their model accuracy as "satisfactory" and note that the model they built would be a worthwhile replacement for existing methods of determining whether to grant someone a loan.

[3] In the second paper, the authors seek to use a real-time binary classification model method to "classify a loan applicant as a good or bad risk."

They start by creating a dataset using featuring engineering, where historical data is transformed into the features they will utilize. Upon training their model, they intend to use it with a "real-time streaming data pipeline."

They create a deep neural network by using an auto-encoder, which compresses an input into fewer bits, and then reconstructs the original input. It is not made clear in the paper how the auto-encoder creates a neural network. The final neural network used contains 18 inputs, six hidden layers, and 18 outputs. The 18 outputs seems strange, I would expect just one output for this machine learning task.

They prep their dataset for the study by converting categorical data to integers and removing some variables due to them being deemed useless. In addition to their neural network, three other classification methods are used for a comparison. The other methods are SVM regression, logistic regression, and lastly a non-linear artificial neural network.

After performing machine learning with all four methods, they averaged the result of four runs for each method. They found that the deep learning neural network had the best result with an accuracy of 87%. Close behind was linear SVM regression at 86%. They conclude that, while the deep learning neural network had the best accuracy, the result may be misleading due to the "accuracy paradox." The authors also intend to tune their program in hopes of getting better results.

## Problem Formulation

The dataset used for analysis contains an anonymous list of 614 individuals along with the lenders' required information for loan approval determination; a brief explanation of that data is provided below in Data Description and Experimental Setup.

The goal is to utilize machine learning techniques to create a model that can determine loan approvals with an accuracy of 80% or greater. 80% was chosen due to the performances observed in the related works. The analysis will utilize only techniques covered in EE600 Seminar: Machine Learning.

Unfortunately, over 20% of the raw dataset includes missing data (i.e. one or more features blank). Preprocessing techniques stated in Preliminaries were used to fill in missing data, which may affect the confidence in our accuracy. Additionally, the data source does not explicitly state where and when the data was collected. Due to the health of the economy, government, and local municipalities, it is impossible to determine where the machine learning model(s) can be applied. Privacy aside, without knowledge of the loan originator(s), it is possible that the data collected is not a representative sample for home loan approvals. Also, the dataset did not collect critical information required by some home loan approval processors in the United States of America: credit score, type of loan (e.g. jumbo, FHA, fixed/adjustable rates), or debt-to-income ratio. The "credit history" feature of the dataset contains only binary values ("0" or "1"), whereas the USA FICO credit system ranges from 300 to 850. Although a binary value is sufficient for the analysis, utilizing the entire FICO range may provide a more robust model.

Regardless, the machine learning model will provide insight as to whether a machine learning algorithm is efficient, and later tweaking can be performed to localize the model(s).

## Preliminaries

Having defined the problem we are intending to solve, several steps will have to be taken before we can begin working with the dataset. The first pre-processing step will be to fill in the missing values. There are multiple ways to handle missing values, including removing entries from the dataset completely. Since 134 out of 615 samples have missing data, this method would diminish the size of our dataset significantly and is not feasible. Other methods include mean and mode imputation, and linear regression. The most appropriate method for this dataset is to use mode imputation for binary features and mean imputation for the continuous features. The second pre-processing step to perform is converting non-numerical values to numerical so that they can be used with our machine learning algorithm. As most of the non-numerical values in our dataset are binary (e.g., male or female) we can simply convert them to ‘0’ and ‘1’. The single non-binary, non-numerical feature has three possible values, so we will use a technique called “dummy coding” and replace it with two separate features, each with a value of either ‘0’ or ‘1’.

Since the dependent variable of our dataset is binary (approved for loan or not), the most appropriate machine learning technique to use is logistic regression. Logistic regression involves using a maximum likelihood estimation to find the optimal parameters of the model. As a comparison, we will also use neural networks and the naïve Bayes method, which is a classification technique where it is assumed that none of the features are related to each other.

## Data Description and Experimental Setup

The raw data contains the information stated below in Table 1 for 614 anonymous individuals:

| **Variable** | **Description** | **Data Format** |
|---|---|---|
| Loan_ID | Unique Loan ID | Key / Unique ID |
| Gender | Male or Female | Text |
| Marriage Status | Married or not married (Y/N) | Text |
| Dependents | Number of Dependents | Text (0, 1, 2, 3+) |
| Education Status | Graduate or Undergraduate | Text |
| Self-Employed | Self Employed (Y/N) | Text |
| Applicant Income | Income per month | Integer |
| Co-applicant Income | Income per month | Integer |
| Loan Amount | Loan amount (thousands) | Integer |
| Loan Amount Term (Duration) | Loan amount duration in months (increments of 12) | Integer |
| Credit History | Does credit history meet guidelines? (0, 1) | Boolean |
| Property Area | Urban, Rural, or Semiurban | Text |
| Loan Status | Output of interest (Y/N) | Text |

Table 1: Breakdown of raw data used for loan approval analysis.

After preprocessing the data, the completed dataset was loaded into MATLAB. From there, 75% of data was designated into a training set, and the remaining 25% into a testing set. Designation was performed by establishing a sequence of randomly ordered numbers from 1 to 614 and drawing the first 461 as the training set.

The data was randomly split 5 times to assess whether randomizing the training and test datasets had any significant influence on the accuracy. For each iteration, logistic regression, neural network, and bayes classification were applied to the training set to produce a machine learning model.

The neural network parameters were alpha = .10, 5 hidden layers, and number of epochs set to a variable number. Each iteration of the neural network would compute until the change in cost fell below .05.

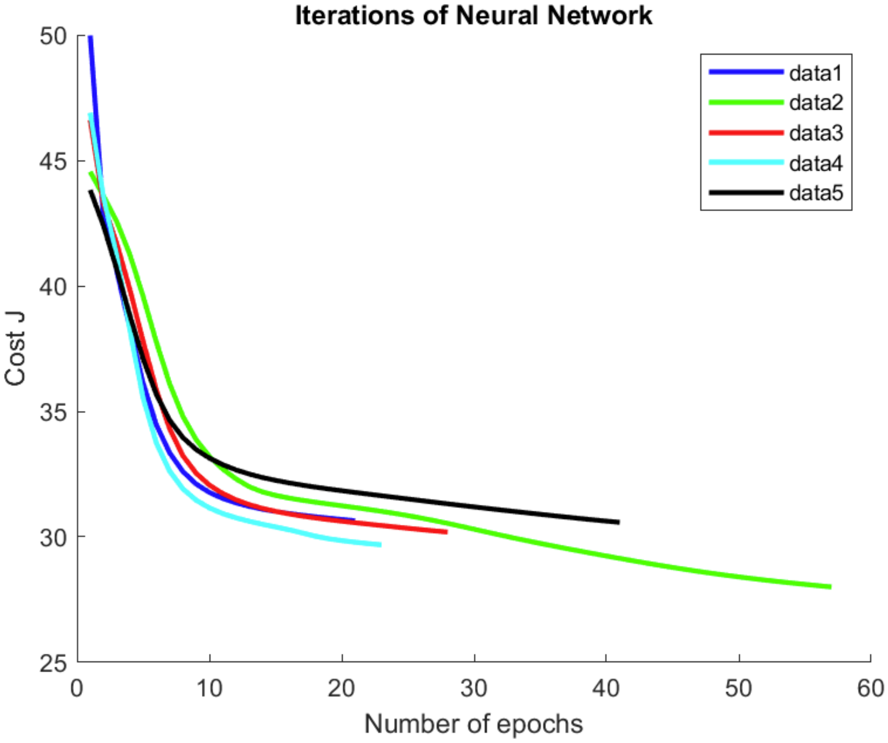

Figure 2: Cost vs. Number of Epochs for Neural Network method

Bayes classification was implemented using the MATLAB function **fitcnb()** from the MATLAB Statistics and Machine Learning Toolbox, which outputs the property DistributionParameters used for classification.

After the training sets were analyzed, all computed model parameters were used with the training set inputs to compute a predicted value, which was compared to the actual test set data. Accuracy was computed for each iteration and the average of all 5 runs serves as the reported value.

## Experimental Results and Analyses

The accuracy of each machine learning model is tabulated below in Table 2. They are all roughly the same; however, the run-time for the neural network algorithm consumed 3X more time than the other two.

| | Logistic Regression | Neural Network | Bayes |
|---|---|---|---|
| Iteration 1 | 0.8270 | 0.8000 | 0.8108 |
| Iteration 2 | 0.8000 | 0.7946 | 0.7730 |
| Iteration 3 | 0.8000 | 0.7892 | 0.7946 |
| Iteration 4 | 0.8216 | 0.8108 | 0.8108 |
| Iteration 5 | 0.8162 | 0.8162 | 0.7892 |
| Average | **0.8130** | **0.8022** | **0.7957** |

Table 2: Accuracy of machine learning model per iteration

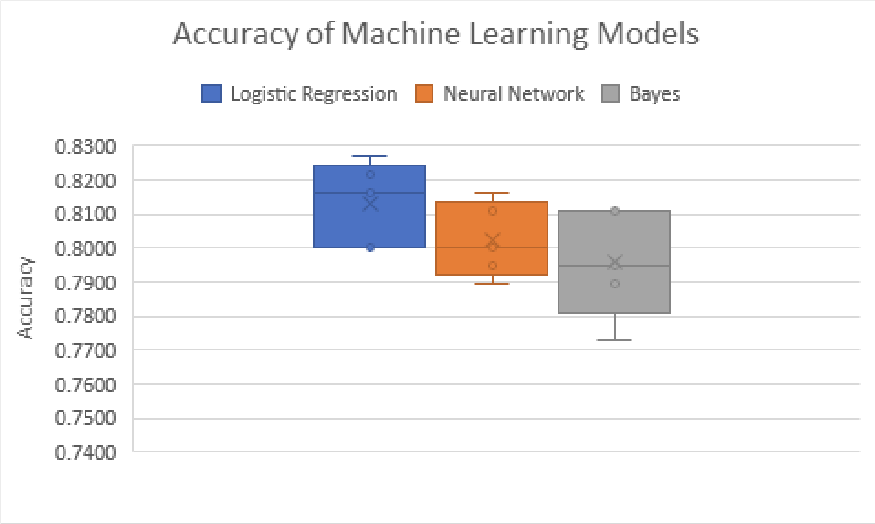

Table 3: Box & whisker plot of machine learning models

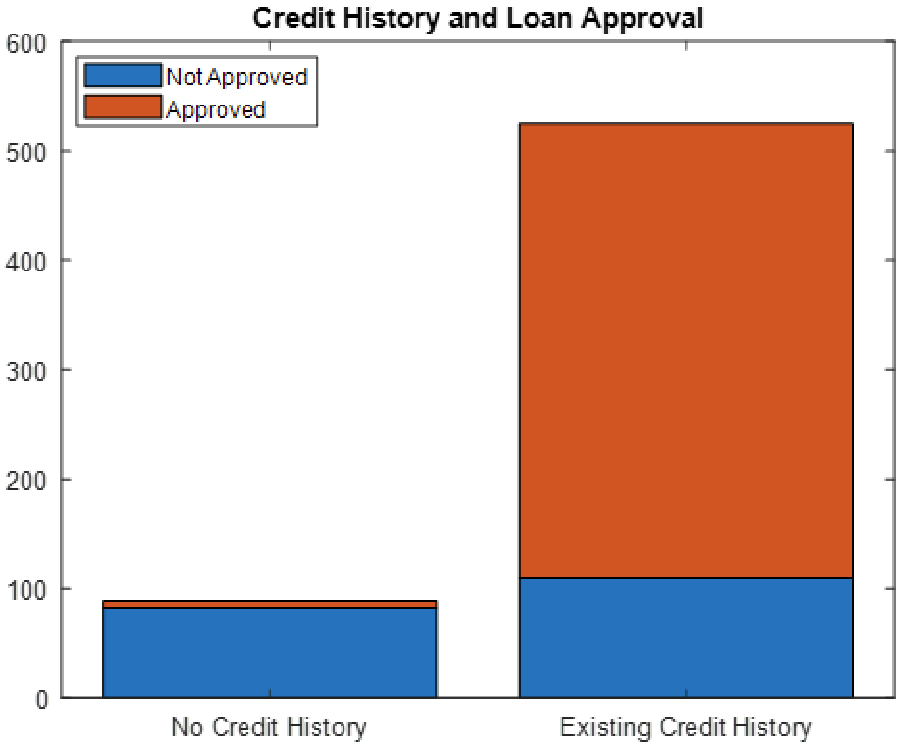

Figure 2: Credit History vs. Loan Approval

According to the beta parameters computed from the logistic regression, the largest parameter was feature 11: credit history. Credit History is a statistically significant coefficient. Figure 2 above shows the dataset's distribution of loan approvals for those with and without credit history. Of the 89 applicants with no credit history, 92% were denied, whereas only 21% of the 525 applicants with credit history were denied. This suggests that applicants with credit history are favored over those without, though it would be helpful to obtain a larger number of applicants in the latter category to further explore this observation. The statement does coincide with the fact that credit history is a critical factor in nearly all aspects of large financial purchases in the USA.

## Conclusion

In this paper, we analyzed the use of several machine learning techniques to determine whether a potential borrower should be approved for a loan. The use of machine learning allows for a lender to very quickly determine whether to approve someone for a loan and supplies the added benefit of putting all borrowers on equal footing. Since everyone is judged based on the same information, there is no potential for a loan officer to turn down an otherwise qualified applicant based on their own personal biases. Further research that could be done on this topic include finding ways to improve accuracy, exploring other potential variables that are not included in this dataset, such as the borrower's current debt, and determining the best interest rate for a particular borrower.

## References

[1] Chamber of Commerce. “Personal Loan Statistics.” *Chamber of Commerce.org*, 2019, https://www.chamberofcommerce.org/personal-loan-statistics. Accessed 11 December 2021.

[2] S, R., Shekhar Jha, P., Raghupathi Vasishtha, I., H, S., & Zafar, N. (2021). Monetary Loan Eligibility Prediction using Machine Learning. *IJESC*, *11*(07), 28403–28406. https://ijesc.org/upload/0e4caa4fba55382053b74cf62fe7e8aa.Monetary%20Loan%20Eligibility%20Prediction%20using%20Machine%20Learning%20(1).pdf

[3] Abakarim, Y., Lahby, M., & Attioui, A. (2018). Towards an efficient real-time approach to loan credit approval using Deep Learning. *2018 9th International Symposium on Signal, Image, Video and Communications (ISIVC)*. https://doi.org/10.1109/isivc.2018.8709173

[4] Schulz, Matt. “Personal Loan Statistics | LendingTree.” *Lending Tree*, 11 March 2021, https://www.lendingtree.com/personal/personal-loans-statistics/. Accessed 11 December 2021.

## Individual Contribution

| Section Title | Name of Contributor |
|---|---|
| Introduction | Paulie |
| Related Work | Peter |
| Problem Formulation | Paulie |
| Preliminaries | Peter |
| Data Description and Experimental Setup | Paulie |
| Experimental Results and Analyses | Paulie & Peter |
| Conclusion | Peter |
| References | Paulie & Peter |
| Individual Contributions | Paulie & Peter |