**Logistic Regression for Loan Prediction**

**https://www.kaggle.com/datasets/vikasukani/loan-eligible-dataset?resource=download**

Eli Dalton, [edalton2@bellarmine.edu](mailto:edalton2@bellarmine.edu)

**ABSTRACT**

In this project my goal was to create a functioning logistic regression model using machine learning with a dataset of choice. Once I found the Loan eligibility dataset, I was specifically looking to predict loan approval based on certain data inputs.

1. **INTRODUCTION**

I found a dataset with several inputs about consumers who were applying for home loans. This was created with data from the Dream House Finance company. Customer information that was used was for customers who had already been approved at a basic eligibility level. For the purpose of this assignment I chose to apply the predictive machine learning classification model to the dataset.

1. **BACKGROUND**
   1. *Data Set Description*

This dataset was found on Kaggle. I chose to go with this dataset because being an accounting major I enjoy numbers, and I have had a decent bit of consumer loan exposure through a couple of past internships, so I thought it would be interesting to work with. This data was collected by Dream Housing Finance company because they were seeking to automate their loan eligibility process. They were hoping to be able to give a real-time answer to customers based on their inputs.

* 1. *Machine Learning Model*

The logistic regression model is a relatively straightforward process. It is made easy by the libraries already available in Python. First the dependent variable is split from the dataset because we are trying to use the independent variables to predict the dependent variable. Once the dataset is split in this way, it is then split into training and test sets, the model is applied to the training set to align it with the data. Then predictions can be made with the test set inputs and the accuracy of the model can be evaluated.

1. **EXPLORATORY ANALYSIS**

The dataset is comprised of the data from 614 unique loans. It contains eleven different independent variables that will be used to predict the dependent variable. Because it was not useful for the purposes here, the Loan ID column was dropped from the dataset. Seven different columns had some missing values, but not any huge percentage of any variable. The categorical variables were filled with the mode of their respective columns, and the numerical data was filled with their respective medians because they were all skewed (See tables below)

**Table 1: Data Types**

|  |  |
| --- | --- |
| *Variable Name* | *Data Type* |
| Loan ID | object |
| Gender | Object |
| Married | Object |
| Dependents | Object |
| Education | Object |
| Self-Employed | Object |
| Applicant Income | Int |
| Co-applicant Income | Float |
| Loan Amount | Float |
| Loan Amount Term | Float |
| Credit History | Float |
| Property Area | Object |
|  |  |

1. **METHODS**
   1. *Data Preparation*

To prepare the data for my model I had to do a few things. First, there were seven different variables that had some missing values. Therefore, I went through and filled in the missing values using the mode of the categorical variables and the median of the numerical variables because they were all skewed so the mean would not work. The Loan ID column was also unnecessary for the purpose of this assignment because it is just an identifier for each unique loan which has no bearing on whether or not a loan is granted so I deleted that column. All of this prepared the data for the logistic regression.

* 1. *Experimental Design*

I performed three different experiments with this data. I ran the model three different times adjusting the split ratio between the training and testing sets each time.

Table X: Experiment Parameters

|  |  |
| --- | --- |
| **Experiment Number** | **Parameters** |
| 1 | Training set = 80%, testing set = 20% |
| 2 | Training set = 90%, testing set = 10% |
| 3 | Training set = 70%, testing set = 30% |

* 1. *Tools Used*

Python was the only programming language used for this assignment, but multiple python libraries were used at various points for the assignment. Numpy, pandas, and seaborn were used along with different functions of sklearn. Pandas was used to read the data and along with numpy they were used to manipulate and preprocess the data. Seaborn was used to visualize the numerical data to determine the best method to fill missing data. Sklearn was used for most of the actual logistic regression functions.

1. **RESULTS**
   1. *Classification Measures and Discussion of Results*

Experiment #1:

Table

Description automatically generatedChart

Description automatically generatedText

Description automatically generated

Experiment #2:

Table

Description automatically generatedChart

Description automatically generatedText

Description automatically generated

Experiment #3:

Table

Description automatically generatedChart, treemap chart

Description automatically generatedText

Description automatically generated

* 1. *Problems Encountered*

I struggled to find a dataset that interested me and would work for this assignment. However, once I settled on this dataset there were not any major problems. The main issue was missing data, which I was able to fill in.

* 1. *Limitations of Implementation*

There are definitely some possible limitations of my model. These limitations primarily exist from the dataset. More data entries could be helpful. Also, there could be other variables not included in this dataset that might have a strong affect on determining whether someone gets approved for a loan. I think this type of model could be very affective for this, but you would just have to make sure all the important variables are there.

* 1. *Improvements/Future Work*

The main thing I would like to do to improve my model in future work would be to perform more experiments to see what the right balance would be for the training and test sets. It would also be interesting to see if changing the random state would have any affect.

1. **CONCLUSION**

From the experiments performed for this assignment, my model is less than ideal and in its current state is not accurate enough to use. However, there was a positive change seen in the accuracy, so it is possible that it might could become usable in the future with more work and adjustments.

**REFERENCES**

No references.

**TABLES**

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated