**ALY6020 Predictive Analytics**

Dr. Thomas Goulding

Module 3

Logistic Regression Models

Jeff Hackmeister

6/7/2025

**ABSTRACT**

This study utilized a logistic regression model for predicting loan application approval. The analysis was conducted on a dataset of 5,000 applications that contained demographic, personal financial, and banking information for each applicant. The objective is to create an automated decision instrument for loan approval and identifying the variables with the highest predictive value.

After data exploration, a model was fit to the data that achieved a 91.6% accuracy when tested against the known data. However, the supplied data was heavily skewed towards denied applications (91.2% of all observations) this led to lower precision (53.3%) and recall (36.4%) metrics.

Overall, the model was highly accurate but extremely conservative in approving loan applications. The variables displaying the highest predictive influence on the model were income, education and family size.

**INTRODUCTION**

To begin the analysis, all necessary packages were imported into a Jupyter notebook, and the provided spreadsheet was read in.

A screenshot of a computer

AI-generated content may be incorrect.

With the data loaded, an exploration of the data began.

A screenshot of a graph

AI-generated content may be incorrect.

To explore the distribution of the variables a series of plots were created. First, was the Personal Loan variable – this is the target variable of the study and is a binary of whether a previous loan application was accepted. From this, it is observed that most applications have been denied in the past. As part of the exploration of the data, distribution plots were created for all variables, these can be found in the appendix of the study.

A graph of a person loan

AI-generated content may be incorrect.

A correlation heatmap was produced to see the influence of each variable on all others in the dataset.

A screenshot of a computer

AI-generated content may be incorrect.

From this, high correlations are witnessed between Income, Credit Card Average (CCAvg) and CD Account and the target variable of Personal Loan.

**LOGISTIC REGRESSION**

With the exploratory analysis completed, a model can be fit to the data. First, the data is prepared by separating the Personal Loan variable from the rest of the dataset and then creating an 80/20 train and test split of each dataset. Once the model is fit, accuracy can be measured using the model.score function. The resulting score is 90.78% accuracy.

A computer code with black text

AI-generated content may be incorrect.

With the model in place, we can using the remaining test observations to make predictions and produce a confusion matrix.

A yellow and purple squares with numbers

AI-generated content may be incorrect.

From this matrix, both a high level of accuracy and skew towards denying applications can be seen with 876 of the 1,000 predictions being true negative, which in this case is a predicted denial that was also an actual denial in the data.

A classification report shows similar results.

A screenshot of a computer

AI-generated content may be incorrect.

The overall accuracy of 91% closely aligns with the original results. The high precision (93%), recall (97%) and F1 score (95%) for the denial cases are very strong. However, the much lower values for the loan approval values indicate a high bias towards denying an application. This is a highly conservative model for loan approvals.

**PREDICTIVE VARIABLES**

When looking at the predictive power of each variable, there are clear high performers.

**A screenshot of a computer

AI-generated content may be incorrect.**

Income, Education and Family (a measure of family size) are the most predictive of all variables. The are also all positive predictors, meaning an increase in one of these variables leads to higher probability of loan approval.

**CONCLUSION**

The analysis of the data and the performance of the model indicates that a highly accurate predictive model can be fit for loan approval. In the current state, the model is highly conservative and will deny over 90% of applications received. The results of this study do not, and can not, make a determination of the appropriateness of that conservative nature. In a period of economic uncertainty, this model could be very useful in limiting the bank’s exposure to potentially risky loans. However, if the economic climate indicates expansion of the overall economy, making adjustments to the model approve a higher percentage of applications may be warranted. In this case, adding additional weight to the income, education and family variables would be a prudent decision to add additional customers while minimizing risk to the institution.

**REFERENCES**

[1] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning : Data Mining, Inference, and Prediction, Second Edition*. Springer New York.

[2] OpenAI. (2024). ChatGPT (Claude 4) [Large language model]. https://claude.ai

‌[3] Steele, B., Chandler, J., & Reddy, S. (2016). *Algorithms for Data Science*. Springer.

**IMAGE APPENDIX**

A graph of a number of bars

AI-generated content may be incorrect.A graph of income distribution

AI-generated content may be incorrect.A graph of age distribution

AI-generated content may be incorrect.A graph of a mortgage distribution

AI-generated content may be incorrect.A graph of a graph

AI-generated content may be incorrect.A graph with a purple and yellow rectangle

AI-generated content may be incorrect.A bar chart with a purple rectangle

AI-generated content may be incorrect.A graph with different colored bars

AI-generated content may be incorrect.A graph of a credit card

AI-generated content may be incorrect.A graph of a bar graph

AI-generated content may be incorrect.A graph of a person loan

AI-generated content may be incorrect.

Model Performance Metrics

A graph of different colored bars

AI-generated content may be incorrect.