Assessing Risk Attributes for Personal Loans in Financial Services through Logistic Regression and Neural Networks

Sean Furlong

### Air Force Institute of Technology

### *Abstract* — This this project is a demonstration into creating logistic regression models and neural network models that can make the determination of potential risk for financial services institutions when issuing loans to clients. This project takes a large data set that has input variables including a client's age, income, and loan amount. The data set also provides and output variable that determines whether that client defaulted on that loan or was able to pay it off in a timely manner. throughout this project, a logistical regression analysis is completed, and a model is created they can predict these potential risk factors with 88% accuracy. the project also creates a neural network through unsupervised learning which can make the prediction for the user with a 68.6% accuracy rating in the learning process. this is done through creating a multi-level neural network with hyperparameters. The conclusion of this project determines that the models in place are effective enough to make predictions.

*Keywords*  — Logistic Regression, Neural Network, Accuracy, precision, recall, Area under curve (AUC)

### I. Introduction & Background

Problem motivation

The financial services industry depends on leveraging risk in order to benefit consumer as well as the institution. In order to properly manage risk, financial institutions need to be able to determine potential factors that define the difference between an ideal candidate for a loan versus a candidate with who potentially would default. According to Deloitte, “effectively managing risk in the current uncertain and volatile environment will demand new capabilities and a rethinking of how risk management operates” [1]. Increased insecurity in an uncertain economy has called for the need for financial institutions to adapt with new methods and techniques for determining potential risk in loan candidates. It is important to be able to determine based on predefined factors whether a candidate will be able to pay off a loan in order to prevent the financial institution from experiencing loss from a potential default. Financial services institutions need to be able to look at a potential candidate and make the determination on whether there would be a strong fit for a loan based on factors such as age income in the amount that alone was taken out for. As stated by Gregory Phalen, “increases in risk can increase or decrease margins, prices, and loan amounts” [2]. Financial institutions have to look at assessing loan candidates’ risk through new and innovative manners in order to ensure that there is maximum profitability for all. The motivation behind this project is to determine what factors contribute to the potential for a default on alone in order to save the financial institution the potential for financial losses. The goal is to identify these factors and create an automated process that will allow for financial institutions to make predictions about whether a candidate is a risk for default.

Using the available data, this project will determine if there is an ideal model for predicting risk in loans for financial services through logistical analysis and classical machine learning algorithms. The Joint Artificial Intelligence Center (JAIC) highlights that, “DOD financial communities are already incorporating various types of automation solutions for routine financial management tasks to streamline operations, reduce errors, and save time [3]. Automation through machine learning can provide the necessary analysis for financial services to make appropriate determinations for loan eligibility. The intention of this project is to determine whether someone will default on their loan based on common factors such as age, income, and the loan amount. The desired output would allow the financial institution to save money and potentially direct customers and clients to loan options that would be a little more accommodating to their current financial needs.

1. *Data Acquisition*

The data collection for this project was extremely important and in order to make the determination of how a financial institution can determine the strength of a loan candidate. This project utilizes the mindset from the Department of Defense when it comes to collecting data suitable for analysis. According to the DOD, “by enabling humans to supervise automated tasks, AI has the potential to reduce the number and costs of mistakes, increase throughput and agility, and promote the allocation of DoD resources to higher-value activities and emerging mission priorities” [4]. This project utilizes a data set found on kaggle.com that has various loan information from financial institutions that determine whether a loan went into default or was paid off. This output was determined from input factors such as the lone candidate's age, income, and the amount of money they were taking out for the loan. This data aligns with the DoD mindset stated earlier because this is information that would be accessible to the average analyst making that determination of whether to provide a loan to somebody.

1. *Data Understanding*

Understanding the composition of this data, will allow for the determination of the best model for both supervised learning and machine learning. In research published by the Institute for Defense Analysis, the authors disclose that “adapted natural language processing and machine-learning techniques to automate the initial data extraction and aggregation, which would have been unmanageable if approached manually” [5]. Determining what type of data is being used is essential in determining what type of analysis will be used in order to get the most efficient form of data extraction. This step can be accomplished through data visualization and examining the columns within the data set.

Table 1. Feature Description

|  |  |
| --- | --- |
| Input Variable | DESCRIPTION |
| clientid  income | Primary Key unique  Annual income of client |
| age | Age of client |
| loan | Loan amount |
| default | Could not pay  0 = paid loan  1 = couldn’t pay in time |

this data set had over 1000 separate data points of clients who took out loans from varying demographics. This allowed for the data to be cleanly divided up into a test data set and the training data set to ensure the accuracy of the model. The input variables, reading from left to right, include the annual income of the client, the age of the client when they took out the loan, and the loan amount. The output variable includes whether the client defaulted on that loan. The categorical zero or one answer from the output variable suggest that this data set would work well with logistic regression.

Upon initial examination of the data, observations can be made that can be further examined during the construction of the model. Logistic Regression is used to estimate the probability that an instance will fall in a certain class as a binary classifier of either 0 or 1 [6]. There are two classes within this data set. The client either defaults on their loan or they do not default on their loan. These classes are unequally represented because most loans were paid off and did not go into default. this is represented in the scatter plots provided that compared different input variables and their impacts on each other in terms of whether a client defaulted or not. Within the following scatterplots the blue dots represented loans that were paid off on time and the orange dots represent loans that went into default for one reason or another.

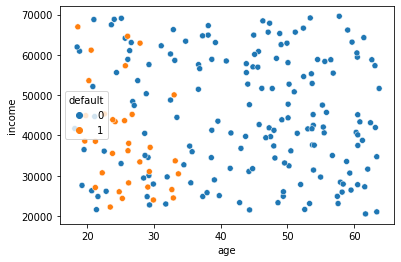


Fig. 1 Scatterplot depicting defaulted loans versus paid loans when comparing age and income.

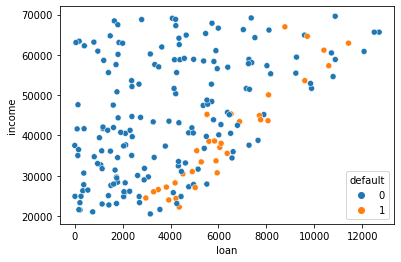


Fig. 2 Scatterplot depicting defaulted loans versus paid loans when comparing loan amount and income.

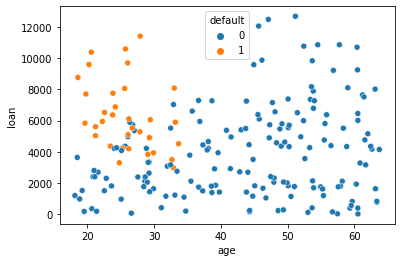


Fig. 2 Scatterplot depicting defaulted loans versus paid loans when comparing age and loan amount.

Based on initial examination, the data demonstrates a level of clustering between the different input variables. This data also seems like it would be ideal 4 unsupervised learning via a neural network. . according to a journal article, “Adaptive Resonance Theory (ART) neural networks uses unsupervised learning technique. The algorithm of the network operation consists in maintaining readiness to learn new patterns while preventing the rejection or modification of previously learned ones” [7]. Neural networks will allow for patterns and trends to be identified and create a model that will allow financial services Analysts to make the determination of whether alone would be appropriate for a particular candidate without having to do manual inputs themselves.

II. Method

This project utilized both the logistic regression, measuring various classification metrics, and neural network that trained the model and provided a measured precision and accuracy that will be examined in further detail.

1. *Data Preparation*

In order to extract the best data for analysis, various data preparation methods were undertaken. The first thing that was done involved removing all null values or blank spaces. Next all values that seemed unrealistic to a particular column were omitted from the data. This included a negative loan amount, in age under 18, and low income that would be interpreted as an outlier within the data. Finally, floor limits were set on certain data points including age at 18 years old, loans above $1000, and income above $10,000. All these steps for data preparation ensured that there was a level of uniformity among the data points so a majority assessment could be made.

1. *Metrics*

For logistic regression, several methods were used in order to determine the validity of the model. This includes precision, recall, and an F1 score. These metrics were instrumental in determining how many results produced true positives, false negatives, and the unbalance nature of the data, respectively. There will also be an evaluation oven LLR P-value which determined the certainty of our results. Lastly, an ROC curve will be used to demonstrate the performance of the logistic regression model at different classification thresholds. When examining the neural network modeling, accuracy precision and recall were vital metrics for comparison between the supervised learning approach and the neural network approach. The under the area curve (AUC) was used to determine the accuracy of the machine learning model.

1. *Classical Modeling*

After running the logistic regression modeling with various comparisons to different input variables. the determination was made that the model that was most effective for predicting whether a candidate would default involved utilizing a model net contained all input variables to include age, income, and loan amount. When the initial regression analysis results were determined for this model, 140 observations were taken as a training model. The log likelihood for this model was -15.302. this number is high, so it was be considered an initial indicator that this was a good fit model to work with. This model also had a log likelihood value that was higher than the log likelihood no value of -57.416. this indicates that the model will be useful to use. The final element of logistic regression results that will be discussed is the LLR P-value. This value was 3.789\*10-18. This number was significantly lower than alpa 0.0 5 and can reject the null hypothesis that the model is not statistically significant.

The final reason that this model was selected involved the P values for the individual input variables. All p values for the model registered at 0.0001. This is less than .05 alpha value which means that these input variables reject the null hypothesis and are therefore statistically significant to the model.

1. *Neural Network Modeling*

After normalizing the data, it was determined that a feature layer, along with two 5-neuron hidden layers would compose the neural network that would create the basis for this model. this model would take the list of metrics and run the regression values through a sigmoid function. This sigmoid function would serve as the activation function for the neural network because it corresponds with a logistic zero or one output. It was then determined that the data would be run over the course of 100 epochs in order to get an accurate assessment of accuracy and loss within the model. Accuracy, precision, and recall would be tracked over the course of these at box did determine what would be the strongest classification threshold value before potentially overfitting the data.

There were hyperparameters put in place after the initial model was run in order to determine if more efficiency could be added to the learning process. These parameters included a learning rate of 0.001 and a batch size of 100.

III. Analysis & Results

After statistical analysis was performed for both the classical model and the neural network model, the determination could be made of their strength. These strengths were measured in relation to the precision, recall, F1, accuracy, and area under curve.

1. *Classical Modeling*

The most effective model that was selected included the model that had all three variables of age, income, and loan amount.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model  No. Observations: 140 |  | | F1 Score | Precision | Recall |
| 0 | |  | 0.92 | 0.95 | 0.89 |
| 1 | |  | 0.76 | 0.69 | 0.85 |
| **Accuracy** | |  | .88 |  |  |
| **Weighted Avg** | |  | .89 | 0.90 | 0..88 |
| **LLR p-value: 3.798e-18** | |  |  |  |  |

Table 2. (example) Logistic Regression Model Results. Classification Report and Key Logistic Regression Results

According to the key measurements, the model had an accuracy score of .88, an overall weighted average F1 score of .89, A precision reading of .90, and a recall of .88. Individual breakdowns of these scores or respective classifications can be seen in the table above. In F1 score of .89 is very close to 1.0. This indicates that there is a high amount of accuracy with this model even with an unequal class bin distribution. A precision of .90 indicates that there is a lot of relevant and strong information meaning that there are many true positives in this model. A recall score of .88 means that there are few false negatives which is another strong indicator that this is a good model to work with.

Another strong metric is the result scene with the receiver operating characteristics curve. the AUC for this curve was 0.87. This is considered strong and this model can do an adequate job and determining positive and negative result rates. The high the high number indicates that this model is good at distinguishing between clients who will default in clients who will be able to pay off their loans.

Line chart

Description automatically generated with low confidence

Figure 4. ROC curve for Logistic Regression Model

The numbers found in the classification report table above were extracted from the confusion matrix that took the test set data ff 60 observations. In this data, 42 results for true positives, 5 results were false negatives, 2 results were false positives, and 11 results were true negatives

[42 5]

[ 2 11]

Table 3. Confusion Matrix

With 88% accuracy and strong precision and recall numbers, the determination can be made that this model is particularly strong for predicting whether a client will default on a loan based on the parameters of income age and loan amount.

1. *Neural Network Modeling*

After running the two-layer neural network with hyperparameters the following behavior was observed through 100 epochs. The network demonstrates that between 30 and 40 epochs, Accuracy and precision rose while recall had a minor fall. After 40 epochs, while accuracy climbed to high levels, the recall dropped to level that would deem the model unusable and precision leveled out better round 0.5. Based on this chart, it can be determined that the most effective neural network was produced at around 40 epics with a learning rate of about 0.68.

Chart, line chart

Description automatically generated

Figure 5. Accuracy, precision, and recall for Neural Network with hyperparameters through 100 epochs

This model accuracy of .68 is also verified in the ROC curve. the AUC for this neural network when comparing true positive rates to false positive rates was 0.686. This means that there is a 68.6% chance that the model will be able to distinguish between a true positive rate and a false positive rate.

Line chart

Description automatically generated with medium confidence

Figure 6. ROC curve for neural network

This is also confirmed in the metric chart that shows overfitting beginning around 40 epochs.

Chart, line chart

Description automatically generated

Figure 7. Metric Chart

1. *Model* *Evaluation*

Based on the criteria that was established earlier in this project, both the logistic regression model and the neural network model have demonstrated that they can adequately predict weather in client has the potential to default on a loan given by a financial services institution.

Both models had all three input variables of age, annual income, in loan amount in order to produce the output variable of weather in individual bold default on that loan. These models were chosen because they both have high levels of accuracy and precision.

The logistic model maintained an accuracy level of 88% while also having precision levels of 90% and a recall of 89%. This demonstrates that a strong model has been created because there is a much larger chance that the model will produce true positives rather than false negatives or false positives. While the neural network was not as strong it also demonstrated a level of reliability that could make it useful to the user. Analysis demonstrated that after 40 epochs, the neural network with hyperparameters produced the strongest model with about 68.6% accuracy that's still maintained a high level of recall and precision.

1. *Model Application (a.k.a. deployment)*

There are several practical uses for this model in the world of financial services. An analyst in the financial services sector could use this model to make that determination on whether somebody should be accepted for a loan based on how old they are, how much money they make a year, and how much they are planning to take out as a loan. These analysts can look at these potential warning signs and potentially tweak or modify certain variables for that client to allow them to safely receive a loan of a different amount or a potentially a different time in that client’s life.

A qualitative application would be to allow a certain client to maybe apply for a mortgage on a home. the financial services firm can determine what amount that person could realistically pay before they might get into financial trouble.

The quantitative applications for this model would be saving the financial services firm a large amount of money and managing risk for the bank. This application would also allow for the client to save their own money from potential financial hardship and allow them to put themselves in the best financial position to succeed.

IV. Conclusion

In conclusion, this project was created to determine a solution for financial services firms that would allow them to decide on clients on whether they would be a potential risk for defaulting on a loan. This project took a large data set that contained loaned information for several clients from unknown banking locations. these loans provided information such as the client's age, the amount of money the client had for an income, in the loan amount that was taken out. The final column was the output variable which demonstrated whether that client defaulted on their loan or was able to pay it off in the determined manner. The logistic model that was created demonstrated the ability to make predictions based on these three input variables with an 88% accuracy level. This information was also developed into a neural network that demonstrated to have high levels of accuracy with hyperparameters in multi-level neural networks. This neural network demonstrated an accuracy of 68.6%.

While these models have demonstrated adequate abilities to make predictions for within the constraints of this project, there is also the need for improvement. This is because when working in financial services and money is involved there can't be a 12% discrepancy in the accuracy of what data is trying to tell you. Future suggestions I would make for further study would include tweaking the hyperparameters in the neural network As well as trying to find potential new variables that could add further inside to the data.

### References

|  |  |
| --- | --- |
| [1] | E. Hida and L. Julian, "The Future of Risk in Financial Services," Deliotte, United Kingdom, 2017. |
| [2] | G. Phalen, "Collateralized borrowing and increasing risk," *Economic Theory,* vol. 63, no. 2, pp. 471-502, 2017. |
| [3] | JAIC Public Affairs, "AI in Defense: DOD's Artificial Intelligence Blog," JAIC, October 2020. [Online]. Available: https://www.ai.mil/blog\_10\_01\_20-jaic\_partners\_with\_diu\_on\_aiml\_models\_to\_resolve\_complex\_finanical\_errors.html. . [Accessed 10 November 2021]. |
| [4] | Department of Defense, "Saummary of the 2018 Department of Defense Artificial Intelligence Strategy," Department of Defense, 2018. [Online]. |
| [5] | L. Odell, "Data Exploration and Management of Defense Finance and Accounting Artifacts," Defense Technical Information Center, 1 January 2018. [Online]. Available: https://apps.dtic.mil/sti/citation/AD1122566. [Accessed 10 November 2021]. |
| [6] | A. Geron, Hands-On Machine Learning with SciKit-Learn, Keras, and Tensorflow, Sebastpol: O'Riley, 2019. |
| [7] | M. Briston and Z. Piotrowski, "Artificial Intelligence Applications in Military Systems and their Influence on Sense of Security of Citizens," Electronics, 6 April 2021. [Online]. Available: https://doi.org/10.3390/electronics10070871. [Accessed 10 November 2021]. |
|  |  |
|  |  |