Mortgage Approval Prediction Using Neural Networks

Bryce Hinkley
UTSA Computer Science
One UTSA Circle, San Antonio,TX 78249
bryce.hinkley@my.utsa.edu

Alex Van Nimwegen
UTSA Computer Science
One UTSA Circle, San Antonio,TX 78249
alex.vannimwegen@my.utsa.edu

***Abstract*–Applying for a mortgage can be a lengthy and at times costly process, which creates a need for a tool to self-evaluate before making an application. Our project uses deep learning in order to predict the likelihood of an individual being accepted for a loan. The methodology will be to use a neural network in order to train on a large quantity of information from the HDMA 2017 financial dataset to achieve an accurate model. The model will utilize both individual information and the macro-information of the region of the applicant due to the large effect that macroeconomic conditions have on the ability to pay off a loan. In order to make the predictions, a deep neural network will be used from the features in the dataset. We can both deem the possibility of making such predictions as well as confirm the role of using macroeconomic data in models. Ethical considerations will be made with demographic data to reflect both the true probability of an applicant paying off a loan, as well as systemic biases that affect approval rates based on demographics.**

***Keywords*–Deep neural networks, loan prediction, macroeconomic**

1. Introduction

For many families, securing a mortgage loan is a significant milestone, and the process can be daunting, especially for those with limited experience or a tight budget. Our objective is to alleviate this stress by designing a tool that helps provide the knowledge necessary so that anybody can apply for a loan and make informed decisions about their financial future. Recent research indicates that the most efficient approach to forecasting mortgages involves the use of a convolutional neural network, along with the integration of both micro and macroeconomic data. Our model aims to generate precise and applicable mortgage predictions while also mitigating systemic biases. The objective of our research is to construct a precise model that can be subsequently employed to evaluate the impact of biases in a particular region on mortgage loans. By incorporating socio-economic information into our loan prediction model, and subsequently re-evaluating the model's performance when such information is omitted, we aim to assess any potential biases that may exist within specific areas of the model.

2. Related Works

The general purpose of our network will be to conduct predictions based on user information. The values to be returned will be the percentages between whether or not the applicant will be approved. In order to decide upon the type of network to use, other research papers suggest the usage of a CNN for our type of problem. They argue that CNNs' speed, early predictive power, and robustness should pave the way for an application in process outcome prediction. They came to this conclusion after comparing both CNNs and Attention-LSTMs neural networks [2]. Research for predicting if an individual should receive a loan with the usage of neural networks also offers insight into how we could use other algorithms in addition to the CNN to make accurate predictions. Other research suggests using the CNN to extract the local spatial features hidden in the results of the Stacking algorithm in order to improve the generalization ability of the model. The researchers used the wrapper method and variance inflation coefficient (VIF) to extract the features from the original data. That data was processed by the Stacking algorithm, then the CNN, and subsequently a Support Vector Machine to make predictions [3]. They found that this gave better results than using the CNN alone. In order to decide on which type of data to use we consulted current research. Some researchers were able to make accurate predictions on loan defaulting by only using the balances of the checking account, savings account, and the credit card, in addition to the daily number of transactions on the checking account, and amount transferred into the checking account [4]. From this they were able to achieve approximately 91% accuracy from their model. Other researchers used both macroeconomic and microeconomic information in order to determine loan risk. They found that macroeconomic information had a significant impact on long-term loans, such as a mortgage, which would suggest that including macroeconomic information would be relevant for our model [1].

3. Proposed Approach

Our approach to creating our neural network model will be to acquire public consumer information that can be used by our model. The data will require a value to use as a target value by our neural network. In this instance, since we are working with predicting the approval of a loan, it would be a rejection and approval value for the loan. This value would be used in order to classify the information within a deep neural network (DNN). Our model infrastructure will be a DNN that will use normalized data from a dataset. Using other preprocessing algorithms, such as the Stacking algorithm in order to improve generalization, as well as the usage of an Attention-DNN to weigh important features more will be included in our experimentation. This approach would give a reasonably accurate tool that would help individuals self-evaluate their likelihood of getting a loan with a fast model that yields percentage-based output for more insight. Our approach differs from other related research papers because the predictions will be made on whether an institution will grant a client a loan. The related research cited attempts to predict if a client will be unable to pay a loan or not, rather than if they will be given a loan.

4. Experiments

4.1 Experiment Setup

At the time of writing, it does not appear that there have been publications on creating a neural network for predicting if someone will be approved or not, and accordingly, there isn't any data that can be used for a direct comparison. However, research on determining the risk involved in giving out mortgage loans gives us another experiment setup to compare with. The dataset used in other research was publicly accessible loan information that features geographic information, such as unemployment, and individual information, such as FICO scores and income [1]. The researchers cleaned the data by removing incorrectly filed information, and entries without critical information, and encoded an additional information indicator for entries with extra information that was non-critical. For measuring performance they used a receiver operating characteristic curve (ROC) and area under the ROC curve (AUC) as a standard measure of predictive accuracy for a binary classifier.

4.2 Dataset

The dataset that will be used for our experiment will be the Home Mortgage Disclosure Act (HMDA) 2017 dataset. The dataset contains all records available, is on a nationwide scope for the United States of America, and uses both plain language labels as well as HDMA codes. The information does contain an approval or rejection value to be used for classifying information. The dataset is available for usage publicly and can be acquired from the consumer finance government website. This dataset has approximately 14 million rows of information with 45 columns, so it has sufficient data for training and testing [5]. The data is skewed with more approved loans than rejected loans with approximately 7 million approvals to 2 million rejections.

The data will be cleaned and features that are populated largely by null values will be dropped prior to measuring any correlations or dropping irrelevant data such as ID numbers. The data will also be put through an undersampling process to balance the data to be split perfectly even between both classes to balance the data to prove the effectiveness of the approach.

After this the data we applied categorical and numerical encoding to each feature in the dataset and applied normalization to all of the numerical values in order to improve generalization and model performance.

We will also compare the performance of models that do not include any macroeconomic data in order to evaluate the performance changes from including this information.

4.3 Metrics

Standard model metrics will be provided such as the validation accuracy, the precision, recall, and F1 score. These can be defined as seen below,

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TN}{TN+FN} \quad (3)$$

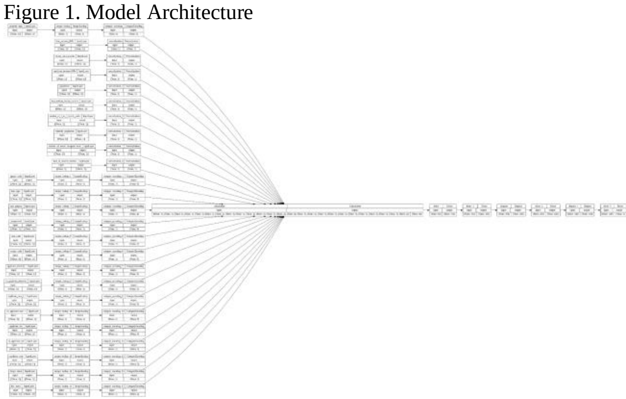
$$F1\,Score = \frac{2 * Recall * Precision}{Recall + Precision} \quad (4)$$

In order to measure the effectiveness of the model, we will use the ROC and AUC metrics in order to measure the broad scale of true and false detections for a binary classifier. We will also use categorical accuracy for the neural network training. Categorical accuracy is used in order to see if the major hot value matches the true value or not.

4.4 Model Architecture

This architecture as seen in figure 1 shows a model that takes in both numerical and categorical features as inputs and concatenates them into a single layer. The encoded categorical features are typically transformed using one-hot encoding or embedding layers to convert them into a numerical representation.

Figure 1. Model Architecture



The concatenated layer is then passed through several dense layers with different activation functions. The first dense layer has 512 units and uses a rectified linear unit (ReLU) activation function, while the second dense layer has 256 units and uses a hyperbolic tangent (tanh) activation function. These activation functions introduce nonlinearity into the model and help to capture complex patterns in the data.

To prevent overfitting, two dropout layers are added after the first and third dense layers. Dropout randomly sets a fraction of the input units to zero during training, forcing the model to learn more robust representations of the data.

After the final dropout layer, another dense layer is added with 128 units and a ReLU activation function. This layer further refines the representations learned by the earlier layers.

Finally, an output layer with a single unit and a linear activation function is added. This output layer produces a continuous output, which makes it suitable for regression problems where the goal is to predict a numerical value.

Overall, this architecture is a feedforward neural network that can handle both numerical and categorical features and includes techniques like dropout to prevent overfitting. It can be trained using stochastic gradient descent and backpropagation to minimize a chosen loss function.

4.5 Training

The model is then trained on the testing data for 20 epochs and uses a batch size of 50. The batch size chosen was selected in order to have a moderate level of generalizability, performance, and a smoother convergence.

5. Results

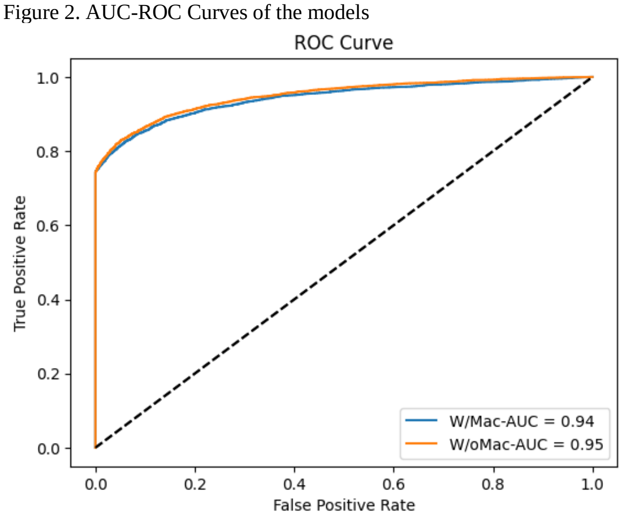
In table 1 the results of the metrics used can be seen for the model. The accuracy is lower compared to other research that yielded 91% accuracy using an optimized model but without macro information data [1]. This shows no accuracy changes from macroeconomic information being included when combined with microeconomic information when making predictions for loan candidacy. However, in model 1 notice that the recall and precision scores change by roughly 6 to 8% depending on whether or not macroeconomic information was included in the model.

Table 1. Model Metric Results

Dataset	Validation Accuracy	Precision	Recall	F1 Score
With macro data	87.9%	98.0%	77.1%	86.4%
Without macro data	88.8%	92.5%	84.3%	88.2%

Seen in figure 2 are the AUC-ROC curves of the two models. The two AUC-ROC values are almost the same. This means that the addition of macroeconomic data did not significantly improve the model's classification performance.

Figure 2. AUC-ROC Curves of the models



This could happen if the macroeconomic data did not provide much additional information or if the neural network model was already able to capture the relevant features from the dataset without the macroeconomic data. Alternatively, it is also possible that the dataset is not large enough to show a noticeable difference between the two models.

6. Conclusion

In our future research, we aim to enhance the model complexity to increase the information gain, thereby leading to higher levels of accuracy. Another objective is to develop a novel model that eliminates demographic information in order to mitigate any potential biases while maintaining our primary goal of accurately predicting mortgage loan approvals. We plan to explore various feature selection methods and model architectures that can further enhance the model's predictive power and robustness. These efforts would contribute to the development of a more accurate and unbiased model and hopefully result in simulating fair and transparent lending practices.

In the end, our study successfully developed a neural network model for predicting mortgage loan approvals, demonstrating the effectiveness of machine learning methods in enhancing lending practices. Through our analysis, we identified that different types of economic information contribute to varying degrees, with both micro and macro economic data playing a significant role in improving the model's predictive power. Despite achieving accurate predictions, our study did not incorporate a strategy to mitigate systemic biases in the data, which remains a crucial area of concern for fair lending practices. Overall, our study contributes to the ongoing efforts towards improving the accuracy and equity of mortgage loan approval processes.

References

[1] Sirignano, J., Sadhwani, A., & Giesecke, K. (2016). Deep learning for mortgage risk. arXiv preprint arXiv:1607.02470.

[2] Weytjens, Hans, and Jochen De Weerdt. "Process outcome prediction: CNN vs. LSTM (with attention)." Business Process Management Workshops: BPM 2020 International Workshops, Seville, Spain, September 13–18, 2020, Revised Selected Papers 18. Springer International Publishing, 2020.

[3] Li, Meixuan, Chun Yan, and Wei Liu. "The network loan risk prediction model based on Convolutional neural network and Stacking fusion model." Applied Soft Computing 113 (2021): 107961.

[4] Kvamme, Håvard, et al. "Predicting mortgage default using convolutional neural networks." Expert Systems with Applications 102 (2018): 207-217.

[5] "Download HMDA data | Consumer Financial Protection Bureau," Consumer Financial Protection Bureau. https://www.consumerfinance.gov/data-research/hmda/historic-data/