Predicting who will get a home mortgage in New York - CS230, Winter 2018

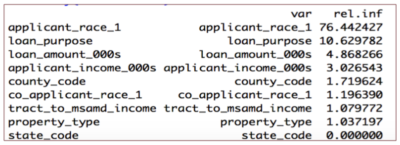
Manjinder Kandola, Aris Konstantinidis, Dasmer Singh

Problem

The Home Mortgage Disclosure Act (HMDA), a US federal law enacted by the Congress in 1975, requires that financial institutions publish data on home mortgage applications to ensure transparency and adequate home financing in all geographic areas. We built a binary classification model that receives as input an individual home mortgage application entry and predicts the outcome of the application with an 80% accuracy.

Model - Features

In order to create the neural network, we used Keras. We started with 140,000 rows each of which have 6 features. Our neural network has 3 hidden layers (as shown in the figure). The first one has 100 nodes, the second has 50 nodes, and the third one has 10 nodes. We came to these number of nodes after rigorous trial and error and based it CNN dimension reduction. Before inputting the data, we normalized it so that the model would converge faster. Parameters were randomly initialized and we used Stochastic Gradient Descent decay with momentum of 0.9. We tried various activation functions and eventually used tanh for all hidden layers because it converged fastest with the best accuracy on test set. We used a sigmoid output for our final response.



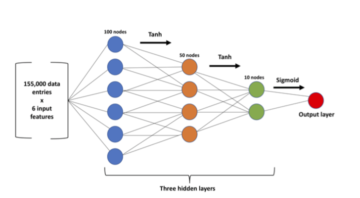
```
                                        var   rel.inf
applicant_race_1           applicant_race_1 76.442427
loan_purpose                   loan_purpose 10.629782
loan_amount_000s           loan_amount_000s  4.868266
applicant_income_000s applicant_income_000s  3.026543
county_code                     county_code  1.719624
co_applicant_race_1     co_applicant_race_1  1.196390
tract_to_msamd_income tract_to_msamd_income  1.079772
property_type                 property_type  1.037197
state_code                       state_code  0.000000
```

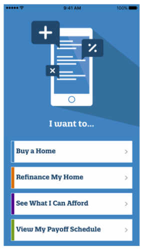
Table 1. List of variables used in building the GBM model with their respective weights

Dataset

The dataset that we used is publicly available on consumerfinance.gov. It consists of 439,654 home mortgage applications that were submitted in New York in 2015. Each row represents an individual application, while columns contain 78 variables. The variables can be grouped into the following main categories: Lender / Institution, Applicant, Loan, Property. We are confident that the data is highly accurate, since it contains all the originally submitted applications without any editing. We split the data as follows: 87% train set, 6.5% dev set, 6.5% test set). The input features that we used are: race, co-applicant race, loan purpose, property type, state code and loan amount.

Network Architecture





	Neural Network	Log	GBM
Training Accuracy	80%	78%	79%
Test Accuracy	74%	72%	72%

Table 2. Summary of accuracy results achieved by each model

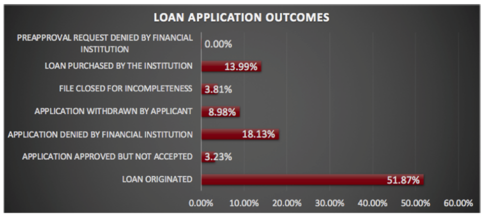


Chart 1. Loan application outcomes

Results

Our model predicted with an accuracy of 80% whether a loan application is going to be approved or not. We iterated our model across various hyperparameters and network structures, including choice of activation function (leaky ReLU vs tanh), # of layers/nodes, choice of cost function, etc. Ultimately, the model as shown in the figure resulted in the highest test accuracy.

Discussion

Although our model reached an acceptable accuracy, it can definitely be improved in the future. Our next steps will include more hyper parameter tuning (dropout probability, learning rate, number of layers, etc.), to find the best combination. Also, we will try to use one-hot encoding instead of real valued numbers in some of our features and see if our model's accuracy will improve. Finally, we might need to get more data for two reasons. Firstly, because the amount of data we used might not be sufficient to reach a higher accuracy in our predictions. Secondly, to test how "unfair" our algorithm is (e.g. whether it discriminates based on race, which was the variable with the highest weight in our model.

Contact

Manjinder Kandola
Stanford Graduate School of Business
Email: mkandola@stanford.edu
Phone: (646) 229 2719

References

1. https://www.consumerfinance.gov/data-research/hmda/explore
2. https://www.kaggle.com/ambarish/eda-home-mortgage-ny-with-feature-analysis
3. https://www.nytimes.com/2017/08/24/upshot/how-redlinings-racist-effects-lasted-for-decades.html
4. https://www.ffiec.gov/hmda/