Machine Learning in Home Equity Risk Management: Unbanked Population Credit Assessment

Yitian Zhang[1,*], Parsa Moghaddamcharkari[2,a]

[1,2]University of Toronto, Ontario, Canada

*Corresponding author: yitian.zhang@mail.utoronto.ca
aparsa.moghaddamcharkari@mail.utoronto.ca

Abstract. This study leverages an imbalanced dataset provided by a home equity company to assess unbanked population's repayment ability. The target variable is whether the client has repayment difficulties, and independent variables include demographic information and credit history. Logistic regression model and other machine learning models are constructed for comparison. It is found that the neural network model has the best overall performance. Also, clients who are reachable by phone, or have been employed for a longer period in the past are less likely to have repayment difficulties. On the other hand, older clients or whose permanent address does not match their contact address or highest education attended is secondary education would have a higher probability of having repayment difficulties.

Keywords: Credit Risk Assessment, Logistic Regression, Random Forest, Gradient Boosted Tree, Neural Network

1 Introduction

Many people struggle to get loans due to insufficient or non-existent credit histories. This study helps financial institutions improve assessment of credit quality of unbanked population to increase market efficiency. Traditional econometrics models have limitations in processing large, imbalanced, and high dimensional data sets. In this case, the predictive performance of logistic regression and other machine learning models including random forest, gradient boosting trees, and neural networks were compared to find the most suitable model for credit quality assessment leveraging imbalanced dataset. Based on importance sore, Wald Statistics, and partial dependence plots, the key factors affecting repayment ability were identified. This study seeks to provide insights for financial institutes to improve assessment of unbanked population's creditworthiness by implementing suitable models and identifying key factors.

G. Guan et al. (eds.), *Proceedings of the 2023 3rd International Conference on Education, Information Management and Service Science (EIMSS 2023)*, Atlantis Highlights in Computer Sciences 16,
https://doi.org/10.2991/978-94-6463-264-4_58

2 Literature Review

There is literature discussing assessment of borrowers' creditworthiness leveraging different techniques. A paper investigated the usage of Convolution Neural Network on Personal Credit Default data to increase predictive performance [1]. Similarly, another research on small and medium enterprises credit risk problem in Turkey leverages

MLP, SVM and KNN classifiers [2]. Combining different machine learning techniques such as clustering and classification models into a hybrid model could also improve characterizing credit behavioral [3]. Comparing different models is one of our interests. A paper discussing the prediction accuracy of logistic regression, SVM, random forest, and neural network on forecasting credit risk concludes that random forest has the best predictive power overall [4]. Another paper argues that using information from decision trees in logistic regression could improve the performance while maintaining interpretability [5]. Data processing is an important component in modeling. A study on optimizing credit banking risk evaluation for home equity loans shows how feature selection improves model performance and computational efficiency when using advanced credit scoring methods [6]. Another study states that proper data mining improves the performance for both traditional econometric model such as logistic regression and machine learning model such as decision trees [7].

3 Methodology

3.1 Data Settings, Cleaning, and Feature Engineering

We leveraged the dataset provided by a company called Home Equity, which provides loans to unbanked populations. The data source contains seven sub-datasets, which contain the target variable - whether the client has repayment difficulties (1 refers to having difficulties, 0 otherwise), historical transactions of repayments and loans, and demographic information. The dataset with duplicated features from other datasets was dropped and the remaining six datasets were merged by client's ID and credit history transaction ID. After merging, some summary statistics such as number of missing values, average, standard deviation were calculated. Constant variables and variables with over 50 % missing values are removed. Missing values for continuous variables were filled with mean, and mode was used to fill in the missing values for categorical variables. Also, a dummy was added for each variable with missing value as a missing flag. The final cleaned dataset contains 102 features across 307,511 data points. Continuous variables with significant outliers were binned into categorical variables using K-means, and all categorical variables were converted to dummies. Lastly, clustering was applied to reduce the number of features. Eventually, there were 60 features remaining for further processing.

3.2 Model Construction

Logistic regression, random forest, gradient boosted tree, extreme gradient boosted tree, and neural network models were constructed following the below procedures:

- Random data splitting into training (75%), and testing sets (25%).
- Under-sampling the training set by removing examples from the training dataset in the majority class.
- Model initial fitting on training set and testing set.
- Hyperparameters tuning using techniques such as cross-validation or grid search to select the best set of hyperparameters.
- Refitting model on training set and testing set using selected hyperparameters.

3.3 Statistical Analysis

In-sample and out-of-sample Performance metrics such as RMSE, MAE, Accuracy, Precision, Recall, AUC- ROC rates, as well as F1-score were calculated to compare model performance. Also, the importance score, Wald Statistics, and partial dependence plots were examined to identify the most important factors and their impact on the dependent variable.

4 Results and Analysis

The regression results for logistic model is summarized in table 1. The estimated coefficients for logistic regression represent the estimated average change in the log-odds of the target variable for a one-unit change in the corresponding predictor variable, while holding all other predictor variables constant. For example， -0.255 for REGION_RATING_CLI means holding all other variables constant, if the rating of the region where the client lives increase by 1 unit, then the log-odds of the client having repayment difficulties is estimated to decreases 0.255 on average. Which means if the client lives in a higher rating region, then the probability of the client having repayment difficulty will decrease.

The in-sample and out-of-sample performance metrics were summarized in table 2 and 3.

Table 1. Logistic Regression Output

Variable Name	Estimated Coef.	Z-statistics	P> |z|	[0.025	0.975]
(const)	-0.6819*** (0.050)	-13.512	0.000	-0.781	-0.583
REGION_RATING_CLIENT	-0.2556*** (0.042)	-6.106	0.000	-0.338	-0.174
FLAG_CONT_MOBILE	-0.4558*** (0.011)	-41.603	0.000	-0.477	-0.434
AMT_CREDIT_SUM_LIMIT_(20.16, 23563.475]	0.1251*** (0.033)	3.840	0.000	0.061	0.189

AMT_RECIVABLE_(62175.128, 913096.639]	0.4547*** (0.034)	13.230	0.000	0.387	0.522
REG_CITY_NOT_LIVE_CITY	0.1213*** (0.011)	11.175	0.000	0.100	0.143
REG_REGION_NOT_LIVE_REGION	0.1695*** (0.013)	13.383	0.000	0.145	0.194
AMT_CREDIT_SUM_DEBT_(-6981558.211, 0.0]	-0.2918*** (0.035)	-8.240	0.000	-0.361	-0.222
NAME_FAMILY_STATUS_Married	-0.1646*** (0.023)	-7.044	0.000	-0.210	-0.119
AMT_CREDIT_MAX_OVERDUE_(-0.001, 861.375]	-0.2848*** (0.028)	-10.108	0.000	-0.340	-0.230
DAYS_BIRTH	0.9088*** (0.050)	18.142	0.000	0.811	1.007
AMT_GOODS_PRICE	-0.3564*** (0.134)	-2.656	0.008	-0.619	-0.093
AMT_CREDIT_SUM_(-0.001, 142204.5]	0.2961*** (0.040)	7.369	0.000	0.217	0.375
NAME_EDUCATION_TYPE_Secondary / secondary special	0.3588*** (0.026)	13.589	0.000	0.307	0.411
AMT_CREDIT_SUM_DEBT_(717779.52, 1417764.15]	0.1861*** (0.037)	4.976	0.000	0.113	0.259
DAYS_EMPLOYED_(-458.0, -144.0]	0.3386*** (0.036)	9.311	0.000	0.267	0.410
DAYS_EMPLOYED_(-1695.0, -1213.0]	0.1962*** (0.037)	5.343	0.000	0.124	0.268
DAYS_EMPLOYED_(-822.0, -458.0]	0.3214*** (0.036)	8.826	0.000	0.250	0.393
AMT_CREDIT_SUM_DEBT_(36405.0, 133663.5]	-0.2768*** (0.040)	-6.959	0.000	-0.355	-0.199
application_train_flag_SK_DPD	-0.1404*** (0.050)	-2.808	0.005	-0.238	-0.042

Note: *p<0.1; **p<0.05; ***p<0.01

Table 2. In Sample Performance Metrics

Model	Hyperparameter Tuning	RMSE	MAE	Accuracy	Precision	Recall	F1-Score	ROC AUC
Logistic	Before	0.4668	0.4358	0.6498	0.6542	0.6355	0.6447	0.6498
	After	0.4669	0.4361	0.6639	0.1404	0.6153	0.2286	0.6494
Gradient Boosted Tree	Before	0.5902	0.3484	0.6516	0.6551	0.6405	0.6477	0.6516
	After	0.5608	0.3145	0.6855	0.6875	0.6563	0.2458	0.6855
XG Boosted Tree	Before	0.5345	0.2857	0.7143	0.7159	0.7105	0.7132	0.7143
	After	0.5571	0.3104	0.6896	0.6916	0.6844	0.6880	0.6896
Random Forest	Before	0.0259	0.0007	0.9992	0.9993	0.9995	0.9993	0.9993
	After	0.4763	0.2269	0.7752	0.7731	0.7695	0.7723	0.7731
Neural Network	Initial Model	0.4312	0.3709	0.7208	0.7223	0.7212	0.7234	0.8042
	Initial Model + Callback	0.4813	0.4334	0.6201	0.7121	0.3943	0.5002	0.7412
	Dropout	0.4432	0.4021	0.7112	0.7109	0.7105	0.7120	0.7903
	Dropout + Callback	0.4921	0.4511	0.6245	0.7504	0.3545	0.4722	0.7211
	Batch Normalization	0.4321	0.3843	0.7266	0.7267	0.7145	0.7264	0.7901
	Batch Normalization +	0.4712	0.4344	0.6509	0.7632	0.4511	0.5704	0.7412

	Callback							
	Dropout + L2 Reg	0.4411	0.3945	0.7194	0.7132	0.7312	0.7234	0.7811

Table 3. Out-of-sample Performance Metrics

Model	Hyperparameter Tuning	RMSE	MAE	Accuracy	Precision	Recall	F1-Score	ROC AUC
Logistic	Before	0.4652	0.4353	0.6643	0.1405	0.6148	0.2287	0.6498
	After	0.4654	0.4356	0.6643	0.1405	0.6148	0.2287	0.6498
Gradient Boosted Tree	Before	0.5842	0.3413	0.6587	0.1399	0.6249	0.2286	0.6433
	After	0.5710	0.3260	0.6740	0.1512	0.6563	0.2458	0.6659
XG Boosted Tree	Before	0.5764	0.3323	0.6677	0.1481	0.6532	0.2414	0.6611
	After	0.5707	0.3257	0.6743	0.1514	0.6568	0.2461	0.6663
Random Forest	Before	0.5989	0.3586	0.6414	0.1353	0.6362	0.2231	0.6390
	After	0.5809	0.3375	0.6625	0.1457	0.6516	0.2382	0.6576
Neural Network	Initial Model	0.4913	0.4367	0.6323	0.6489	0.6223	0.6356	0.6833
	Initial Model + Callback	0.4855	0.4317	0.6245	0.7190	0.3956	0.5091	0.7098
	Dropout	0.4767	0.4391	0.6557	0.6312	0.6368	0.6345	0.732
	Dropout + Callback	0.4811	0.4409	0.6107	0.7563	0.3327	0.4615	0.7134
	Batch Normalization	0.4717	0.4219	0.6446	0.6565	0.6234	0.6423	0.6976
	Batch Normalization + Callback	0.4845	0.4366	0.6374	0.7223	0.4267	0.5332	0.7112
	Dropout + L2 Reg	0.4734	0.4334	0.6558	0.6597	0.6620	0.6506	0.7034

It is found that Random Forest has the highest in-sample predictive performance metrics (Accuracy, Pre- cision, Recall, F1-Score, ROC AUC) and the lowest error metrics (RMSE, MAE) before hyperparameter tuning. Meanwhile its predictive performance metrics decreased significantly (especially for Precision and F1-score) when fitted on testing set. There was a similar trend in all other models that hyperparameter tun- ing brings in-sample and out-of-sample performance closer. Neural network has the most stable performance between training and testing set and produces the highest Precision and F1-Score on the testing sample. Additionally, 'Dropout + L2

Regularization' combination gave the highest predictive performance and the lowest error compared with other combinations of settings for neural network. Lastly, all models other than neural network overestimate the probability of client's having repayment difficulties as those models have a relatively low precision score.

The top 10 variables with the highest importance score (Wald statistics for logistic regression model) were compared across all five models. The following five variables in table 4 were found to have the highest overall importance among all models' results.

Table 4. Variables Appearing Most Frequently in the Top 10

Variable Name	Description
FLAG_CONT_MOBILE	Was mobile phone reachable (1=YES, 0=NO)
DAYS_BIRTH	Client's age in days at the time of application

REG_REGION_NOT_LIVE_REGION	Flag if client's permanent address does not match contact address (1=different, 0=same, at region level)
NAME_EDUCATION_TYPE_Secondary/secondary special	Whether the highest education the client achieved is secondary (1=YES, 0=NO)
DAYS_EMPLOYED	How many days before the application the person started current employment

Lastly, partial dependence plots show that FLAG_CONT_MOBILE and DAYS_EMPLOYED have a negative relationship with probability of clients' having repayment difficulties, while

NAME_EDUCATION_TYPE_Secondary/secondary special, DAYS_BIRTH, and REG_REGION_NOT_LIVE_RE have a positive relationship with probability of clients' having repayment difficulties.

5 Discussion and Conclusion

Given the performance metrics of the five models, it is found that the neural network model has the most consistent and best generalized performance in assessing credit quality assessment. Neural networks have higher flexibility and better ability to learn complex patterns and trends due to the architectural structure. Networks can map nonlinear relationships among data whereas some other more traditional techniques such as logistic regression are incapable to do so. Furthermore, networks can extract meaningful results from complex data. This means that the model could extract relevant features which is very important when working with complicated and high dimensional data [8]. This is mainly contributed by network models' advantage of extracting features autonomously whereas traditional techniques process feature selection manually [1].

Again, we know that if the client is reachable by phone or has been employed for a longer period in the past, it's less likely for the client to have repayment difficulties. On the other hand, if the client's is older or the client's permanent address does not match contact address or the highest education attended is secondary, then the client has a higher probability of having repayment difficulties. This is partially consistent with the logistic regression results. Mobile phones, as the primary means of communication, are closely intertwined with daily life. In this case, if someone cannot be reached by phone, it is likely they purposely avoid answering debt collecting calls from loan companies as they might already have repayment difficulties. A study demonstrated that mobile phone usage data can be used to make predictions and find the best classification method for credit scoring even if the dataset is Small [9]. Similarly, some individuals may intentionally provide false addresses to prevent lending companies from physically locating and causing trouble for them due to debt collection. Also, if the client is older, then there might be a probability of existing credit issues as the older population usually have a rich credit history and are not unbanked. It might be the case that the older client already has low credit quality and cannot access bank backed loans, thus, they could only look for a company providing low quality home equity loans. Also, clients who work for a longer time in the past prove their working ability and repayment ability using salary. Lastly, limited

education can also restrict job skills and income, thereby affecting the ability to repay debts.

Based on the previous findings, we suggest home equity companies leverage neural networks when facing imbalanced data of unbanked population. Also, there are some key variables to focus on when gathering client's information:

- Is the client reachable by phone?
- Does the client's contact address match
- permanent address?
- Client's age, highest education achieved
- and historical employment years.

To sum up, our study leveraged datasets with high dimensionality of features and rich data points. At the same time, the main limitation of our study is that the dataset is significantly imbalanced. Even though we used under-sampling, the results could still be biased. The subsequent course of action involves exploring alternative datasets that offer more comprehensive, representative features and balanced data.

References

1. Zhou, X., Zhang, W., & Jiang, Y. (2020). *Personal credit default prediction model based on Convolution Neural Network.* Mathematical Problems in Engineering, 2020, 1–10. https://doi.org/10.1155/2020/560839 2
2. Derelioğlu, G. & Gü rgen, F.(2011). *Knowledge discovery using neural approach for SME's Credit Risk Analysis Problem in Turkey*. Expert Systems with Applications, 38(8), 9313–9318.https://doi.org/10.1016/j.eswa.2011.01012
3. Wallis, M., Kumar, K., & Gepp, A. (2022). *Credit rating forecasting using Machine Learning Tech- niques.*
Research Anthology on Machine Learning Techniques, Methods, and Applications, 734–752. https://doi.org/10.4018/978-1-6684-6 291-1.ch039
4. Shiv, S. J., Murthy, S., & Challuru, K. (2018). *Credit risk analysis using machine learn- ing techniques*. 2018 Fourteenth International Conference on Information Processing (ICINPRO). https://doi.org/10.1109/icinpro43533. 2018.9096854
5. Dumitrescu, E., Hu´e, S., Hurlin, C., & Tokpavi, S. (2022). *Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects*. European Journal of Operational Research, 297(3), 1178–1192. https://doi.org/10.1016/j.ejor.2021.0 6.053
6. P´erez-Mart´ın, A., P´erez-Torregrosa, A., Rabasa, A., & Vaca, M. (2021). *Feature selection to optimize credit banking risk evaluation decisions for the example of home equity loans*. Prime Archives in Applied Mathematics.
https://doi.org/10.37247/paam2ed.2. 2021.6
7. Yap, B. W., Ong, S. H., & Husain, N.
H. (2011). *Using data mining to improve assessment of creditwothiness via credit scoring models*. Expert Systems with Applications, 38(10), 13274–13283. https://doi.org/10.1016/j.eswa.2011.0 4.147
8. Egwu, N., Mrziglod, T.,& Schuppert, A. (2022). *Neural network input feature selection using structured L2 − Norm Penalization. Applied Intelligence.*

https://doi.org/10.1007/s10489-022-03539-8
9. Shema, A. (2019). *Effective credit scoring using limited mobile phone data.* Proceedings of the Tenth International Conference on Information and Communication Technologies and Development, 2019, 1–10. https://doi.org/10.1145/3287098.328 7116