Deliquency Prediction Using ML

{ Mentor: Mark Williams, Mentee: Mengjia Tan, Chuyi Wang} BU MSMFT Class of 2024

Objectives

We develop a model to predict consumer default based on machine learning. Our model predicts favorable creditworthiness for individuals with subpar FICO scores. We believe such a model would be helpful in the bank's policies design to reduce customer default and accelerate the credit scoring efficiency. Below are some main challenges in this model

1. Data resource
2. Feature Selection
3. Machine Learning models

Introduction

Despite the ubiquitous use of credit scores in everyday financial transactions, little is known about the performance of these models in assessing default risk at the individual level. This model proposes a novel approach to predicting consumer default based on 10 ML models (such as XGB, Logistic Regression, Random Forest, Extra Tree,etc) with the best AUC score. Our methodology outperforms traditional regression as our model relies on machine learning, which could account for complex non-linear patterns of interaction among variables affecting the outcome of interest.

Background

Figure 1: Aproved Loan and Written-off Probability

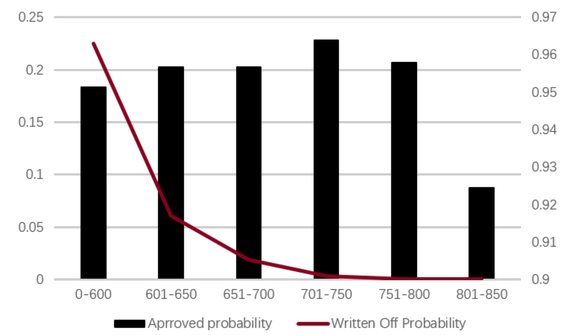
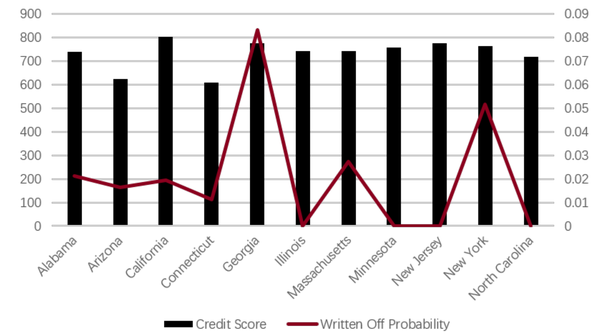


Figure 2: Credit Score of Different States



Methodologies

Data Pre-processing

- NaN, categorical, strings inputs
- Outlier Detection (data that lies in extreme quantile/ unreasonable inputs such as debt ratio greater than 1)
- Imbalanced Data
- Data Binning and Feature Selection

Machine Learning training

- Trained pre-processed data into following Classifiers: Extra Trees, Random Forest, Decision Tree, K Neighbors, Extreme Gradient Boosting, Light Gradient Boosting Machine, Gradient Boosting, Ada Boost, Ridge, Quadratic Discriminant Analysis, Linear Discriminant Analysis, Logistic Regression Naive Bayes, SVM, Dummy.

Final Prediction

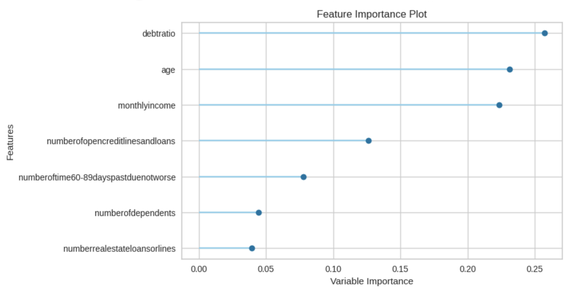
- We choose the model with the highest AUC score to predict delinquency probability.

Feature Selection

Figure 5: ML Scores Grid

feature	n_bins	iv	js	gini
revolving credit limit	11	1.085953	0.125257	0.546613
age	8	0.262774	0.031702	0.269925
30-59days good	3	0.506392	0.0593	0.299516
debt ratio	6	0.0425	0.005275	0.102624
income	9	0.14587	0.018	0.213168
loan accounts	8	0.141552	0.016958	0.157508
90days late	2	0.669932	0.071576	0.2637
real estate loan	3	0.076271	0.009503	0.140315
60-89days good	2	0.403839	0.044459	0.187343
dependents	4	0.028402	0.003542	0.088847

Figure 6: Feature importance plot

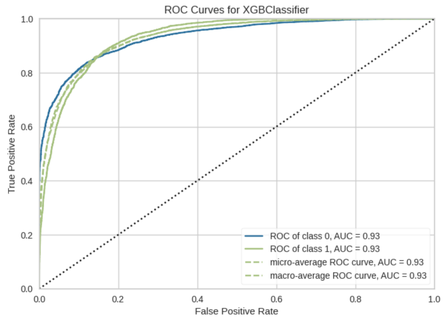


Conclusion

Figure 3: ML Model results

Model	Accuracy	AUC	Recall	Precision	F1 Score	Kappa	MCC	Training Time (s)
ET	0.9939	1.0000	1.0000	0.9880	0.9940	0.9878	0.9879	6.8790
RF	0.9903	1.0000	1.0000	0.9809	0.9904	0.9806	0.9807	10.9700
DT	0.9658	0.9658	1.0000	0.9360	0.9669	0.9316	0.9338	0.6810
KNN	0.9098	0.9671	1.0000	0.8472	0.9172	0.8195	0.8332	0.9240
XGB	0.8317	0.9114	0.8566	0.8159	0.8358	0.6634	0.6642	1.3790
LGBM	0.7482	0.8368	0.7498	0.7475	0.7486	0.4964	0.4965	2.4160
GBDT	0.6949	0.7726	0.6828	0.6998	0.6912	0.3898	0.3899	10.2350
AdaBoost	0.6878	0.7563	0.6585	0.6995	0.6784	0.3756	0.3763	3.2710
Ridge	0.6631	0.0000	0.6354	0.6727	0.6535	0.3261	0.3267	0.4090
LDA	0.6631	0.7290	0.6354	0.6726	0.6535	0.3261	0.3266	0.5120
QDA	0.6576	0.7280	0.5148	0.7209	0.6002	0.3152	0.3291	0.4120
Logistic	0.6566	0.7241	0.5826	0.6839	0.6291	0.3131	0.3167	2.4740
NB	0.6541	0.7222	0.5367	0.7018	0.6079	0.3083	0.3174	0.6970

Figure 4: XGBoost AUC-ROC PLot



- Highlights the models with excellent scores. The best one is Extra Tree. To prevent overfitting, we use grid search to optimize the hyperparameter.
- By utilizing XGBoost Classifer, we can reach 73.47% accuracy, and 96.59% AUC score, which proves machine learning can be powerful in predicting delinquency probability.
- Feature Selection is essential for such a PD model, thus choosing appropriate techniques is significant. When the database is huge, we might need to consider Neural Networks instead of classification algorithms.
- Nunc at convallis urna. isus ante. Pellentesque condimentum dui. Etiam sagittis purus non tellus tempor volutpat. Donec et dui non massa tristique adipiscing.

References

Hand, D. J., amp; Henley, W. E. (1997). Statistical Classification Methods in Consumer Credit scoring: A Review. Journal of the Royal Statistical Society Series A: Statistics in Society, 160(3), 523–541. https://doi.org/10.1111/j.1467-985x.1997.00078.x

Future Research

Building on our work, we plan to combine different models to improve our predictions. We aim to integrate non-traditional data, such as state and marital status, into our research. We'll also use other datasets, like DFCU, to test our model. Lastly, we aim to create a model that can predict expected losses.

This approach will give us a better understanding of financial risks and help us make better decisions.

Contact Email mtan818@bu.edu, Chuyiwww@bu.edu