Risk Models and ML Methods

Version 2.0

Dr. Ye Luo
HKU Business School

Feb 2025

Outline of Today

- Classification in financial Risks.
- Logistic Regression, SVM.
- Python Implementation
- Real data example.

Topics: Credit risk model for SME and individuals (with CCB)

Banking & Finance

China's small business lending push could equal US$418 billion in new loans, S&P Global says

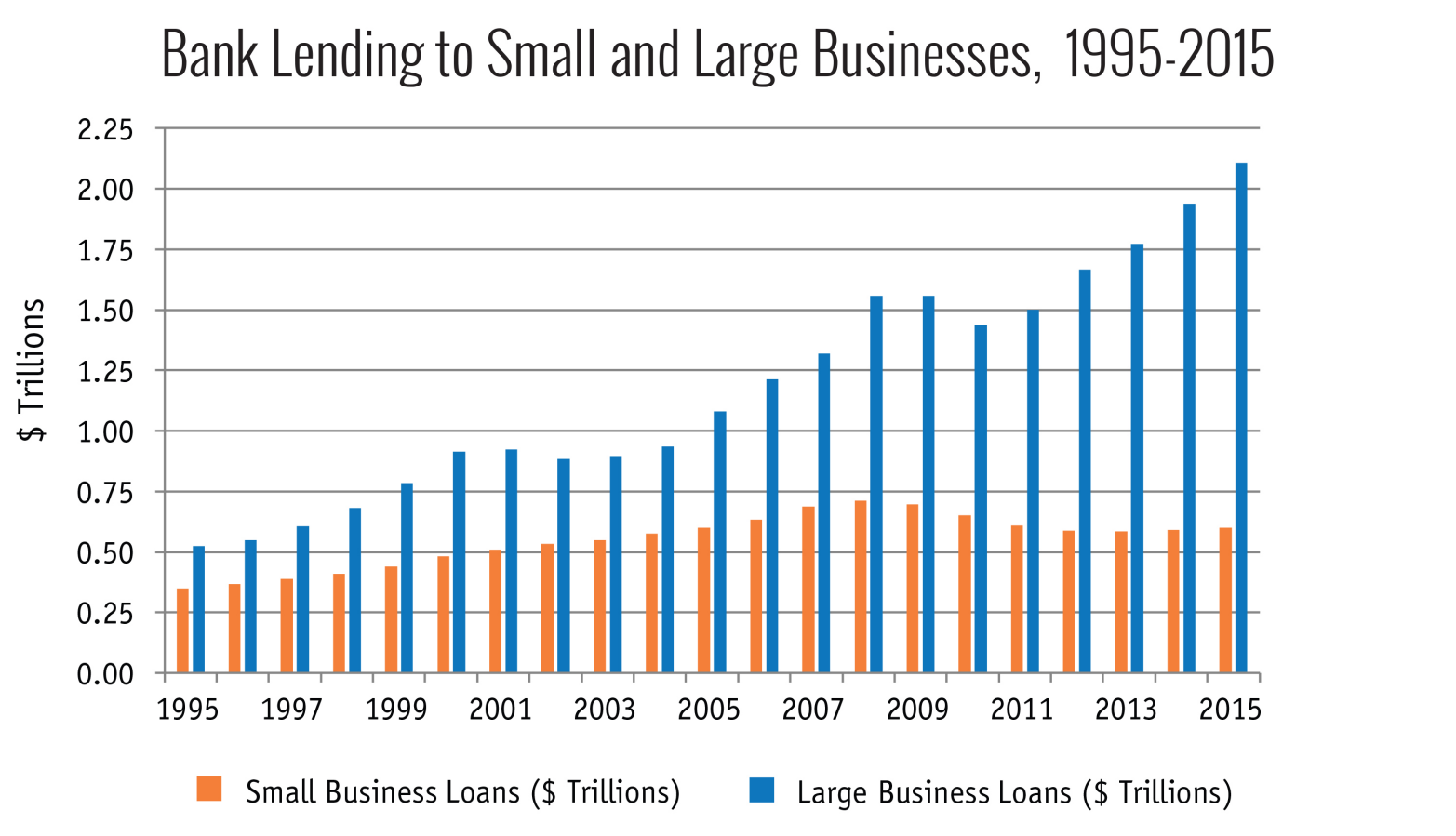
- Lending push could lead to higher credit costs and weaker asset quality, but the affect on credit quality is limited
- Loans to micro and small businesses in China totalled about 9.36 trillion yuan at the end of 2018

Chad Bray
Published: 6:51pm, 15 Apr, 2019

Why SME-lending?

- Economic downturn requires relaxation of SME financial constraints.
- SME-lending is quite low in volume among all lendings in the globe.
- The obstacle is the technique to adequately estimate default risks. You can also do collateral without screening, but it will be difficult to liquidate the collateral in bad markets - may leads to insolvency of financial institutions.
- Lack of data, bad quality of data of SME, fraud/fake data, unreliable data, white gloves, lack of financial monitoring, etc.

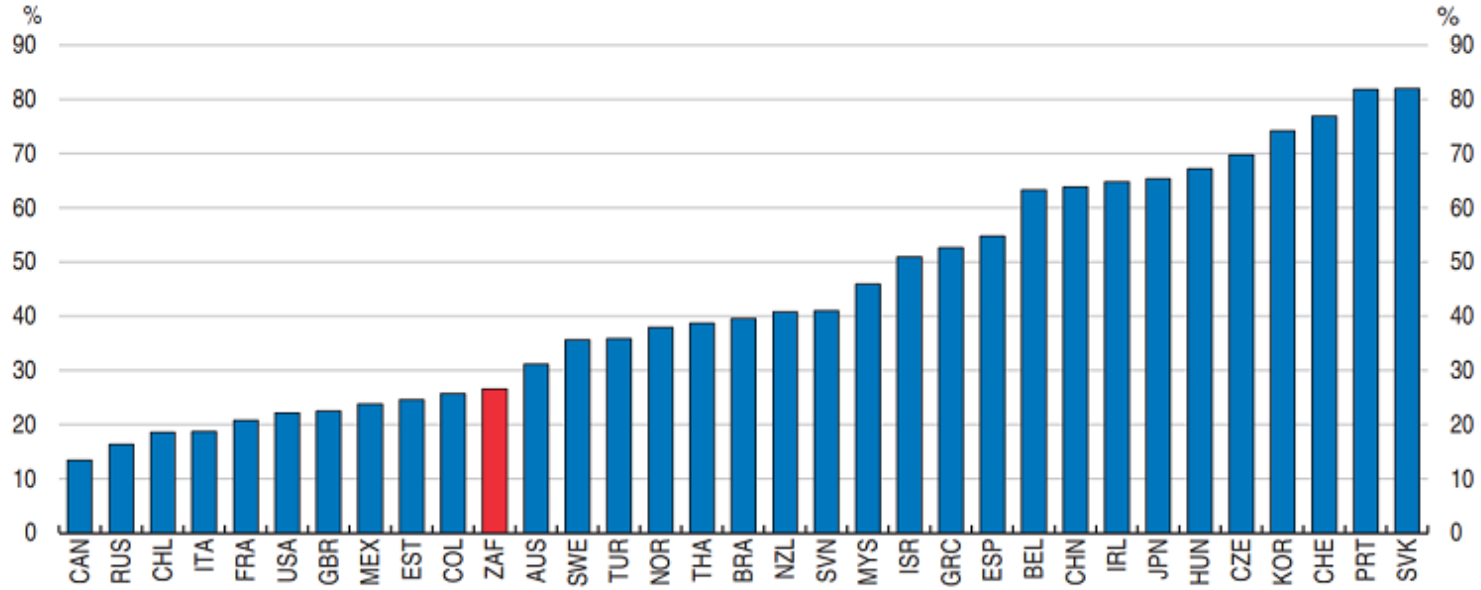
Topics:US-lending ratio



Topics:Global-lending ratio

Figure 34. **SME lending is relatively low**

As a percentage of total business lending, 2015 or latest

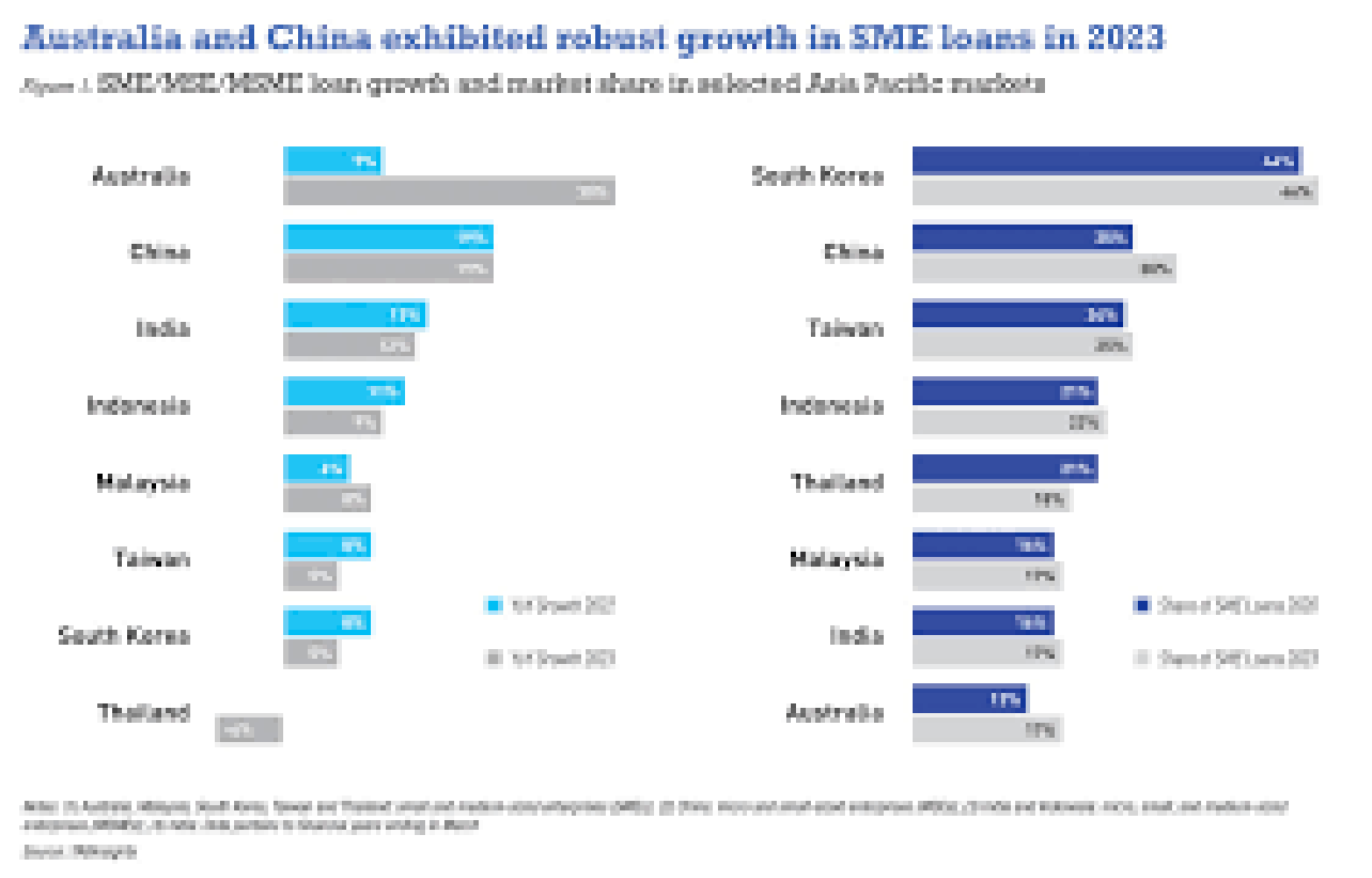


Note: Definitions differ across countries. Data for South Africa are for 2016.

Source: South African Reserve Bank; OECD (2017), *Financing SMEs and Entrepreneurs 2017: An OECD Scoreboard*; OECD calculations.

StatLink http://dx.doi.org/10.1787/888933553100

Topics:Global-lending ratio in 2023



Topics:China small banks lending ratio

Topics:Big 5 lending ratio

Topics:SME Lending in 2023

新华社 XINHUA NEWS 新华财经 XINHUA FINANCE | 面包财经

增速排名	证券简称	较上年末增速	普惠型小微企业贷款余额（亿元）
1	兴业银行	13.92%	4,604.22
2	中信银行	13.88%	5,078.92
3	光大银行	13.58%	3,467.66
4	招商银行	12.50%	7,631.29
5	浙商银行	9.34%	3,031.90
6	平安银行	9.30%	5775.73
7	华夏银行	8.20%	1,741.02
8	民生银行	8.10%	5935.32
9	浦发银行	6.24%	4000.84

数据来源: 新华财经、面包财经、公司公告

Topics:SME Lending in 2023 versus 2022

新华社 XINHUA NEWS 新华财经 XINHUA FINANCE | 面包财经

排名	证券简称	2023年占比	2022年	较上年末变动（百分点）	变动排名
1	浙商银行	18.41%	18.24%	0.17	9
2	平安银行	16.79%	15.87%	0.93	1
3	民生银行	13.52%	13.26%	0.26	8
4	招商银行	12.01%	11.21%	0.80	2
5	中信银行	9.44%	8.66%	0.78	3
6	光大银行	9.26%	8.55%	0.72	4
7	兴业银行	8.82%	8.11%	0.71	5
8	浦发银行	8.04%	7.68%	0.35	6
9	华夏银行	7.43%	7.08%	0.35	7

数据来源: 新华财经、面包财经、公司公告

Interest rate table in Chinese Money Market

Institutions	Interest Rate
Large Banks	5.1% - 5.5%
Regional Banks	6.5% - 7%
Fintech	15%
Others (informal lending)	$> 36\%$

some statistics for comparison

- Total number of SME in China: $80 \sim 120$ Million. Active SME: $60 - 70$ Million.
- Coverage: Banks, Big 5: < 1.5 Million out of around 10 Million customers.
- Internet platform: Alibaba: 300 Billion RMB Loan Balance for 200K T-MALL and 10M Taobao sellers.
- Chinese total credit market per year: ~ 10 Trillion RMB.

Feature of SME financing

- It must be automated. There are just too many SMEs to be financed. Bank's manpower is limited.
- It must be data driven - risk models play an important role in the process.
- Improving the data collection process: quality of data, variety of data, frequency of observations.
- Improving the models/algorithms: better feature engineering, better models and faster algorithms, invention of new algorithms.
- Automation of processes in the traditional financial agency is becoming a big business.

Classification Problems in Finance

- Predicting risks (default prediction, fraud detection).
- Simple Picture: outcome variable is binary. Complex picture: banks actually put many binary labels on you: M+1: early warning, M+2: 2 months delay of payments, M+3: default.
- You may pick one to start with, such as M+3.
- In real data: $M+1-->M+2:50\%$, $M+2-->M+3:80-85\%$.

Risk models in financial industry

- First Generation: E.Altman's Z-score (1968):
  Z-score = 1.2 working capital/total assets+1.4 retained earnings/total assets + earnings before interest and tax/ total assets + market value of equity/ total liabilities + sales/ total assets.
  A score below 1.8 means it is likely to go bankruptcy.
- Many firms and auditing agencies are still using Z-score.
- Not good for SMEs because the traditional financial measures are not accurate.

Risk models in financial industry

- 2nd generation: logistic regression(1980s):
  Probability of default = Logit(-Score), where $Score = X\beta$ for some characteristics X. $Logit(u) = \exp(u)/(1 + \exp(u))$.
- Model is estimated from data, it is the workhorse model in the industry so far.
- A majority of Chinese banks are using this generation of models, established at around 2000-2010s.

- Multi-dimensional Predictors X - characterize borrower's features.
- Typical features: finance related measurement: cash flow/income, debt, etc. Demographics: age, gender, education, marriage, homeowner, etc. Alternative measurements: digital footprints, financial behavior, self-disciplinary measurements, etc.
- A scalar label $Y \in \{0, 1\}$, indicating for default/non-default in the record.
- Task: Predict Y from X, e.g., $Pr(Y = 1|X) = f(X, \beta)$, estimation of β given a functional form f.

Maximum Likelihood Method

- The likelihood of observing Y given X is:

$$L(Y|X,\beta) := 1(Y=1)f(X,\beta) + 1(Y=0)(1-f(X,\beta)).$$

- We maximize the log-likelihood over sample (X_i, Y_i), $i = 1, 2, ..., n$:

$$\hat{\beta} = \text{argmax}_\beta \sum_{i=1}^{n} \log L(Y_i|X_i,\beta).$$

- Why taking the logarithm? The likelihood of parameter β given the sample (assuming independence) is that:

$$\Pi_{i=1}^{n} L(Y_i|X_i,\beta).$$

- Multiplication is difficult to deal with - use logarithm to turn it into a sample mean.
- The L function here can be interpreted as $-1\times$ loss function. Minimizing loss function is equivalent to maximizing log-likelihood.

The workhorse model: logistic regression

- PD_i = probability of default for i^{th} observation. X_i = characteristics/features. Only observe actual default behavior $D_i \in \{0, 1\}$.
- $D_i = 0$: non-default. $D_i = 1$: default.
- Log Likelihood of data $L(\beta) := \sum_{i=1}^{n} \log Pr(D_i|X_i)$, where D_i = default behavior, binary (observed).
- $D_i = 1(X_i\beta + \epsilon_i < 0)$. When $\epsilon \sim Logit(\cdot)$, $Pr(D_i = 0|X_i) = \frac{\exp(-X_i\beta)}{1+\exp(-X_i\beta)}$. $Pr(D_i = 1|X_i) = 1 - Pr(D_i = 0|X_i)$.
- Maximize log-likelihood (MLE): $\hat{\beta} = argmax_\beta L(\beta)$.

Maximum Likelihood Method

- This approach is referred as maximum likelihood approach. Namely, with a model $Pr(Y|X,\beta)$ where β is the parameter of interest, the Maximum Likelihood Estimator (MLE) is referred as:

$$\hat{\beta}_{MLE} = argmax_{\beta} \sum_{i=1}^{n} \log Pr(Y|X,\beta).$$

- For linear regression: $Y_i = X_i\beta + \epsilon_i$. Assuming that $\epsilon_i \sim N(0,\sigma^2)$ as a normal random variable. The likelihood of observing Y_i conditional on X_i is:

$$Pr(Y_i|X_i,\beta) = \frac{1}{\sqrt{2\pi}\sigma} \exp^{-\frac{(Y_i - X_i\beta)^2}{2\sigma^2}}.$$

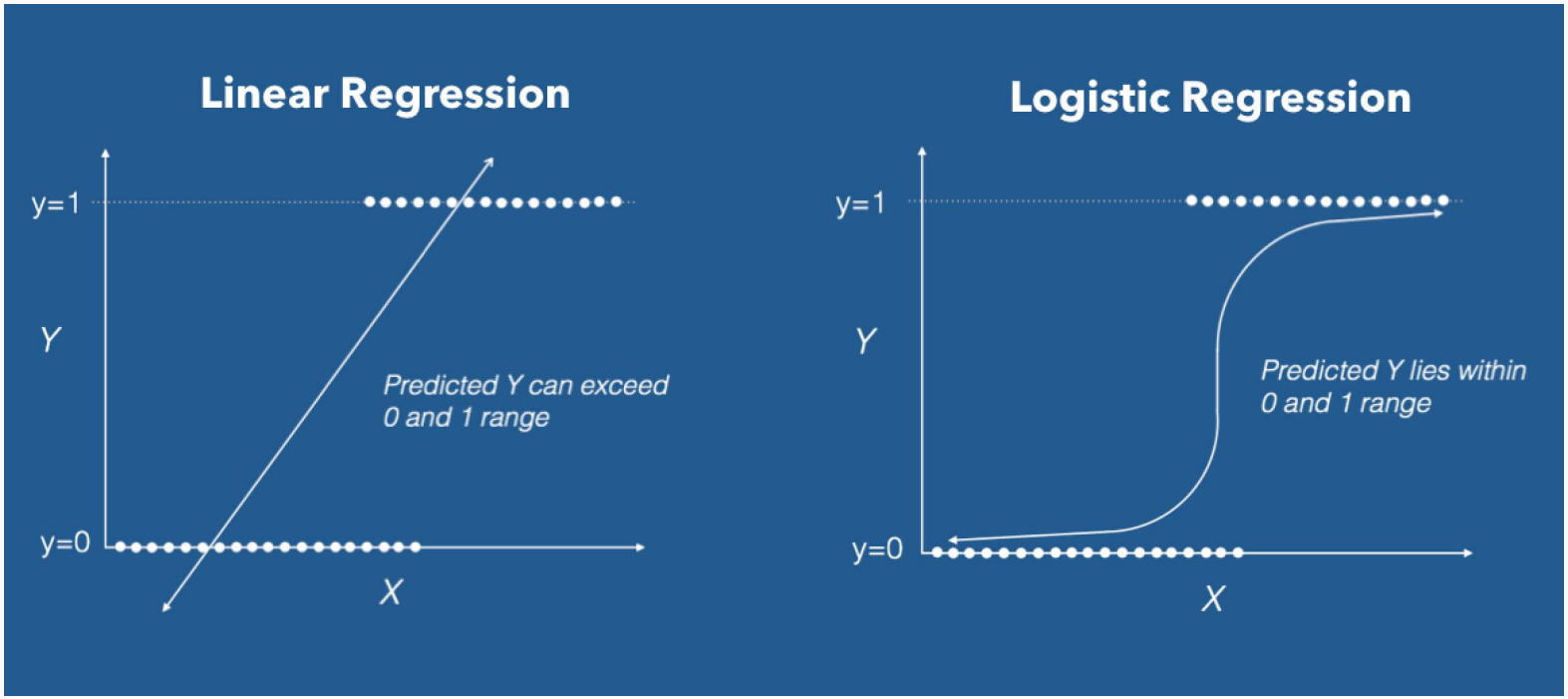
- The maximum log-likelihood yields:

$$\begin{aligned}\hat{\beta}_{MLE}, \hat{\sigma}_{MLE} &= argmax_{\beta,\sigma} \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi}\sigma} \exp^{-\frac{(Y_i - X_i\beta)^2}{2\sigma^2}} \\ &= argmax_{\beta,\sigma}\, n \log\left(\sqrt{2\pi}\sigma\right) - \sum_{i=1}^{n} (Y_i - X_i\beta)^2/\sigma^2.\end{aligned}$$

MLE and Related Methods

- Fixing σ, the MLE approach is equivalent to linear regression in this case - what is the key condition?
- Typically MLE requires modelling of the distribution of the residual ϵ.
- For binary classification problem: Logit: $Pr(\epsilon < z) = \frac{\exp(z)}{1+\exp(z)}$.
- Probit: $Pr(\epsilon < z) = \Phi(z/\sigma)$, where $\Phi(x) = \int_{-\infty}^{x} \phi(x)$, with $\phi(x) = \frac{1}{\sqrt{2\pi}} \exp^{-\frac{x^2}{2}}$ is the density function of a standard normal random variable.

Logistic Regression



sklearn.linear_model.LogisticRegression

In Package sklearn.linear_model, there is a class called LogisticRegression.

class sklearn.linear_model.LogisticRegression(penalty='l2', dual=False, tol=0.0001, C=1.0, fit_intercept=True, intercept_scaling=1, class_weight=None, random_state=None, solver='warn', max_iter=100, multi_class='warn', verbose=0, warm_start=False, n_jobs=None, l1_ratio=None)

Key inputs: penalty functions, solver. You can set others to be the default values.

Parameters of input

- penalty functions: penalty : str, ‘l1’, ‘l2’, ‘elasticnet’ or ‘none’, optional (default=’l2’). For small problems, you may just set penalty as none. Default penalty is ’l2’.
- solver : str, {‘newton-cg’, ‘lbfgs’, ‘liblinear’, ‘sag’, ‘saga’}, optional (default=’liblinear’).
- newton-cg, lbfgs, sag, saga works for multi-classes.
- liblinear does not handle ’l1’ penalty.
- LogisticRegression.fit(X,Y)

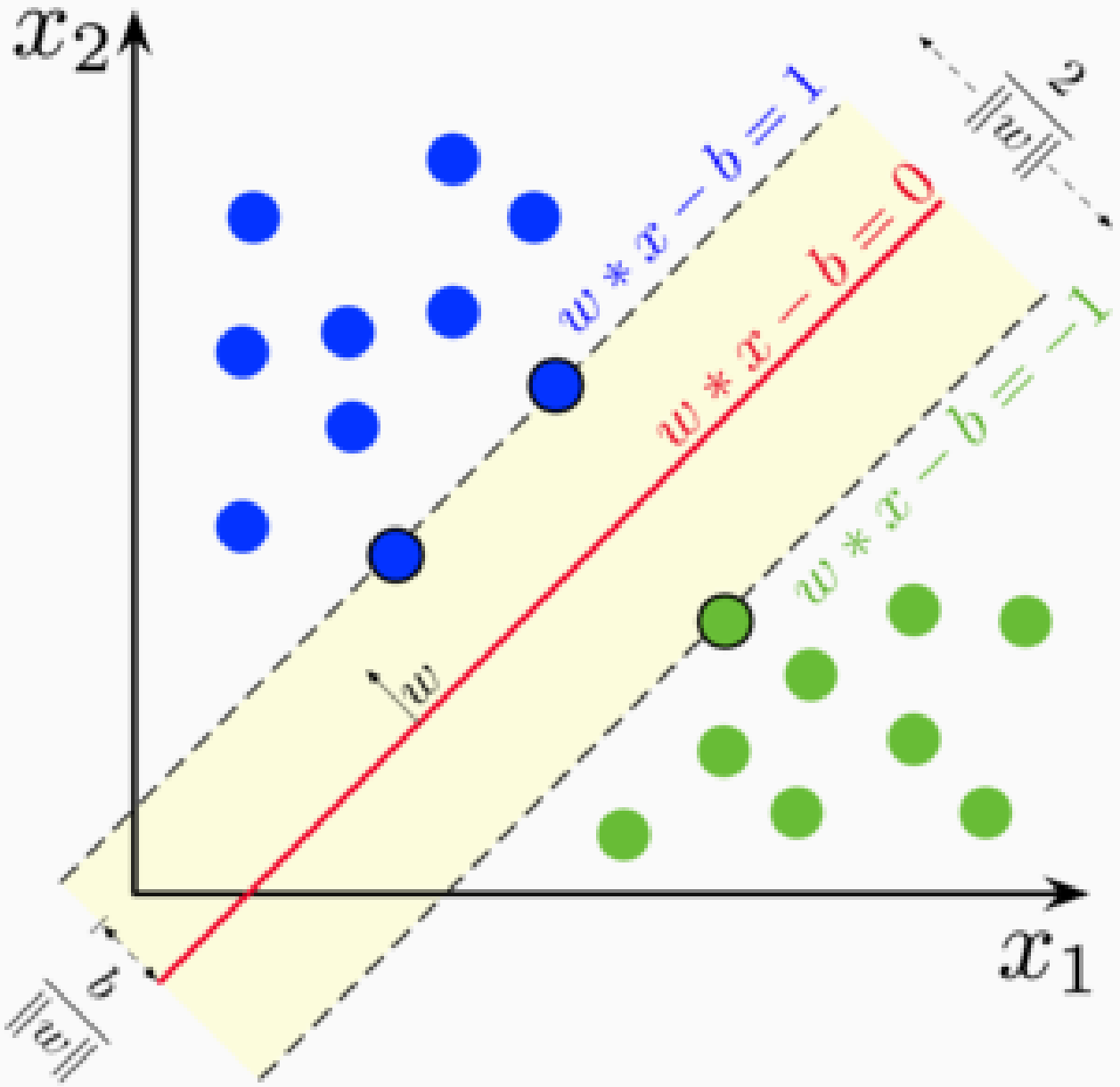
Actually implement it in python

- call a class: logisticRegr = LogisticRegression()
- Train models on (x_train, y_train):
  logisticRegr.fit(x_train, y_train)
- Predicting multiple outcomes on test data:
  logisticRegr.predict(x_test[0:10]).
- Report mean accuracy on the give data set.
  logisticRegr.score(x,y):

Support Vector Machines

- Generation 2.5: Support Vector Machines(SVM)(1990s-2000s).
- Relatively more advanced than logistic regressions.
- Graphically, the SVM tries to learn the maximum gap between classes.
- First introduce linear SVM.

SVM



How SVM works?

- Try to locate the boundary points between groups.
- Draw a separation hyper-plane in the middle of the gap.
- Relatively more robust than logistic regression in high dimensional settings.
- Can deal with unbalanced sample (small sample of bad points) relatively well.
- State of Art Machine Learning method before Deep Learning.

Mathematical Formulations

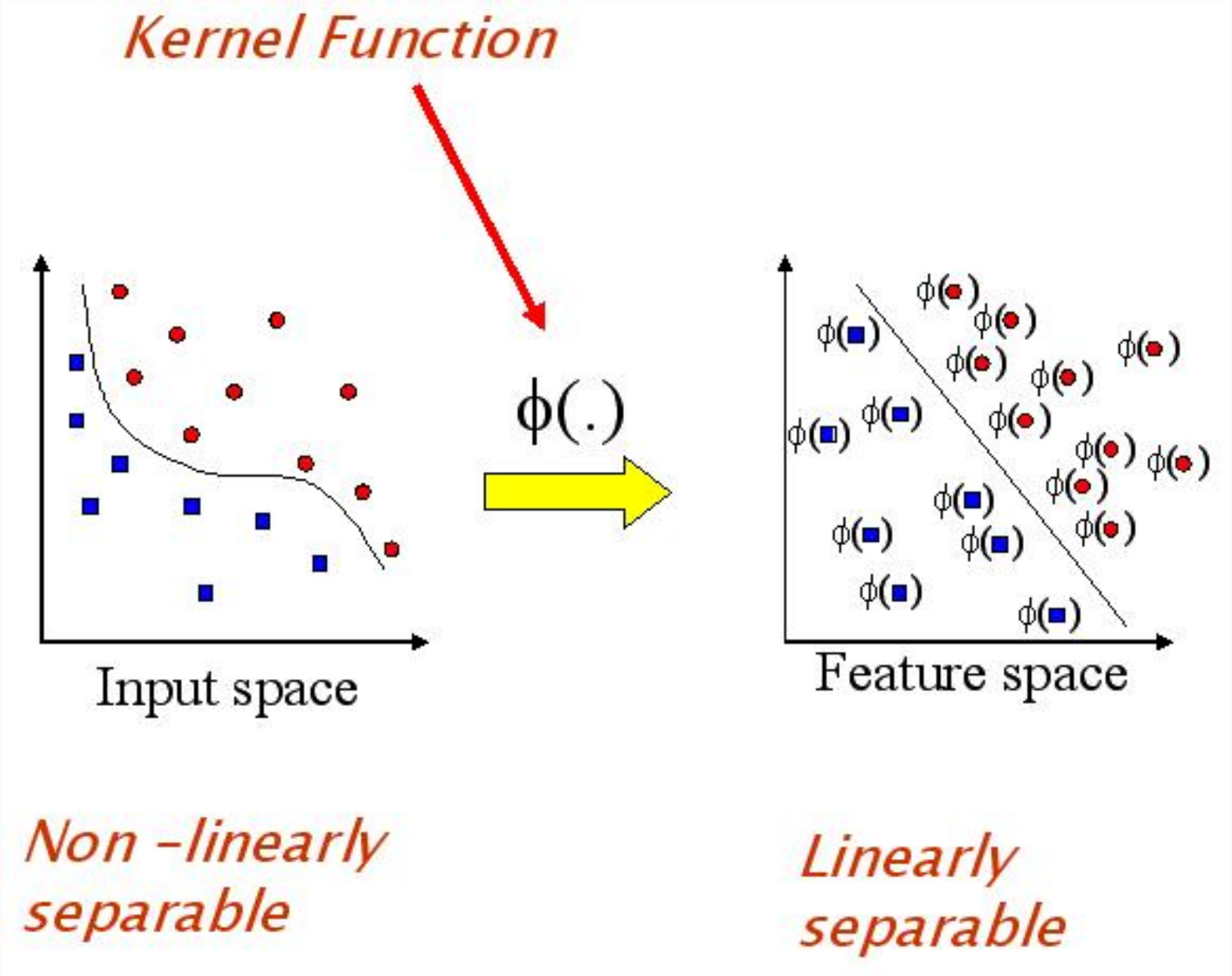
- A model $w'x + b > 0$ classifies $y = 1$, $w'x + b < 0$ classifies $y = -1$.
- Find a set of w, b such that: $y_i(w'x_i + b) \geq 1$ for all $i = 1, 2, ..., n$.
- Use the minimum scale w.
- $\min_{w,b} \frac{1}{2}||w||^2$ subject to:

$$y_i(w'x_i + b) \geq 1$$

for all $i = 1, 2, ..., n$.

Kernel SVM

- The SVM's separating hyper-plane is linear in features x. This assumes linear boundary - may not be correct in practice.
- Given a feature mapping $\phi : x \to \phi(x)$, consider a kernel function $K(x, z) = \phi(x)'\phi(z)$.
- Suppose $x = [x_1, x_2, x_3]$. For $K(x, z) = (x'z)^2$ (polynomial), it is equivalent to have $\phi(x)$ as:
  $\phi(x) = [x_1x_1, x_1x_2, x_1x_3, x_2x_1, x_2x_2, x_2x_3, x_3x_1, x_3x_2, x_3x_3]$.
- Kernel-SVM = SVM with non-linear features of x.
- Replace x with $\phi(x)$ when performing classification.


Kernel Function
ϕ(.)
Input space
Feature space
Non -linearly separable
Linearly separable

Implementation of SVM

- from sklearn import svm
- Data: X,y.
- Use "SVC" function in the svm model. Arguments in SVC include: svm.SVC(C=1.0, cache_size=200, class_weight=None, coef0=0.0, decision_function_shape='ovr', degree=3, gamma='scale', kernel='rbf', max_iter=-1, probability=False, random_state=None, shrinking=True, tol=0.001, verbose=False)
- clf=svm.SVC()
- clf.fit(X,y)

A few key inputs

- class_weight: how to weight the data. Default = None, Choose option "balanced" if you have unbalance sample. The algorithm will reweight the data for you.
- kernel: choose to use kernel function to produce non-linear boundary estimation.
- Linear: $< x, x' >$ - this produces linear SVM. polynomial: $(\gamma < x, x' > + r)^d$, rbf: Gaussian kernel $\exp\left(-\gamma ||x - x'||^2\right)$. γ and d must be specified.

Unbalanceness of sample

- Usually in finance, "bad" samples (default) are quite rare.
- 1-2% non-performing loans in banking industry for major banks - you have way more good sample than bad sample.
- Idea: would rather to kill more bad sample at expense of killing more good sample.
- Re-balance the weights - overweight bad samples when training a model.

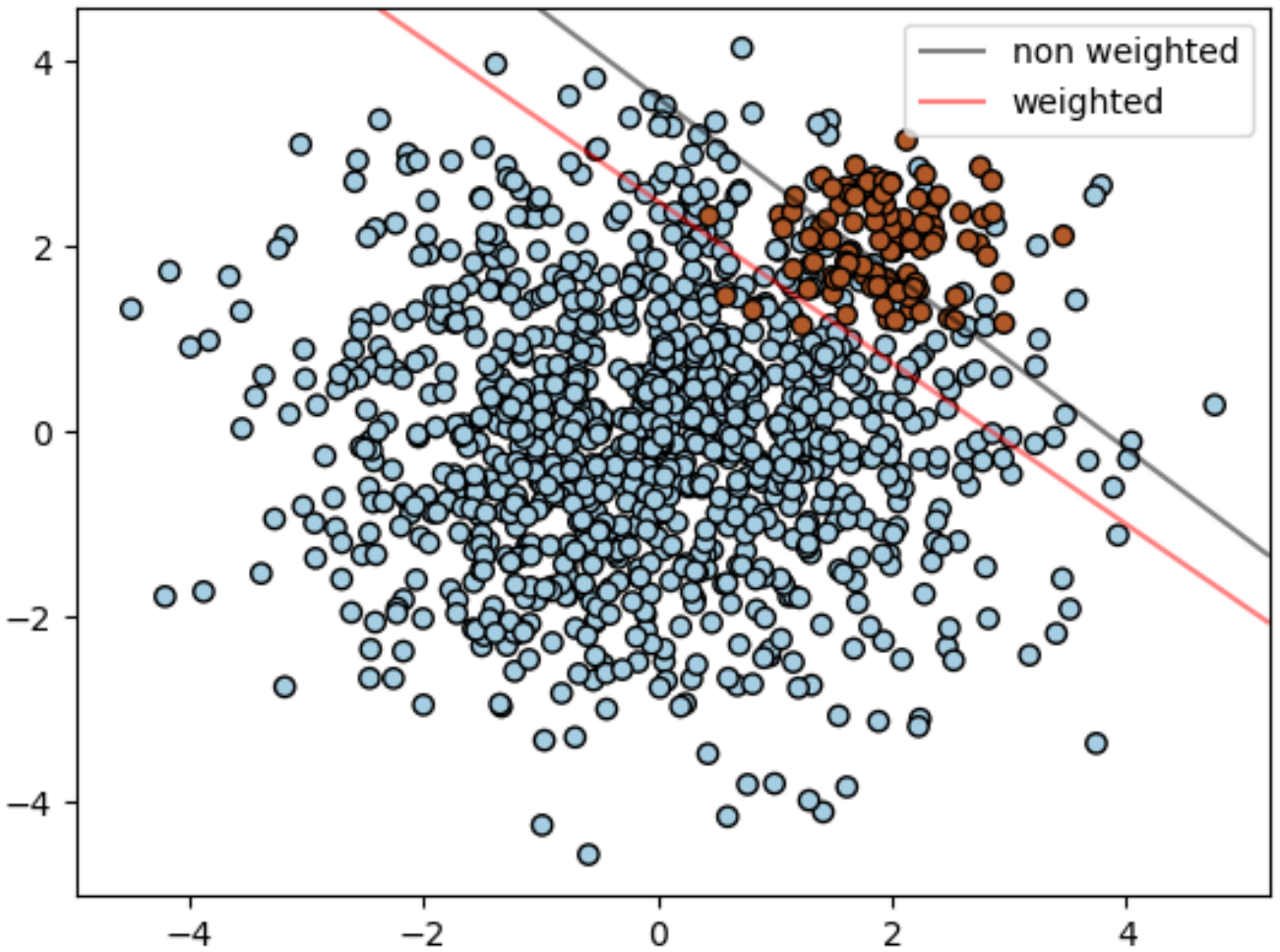
Unbalanceness of sample in MLE setup

- In MLE setup, suppose X_i, Y_i are observed. Assign weight $w_i := 1$ if $Y_i = 0$, and $w_i := K$ if $Y_i = 1$.
- K is set such that $K = \frac{\text{number of good sample}}{\text{number of bad sample}}$.
- Run weighted MLE:

$$\hat{\beta} = \text{argmax}_\beta \sum_{i=1}^{n} w_i \log L(Y_i | X_i, \beta).$$

- We can use similar re-weighting ideas in other applications.

SVM-unbalanced sample



How to measure the effectiveness of a risk model?

- Any risk model will assign a function $g : X \mapsto \{0, 1\}$. X: a vector of predictors.
- For example, given a logistic regression model $f : X \mapsto [0, 1]$ that predicts the likelihood of default. By employing a cutoff threshold $c \in (0, 1)$, the $g(\cdot)$ function can be the indicator function

$$g := 1(f(X) > c).$$

- Such rule is equivalent to use "score card" to determine loan decisions in the banking industry.

Positive Measures

Null hypothesis: individual i is negative (means non-default in financial application).

- Given a decision rule $g(\cdot)$, we define:
- True positive (TP): the model detects positive sample $(Y_i = 1, g(X_i) = \hat{Y}_i = 1)$ as "positive".

$$TP = \sum_{i=1}^{N} 1(g(X_i) = 1)1(Y_i = 1).$$

- False positive (FP): the model detects negative sample $(Y_i = 0, g(X_i) = \hat{Y}_i = 1)$ as "positive".

$$FP = \sum_{i=1}^{N} 1(g(X_i) = 1)1(Y_i = 0).$$

Negative Measures

Null hypothesis: individual i is negative (means non-default in financial application).

- True negative (TN): the model detects negative ($Y_i = 0, g(X_i) = \hat{Y}_i = 0$) as "negative".

$$TN = \sum_{i=1}^{N} 1(g(X_i) = 0)1(Y_i = 0).$$

- False negative (FN): the model detects positive ($Y_i = 1, g(X_i) = \hat{Y}_i = 0$) as "negative".

$$FN = \sum_{i=1}^{N} 1(g(X_i) = 0)1(Y_i = 1).$$

Discussions

- TP: will be rejected correctly - less non-performing loans.
- FP: will be rejected - less revenue.
- TN: will be accepted - more revenue.
- FN: will be accepted - non-performing loans (mostly affects default rates).

Measurements

- Sensitivity, Recall, or True positive rate (TPR). FNR: false negative rate, miss rate.

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN} = 1 - FNR.$$

- Specificity, selectivity or True negative rate (TNR). FPR: false positive rate.

$$TNR = \frac{TN}{N} = \frac{TN}{TN + FP} = 1 - FPR.$$

Measurements-2

- Precision or positive predictive value (PPV). FDR: false discovery rate

$$PPV = \frac{TP}{TP + FP} = 1 - FDR.$$

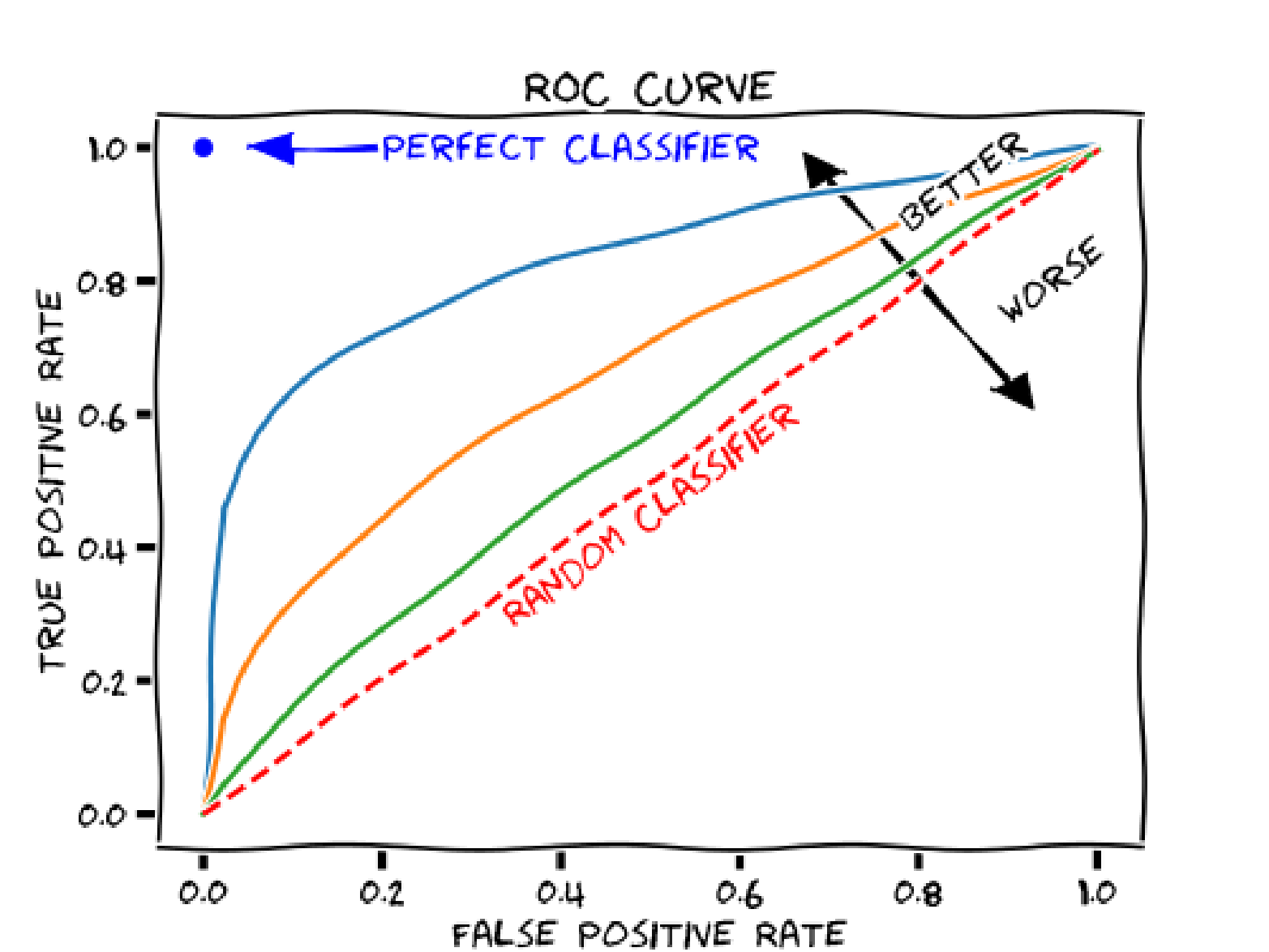
- Negative prediction rate (NPV). FOR: False omission rate.

$$NPV = \frac{TN}{TN + FN} = 1 - FOR.$$

Evaluating a probabilistic model

- $f(X)$ is a model that predicts the default probability.
- Can use a cut-off c to make decisions.
- When you move c from 0 to 1, what changes?
- $FPR =?$ when $c = 1$? $FPR =?$ when $c = 0$?

Receiver operating characteristics



Discussion

- ROC is the boundary of decision rule - pick one point according to bank's interest.
- 10% or 20% cut-off is often used.
- AUC: Area under the curve of ROC- measures quality of ROC. The bigger the better.
- Most "good" risk models have AUC around 0.75-0.85.

Real Example: Consumer credit-risk models via machine-learning algorithms, Andrew W. Lo

- Investigated on consumer credit-risk models. ML significantly improve classification rates of credit-card holder delinquencies and defaults, with linear logistic regression at 85%.
- Using conservative assumptions for the costs and benefits of cutting credit lines based on machine-learning forecasts, they estimate the cost savings to range from 6% to 25% of total losses.
- Data: U.S. credit bureau 2005-2009, banks, etc.

Lo data

Transaction data	
Transaction count	*By category (cont.)*
Total inflow	
Total outflow	Hotel expenses
	Travel expenses
By Channel:	Recreation
	Department store expenses
ACH (count, inflow and outflow)	Retail store expenses
ATM (count, inflow and outflow)	Clothing expenses
BPY (count, inflow and outflow)	Discount store expenses
CC (count, inflow and outflow)	Big box store expenses
DC (count, inflow and outflow)	Education expenses
INT (count, inflow and outflow)	Total food expenses
WIR (count, inflow and outflow)	Grocery expenses
	Restaurant expenses
By Category	Bar expenses
	Fast food expenses
Employment inflow	Total rest/bars/fast-food
Mortgage payment	Healthcare related expenses
Credit-card payment	Fitness expenses
Auto loan payment	Health insurance
Student loan payment	Gas stations expenses
All other types of loan payment	Vehicle expenses
Other line of credit payment	Car and other insurance
Brokerage net flow	Drug stores expenses
Dividends net flow	Government
Utilities payment	Treasury
TV	Pension inflow
Phone	Collection agencies
Internet	Unemployment inflow

Lo data

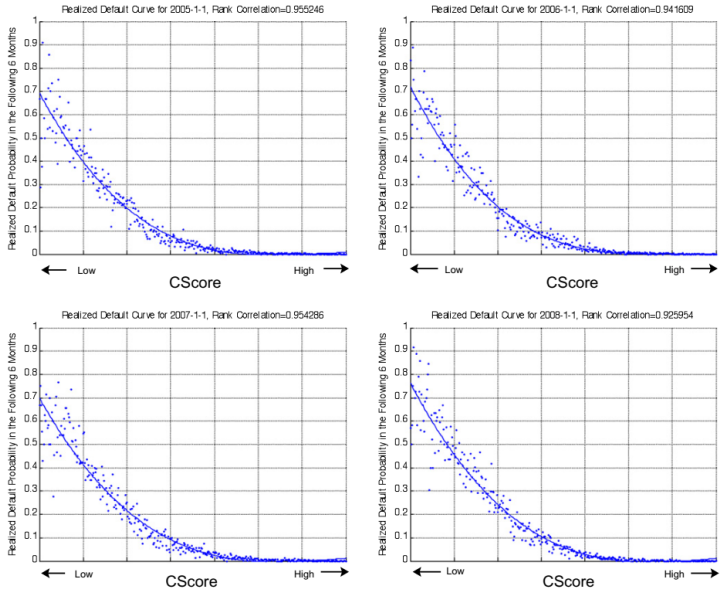


Fig. 2. Relationship between CScore and subsequent default frequency (over the next 6 months) for January of 2005, 2006, 2007, 2008. Second-order polynomial approximations are also shown (solid lines).

features

Model inputs	
Credit bureau data	*Transaction data (cont.)*
Total number of trade lines	Total expenses at discount stores
Number of open trade lines	Total expenses at big-box stores
Number of closed trade lines	Total recreation expenses
Number and balance of auto loans	Total clothing store expenses
Number and balance of credit cards	Total department store expenses
Number and balance of home lines of credit	Total other retail store expenses
Number and balance of home loans	
Number and balance of all other loans	Total utilities expenses
Number and balance of all other lines of credit	Total cable TV & Internet expenses
Number and balance of all mortgages	Total telephone expenses
Balance of all auto loans to total debt	Total net flow from brokerage account
Balance of all credit cards to total debt	Total net flow from dividends and annuities
Balance of all home lines of credit to total debt	
Balance of all home loans to total debt	Total gas station expenses
Balance of all other loans to total debt	Total vehicle related expenses
Balance of all other lines of credit to total debt	
Ratio of total mortgage balance to total debt	Total lodging expenses
	Total travel expenses
Total credit-card balance to limits	
Total home line of credit balances to limits	Total credit-card payments
Total balance on all other lines of credit to limits	Total mortgage payments
	Total outflow to car and student loan payments
Transaction data	
	Total education related expenses
Number of Transactions	
Total inflow	*Deposit data*
Total outflow	
Total pay inflow	Savings account balance
	Checking account balance
Total all food related expenses	CD account balance
Total grocery expenses	Brokerage account balance
Total restaurant expenses	
Total fast-food expenses	
Total bar expenses	

Models

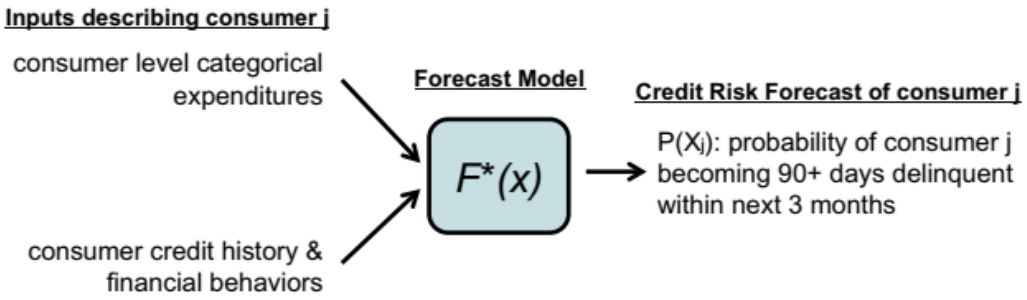


Fig. 10. A summary of the machine-learning algorithm used to construct the consumer credit-risk model.

Average Scores among groups

Training window		Prediction date	Evaluation window		Average predicted probability of 90 + delinquency on credit card in the next 3 months		
Start	End		Start	End	Among all customers	Among customers going 90 + days delinquent	Among customers NOT going 90 + days delinquent
Feb-08	Apr-08	Apr-08	May-08	Jul-08	0.7	10.3	0.7
Mar-08	May-08	May-08	Jun-08	Aug-08	0.4	8.0	0.4
Apr-08	Jun-08	Jun-08	Jul-08	Sep-08	0.5	9.7	0.4
May-08	Jul-08	May-08	Aug-08	Oct-08	0.4	8.6	0.3
Jun-08	Aug-08	Aug-08	Sep-08	Nov-08	0.4	9.6	0.4
Jul-08	Sep-08	May-08	Oct-08	Dec-08	0.5	9.5	0.4
Aug-08	Oct-08	Oct-08	Nov-08	Jan-09	0.5	10.1	0.4
Sep-08	Nov-08	May-08	Dec-08	Feb-09	0.5	10.6	0.5
Oct-08	Dec-08	Dec-08	Jan-09	Mar-09	0.5	8.6	0.5
Nov-08	Jan-09	May-08	Feb-09	Apr-09	0.5	10.2	0.5

Classifier Threshold = 10%

Actual Outcome	Model Prediction: Good	Model Prediction: Bad
Good	95.16%	2.27%
Bad	0.32%	2.25%

Classifier Threshold = 20%

Actual Outcome	Model Prediction: Good	Model Prediction: Bad
Good	96.37%	1.06%
Bad	0.44%	2.13%

Classifier Threshold = 30%

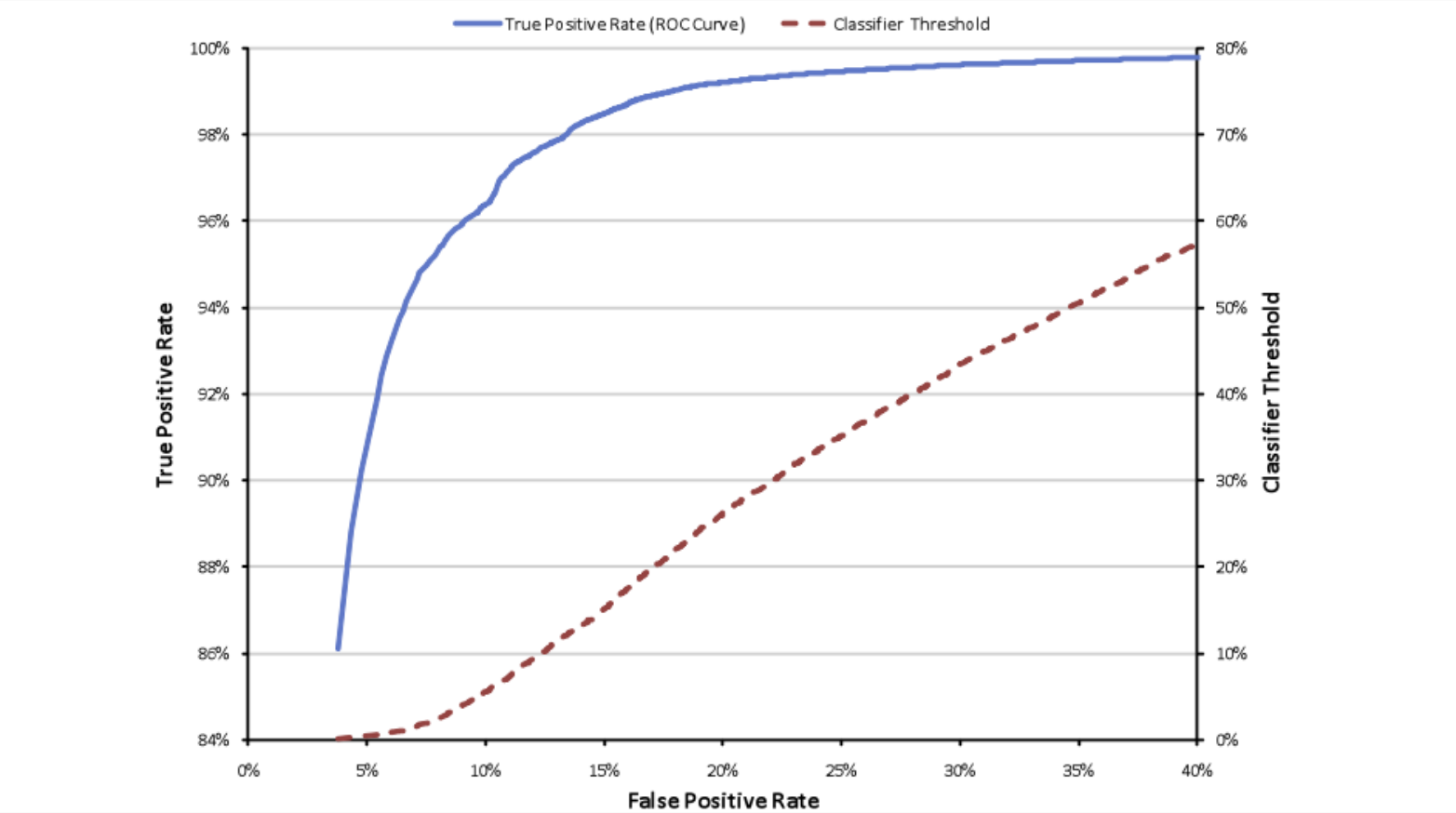
Actual Outcome	Model Prediction: Good	Model Prediction: Bad
Good	96.78%	0.65%
Bad	0.57%	2.00%

Classifier Threshold = 50%

Actual Outcome	Model Prediction: Good	Model Prediction: Bad
Good	97.14%	0.29%
Bad	0.89%	1.68%

Fig. 15. Confusion matrices of machine-learning forecasts of 90-days-or-more delinquencies for four classification thresholds. Rows correspond to actual customer types ("Bad" is defined as 90-days-or more delinquent). The numerical example is based on the December 2008 model forecast for the 3-month forecast horizon from January to March 2009.

Test Curves


True Positive Rate (ROC Curve)
Classifier Threshold
100%
98%
96%
94%
92%
90%
88%
86%
84%
True Positive Rate
80%
70%
60%
50%
40%
30%
20%
10%
0%
Classifier Threshold
0%
5%
10%
15%
20%
25%
30%
35%
40%
False Positive Rate

Case Study: Chinese Credit Lending Market and Risk Models

- SME behavior is hard to capture - lack of data.
- Corporate credit system is not well constructed by far.
- Many SMEs' lack of even digital footprint, leading to difficulties to measure their risks.
- Utility such as electricity consumption, health care, etc., are dispersed in different departments of the government.

The solution plan

- It's difficult to characterize firm risk directly. But SME's risk is closely related to SME owner's risk.
- The financial institutions have strong information on personal accounts, including deposits, investments, loans, debts, and etc.
- The data can be divided into strong financial information, and behavior information.

Related Economic and Financial Theory

- Performance coherence: the entrepreneur who takes more risks in their personal lifes also tend to take more risks in managing their firms.
- Self-discipline: More self-disciplinary person in their personal life tends to be more self-disciplinary when managing their firms.
- Investment-Consumption equilibrium: The change of consumption is related to the change of asset returns or risks.

Data and Modeling

- Our data is based on a random sample of about 100,000 lending records to small enterprises between 2018 and 2019 from a large national bank in China.
- or each business, we use high-frequency records from its owner's personal banking to construct 26 entrepreneur personal features that span four key dimensions:
  1. basic sociodemographic characteristics such as gender, age, and education;
  2. consumption behaviors such as the amount, type, and intra-day timing of spending;
  3. financial behaviors such as the level and allocation of financial assets;
  4. credit activities such as credit limits, utilization and delinquency.

Summary of Stats

Table I
Summary statistics

This table presents summary statistics for our sample. Panel A provides descriptive statistics on business-specific features. Panel B provides descriptive statistics on entrepreneur personal features.

Variables	Mean (st.d.)	Variables	Mean (st.d.)	Variables	Mean (st.d.)	Variables	Mean (st.d.)
Panel A. Business-specific Features							
A1. Credit records		**A2. Financial features**		**A3. Basic characteristics**			
No. related business entities	1.4 (3)	Business account balance	179,000 (389,000)	Age (years)	5 (4)		
Bad business credit (%)	5	Business net cash flow (RMB)	8,000 (45,000)				
Non-performing business debt (10K RMB)	0.06 (19)						
Business debt (10K RMB)	12 (26)						
Panel B. Entrepreneur personal features							
B1. Sociodemographic characteristics		**B2. Consumption behaviors**		**B3. Finance behaviors**		**B4. Credit behaviors**	
Age (years)	41 (8)	Consumption ratio (%)	10 (40)	Personal account balance (RMB)	203,000 (344,000)	Length of overdue conditional on overdue (month)	1 (0.2)
Female (%)	24	Discretionary consumption (%)	17 (22)	Stock market participation (%)	20	Amount overdue (RMB)	900 (4,000)
Education (postgraduate) (%)	5	Housing (%)	11 (20)	Share of stock investment (%)	2 (10)	Credit card delinquency (%)	0.6
Education (bachelor) (%)	60	Traveling (%)	0.4 (2)	Share of stock investment conditional on participation (%)	12 (20)	Total credit limit (RMB)	345,000 (490,000)
Education (junior college) (%)	10	Gambling (%)	0.01 (0.3)	Length of relationship (month)	60 (42.7)	Max credit limit (RMB)	47,000 (75,000)
Married (%)	89	Cash spending (%)	8 (11)			Min credit limit (RMB)	9,000 (14,000)
Birthplace level (large) (%)	12	Work hour consumption (%)	25 (15)			Credit utilization (RMB)	9,000 (22,000)
Local resident (%)	14	Latest hour (hour)	20 (1.0)			No. creditors	3 (2.2)

Pairwise-Correlation

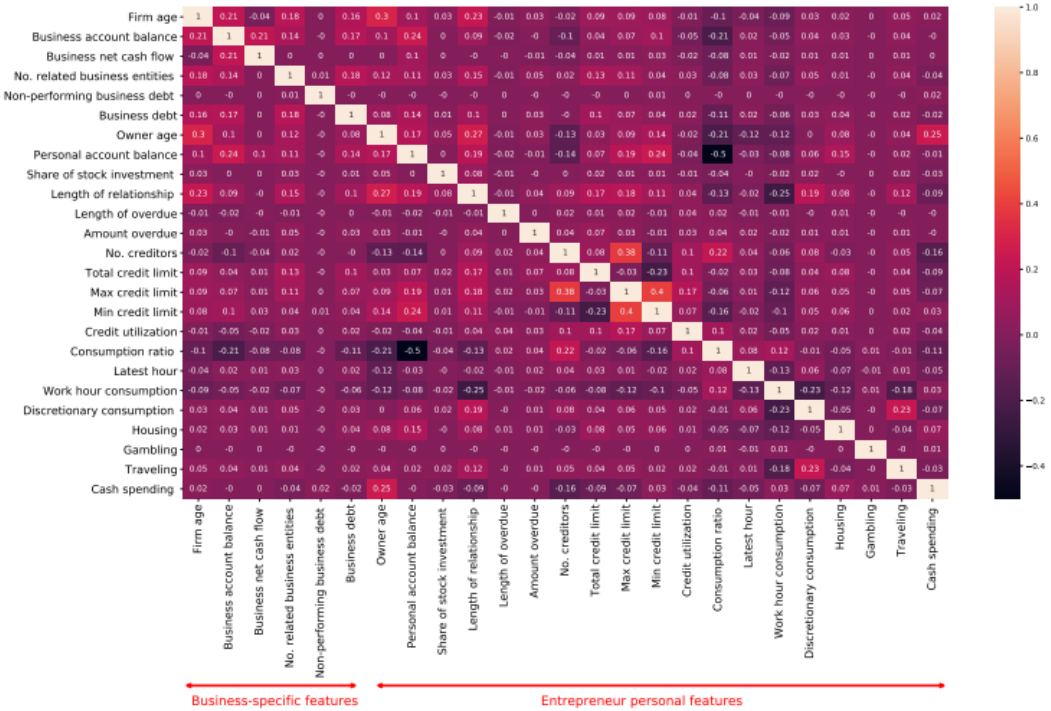


Figure 1. Pairwise Correlation Matrix

This figure presents Pearson correlation coefficients between the numeric variables, including 6 business-specific features and 19 entrepreneur personal features in the sample.

Benchmark model

Table II
Benchmark model: Firm characteristics and small business credit outcomes

This table estimates a standard logistic regression model where the dependent variable ($Default$ (0/1)) is equal to one if the small business loan is in delinquency for over 30 days. Column 1 and 2 estimate the predictive power of firm credit records. Column 3 and 4 estimate the predictive power of firm financial features. Column 5 uses both the firm credit records and firm financial features. Standard errors are clustered by month, firm industry, and firm location. *, **, and *** represent statistical significance at the 10%, 5%, and 1% levels, respectively.

	(1) Firm credit records		(2) Firm credit records & FEs		(3) Firm financial features		(4) Firm financial features & FEs		(5) Firm credit records & Firm financial features & FEs	
Variables	Coef.	St.D.	Coef.	St.D.	Coef.	St.D.	Coef.	St.D.	Coef.	St.D.
No. related business entities	0.001	(0.009)	0.013	(0.009)					0.020 **	(0.010)
Bad business credit	0.274 ***	(0.051)	0.317 ***	(0.051)					0.233 ***	(0.052)
log (Non-performing business debt)	0.244 ***	(0.080)	0.251 ***	(0.083)					0.239 ***	(0.084)
log (Business debt)	0.023	(0.017)	0.040 **	(0.018)					0.111 ***	(0.018)
log (Business account balance)					-0.242 ***	(0.009)	-0.238 ***	(0.010)	-0.251 ***	(0.010)
log (Business net cash flow)					-0.011 ***	(0.004)	-0.014 ***	(0.004)	-0.013 ***	(0.004)
Firm age	No		Yes		No		Yes		Yes	
Firm size fixed effects	No		Yes		No		Yes		Yes	
Month fixed effects	Yes		Yes		Yes		Yes		Yes	
Firm industry fixed effects	Yes		Yes		Yes		Yes		Yes	
Firm location fixed effects	Yes		Yes		Yes		Yes		Yes	
Adj. R2	**0.038**		**0.043**		**0.069**		**0.071**		**0.077**	
In-sample AUC	**0.656**		**0.670**		**0.727**		**0.729**		**0.739**	
Difference to AUC=50%	**0.156 *****		**0.170 *****		**0.227 *****		**0.229 *****		**0.239 *****	

Adding additional features

Table IV
Overall predictive power of entrepreneur personal features

This table estimates a standard logit model where the dependent variable ($Default$ (0/1)) is equal to one if the loan is in delinquency over 30 days. Column 1 estimates the predictive power of all entrepreneur personal behaviors. Column 2 include both the business-specific features and all personal behaviors. Column 3 and Column 4 further include fixed effects. Standard errors are clustered by month, industry, and location. *, **, and *** represent statistical significance at the 10%, 5%, and 1% levels, respectively.

	(1) All personal behaviors		(2) Business-specific features & All personal behaviors		(3) All personal behaviors & Fixed effects		(4) Business-specific features & All personal behaviors & Fixed effects	
Variables	Coef.	St.D.	Coef.	St.D.	Coef.	St.D.	Coef.	St.D.
No. related business entities			0.034 ***	(0.010)			0.035 ***	(0.010)
Bad business credit			0.189 ***	(0.053)			0.191 ***	(0.054)
log (Non-performing business debt)			0.194 **	(0.098)			0.182 *	(0.103)
log (Business debt)			0.162 ***	(0.020)			0.194 ***	(0.021)
log (Business account balance)			-0.171 ***	(0.012)			-0.178 ***	(0.013)
log (Business net cash flow)			-0.007 *	(0.004)			-0.006	(0.004)
log (Age)	1.055 ***	(0.162)	1.163 ***	(0.164)	1.001 ***	(0.164)	1.077 ***	(0.167)
Gender (female)	0.025	(0.033)	0.025	(0.033)	0.023	(0.034)	0.025	(0.033)
Education (postgraduate)	0.110	(0.087)	0.140	(0.091)	0.079	(0.088)	0.093	(0.091)
Education (bachelor)	0.051	(0.045)	0.056	(0.045)	0.038	(0.046)	0.041	(0.046)
Education (junior college)	-0.123	(0.082)	-0.134	(0.083)	-0.096	(0.083)	-0.100	(0.084)
Marital (married)	-0.151 ***	(0.039)	-0.145 ***	(0.039)	-0.139 ***	(0.040)	-0.130 ***	(0.039)
Birthplace level (large)	0.001	(0.040)	0.037	(0.040)	0.026	(0.041)	0.046	(0.041)
Local resident	-0.043	(0.035)	-0.078 **	(0.036)	-0.058	(0.037)	-0.076 **	(0.037)
Consumption ratio	0.994 ***	(0.101)	0.904 ***	(0.101)	0.911 ***	(0.105)	0.811 ***	(0.105)
Discretionary consumption	0.536 ***	(0.111)	0.525 ***	(0.114)	0.514 ***	(0.111)	0.504 ***	(0.115)
Housing	-1.084 ***	(0.198)	-1.105 ***	(0.203)	-1.090 ***	(0.202)	-1.108 ***	(0.207)
Traveling	2.691 *	(1.399)	2.513 *	(1.426)	2.838 **	(1.142)	2.645 *	(1.443)
Gambling	5.358 *	(2.917)	5.032 *	(2.588)	6.725 **	(2.830)	6.334 **	(2.625)
Cash spending	0.766 ***	(0.269)	0.853 ***	(0.267)	0.424	(0.284)	0.514 *	(0.283)
Work hour consumption	0.866 ***	(0.220)	0.871 ***	(0.220)	0.893 ***	(0.226)	0.949 ***	(0.227)
Latest hour	0.239 ***	(0.029)	0.232 ***	(0.029)	0.230 ***	(0.030)	0.227 ***	(0.030)
log (Personal account balance)	-0.288 ***	(0.019)	-0.297 ***	(0.019)	-0.309 ***	(0.020)	-0.325 ***	(0.021)
Stock market participation	-0.164 ***	(0.047)	-0.158 ***	(0.048)	-0.145 ***	(0.048)	-0.140 ***	(0.048)
Share of stock investment	-0.933 *	(0.510)	-0.887 *	(0.508)	-0.937 **	(0.512)	-0.922 *	(0.510)
log (Length of relationship)	-0.107 ***	(0.019)	-0.115 ***	(0.019)	-0.096 ***	(0.019)	-0.102 ***	(0.020)
Length of overdue	0.363 ***	(0.114)	0.373 ***	(0.115)	0.370 ***	(0.130)	0.381 ***	(0.130)
log (Amount overdue)	0.037 ***	(0.008)	0.035 ***	(0.008)	0.037 ***	(0.008)	0.035 ***	(0.008)
Credit card delinquency	0.934 ***	(0.251)	0.948 ***	(0.249)	0.904 ***	(0.252)	0.898 ***	(0.253)
log (Total credit limit)	-0.024 ***	(0.009)	-0.027 ***	(0.009)	-0.027 ***	(0.009)	-0.030 ***	(0.009)
log (Max credit limit)	-0.075 ***	(0.010)	-0.065 ***	(0.010)	-0.073 ***	(0.010)	-0.065 ***	(0.010)
log (Min credit limit)	0.022 *	(0.012)	0.020 *	(0.012)	0.021 *	(0.012)	0.020 *	(0.012)
log (Credit utilization)	0.035 ***	(0.006)	0.029 ***	(0.006)	0.035 ***	(0.006)	0.030 ***	(0.006)
No. creditors	0.213 ***	(0.012)	0.187 ***	(0.012)	0.219 ***	(0.012)	0.196 ***	(0.012)
Firm age	No		Yes		No		Yes	
Firm size fixed effect	No		Yes		No		Yes	

Out-of-sample roc

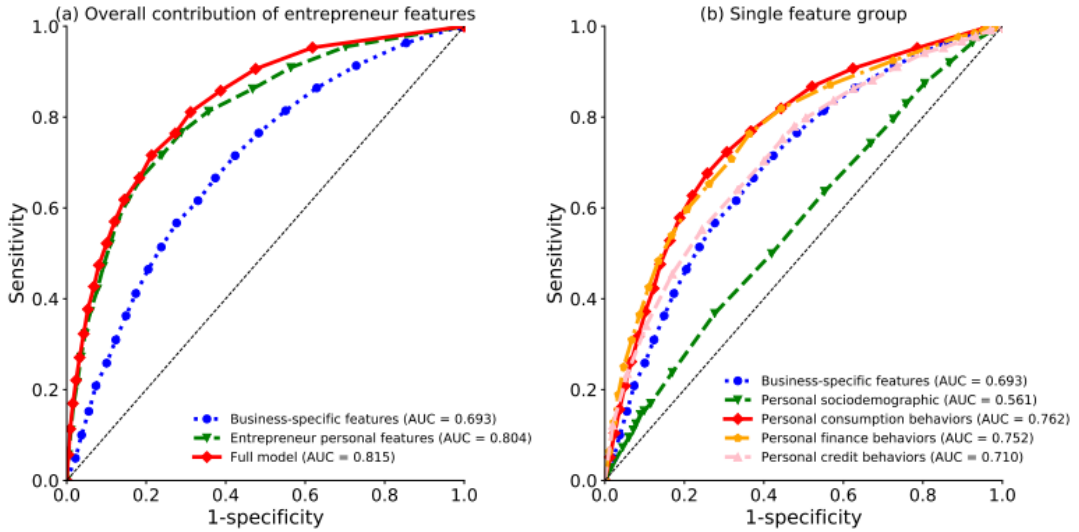


Figure 2. Out-of-sample ROC curves under standard logistic regression model

This figure presents the contribution of entrepreneur personal features to predicting business credit performance by providing the receiver operating characteristics curves (ROC curves). Subfigure (a) compares the predictive power of business-specific features and entrepreneur personal features, and Subfigure (b) compares the predictive power of each dimension of entrepreneur personal behaviors along with the benchmark model.

ML-Comparison

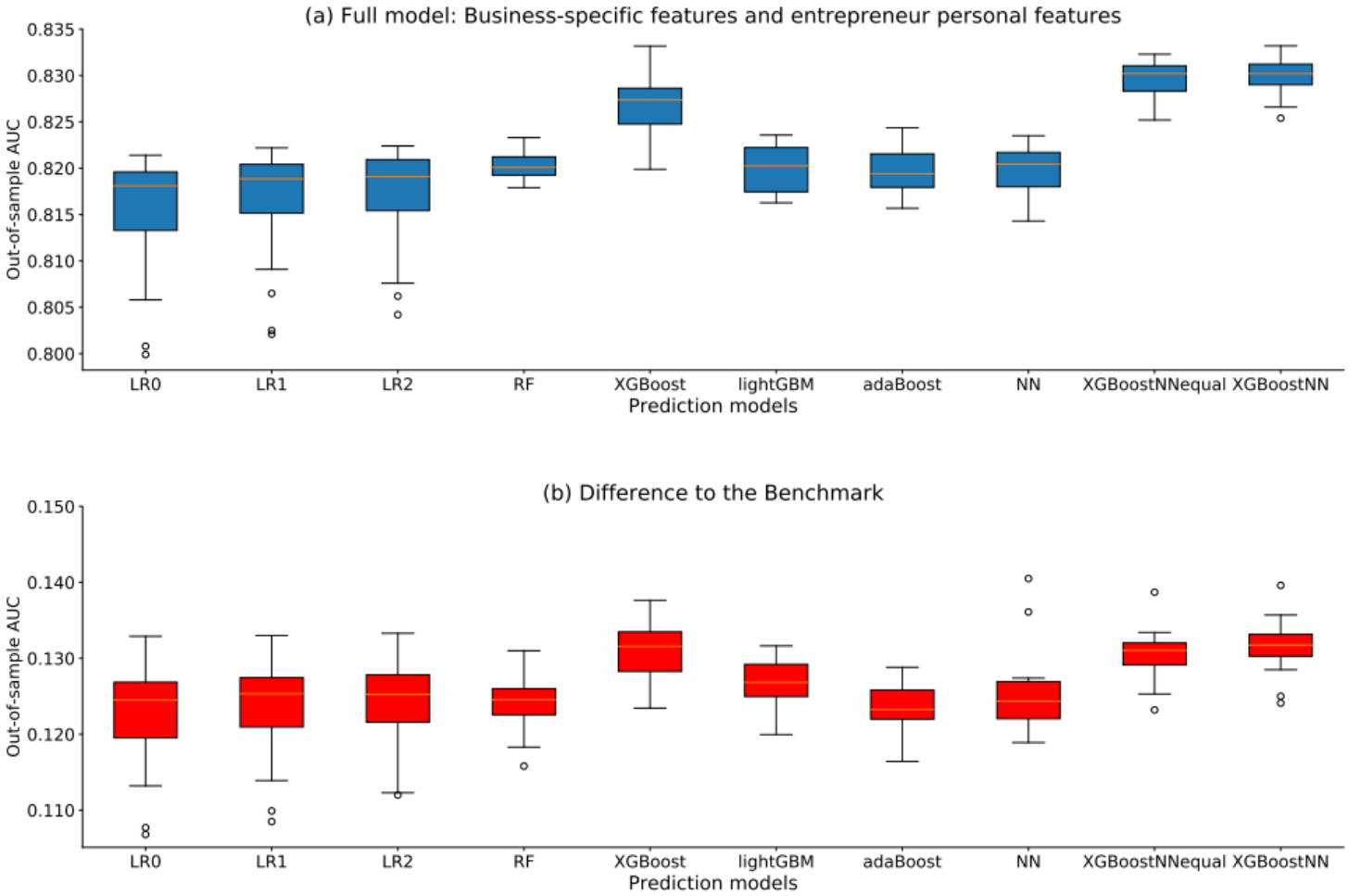


Figure 4. Comparison of out-of-sample AUC across models

Default-Comparison

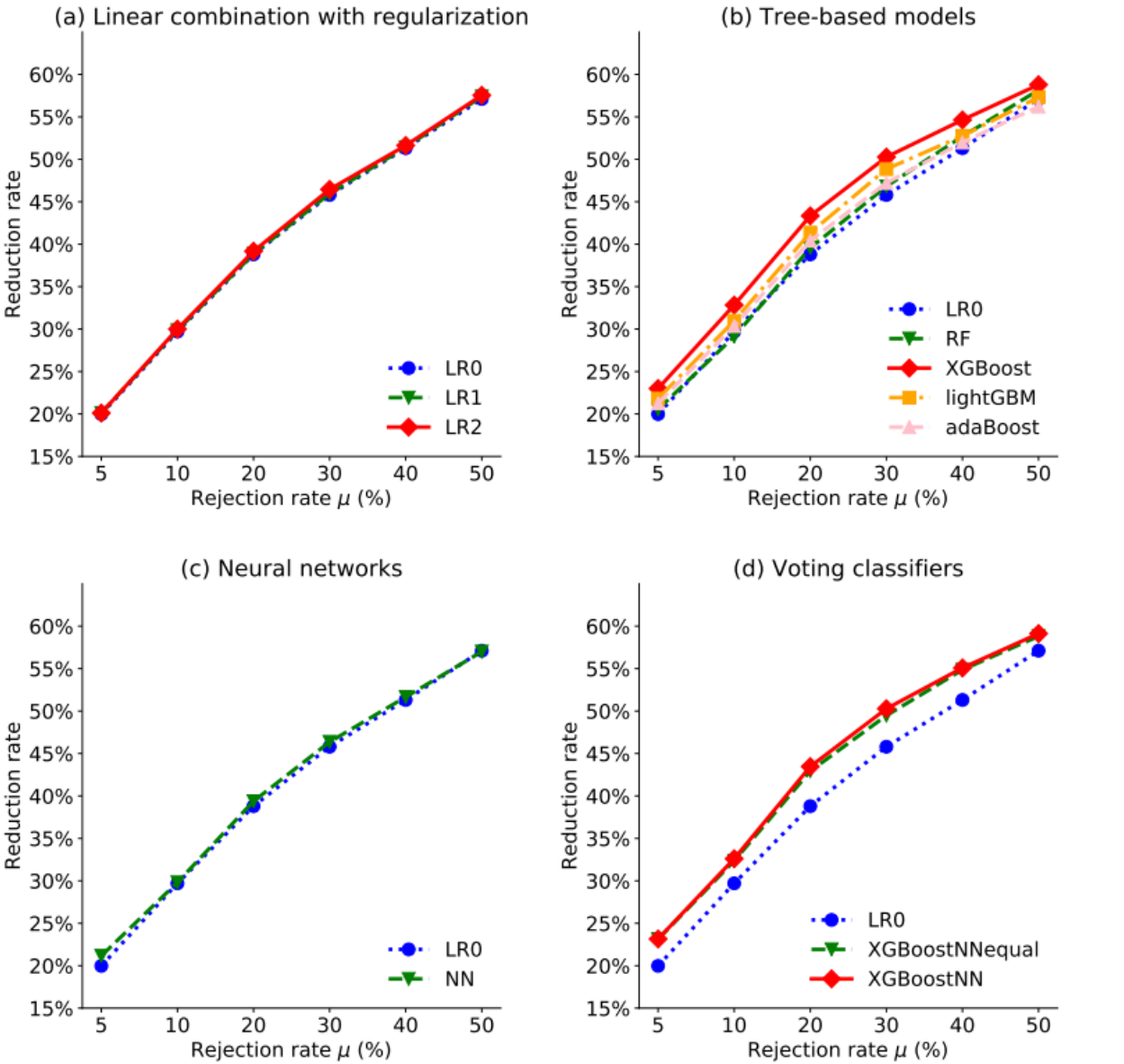


Figure 5. Relative default reduction rate across models

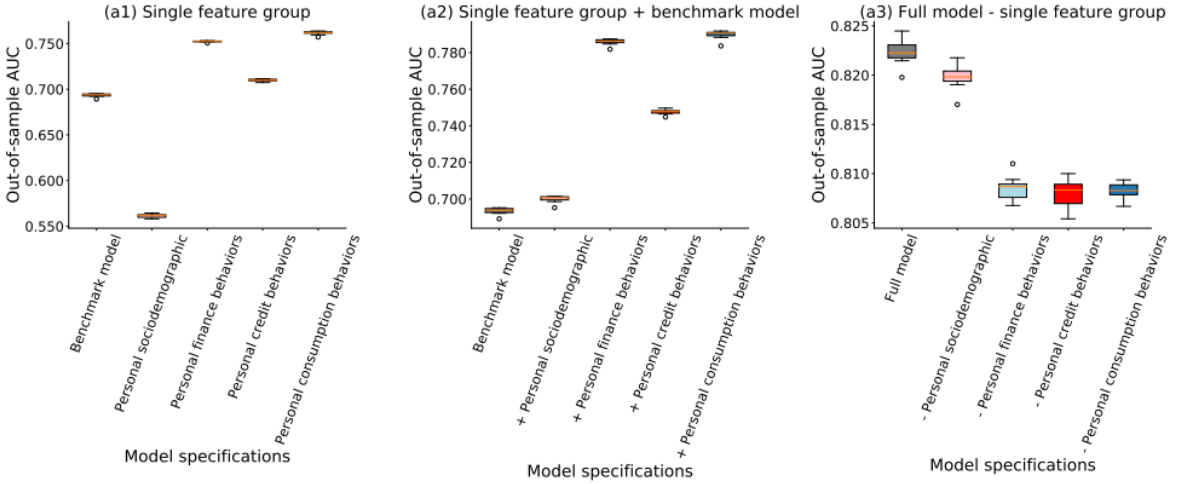
Prediction-Power-Comparison

Table V
Predictive power across machine learning models

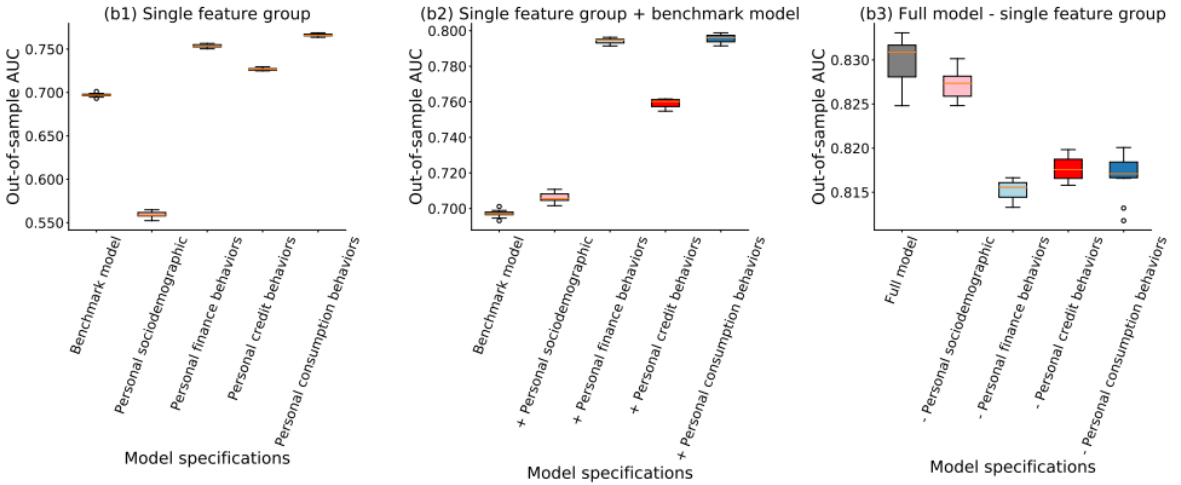
This table compares the prediction performance (measured by AUC, the area under curves) of ten machine learning models. Column 1 reports the AUCs of the benchmark model that uses only business-specific features. Column 2 reports the AUCs of the full model that uses both the business-specific features and entrepreneur personal features. Column (3) reports the difference in AUC between Columns 2 and 1. Panel (a) shows the out-of-sample results, and Panel (b) shows the out-of-sample/out-of-time results. Numbers in parenthesis are the standard deviations of AUCs.

Model	(1) Business-specific features	(2) Business-specific features & entrepreneur personal features	(3) Difference in AUC
Panel A. out-of-sample			
Logistic regression (LR)	69.30% (0.21%)	81.53% (0.67%)	12.23%
I. Linear combination with regularization			
Logistic regression with l_1 regularization (LR1)	69.30% (0.21%)	81.62% (0.64%)	12.32%
Logistic regression with l_2 regularization (LR2)	69.30% (0.21%)	81.68% (0.58%)	12.38%
II. Tree-based models			
Random forests (RF)	69.61% (0.30%)	82.02% (0.15%)	12.41%
eXtreme Gradient Boosting (XGBoost)	69.61% (0.20%)	82.68% (0.32%)	13.07%
Light Gradient Boosting Machine (lightGBM)	69.34% (0.19%)	82.02% (0.25%)	12.68%
Adaptive Boosting (AdaBoost)	69.65% (0.21%)	81.99% (0.26%)	12.34%
III. Neural networks			
Dropout Neural Networks (NN)	69.47% (0.42%)	81.99% (0.25%)	12.52%
IV. Voting classifier			
XGBoost + Dropout NN with equal voting weights (XGBoostNNEqual)	69.91% (0.23%)	82.96% (0.20%)	13.05%
XGBoost + Dropout NN with different voting weights (XGBoostNN)	69.86% (0.27%)	83.01% (0.21%)	13.15%
Panel B. out-of-sample/out-of-time			
Logistic regression (LR)	69.80% (0.71%)	81.00% (1.09%)	11.20%
I. Linear combination with regularization			
Logistic regression with l_1 regularization (LR1)	69.81% (0.71%)	81.45% (1.00%)	11.65%
Logistic regression with l_2 regularization (LR2)	69.82% (0.71%)	81.54% (0.96%)	11.73%
II. Tree-based models			
Random forests (RF)	69.72% (1.38%)	82.50% (0.60%)	12.78%
eXtreme Gradient Boosting (XGBoost)	70.08% (0.86%)	82.26% (0.64%)	12.18%
Light Gradient Boosting Machine (lightGBM)	69.90% (0.94%)	81.47% (0.74%)	11.57%
Adaptive Boosting (AdaBoost)	70.18% (2.37%)	81.83% (0.61%)	11.65%
III. Neural networks			
Dropout Neural Networks (NN)	68.52% (2.22%)	81.96% (0.68%)	13.44%
IV. Voting classifier			
XGBoost + Dropout NN with equal voting weights (XGBoostNNEqual)	70.09% (1.05%)	82.88% (0.57%)	12.80%
XGBoost + Dropout NN with different voting weights (XGBoostNN)	70.03% (0.94%)	82.80% (0.62%)	12.77%

Feature-Comparison



(a) Standard logistic regression



(b) XGBoostNN

Figure 6. Contribution of each feature group of entrepreneur personal behaviors

Feature-Importance

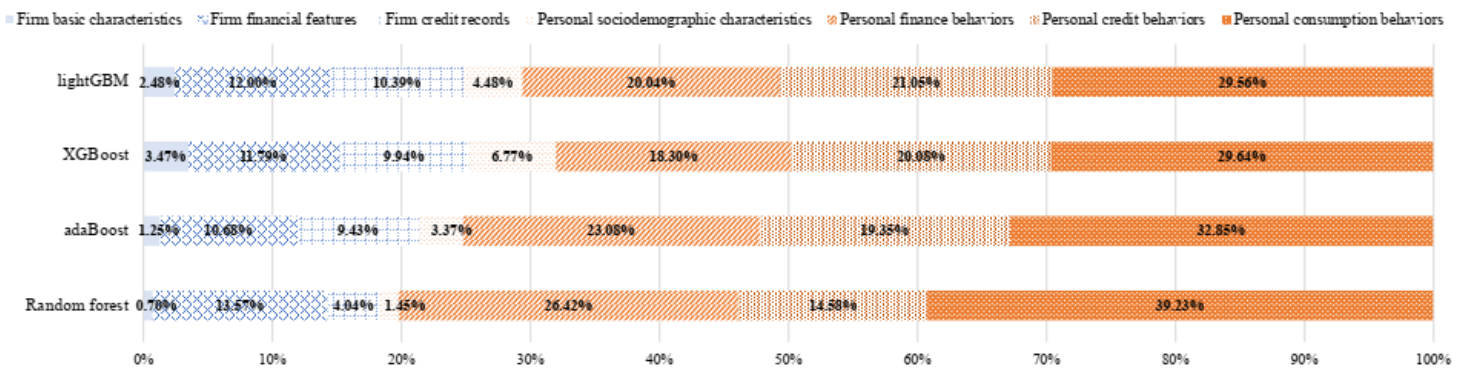
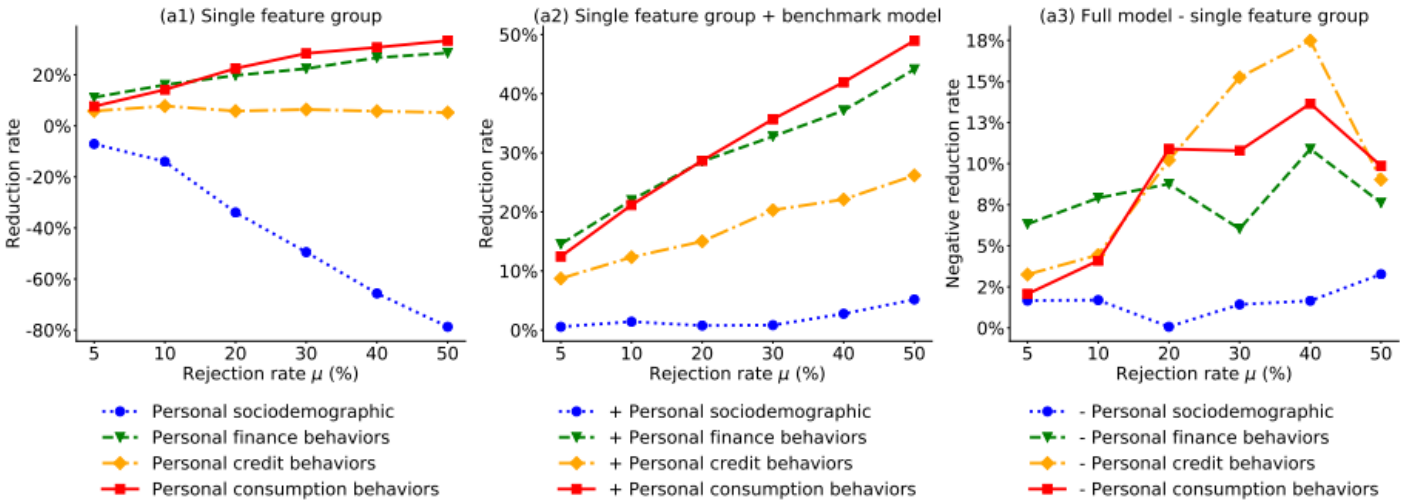
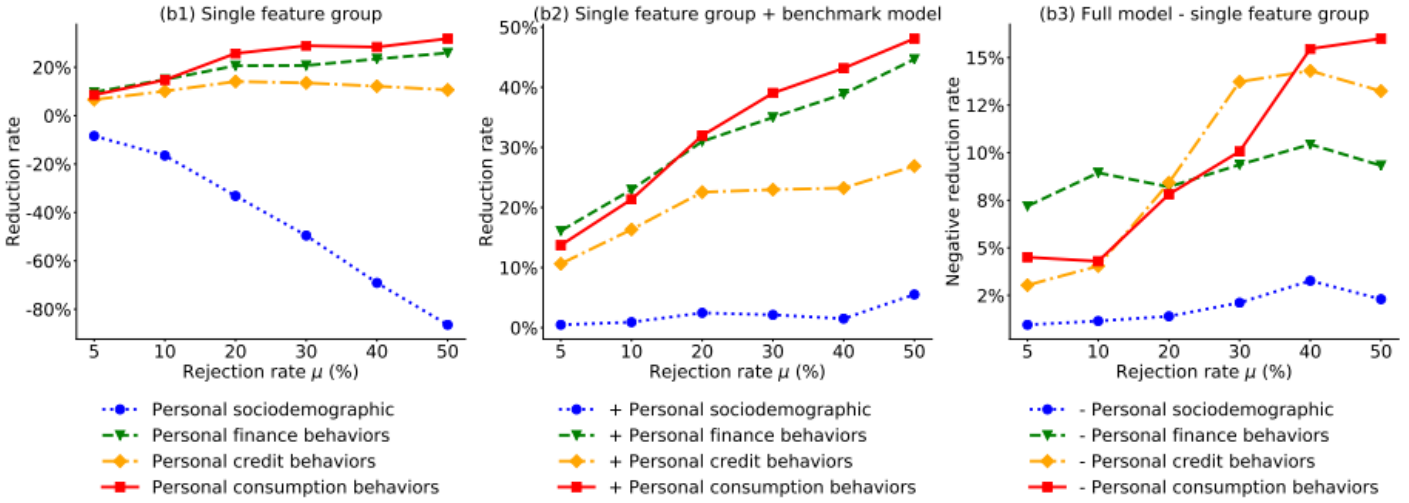


Figure 8. Feature importance of feature groups

Reduction in Default by features



(a) Standard logistic regression



(b) XGBoostNN

Figure 7. Default reduction by each feature group of entrepreneur personal behaviors

Thank you for your participation!