Disagreement between Human and Machine Predictions

Oct 2nd, 2020
@Keio University
Applied Economics Workshop

Daisuke Miyakawa (Hitotsubashi)
Kohei Shintani (Bank of Japan)

*This is the joint work with TOKYO SHOKO RESEARCH, LTD. (TSR) and conducted during one of the authors staying at BOJ-IMES. Usual disclaimer is applied.

Background

- Prediction tasks
  - E.g., firm exit, financial markets, macro, etc.
  - Better prediction ⇒ Better decision
- Machine learning (ML) methods
  - Using high dimensional information “mainly” for prediction
  - Varian ’14, Mullainathan & Spiess ’17, Athey ‘19

Background (cont'd)

- Use ML for prediction
  - Successful
    - Labor: Chalfin et al. '16
    - Public: Kleinberg et al. '18, Bazzi et al. '19, Lin et al. '20
    - Medical: Patel et al. '19, Mei et al. '20
    - Financial: Agrawal et al. '18
  - "ML $\succ$ Human" on average ($\Leftrightarrow$ They disagree)

Research question

- Any **systematic pattern** in the **disagreement**?

  - Informative to understand human **AND** machine errors
    - E.g., informational opaqueness
    - Can "ML $\prec$ Human" be the case?
      - $\Rightarrow$ ***Yes*** (economist view): Signal extraction from soft info
      - $\Rightarrow$ ***No*** (psychologist view): Noisy prediction
        - $\Leftrightarrow$ Kleinberg et al. '18: ML $\succ$ "Predicted" judge $\succ$ Judge

  - Useful for **task allocation**
    - General computerization: Frey & Osborne '13
    - Automation: Acemoglu & Restrepo '18

What we are doing

A) Construct a ML-based prediction model

B) Measure the disagreement b/w ML & Human

C) Examine how opaqueness works as its determinants

D) Do a counterfactual exercise for task allocation

What we are ***NOT*** doing

A) Inventing a new ML algorithm

B) Studying other than business enterprises

C) Studying other than credit rating

D) Causal impact of the introduction of ML score

- Paravisini & Schoar ‘15, Hoffman et al.’18

Key takeaways

- “ML $\succ$ Human” on average
  - Highly robust against many concerns

- “ML $\succ$ Human $\succ$ Predicted human”
  - ≠ Kleinberg et al. (***QJE*** ‘18) and supporting economists’ view

- Relative performance of H/M ↑ as firms opaqueness↑
  - Highly robust against many concerns

- “ML $\prec$ Human” could be the case when…
  i. Firms are very opaque
  ii. Type I error is more concerned (than Type II error is)

Contribution

- First to study H-M disagreement in social science
  - Raghu et al. '19: Algorithmic triage for diabetic retinopathy
    (≠ Anderson et al. '17, McIlroy-Young '20 for "chess")

- This is mainly because...
  - Data limitation on human prediction
  - Data limitation on target attributes
  - Data limitation on "human" (⇒ severe omitted variable issues)
    ⇔ E.g., Kleinberg et al. '18: No judge attributes
  - Selection label problem
    ⇒ Not the case in our data

⇒ **When we should/shouldn't use ML?** (≠ Luca et al. '16)

Organization

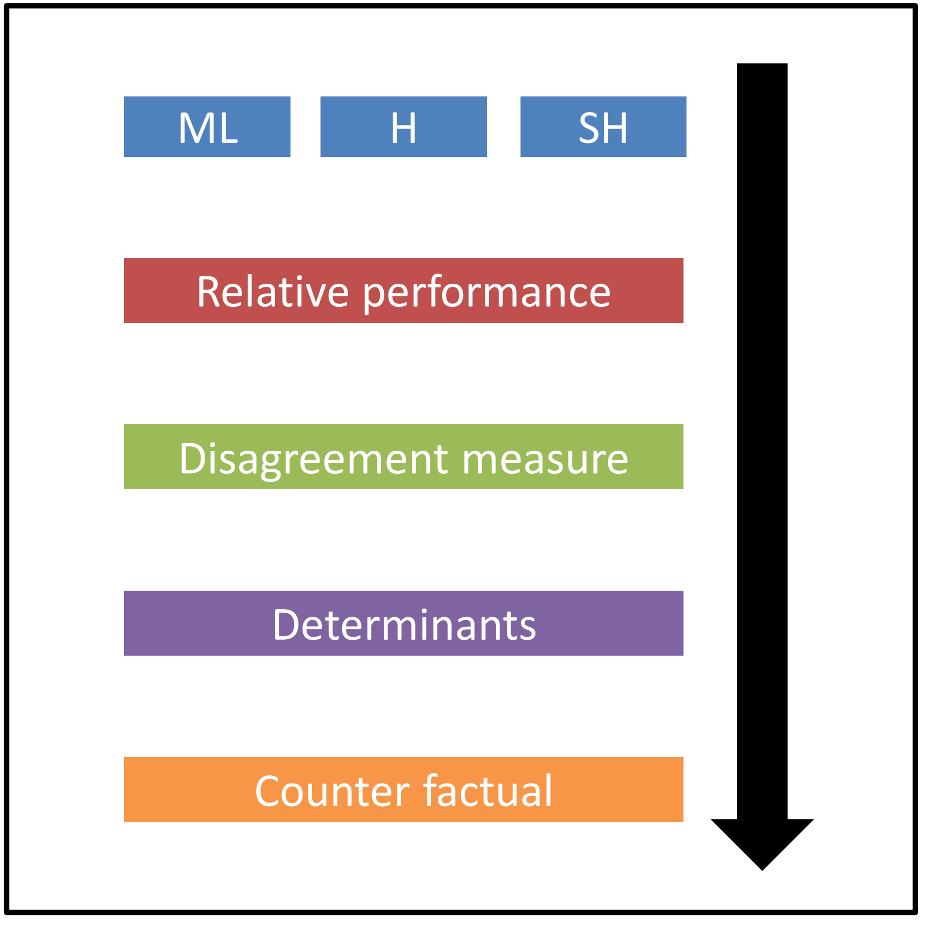
1. **Theoretical illustration**

2. Methodology

3. Data

4. Results

5. Summary



ML H

1. Theoretical illustration

- Ground truth for an instance f : $a(f)$

- Prediction: $m(f)$ by M & $h(f, i)$ by H(i)

- Prediction errors

$$\Theta(f) = L(a(f), m(f))\text{: M}$$

$$\Omega(f, i) = L(a(f), h(f, i))\text{: H}$$

1. Theoretical illustration

ML | H | SH

- Relative error rate of H to M: Our main interest

$$Proxy_{f,i} = \Omega(f,i) - \Theta(f).$$

- Structure human's prediction & proxy: Also examined
  - Human prediction solely ∝ observable info

$$\Omega_h(f)$$

$$Proxy'_{f,i} = \Omega(f,i) - \Omega_h(f)$$

ML H

1. Theoretical illustration

- "Ultimate" goal:

$$\min_{S,T} \sum_{f \in S} \Theta(f) + \sum_{f \in T} \Omega(f, i)$$
$$\text{s.t. } S \cup T = U; S \cap T = \emptyset$$

$\Rightarrow$ (S^*, T^*) as a function of (f, i)

$\Rightarrow$ Main interest: Info opaqueness as the determinants + other control variables

$\Leftrightarrow$ We achieve this through CF exercises

Organization

1. Theoretical illustration
2. **Methodology**
3. Data
4. Results
5. Summary

2-1. Method: ML-prediction

ML

- Target of the prediction (outcome):
  - *1(Dynamics)* in default & voluntary closure & sales growth

- Predictors
  - #(independent variables) > 200: Observed before the dynamics

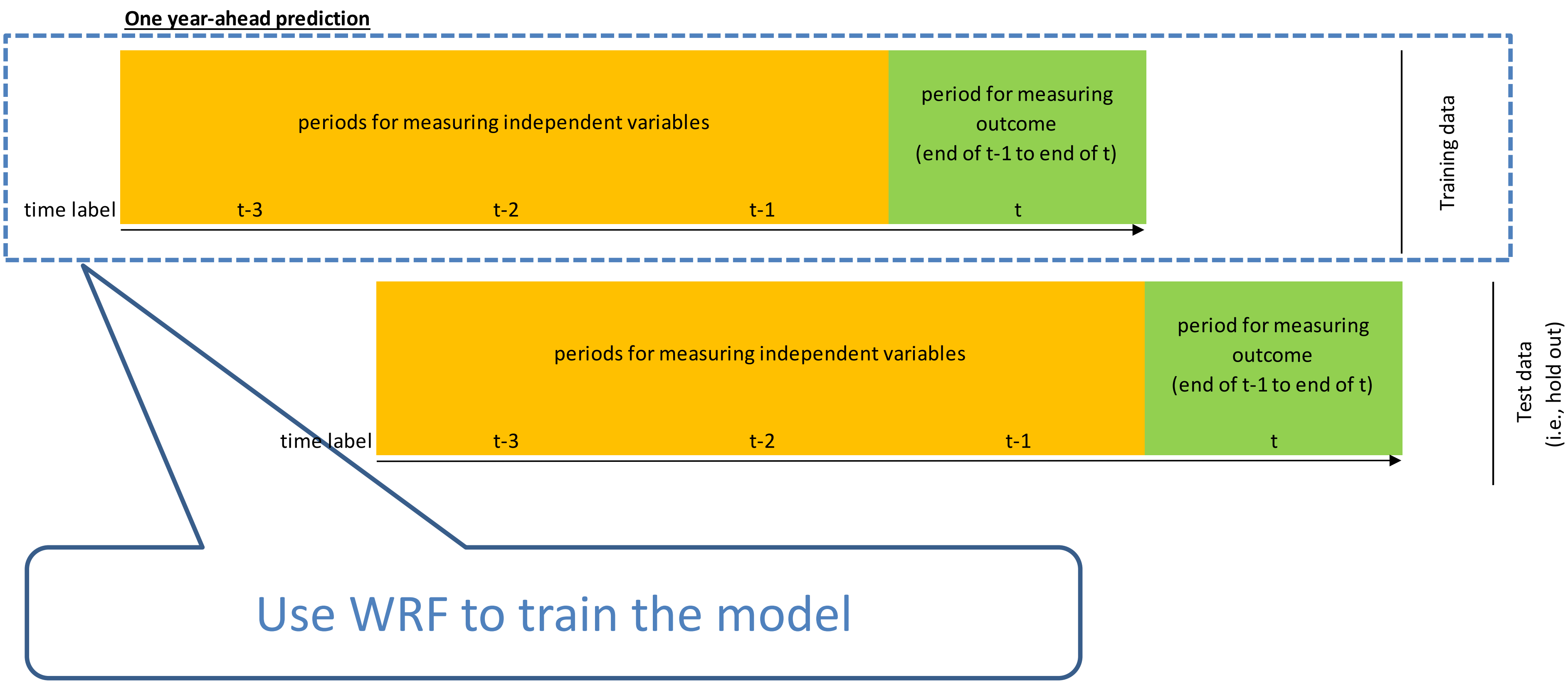
  - 6 groups of variables
    - Firms' basic attributes (***firmown***)
    - Detailed financial statement information (***kessan***)
    - Geographical/industry information (***geo/ind***)
    - Bank relation (***bank***)
    - Customer-supplier relation (***network***)
    - Shareholder information (***share***)

2-1. Method: ML-prediction

ML

□ “Training”

Prediction w/ machine learning（weighted random forest: WRF）



ML

＜Random forest＞

☐ Tree prediction

- Category (outcome) & attributes
- Discretize
- Compute the information gain associated with the “creation” of a splitting rule (i.e., “edge”) at each node
  - Criterion: Entropy, Gini
- Root (starting point)

  →At each node, create a tree/edge by referring to the best splitting rule among all the attributes and the thresholds

  →Repeat→・・・→Terminal node (“leaf” only consisting of P/N)

☐ Random forest

- Bootstrap the data and do the tree prediction for each data
- Assemble (e.g., majority) the decisions and decide the tree

ML

＜”Weighted” Random forest＞

□ Chen et al. (2004)

- Imbalance problem

- (i) Sampling technique

- (ii) Penalizing misclassification ← Weighted R.F.

  - Weighting minority class more during the search of tree structure

  - Weighting the leaf corresponding to the minority class when deciding the final tree structure

  - Class weight (hyper-parameter) is determined through out-of-bag estimate（i.e., accuracy test based on the data not sued in bootstrapping)

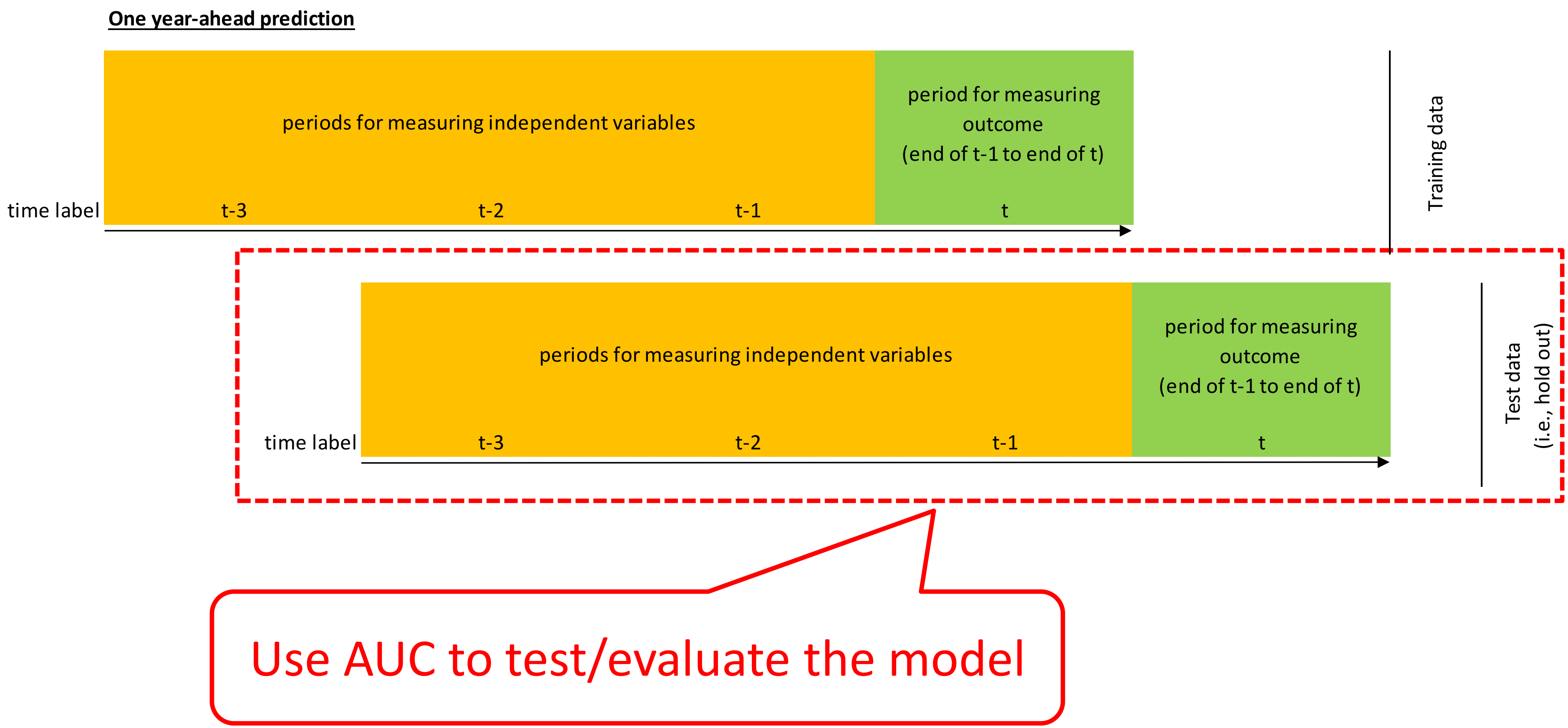
2-1. Method: ML-prediction

ML

Relative performance

- "Test" using hold-out data

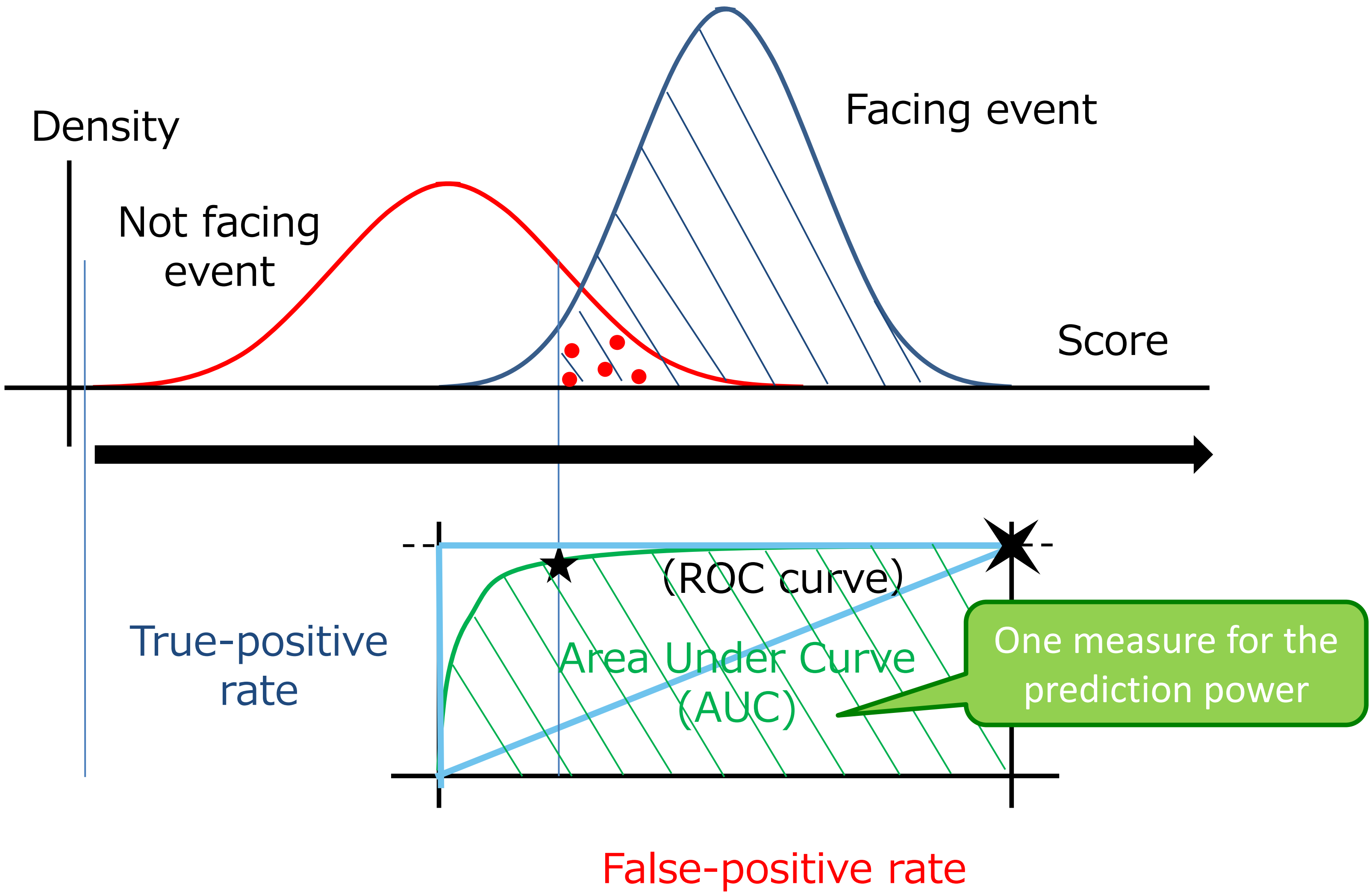
Evaluate the prediction power ⇒ ROC curve, and AUC



＜Evaluation: ROC curve & AUC＞

ML

Relative performance



2-2. Method: Human-Prediction

H

- Target of the prediction (outcome):
  - ***1(Dynamics)*** in default & voluntary closure & sales growth

- Predictors
  - Human
    - Widely used creditworthiness score: ***fscore***
    - Also, use the sub-scores for ***fscore***

      ⇒ 4 sub-scores: CEO, growth opportunity, stability, openness
  - Calibrate by Probit (with oversampled positive data)

Some immediate concerns

ML H SH

a. Credit ratings as human prediction?
   - Mixture of rule-based scoring & discretion
   - Also, compare it with “structure” human

b. Same information used by Human & ML?
   - Trying to make it comparable by reducing the info for ML
   - Still, room for Human to use soft/private info (our interest)

c. Omitted payoff bias?
   - Use sub-scores

Some immediate concerns

ML H

d. (Calibrated) score?
   - Rank-based analysis

e. Other ML methods (LASSO and XGB)?

f. Structural change?
   - ML $\succ$ H on average is confirmed for all test years

2-3. Method: “Structured” Human

SH

- Construct a model for replicating human decision (SH)
  - WRF
  - Economist view vs. psychologist view
  - We can specify the information set used for the prediction

    ⇒ Use this prediction instead of ML in our analysis

- Target of the prediction (outcome):
  - ***fscore***

- Predictors
  - #(independent variables) > 200
  - The 6 groups of variables
    - ***firmown***, ***kessan***, ***geo/ind***, ***bank***, ***network***, ***share***

2-4. Method: Disagreement

- *Proxy*: Measure the disagreement
  - Predict firms' outcome with test data by M & H & SH
    - Predicted outcomes for each company (between 0 and 1)
    - Larger means the company is more likely to face an event
    - *"t"* is addeted to the subscript
  - Normalize predicted outcomes for each model

$$Outcome_{f,t}^{ML} \ \& \ Outcome_{f,i,t}^{H} \ \& \ Outcome_{f,t}^{SH}$$

2-4. Method: Disagreement

Disagreement measure

- *Proxy*: Measure the disagreement

  - Large ⇔ M or SH $\succ$ H

  - M vs H

$$Proxy_{f,i,t} = Outcome^{ML}_{f,t} - Outcome^{H}_{f,i,t} \text{ for exit firms}$$
$$= Outcome^{H}_{f,i,t} - Outcome^{ML}_{f,t} \text{ for non-exit firms}$$

  - SH vs H

$$Proxy'_{f,i,t} = Outcome^{SH}_{f,t} - Outcome^{H}_{f,i,t} \text{ for exit firms}$$
$$= Outcome^{H}_{f,i,t} - Outcome^{SH}_{f,t} \text{ for non-exit firms}$$

2-5. Method: Determinants

Identifying the determinants

- Firm-Analyst-time level Panel estimation:

$$Proxy_{f,i,t} = G\left(\boldsymbol{O}_{f,t}, \boldsymbol{F}_{f,t}, \boldsymbol{I}_{i,t}, \boldsymbol{Z}_{i,t}\right) + \boldsymbol{\eta}_{f,i,t} + \varepsilon_{f,i,t}$$

where

$\boldsymbol{O}_{f,t}$: Firm (i.e., target of scoring)' informational opaqueness

$\boldsymbol{F}_{f,t}$: Firm (i.e., target of scoring)-attribute

$\boldsymbol{I}_{i,t}$: Analyst (i.e., human making score)- attribute

$\boldsymbol{Z}_{i,t}$: Team- attribute

$\boldsymbol{\eta}_{f,i,t}$: Fixed-effects

Organization

1. Theoretical illustration
2. Methodology
3. **Data**
4. Results
5. Summary

3-1. Data: Overview

- TSR data: 1M+ firms/year

  Similar to D&B in the U.S.

  - KJ: Basic firm attributes, bank relation, shareholding
  - SK: Supply chain network information
  - KESSAN: Financial statement information
  - Firm-Analyst table & HR data
  - Exit frag: Default, voluntary exit
  - t = 2010-1016 (t = 2017- in lockbox)

- Split the data to training & test (i.e., hold-out) data
  - One-year ahead predictions
  - Also, setting up the “lock box”

3-2. Data: Selection label problem?

- One typical issue in the comparison of prediction power

  - Outcomes might be recorded for a limited #(obs), which makes it difficult to compare machine- and human predictions

    - E.g., crime record is recorded only for released defendants
      ⇔ Kleinberg et al. '18

    - E.g., teaching performance is recorded only for hired teachers
      ⇔ Jacob et al. '18

  - We **do not have** this issue as TSR put scores for all firms and we can observe the default for all those firms

3-3. Data: Summary

Variable	Definition	#samples	min.	25%tile	median	mean	75%tile	max	sd
Disagreement									
$Proxy_{f,i,t}$	Relative performance of machine predictions for firm *f*. The larger (smaller) value means that machine (analyst *i*) can predict outcome better.	3,983,158	-5.066	-0.95	-0.09	0.00	0.89	5.62	1.29
$structured\ fscore_{f,t}$	Firm *f*'s hypothetical *fscore* considered as analysts could use only hard information for predictions. It is calculated as a replication of *fscore* by machine prediction method.	3,983,158	19.300	43.27	46.19	46.82	49.66	80.95	5.26
Number of available variables									
$\#(available\ variables)_{f,t}$	The number of firm *f*'s hard information available for predictions.	3,983,158	10	38.00	80.00	91.02	132.00	276	60.42
Firm Characteristics									
$\log(sales_{f,t})$	The logarithm of firm *f*'s gross sales.	3,983,158	0.000	10.29	11.29	11.37	12.41	23.92	1.86
$\log(sales_{f,t})-\log(sales_{f,t-1})$	Log change in firm *f*'s gross sales.	3,983,158	-14.230	-0.06	0.00	0.01	0.07	12.73	0.36
$\#(industry)_{f,t}$	The number of industry codes which are assigned to firm *f*. It takes values from 1 to 3.	3,983,158	1	1.00	2.00	1.92	3.00	3	0.85
$priority_{f,t}$	Firm *f*'s relative importance for analysts.	3,810,937	0	0.00	2.00	14.76	8.00	41,290	75.80
$fscore_{f,t}$	A score that summarizes an overall performance of firm *f* provided by TSR. It takes values from 0 to 100.	3,983,158	0	43.00	46.00	46.82	50.00	88	5.91
Analyst Characteristics									
$\#(tenure\ years)_{i,t}$	Analyst *i*'s length of serveice.	3,503,183	0.003	3.59	8.05	10.51	15.38	43.620	8.67
$\#(assigned\ companies)_{i,t}$	The number of companies for which analyst *i* is responsible to make *fscore*.	3,810,987	1	610	939	1,516	1,862	11,570	1,684.70
$industry\ experience_{f,i,t}$	The number of companies (1) having the same industry codes as firm *f*, and (2) having been responsible for analyst *i* to make *fscore* for recent 3 years.	3,810,987	1	27.00	85.00	263.60	271.00	6,241	515.25
Team Characteristics									
$\#(team\ members)_{i,t}$	The number of colleagues belonging to the same division as analyst *i*.	3,495,647	0	8.00	13.00	15.02	20.00	119	9.70
Average $\#(tenure\ years)_{i,t}$	Average length of service across team members including analyst *i*.	3,466,648	0.504	7.50	9.76	10.35	12.72	37.19	4.18
Average $industry\ experience_{f,i,t}$	Average industry experience across team members including analyst *i*.	3,466,648	0	25.67	60.33	117.60	162.30	883.00	136.57
Average $\#(assigned\ companies)_{i,t}$	Average number of assigned companies across the team members including analyst *i*.	3,466,648	1	920.20	1,230.00	1,407.00	1,877.00	3,543	679.30

Organization

1. Theoretical illustration
2. Methodology
3. Data
4. **Results**
5. Summary

4-1. Result: ML ≻ Human?

Relative performance

- Default & Closure
- Economist vs. psychologist
  - Default: Econ
  - Closure: Psy

Table 2: AUC

Test data: $t = 2013$

Model	default	voluntary closure
Human	0.634 (0.0049)	0.719 (0.0030)
Machine	0.793 (0.0041)	0.828 (0.0024)
Structured human	0.617 (0.0046)	0.749 (0.0027)
Machine & *fscore*	0.807 (0.0040)	0.829 (0.0023)
Machine with small information	0.777 (0.0044)	0.829 (0.0024)

Test data: $t = 2014$

Model	default	voluntary closure
Human	0.639 (0.0052)	0.729 (0.0031)
Machine	0.780 (0.0045)	0.828 (0.0024)
Structured human	0.622 (0.0049)	0.757 (0.0028)
Machine & *fscore*	0.794 (0.0043)	0.830 (0.0024)
Machine with small information	0.765 (0.0048)	0.829 (0.0024)

Test data: $t = 2015$

Model	default	voluntary closure
Human	0.653 (0.0055)	0.737 (0.0031)
Machine	0.786 (0.0045)	0.833 (0.0024)
Structured human	0.638 (0.0052)	0.766 (0.0028)
Machine & *fscore*	0.799 (0.0044)	0.835 (0.0024)
Machine with small information	0.768 (0.0050)	0.834 (0.0025)

Test data: $t = 2016$

Model	default	voluntary closure
Human	0.663 (0.0053)	0.748 (0.0031)
Machine	0.773 (0.0045)	0.841 (0.0025)
Structured human	0.648 (0.0050)	0.776 (0.0027)
Machine & *fscore*	0.789 (0.0044)	0.843 (0.0025)
Machine with small information	0.758 (0.0049)	0.843 (0.0024)

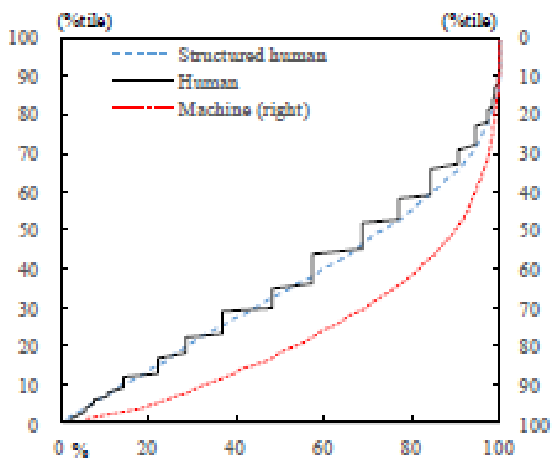
4-2. Result: H vs. SH?

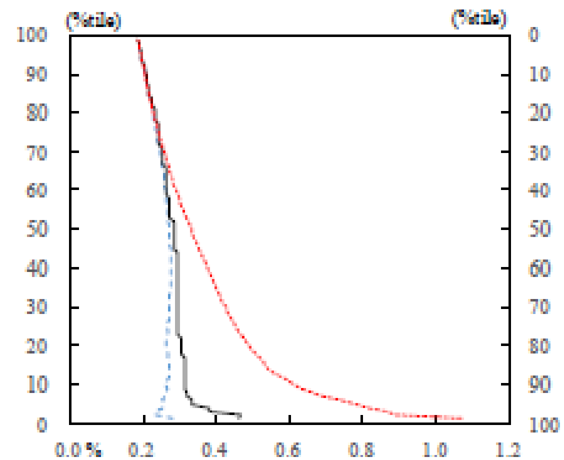
Relative performance

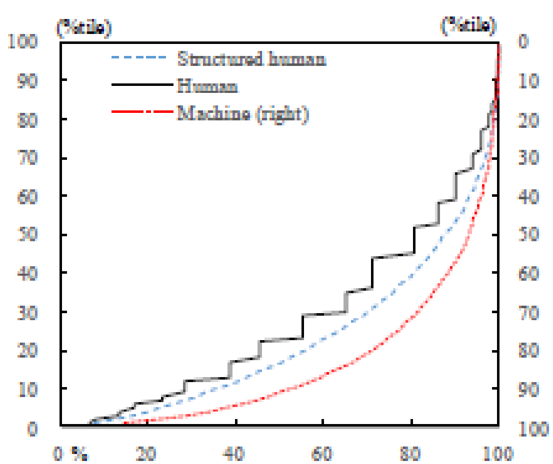
- Econ view is supported for default (not for closure)

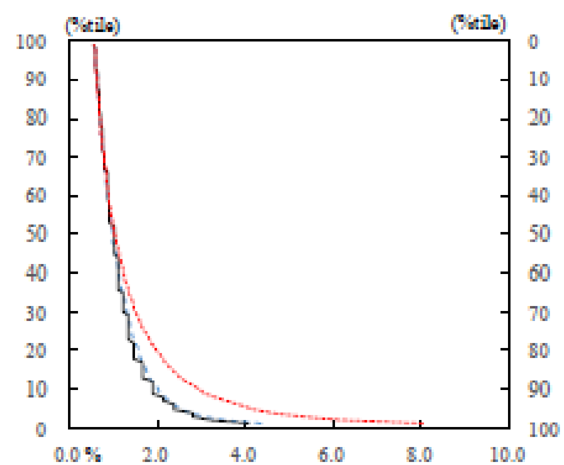
$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$









4-3. Result: Determinants

Determinants

- Higher opaqueness ⇒ M $\prec$ H
- Same pattern for SH $\prec$ H

	default						*voluntary closure*					
	Machine vs. Human			***SH vs. Human***			***Machine vs. Human***			***SH vs. Human***		
	Coef.	S.E.		Coef.	S.E.		Coef.	S.E.		Coef.	S.E.	
Number of available variables												
#(*available variables*) $_{f,t}$	0.566	0.001	***	0.041	0.000	***	0.485	0.001	***	0.031	0.000	***

(All the attributes $\boldsymbol{F}_{f,t}, \boldsymbol{I}_{i,t}, \boldsymbol{Z}_{i,t}$ are controlled)

	default: Machine vs. Human	default: SH vs. Human	voluntary closure: Machine vs. Human	voluntary closure: SH vs. Human
Firm fixed-effect	yes	yes	yes	yes
Analyst fixed-effect	yes	yes	yes	yes
Year fixed-effect	yes	yes	yes	yes
#(obs)	3,238,817	3,238,817	3,238,817	3,238,817
F	14,314.100	3,591.740	12,417.240	3,908.300
Adj. R-squared	0.879	0.789	0.831	0.777
Within R-squared	0.071	0.019	0.062	0.020

4-4. Result: Determinants

Robustness

- M vs. ground truth & H vs. ground truth (Table A1)
- Rankings based analysis:
  - Difference in ranking (Table A2)
  - A dummy variable taking the value of one if $Proxy_{f,i,t}$ is positive and zero otherwise (Table A3)
  - 1 to 10 variables, depending on the level of $Proxy_{f,i,t}$ (Table A4).
- Replace analyst-level fixed effect with analyst-year-level fixed effect (Table A5)
- Employ one of the sub-scores of *fscore*, which represents the "stability" of each firm, instead of the total *fscore* (Table A6)
- AUC estimation and proxy estimation based on the two alternative methods (i.e., LASSO and extreme gradient boost) (Table A8, A9)

4-5. Result: Determinants

Determinants

- Growth?
  - **1**(sales growth > Industry average + 1 std. dev.)

	Machine vs. Human		*SH vs. Human*	
	Coef.	S.E.	Coef.	S.E.
Number of available variables				
#(*available variables*) $_{f,t}$	0.196	0.003 ***	0.037	0.000 ***

(All the attributes $\boldsymbol{F}_{f,t}, \boldsymbol{I}_{i,t}, \boldsymbol{Z}_{i,t}$ are controlled)

	Machine vs. Human	SH vs. Human
Firm fixed-effect	yes	yes
Analyst fixed-effect	yes	yes
Year fixed-effect	yes	yes
#(obs)	3,037,588	3,037,588
F	4,799.540	650.920
Adj. R-squared	0.590	0.639
Within R-squared	0.026	0.004

4-6. Result: Task allocation

Counter factual

- Orthogonalize $\boldsymbol{O}_{f,t}$ to...
  - Firm's sales, sales growth, industry classification

- Then, make 5 (equal #obs) sub-groups accounting for
  - Highly Opaque
  - Opaque
  - Average
  - Transparent
  - Highly transparent

- Then, count # of TN, FN, TP, FP based on M & H

4-6. Result: Task allocation

Counter factual

- Firms actually do NOT exit (many)
  - H can reduce type I error for opaque firms

	Prediction for default			*Prediction for voluntary closure*		
	M = default H = not default (1)	M = not default H = default (2)	(2)/(1)	M = closure H = not closure (1)	M = not closure H = closure (2)	(2)/(1)
~20 %tile	49,117	23,068	0.47	25,206	19,453	0.77
20~40 %tile	36,094	54,446	1.51	28,326	23,667	0.84
40~60 %tile	37,362	46,368	1.24	28,370	28,134	0.99
60~80 %tile	33,409	39,218	1.17	20,249	30,962	1.53
80 %tile~	11,652	30,608	2.63	8,026	34,406	4.29

4-6. Result: Task allocation

Counter factual

- Firms actually exit (a few)
  - It is accompanied by larger type II error

	Prediction for default			*Prediction for voluntary closure*		
	M = default H = not default (3)	M = not default H = default (4)	(3)/(4)	M = closure H = not closure (3)	M = not closure H = closure (4)	(3)/(4)
~20 %tile	88	21	4.19	140	51	2.75
20~40 %tile	82	40	2.05	195	42	4.64
40~60 %tile	86	37	2.32	231	43	5.37
60~80 %tile	74	37	2.00	174	54	3.22
80 %tile~	38	27	1.41	72	45	1.60

Organization

1. Theoretical illustration
2. Methodology
3. Data
4. Results
5. **Summary**

5. Summary

- ML outperforms Human-prediction on average
- Yet, human-prediction could outperform for opaque firms due to the employment of soft info
  - # of exit firms are much smaller than that of non-exit firms
  - Type I error↓ overwhelms Type II error ↑ in terms of AUC

⇒ **When we should/shouldn't use ML** (≠ Luca et al. '16)

⇒ Other fields and issues (e.g., financial MKT)

X1: Grid search results

exit_default
(train for *t* = 2016, model 15)

		min.node.size				
		10	100	1,000	10,000	100,000
Mtry	1	0.705 <0.700>	0.703 <0.700>	0.706 <0.702>	0.711 <0.707>	0.695 <0.687>
	5	0.696 <0.688>	0.696 <0.688>	0.702 <0.698>	0.769 <0.765>	0.751 <0.747>
	14	0.689 <0.685>	0.687 <0.684>	0.715 <0.707>	0.773 <0.773>	0.769 <0.760>
	73	0.729 <0.716>	0.726 <0.718>	0.709 <0.710>	0.765 <0.764>	0.773 <0.766>

Note: upper value is ROC for training data, lower is AUC for test.

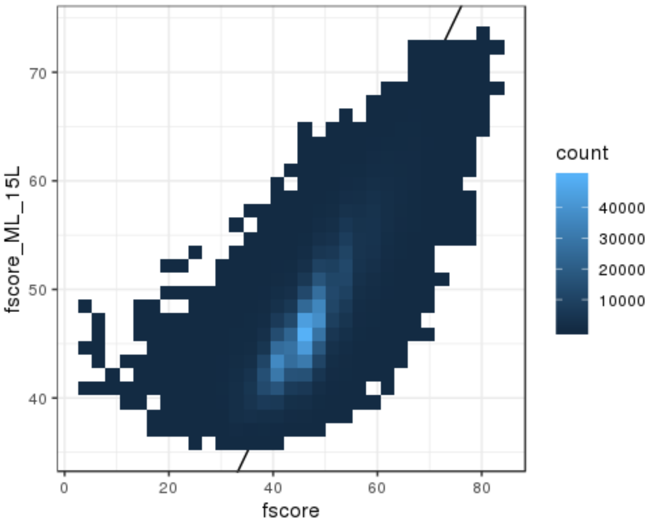
score
(train for *t* = 2016, model 15)

		min.node.size				
		10	100	1,000	10,000	100,000
mtry	1	5.143 (0.342) <5.126>	5.145 (0.338) <5.124>	5.153 (0.338) <5.132>	5.171 (0.330) <5.157>	5.309 (0.275) <5.259>
	5	3.729 (0.622) <3.716>	3.754 (0.620) <3.740>	3.841 (0.609) <3.824>	4.047 (0.577) <4.013>	4.551 (0.478) <4.467>
	14	3.358 (0.681) <3.352>	3.379 (0.678) <3.371>	3.476 (0.662) <3.460>	3.705 (0.624) <3.672>	4.231 (0.531) <4.155>
	73	3.317 (0.686) <3.313>	3.314 (0.687) <3.309>	3.384 (0.674) <3.374>	3.574 (0.639) <3.547>	4.049 (0.540) <3.999>

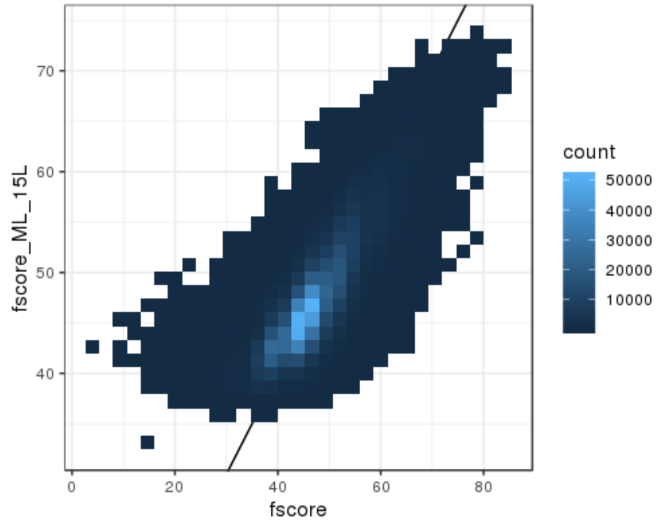
Note:
upper value is RMSE for training data, middle is R-squared, lower is RMSE for test.

X2: Predicted H

test for t = 2014



test for t = 2015



X3: Model configuration

	Model (set of variables use for prediction) pattern						
	1	8	15	17	18	19	20
	Estimation method						
Variable group	Probit	WRF	WRF	WRF	WRF	WRF	WRF
Fscore	○	○					
Firm own		○	○	△		△	
Financial statement		○	○	△	△		
geo/ind		○	○				
Bank		○	○	○	○	○	○
Network		○	○	△	△	△	△
Shareholder		○	○	△	△	△	△

Note: △ indicates smaller set of variables is applied compared to ○. Blank means no variables are in the model.

Thank you and comments are welcome!

<Contact Information>

Daisuke Miyakawa:
Associate Professor
Hitotsubashi University Business School (HUB)
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8439 Japan
E-mail: dmiyakawa@hub.hit-u.ac.jp
Web: https://sites.google.com/site/daisukemiyakawaphd/

Kohei Shintani:
Director and Senior Economist
Institute for Monetary and Economic Studies
Bank of Japan
2-1-1 Nihombashi-Hongokucho, Chuo-ku, Tokyo 103-8660 Japan
E-mail: kouhei.shintani@boj.or.jp