IFC-Bank of Italy Workshop on "Machine learning in central banking"

19-22 October 2021, Rome / virtual event

Disagreement between human and machine predictions[1]

Daisuke Miyakawa, Hitotsubashi University Business School, Japan,
and Kohei Shintani, Bank of Japan

[1] This presentation was prepared for the Workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the central banks and other institutions represented at the event.

Disagreement between Human and Machine Predictions

By DAISUKE MIYAKAWA AND KOHEI SHINTANI*

In this study, we show that human predictions of firm exits disagree with machine predictions. First, human predictions generally underperform machine predictions. Second, the performance of human relative to machine predictions improves for firms with less observable information that is possibly due to the unstructured information that only humans can use. Specifically, under the environment where the number of exiting firms is much smaller than that of non-exiting firms, the reduction in type I errors from reallocating prediction tasks to humans instead of machines for opaque firms leads to better performance of predictions. (93 words)

Keywords: Machine Prediction; Human Prediction; Disagreement

JEL classification: C10, C55, G33

* Miyakawa (corresponding author): Associate Professor, Hitotsubashi University Business School, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8439 JAPAN. E-mail: dmiyakawa@hub.hit-u.ac.jp. 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.

I. Introduction

Prediction is an important task in both private business and public policy. Recent advances in prediction techniques, such as machine learning, have helped make the performance of prediction tasks more reliable than those dependent upon human judgment and classical parametric models. The practical application of these new prediction techniques has been the focus of recent academic, policy, and business discussions (Varian 2014; Mullainathan and Spiess 2017; Athey 2019). A number of fields have already reported successful applications of these techniques such as labor markets (Chalfin et al. 2016), public services (Kleinberg et al. 2018; Bazzi et al. 2019; Lin et al. 2020), medical services (Patel et al. 2019; Mei et al. 2020), and the financial industry (Agrawal et al. 2018).

The growing employment of these powerful prediction techniques naturally raises the question of the ways in which machine predictions disagree with and outperform human predictions. This question is particularly relevant given the number of recent studies which argue that technological advances will lead either to the replacement of human labor with machines in certain types of jobs (e.g., Frey and Osborne 2017) or to the reallocation of human resources to other types of jobs (e.g., Autor et al. 2003; Acemoglu and Autor 2011; Acemoglu and Restrepo 2018). To understand the ways in which machines outperform humans in predictions, we identify the cases in which human predictions outperform machine predictions.

While this question has started to be examined in fields like medical studies (e.g., Raghu et al. 2019), it has not yet been investigated in the context of social sciences.

The goal of this study is to use the context of firm exits to show the patterns of disagreement between human predictions and machine predictions and each predictor's relative performance. First, following the medical studies, we test the relative performance of predictions based on machine learning techniques and those based on human judgment for two modes of firm exits: corporate default and voluntary closure. Second, we identify the systematic patterns of disagreements between human and machine predictions for those events. The disagreement between them is measured by the performance of the machine relative to that of the human. Thus, we can see not only whether humans and machines disagree but also, more importantly, the ways in which they disagree. Suppose a firm is actually found to default ex post. Ex-ante human and machine predictions could differ. As reported by Kleinberg et al. (2018) in the context of judicial bail decisions, machine predictions outperform human predictions more often. Nonetheless, the relative performance of human predictions may be better in specific circumstances, such as default predictions for informationally opaque firms. Given this conjecture, we find that the relative performances of human and machine predictions are conditional on the characteristics of their prediction targets: firms. Third, after confirming the conjecture, we implement a set of counterfactual exercises that reallocate the

predictions for firms with specific characteristics to humans instead of machines and see how overall performance of predictions varies.

To the best of our knowledge, this study is the first to explicitly examine the systematic patterns of disagreement between human and machine predictions in the context of social science and to use these systematic patterns to improve the overall performance of predictions.[1] We take advantage of our access to a huge volume of firm-level high-dimension panel data collected by one of the largest Japanese credit reporting agencies, together with the prediction results of anonymous professional analysts who work for the agency. These comprehensive datasets provide us with an ideal research ground on which we can construct a machine prediction model to compare its predictions with human predictions and show how they disagree and perform.

The empirical findings are summarized as follows: First, machines have better average performance in predicting firm exits than humans have, which is consistent with the results reported by the studies in another field (e.g., Kleinberg et al. 2018). Second, the performance of human predictions relative to that of machine predictions improves as the availability of information on firm characteristics declines. This improvement could be the case when human predictions effectively

[1] Anderson et al. (2017) report in the domain of chess that human decisions tend to be wrong for more difficult instances of chess. Their study shares the motivation with ours in the sense that both characterize the determinants of the performance of human decisions. The difference is that we compare human predictions not only with the ground truth (i.e., firm exits which we observe ex post), which is done in Anderson et al. (2017), but also with machine predictions.

use unstructured information in their predictions. The research has referred to this kind of unstructured information as "soft information" (e.g., Liberti and Petersen 2019). Examples of soft information include CEO's management ability, the prospects of future product development, and so on. It is difficult to record all of this highly qualitative information as structured (i.e., "hard") information in, for example, firms' financial statements or other documents.

Therefore, we compare the human predictions recorded in our dataset not only to machine predictions but also to the part of the human predictions solely correlated with structured information.[2] As the latter structured human predictions do not rely on unstructured information, the comparison between them and the original identifies to what extent humans used unstructured information in their predictions. Similar to the comparison between the original human predictions and machine predictions, we find that the performance of human predictions relative to that of structured human predictions improves as the availability of information on firm characteristics declines. We also separately regress the performance of human and machine predictions on various characteristics including firm attributes and confirm that the negative marginal effects associated with lower availability of information is more sizable for machine predictions than for human predictions.

[2] A similar attempt to replicate human decisions was done in the context of chess (e.g., McIlroy-Young et al. 2020).

Given the empirical finding that the availability of observable information is a key driver in the disagreement between human and machine predictions and their relative performance, we implement a set of counterfactual exercises that reallocate predictions to professional analysts from machines that depends on how much information is available for each firm. The "improvement" in the relative performance of human predictions along with the change in specific firm characteristics does not guarantee that the "level" of conditional performance of human predictions is higher than that of machine predictions. In this sense, our counterfactual exercises are useful in confirming whether there could be any cases in which humans outperform machines when making predictions in the level of prediction performance.

As a main characteristics of firms, we pay attention to the number of available variables for each firm, which is closely related to the opaqueness of the firms. We orthogonalize the number of available variables to other firm characteristics such as size, past growth trend, and industry fixed effects so that we can extract the variation in the information opaqueness independent of those characteristics. Using this orthogonalized variable accounting for the information opaqueness, we classify firms into five categories that range from firms with the least information, little information, average information, more information, and the most information. For most of the cases except for firms with the least information, machine predictions outperform human predictions in terms of both type I and type II errors.

Nonetheless, we also find that reallocating predictions on firms with the least information to humans instead of machines leads to a sizable reduction in the type I error. To illustrate, for firms with the least information, the number of actual non-exiting firms predicted as "exit" by machines but "non-exit" by humans is larger than the number of actual non-exiting firms predicted as "non-exit" by machines but "exit" by humans. Thus, reallocating predictions on those firms to humans instead of machines reduces the number of false-positives, and the type I error becomes smaller. However, the reallocation of the predictions on these firms is also accompanied by a larger type II error; that is, the number of actual exiting firms predicted as "exit" by machines but "non-exit" by humans is larger than the number of actual exiting firms predicted as "non-exit" by machines but "exit" by humans. These results mean that reallocating predictions to humans instead of machines also reduces the number of true-positives, and thus the type II error increases. As the number of exit firms are much smaller than that of non-exit firms, the reduction in the type I error achieved by reallocating predictions on those opaque firms to humans instead of machines overwhelms the increase in the type II error. This is the mechanics in which the relative performance of human predictions to that of machine predictions improves as the availability of information on firm characteristics declines.

These results jointly show the usefulness of powerful machine prediction techniques for practical purposes and highlight a subtle feature of human prediction

in the context of exit prediction. Overall, most of the prediction work for firm exits can be assigned to machines. Nonetheless, under specific circumstances, such as when the prediction targets are informationally opaque and the user of the resulting predictions is more concerned about the type I error than the type II error due to, for example, the imbalance between the numbers of exit and non-exit firms, then there is still room for human predictions to outperform machine predictions. Although we are not dealing with individuals as the subjects of predictions in the present study, these results support Gebru (2020) who reports that automated facial analysis systems tend to have lower prediction power for individuals with specific characteristics (e.g., dark-skinned women). Regardless of what types of subject that are the targets of the prediction, understanding under which cases machines could be wrong is useful.

The rest of the study proceeds as follows: Section II presents the theoretical underpinning of our empirical study, which follows Raghu et al. (2019). In Section III, we explain our empirical methodology and give a brief account of the institutional background related to the prediction of firm exits. Section IV gives details on the data used for our study. In Section V, we present and discuss the empirical results. Section VI concludes.

II. Conceptual Framework

In this section, we present the conceptual framework that represents the disagreement between human and machine predictions and their relative performance. Suppose there is a prediction f for a specific outcome. We set predictions for firms' default or voluntary closure as our prediction f. The f is accompanied by a set of attributes. It consists of, for example, the amount of available information associated with the firms as well as other firm characteristics in their financial statement. The f has the actual outcome $a(f)$ that we refer to as a ground truth. This ground truth only exists ex post when we observe whether the firm defaults or not within specific periods of time. For f, a prediction machine has its own prediction denoted by $m(f)$. Similarly, a professional analyst i with a set of individual attributes has its own prediction for f. We name this analyst's prediction $h(f, i)$. Using these items, we can first define the prediction error $\Theta(f)$ of machines for an f as follows:

$$\Theta(f) = L(a(f), m(f)). \tag{1}$$

Second, we can define the prediction error $\Omega(f, i)$ of humans for f by an analyst i as follows:

$$\Omega(f, i) = L(a(f), h(f, i)). \tag{2}$$

Suppose we have a set of predictions U. What we ultimately want to solve is an allocation problem of U for machines (i.e., S) or analysts (i.e., T). Such an optimization problem can be formulated as follows:

$$\min_{S,T} \sum_{f \in S} \Theta(f) + \sum_{f \in T} \Omega(f,i) \text{ s.t. } S \cup T = U; S \cap T = \emptyset. \quad (3)$$

This problem is called "an algorithmic triage" in Raghu et al. (2019). To solve this problem, we obtain the best assignment (S^*, T^*) as a function of (f, i). This optimal assignment function tells us whether we should assign a specific prediction f to the machine or to an analyst i.[3] In this paper, we specifically aim to identify $\Theta(f)$ and $\Omega(f,i)$ so that we can understand the sources of the disagreement and further solve the algorithmic triage problem as a counterfactual exercise.

For this purpose, we define an additional function $Proxy_{f,i}$ as follows:

$$Proxy_{f,i} = \Omega(f,i) - \Theta(f). \quad (4)$$

As $\Theta(f)$ and $\Omega(f,i)$ denote the prediction errors of the machine and the analyst, the relative performance of the human prediction becomes higher as $Proxy_{f,i}$

[3] Although the current setup does not contain any constraints for the optimization problem, realistic constraints such as a maximum number of instances a professional analyst can take care of could be introduced to the problem. Such a problem is a classic example of a matching problem.

becomes smaller. As we explicitly demonstrate in the following sections, this $Proxy_{f,i}$ accounts not only for the disagreement between human and machine predictions but also for their relative performance.

While the current setup suffices to study the systematic disagreement between human and machine predictions, further decomposition of $\Omega(f,i)$ into those correlated with structured information and the rest of the components is useful for understanding the source of the disagreement between human and machine predictions. Let $\Omega_h(f)$ account for the error component of the human prediction correlated with structured observable attributes of f. Using this decomposition, we can define another measure for disagreement between the human prediction and the structured human prediction that relies solely on hard information.

$$Proxy'_{f,i} = \Omega(f,i) - \Omega_h(f). \tag{5}$$

Suppose $Proxy'_{f,i}$ becomes smaller as the change in an attribute of the instance f (e.g., the amount of available information decreases). This change means the relative performance of the human prediction to the structured human prediction becomes better due to the change in the attribute. In this illustration, the volume of structured information becomes smaller, the room for analysts to effectively utilize unstructured information for prediction becomes larger. This comparison between human and structured human predictions highlights the reason why human

predictions can surpass machine predictions, with the latter relying only on structured information.

III. Empirical Strategies

This section first presents the way that we construct a machine learning prediction model for firm exits. Then, we explain how to identify the determinants of disagreement and the relative performance of human and machine predictions.

A. Machine Prediction

To obtain machine predictions, we construct a standard machine learning method. Our particular problem with predicting relatively rare firm exits falls into the class of "imbalanced label predictions." Following the literature, we apply a weighted random forest and a minority-class oversampling method.[4] Random forest models aggregate many individual decision tree models that are each trained on a randomly selected samples and predictors from the training data. To predict rare events, Chen et al. (2004) develop an extension of the random forest, called a weighted random forest. Logically, the method weighs data corresponding to a minority event (e.g., a firm exit) much more heavily than that corresponding to a majority event (e.g., non-exit).

[4] We also use other machine learning techniques such as LASSO and extreme gradient boost to construct prediction models and confirm the robustness of our results. All the results are in the online appendix.

In our baseline exercise, we train models by using outcome variables from the end of year $t-1$ to the end of year t and the predictors available for the periods from year $t-3$ to $t-1$. Then, we conduct out-of-sample predictions of the realization of the outcome variables from the end of year t to the end of year $t+1$ by using the information available over the periods from year $t-2$ to t.

We use the receiver operating characteristic (ROC) curve to evaluate the predictive performance of the model. To implement the prediction of a binary exit outcome, we need a specific threshold. When a predicted score surpasses the threshold, it indicates a positive binary outcome. For a given trained model, the ROC curve plots the true and false positive rates that correspond to the variation in this threshold value. Without any predictors (i.e., random guesses), the curve should follow a 45-degree line, and curves that are closer to the top-left corner are desirable (maximize true positive rate and minimize false positive rate). Following convention, we summarize the ROC curve with the area under the curve (AUC).

B. Human Prediction

"fscore"—Credit reporting agencies examine and predict firm exits as these outcomes are of great interest to business and government entities. Examples of such credit reporting agencies include Dunn and Bradstreet in the US, Experian in European countries, and Tokyo Shoko Research (TSR) in Japan. By providing structured information such as financial statements to their clients, credit reporting

agencies typically calculate and publish a credit rating score, which we call "*fscore*", to summarize the overall performance of a firm. This score is typically constructed from both structured information on firm characteristics and from the contents of in-depth interviews on firm's characteristics, reputation, growth opportunity, and so on (i.e., unstructured information). The score is constructed by a professional analyst and assigned to each firm in each year. As in financial institutions such as banks, the agency evaluates each analyst on the performance of their predictions of this $fscore$. Thus, analysts have a reasonable incentive to produce good predictions.

These agencies typically rely on their own (often confidential) algorithm to construct the scores. While a part of the score systematically depends on structured information, a large part of the score reflects professional analysts' subjective evaluation of the targeted firm. To illustrate, according to the publicly available information, a score given by TSR (max: 100 points) is the summation of (i) the capability of the firm (max: 20 points) based on business attitude, experience, and asset condition; (ii) the growth possibility (max: 25 points) based on past sales growth, growth of profits, and characteristics of the products; (iii) stability (max: 45 points) based on the firm's age, stated-capital, financial statement information, room for collateral provision, and real and financial transaction relationships; and (iv) the firm's reputation (max 10 points) based on the level of disclosure and overall reputation. Most of these items are rarely recorded as structured information

but largely as unstructured information. Given this institutional background, we use the *fscore* assigned by TSR as the output of human predictions.

We use this score and the ex-post record of exit to run a weighted Probit estimation that has the exit indicator on the left hand-side and only $fscore$ on the right hand-side of the estimated equation. Through this equation, we transform the $fscore$ into a value between 0 and100 as the score associated with the occurrence of the firm exit and use it as the result of human prediction.[5]

Can we really use fscore as a human prediction? There could be several immediate concerns about using the *fscore* as the output of human predictions. First, this score might also be constructed by some machine algorithms. If this is the case, the comparison between human and machine predictions becomes merely a comparison of two algorithms. However, we also try to separate out the analysts' predictions correlated with structured information from the original *fscore*. Using this framework, we can explicitly study the difference between predictions based on structured information and those based on unstructured information, the latter of which can be handled only by human analysts.

[5] We should note that due to the weighting procedure for a minority-class oversampling, the output obtained by WRF and this Probit estimation are not exactly the exit probability in the data. Instead it is the probability of exits in the balanced sample consisting of equal numbers of exits and non-exits. Given there is no problem for us to use these probabilities as far as the machine outputs are constructed in the comparable way, we use them in the following empirical analyses. We also construct a ranking based on the outputs obtained by WRF and the Probit estimation and use it for our empirical analysis. The results of which are reported in the online appendix.

Second, machine predictions can take into full account higher dimensions of information than human analysts can. When this is the case, the comparison between *fscore* and machine prediction might account only for the difference between the two different datasets used by humans and machines. Although we think the ability to handle different volumes of information itself is one aspect of the difference between humans and machines and thus worth examining, we also try to compare human and machine predictions on an equal footing in terms of the volume of structured information.

Third, the target of machine and human predictions might not be exactly the same. This issue is called an omitted payoff bias in the literature (Chalfin et al. 2016). As we will detail in the next section, we construct machine learning-based prediction models explicitly targeting one of the two modes of firm exits (i.e., default and voluntary closure), while the *fscore* summarizes the overall performance of a firm. Although the *fscore* is typically used in credit risk management and thus largely accounts for the prospects of firm exits, it is better to have human predictions more directly connected to firm exits.[6] For this purpose, we employ not only the overall firm performance score but also the sub-scores corresponding to the financial stability of firms as human predictions.

[6] TSR guidelines provide the following categorization of *fscore* ranges: (a) caution required (scores 29 and under), (b) medium caution required (scores between 30 and 49), (c) little caution required (scores between 50 and 64), (d) no specific concern (scores between 65 and 79), and (e) no concern at all (scores 80 and above).

Apart from these concerns, the external validity of the results is also important. Disagreements between human and machine predictions may be important in other situations, such as the comparison between machines and investors who put more emphasis on the "upside" of a firm's performance rather than the downside. To address these concerns, we implement the same set of analyses for firms' sales growth and assess the robustness of our results regarding firm exits.

Structured human prediction—As already noted, *fscore* is likely to account for both structured and unstructured information. While it is still informative to compare the original *fscore* with the machine score, we also extract the component of *fscore* associated only with the unstructured information. For this purpose, we construct a machine learning prediction model for *fscore* by using the same right hand-side variables as we use to construct the machine prediction model. This "structured" *fscore* accounts only for the part of *fscore* correlated with the structured information. We use this predicted score and the actual record of exits to run a weighted Probit estimation to transform the structured *fscore* into the probability that is associated with the occurrence of the firm exits.

C. Measurement of "disagreement"

We measure the disagreement between human and machine predictions for a specific exit mode of firm f in year t. We standardize the machine scores of exits,

the calibrated *fscore* by a weighted Probit estimation, and the calibrated structured *fscore* as a mean zero and the standard deviation as one. By using these standardized scores for machines (ML), analysts (H), and structured humans (*SH*) that are denoted by $Outcome$, we compute a variable $Proxy$ for a firm (f), analyst (i), and a time (t), which was conceptualized in the previous section, as the following definition:

$$(6) \qquad Proxy_{f,i,t} = Outcome^{ML}_{f,t} - Outcome^{H}_{f,i,t} \text{ for exit firms,}$$
$$= Outcome^{H}_{f,i,t} - Outcome^{ML}_{f,t} \text{ for non-exit firms,}$$

$$(7) \qquad Proxy'_{f,i,t} = Outcome^{SH}_{f,t} - Outcome^{H}_{f,i,t} \text{ for exit firms,}$$
$$= Outcome^{H}_{f,i,t} - Outcome^{SH}_{f,t} \text{ for non-exit firms.}$$

Due to the way we compute $Proxy$, this measure of the disagreement becomes larger when the machine or structured human produces better predictions than the human does.

D. Identifying the determinants of "disagreement"

Once we obtain a measurement of $Proxy$, we can estimate the relationship between $Proxy$ and a linear function $G(\cdot)$ of various explanatory variables that consist of informational opaqueness of firms ($\boldsymbol{O}_{f,t}$), their attributes ($\boldsymbol{F}_{f,t}$), analyst attributes

($\boldsymbol{I}_{i,t}$), and team attributes ($\boldsymbol{Z}_{i,t}$) as well as various configurations of fixed effects ($\boldsymbol{\eta}_{f,i,t}$):

(8) $Proxy_{f,i,t} = G\left(\boldsymbol{O}_{f,t}, \boldsymbol{F}_{f,t}, \boldsymbol{I}_{i,t}, \boldsymbol{Z}_{i,t}\right) + \boldsymbol{\eta}_{f,i,t} + \varepsilon_{f,i,t}$ for $t = 2013, \cdots, 2016$.

In the baseline estimation, we use a firm-level fixed effect, analyst-level fixed effect, and a year-level fixed effect for $\boldsymbol{\eta}_{f,i,t}$, while alternative configurations of fixed effects are also used for the robustness check.

IV. Data

In this section, we provide the details of the data used in our empirical analysis. All the data were obtained from TSR through its joint research contract with Hitotsubashi University. We use multiple datasets to construct a machine prediction model for firm exits to estimate the determinants of $Proxy_{f,i,t}$ and to implement counterfactual exercises.

A. Firm-level panel data

One of our main data sources is an annual-frequency panel of Japanese firm data from t=2010 to 2016 that provide information on firms' financial statements and basic details such as industry classification, firm characteristics, precise geographic location, and age. The year identifier t accounts for the timing of collection and

means that t consists of the data extracted as of the end of December of the year t from TSR. Given a large portion of Japanese firms use an accounting period that ends on March 31, the file labeled t =2012, for example, consists of a large amount of firm information that corresponds to the accounting period up to the end of March 2012. The original data cover around three million firms per year. We use the data that cover around one million firms which provide the information we need for our empirical analysis such as the latest financial statement. According to the Japanese Small and Medium Size Enterprises Agency, there are around three to four million active companies in Japan. The TSR data account for around one-third of that firm population. One point of note is that the sample selection is tilted toward some specific industries, such as construction companies.

These firm-level panel data are accompanied by three types of relational information regarding real and financial partners. First, this information contains a list of up to 10 lender banks. Second, the information also covers firm-to-firm trade. It lists up to 48 customer and supplier firms for each company. In addition to the list of each target firm's trade partners, we also use the trade relationship reported by those trade partners. As there are many trade relationships not reported by the targeted firms but only by their trade partners, this operation significantly extends the list of trade partners. Third, the data also contain the list of shareholders.

B. Predictions

We consider the two exit outcomes to be predicted one-year ahead: firm default and voluntary closure. The explanatory variables and outcome variable used in constructing a machine prediction model are defined for separate time intervals; explanatory variables from 2010 to 2012 to predict the outcome for the one-year window from the end of 2012 to the end of 2013, explanatory variables from 2011 to 2013 to predict the outcome from the end of 2013 to the end of 2014, and so on. The latest data are the explanatory variables from 2014 to 2016 that are used to predict the outcome from the end of 2016 to the end of 2017.

We measure defaults and voluntary closures as the firms that exited the market for these reasons during the one-year window. Then, we separately prepare two dummy variables that equal one if firms exited through either default or voluntary closure.

C. Firm attributes

To construct a machine prediction model of firm exits, we use the following six categories of attributes of firms: basic characteristics (***firm own***), detailed financial statement information (***financial statement***), geography and industry-related variables (***geo/ind***), firm-bank borrowing relationship variables (***bank***), supply chain network variables (***network***), and shareholder-subsidiary relationship

variables (***shareholder***). All the variables categorized in each group are summarized in the online appendix.

We set up the two prediction models for each one of the exit modes using these six groups of firm attributes together with the differenced and double-differenced variables of those variables.[7] We create a set of dummy variables to deal with missing variables that equals one if the corresponding variable is missing for a firm and zero otherwise. When a missing dummy variable equals one, we use zero for the original missing record.

D. Potential determinants of disagreement

To estimate the determinants of the disagreement between human and machine predictions, we set up the following three groups of variables: the amount of available information, firm attributes, and analyst/team attributes.

Number of available variables—As the most important potential determinant in our analysis, which is denoted by $\boldsymbol{O}_{f,t}$, we use the number of variables available (#(*available variables*)) for each firm in the dataset. This number accounts for the opaqueness of firms. When this number is small, both humans and machines can use only a limited amount of structured information. As humans can also utilize

[7] In our data, the predictors and the ex-post outcomes accounting for firm exits are observable. In this sense, our analysis does not suffer from the selective label problem that some of the ex-post outcomes cannot be observed due to selection (Lakkaraju et al. 2017).

soft information, the estimated coefficient associated with #(*available variables*) shows how effectively humans use such soft information in their predictions.

Firm attributes—We use a subset of variables that we used to construct the machine prediction model as the potential determinants, which we denote $\boldsymbol{F}_f$. The list consists of the logarithm of firm sales, its difference, the dummy variable for listed status, and the number of industries the targeted firms operate in. We use this list of variables as they are less prone to missing data.[8] In addition to these variables, we also use the information that relates to the task priority of each firm (*priority*) inside the credit reporting agency that is denoted by a number, with a larger number corresponding to a higher priority. The dataset includes the firm-level panel data of *fscore*. The number is computed as the sum of the four sub-scores that represent the ability of the firm, growth possibility, stability, and reputation. In the following empirical analysis, we use both the *fscore* and the decomposition of each component.

Analyst/Team attributes—We also use the attributes $\boldsymbol{I}_i$ of the analysts. To measure $\boldsymbol{I}_i$, at each data point, we use the attributes of the analysts working for TSR as stored in their anonymized background information. As analysts enter and exit the pool of

[8] Note that the existence of missing data in specific variables can be taken care of by introducing dummy variables that account for the missing data in the non-parametric model such as the random forest we use for constructing the prediction model. Contrary to this, the parametric model such as the panel estimation used for identifying the determinants of the disagreement cannot take care of the missing variables well.

TSR employees, the data become unbalanced panel data. This dataset is accompanied by a table that lists the firms assigned to each analyst at each data point that we use to relate analysts to firms. The dataset allows us to identify the list of assigned firms in each year and the tenure of each analyst. The former information allows us to calculate the number of firms assigned to each analyst and any previous exposure of an analyst to other firms in the industry of the targeted firms, which can be interpreted as the industry expertise of the analyst.

The dataset also allows us to measure the characteristics associated with the team each analyst belongs to, which is denoted by $\boldsymbol{Z}_{i,t}$. First, we measure the size of the team by counting the number of analysts in each department. Second, we measure the average tenure of all members of the team. Third, we measure the average number of firms assigned to the analysts in the team. Fourth, we also measure the average industry expertise of all the analysts in each team.

We understand that this analyst and team information is endogenous as the assignments of analysts to teams and to targeted firms are not random. Thus, we treat these variables simply as control variables in the regression of the determinants for $Proxy_{f,i,t}$ and do not intend to establish any causal relation between these variables and $Proxy_{f,i,t}$.

Table 1 summarizes the variables used to estimate the determinants of the disagreement between human and machine predictions, together with the *fscore*, structured *fscore*, and $Proxy_{f,i,t}$.

Table 1: Summary statistics

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

V. Empirical Results

A. Prediction performance

The following four panels in Table 2 show the AUCs and their standard errors of the five prediction models for the years 2013 to 2016. The first and second rows show the performance of human predictions and machine predictions, respectively. The third row is for the structured human predictions. The fourth and fifth rows show the performances of machine predictions with different sets of independent variables. The fourth row is the case where we add *fscore* to the list of independent

variables used to construct a machine prediction model. The fifth row corresponds to the case where we use only a small set of independent variables to construct a machine prediction model.[9] This smaller set is used to compare human and machine predictions on an equal footing in terms of the volume of structured information.

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)

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)

[9] As the smaller set of variables, we use all the ***firm own*** variables except for dividend-related variables; ***financial statement*** variables that represent total assets, profit, and EBITDA all the ***bank*** variables; ***network*** variables that represent only customers and suppliers with direct links; and ***shareholder*** variables in direct shareholding relations.

Machine with small information	0.768 (0.0050)	0.834 (0.0025)	Machine with small information	0.758 (0.0049)	0.843 (0.0024)

Note: Each number represents the AUC and the number in the parentheses is its standard error.

First, the tables show that the AUC of machine predictions (the second row) is significantly higher than that of human predictions (the first row) given the size of the standard errors of those AUCs. This is the case even when we use the smaller set of independent variables (the fifth row). Thus, machine predictions outperform human predictions on average.

Second, in the case of predicting default, humans outperform structured humans (the first and third rows). We also find that *fscore* makes an additional contribution to the overall performance of the machine predictions (the second and fourth rows). These results contrast with the findings of Kleinberg et al. (2018). In their empirical analysis of judicial bail decisions, they report that the structured human does a better job of predicting risky criminals than the judge. They claim that the "psychologist's view" in which humans make noisy predictions overwhelms the "economist's view" in which humans can use soft information to make a better prediction. Our result shows that at least in our setup for default predictions, the economist's view should be more reliable. Furthermore, as for predicting voluntary

closure, the structured human does a better job than the human does, which is consistent with the psychologist's view.[10]

B. Determinants of disagreement

Table 3 summarizes the results of the panel estimation associated with default and voluntary closure. All the coefficients are shown in the percent point (i.e., the estimated coefficients times 100).

Table 3: Baseline estimation

	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
Firm characteristics								
log(*sales* $_{f,t}$)	-18.545	0.127	3.987	0.028	-8.511	0.111	5.036	0.030
log(*sales* $_{f,t}$) - log(*sales* $_{f,t-1}$)	13.015	0.097	-0.618	0.022	5.205	0.086	-0.521	0.023
listed $_{f,t}$	-2.105	2.758	0.605	0.621	-18.931	2.429	-6.351	0.662
#(*industry*) $_{f,t}$	-3.009	0.159	-0.084	0.036	0.097	0.140	-0.129	0.038
priority $_{f,t}$	0.001	0.000	0.000	0.000	0.002	0.000	-0.000	0.000
Analyst characterstics								
#(*assigned companies*) $_{i,t}$	-0.001	0.000	-0.000	0.000	-0.001	0.000	-0.000	0.000
industry experience $_{f,i,t}$	-0.004	0.000	0.000	0.000	-0.001	0.000	0.001	0.000
Team characteristics								
#(team members) $_{i,t}$	0.081	0.012	-0.001	0.003	0.106	0.010	-0.001	0.003
Average #(*tenure years*) $_{i,t}$	0.136	0.016	-0.008	0.004	-0.008	0.014	-0.006	0.004
Average industry experience $_{f,i,t}$	0.014	0.001	0.000	0.000	0.001	0.001	0.000	0.000
Average #(*assigned companies*) $_{i,t}$	-0.001	0.000	-0.000	0.000	-0.002	0.000	-0.000	0.000
Constant	152.997	1.512	-49.111	0.340	54.692	1.331	-59.965	0.363
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	

[10] In the online appendix, we examine the recall and precision measures for machine, human, and structured human predictions over different thresholds of prediction.

From the columns labeled as "*Machine vs. Human*", regardless of whether we use default or voluntary closure as the prediction target, we find that the prediction performance of humans relative to machines becomes better for firms with less observable information on their attributes (i.e., lower values for #(*available variables*)). Thus, for more opaque firms, the relative performance of human predictions to machine predictions improves.

Why do analysts perform better in the case of opaque firms? One conjecture is that analysts are using unstructured information that by definition, cannot be used in machine predictions. To confirm this conjecture, we also run the panel regression for $Proxy'_{f,i,t}$ that is defined by replacing $Outcome^{ML}_{f,t}$ with $Outcome^{SH}_{f,i,t}$. This regression characterizes under what conditions human predictions outperform those of the structured humans. The results in the columns labeled as "*SH vs. Human*" show a similar pattern to that in "*Machine vs. Human*", that is, the relative power of human predictions compared with structured human predictions becomes higher as the amount of available information becomes smaller.[11]

We also separately regress the performance of human and machine predictions on the same set of attributes. From the estimation results (reported in the online appendix), we confirm that the negative marginal effect associated with lower

[11] We also find that the marginal effect of the available information on the relative performance of human predictions compared to that of structured human predictions is much smaller than that for humans vs. machines. This difference means that the sensitivity of the structured human predictions to the level of available information is much lower than that of machine predictions.

availability of information is greater for machine predictions than for human predictions. This effect could be the case again when humans effectively use unstructured information to make predictions.

To check the robustness of the results and address the concerns we raised in the previous section, we first use alternative methods of measuring the disagreement between human and machine predictions. As detailed above, we are using the ex-post record of firm exits to obtain the probabilities of exit that are measured by *fscore* and the structured *fscore*. As the transformation of *fscore* to the probability is simply a monotonic transformation and does not change the order of the score, it does not affect the comparison of human and machine predictions. Nonetheless, in reality, such an ex-post record of exit that is used to calibrate *fscore* to probability is not attainable in the process of human predictions. Thus, we also construct a set of "rankings" based on the machine prediction, *fscore*, and the structured *fscore*. In this ranking of prediction outcomes, we do not need to refer to the ex-post default records for the purpose of calibration. Second, we also define a dummy variable that is equal to one if $Proxy_{f,i,t}$ is positive and zero otherwise. We use this dummy variable and run a linear probability model with the abovementioned fixed effects and conditional logit model with firm-level fixed effects. We also set 1 to 10 variables depending on the level of $Proxy_{f,i,t}$ and run an ordered-logit estimation without fixed effects. Third, we replace the analyst-level fixed effect with the analyst-year-level fixed effect so that we can take complete account of analyst-level

unobservable factors that vary over time and that subsume team-level time-variant unobservable factors. These are not likely to be captured by the limited number of explanatory variables $\boldsymbol{I}_i$ and $\boldsymbol{Z}_{i,t}$. Fourth, we use one of the sub-scores of *fscore*, which represents the stability of a firm, instead of the total *fscore*, so that the target of human predictions becomes more comparable to that of machine predictions. Fifth, instead of weighted random forest, we use LASSO or extreme gradient boost for producing machine predictions. All the results are shown in the online appendix and are consistent with the results in Table 3.

C. Counterfactual exercises

Can we use the empirical findings presented in the previous section to improve the overall performance of predictions on firm exits? Given that the performance of humans relative to machines improves for more opaque firms, then agencies will naturally assign these firms to humans and firms with greater information to machines.

Based on this conjecture, we split the sample into five subsamples according to the number of observable variables. We aim at setting up multiple groups for which the relative performance of humans differs from that of machines. To construct subgroups purely tied up to the number of observable variables, we regress #(available variables) to a firm's sales, growth, and industry classification that are significant in the estimation of $Proxy_{f,i,t}$ and take out the residual. Then, we use

this residual to sort the firms and construct five subsamples so that we can set up five groups of firms depending on the level of #(available variables) that is orthogonal to other firm attributes.

In each subsample, we evaluate the performances of human and machine predictions. By comparing, for example, the number of false negatives based on machine predictions *(ML)* to those based on human predictions *(H)* for the same set of firms, we can describe what happens to the prediction performance for the subsample by reallocating predictions to humans instead of machines.

Table 4: Reallocation of predictions instances

(a) Firms actually do *NOT* exit ex post

	Prediction for default			*Prediction for voluntary closure*		
	ML = default *H* = not default (1)	*ML* = not default *H* = default (2)	(2)/(1)	*ML* = closure *H* = not closure (1)	*ML* = 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

(b) Firms actually do exit ex post

	Prediction for default			*Prediction for voluntary closure*		
	ML = default *H* = not default (3)	*ML* = not default *H* = default (4)	(3)/(4)	*ML* = closure *H* = not closure (3)	*ML* = 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

Note: ML and *H* denote the predictions of machines and humans, respectively.

The two panels in Table 4 summarize the number of false-positive, false-negative, true-positive, and true-negative cases for the five subsamples. We treat the top 30% of firms in terms of the prediction score as the firms predicted to exit.[12]

For example, the columns marked (1) in panel (a), show the number of false-positives for machine predictions and true-negatives for human predictions, as these columns show the number of firms that do *not* exit ex post. Conversely, the columns marked (2) in panel (a) show the number of true-negatives for machine predictions and false-positives for human predictions for firms that do not exit ex

[12] For robustness check, we vary this prediction threshold (i.e., the top 30% in this baseline exercise) from the top 50% to the top 20% and confirm the results do not change.

post. Panel (b) in Table 4 summarizes the number in the same manner but for the firms that actually *do* exit ex post.

Comparing the numbers in each column, we can see how type I and type II errors vary depending on whether the predictions are allocated to machines or to humans. In six out of the 10 rows in Panel (a), the number in column (1) is smaller than that in column (2), while in Panel (b), all the numbers in column (3) are larger than those in column (4).

First, these results mean that the type II error is always smaller in machine predictions than in human predictions regardless of the level of available information. Even for the firms with the least information, human predictions cannot outperform machine predictions. Second, in the case of the firms with the least information (i.e., the first raw labeled as "~20%tile"), it is still possible to reduce the number of false-positives, and thus reduce the type I error, by reallocating the default predictions to humans instead of to machines (i.e., the number of false-positives is reduced from 49,117 to 23,068). In the case of voluntary closure, we can also achieve a smaller type I error for firms with the least, little, and average amounts of information (i.e., the first, second, and third raws labeled "~20%tile", "20~40%tile", and "40~60%tile") by reallocating the default predictions to humans instead of machines.

However, a reallocation of predictions is accompanied by a larger type II error, as shown above. The numbers in column (3) are always larger than those in column

(4) that indicates the reallocation of predictions always increases the number of false-negatives. As one interesting result, we also find that in the case of default predictions, the ratio is larger as we move from the subsample with the least information to that with the largest amount. This pattern is inconsistent with the positive coefficient obtained in our estimation of $Proxy_{f,i,t}$. This is the case simply because, in our data, the number of exits is much smaller than that of non-exits. In other words, the performance of human predictions relative to machine predictions with respect to the level of available information is driven by human predictions correctly predicting non-exit firms.

These results reconfirm the usefulness of machine prediction techniques in the context of exit predictions. There is however room for human predictions to outperform machine predictions under specific circumstances, such as when the prediction targets are informationally opaque or when the user of the prediction results is more concerned with a type I error than a type II error due to, for example, the imbalance between the numbers of exit and non-exit firms.

D. Growth prediction

We have so far focused on exit predictions. What happens if we focus on the upside of firm dynamics instead? We repeat the same analyses by considering firm growth as the target of our predictions. We define growth in sales as a rate of one standard deviation higher than the industry average defined in two digits over the one-year

window used to measure the outcome. Then, we prepare a dummy variable that equals one if firms experience a growth rate higher than these criteria.

As predictions for upside events are the opposites of downside predictions, we conjecture that while overall performance is still higher for machine predictions than human predictions, and the relative performance of human predictions also improves when the available information is smaller as we have described, the source of this better performance is not from a lower type I error but from a lower type II error. In other words, analysts more correctly predict growth for actually growing firms based on less information. As presented in the online appendix, this is indeed the case. Although the levels of type I and type II errors are always higher in the case of human predictions, relative prediction performance of analyst to machine improves for actually growing firms as available information becomes smaller.

VI. Conclusion

We empirically examine the relative performance of machine and human subjective predictions for firm exits. Using a huge volume of firm-level high-dimension panel data, we find that human predictions are not as accurate as machine predictions on average. As for predicting the exits of informationally opaque firms, the relative performance of human predictions improves.

One important point is that when using machine predictions in practice, Luca et al. (2016) claim that they cannot ensure automated decision-making as it is necessary to take into account the various dimensions of the problems under consideration. This study provides evidence that accounting for the conditions under which a prediction is to be assigned to a machine is also necessary. Our findings cast light on the circumstances and the extent to which tasks should be allocated either to machines or to humans.

Future extensions of the present study may benefit from the inclusion of additional explanatory variables as determinants of $Proxy$. A large-sized aggregate-level shock, such as a market downturn or a natural disaster, could have a marginal effect on each determinant of $Proxy$. Understanding potentially relevant shocks is useful in considering how we should allocate prediction tasks to machines and humans under specific circumstances. Such an additional analysis will help us to understand both the nature of human error and how humans and machines can work together to provide accurate predictions.

REFERENCES

Acemoglu, Daron, and David Autor. 2011. "Skills, Tasks and Technologies: Implications for Employment and Earnings." In *Handbook of Labor Economics* Vol. 4B, edited by Ashenfelter, Orley, and David Card, 1043-1171. Amsterdam: North-Holland.

Acemoglu, Daron, and Pascual Restrepo. 2018. "Artificial Intelligence, Automation and Work." NBER Working Paper No. 24196.

Agrawal, Ajay, Joshua Gans, and Avi Goldfarb. 2018. *Prediction Machines: The Simple Economics of Artificial Intelligence*. Harvard Business Review Press.

Anderson, Ashton, Jon Kleinberg, and Sendhil Mullainathan. 2017. "Assessing Human Error Against a Benchmark of Perfection." *ACM Transactions on Knowledge Discovery from Data (TKDD)* 11 (4), 45:1-25.

Athey, Susan. 2019. "The Impact of Machine Learning on Economics." In *The Economics of Artificial Intelligence: An Agenda*, edited by Agrawal, Ajay, Joshua Gans, and Avi Goldfarb, chapter 21. University of Chicago Press.

Autor, David H., Frank Levy, and Richard J. Murnane. 2003. "The Skill Content of Recent Technological Change: An Empirical Exploration." *Quarterly Journal of Economics* 118 (4): 1279-1333.

Bazzi, Samuel, Robert A. Blair, Christopher Blattman, Oeindrila Dube, Matthew Gudgeon, and Richard Merton Peck. 2019. "The Promise and Pitfalls of Conflict Prediction: Evidence from Colombia and Indonesia." NBER Working Paper No. 25980.

Chalfin, Aaron, Oren Danieli, Andrew Hillis, Zubin Jelveh, Michael Luca, Jens Ludwig, and Sendhil Mullainathan. 2016. "Productivity and Selection of Human Capital with Machine Learning." *American Economic Review* 106 (5): 124-27.

Chen, Chao, Andy Liaw, and Leo Breiman. 2004. "Using Random Forest to Learn Imbalanced Data." Technical Report 666 Statistics Department of University of California at Berkley.

Frey, Carl Benedikt, and Michael A. Osborne. 2017. "The Future of Employment: How Susceptible Are Jobs to Computerisation? *Technological Forecasting and Social Change* 114: 254-280.

Gebru, Timnit. 2020. "Race and Gender." In *The Oxford Handbook of Ethics of AI* ch. 13, edited by Dubber, Markus D., Frank Pasquale, and Sunit Das, 251–269. Oxford University Press.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. "Human Decisions and Machine Predictions." *Quarterly Journal of Economics* 133 (1): 237-293.

Lakkaraju, Himabindu, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. "The Selective Labels Problem: Evaluating Algorithmic." *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* August 2017: 275-284.

Liberti, José María, and Mitchell A. Petersen. 2019. "Information: Hard and Soft." *Review of Corporate Finance Studies* 8 (1): 1-41.

Lin, Zhiyuan "Jerry", Jongbin Jung, Sharad Goel, and Jennifer Skeem. 2020. "The Limits of Human Predictions of Recidivism." *Science Advances* 6 (7).

Luca, Michael, Jon Kleinberg, and Sendhil Mullainathan. 2016. "Algorithms Need Managers, Too." *Harvard Business Review* 94 (1/2): 96-101.

McIloroy-Young, Reid, Siddhartha sen, Jon Kleinberg, and Ashton Anderson. 2020. "Aligning Superhuman AI with Human Behavior: Chess as a Model System." *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* August 2020: 1677-1687.

Mei, Xueyan, Hao-Chih Lee, Kai-yue Diao, Mingqian Huang, Bin Lin, Chenyu Liu, Zongyu Xie, Yixuan Ma, Phillip M. Robson, Michael Chung, Adam Bernheim, Venkatesh Mani, Claudia Calcagno, Kunwei Li, Shaolin Li, Hong Shan, Jian Lv, Tongtong Zhao, Junli Xia, Qihua Long, Sharon Steinberger, Adam Jacobi, Timothy Deyer, Marta Luksza, Fang Liu, Brent P. Little, Zahi A. Fayad, and Yang. 2020. "Artificial Intelligence-enabled Rapid Diagnosis of Patients with COVID-19." *Nature Medicine*.

Mullainathan, Sendhil, and Jann Spiess. 2017. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives* 31 (2): 87-106.

Patel, Bhavik N., Louis Rosenberg, Gregg Willcox, David Baltaxe, Mimi Lyons, Jeremy Irvin, Pranav Rajpurkar, Timothy Amrhein, Rajan Gupta, Safwan Halabi, Curtis Langlotz, Edward Lo, Joseph Mammarappallil, A. J. Mariano, Geoffrey Riley, Jayne Seekins, Luyao Shen, Evan Zucker, and Matthew P. Lungren. 2019. "Human-machine Partnership with Artificial Intelligence for Chest Radiograph Diagnosis." *npj Digital Medicine* 2.

Raghu, Maithra, Katy Blumer, Greg Corrado, Jon Kleinberg, Ziad Obermeyer, and Sendhil Mullainathan. 2019. "The Algorithmic Automation Problem: Prediction, Triage, and Human Effort." arXiv:1903.12220.

Varian, Hal R. 2014. "Big Data: New Tricks for Econometrics." *Journal of Economic Perspective* 28 (2): 3-28.

The research data for this article

The data used in this study are proprietary, and we gained access to the data through a joint research contract between Miyakawa's institute and TSR. Therefore, we cannot provide the data to the journal.

Acknowledgments

This study was conducted as a part of a joint research project of Hitotsubashi University and Tokyo Shoko Research Ltd. (TSR). We thank Hiroshi Fujiki, Kaoru Hosono, Kentaro Iwatsubo, Kiyoshi Izumi, Koichi Kume, Makoto Nirei, Kazuhiko Ohashi, Arito Ono, Shigenori Shiratsuka, Daisuke Tsuruta, Hirofumi Uchida, Iichiro Uesugi, Toshiaki Watanabe, Yasutora Watanabe, and Isamu Yamamoto. We also want to thank the seminar participants at SigFin, Hitotsubashi University, and RIETI; the participants at the 2019 Finance Workshop held by the Institute for Monetary and Economic Studies; the Bank of Japan; 2020 JEA spring meeting; 28th NFA annual meeting; and the staff of the Bank of Japan for helpful suggestions. This study was supported by the Grant-in-Aid for Scientific Research (B) of the Japan Society for the Promotion of Science [grant number 17H02526]. Further, it was prepared in part while Miyakawa was a visiting scholar at the Institute for Monetary and Economic Studies, Bank of Japan. The views expressed in this study are those of the authors and do not necessarily reflect the official views of the Bank of Japan.

For Online Publication

Online Appendix A

The list of variables we use to construct the machine learning prediction model is as follows:

Firm-own characteristics (*firm own*): As variables that represent the firms' own characteristics, we use size as measured by the logarithm of sales and the change in sales from the previous period, profit-to-sales ratio and any change from the previous period, the status of dividend payments (paid or not) and any change from the previous period, whether the firm is listed or not, the number of employees, the logarithm of stated capital, and dummy variables that represent industry classification (note: multiple industry codes are recorded). We also use firm age, owner age, and the number of establishments.

Firms' financial statement information (*financial statement*): We set up a number of financial variables used in the literature to represent firms' detailed financial statement information.[13]

[13] The list of "***financial statement***" variables consists of the following items: Logarithm of total assets, cash-to-total assets ratio, liquid assets-to-total assets ratio, tangible assets-to-total assets ratio, receivables turn-over, inventory turn-over, total liability-to-total assets ratio, liquid liability-to-total assets ratio, bond-to-total liability ratio, bank borrowing-to-total liability ratio, bank short borrowing-to-total bank borrowing ratio, payables turn-over, interest coverage ratio, liquid assets-to-liquid liability ratio, fixed compliance ratio, fixed ratio, working capital turn-over, gross profit-to-sales ratio, operating profit-to-sales ratio, ordinary profit-to-sales ratio, net profit before tax-to-sales ratio, logarithm of EBITDA, logarithm of EBITDA-to-sales ratio, special income-to-sales ratio, special expenses-to-sales ratio, and labor productivity.

Industry and geographical information (*geo/ind*): We set up the following two groups of variables to represent the industry and area to which the firms belong. First, we construct the variables measuring the average sales growth of firms located in the same city as the targeted firms. Second, we compute the average sales growth of firms belonging to the same industry that are classified at the 2-digit level.

Lender banks information (*bank*): As variables that represent the firms' borrowing relationships with lender banks, we construct a dummy variable to represent a change in main lenders (i.e., top lender bank) or in the number of lender banks.

Supply-chain linkage information (*network*): We construct the following two groups of variables to represent the supply chain network. First, we compute widely used network metrics for each firm by using the network information on the supply chain. The metrics consist of degree centrality; eigenvector centrality; egonet eigenvalue; co-transaction; and the number of transaction partners, both direct (i.e., customers and suppliers) and indirect (e.g., suppliers' suppliers, and customers' suppliers). Second, we construct a number of variables that represent the characteristics of transaction partners. To summarize this information, we use the average, maximum, minimum, and the sum of *fscore* associated with each transaction partner. Note that while the network metrics cover both direct and

indirect transaction partners, the transaction partners' characteristics only cover direct transaction partners.

Shareholder linkage information (*shareholder*): We set up similar variables to those for the supply chain network as predictors of shareholder information.

Online Appendix B

We list the tables and figures referred to in the study for the robustness check. First, we show an alternative way to compare the prediction power of machines, humans, and structured humans (Figure A1). We can confirm that machine predictions outperform human predictions on average. Regarding the comparison between human predictions and those of the structured human predictions, human predictions are more precise in the case of default predictions, while the structured human predictions are better in terms of voluntary closure. Second, instead of estimating the determinants of $Proxy_{f,i,t}$, we estimate separately the determinants of $Proxy^{m}_{f,t}$ and $Proxy^{h}_{f,i,t}$, that represent the prediction performances of machines and humans, respectively. Comparing the estimated coefficients associated with the independent variables, we can see how the respective prediction powers of machines and humans vary according to the change in determinants (Table A1).

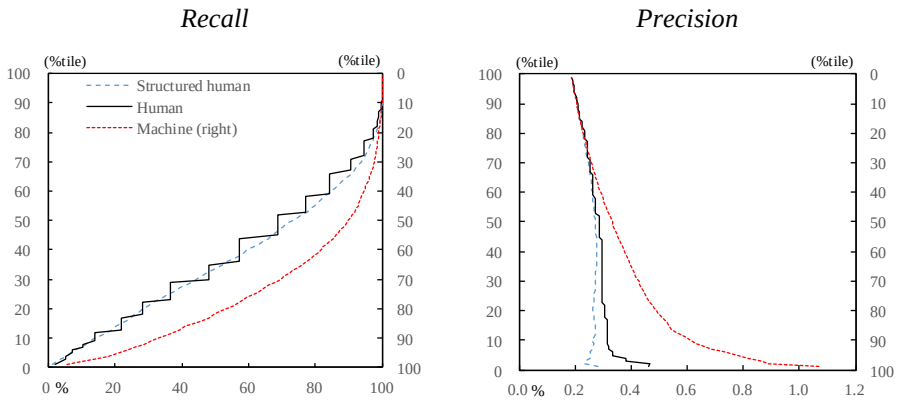
(A1) $$Proxy^{m}_{f,t} = Outcome^{ML}_{f,t} - 1 \text{ for exit firms,}$$
$$= 1 - Outcome^{ML}_{f,t} \text{ for non-exit firms,}$$

(A2) $$Proxy^{h}_{f,i,t} = Outcome^{H}_{f,i,t} - 1 \text{ for exit firms,}$$
$$= 1 - Outcome^{H}_{f,i,t} \text{ for non-exit firms.}$$

Third, we construct a set of rankings based on the machine prediction, *fscore*, and structured *fscore* and repeat the same estimation for the disagreement (Table A2). Fourth, we also define a dummy variable that equals one if $Proxy_{f,i,t}$ is positive and zero otherwise. Then we run a linear probability model and conditional logit model (Table A3). We also set 1 to 10 variables, which depend on the level of $Proxy_{f,i,t}$, and run an ordered-logit estimation (Table A4). Fifth, we replace the analyst-level fixed effect with the analyst-year-level fixed effect (Table A5). Sixth, we use one of the sub-scores of *fscore*, which represents the stability of each firm, instead of the total *fscore*, so that the target of human predictions becomes plausibly more comparable to that of machine predictions (Table A6). Seventh, we summarize the results of the proxy estimation and counterfactual exercise representing firm growth (Table A7). Eighth, we repeat the AUC estimation and proxy estimation based on the two alternative methods (i.e., LASSO and extreme gradient boost) (Table A8, A9). All the results are consistent with the ones we presented in the study.

Figure A1: Recall and precision measures over different thresholds

Default (test year: *t*=2016)



Voluntary closure (test year: *t*=2016)

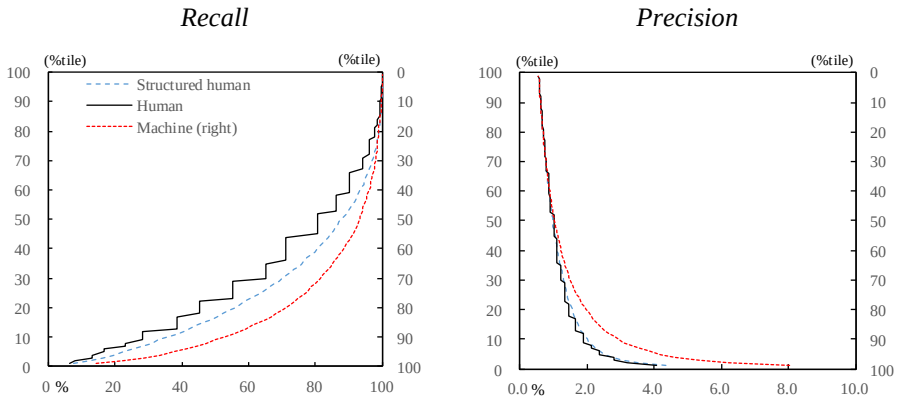


Table A1: Prediction performance of machines and humans

	default				***voluntary closure***			
	Machine		***Human***		***Machine***		***Human***	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
Number of available variables								
#(*available variables*) $_{f,t}$	0.102	0.000	0.008	0.000	0.118	0.000	0.012	0.000
Firm characteristics								
log(*sales* $_{f,t}$)	2.318	0.020	5.024	0.014	6.461	0.021	7.493	0.021
log(*sales* $_{f,t}$) - log(*sales* $_{f,t-1}$)	1.701	0.015	-0.440	0.011	0.231	0.017	-0.760	0.016
listed $_{f,t}$	2.477	0.443	2.621	0.303	-1.838	0.481	2.168	0.467
#(*industry*) $_{f,t}$	-0.502	0.025	0.099	0.017	0.244	0.027	0.202	0.027
priority $_{f,t}$			0.000	0.000			0.000	0.000
Analyst characterstics								
#(*assigned companies*) $_{i,t}$			0.000	0.000			0.000	0.000
industry experience $_{f,i,t}$			-0.000	0.000			-0.000	0.000
Team characteristics								
#(team members) $_{i,t}$			0.002	0.001			-0.005	0.002
Average #(*tenure years*) $_{i,t}$			0.014	0.002			0.016	0.003
Average industry experience $_{f,i,t}$			-0.000	0.000			0.000	0.000
Average #(*assigned companies*) $_{i,t}$			0.000	0.000			0.000	0.000
Constant	29.191	0.226	-4.012	0.166	-19.798	0.245	-28.631	0.256
Firm fixed-effect	yes		yes		yes		yes	
Analyst fixed-effect	yes		yes		yes		yes	
Year fixed-effect	yes		yes		yes		yes	
#(obs)	3,756,803		3,238,817		3,756,803		3,238,817	
F	53,485.400		15,304.020		78,182.190		14,025.710	
Adj R-squared	0.815		0.897		0.876		0.866	
Within R-squared	0.092		0.075		0.129		0.069	

Table A2: Rank-based disagreement estimation

	Machine vs. Human			
	default		***voluntary closure***	
	Coef.	S.E.	Coef.	S.E.
Number of available variables				
#(*available variables*) $_{f,t}$	1,607.929	4.271	1,527.788	3.784
Firm characteristics				
log(*sales* $_{f,t}$)	-58,115.530	374.526	-25,088.000	331.840
log(*sales* $_{f,t}$) - log(*sales* $_{f,t-1}$)	37,273.310	287.922	16,041.170	255.107
listed $_{f,t}$	27,956.380	8,164.855	-34,210.110	7,234.288
#(*industry*) $_{f,t}$	-8,595.519	471.108	620.723	417.415
priority $_{f,t}$	5.258	1.144	8.109	1.013
Analyst characterstics				
#(*assigned companies*) $_{i,t}$	-1.894	0.313	-3.357	0.277
industry experience $_{f,i,t}$	-11.528	0.604	-6.217	0.535
Team characteristics				
#(team members) $_{i,t}$	268.315	34.572	346.771	30.632
Average #(*tenure years*) $_{i,t}$	384.545	48.371	-63.242	42.858
Average industry experience $_{f,i,t}$	39.630	2.346	-2.152	2.079
Average #(*assigned companies*)	-2.936	0.437	-5.742	0.387
Constant	470,115.500	4,475.366	125,805.500	3,965.298
Firm fixed-effect	yes		yes	
Analyst fixed-effect	yes		yes	
Year fixed-effect	yes		yes	
#(obs)	3,238,817		3,238,817	
F	13,426.970		13,873.310	
Adj. R-squared	0.876		0.820	
Within R-squared	0.067		0.069	

Table A3: Dummy variable measure for disagreement

(1) Linear probability model

	Machine vs. Human			
	default		***voluntary closure***	
	Coef.	S.E.	Coef.	S.E.
Number of available variables				
#(*available variables*) $_{f,t}$	0.157	0.001	0.265	0.001
Firm characteristics				
log(*sales* $_{f,t}$)	-5.664	0.076	-3.578	0.085
log(*sales* $_{f,t}$) - log(*sales* $_{f,t-1}$)	4.064	0.059	2.315	0.065
listed $_{f,t}$	2.856	1.664	-7.332	1.849
#(*industry*) $_{f,t}$	-1.350	0.096	0.042	0.107
priority $_{f,t}$	0.001	0.000	0.002	0.000
Analyst characterstics				
#(*assigned companies*) $_{i,t}$	-0.000	0.000	-0.001	0.000
industry experience $_{f,i,t}$	-0.001	0.000	-0.000	0.000
Team characteristics				
#(team members) $_{i,t}$	0.041	0.007	0.041	0.008
Average #(*tenure years*) $_{i,t}$	0.005	0.010	0.005	0.011
Average industry experience $_{f,i,t}$	0.006	0.000	0.000	0.001
Average #(*assigned companies*) $_{i,t}$	-0.001	0.000	-0.001	0.000
Constant	93.738	0.912	59.737	1.014
Firm fixed-effect	yes		yes	
Analyst fixed-effect	yes		yes	
Year fixed-effect	yes		yes	
#(obs)	3,238,817		3,238,817	
F	3,135.790		6,343.690	
Adj. R-squared	0.721		0.659	
Within R-squared	0.016		0.033	

(2) Conditional logit model

	Machine vs. Human			
	default		***voluntary closure***	
	Coef.	S.E.	Coef.	S.E.
Number of available variables				
#(*available variables*) $_{f,t}$	1.942	0.013	2.587	0.012
Firm characteristics				
log(*sales* $_{f,t}$)	-87.264	1.207	-42.894	1.011
log(*sales* $_{f,t}$) - log(*sales* $_{f,t-1}$)	65.887	0.962	28.807	0.783
listed $_{f,t}$	45.617	25.010	-82.705	20.077
#(*industry*) $_{f,t}$	-20.860	1.326	-6.271	1.235
priority $_{f,t}$	0.095	0.014	0.072	0.008
Analyst characterstics				
#(*assigned companies*) $_{i,t}$	0.000	0.001	0.000	0.000
industry experience $_{f,i,t}$	0.006	0.001	-0.002	0.001
Team characteristics				
#(team members) $_{i,t}$	0.425	0.071	0.409	0.065
Average #(*tenure years*) $_{i,t}$	-0.241	0.114	-0.067	0.104
Average industry experience $_{f,i,t}$	0.022	0.006	-0.104	0.005
Average #(*assigned companies*) $_{i,t}$	-0.003	0.001	-0.002	0.001
Constant				
Firm fixed-effect	yes		yes	
Analyst fixed-effect	no		no	
Year fixed-effect	no		no	
#(obs)	736,498		922,303	
Log-likelihood	-259,176.670		-315,385.000	
χ-squared	30,953.570		57,174.730	

Table A4: Ordered logit estimation

	Machine vs. Human			
	default		***voluntary closure***	
	Coef.	S.E.	Coef.	S.E.
Number of available variables				
#(*available variables*) $_{f,t}$	1.214	0.005	2.262	0.005
Firm characteristics				
log(*sales* $_{f,t}$)	-171.686	0.244	-22.596	0.210
log(*sales* $_{f,t}$) - log(*sales* $_{f,t-1}$)	103.072	0.390	26.065	0.366
listed $_{f,t}$	542.157	6.472	-103.528	5.877
#(*industry*) $_{f,t}$	-48.697	0.389	-1.500	0.385
priority $_{f,t}$	0.086	0.003	0.010	0.002
Analyst characterstics				
#(*assigned companies*) $_{i,t}$	0.001	0.000	-0.001	0.000
industry experience $_{f,i,t}$	0.047	0.001	0.032	0.001
Team characteristics				
#(team members) $_{i,t}$	2.314	0.028	2.805	0.028
Average #(*tenure years*) $_{i,t}$	-0.375	0.049	-0.498	0.049
Average industry experience $_{f,i,t}$	0.255	0.002	0.297	0.002
Average #(*assigned companies*) $_{i,t}$	-0.030	0.000	-0.041	0.000
Constant				
Firm fixed-effect	no		no	
Analyst fixed-effect	no		no	
Year fixed-effect	no		no	
#(obs)	3,466,611		3,466,611	
Log-likelihood	-6,008,220.100		-6,508,573.100	
χ-squared	621,072.400		253,758.480	

Table A5: Alternative fixed-effects specification

	Machine vs. Human			
	default		***voluntary closure***	
	Coef.	S.E.	Coef.	S.E.
Number of available variables				
#(*available variables*) $_{f,t}$	0.571	0.001	0.482	0.001
Firm characteristics				
log(*sales* $_{f,t}$)	-19.063	0.125	-8.293	0.111
log(*sales* $_{f,t}$) - log(*sales* $_{f,t-1}$)	13.213	0.096	5.074	0.085
listed $_{f,t}$	-4.449	2.732	-19.247	2.412
#(*industry*) $_{f,t}$	-3.538	0.158	0.002	0.140
priority $_{f,t}$	0.000	0.000	0.002	0.000
Analyst characterstics				
#(*assigned companies*) $_{i,t}$				
industry experience $_{f,i,t}$	0.001	0.000	0.000	0.000
Team characteristics				
#(team members) $_{i,t}$				
Average #(*tenure years*) $_{i,t}$				
Average industry experience $_{f,i,t}$	0.017	0.001	0.000	0.001
Average #(*assigned companies*) $_{i,t}$				
Constant	157.847	1.465	49.298	1.293
Firm fixed-effect	yes		yes	
Analyst-Year fixed-effect	yes		yes	
Year fixed-effect	yes		yes	
#(obs)	3,238,266		3,238,266	
F	22,197.050		18,409.250	
Adj. R-squared	0.882		0.834	
Within R-squared	0.073		0.061	

Table A6: Using sub-score as human predictions

	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*) $_{f,t}$	0.637	0.002	0.018	0.000	0.519	0.002	0.018	0.000
Firm characteristics								
log(*sales* $_{f,t}$)	5.178	0.191	3.120	0.044	13.864	0.166	3.240	0.044
log(*sales* $_{f,t}$) - log(*sales* $_{f,t-1}$)	17.783	0.142	-2.203	0.033	13.444	0.123	-2.283	0.033
listed $_{f,t}$	8.962	3.434	4.606	0.787	-9.880	2.974	4.304	0.787
#(*industry*) $_{f,t}$	-2.132	0.227	0.090	0.052	1.092	0.197	0.086	0.052
priority $_{f,t}$	0.000	0.000	0.000	0.000	0.001	0.000	-0.000	0.000
Analyst characterstics								
#(*assigned companies*) $_{i,t}$	-0.002	0.000	0.000	0.000	0.000	0.000	0.001	0.000
industry experience $_{f,i,t}$	-0.003	0.000	0.001	0.000	0.002	0.000	0.001	0.000
Team characteristics								
#(team members) $_{i,t}$	0.028	0.019	-0.017	0.004	0.026	0.017	-0.018	0.004
Average #(*tenure years*) $_{i,t}$	0.080	0.026	-0.046	0.006	-0.078	0.022	-0.047	0.006
Average industry experience $_{f,i,t}$	0.026	0.001	-0.002	0.000	-0.005	0.001	-0.002	0.000
Average #(*assigned companies*) $_{i,t}$	0.001	0.000	0.000	0.000	-0.001	0.000	0.000	0.000
Constant	-132.004	2.359	-38.266	0.540	-212.930	2.044	-39.522	0.540
Firm fixed-effect	yes		yes		yes		yes	
Analyst fixed-effect	yes		yes		yes		yes	
Year fixed-effect	yes		yes		yes		yes	
#(obs)	2,199,518		2,199,518		2,199,518		2,199,518	
F	10,515.140		719.200		11,101.810		752.040	
Adj. R-squared	0.825		0.712		0.830		0.718	
Within R-squared	0.081		0.006		0.085		0.006	

Table A7: Growth prediction

(1) *Proxy* estimation

	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
Firm characteristics				
log(*sales* $_{f,t}$)	-50.833	0.229	-0.166	0.039
log(*sales* $_{f,t}$) - log(*sales* $_{f,t-1}$)	14.032	0.174	-0.439	0.030
listed $_{f,t}$	-24.028	4.837	3.056	0.830
#(*industry*) $_{f,t}$	-1.239	0.281	0.036	0.048
priority $_{f,t}$	0.005	0.001	0.000	0.000
Analyst characterstics				
#(*assigned companies*) $_{i,t}$	-0.000	0.000	-0.000	0.000
industry experience $_{f,i,t}$	0.003	0.000	0.000	0.000
Team characteristics				
#(team members) $_{i,t}$	-0.167	0.021	-0.008	0.004
Average #(*tenure years*) $_{i,t}$	-0.357	0.029	-0.014	0.005
Average industry experience $_{f,i,t}$	-0.017	0.001	0.000	0.000
Average #(*assigned companies*) $_{i,t}$	0.001	0.000	-0.000	0.000
Constant	574.761	2.737	-0.627	0.470
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	

(2) Counterfactual exercise

(a) Firms that actually do not grow ex post

	M = growth H = not growth (1)	M = not growth H = growth (2)	(2)/(1)
~20 %tile	12,799	30,678	2.40
20~40 %tile	15,822	38,401	2.43
40~60 %tile	18,513	31,610	1.71
60~80 %tile	25,171	22,727	0.90
80 %tile~	34,835	11,263	0.32

(b) Firms that actually grow ex post

	M = growth H = not growth (3)	M = not growth H = growth (4)	(3)/(4)
~20 %tile	1765	791	2.23
20~40 %tile	2170	978	2.22
40~60 %tile	2660	883	3.01
60~80 %tile	3599	760	4.74
80 %tile~	5308	401	13.24

Table A8: AUCs of alternative prediction models for default

Test data: t = 2013

Model	LASSO	XGBoost
Human	0.634 (0.0049)	
Machine	0.783 (0.0042)	0.807 (0.0039)
Structured human	0.529 (0.0047)	0.598 (0.0046)
Machine & *fscore*	0.806 (0.0040)	0.823 (0.0037)
Machine with small information	0.746 (0.0046)	0.783 (0.0043)

Test data: t = 2014

Model	LASSO	XGBoost
Human	0.639 (0.0052)	
Machine	0.774 (0.0047)	0.787 (0.0044)
Structured human	0.537 (0.0051)	0.558 (0.0096)
Machine & *fscore*	0.798 (0.0044)	0.815 (0.0042)
Machine with small information	0.740 (0.0051)	0.768 (0.0049)

Test data: t = 2015

Model	LASSO	XGBoost
Human	0.653 (0.0055)	
Machine	0.774 (0.0049)	0.804 (0.0044)
Structured human	0.547 (0.0053)	0.500 (0.0115)
Machine & *fscore*	0.804 (0.0046)	0.818 (0.0044)
Machine with small information	0.735 (0.0054)	0.772 (0.0050)

Test data: t = 2016

Model	LASSO	XGBoost
Human	0.663 (0.0053)	
Machine	0.779 (0.0049)	0.786 (0.0046)
Structured human	0.563 (0.0054)	0.516 (0.0111)
Machine & *fscore*	0.803 (0.0046)	0.810 (0.0045)
Machine with small information	0.738 (0.0054)	0.767 (0.0049)

Note: Each number represents the AUC, and the number in the parentheses is its standard error.

Table A9: Proxy estimation based on alternative prediction models

(1) LASSO

	Machine vs. Human		***SH vs. Human***	
	Coef.	S.E.	Coef.	S.E.
Number of available variables				
#(*available variables*) $_{f,t}$	0.495	0.002	0.150	0.001
Firm characteristics				
log(*sales* $_{f,t}$)	-12.859	0.146	10.266	0.082
log(*sales* $_{f,t}$) - log(*sales* $_{f,t-1}$)	17.666	0.113	-1.179	0.063
listed $_{f,t}$	59.775	3.193	4.973	1.792
#(*industry*) $_{f,t}$	-4.934	0.184	-0.769	0.103
priority $_{f,t}$	0.007	0.000	0.001	0.000
Analyst characterstics				
#(*assigned companies*) $_{i,t}$	-0.001	0.000	-0.001	0.000
industry experience $_{f,i,t}$	-0.001	0.000	-0.000	0.000
Team characteristics				
#(team members) $_{i,t}$	0.112	0.014	0.009	0.008
Average #(*tenure years*) $_{i,t}$	0.123	0.019	0.016	0.011
Average industry experience $_{f,i,t}$	0.009	0.001	-0.005	0.001
Average #(*assigned companies*) $_{i,t}$	-0.001	0.000	-0.001	0.000
Constant	97.460	1.750	-130.928	0.982
Firm fixed-effect	yes		yes	
Analyst fixed-effect	yes		yes	
Year fixed-effect	yes		yes	
#(obs)	3,238,817		3,238,817	
F	9,181.380		4,103.740	
Adj. R-squared	0.841		0.832	
Within R-squared	0.047		0.021	

(2) Extreme gradient boost

	Machine vs. Human		***SH vs. Human***	
	Coef.	S.E.	Coef.	S.E.
Number of available variables				
#(*available variables*) $_{f,t}$	0.449	0.003	0.075	0.004
Firm characteristics				
log(*sales* $_{f,t}$)	0.298	0.264	2.947	0.348
log(*sales* $_{f,t}$) - log(*sales* $_{f,t-1}$)	12.878	0.203	-0.930	0.268
listed $_{f,t}$	-5.342	5.763	-24.407	7.592
#(*industry*) $_{f,t}$	-3.276	0.333	-5.364	0.438
priority $_{f,t}$	-0.051	0.001	-0.123	0.001
Analyst characterstics				
#(*assigned companies*) $_{i,t}$	0.002	0.000	-0.001	0.000
industry experience $_{f,i,t}$	-0.008	0.000	0.010	0.001
Team characteristics				
#(team members) $_{i,t}$	0.768	0.024	0.392	0.032
Average #(tenure years) $_{i,t}$	0.508	0.034	0.139	0.045
Average industry experience $_{f,i,t}$	-0.035	0.002	-0.020	0.002
Average #(assigned companies) $_{i,t}$	-0.005	0.000	-0.006	0.000
Constant	-52.916	3.159	-27.909	4.161
Firm fixed-effect	yes		yes	
Analyst fixed-effect	yes		yes	
Year fixed-effect	yes		yes	
#(obs)	3,238,817		3,238,817	
F	2,886.910		1,230.400	
Adj. R-squared	0.506		-0.042	
Within R-squared	0.015		0.007	

Disagreement between Human and Machine Predictions

Oct 21st, 2021

IFC and Bank of Italy Workshop on

"Data Science in Central Banking"

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
- Use ML for prediction
  - Successful in general
    - 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 ≻ Human" on average (⇔ They disagree)

Our 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 do

A) Construct a ML-based prediction model
- Massive size of firm-level data w/ high dimension information
- Various outcomes (default + voluntary exit + sales growth)

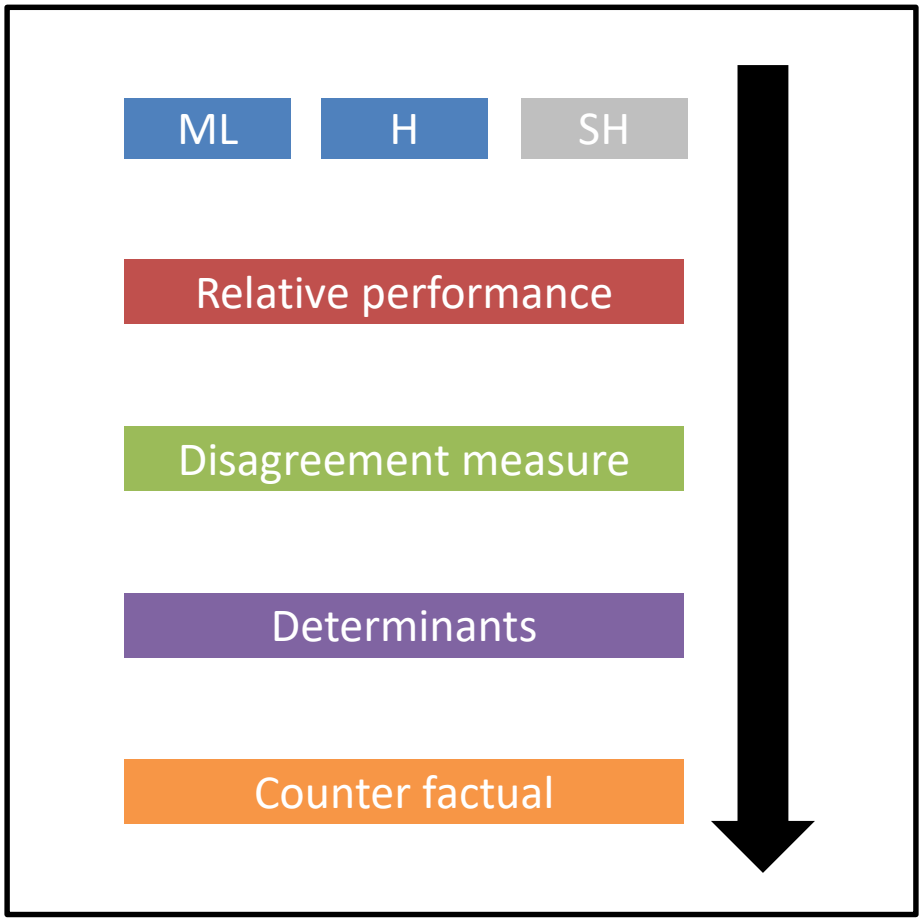
B) Measure the disagreement b/w ML & Human
- Human = Credit rating made by analysts
- Vs. Machine or "Structured" human
- "*Proxy*" ↑ (↓) ⇔ ML works better (worse)

C) Examine how opaqueness works as its determinants
- Firms' informational opaqueness
- Controlling for various attributes as much as possible

D) Do a counterfactual exercise for task allocation
- Improve prediction power by allocating tasks to M & H

Organization of the paper

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



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)

Method: Disagreement

Disagreement measure

- *Proxy*: Measuring the "disagreement"
  - Predict firms' outcome with test data by M & H & Structured H
    - 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

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

Method: Disagreement

Disagreement measure

- *Proxy*: Measure the disagreement

  - Large ⇔ M or SH ≻ H

  - M vs H

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

  - Structured H vs H

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

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

4-3. Result: Determinants

Determinants

- Higher opaqueness ⇒ M ≺ H
- Same pattern for SH ≺ 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)

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

Key takeaways

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

- "ML $\succ$ Human $\succ$ Predicted human"
  - $\neq$ 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)

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
Bank of Japan
2-1-1 Nihombashi-Hongokucho, Chuo-ku, Tokyo 103-8660 Japan
E-mail: kouhei.shintani@boj.or.jp