Model Selection for Optimal Screening with AI in Hiring

DRAFT: Do not copy, cite, or distribute without permission of the authors

Achyuta Adhvaryu, Jean-François Gauthier,
Yuqing Gu, Anant Nyshadham, Jorge Tamayo*

May 24, 2022

Abstract

The cost of making bad hiring decisions leads many employers to use personality tests as a first round screening. Problems arise when different misclassification errors incur different penalties. When facing a limited labor supply or high benefit-cost ratio of the vacancy, firms tend to avoid dropping potential candidates even if this means they have to include some unsuitable candidates and vice versa. Thus, it is important to design a tool that could balance these two types of errors based on both market tightness and the benefit-cost ratio of the position. In this paper we introduce a cost-sensitive machine learning framework that minimizes hiring cost under different conditions. We empirically show our cost-sensitive learning approach could achieve a lower hiring cost compared with other methods.

Keywords: Hiring, Cost-Sensitive, Personality, Performance, Machine Learning

*Adhvaryu: University of Michigan, NBER, & BREAD, 701 Tappan Ave, Ann Arbor, MI 48109 (email: adhvaryu@umich.edu); Gauthier: Boston College, 140 Commonwealth Avenue, Maloney Hall, 333A, Chestnut Hill, MA, 02467 (email: gauthija@bc.edu); Gu: University of Michigan; Nyshadham: University of Michigan & NBER, 701 Tappan Ave, Ann Arbor, MI 48109 (email: nyshadha@umich.edu); Tamayo: Harvard University, Harvard Business School, Morgan Hall 292, Boston, MA, 02163 (email: jtamayo@hbs.edu).

1 Introduction

Studies show that personality measures can predict occupational performance in a wide variety of jobs. Hogan and Holland (2003) summarized that well-constructed measures of normal personality are valid predictors of a wide range of occupational performance metrics. For example, Dollinger and Orf (1991) found that conscientiousness is a successful predictor of college student's course grade. Barrick and Mount (1991) found that three personality dimensions, emotional stability, openness, and agreeableness, explained 28% of the variation in participants' management performance. Salgado and Rumbo (1997) found that Big Five traits are strong predictors of job performance for financial service managers. Previous studies have paid attention to how personality traits predict managers' or students' performances. Here, we focus on the predictive power of personality tests on the performance of online customer service workers. We conducted a personality survey via Mturk[1] and then gave these surveyed workers three types of customer service tasks: objective tasks, judgment tasks, and review tasks and scored their performance on each of them. We then use machine learning methods to predict the score workers obtained in customer service tasks according to their personalities.

Our aim is to understand how firms can make use of psychometric screening in their hiring processes in order to lower hiring costs. We consider firms that receive applications for vacancies from a large number of candidates and that aim to reduce the size of this pool by way of machine learning screening. Applicants are *a priori* too similar to distinguish qualified from unqualified ones by simply looking at their résumés. To tell these two types of applicants apart, an interview is needed.[2] Firms have idiosyncratic interview costs and benefits of filling a vacancy. Moreover, qualified job seekers may be few and far between depending on the state of the economy. Hence, a firm that wants

[1] Amazon Mechanical Turk (MTurk) is a website for businesses to hire remote workers in order to do specific tasks.

[2] We assume that firms are able to separate qualified from unqualified applicants with probability one at the interview stage.

to build an efficient screening framework needs to take these elements into account. In fact, we show that these elements play a key role in the choice of machine learning model to use, with the model choice often changing with the state of the economy for the same firm.

The abundance of quality applicants, or lack thereof, is particularly interesting and novel in the machine learning literature. When the market is tight and qualified job seekers are rare, a screening algorithm that is too restrictive may falsely reject too many applicants and fail to escalate the few truly qualified applicants that applied for the vacancy. Instead, when qualified applicants are abundant, a more restrictive algorithm may be just what the firm needs as it will escalate enough, but not too many, applicants to the interview round where a sufficient number of truly qualified applicants will be discovered and hired. Because rejecting qualified applicants by mistake (False Negative) and wrongfully interviewing unqualified applicants (False Positive) bear different costs, one can't simply balance these two types of misclassification errors to yield an efficient framework. Indeed, a framework must be sensitive to the differences in cost entailed by the different type of errors embedded in machine learning models.

Cost-sensitive imbalanced machine learning has been a widely studied problem in some other branches of the literature (Elkan, 2001; Zadrozny et al., 2003; He and Garcia, 2009; Buda et al., 2017), as many real-world problems have specific costs relating to each type of misclassification errors. For example, cost-sensitive learning has been used in credit scoring (Xia et al., 2017), where the cost of incorrectly loaning money to a defaulting individual is different to the cost of not loaning money to a person who will never default; in software defect prediction (Liu et al., 2014), where misclassifying defect prone components leads to a higher cost than misclassifying non-defect prone components; in churn modelling, where failing to identify a profitable or unprofitable churner has different economic implications. One interesting application that has received much discussion is the credit card fraud detection problem, which has similarities with our study. Banks screen

on transactions to decide whether to classify some transactions as being suspicious that can then be further investigated. Banks have to pay the investigation cost if they falsely tag normal transactions as being suspicious and bear an often large money loss if they fail to investigate a fraudulent transaction. Therefore, missing a fraudulent transaction can be much more costly than paying the investigation cost on a normal transaction misclassified as being potentially fraudulent. Banks have to design screening frameworks to minimize the expected total cost under different conditions. Studies show that many machine learning techniques can be used to minimize this expected cost. Sahin et al. (2013) designed a cost-sensitive decision tree based on cost related entropy. Stolfo et al. (2000) found that AdaBoost model is predictive in cost-sensitive problems. Besides, neural networks (Maes et al., 2002), Bayesian learning (Maes et al., 2002), support vector machines (Bahnsen et al., 2013) are also widely used algorithms that have found success in cost-sensitive scenarios. In this paper we partition the dataset in three subsets: a training set, a validation set, and a testing set. We use the training data in order to fit the different models, the validation data to choose the cost-minimizing model, and the testing set to report the value of the cost function for the chosen model. In addition to proposing particular algorithms, Bahnsen et al. (2013) suggested a Bayes Minimization Risk function to minimize expected total cost in credit card fraud detection problems. In our paper, we design a cost matrix that takes into account both firm idiosyncratic interview cost and benefit of filling a vacancy, as well as the labor market tightness and construct a Bayes Minimization Risk function accordingly. Then we propose associated classification thresholds that can be used reach the lowest cost for each algorithm.

The remainder of the paper is organized as follows: In section 2 we introduce our main methodology and explain our model selection scheme. In section 3, we describe our dataset. In section 4 we present the main empirical strategy used in the work, and in Section 5 we present the paper's main results. Lastly, in Section 7 we draw our main conclusions for the work.

2 Cost-Sensitive Design: Optimal classification threshold

2.1 Model

In the hiring market, firms post some vacancies and receive a large number of applicants for a limited number of positions. Since it is costly for firms to screen every single participant through interviews, employers may benefit from carefully designed psychometric-screening procedures to select the most qualified applicants to interview.[3] If psychometric screening is cheaper than interviews, then the net total hiring cost for a particular vacancy is likely to fall for two reasons:[4] first, psychometric screening can help identify qualified individuals who may have been ignored otherwise, and second, it may reduce the number of interviews to conduct. In this section, we first describe the problem a firm faces using psychometric screening to reduce the number of applicants and then interview the retained applicants. Next, we explain the data requirements needed to build a psychometric-screening machine learning algorithm. Finally, we provide a strategy to choose between algorithms. We show that the choice depends on the cost structure and the tightness of the labor market.

Here, we consider the psychometric screening process as a binary classification problem, where the screening procedure aims to predict whether an applicant is qualified for a task or not. For simplicity, we assume that firms can tell whether an applicant is qualified or not without error at the interview stage. However, the psychometric screening is inherently imperfect; thus, firms need to consider the two types of errors present in any classification algorithms, namely false positive and false negative, when designing such processes. A false negative occurs when a qualified applicant is not identified as such

[3]Certain psychometric characteristics can be correlated with certain demographics such as race and gender. Psychometric screening could potentially exacerbate disparities if used in an inappropriate manner. It goes without saying that sensitive characteristics such as gender, race, religion, sexual orientation, should not be used to train the screening algorithms. Second, a conscientious employer interested in using psychometric screening could do so *within* demographic groups to identify the most qualified individuals within each group and interview all retained individuals.

[4]The net hiring cost, it the benefit of hiring a person minus the cost of hiring that person.

by the algorithm and the applicant is not escalated to the interview round. Instead, a false positive means that an unqualified applicant is identified as being qualified and is escalated to the interview round.

These two errors will have different costs for the employer. False positives imply that an employer interviews an unqualified applicant and rejects that applicant post interview while bearing the interview cost. False positives, instead, mean that the firm misses out on qualified workers. We assume that the cost of missing out on a qualified worker depends on the tightness of the labor market. That is, it is more costly to fail to interview a qualified applicant when qualified individuals rarely apply to a vacancy than when quality applicants are relatively abundant.

Therefore, a good algorithm must take into account the relative cost of the classification errors in order to minimize the net total of hiring cost. Here, we assume that the cost of having an applicant do a psychometric survey is negligible to the firm. For simplicity, we also assume that all truly qualified applicants who make it to the interview round are hired.

The net total hiring cost then depends on the employer's idiosyncratic incremental benefit of filling a vacancy with a qualified worker instead of an unqualified worker, b, and its interview cost per applicant, t, as well as the relative abundance of quality applicants in local labor market. We assume that the number of qualified individuals apply follows a Poisson process. Let $q(\theta)$ be the (Poisson) arrival rate for a vacancy normalized to be between 0 and 1, where θ is the market tightness defined as the ratio of vacancies, v, to the number of workers unemployed in the local labor market, u.[5] A Poisson process means that in some given short time interval Δt, the probability that one vacancy will be filled with one matched worker is $q(\theta)\Delta t$. This probability increases with the unemployment rate μ and with the length of time interval Δt.

For example, if a single qualified applicant applies per period on average, then $q(\theta) =$

[5]We can parametrize the arrival rate as $q(\theta) = \eta\theta^{-\beta}$ where η is a measurement of job search frictions and β is the matching function elasticity (Moscarini, 2005; Rogerson et al., 2005).

1. Instead, if one quality applicant applies every 2 periods on average, then $q(\theta) = 0.5$. Then, we can interpret $q(\theta)$ as the probability of seeing a qualified applicant per period. As a result, $\lambda = 1 - q(\theta)$, is the probability of losing the benefit associated with hiring a qualified applicant if such an individual is wrongfully rejected in the psychometric screening phase. When the screening algorithm correctly classifies an individual as being qualified, the employer interviews this worker and the applicant is hired providing a benefit, b, to the firm. Table 1 summarizes the possible classifications of an applicant and the cost to the firm for each of these cases.

Note that the labor market conditions affect the screening process by affecting the opportunity cost of losing a potential talented employee. To be specific, if the job market is full of vacancies with relatively less job seekers or search frictions are large, it will be difficult for firms to find another suitable candidate if they wrongfully reject one. In other words, the cost of a false negative will be larger in a tight labor market. Take two extreme cases as examples. When $q(\theta) = 0$ and $\lambda = 1$, a representative firm will never have the chance to meet such person if the firm turns down a qualified candidate so the firm will lose the benefit of hiring a qualified individual (b) with probability one. On the other hand, when $q(\theta) = 1$ and $\lambda = 0$, firms are sure to get an application from a qualified worker later. In this case, the cost of wrongfully rejecting a qualified worker is negligible.

The specification above allows us to write the net expected cost for applicant i as:

$$Total\ Cost = TP \times (t - b) + FP \times t + FN \times \lambda b, \tag{1}$$

where TP, denotes True Positives, FP, False Positives, TN, True Negatives, and, FN, False Negatives. Note that $TP = 1$ if the applicant is truly qualified and identified as such by the algorithm and 0 otherwise, $FP = 1$ if the applicant is not qualified, but wrongly classified as being qualified by the algorithm and 0 otherwise. Finally, $FN = 1$ if the applicant is unqualified and correctly identified as such by the algorithm and 0 otherwise.

Here, we assume that the employer hires all truly qualified applicant that make it to the interview stage. This is consistent with large firms with a large number of entry-level jobs facing an even larger pool of applicant such as warehouse personnel at *Amazon* or drivers at *Uber*. We can easily imagine that it may be hard to predict an applicant's suitability for these jobs from their résumé alone, since many may not have experience in these occupations for example.[6] Following machine learning literature conventions, we can also write the cost function above for individual i as a loss function $L(p_i, y_i)$:

$$L(p_i, y_i) = y_i \left[p_i(t - b) + (1 - p_i)b\lambda \right] + (1 - y_i)p_i t, \tag{2}$$

where y_i equals 1 if worker i is qualified for the position and 0 otherwise. Similarly, p_i equals 1 if i is predicted to be qualified by the algorithm and 0 otherwise.

2.2 Optimal threshold

In the psychometric screening phase, we aim to predict whether an applicant is qualified for the position based on their psychometric profile and experience. In the present paper, we rely on soft classification algorithms and obtain the class conditional probabilities, that is, the probability that a given individual is qualified or not.[7] Then, we choose a classification threshold that minimizes the expected cost of making classification decision. A classification threshold is a performance level (in estimated probability of being qualified) above which an individual is deemed to be qualified for a task. For example, imagine that

[6] Suppose that the firm has ν vacancies to fill. Let $\#TP$, and $\#FN$ be the number of true positives and false negative respectively, and assume that all truly qualified workers have a uniform change of being hired post interview, then $Total\ Cost = TP \times (t - \frac{b\nu}{\#TP}) + FP \times t + FN \times \lambda \frac{b\nu}{\#TP+\#FN}$. Here, $\frac{\nu}{\#TP}$ is the probability of hiring this applicant if they are found to be truly qualified in the interview stage. $\frac{\nu}{\#TP+\#FN}$ is the probability that a qualified person is hired if all qualified individuals make it to the interview round.

[7] Among numerous classifiers, some are hard classifiers while some are soft ones. Soft classifiers explicitly estimate the class conditional probabilities and then perform classification based on estimated probabilities whereas hard classifiers directly target on the classification decision boundary without producing the probability estimation. These two types of classifiers are based on different philosophies and each has its own merits (Liu et al., 2011). Soft classification provides more information than hard classification and consequently it is desirable in our situation where the probability estimation is useful (Wang et al., 2007).

a firm has a performance metric for particular task ranging between 0 and 100%, where 100% means that someone is able to perform this task without mistake. A firm may want to hire applicants that it believes will do a task correctly at least 80% of the time. This threshold can be set extraneously, by the employer, but the threshold can also depend on the cost, benefit and labor market conditions to minimize the total expected hiring cost.

To find the cost minimizing productivity threshold, we first compute the expected cost, or *risk*, of escalating someone to the interview given the information set we have about the applicant, X_i, and the expected cost of rejecting an applicant in the psychometric screening round. The former can be written as:

$$R(p_i = 1|X_i) = L(p_i = 1, y_i = 1)P(y_i = 1|X_i) + L(p_i = 1, y_i = 0)P(y_i = 0|X_i). \quad (3)$$

Similarly, we can write the expected cost of rejecting an applicant in the first round as:

$$R(p_i = 0|X_i) = L(p_i = 0, y_i = 0)P(y_i = 0|X_i) + L(p_i = 0, y_i = 1)P(y_i = 1|X_i). \quad (4)$$

In the equations above, $P(y_i = 1|X_i)$ and $P(y_i = 0|X_i)$ are the probabilities that applicant i is truly qualified or not given their psychometric profile and experience, X_i. $L(p_i, y_i)$ is the loss function specified in Equation (1). The expected cost of predicting that applicant i is qualified given X_i, is given by $R(p_i = 1|X_i)$. Similarly, the expected cost of predicting that the applicant is unqualified given X_i, is $R(p_i = 0|X_i)$.

Given the psychometric and experience information, the optimal strategy is to promote applicant i to the interview stage if the expected cost of doing so exceeds the expected cost of rejecting the applicant in the psychometric screening phase. That is:

$$R(p_i = 1|x_i) < R(p_i = 0|x_i)$$

By using (1), (2), and (3), we obtain the following optimal threshold:

$$P(y_i = 1|X_i) > \frac{t}{(1+\lambda)b}$$

Thus, to minimize the total expected cost, individuals should be promoted to the interview stage if their conditional probability of being qualified exceeds $\frac{t}{(1+\lambda)b}$. This means an employer should interview more applicant if the interview cost is low, when qualified applicants are less abundant, and/or when the benefit of filling a vacancy is high.

3 Experiment and Data

We recruited workers on Amazon's online work platform, Mechanical Turk. Every worker was asked to complete an extensive psychometric survey in which we measured a wide array of soft and cognitive skills, as well as demographic information. Workers who participated in the study and completed the survey were then ask to perform tasks designed to mimic aspects of customer support occupations. We detail the survey and the tasks below. For the purpose of this study, we retain data from the 253 participants who completed all questions of the survey and attempted all tasks.

3.1 Survey Protocol

Participants were first asked to complete an extensive demographic and psychometric survey. They received \$2 for attempting the survey and an additional \$5 for answering most questions which implies \$8.5/hour pay on average.

The demographic section measures variables that may be more easily observed from CV's and résumés such as age, education, general experience and experience in customer services, and English proficiency.

The psychometric sections of the survey includes measures of skills, traits, and be-

haviors spanning cognitive and noncognitive abilities.[8] The survey consists of several different modules intended to measure both traditional dimensions of worker skill found to be associated with performance and additional modules on personality and risk and time preferences. Table 2 presents summary statistics for these measures. The different psychometric modules can be organized in the following categories:

- **Soft skills**: Recent empirical studies have begun to document the incremental importance of soft skills for earnings and productivity (Borghans et al., 2008; Heckman and Kautz, 2012). We include the Big 5 personality trait modules capturing conscientiousness, openness, extraversion, emotional stability or neuroticism, and agreeableness. We also measure the participants' self-esteem, grit and resilience, general motivation, and autonomy. We measure the participants' ability to read emotions and how strong they believe having the ability to control events or enact changes.[9]

- **Cognitive skills**: The literature on return to cognitive skills in productivity and earnings is long-standing and well-established (Boissiere et al., 1985; Bowles et al., 2001). To inform cognitive skills we use Raven's Progressive Matrices module, which is a common psychometric test capturing abstract reasoning and fluid intelligence.

- **Risk and time preferences**: In addition to the measures of soft and cognitive skills above, we measure participants' aversion towards risk and their patience.

3.2 Performance outcomes

Participants were told that they would be offered further and better-paid work (on average) upon completion of the survey.[10] The work in question consisted of tasks representa-

[8] see Boissiere et al. (1985); Rosenberg (1965); Borghans et al. (2008); Heckman and Kautz (2012); Duckworth and Steinberg (2015)

[9] The former skills are captured by the Reading-the-mind-in-the-eyes and locus of control psychometric modules.

[10] Participants could read the following at the beginning of the survey: *If you properly complete the survey (i.e., provide reasonable answers to all questions), you will also be invited to the second part of the study. The second*

tive of certain aspects of the work done in customer service occupations. We constructed three sets of tasks. In the first tasks, participants have to *objectively* report product information. In the second set of tasks, participants have to *judge* which product satisfies the requests of hypothetical customers. In the last tasks, have to *review* actual negative reviews written about a product and write a hypothetical *reply* to the customers. We detail the different tasks further below.

3.2.1 Objective tasks

In the objective tasks, participants are presented the product page of certain products on Amazon. They are then asked to list information about the products such as the price and sales price, the number of items left in stock, and so on. Performance is measured by the proportion of correct information listed by the participant to the total number of queries.

3.2.2 Judgment tasks

In these tasks, participants are presented information for 3 competing products listed on Amazon. For example, three different sleeping bags from three different brands. Then, in each task, they see five hypothetical requests from customers and have to suggest which product or products (if any) satisfy the customer requests. Below is an example of a request:

> *I'm 6 feet tall and my current sleeping bag is too small. I want a bag that will fit me and that will keep me warm bellow* $15°F$.

The participants have to find which product(s) (if any) fits the length and temperature requirements of the customer. Performance is assessed by the number of correctly satisfied requests to the number of requests in a given task.

part will pay significantly more on average than this survey. It will pay on average $10.40 per hour and up to $11.75 per hour depending on performance. The second part of the study consists of a set of customer service style tasks performed online in a similar survey environment.

3.2.3 Review/Reply tasks

In these tasks, participants are shown the product page of a product on Amazon. They are then shown four negative reviews that were actually left by customers who purchased the product. The participants are instructed to reply in a way that they believe to be likely to convince the review writer to continue buying from the product manufacturer in the future.[11] The participants have to decide which customers to write back to and follow specific rules in doing so. For example, they have to reply to at most three and at most two customers. Each reply has to respect a word count. Up to one customer may be offered a refund, and if so, the participant should explain that they are offering a refund and are escalating their case to a manager for processing.

Performance is assessed in two ways. We first compute the percentage of objective guidelines that were respected. Then, we hired three research assistants to score the quality of two replies chosen at random for each task and for each participant. To do so, the RA's had to suppose they were the one who wrote the initial comment, read the participant's reply, and answer to what degree they agreed with the three following statements using a five-point scale:

1. My concerns were heard and understood by the customer services.
2. The response from the customer service representative fully addressed my complaint(s).
3. I would buy from this manufacturer again and/or recommend the manufacturer's product to a friend.

We compute the total score on every reply for each RA and take the average total score across the three RA's as the final measure of quality.

[11] They were shown example of good and bad replies to reviews left on another product when we introduce that task, but we did not explain what made the replies good or bad.

In each task type, participants had to do at least 2 tasks that all had multiple questions. We present the densities of the outcomes in Figure 1. We see that there is substantial variation in performance across tasks.

4 Empirical Strategy

In this section, we formally bring the model's predictions to the data. The model predicts that applicants should be promoted to the interview stage if the conditional probability of being qualified exceeds the optimal cutoff, $\frac{t}{(1+\lambda)b}$, or the cutoff chosen by the firm.

In this section, we recover two exogenous parameters, λ, and the optimal cutoff needed for the machine learning algorithms. These two parameters should be determined by outside labor market conditions and firm preferences rather than the algorithms themselves. Here we present one way for firms to choose the parameters.

4.1 Labor market index (λ)

As we noted previously, λ is the labor market tightness index which measures the difficulty of finding an alternative qualified worker for a vacant position. Intuitively, as λ increases, qualified workers are more difficult to find and the expected cost of excluding an applicant from the interview round increases. λ depends on the number of qualified candidates in the market, the size of the labor supply, and the labor demand as measured by the number of vacancies.

In the labor market, there are vacancies needing to be filled and candidates looking for jobs. Intuitively, when the number of vacancies is large relative to the number of job seekers, then firms have to compete more intensely for qualified workers as they are relatively limited. As a result, the probability of receiving an application form a qualified individual is lower. If there are relatively few vacancies relative to the number of seekers, the probability of receiving an application from a qualified person is higher. Therefore,

the probability of qualified arrivals during a unit of time, q, is a function of market tightness, θ, defined as the ratio between the number of vacancies in the entire market and the number job seekers in the market.[12] Suppose that $q(\theta) = 0$ and a firm receives an application from a qualified person. Then, the firm will never have the opportunity to find another qualified worker if they wrongfully reject this applicant. This means the cost to the firms for a false negative prediction (i.e. wrongfully rejecting a qualified applicant) is equal to the benefit the firm would have received from hiring this foregone quality applicant, relative to hiring an unqualified applicant.

When $q(\theta) = 1$, the firm has a 100% change of receiving an application form another qualified workers if they were to reject the current applicant. In this case, the cost of wrongfully rejecting a quality applicant is close to 0.

Following this logic, if $q(\theta)$ increases by 1%, meaning that the firm is 1% more likely to find another quality applicant during the current period, the cost of a false negative to the firm will drop by 1%. This implies a negative linear relationship between λ and $q(\theta)$.

The tightness of a market varies over time as some firms and individuals start and end their job search process every period. Here, we consider the screening problem of a firm at a specific point in time. For simplicity, we use a constant λ averaged across several time periods and across all markets. This allows us to get a general idea of the cost for firms in the U.S labor market. That is, we let λ be the following:

$$\lambda = \frac{1}{T}\Sigma_{t=1}^{T}(1 - q(\theta_t)) \tag{5}$$

Next, we follow Shimer (2007) and recover the arrival rate of qualified workers during a unit of time, $q(\theta_t)$. Shimer (2007) develops the matching function measuring the transition rate from unemployment to employment as a function of vacancy-unemployment ratio, $q(\theta_t)$. In his approach, workers and jobs are randomly assigned to labor markets

[12] To be specific, we use Poisson arrival rate to model the arrival of qualified workers as Poisson process is the most commonly used model for random, mutually independent message arrivals.

and there exist markets with unemployment. In some markets there are more job seekers and vacancies, and in others, there are more vacancies than job seekers. He allows for frictions creating a mismatch between workers and firms in the labor market. Shimer's model is consistent with the U.S. Beveridge curve and yields a Cobb-Douglas matching function of the following form:

$$q(\theta_t) = a\theta_t^{-b}$$

, where θ_t is the tightness of the market defined as $\theta_t = \frac{v_t}{u_t}$. v_t is the vacancy rate in period t, and u_t is the unemployment rate in period t.[13] The parameter a represents the degree of search frictions, and b is the matching function elasticity. In Shimer's job flow model estimation, the matching function is derived as:

$$q(\theta_t) = 0.551\theta_t^{0.214} \tag{6}$$

In order to recover λ, we plug in the unemployment rate and the vacancy rate into equations (5) and (6). We construct unemployment rates as the ratio between unemployed workers and the number of employed and unemployed workers. We construct vacancy rates as the ratio of job openings to the number of employed workers and job openings. Employment and unemployment data is obtained from the Bureau of Labor Statistics (BLS) and the Current Population Survey (CPS), and the job opening data is obtained from the Job Openings and Labor Turnover Survey (JOLTS).[14] We focus on the pre-Covid period of Jan 2010 to Dec 2019.[15] Following this strategy, we get $\lambda = 0.5$ under Shimer's model. We also show results under less frictions with $\lambda = 0.25$ and more frictions with

[13] This should be the aggregate job searching rate instead of total unemployment rate in matching function. In actual use, however, the unemployment rate is common.

[14] We use seasonally adjusted data to eliminate trend effects.

[15] We want to get the most recent 10 years of data which should be enough to get a general sign of how well the overall labor market matched. We avoid Covid years because this is a natural shock and will disrupt the general view.

$\lambda = 0.75$.[16]

Following the same procedure for different markets would allow idiosyncratic firms to recover values of λ for their relevant market. As a starting point, firms that find it difficult to find applicants for a job may set $\lambda = 0.75$. If finding applicants is relatively easy, they may set $\lambda = 0.25$. When finding applicants for a job is neither difficult nor easy, they may set $\lambda = 0.5$.

4.2 Cutoffs

In order to train the classification models, we first need a sample of candidates in which we know who is qualified or not. Firms can observe a sample of existing workers with varying degrees of qualification, their performance, and their characteristics. They can, then, train the models using this sample. Once the models are trained, the characteristics of a new candidate can be inputted into the trained models which will predict whether this candidate is qualified or not. In the context at hand, we have candidates and their performance on different tasks and we need to find a cutoff for each task to decide which candidate is qualified and which is not. We select the cutoffs in a way that the proportion of workers said to be qualified in the sample reflects the proportion of truly qualified workers in the market.

Given the arrival rate of qualified workers, $\eta = q(\theta)$, which measures the average number of applicants arriving during a unit of time, we want to know the proportion of qualified workers in the whole pool of applicants. That is, with the distributions of productivity in a given task, we want to set a cutoff, above which workers will be considered qualified and below which they will be considered unqualified. This way, we obtain a classification problem for each task. For simplicity, we suppose that 1 worker arrives in each period. Therefore, it takes N periods to allow all workers to apply, where N is the

[16] 0.5 is what we estimate from the model. As we mentioned before, this just served as a general example for firms to consider. 0.25 and 0.75 are two arbitrary number we pick to test our model for less frictional and more frictional cases. Firms should set their own λ according to the labor market condition they face.

total number of workers looking for jobs.

The probability that n qualified workers apply within N periods, P_n, can be expressed as:

$$P_n = \frac{(\eta \cdot N)^n}{n!} e^{-\eta N}. \tag{7}$$

It can be showed that the number of qualified candidates in the sample of applicants is given by:

$$n = \arg\max_{n \in Z} Pn = [\eta N] = [(1 - \lambda)N].$$

Table 3 presents the distribution for performance in the different tasks. When a given λ is selected, then one should only consider the top $1 - \lambda$ candidates as being qualified under the current procedure. Recall that doing so will minimize the hiring cost in expectation. For example, if $\lambda = 0.75$ then the top 25% of performers should be considered as qualified. Alternatively, a firm may use a threshold that fits its productivity requirements, but may not lower the interview cost. For example, a firm with high productivity standards may consider only the top 10% of performers as being qualified, regardless of the market tightness.

5 Results

In this section, we present a strategy to select machine learning models in order to minimize the hiring cost of a firm.

5.1 Productivity threshold

We first show the implications of choosing productivity thresholds that depend on the abundance of quality job seekers in the market, relative to an exogenous threshold.

Let's consider the optimal market-tightness-driven threshold to a non-optimal thresh-

old of 0.5 chosen exogenously. That is, a threshold where applicants with predicted probability of being qualified that exceeds 50% are promoted to the interview round. By virtue of not taking into account the abundance of workers, this exogenous threshold will not minimize the interview cost.

To be specific, we can write whether an applicant is promoted to an interview or not under the suboptimal (S) and optimal (O) thresholds as follows:

$$\text{Promoted to Interview}^S = \begin{cases} 1 & \text{if } P(y_i = 1|x_i) > 0.5 \\ 0 & \text{if } P(y_i = 1|x_i) \leq 0.5 \end{cases} \tag{8}$$

$$\text{Promoted to Interview}^O = \begin{cases} 1 & \text{if } P(y_i = 1|x_i) > \frac{t}{(1+\lambda)b} \\ 0 & \text{if } P(y_i = 1|x_i) \leq \frac{t}{(1+\lambda)b} \end{cases} \tag{9}$$

In equations (8) and (9), $P(y_i = 1|x_i)$ represents the probability that an applicant is truly qualified estimated by a given model (conditional on the variables included in the model.) For example, the first line of equation (9 indicates that an applicant is promoted to an interview if their probability of being truly qualified estimated by the model exceeds $\frac{t}{(1+\lambda)b}$.

When $\lambda = 0.25, b/t = 1.5$, the optimal classification threshold becomes $\frac{t}{(1+\lambda)b} = 0.83$ while the suboptimal threshold remains unchanged, i.e., 0.5.

At these parameter values, quality applicants arrive relatively frequently. Therefore, there is less of a need to interview and hire individuals with a low probability of being truly qualified. On the other hand, when $\lambda = \frac{2}{3}, b/t = 1.5$, the optimal threshold becomes 0.4, reflecting the fact that quality applicants are rarer and more applicants need to be interviewed. The selection result of each model with and without model search is listed in Table 4.

We find that going from the suboptimal threshold (0.5) to the optimal threshold (0.83),

the number of false positives decreases. This is because the probability of being misclassified as qualified is 34% smaller (17% vs 50%). On the other hand, since we are more selective under the optimal threshold, the number of true positive has to fall reflecting the compromise between true positives and false positives in classification models. This compromise benefits the firm however. Since truly qualified applicants are fairly frequent, it is worth failing to interview all truly qualified applicants. Indeed, we can see that the total interview cost is generally lower under the optimal threshold.

5.1.1 Model Selection

In the exercise we partition the data into three datasets: a training set, a validation set, and a testing set. We fit the models using the training set. The validation set allows us to select versions of each models that optimises our objective. In particular, we use three model-selection strategies: (1) we select models by minimizing the hiring cost function presented before, (2) we select models by maximizing using an F_β statistic, (3) we select models with the highest true positive rates. We report the cost implied by each model using the testing set. We repeat the exercise for different market tightness values, λ and for different values of the benefit of filling a vacancy to the interview cost, b/t. We present the results in Tables 5-8. The F_β statistic is an object commonly used in the machine learning literature. It strikes a balance between (1) the number of true positives, and (2) the number of false positives and false negatives since one can't increase (1) without also increasing (2). We provide further details about this statistic in Appendix 8.2.

Tables 5-8 indicate that the best models are not always the same for all scenarios. When outside labor market condition (λ) or the cost structure of firms (b/t) change, the optimal machine learning model often changes as well. The intuition is that different models put different weights on true positives and true negatives. As a result, when truly qualified applicants become rarer or their value/cost to the firm changes, different models will do a better job at reducing the hiring cost for the firms.

The selection criteria chosen also matters. Postulating the cost function and selecting models on that metric will, of course, minimize the hiring cost. A firm may instead choose a model that yields, say, the highest number of true positives or maximizes an F statistic. However, we show that doing so does not always lead to selecting the cost-minimizing model. The reason is that these metrics focus only on some aspects of the total hiring cost.

For example, when focusing on the number of true positives, the firm has a larger pool of qualified applicants to draw from. However, to find all these qualified folks, the firm needs to live with higher false positives and/or false negatives rates that increase mechanically. Indeed, to increase the number of true positives, firm will need to interview more people which will necessarily include less qualified applicants or interview applicants with very high probability of being qualified, thereby forgoing to interview applicants with a lower probability of being qualified that may, nevertheless, be qualified.

As false positives and false negatives pose different cost to firms, they will need to adjust the classification thresholds in order to minimize total cost. For example, when qualified workers are difficult to find or very valuable (when b/t is large or λ is small), firms may benefit from interviewing more applicants some of whom may be unqualified to ensure that they will not miss out on any qualified applicants.

Our model selection based on the total cost function leads the firms to select the profit-maximizing model (if correctly specified) since it accounts for the labor market condition as well as the cost and benefit of hiring by individual firms.

5.1.2 Model Evaluation

Next, we study the performance of the various machine learning models under different labor market conditions and different cost and benefit of filling a vacancy. We present the results in Figures 3 through 7. On the vertical axis, we plot the total hiring cost for every model measured in standard deviation units from the average model cost to easily

compare model performance across parameter values. On the horizontal axis we plot the labor market tightness and evaluate the models at $\lambda = 0.2$, $\lambda = 0.5$, and $\lambda = 0.75$. Figure 3 panel (a), (b), and (c) show the results for the objective task evaluated at $b/t = 1.5$, $b/t = 2$, and $b/t = 4$, respectively. The following figures repeat the exercise for the other tasks.

The first main takeaway is that no single model always dominate the others whether it be within task or across tasks. This goes back to the point we made earlier that models put different weights on TP, FP, and FN. Therefore, as quality workers become rarer or the benefit of hiring to the cost of interview ratio changes, a firm may benefit from choosing a different model that will weigh these elements in a way that reflect the changes.

Second, the differences in cost are large. The difference between the lowest-cost model and the second-lowest cost model often exceeds one standard deviation. This means that the algorithm choice for a large employer can be critical.

Third, some models tend to do better than others. For example Random Forest and Logistic Regressions tend to yield among the lowest cost, while AdaBoosting tends to yield a higher cost.

5.2 Parameter Analysis

In Table 13 we present the raw total hiring cost for different values of the parameters. As we see from the table, the total cost for each task under the same b/t increase as λ increases. The larger λ is, the tighter is the market. This parameter enters our cost function as the multiplier for the cost of false negatives. Also, the increase of λ represents higher classification threshold given by equation (19) (clarify this). Intuitively, when λ increases, the number of qualified workers in the potential pool decrease which means it is difficult for firms to find another qualified applicant if they wrongly dismiss one in the screening phase. In this case, it is costly to make wrong predictions and it is more difficult to make right prediction as λ rises. As a result, the total cost increases. Moreover, the number of

true positive predictions falls as λ increases. Hence, algorithms that put a high weight on true positive rates become less desirable as λ increases. With higher values of λ, models that jointly minimize false positive and false negative rates tend to perform better.

In Figure 8 we plot the total hiring costs as the values of both λ and b/t change. When the value of λ increases, then the cost tend to increase as we mentioned above. Also, when the value of b/t decreases, then the benefit/interview cost is low, so the overall cost tends to decrease. The lowest costs are achieved when the arrival rate is high and the benefit/interview cost ratio is low. This situation can arise when unemployment is relatively high so that many quality applicants are looking for jobs and employers have streamlined their interview processes lowering the interview cost for example.

6 Fair Machine Learning

The use of risk assessments by machine learning in sensitive areas such as criminal justice, hiring procedures or loan applications has been subject to intense scrutiny in the recent years (Berk et al., 2021). Such machine learning algorithms can be discriminatory towards protected groups, when variables for race and gender are included for example, as the data used to train this models could encode past discriminatory decisions. For example, if women were historically and systematically offered risky loans and men safer loans, then women would appear to default more often. As a result, predicting default risk from historical data using gender as a variable in the model can mechanically predict that women default more often. However, if past default was rooted in unequal treatments, the algorithm would repeat and could exacerbate the issue.

One of the proposed methods for alleviating discrimination and getting fairer machine learning models is anti-classification (Corbett-Davies and Goel, 2018). Anti-classification is when machine learning models do not consider protected attributes (such as gen-

der, race, or its proxies) in order to make a decision. In this work we implement anti-classification by changing the value of all protected features in the testing set to the mean of that protected feature. We can see in Tables 9, 10, 11, 12 the performance of the model selection technique with this anti-classification procedure and several values of λ and b/t. This important consideration will often lead to a different choice of model.

7 Conclusion

Firms rely more and more on personality traits to screen workers. We provide a framework that firms can use to streamline this process by way of carefully implemented machine learning algorithms. We show that minimizing the hiring cost is not straightforward but that choosing the right machine learning model can greatly reduce cost. To do so requires taking into account not only the idiosyncratic interview cost and benefit of filling vacancies by firms, but also the market tightness that governs the abundance of qualified workers in the economy.

References

Bahnsen, A. C., A. Stojanovic, D. Aouada, and B. Ottersten (2013). Cost sensitive credit card fraud detection using bayes minimum risk. In *2013 12th International Conference on Machine Learning and Applications*, Volume 1, pp. 333–338.

Barrick, M. R. and M. K. Mount (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology 44*(1), 1–26.

Berk, R., H. Heidari, S. Jabbari, M. Kearns, and A. Roth (2021). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research 50*(1), 3–44.

Boissiere, M., J. B. Knight, and R. H. Sabot (1985). Earnings, schooling, ability, and cognitive skills. *The American Economic Review 75*(5), 1016–1030.

Borghans, L., A. L. Duckworth, J. J. Heckman, and B. ter Weel (2008, February). The economics and psychology of personality traits. Working Paper 13810, National Bureau of Economic Research.

Bowles, S., H. Gintis, and M. Osborne (2001). The determinants of earnings: A behavioral approach. *Journal of economic literature 39*(4), 1137–1176.

Buda, M., A. Maki, and M. Mazurowski (2017, 10). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks 106*, 1–23.

Corbett-Davies, S. and S. Goel (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. *ArXiv abs/1808.00023*, 1–25.

Dollinger, S. J. and L. A. Orf (1991). Personality and performance in "personality": Conscientiousness and openness. *Journal of Research in Personality 25*(3), 276–284.

Duckworth, A. L. and L. Steinberg (2015). Unpacking self-control. *Child Development Perspectives 9*(1), 32–37.

Elkan, C. (2001). The foundations of cost-sensitive learning. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'01, San Francisco, CA, USA, pp. 973–978. Morgan Kaufmann Publishers Inc.

He, H. and E. A. Garcia (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering 21*(9), 1263–1284.

Heckman, J. J. and T. D. Kautz (2012, June). Hard evidence on soft skills. Working Paper 18121, National Bureau of Economic Research.

Hogan, J. and B. Holland (2003, February). Using theory to evaluate personality and job-performance relations: a socioanalytic perspective. *The Journal of applied psychology 88*(1), 100—112.

Liu, M., L. Miao, and D. Zhang (2014). Two-stage cost-sensitive learning for software defect prediction. *IEEE Transactions on Reliability 63*(2), 676–686.

Liu, Y., H. H. Zhang, and Y. Wu (2011). Hard or soft classification? large-margin unified machines. *Journal of the American Statistical Association 106*(493), 166–177. PMID: 22162896.

Maes, S., K. Tuyls, B. Vanschoenwinkel, and B. Manderick (2002, 08). Credit card fraud detection using bayesian and neural networks.

Moscarini, G. (2005). Job matching and the wage distribution. *Econometrica 73*(2), 481–516.

Rijsbergen, C. J. V. (1979). Retrieval effectiveness.

Rogerson, R., R. Shimer, and R. Wright (2005, December). Search-theoretic models of the labor market: A survey. *Journal of Economic Literature 43*(4), 959–988.

Rosenberg, M. (1965). *Society and the Adolescent Self-Image*. Princeton University Press.

Sahin, Y., S. Bulkan, and E. Duman (2013). A cost-sensitive decision tree approach for fraud detection. *Expert Systems with Applications 40*(15), 5916–5923.

Salgado, J. F. and A. Rumbo (1997). Personality and job performance in financial services managers. *International Journal of Selection and Assessment 5*(2), 91–100.

Shimer, R. (2007). Mismatch. *American Economic Review 97*(4), 1074–1101.

Stolfo, S. J., Wei Fan, Wenke Lee, A. Prodromidis, and P. K. Chan (2000). Cost-based modeling for fraud and intrusion detection: results from the jam project. In *Proceedings DARPA Information Survivability Conference and Exposition. DISCEX'00*, Volume 2, pp. 130–144.

Wang, J., X. Shen, and Y. Liu (2007, 11). Probability estimation for large-margin classifiers. *Biometrika 95*(1), 149–167.

Xia, Y., C. Liu, and N. Liu (2017). Cost-sensitive boosted tree for loan evaluation in peer-to-peer lending. *Electronic Commerce Research and Applications 24*, 30–49.

Zadrozny, B., J. Langford, and N. Abe (2003). Cost-sensitive learning by cost-proportionate example weighting. In *Third IEEE International Conference on Data Mining*, pp. 435–442.

Table 1: Cost Table related to the hiring problem.

	Predictions of the Algorithm	
	Qualified	**Unqualified**
Truly Qualified Net cost	True Positive (TP) $t - b$	False Negative (FN) λb
Truly Unqualified Net cost	False Positive (FP) t	False Negative (TN) 0

Note: Table 1 assigns costs to all the possible classifications a given model may assign to a given individual. If a truly qualified individual is classified as qualified, then the cost of the interview is $t - b$. If a truly qualified individual is classified as non-qualified then the cost of not interviewing is λb. If a truly unqualified individual is classified as qualified then interviewing them has a cost t. If a truly unqualified individual is classified as unqualified then the cost of not interviewing him is 0.

Table 2: Summary of Statistics

	Mean	SD
Personality		
Self esteem	14.478	34.514
Locus of controls	-7.793	7.471
Motivation	19.880	18.652
Grit	11.312	27.210
Resilience	5.547	20.832
Conscientiousness	15.730	27.032
Openness	51.849	24.338
Extraversion	-2.540	31.961
Agreeableness	29.506	26.687
Emotional stability	-28.220	33.817
Auto	48.786	24.772
Reading the mind in the eyes		
Reading score	25.281	6.552
Success of reading	0.723	0.187
Raven test		
Raven score	46.553	10.016
Success of raven test	0.428	0.116
Risk and time preferences		
Risk averse index	3.455	1.478
Discount index	1.435	0.771
Observables		
Age	35.395	10.323
Education	14.024	10.416
Experience	7.356	6.974
Experience of customer service	4.324	1.300
English	4.846	0.412
Task		
Objective task	49.340	7.314
Judgment task	5.202	1.564
Review	25.195	9.489
Review obj criteria	4.818	1.949
Review general score	14.329	5.444
N	253	

Note: Table 2 presents the means and standard deviations of the psychometric variables measured to distinguish between qualified and unqualified workers. There are 6 groups of variables: Personality, which relates to the personality traits of a given individual; Reading the mind in the eyes, which is a test that measures social intelligence; Raven Test, a test used to measure abstract reasoning capabilities; Risk and time preferences; Observables, which relate to some individual characteristics relating to age, knowledge and experience; and Task which measures the ability of a worker to perform a given task.

Table 3: Score Distribution for the different tasks.

Objective		Judgment		Review		Review Objective Criteria		Review General Score	
Score	Cumulative	Score	Cumulative	Score	Cumulative	Score	Cumulative	Score	Cumulative
45	22.92	3	13.83	23.33	24.51	1	10.67	13.33	24.90
46	26.88	4	34.78	23.67	25.69	2	14.62	13.50	25.30
51	47.83			28.00	47.83	3	19.76	16.83	48.62
52	56.52	5	56.52	28.33	50.20	4	26.48	17.33	50.20
53	64.43	6	77.47	31.00	73.52	5	37.15	17.58	73.52
54	75.49	7	93.28	31.33	76.68	6	100.00	17.67	75.89

Note: Table 3 presents the cumulative percentage for the score distribution of the different performed tasks: Objective task, Judgment task, Review task, Review Objective Criteria, and Review General task.

Table 4: Parameter Setting $\lambda = 0.25, b/t = 1.5$. Objective Task

Metric	TP		FP		FN		TC	
Model	Y	N	Y	N	Y	N	Y	N
Random Forest	28	28	3	3	0	0	-11	-11
Logistic Regression	18	18	3	4	3	3	-4.875	-3.875
Neural Network	27	27	0	0	2	2	-12.75	-12.75
k-Nearest Neighbors	25	25	2	2	0	0	-10.5	-10.5
Support Vector Machine	24	24	3	3	1	1	-8.625	-8.625
Stochastic Gradient Descent	23	23	4	4	4	4	-6	-6
Gradient Boosting	25	25	3	3	1	1	-9.125	-9.125
Ada Boosting	3	21	1	4	21	3	7.375	-5.375

Note: Table 4 presents the performances in True Positives (qualified workers classified as qualified), True Negatives (unqualified workers classified as unqualified), False Negatives (qualified workers predicted as unqualified) and the Total Cost (Equation 2.1) for the different models with different classification thresholds, with the models with superscript N relate to the threshold in Equation 8, and the models with superscript Y relate to the threshold in Equation 9. The workers are classified according to their objective task score.

Table 5: Parameter Setting $\lambda = 0.75, b/t = 2$

	Performance				Model Selection
	TP	TP rate	F	TC	Best Model
Panel A : **Objective Task**					
$\mathbf{h}^{Profit}$	11	0.92	0.671	2.5	Neural Network
$\mathbf{h}^{F}$	11	0.92	0.71	2.5	Neural Network
$\mathbf{h}^{TP}$	12	1.00	0.60	8	ADA Boosting
Panel B : **Judgment Task**					
$\mathbf{h}^{Profit}$	14	1.00	0.729	-1	Random Forest
$\mathbf{h}^{F}$	14	1.00	0.729	4	Random Forest
$\mathbf{h}^{TP}$	14	1.00	0.66	4 5	Neural Network
Panel C: ***Review Task***					
$\mathbf{h}^{Profit}$	5	0.714	0.531	3	Random Forest
$\mathbf{h}^{F}$	5	0.714	0.531	3	Random Forest
$\mathbf{h}^{TP}$	7	1.00	0.411	14	ADA Boosting
Panel D: ***Review General Score***					
$\mathbf{h}^{Profit}$	14	0.875	0.729	-7	k-Nearest Neighbors
$\mathbf{h}^{F}$	14	0.875	0.729	-7	k-Nearest Neighbors
$\mathbf{h}^{TP}$	16	100	0.74	-4	Neural Network
Panel E : **Review Objective Criteria**					
$\mathbf{h}^{Profit}$	1	0.167	0.227	7.5	Gradient Boost Classifier
$\mathbf{h}^{F}$	1	0.5	0.3	14.5	Random Forrest
$\mathbf{h}^{TP}$	3	0.5	0.3	14.5	Random Forrest

Note: Table 5 presents the performances in True Positives (qualified workers classified as qualified), True Positive rate (percentage of true positives over total qualified workers, representing the percentage of correctly classified qualified workers), F which is the F^*_β measure, TC which is the total cost as in Equation 2.1, h^{Profit} which represents the name of the classifier that gives the best prediction according to lower values of Equation 2.1, h^F which reports the name of the classifier with highest F_β measure, and h^{TP} which reports the name of the classifier with highest TP predicted.

Table 6: Parameter Setting $\lambda = 0.25, b/t = 4$

	Performance				Model Selection
	TP	TP rate	F	TC	Best Model
Panel A : **Objective Task**					
$\mathbf{h}^{Profit}$	24	1.00	0.945	-75	Gradient Boost Classifier
$\mathbf{h}^{F}$	24	1.00	0.945	-75	Gradient Boost Classifier
$\mathbf{h}^{TP}$	24	1.00	0.912	-73	Random Forest
Panel B : **Judgment Task**					
$\mathbf{h}^{Profit}$	24	1.00	0.896	-72	Random Forest
$\mathbf{h}^{F}$	24	1.00	0.896	-72	Random Forest
$\mathbf{h}^{TP}$	24	1.00	0.896	-72	Random Forest
Panel C: ***Review Task***					
$\mathbf{h}^{Profit}$	24	1.00	0.905	-67	Random Forest
$\mathbf{h}^{F}$	24	1.00	0.905	-67	Random Forest
$\mathbf{h}^{TP}$	24	1.00	0.905	-67	Random Forest
Panel D: ***Review General Score***					
$\mathbf{h}^{Profit}$	24	1.00	0.857	-64	Support Vector Machine
$\mathbf{h}^{F}$	24	1.00	0.857	-64	Support Vector Machine
$\mathbf{h}^{TP}$	24	1.00	0.74	-65	Random Forest
Panel E : **Review Objective Criteria**					
$\mathbf{h}^{Profit}$	24	1.00	0.872	-65	Logistic Regression
$\mathbf{h}^{F}$	24	1.00	0.872	-65	Logistic Regression
$\mathbf{h}^{TP}$	24	1.00	0.827	-65	Logistic Regression

Note: Table 6 presents the performances in True Positives (qualified workers classified as qualified), True Positive rate (percentage of true positives over total qualified workers, representing the percentage of correctly classified qualified workers), F which is the F^*_β measure, TC which is the total cost as in Equation 2.1, h^{Profit} which represents the name of the classifier that gives the best prediction according to lower values of Equation 2.1, h^F which reports the name of the classifier with highest F_β measure, and h^{TP} which reports the name of the classifier with highest TP predicted.

Table 7: Parameter Setting $\lambda = 0.75, b/t = 1.5$

	Performance				Model Selection
	TP	TP rate	F	TC	Best Model
Panel A : **Objective Task**					
$\mathbf{h}^{Profit}$	8	0.533	0.55	9.875	Logistic Regression
$\mathbf{h}^{F}$	8	0.533	0.55	9.875	Logistic Regression
$\mathbf{h}^{TP}$	16	1.00	0.652	9.5	ADA Boosting
Panel B : **Judgment Task**					
$\mathbf{h}^{Profit}$	14	0.875	0.743	3.25	Random Forest
$\mathbf{h}^{F}$	14	0.875	0.743	3.25	Random Forest
$\mathbf{h}^{TP}$	16	1.00	0.694	7.00	Support Vector Machine
Panel C: ***Review Task***					
$\mathbf{h}^{Profit}$	2	0.22	0.248	11.875	k-Nearest Neighbors
$\mathbf{h}^{F}$	2	0.22	0.248	11.875	k-Nearest Neighbors
$\mathbf{h}^{TP}$	9	1.00	0.488	15.5	ADA Boosting
Panel D: ***Review General Score***					
$\mathbf{h}^{Profit}$	23	1.00	0.891	-5.5	Random Forest
$\mathbf{h}^{F}$	23	1.00	0.891	-5.5	Random Forest
$\mathbf{h}^{TP}$	23	1.00	0.891	-5.5	Random Forest
Panel E : **Review Objective Criteria**					
$\mathbf{h}^{Profit}$	1	1.00	0.25	4.875	Random Forest
$\mathbf{h}^{F}$	1	1.00	1.00	4.875	Random Forest
$\mathbf{h}^{TP}$	4	1.00	1.00	26	Ada Boosting

Note: Table 7 presents the performances in True Positives (qualified workers classified as qualified), True Positive rate (percentage of true positives over total qualified workers, representing the percentage of correctly classified qualified workers), F which is the F^*_β measure, TC which is the total cost as in Equation 2.1, h^{Profit} which represents the name of the classifier that gives the best prediction according to lower values of Equation 2.1, h^F which reports the name of the classifier with highest F_β measure, and h^{TP} which reports the name of the classifier with highest TP predicted.

Table 8: Parameter Setting $\lambda = 0.25, b/t = 1.5$

	Performance				Model Selection
	TP	TP rate(%)	F	TC	Best Model
Panel A : ***Objective Task***					
$\mathbf{h}^{Profit}$	28	1.00	0.95	-12	Random Forest
$\mathbf{h}^{F}$	28	1.00	0.95	-12	Random Forest
$\mathbf{h}^{TP}$	28	1.00	0.95	-12	Random Forest
Panel B : ***Judgment Task***					
$\mathbf{h}^{Profit}$	28	1.00	0.905	-10	Random Forest
$\mathbf{h}^{F}$	28	1.00	0.905	-10	Random Forest
$\mathbf{h}^{TP}$	28	1.00	0.905	-10	Random Forest
Panel C: ***Review Task***					
$\mathbf{h}^{Profit}$	22	0.956	0.873	-6.625	Random Forest
$\mathbf{h}^{F}$	21	0.913	0.873	-4.75	Stochastic Gradient Descent
$\mathbf{h}^{TP}$	22	0.956	0.873	-6.625	Random Forest
Panel D: ***Review General Score***					
$\mathbf{h}^{Profit}$	22	1.00	0.796	-3	Random Forest
$\mathbf{h}^{F}$	22	1.00	0.796	-3	Random Forest
$\mathbf{h}^{TP}$	21	0.954	0.796	-3.124	Logistic Regression
Panel E : **Review Objective Criteria**					
$\mathbf{h}^{Profit}$	19	0.905	0.846	4.75	k-Nearest Neighbors
$\mathbf{h}^{F}$	19	0.905	0.846	4.75	k-Nearest Neighbors
$\mathbf{h}^{TP}$	18	0.857	0.857	-4.875	Stochastic Gradient Descent

Note: Table 8 presents the performances in True Positives (qualified workers classified as qualified), True Positive rate (percentage of true positives over total qualified workers, representing the percentage of correctly classified qualified workers), F which is the F^*_β measure, TC which is the total cost as in Equation 2.1, h^{Profit} which represents the name of the classifier that gives the best prediction according to lower values of Equation 2.1, h^F which reports the name of the classifier with highest F_β measure, and h^{TP} which reports the name of the classifier with highest TP predicted.

Table 9: Parameter Setting $\lambda = 0.75, b/t = 2$

	Performance				Model Selection
	TP	TP rate	F	TC	Best Model
Panel A : **Objective Task**					
$\mathbf{h}^{Profit}$	11	0.92	0.73	-1	Logistic Regression
$\mathbf{h}^{F}$	11	0.92	0.73	-1	Logistic Regression
$\mathbf{h}^{TP}$	11	0.92	0.73	-1	Logistic Regression
Panel B : **Judgment Task**					
$\mathbf{h}^{Profit}$	9	0.82	0.57	8	k-Nearest Neighbors
$\mathbf{h}^{F}$	9	0.82	0.57	8	k-Nearest Neighbors
$\mathbf{h}^{TP}$	11	1.00	0.57	10	Neural Network
Panel C: ***Review Task***					
$\mathbf{h}^{Profit}$	8	1.00	0.45	16	Stochastic Gradient Descent
$\mathbf{h}^{F}$	8	1.00	0.45	16	ADA Boosting
$\mathbf{h}^{TP}$	8	1.00	0.45	16	ADA Boosting
Panel D: ***Review General Score***					
$\mathbf{h}^{Profit}$	4	0.57	0.49	6.5	k-Nearest Neighbors
$\mathbf{h}^{F}$	7	1.00	0.41	18	ADA Boosting
$\mathbf{h}^{TP}$	7	1.00	0.41	18	ADA Boosting
Panel E : **Review Objective Criteria**					
$\mathbf{h}^{Profit}$	18	0.90	0.80	-7	Gradient Boost Classifier
$\mathbf{h}^{F}$	18	0.90	0.80	-7	Gradient Boost Classifier
$\mathbf{h}^{TP}$	20	1.00	0.83	-10	Neural Network

Note: Table 9 presents the performances in True Positives (qualified workers classified as qualified), True Positive rate (percentage of true positives over total qualified workers, representing the percentage of correctly classified qualified workers), F which is the F^*_β measure, TC which is the total cost as in Equation 2.1, h^{Profit} which represents the name of the classifier that gives the best prediction according to lower values of Equation 2.1, h^F which reports the name of the classifier with highest F_β measure, and h^{TP} which reports the name of the classifier with highest TP predicted.

Table 10: Parameter Setting $\lambda = 0.25, b/t = 4$

	Performance				Model Selection
	TP	TP rate	F	TC	Best Model
Panel A : **Objective Task**					
$\mathbf{h}^{Profit}$	26	1.00	0.94	-75	Gradient Boost Classifier
$\mathbf{h}^{F}$	26	1.00	0.94	-75	Gradient Boost Classifier
$\mathbf{h}^{TP}$	26	1.00	0.89	-72	Random Forest
Panel B : **Judgment Task**					
$\mathbf{h}^{Profit}$	29	1.00	0.95	-84	Random Forest
$\mathbf{h}^{F}$	29	1.00	0.95	-84	Random Forest
$\mathbf{h}^{TP}$	29	1.00	0.95	-84	Random Forest
Panel C: ***Review Task***					
$\mathbf{h}^{Profit}$	26	1.00	0.89	-72	Random Forest
$\mathbf{h}^{F}$	26	1.00	0.89	-72	Random Forest
$\mathbf{h}^{TP}$	26	1.00	0.89	-72	Random Forest
Panel D: ***Review General Score***					
$\mathbf{h}^{Profit}$	25	1.00	0.89	-69	Logistic Regression
$\mathbf{h}^{F}$	25	1.00	0.89	-69	Logistic Regression
$\mathbf{h}^{TP}$	25	1.00	0.87	-68	Random Forest
Panel E : **Review Objective Criteria**					
$\mathbf{h}^{Profit}$	26	0.93	0.93	-74	Gradient Boost Classifier
$\mathbf{h}^{F}$	26	0.93	0.93	-74	Gradient Boost Classifier
$\mathbf{h}^{TP}$	28	1.00	0.93	-80	Random Forest

Note: Table 10 presents the performances in True Positives (qualified workers classified as qualified), True Positive rate (percentage of true positives over total qualified workers, representing the percentage of correctly classified qualified workers), F which is the F^*_β measure, TC which is the total cost as in Equation 2.1, h^{Profit} which represents the name of the classifier that gives the best prediction according to lower values of Equation 2.1, h^F which reports the name of the classifier with highest F_β measure, and h^{TP} which reports the name of the classifier with highest TP predicted.

Table 11: Parameter Setting $\lambda = 0.75, b/t = 1.5$

	Performance				Model Selection
	TP	TP rate	F	TC	Best Model
Panel A : **Objective Task**					
$\mathbf{h}^{Profit}$	3	0.42	0.34	11	Logistic Regression
$\mathbf{h}^{F}$	3	0.42	0.34	11	Logistic Regression
$\mathbf{h}^{TP}$	7	1.00	0.37	21.5	ADA Boosting
Panel B : **Judgment Task**					
$\mathbf{h}^{Profit}$	7	0.875	0.44	15.625	k-Nearest Neighbors
$\mathbf{h}^{F}$	7	0.875	0.44	15.625	k-Nearest Neighbors
$\mathbf{h}^{TP}$	8	1.00	0.41	20	ADA Boosting
Panel C: ***Review Task***					
$\mathbf{h}^{Profit}$	0.0	0.0	0.0	8.75	Support Vector Machine
$\mathbf{h}^{F}$	6	1.00	0.34	22	ADA Boosting
$\mathbf{h}^{TP}$	6	1.00	0.34	22	ADA Boosting
Panel D: ***Review General Score***					
$\mathbf{h}^{Profit}$	0.0	0.0	0.0	10.125	Neural Network
$\mathbf{h}^{F}$	9	1.00	0.45	18.5	ADA Boosting
$\mathbf{h}^{TP}$	9	1.00	0.45	18.5	ADA Boosting
Panel E : **Review Objective Criteria**					
$\mathbf{h}^{Profit}$	21	0.95	0.84	-2.375	Random Forest
$\mathbf{h}^{F}$	21	0.95	0.84	-2.375	Random Forest
$\mathbf{h}^{TP}$	22	1.00	0.82	-1	ADA Boosting

Note: Table 11 presents the performances in True Positives (qualified workers classified as qualified), True Positive rate (percentage of true positives over total qualified workers, representing the percentage of correctly classified qualified workers), F which is the F^*_β measure, TC which is the total cost as in Equation 2.1, h^{Profit} which represents the name of the classifier that gives the best prediction according to lower values of Equation 2.1, h^F which reports the name of the classifier with highest F_β measure, and h^{TP} which reports the name of the classifier with highest TP predicted.

Table 12: Parameter Setting $\lambda = 0.25, b/t = 1.5$

	Performance				Model Selection
	TP	TP rate(%)	F	TC	Best Model
Panel A : ***Objective Task***					
$\mathbf{h}^{Profit}$	26	1.00	0.92	-10	Random Forest
$\mathbf{h}^{F}$	26	1.00	0.92	-10	Random Forest
$\mathbf{h}^{TP}$	26	1.00	0.92	-10	Random Forest
Panel B : ***Judgment Task***					
$\mathbf{h}^{Profit}$	26	0.9	0.92	-9.875	Logistic Regression
$\mathbf{h}^{F}$	26	0.9	0.92	-9.875	Logistic Regression
$\mathbf{h}^{TP}$	26	0.9	0.92	-9.875	Logistic Regression
Panel C: ***Review Task***					
$\mathbf{h}^{Profit}$	25	1	0.89	-8.5	Random Forest
$\mathbf{h}^{F}$	25	1	0.89	-8.5	Random Forest
$\mathbf{h}^{TP}$	25	1	0.89	-8.5	Random Forest
Panel D: ***Review General Score***					
$\mathbf{h}^{Profit}$	22	0.95	0.8	-6.625	Random Forest
$\mathbf{h}^{F}$	22	0.95	0.8	-6.625	Random Forest
$\mathbf{h}^{TP}$	22	0.95	0.8	-6.625	Random Forest
Panel E : **Review Objective Criteria**					
$\mathbf{h}^{Profit}$	27	0.96	0.92	-10.125	Gradient Boosting
$\mathbf{h}^{F}$	27	0.96	0.92	-10.125	Gradient Boosting
$\mathbf{h}^{TP}$	27	0.96	0.92	-10.125	Gradient Boosting

Note: Table 12 presents the performances in True Positives (qualified workers classified as qualified), True Positive rate (percentage of true positives over total qualified workers, representing the percentage of correctly classified qualified workers), F which is the F^*_β measure, TC which is the total cost as in Equation 2.1, h^{Profit} which represents the name of the classifier that gives the best prediction according to lower values of Equation 2.1, h^F which reports the name of the classifier with highest F_β measure, and h^{TP} which reports the name of the classifier with highest TP predicted.

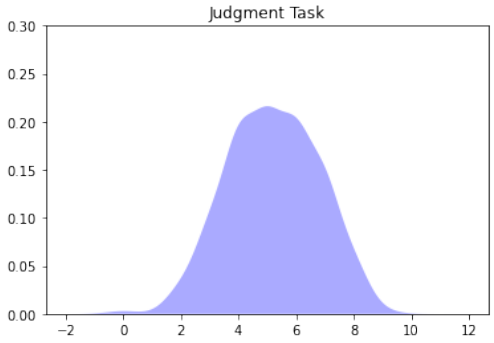
Table 13: Total cost for different values of λ and b/t.

	Total Hiring Cost		
λ / b/t	0.25	0.5	0.75
Objective			
1.5	-8.25	-0.75	13.5
2	-21.0	0.0	9.5
4	-74	-22	-14
Judgment			
1.5	-2.5	-2.5	6.375
2	-15.5	-7	-6
4	-72	-46	-20
Review Score			
1.5	-4.25	5.25	12.25
2	-16.5	-8	14
4	-55	-31	-12
Review obj Criteria			
1.5	-7.375	0	0.875
2	-23.5	-12	-5.5
4	-71	-44	-51
Review General Score			
1.5	-1.125	5.5	9.625
2	-17	-12	17.5
4	-65	-34	-16
N		253	

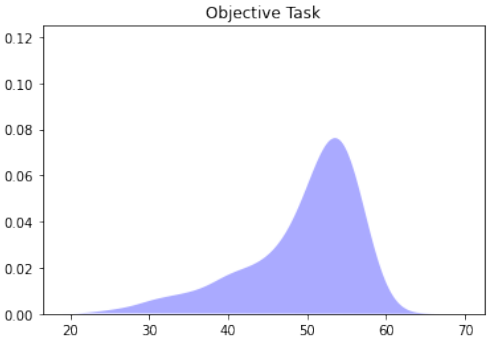
Note: Table 13 presents the performances in Total Cost (Equation 2.1) for the best model for each of the combination of λ and b/t and each of the classification tasks proposed in this work.

Figure 1: Densities for Outcomes

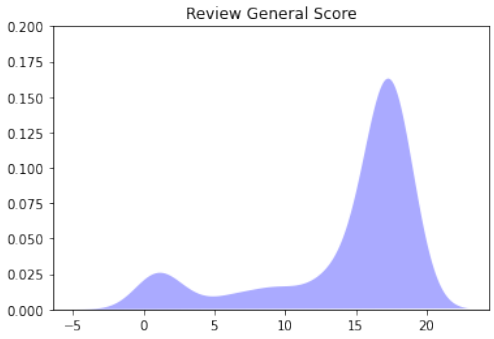
(a) Judgment Task



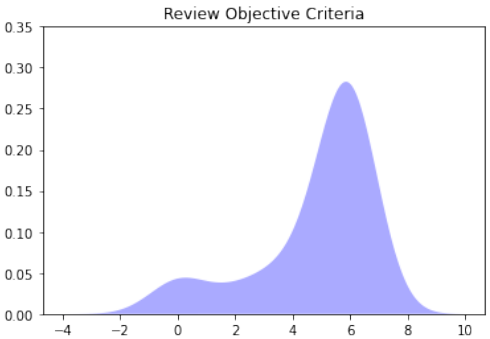
(b) Objective Task



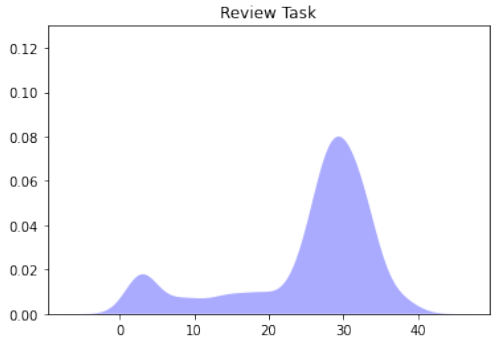
(c) Review General Score



(d) Review Objective Criteria



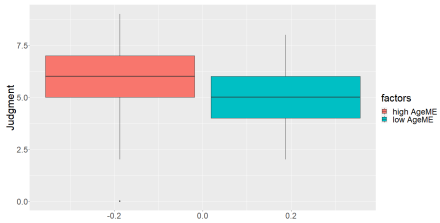
(e) Review Task



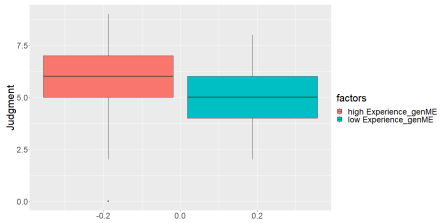
Note: Figure 1 shows the distributions of all the different outcomes included in our study. In the *Objective* tasks, participants were shown product pages from Amazon and were ask to report information about the product such as the price, the number of item left in stock, the rating of the product. In the *Judgment* tasks, they were presented with hypothetical customer requests about different products, like sleeping bags, and had to select the product(s) that satisfied the request of the customer from a set of products. For example, a client may request look for a sleeping bag of a particular size and temperature and participants had to find which of three candidate sleeping bags if any would fit the parameters set by the client. In the *Review* tasks, participants were shown 4 actual (bad) reviews that customers left on certain products on Amazon, and they then had to decide to which customer(s) they wanted to reply to and write hypothetical replies within a set number of words. In each task type, participants had to do at least 2 tasks that had multiple questions.

Figure 2: Boxplots

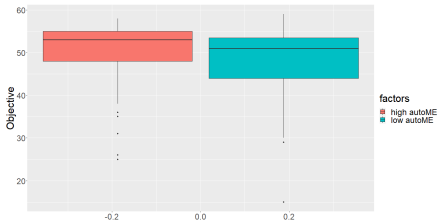
(a) Judgment task



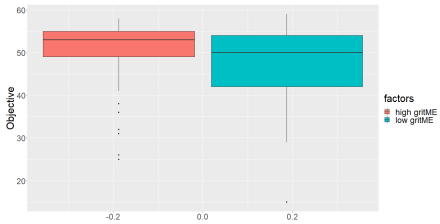
(b) Judgment task



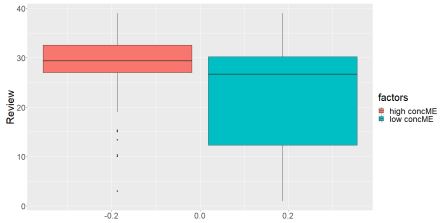
(c) Objective task



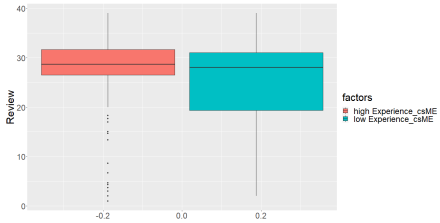
(d) Objective task



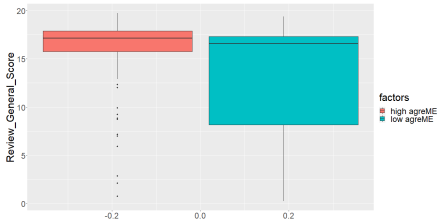
(e) Review task



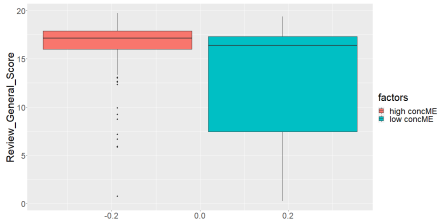
(f) Review task



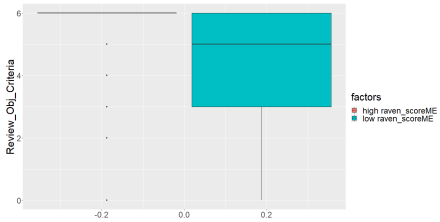
(g) Review General Score



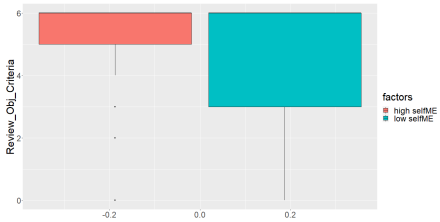
(h) Review General Score



(i) Review Objective Criteria

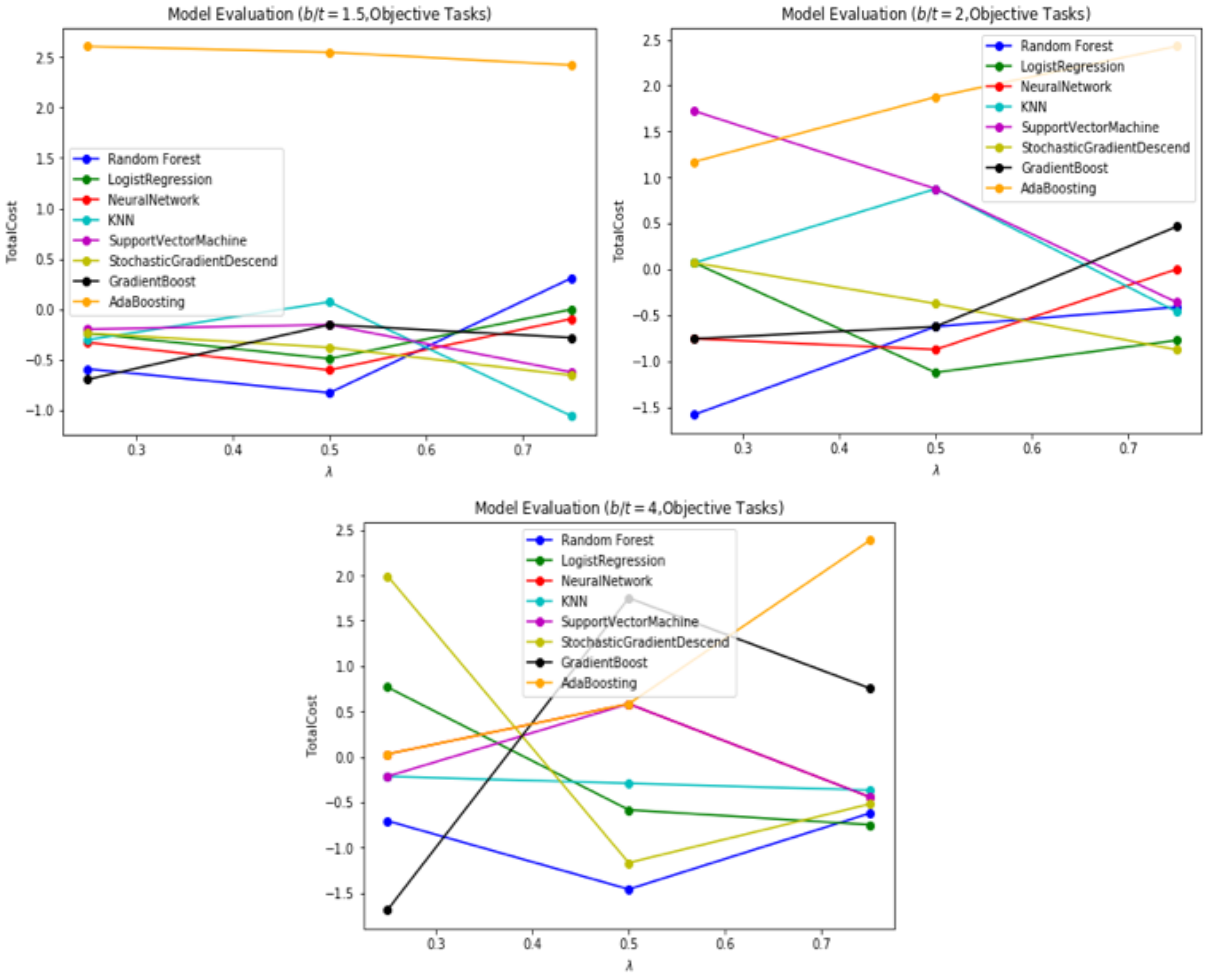


(j) Review Objective Criteria



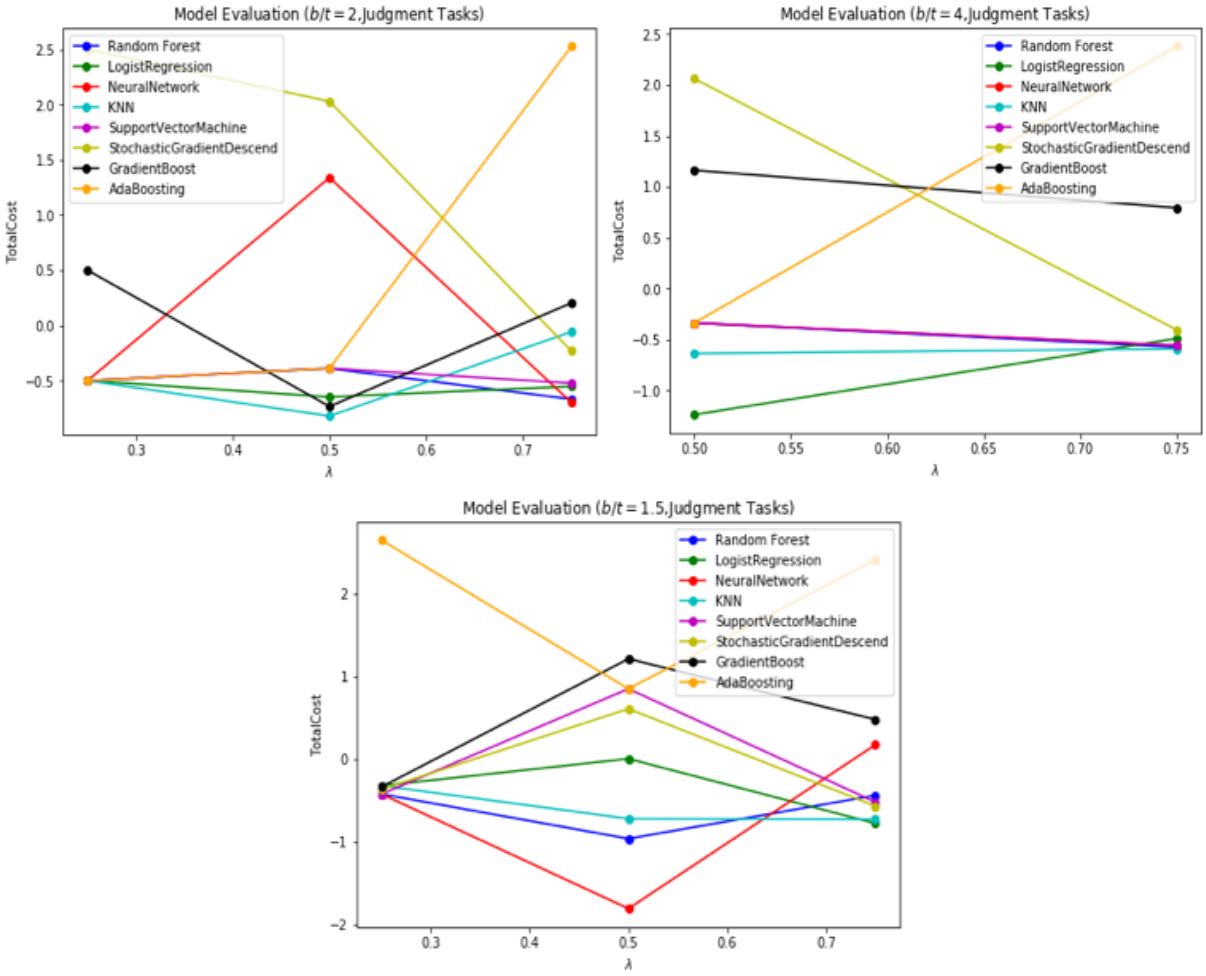
Note: Figure 2 we split the different outcomes at the median of each measure across workers and construct boxplots of each outcome for each sumsamble (above and below median)

Figure 3: Model Evaluation for Objective Task



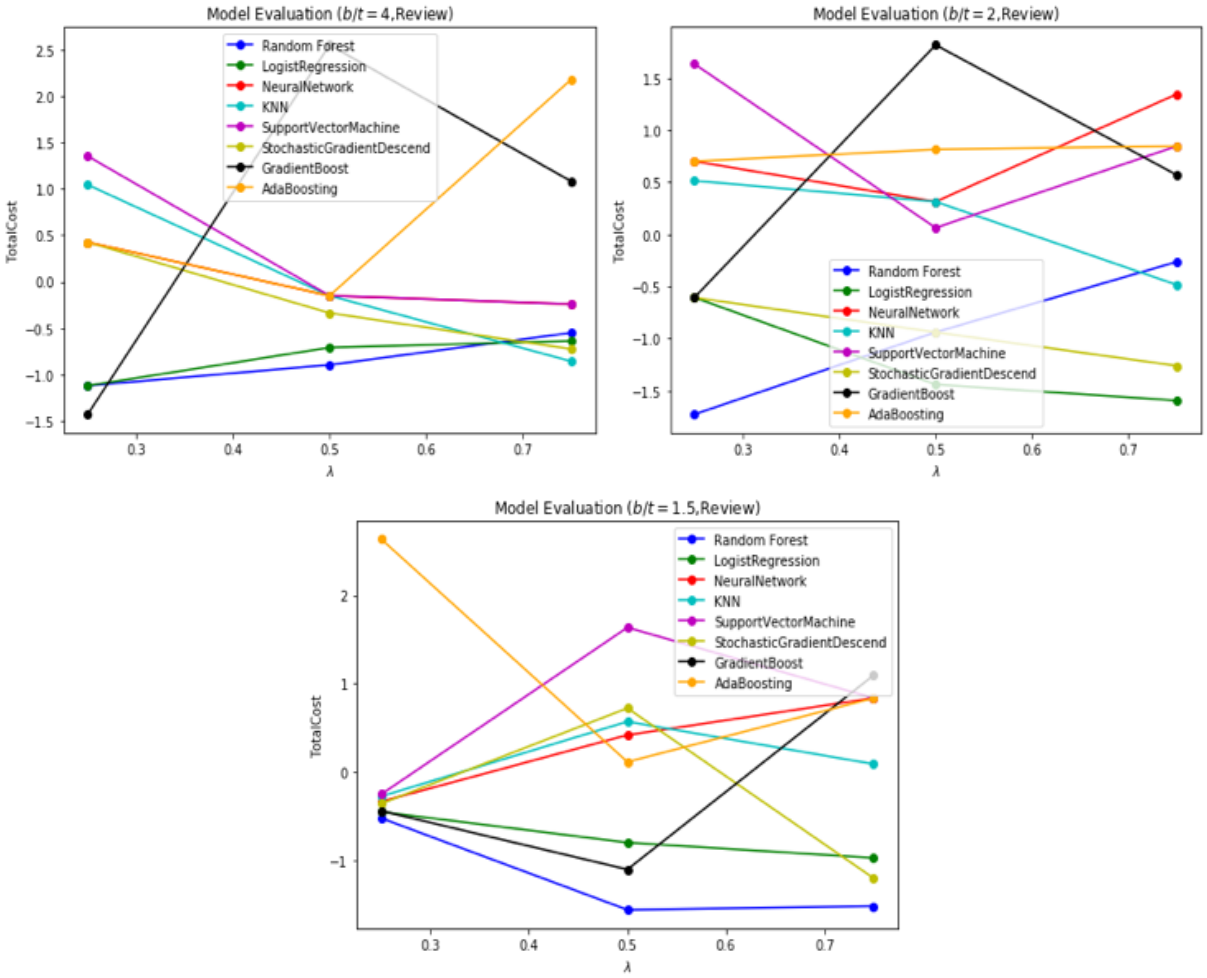
Note: Figure 3 presents the Total Cost in the test set for various different classification models, using different values of b/t and *lambda* for the Objective Task. The Blue line represents the cost for a random forest, the dark green line represents the cost for a logistic regression, the red line represents the cost for a neural network, the teal line represent the cost for a K-Nearest Neighbor classifier, the purple line represent the cost for a support vector machine, the light green line represents the cost for a stochastic gradient descent classifier, the black line represent the cost for a Gradient Boosting classifier and the orange line represents the cost for a AdaBoost classifier.

Figure 4: Model Evaluation for Judgment Task



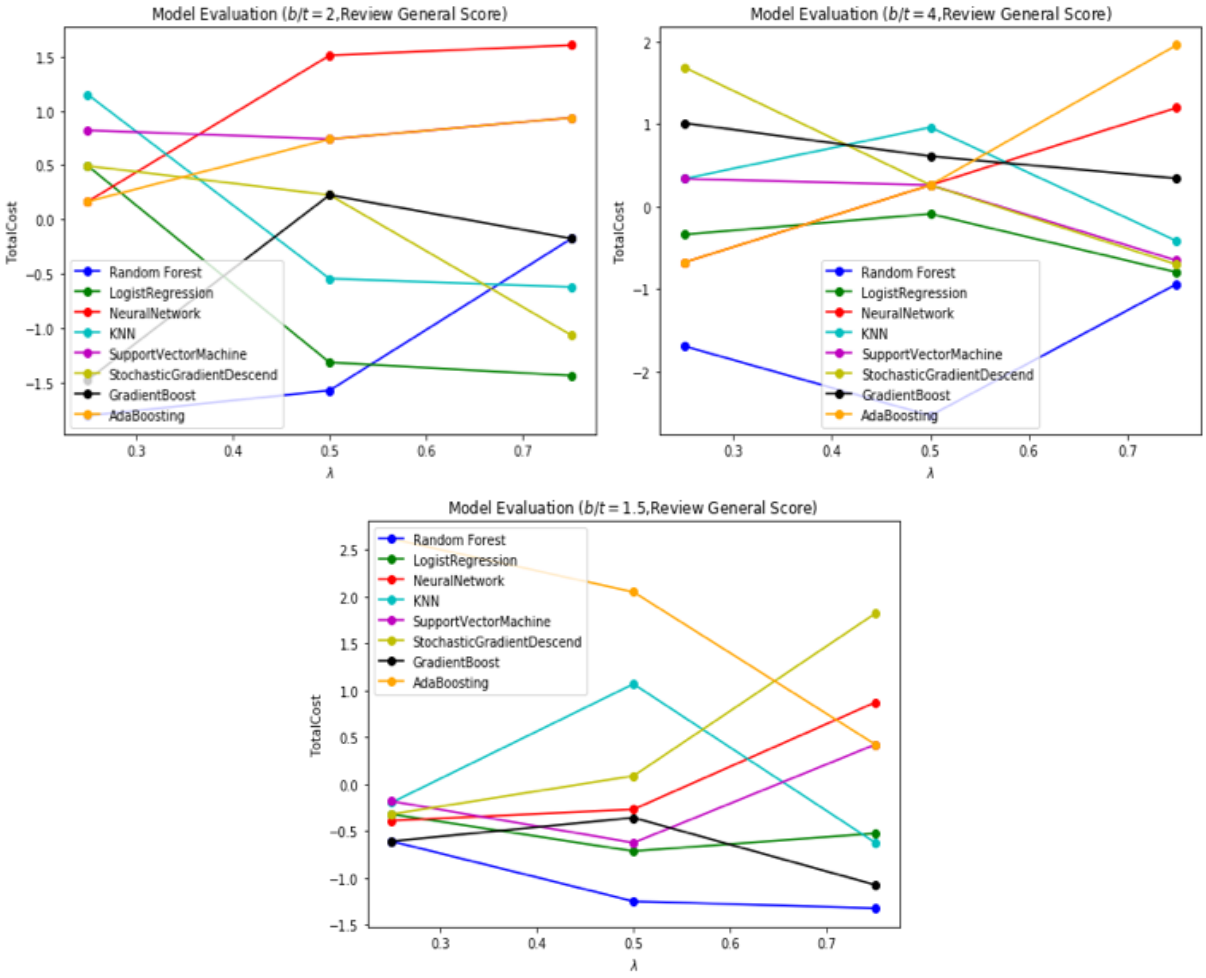
Note: Figure 4 presents the Total Cost in the test set for various different classification models, using different values of b/t and *lambda* for the Judgment Task. The Blue line represents the cost for a random forest, the dark green line represents the cost for a logistic regression, the red line represents the cost for a neural network, the teal line represent the cost for a K-Nearest Neighbor classifier, the purple line represent the cost for a support vector machine, the light green line represents the cost for a stochastic gradient descent classifier, the black line represent the cost for a Gradient Boosting classifier and the orange line represents the cost for a AdaBoost classifier.

Figure 5: Model Evaluation for Review Task



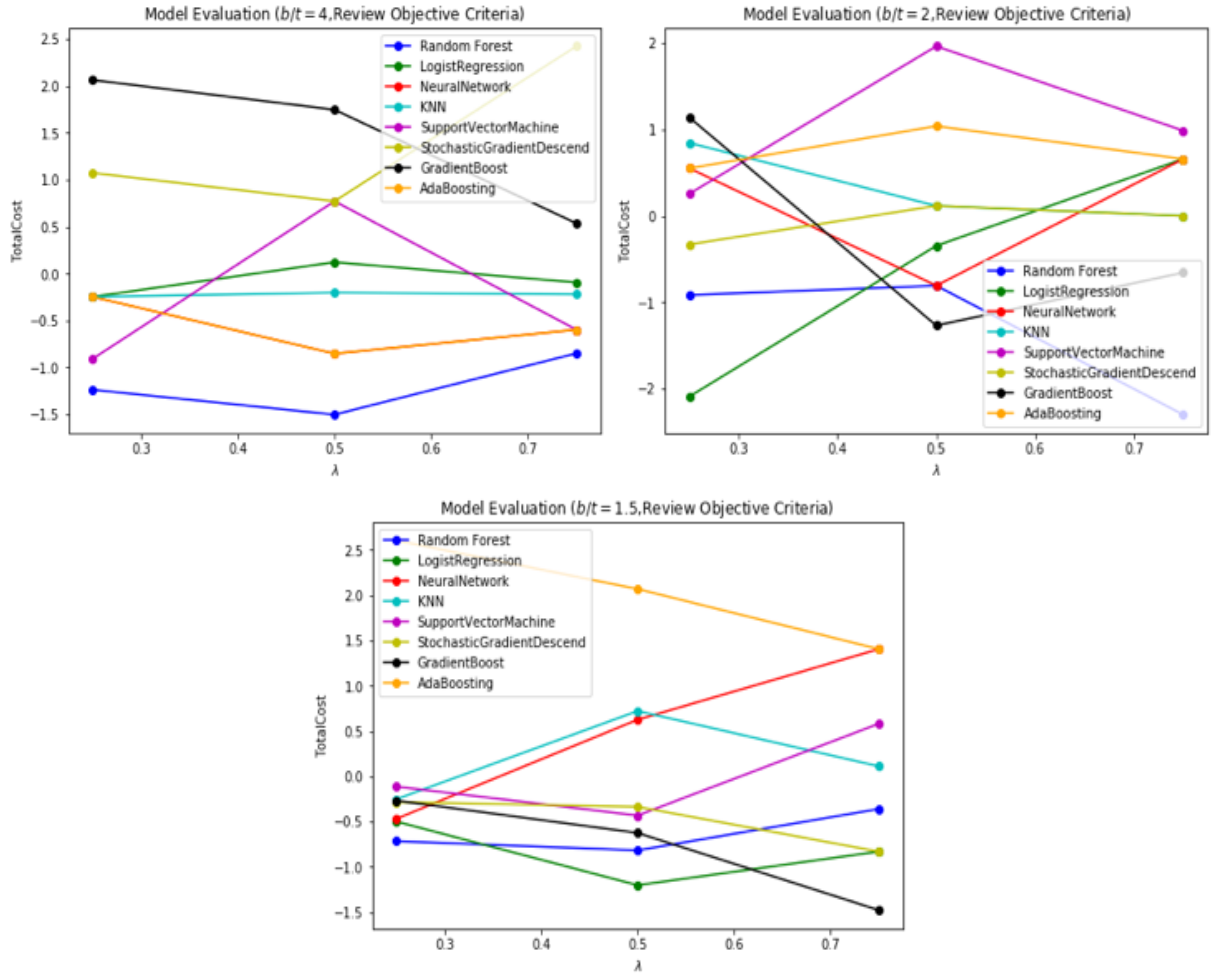
Note: Figure 5 presents the Total Cost in the test set for various different classification models, using different values of b/t and *lambda* for the Review Task. The Blue line represents the cost for a random forest, the dark green line represents the cost for a logistic regression, the red line represents the cost for a neural network, the teal line represent the cost for a K-Nearest Neighbor classifier, the purple line represent the cost for a support vector machine, the light green line represents the cost for a stochastic gradient descent classifier, the black line represent the cost for a Gradient Boosting classifier and the orange line represents the cost for a AdaBoost classifier.

Figure 6: Model Evaluation for Review General Score



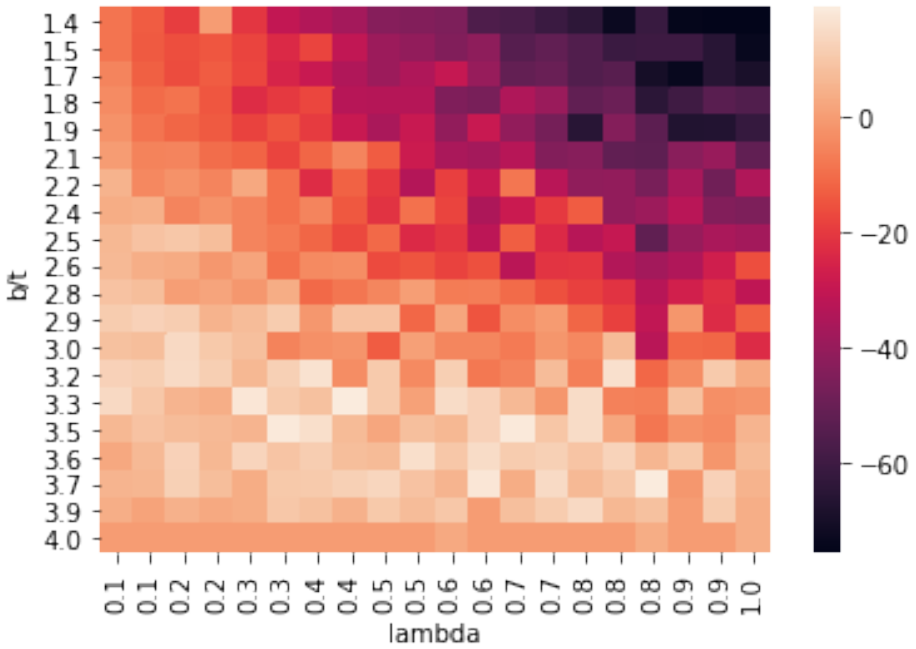
Note: Figure 6 presents the Total Cost in the test set for various different classification models, using different values of b/t and *lambda* for the Review General Score. The Blue line represents the cost for a random forest, the dark green line represents the cost for a logistic regression, the red line represents the cost for a neural network, the teal line represent the cost for a K-Nearest Neighbor classifier, the purple line represent the cost for a support vector machine, the light green line represents the cost for a stochastic gradient descent classifier, the black line represent the cost for a Gradient Boosting classifier and the orange line represents the cost for a AdaBoost classifier.

Figure 7: Model Evaluation for Review Objective Criteria



Note: Figure 7 presents the Total Cost in the test set for various different classification models, using different values of b/t and *lambda* for the Review Objective Criteria. The Blue line represents the cost for a random forest, the dark green line represents the cost for a logistic regression, the red line represents the cost for a neural network, the teal line represent the cost for a K-Nearest Neighbor classifier, the purple line represent the cost for a support vector machine, the light green line represents the cost for a stochastic gradient descent classifier, the black line represent the cost for a Gradient Boosting classifier and the orange line represents the cost for a AdaBoost classifier.

Figure 8: Total hiring costs for different values of λ and b/t.



Note: Figure 8 presents total cost on the test set for the best classifier of the ones considered in this work for many different configurations of λ and b/t.

8 APPENDIX

8.1 Search and Matching models

8.2 F_β

Definition: The relative importance a user attaches to precision and recall is the P/R ratio at which $\delta E/\delta R = \delta E/\delta R$, where $E = E(P, R)$ is the measure of effectiveness based on precision and recall.

Intuition: The way to quantifying importance is to specify the P/R ratio at which the user is willing to trade an increment in precision for an equal loss in recall.

Rijsbergen (1979) uses E (for 'effectiveness measure'), which is just $1 - F$ and the explanation is equivalent whether we consider E or F.

$$F = \frac{1}{(\frac{\alpha}{P} + \frac{1-\alpha}{R})} \tag{10}$$

$$\delta F/\delta P = \delta F/\delta R \tag{11}$$

$$\frac{\alpha}{P^2} = \frac{1-\alpha}{R^2} \rightarrow \frac{R}{P} = \sqrt{\frac{1-\alpha}{\alpha}} =: \beta \tag{12}$$

$$F_\beta = \frac{(1+\beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall} \tag{13}$$

We need to find the right weight, β, that satisfies

The harmonic mean is more intuitive than the arithmetic mean when computing a mean of ratios. Suppose that you have a finger print recognition system and its precision and recall be 1.0 and 0.2, respectively. Intuitively, the total performance of the system should be very low because the system covers only 20% of the registered finger prints, which means it is almost useless. The arithmetic mean of 1 and 0.2 is 0.6 whereas the harmonic mean of them is 0.333 is a more reasonable score than the arithmetic mean (0.6)

β is a parameter that controls a balance between P and R. When $\beta = 1$, F_1 comes to be equivalent to the harmonic mean of P and R. If $\beta > 1$, F becomes more recall-oriented and if $\beta < 1$, it becomes more precision oriented, e.g., $F_0 = P$. While it seems that van Rijsbergen did not define the formula of the F-measure per se, the origin of the definition of the F-measure is van Rijsbergen's E (effectiveness) function (Rijsbergen, 1979):

8.2.1 F_β index

The direct evaluations we can get from the output of the models are: TP (True Positive), FP (False Positive), FN (False Negative) and TN (True Negative). As defined in statistical learning, we define precision and recall as:

$$\begin{cases} Precision = \frac{TP}{TP+FP} \\ \\ Recall = \frac{TP}{TP+FN} \end{cases} \tag{14}$$

We want the model that has highest precision and recall. To do this, we can either increase TP points or reduce FN and FP. According to 'No Free Lunch Theorem', there isn't a model that could boost in both precision and recall because there is always a trade off. Let's combine these two evaluations in the weighted harmonic mean and pick the most commonly used F score as a one-way criteria.

$$F = \frac{1}{(\frac{\alpha}{P} + \frac{1-\alpha}{R})} \tag{15}$$

We can see that a higher score on TP points means both higher precision and recall and higher precision or recall will lead to higher F. In general TP can't increase without decreasing FP, we can only balance precision and recall rate to achieve the highest F. Let's take the partial derivatives w.r.t P and R and set them equal to achieve balance.

$$\delta F/\delta P = \delta F/\delta R \tag{16}$$

Then we get the maximal F condition:

$$\frac{\alpha}{P^2} = \frac{1-\alpha}{R^2} \rightarrow \frac{R}{P} = \sqrt{\frac{1-\alpha}{\alpha}} \tag{17}$$

Let $\beta = \sqrt{\frac{1-\alpha}{\alpha}}$, then F_β can be defined as:

$$F_\beta = \frac{(1+\beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall} \tag{18}$$

and when $\frac{R}{P} = \beta$, F_β reaches its maximum. In other words, we prefer models with high TP. And if TP can't increase at the determent of FP, we want to balance FP errors and FN errors such that $\frac{R}{P} = \beta$.

We want the model with largest F has the lowest cost so we have to find the relationship between β and our cost parameters (t, b and λ).First let's express cost function in terms of TP (True Positive), FP (False Positive), FN (False Negative) and TN (True Negative).

$$TotalCost = TP \times Cost_{TP} + FP \times Cost_{FP} + FN \times Cost_{FN} + TN \times Cost_{TN} \tag{19}$$

We could also write cost function w.r.t P (precision), R (recall) and TP:

$$TotalCost = TP \times [(t-b) + (\frac{1}{P} - 1)t + (\frac{1}{R} - 1)b\lambda] \tag{20}$$

We could see models with larger precision and recall will have lower cost. If the cost is negative, models with larger TP will have lower cost and vise versa.

In order to find β, let's fixed TP and leave precision and recall as independent variables to minimize total cost. In equilibrium:

$$\delta Cost/\delta P = \delta Cost/\delta R \tag{21}$$

We could get:

$$\frac{R}{P} = \sqrt{\frac{\lambda b}{t}} \tag{22}$$

This means if total cost is negative, model with highest number of TP points will have lower cost. If TP rates are similar among models, the one that could balance FP and FN error in way that $\frac{R}{P} = \sqrt{\frac{\lambda b}{t}}$ will have lowest cost. If we set $\beta = \frac{R}{P} = \sqrt{\frac{\lambda b}{t}}$, given TP rates are similar among models, the one with largest F_β will have lowest total cost. If total cost is negative, both F_β and cost minimization prefer higher TP and balance FP and FN errors in the same direction. If total cost is positive, F_β is not a good indicator for cost minimization.

8.3 Feature Selection–LASSO

As seen in Figure 9, for objective task and judgment task, if the cutoff is too large or too small, there will be too much noise and the personality information has little prediction power. However, if we pick a moderate cutoff, around top 30% to top 70%, we could identify which group of variables is most related to the screening qualified workers.Remember that in our review task, more than 60% of the task takers have full mark so only full mark cutoff will enable us to identify qualified worker by personality information. Objective task is the easiest, the work experience and customer support experience is not needed. For judgment task which is more difficult, experience is informative in predicting scores.

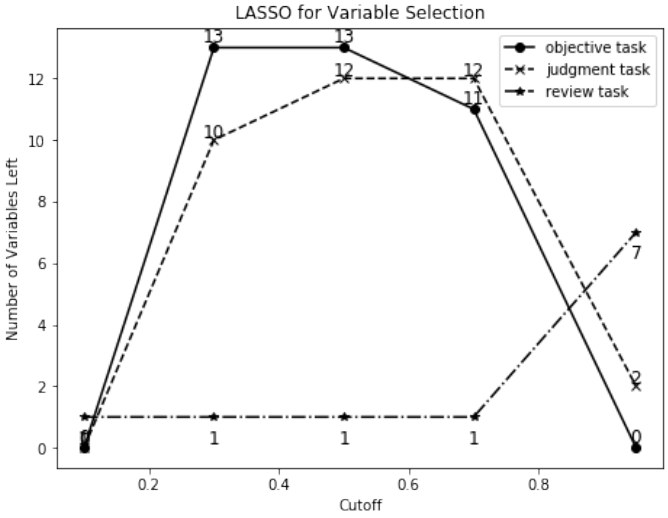
Objective task is the easiest task to predict. In predicting it, as seen in Table 14, some ability-related characteristics, like the working experience and education level has negative correlation with task score while some personality related to persistence has positive correlation, like conscientiousness, locus of control, resilience and successes. For judgment task which is much more difficult, work experience has positive relationship with final score. Some personality types, like extraversion and motivation seems to have negative relationship. However, personality traits like emotional stability, agreeableness, risk aversion, and scores on Ravens tests are useful to predict qualified workers for judgment task. For review task, English score is important, which was not the case on the other

Table 14: LASSO coefficients for different tasks.

Variables	Coefficient (cutoff)		
	Objective task ($\geq$ 52)	Judgment task ($\geq$ 6)	Review task ($\geq$ 2)
Motivation	0	-0.023	0
Conscientiousness	0.011	0.009	0.020
Extraversion	-0.030	-0.007	0
Agreeableness	0.027	0.014	0
Emotional stability	0	0.015	0.004
Openness	0	0	0
Locus of control	0.047	0	0.010
Self-esteem	0	0	0
Grit	0.043	0.020	0
Autonomous function index	0	0	-0.002
Resilience	0.003	0	0
Risk aversion	0	0.058	0
Discount aversion	0.003	0	0
Reading mind in eye	0	0	0
Successes of reading mind tests	0.123	0	0.162
Raven's progressive metrics test	0.090	0.085	0
Successes of raven tests	0.007	0.014	0
Work experience	-0.002	0.042	0
Customer support experience	-0.010	0	-0.004
Education	-0.025	0	0
English score	0	0	0.020

Note: Table 14 presents the LASSO coefficients for the different variables on the different classification tasks.

Figure 9: LASSO for Feature Selection



Note: Figure 9 presents total number of variables left with the LASSO procedures for the objective task, judgment task and review task, using different cutoffs.

two tasks. Also, emotional score, self-esteem and success help predict qualified workers and autonomous and customer support experience related negatively with worker performance for this task.