# AI and AI-Human based Screening and Selection for Salesperson Hiring using Interview Videos

Ishita Chakraborty
University of Wisconsin-Madison, ishita.chakraborty@wisc.edu

Khai Chiong
University of Texas-Dallas, khai.chiong@utdallas.edu

Howard Dover
University of Texas-Dallas, howard.dover@utdallas.edu

K.Sudhir
Yale School of Management, k.sudhir@yale.edu

We consider the problem of AI and AI-human based screening (eliminating bottom candidates) and selection (hiring top candidates) in salesforce hiring. Using videos of structured interviews of candidates and judgements on standard performance criteria by multiple recruiting experts, we develop an AI prediction model of salesforce "skill" by extracting theory-relevant objective measures of interviewee performance embedded in videos. Using the model, we address two issues: First, to aid interpretability of the AI model, we assess what mode of unstructured data from the interview (text, audio and video information) and what specific behaviors (e.g, certain body language or style of conversation) drive AI performance. We find that "interactive conversations" as measured by the number of back and forth between buyer and seller from video data and "willingness to listen" as measured by share of buyer speaking time and the ability to handle buyer objections are most predictive of interviewee sales skills. Second, while research has suggested that AI can serve to augment human decision making, there is relatively limited work on how AI-human hybrid models together improve performance on tasks. We use a Bayesian approach to combine AI prediction with human judgments and we assess predictive performance of the hybrid AI-human model with respect to a pure human panel benchmark. We find that human judgement has limited incremental benefit over AI for screening but improves selection significantly. Further, human input is most important for judging the first 2-3 minutes of the interaction. This suggests a cost-effective way to deploy AI in sales hiring—use it exclusively for screening but augment it with human judgement for selection especially for evaluating the early stages.

*Key words*: Human-AI, Video Analytics, B2B, Salesforce, Machine Learning

## 1. Introduction

The role of automation versus human judgement in different areas has been a widely debated topic across labor economics, HR, sociology as well as in the popular press. Sales recruitment is one area that could benefit tremendously from some degree of automation — recruitment and training is a key spending area for several people-intensive industries like Sales and Marketing, Advertising and Information Technology. A Deloitte 2016 study finds that the cost-per-hire metric has gone as high as $4000.[1] Thus, there is a huge potential cost saving opportunity through automation in the recruitment space. However, as Chapman and Webster (2003) note, though there is a wide interest in organizations to adopt AI technologies at different stages of the hiring process, there is still limited understanding of what factors impact the success of these technological interventions. A typical hiring cycle consists of several stages — sourcing, screening, selection and finally on-boarding and training. [2] The goal of the sourcing stage is to identify a good pool of potential candidates to screen and select from. Next, the screening stage aims to identify certain unobserved human capital relevant to job fit from a set of proxy cues in a reasonably short time to further reduce the applicant pool (Huang and Cappelli 2006). Thus, the main objective here is to answer "who to eliminate at this stage". Finally, this culminates into a selection stage whose goal is to choose the most suitable employees for a job based on information that is most likely to predict future job performance and elicited in a systematic manner (Farr and Tippins 2013). The focus of this study is screening and selection; which accordingly to a large section of recruiters is one of the toughest stages of the hiring process. [3]

[1] https://www.prnewswire.com/news-releases/bersin-by-deloitte-us-spending-on-recruitment-rises-driven-by-increased-competition-for-critical-talent-300070986.html

[2] https://oorwin.com/blog/6-stages-of-the-recruitment-life-cycle.html

[3] 52% of recruiters say that the toughest part about recruitment is screening candidates from a vast pool of potential talent—https://ideal.com/ai-recruiting/

Screening and selection are essentially sorting problems which require the hiring panel to sort candidates into the bottom/top $X$ percentile based on potential suitability for the job and organization. There are both objective and subjective elements of this decision. While the objective elements are based on a score derived from a well-defined testing process, the subjective components are based on more loosely-defined and ill-structured pieces of information (Yakubovich and Lup 2006). As Vincent (2021) note, though AI can be more objective (data-driven) and faster, humans (especially experts) bring in intuition and experience that can lead to high-quality decisions especially in loosely-defined outcome spaces i.e. subjective areas of decision making. Thus theoretically, it seems reasonable to conjecture, that this is an area which is well-suited to combine AI and human intelligence as it involves making a comparatively ambiguous and ill-structured decision in a timely and cost-efficient manner. A combination of human and AI intelligence could improve hiring outcomes such as screening and selection accuracy while reducing the cost of hiring. Besides, there are several reservations and challenges in using pure AI-based systems for HR functions; such as the need for the process to be fair, accountable and explainable and there is a lack of large-scale high-quality training data to train fully automated models (Tambe et al. 2019).[4]

Human-AI hybrid models have been proposed in several domains (Kamar 2016) where automation can improve outcomes but cannot totally replace human judgement, e.g, self-driving cars (Ning et al. 2021), teaching (Holstein et al. 2020) and sales coaching (Luo et al. 2021) but the implementation and effectiveness of these models can be highly context-specific. In the domain of organizational decision-making in particular, there are some

[4] Any hiring tool needs to fulfil the criteria enlisted in the Uniform guidelines on employee selection procedures—https://www.govinfo.gov/content/pkg/CFR-2017-title29-vol4/xml/CFR-2017-title29-vol4-part1607.xml to be considered a valid selection device

theoretical studies that envision such models, e.g., Shrestha et al. (2019). However there is no empirical work that builds and tests the relative performance of hybrid models compared to a pure human or pure AI approach. The aim of this paper is to fill this gap in literature by constructing and testing a video-data based AI and AI-human hybrid model for recruitment screening. Our focus is in the area of sales person recruitment (a $15bn industry). Our AI model is trained using data from real one-one interviews between prospective salespeople and senior industry recruiters conducted during an internship recruitment event in a large US university. The interviews are structured in the format of National Collegiate Sales Competitions. Every interview was rated by a panel of 9-10 human judges which was balanced in terms of gender, experience and industries.Having structured interviews and being evaluated by a balanced panel of industry experts helps to control for a wide range of biases well-documented in employee selection literature (Campion et al. 1997, Pulakos and Schmitt 1995).

We first construct a pure AI model using theory-driven objective metrics of performance captured from the data — content, linguistic style and audio-visual style (voice, body language). Since this model is fully interpretable, it helps to throw light on what aspects of a candidate's performance drive the decisions of the AI model. Interestingly, we find that at the screening stage (selecting worst candidates), the model places highest importance on audio-visual characteristics like having an energetic voice and animated body language. However for selection, the model focuses on more intricate aspects of linguistic style like whether the candidate is being precise and certain as well as making quantitative arguments. We also find that the performance during introduction and question/answering have the most impact on the success of the candidate.

While the AI models achieve reasonable performance compared to the full human panel benchmark, we now ask if we can improve it further by incorporating a small component

of human judgement. We construct a hybrid model that combines AI scoring with human judges' score in a Bayesian fashion. This is similar in principle to the sequential hybrid model described in Shrestha et al. (2019). We find that the hybrid model outperforms pure AI model and reaches very close to human panel judgement both for screening as well as selection. However, only in screening stage, the gains from accuracy are large enough to offset the cost of additional human intervention. Even for selection, there is not much incremental gain in accuracy when we incorporate more than 3 human judges. Finally, we show that the maximum benefit from including human judgement occurs when it is done for the early stages of the interview. This suggests a cost effective manner to deploy AI in video-based recruitment — use pure AI for screening and a human-AI hybrid for selection wherein the human experts only need to judge the initial 2-3 minutes of the interview.

## 2. Literature Review

This paper is connected to three strands in literature — the computer science literature that studies optimal ways to integrate AI and human judgement, the management literature that looks at integration of AI within organization decision-making and the marketing and sales literature on influence tactics and sales recruitment.

### 2.1. Role of AI and Humans in Hiring

While AI is well-suited for structured and well-defined problems (e.g., prediction), it needs to be combined with human intellect and intuition for ambiguous tasks (Vincent 2021, Agrawal et al.). Hiring is a perfect setting where the problem is somewhat loosely defined yet decisions need to be quick and cost-efficient. This goes back to the classic trade-off in decision theory, where accuracy and speed of decision-making is found to be inversely proportional and there are several studies that have focused on understanding how to make high quality decisions quickly (Eisenhardt 1989, Perlow et al. 2002). One might

expect that having AI-assisted screening can reduce time and effort at the cost of the intuition and private information of a human judge, however, intuition is an unconscious and imperfect process and humans often struggle to explain the decisions that are taken based on intuition (Dane and Pratt 2007, Bertrand and Mullainathan 2004, Rivera 2012, Bendick Jr and Nunes 2013). Thus, it remains an empirical question to understand whether or not AI-based or human-only approach gives better outcomes, is more robust and more explainable and whether a hybrid model can outperform either or both.

There is some evidence in the favor of AI e.g., Hoffman et al. (2018) who find that managers who hire against test recommendations end up with worse hires on an average; Cowgill (2018) who find that that algorithms would generally be more reliable at screening especially if the initial training data had considerable noise and Autor and Scarborough (2008) who find that standardized job testing is not harmful for minority workers inspite of the fact that these groups tend to get lower scores on several standardized tests (Hartigan and Wigdor 1989). However, most of these studies define AI as either standardized questionnaires (e.g., personality tests) or at most resume-based screening using some text analysis and predictive modeling (e.g., SVM). Likewise,in the industry, most of the commercial software available for automated hiring focuses mostly on the resume screening stage (Raghavan et al. 2020) and majority of vendors focus on question-based screening which try to replicate existing questionnaires in psychology and organizational behavior. Few companies have experimented with video data which mostly consists of recorded video resumes which are typically 2-3 minutes in duration. In the contrary, our focus is on video-based screening and we use longer duration (20 min) videos which involve a two-sided interaction (in the form of a sales pitch). To the best of our knowledge, the pros and cons of a video-based AI-screening tool has not been studied and there is no study that compares

the performance of a hybrid against the pure AI or pure human approach in the context of salesperson recruitment. (Cowgill 2018) notes, "Counterfactual comparisons between algorithms and other decision-making methods are rare". Our paper bridges this gap in the existing literature by developing and comparing the outcomes of a hybrid AI-human model with those of a pure AI or pure human approach. Video screening is very different from traditional resume-based screening which relies only on certain keywords and phrases to make a decision. In a video, several other modalities (voice, image) are available to the interviewer which can both improve decision making (by providing a richer feature space to train models) but can also result in humans as well as AI models to place too much weight on few or irrelevant audio-visual characteristics.

A promising opportunity in AI is developing systems that can partner with people to accomplish tasks in ways that exceed the capabilities of either individually (Kamar 2016, Bansal et al. 2019). While there are some theoretical papers that envision such models (Shrestha et al. 2019), to the best of our knowledge, this is the first paper in marketing to construct and test the performance of a AI-human hybrid model in the context of salesperson interviewing.

### 2.2. Interpretability: What drives success in sales interviews

Interpretability is a good-to-have feature in any machine learning model (Doshi-Velez and Kim 2017, Vellido 2020) but especially critical for one's used for hiring as these outcomes have long term impact on an individual's life, a firm's profit and societal welfare. There have been some instances of black-box hiring tools leading to extremely harmful social outcomes.[5] Hence, we make our AI model fully interpretable and in line with the past

[5] https://www.washingtonpost.com/technology/2019/10/22/ai-hiring-face-scanning-algorithm-increasingly-decides-whether-you-deserve-job/

research on sales influence tactics. Instead of a bottom-up approach of creating a laundry list of features and then selecting few using feature importance studies, we start by developing a theoretical framework for factors that should impact the success of a candidate in sales recruitment drawing from a rich literature in personal selling and persuasion. Starting from Aristotle's treatise that identifies three main channels of persuasion — logos (logic), pathos (emotions) and ethos (value system), modern-day scholars have come up with more nuanced ways to define influence tactics (e.g., argumentation schemes from Walton et al. (2008), the IBQ questionairre from Yukl et al. (2008) or principles of persuasion from Cialdini and Cialdini (2007)). Literature in personal selling (Sheth 1976, Frazier and Summers 1984, Spiro and Weitz 1990) has identified the importance of content, style and interactivity as the most important factors determining sales success. However, this literature is mostly survey based and though they could identify high-level influence tactics, they could not break it down into specific behaviors. It is extremely hard to precisely measure constructs like *linguistic or audio-visual style* in the absence of recorded videos of buyer-seller interactions. Likewise, research on interactivity and adaptability has been largely inconclusive (Churchill et al. 1975, Evans 1963) due to the absence of similarity measures beyond the demographic characteristics of buyers and sellers. Our granular video-based data allows us to construct features that can capture low-level behaviors associated with these high-level tactics. This is an important contribution as it helps to generate more actionable insights for salesperson training.

There are some recent studies in computer science (Longpre et al. 2019, Shmueli-Scheuer et al. 2019) that study persuasion using audio-visual data. However, these studies primarily focus on one-sided and single-shot instances of persuasion for example an advertisement or a call for action. On the other hand, our setting is a two-way, multi-stage buyer-seller

interaction where interactivity and adaptability are key factors over and above message content and delivery. Manzoor et al. (2020) show the impact of reputation (or ethos) on persuasion outcomes controlling for content and stylistic linguistic factors. However, in our context, reputation does not play a role as all the interviewees are students and this is a one-time interaction. Thus, our interpretable model of sales persuasion can add to this small but growing literature that studies persuasion using multi-modal data.

To summarize, our key contributions are the following — we construct and test the first human-AI hybrid models in the domain of salesperson recruitment. In the process, we test several theoretical predictions about the performance of such hybrid models. Second, unlike past literature where the AI component has typically consisted of only questionnaires or text-based models, our AI model is constructed from multi-modal video data that comprises of text, voice as well as visual components. Finally, our model is fully interpretable which allows us to understand which elements in a sales pitch are most important in driving success. Our findings have implications for both salesperson recruitment as well as training.

## 3. Background and Data

The dataset consists of 195 videos of in-person sales interviews in the format of a National Collegiate Sales Competition (https://www.ncsc-ksu.org/). Every interview is a sales-pitch roleplay where the student plays the role of a seller and persuades the corporate buyer (interviewer) to buy a certain product (the product is a subscription for a well-known CRM product). Each sales interview is then scored by a panel of industry judges comprising of sales executives from Fortune 500 companies.

Each video is approximately 15 minutes long. There are 29 distinct interviewers, who are sales executives from different companies and have varying levels of experience (from managers to VPs and CEOs). On average, each of these interviewers engages with 8 different candidates in separate interviews. In addition, each sales interview is evaluated by

a panel of 9 different judges. Overall, 261 expert judges have scored these videos resulting in 1752 unique evaluations. Thus, each buyer conducts multiple interviews and every judge evaluates multiple interviews. This gives us a unique panel dataset which allows to estimate interviewer and judge specific effects. Note that each student candidate is only interviewed once. Table 1 shows the distribution of demographics for students, interviewers and judges. As seen in the table, we have a good balance of gender and experience among the buyers and judges.

Firms that are participating in this process are looking to identify salesforce talent, whereas students are seeking for a position in the B2B salesforce industry. As a result, both buyers and sellers are motivated by real-world incentives to perform well at the role-plays. See Fig 1 for a snapshot of the interview setting.

**Table 1 Interviewee and Recruiter Demographics**

| | N | Gender | |
|---|---|---|---|
| **Students (Interviewee)** | 195 | 96 female, 99 male | |
| | N | Gender | Experience |
| **Recruiters (Interviewers)** | 29 | 13 female, 16 male | 9 High , 20 low ($\leq$10 yrs) |
| | N | Gender | Experience |
| **Judges** | 291 | 99 female, 192 male | 134 High, 157 low ($\leq$10 yrs) |
| **Evaluations** | | 1752 | |

The judges score these interactions using the National Collegiate Sales judgement criteria (See Fig 2). Our data comprises of all the judge scores as well as the actual videos of the buyer-seller interaction. We use AWS Transcribe to extract time-stamped text transcripts of this exchange. [6] Thus, the full dataset is multi-modal — it includes video (which has

[6] The transcription software is not 100% accurate, we use student RAs to verify some of these transcripts randomly. We noticed some common discrepancies — the software usually got some highly context-specific terms wrong (e.g., CRM, Salesforce, quotas). We then specifically searched these terms and corrected the transcriptions.

both audio and visual components) as well as text (from the transcripts) and an associated judge scoring matrix. The judges rate the sellers based on the following stages of the interaction:

1. *Approach* This is the stage where the seller greets the buyer and tries to engage in small talk in order to understand the buyer better and build a rapport. This stage typically comprises of the first 2-3 mins of a conversation

2. *Need Identification* This stage follows the introduction stage. Here, the seller digs deeper into the problem areas for the buyer's organization and tries to uncover implicit and explicit needs. Sellers typically spend 6-8 minutes for need identification

3. *Presentation* This is where sellers present their product to the buyers, usually they would give a demo of the software and describe key features and how it can meet the needs of their organization. This stage typically consists of 2-5 minutes.

4. *Objection Handling* In this stage, the buyer raises objections about the product and the seller has to put forth counter arguments. The length of this stage is normally 4-5 minutes, however sometimes this continues for longer and the seller runs out of time.

Fig 2 shows the scorecard that the judges use for rating the videos. There are sub-questions under each of the stages of the interaction (approach, needs identification, presentation, objection handling). Every question is scored on a scale of 1-10, however, the minimum score across questions is 7 in the data. There are unequal number of questions in each stage, for example the Demo has 6 questions and Objection handling has 3, hence the max total score could differ across stages. The decision to select or not is based on the Overall Total Score which is derived from the stage-wise scores. Table 2 shows the distribution for different score components in the scorecard. We can see that the highest variance is in the scoring of the demo and need identification stages.

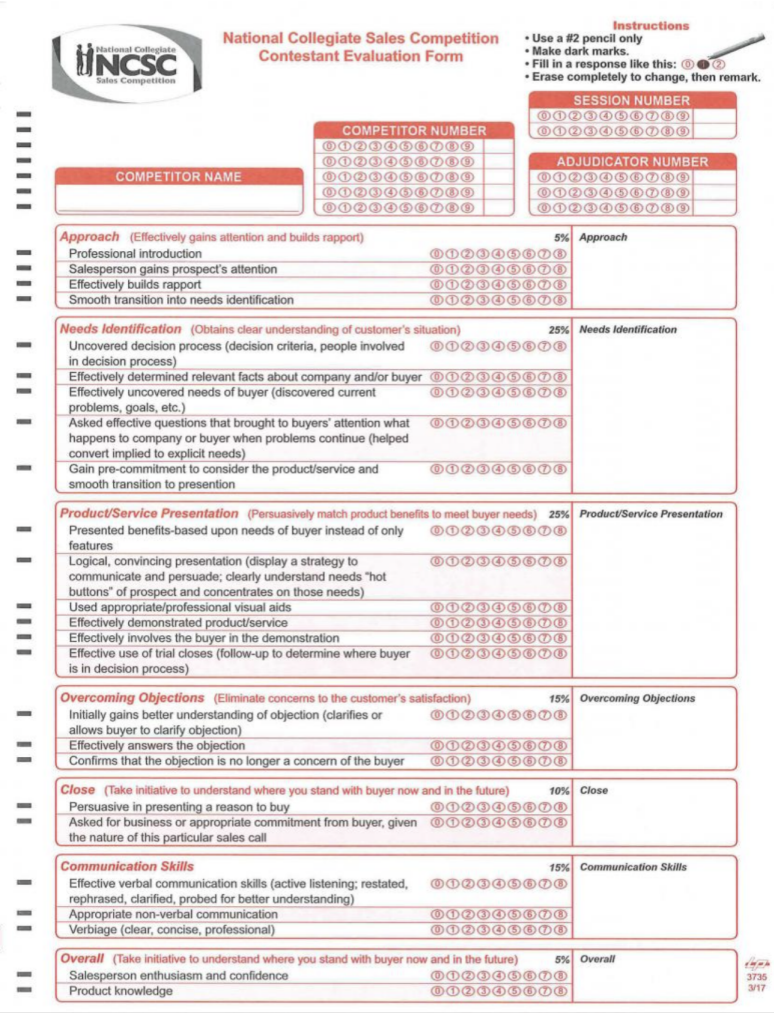
**Table 2 Scorecard: Summary Statistics**

| | | Stages | | | | Entire interview | | |
|---|---|---|---|---|---|---|---|---|
| | **Overall** | **Approach** | **Need** | **Objection** | **Demo** | **Communication** | **Confidence** | **Persuasion** |
| **Max** | 100 | 30 | 50 | 30 | 60 | 30 | 10 | 10 |
| **Min** | 70 | 21 | 35 | 21 | 42 | 21 | 7 | 7 |
| **Mean** | 87 | 27 | 44 | 25 | 51 | 27 | 9 | 8.4 |
| **Median** | 86 | 26 | 43 | 25 | 52 | 27 | 9 | 8 |
| **SD** | 6.6 | 2.17 | 3.6 | 2.6 | 4.6 | 2.3 | 0.89 | 1.01 |

**Figure 1 Sellers: student job seeker (51% M: 49% F). Buyers: salespeople/recruiters (56% M: 44% F)**

## 4. Model

We first state the problem of the hiring firm, and the outcome variable of interest. The hiring company is interested in screening (identify the inferior candidates below a threshold) and selection (find the top candidates). To screen and select candidates, the hiring company ranks these candidates according to a real-valued score. This score should ideally reflect the actual ability of the candidate, separating out the effect of the interviewers and judges. However, in order to tease out the interviewer and judge effect, we require a panel dataset where we observe multiple interviews conducted/evaluated by a particular interviewer/judge. Fortunately in our setting, we have access to this type of data. However, in many real-life scenarios, firms may not always have this type of panel data. In these scenarios, the next best option is to simply use the average score of the hiring panel. We call this Human Average Heuristic. For training and testing our AI models, we use the unbiased estimate of candidate ability (derived from a fixed effects model) as that is a


National Collegiate Sales Competition
Contestant Evaluation Form

Instructions
- Use a #2 pencil only
- Make dark marks.
- Fill in a response like this:
- Erase completely to change, then remark.

SESSION NUMBER
COMPETITOR NUMBER
ADJUDICATOR NUMBER
COMPETITOR NAME

| | | |
|---|---|---|
| **Approach** (Effectively gains attention and builds rapport) | 5% | Approach |
| Professional introduction | | |
| Salesperson gains prospect's attention | | |
| Effectively builds rapport | | |
| Smooth transition into needs identification | | |
| **Needs Identification** (Obtains clear understanding of customer's situation) | 25% | Needs Identification |
| Uncovered decision process (decision criteria, people involved in decision process) | | |
| Effectively determined relevant facts about company and/or buyer | | |
| Effectively uncovered needs of buyer (discovered current problems, goals, etc.) | | |
| Asked effective questions that brought to buyers' attention what happens to company or buyer when problems continue (helped convert implied to explicit needs) | | |
| Gain pre-commitment to consider the product/service and smooth transition to presention | | |
| **Product/Service Presentation** (Persuasively match product benefits to meet buyer needs) | 25% | Product/Service Presentation |
| Presented benefits-based upon needs of buyer instead of only features | | |
| Logical, convincing presentation (display a strategy to communicate and persuade; clearly understand needs "hot buttons" of prospect and concentrates on those needs) | | |
| Used appropriate/professional visual aids | | |
| Effectively demonstrated product/service | | |
| Effectively involves the buyer in the demonstration | | |
| Effective use of trial closes (follow-up to determine where buyer is in decision process) | | |
| **Overcoming Objections** (Eliminate concerns to the customer's satisfaction) | 15% | Overcoming Objections |
| Initially gains better understanding of objection (clarifies or allows buyer to clarify objection) | | |
| Effectively answers the objection | | |
| Confirms that the objection is no longer a concern of the buyer | | |
| **Close** (Take initiative to understand where you stand with buyer now and in the future) | 10% | Close |
| Persuasive in presenting a reason to buy | | |
| Asked for business or appropriate commitment from buyer, given the nature of this particular sales call | | |
| **Communication Skills** | 15% | Communication Skills |
| Effective verbal communication skills (active listening; restated, rephrased, clarified, probed for better understanding) | | |
| Appropriate non-verbal communication | | |
| Verbiage (clear, concise, professional) | | |
| **Overall** (Take initiative to understand where you stand with buyer now and in the future) | 5% | Overall |
| Salesperson enthusiasm and confidence | | |
| Product knowledge | | |

3735
3/17

**Figure 2 NCSC scoring criteria**

better indicator of true quality. [7] However, for the purpose of comparison, we benchmark our AI and hybrid models to the Human average Heuristic as that is most widely used in practice. Next we describe both the fixed effects model (to estimate candidate's true potential) as well as the human average heuritic in more detail.

[7] In Chakraborty et al. (2021) we show how the fixed effects benchmark is superior to the simple average benchmark in terms of achieving greater gender balance in the selected sample.

### 4.1. Estimate of candidate's ability: fixed effects model

We have a panel dataset where an interviewer conducts multiple interviews and a judge evaluates multiple sales interactions. We can use this panel structure to separate a candidate's actual potential $\theta$ from any judge or interviewer-specific effects.

Let $i$ denote the candidate salesperson who is being evaluated for the job. Let $j$ denote the interviewer or buyer who is conducting the interview. Let $h$ denote a human judge who is scoring the interview. The hiring company wishes to screen and rank interviewees based on their true abilities, separating out interviewer's effect and judge's biases.[8]

Let $S_{ijk}$ denote the score given to the candidate $i$ by a human judge $j$, when the interviewer is $k$:

$$S_{ijk} = \theta_i + \gamma_j + \delta_k + \epsilon_{ijk} \tag{1}$$

We take the score $S_{ijk}$ as the *Total Score*, summing up all the components of the scoring criteria described in Figure 2.

Here $\theta_i$ is the candidate $i$'s ability, $\gamma_j$ is the effect of judge $j$, and $\delta_k$ is the effect of the interviewer $k$. The judge's effect can have two components, a judge-specific fixed effect $\alpha_j$, and $\boldsymbol{H}_j$, a vector of demographic variables such as gender, experience etc.

$$\gamma_j = \alpha_j + \boldsymbol{H}_j \boldsymbol{\beta}_2$$

We estimate Equation 1 using linear least squares, obtaining $\hat{\theta}_i$ which is the candidate's true ability separating out the effect of judges and interviewers.

[8] A candidate might score higher just because he or she is matched with certain interviewers or evaluated by judges who favor a certain style

### 4.2. Human Average Heuristic

Here we describe the actual practice of the National Colleagiate Sales Competition where they rank candidates based on *simple average of the Total Score*. From the full panel of 291 judges, we have 1,752 such *Total Score*. A simple average is then used to collapse the scores such that each candidate is associated with one score. Below we summarize the procedure:

1. For each candidate $i$, judge $j$, the *Total Score* $S_{ij}$ is obtained by by taking a weighted sum of different components according to the weights given in Figure 2.
2. For each candidate $i$, calculate the average score $S_i$ by taking a simple average of $S_{ij}$, that is, $S_i = \frac{1}{|\mathcal{J}_i|}\sum_{j\in\mathcal{J}_i} S_{ij}$, where $\mathcal{J}_i$ is the list of judges who scored $i$.
3. Rank candidates according to $S_i$. For *Screening*, we identify all the candidates whose ranking is below the 25th percentile. For *Selection*, we identify all candidates whose rankings are above the 75th percentile.

This *Human Average Heuristic* will serve as a benchmark against which we measure the improvement of our AI model.

### 4.3. The AI model

The goal of the AI model is to rank candidates for screening and/or selection. Once trained, the AI model can screen and select candidates based on interview videos without additional scoring by humans. Of course, we still need a human interviewer to act as a buyer, and the interview itself needs to be recorded.

The training of the AI model consists of predicting $\theta_i$ (candidate's fixed effect) as a function of $\boldsymbol{X}_i$, where $\boldsymbol{X}_i$ is a vector of video features, as in Equation 2 below. $\boldsymbol{X}_i$ includes textual and audio-visual elements such as rate of hand gestures, voice tonality, conversational interactivity, etc, which we will describe in the subsequent sections.

$$\theta_i = f(\boldsymbol{X}_i) \tag{2}$$

We have estimated the candidate's actual potential $\theta_i$ using Equation 1. Now, training the AI model involves finding the relationship between $\hat{\theta}_i$ (target variable) and $\boldsymbol{X}_i$ (feature set) in the training dataset. We experiment with different types of non-linear estimators like Random Forest Regressor and SVM. Once trained, this model is used to predict $\hat{\theta}_i$ from the features extracted from the testing dataset. We then use $\hat{\theta}_i$ to rank candidates. This ranking determines if the candidate falls below a certain threshold for screening, or above a threshold for selection. These scoring thresholds are determined by the hiring company's objectives.

### 4.4. Performance Measures

The predictive accuracy of our different models will be assessed in terms of its ability to classify a candidate into their correct percentile class. Thus, while the original problem involved a continuous dependent variable (score), screening and selection (classifying to bottom/top X percentile) is a binary classification problem. A correct classification decision in the context of screening/selection means the candidate is correctly classified into bottom/top X percentile. Table 3 shows different types of misclassifications that could occur. A false positive in the screening stage means a good candidate is rejected whereas a false negative means a bad candidate passes the screening and moves to the next round. At this stage, the cost of false negatives is usually higher as one needs to spend more managerial time and cost to evaluate them further. In the selection stage, both types of misclassifications could be equally harmful.

The most common accuracy metric is the hit rate which is defined as Hit Rate $= (tp + tn)/(tp + tn + fp + fn)$ where $tp, tn, fp, fn$ stand for true positive, true negative, false positive and false negative respectively. However, this is not an adequate metric when there is a large class imbalance problem which is a characteristic of the screening and selection

problem by definition. For example in the selection problem for Top 10/25 percentile, only a small fraction of data will belong to the positive class (selected). Hence the most appropriate metric for this problem is Balanced Accuracy which is an average of the sensitivity or true positive rate and specificity or true negative rate.

$$\text{Balanced Accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2}$$

$$\text{Sensitivity} = \frac{tp}{tp + fn}$$

$$\text{Specificity} = \frac{tn}{tn + fp}$$

**Table 3 Interpretation of Misclassification**

| | Screening | Selection |
|---|---|---|
| False Positive | Good candidate rejected | Bad Candidate selected |
| False Negative | Bad candidate moves to next round | Good candidate rejected |

We calculate and report Balanced Accuracy for both screening and selection. For the task of screening, we additionally define the *Screening Error Rate* as the proportion of candidates who are ranked in the top 25th percentile according to the ground truth ($\theta_i$ in Equation 1) but who are screened out (ranked below 25th percentile) according to the predictive models, which can be either Hybrid, AI or Human Average Heuristic. Given that the goal of screening is to ensure better selection, this is an even better indicator of screening success compared to Balanced Accuracy alone.

We now describe the process of feature selection for the AI model.

### 4.5. Feature Engineering for AI model

Based on past literature, we hypothesize that four types of features impact persuasion: content of the message, linguistic style, audio-visual style and interactivity between the persuader and persuadee. We now briefly explain each of the feature groups and how they would be extracted from our data. Figure 3 is a broad overview of the process.

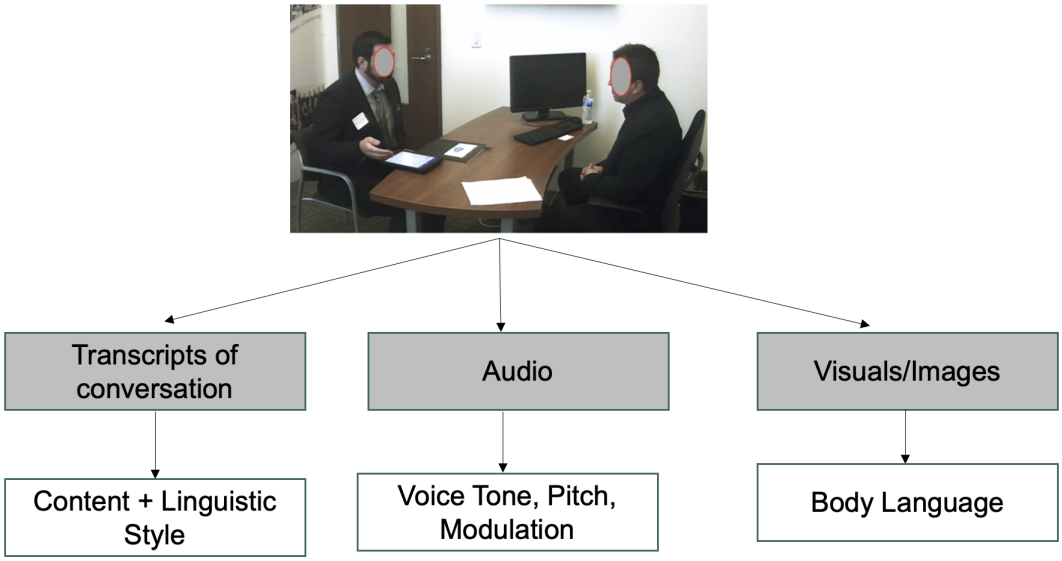



**Figure 3 Feature Extraction from Video data**

1. Content: This comprises of the substantive, topic-oriented aspects of the message excluding any stylistic elements. We run an LDA topic model on our entire conversation transcript data and find 4 important topics that are discussed in these interactions. They are closely related to the four different stages of the conversation. These include topics around *Greeting and Pleasantries*. These generally occur during the beginning of the conversation and involve words and phrases like "Hey", "Good Morning", "weather". As the conversation progresses, the other topics can be categorized into *Busines*, *Technology* and *Pricing*. The *Business* topic is related to the buyer's business problem and includes words like "salesperson", "manager", "growth" and "profitability". The *Technology* topic measures technical aspects of the product (as the scenario involves selling a software product) and consists of words like "app","cloud", "infrastructure" to name a few. *Pricing* topic talks about the cost of the setup and payment terms and conditions. In Table 12 in the Appendix, we summarize the top words for each of the topics. The content-related features include the proportion of each topic mentioned by a seller during the interaction.

2. Style (Linguistic): We capture the linguistic style of both the interviewer as well as interviewee by tagging the conversation transcripts for 23 LIWC (Pennebaker et al. 2001) categories that are most relevant to this context — e.g., verbosity (word-count, words per sentence), positive and negative emotion words, words signifying certainty as well as tentativeness (e.g., absolutely, perfectly, obviously, maybe, could be, try), assent words (yeah, yes, of course) to name a few. This results in a large feature set of linguistic elements which are hard to interpret, hence we perform a Principal Components Analysis on Interviewer and Interviewee linguistic elements to derive 6 types of language styles from these 23 dimensions. In Table 4, we elaborate on each of these styles.

**Table 4 Linguistic Styles: Seller and Buyer**

| **Buyer** | **Description** |
|---|---|
| Type 1 | Precise, Quantitative, Asking Questions |
| Type 2 | Authentic, Disagreeing, Slightly Verbose, Individualistic |
| Type 3 | Analytical, Talks about Business/Money |
| **Seller** | |
| Type 1 | Asking Questions, Less Quantitative, Verbose |
| Type 2 | Business-Oriented, Authentic, Showing power, Certainty |
| Type 3 | More Quant, Slightly Informal |

3. Style (Audio-Visual): Voice and body language (kinesics, proxemics) are also shown to be important elements that influence how a message is interpreted over and above the actual content and linguistic style. In voice, style is mainly conveyed by energy, pitch, speech rate and voice modulation. We use the PyAudio library in Python to extract these low-level voice features. Visual elements that can communicate style include hand movements (velocity and amplitude) as well as body postures. We use OpenPose[9] to extract low-level features such as the pixel locations of various body

[9] Wei et al. (2016), Cao et al. (2017), Simon et al. (2017), Cao et al. (2019)

parts in each video frame (elbows, wrists, neck, shoulders, hip, nose, ears, etc). From these low-level features, we compute higher-level body-language features such as the average velocity of hand-movements, average distance between hands, average torso angles, frequency of head movements. From these higher-level audio-visual features, wee are able to identify three distinct styles through PCA analysis (5).

**Table 5 Audio Visual Style: Seller and Buyer**

| Buyer/Seller | Description |
|---|---|
| Type 1 | Good voice modulation, bright voice and energetic body movements |
| Type 2 | Energetic voice, stable body language |
| Type 3 | Less energetic voice and body movements |

4. Interactivity: This measures whether the dialogue involves active participation from both seller and buyer. A highly interactive conversation would include higher number of turns and would not have long monologues from either the buyer or seller.

In Table 6, we describe some of the important textual, audio-visual and interactivity related features.

### 4.6. Human-AI hybrid model

Given AI predictions, how much can we improve on the AI model by collecting and combining human recommendations to form a Human-AI hybrid system? We model an additional human recommendation as observing a noisy signal of the true value $\theta_i$.

Let $f(\theta_i|\boldsymbol{X}_i)$ denote the distribution of A.I. predictions using video features $\boldsymbol{X}_i$. There are various ways of estimating $f(\theta_i|\boldsymbol{X}_i)$, one way is to use bootstrapping to non-parametrically recover the distribution of A.I predictions. We can also model $f(\theta_i|\boldsymbol{X}_i)$ as $\mathcal{N}(g(\boldsymbol{X}_i), V(\boldsymbol{X}_i)))$, where $g(\boldsymbol{X}_i)$ is the point-prediction corresponding to Equation 2 using random forest or SVM, and $V(\boldsymbol{X}_i)$ is the standard errors for random forests calculated using the jackknife method of Wager et al. (2014).

**Table 6 Textual, Audio, Visual and Interactivity Features**

| Type | Modality | Feature | Explanation |
|---|---|---|---|
| Content | Text | Topic Proportion | Which topics interviewee prioritises |
| Linguistic Style | Text | Words per sentence | This measures how verbose the sentences are |
| Linguistic Style | Text | Complex words | Using words that have more than 6 letters |
| Linguistic Style | Text | Parts of speech | Proportion of verbs, aux verbs, adjectives and prepositions |
| Linguistic Style | Text | Individualistic | "I", "I'm", "me", "my" |
| Linguistic Style | Text | Collaborative | "we", "us", "both" |
| Linguistic Style | Text | Assent | "yes", "obviously", "yeah" |
| Linguistic Style | Text | Filler Words | "hmm", "aah","huh" |
| Linguistic Style | Text | Certainty | "ofcourse", "definitely", "absolutely","confident", "sure" |
| Linguistic Style | Text | Tentativeness | "maybe", " could be", " likely" |
| Linguistic Style | Text | Emotional Words | Positive and Negative Sentiments |
| Linguistic Style | Text | Politeness | "please", "Thanks", "grateful" |
| Linguistic Style | Text | Questions | Question Marks, "how","when", "Where" |
| Linguistic Style | Text | Thinking/Insight | "think", "most likely" |
| Linguistic Style | Text | Cause | "because", "reason", "causes", "therefore" |
| Linguistic Style | Text | Difference | "but", "yet"," disagree" |
| Linguistic Style | Text | Achieve | "success", "win", "launch", "therefore" |
| Voice | Audio | Energy | Measures how energetic or enthusiastic the seller is |
| Voice Style | Audio | Speech Rate | No of words per sec |
| Voice Style | Audio | Brightness | The Spectral Centroid measures whether the voice is bright or dull |
| Voice Style | Audio | Voice Modulation | No of times the voice signal goes from low to high and vice versa |
| Body Language | Visual | Hand Movement velocity | Right Hand/ Left Hand |
| Body Language | Visual | Distance between hands | Open versus closed Hand gestures |
| Body Language | Visual | Face Movement Velocity | Nodding, being responsive |
| Body Language | Visual | Torso angle | Upright versus stooping posture |
| Body Language | Visual | Torso movement | Changes in Posture |
| Interactivity | Text/Audio | Turns | No of turns in the buyer-seller conversation |
| Interactivity | Audio | Buyer_Max_Monologue | The maximum duration (in secs) during which the buyer talks uninterrupted |
| Interactivity | Audio | Seller_Max_Monologue | The maximum duration (in secs) during which the seller talks uninterrupted |
| Interactivity | Audio | buyer_Seller_SOV | Ratio of time buyer speaks to seller speaks |

Let $f(h_{ij}|\theta_i, \boldsymbol{X}_i)$ denote the distribution of judges' scores. Upon obtaining a human score $h_{ij}$, we update the AI predictions via Bayesian updating: $f(\theta_i|h_{ij}, \boldsymbol{X}_i) \propto f(h_{ij}|\theta_i, \boldsymbol{X}_i)f(\theta_i|\boldsymbol{X}_i)$. To illustrate the AI-human hybrid model, consider a simpler case where:

$$f(\theta_i|\boldsymbol{X}_i) \sim \mathcal{N}(g(\boldsymbol{X}_i), \sigma_0^2) \tag{3}$$

$$f(h_{ij}|\theta_i, \boldsymbol{X}_i) \sim \mathcal{N}(\theta_i, \sigma_1^2) \tag{4}$$

Equation 4 says that conditional on the actual ability of the candidate being $\theta_i$, the judge's score is unbiased and centered around $\theta_i$. Moreover, the precision of the judge's score does not vary by candidates. Here, $f(h_{ij}|\theta_i, \boldsymbol{X}_i)$ is the distribution in which the human interventions are drawn from, which is a noisy signal of $\theta_i$.

Upon receiving $h_{ij}$, one updates the AI score in a Bayesian fashion to arrive at the posterior distribution $f(\theta_i|h_{ij}, \boldsymbol{X}_i)$ whose mean is:

$$\text{AI-human hybrid estimate } = \left(\frac{g(\boldsymbol{X}_i)}{\sigma_0^2} + \frac{h_{ij}}{\sigma_1^2}\right) \Big/ \left(\frac{1}{\sigma_0^2} + \frac{1}{\sigma_1^2}\right)$$

or equivalently, we see that the hybrid estimate is just a weighted average of the AI prediction and the human intervention:

$$\lambda g(\boldsymbol{X}_i) + (1-\lambda)h_{ij} \quad \text{for } \lambda \in (0,1)$$

More generally, judges' scores can be a biased signal of $\theta_i$.

$$f(h_{ij}|\theta_i, \boldsymbol{X}_i) \sim \mathcal{N}(\theta_i + \boldsymbol{H}_j\boldsymbol{\alpha}, \sigma^2 V(\boldsymbol{X}_i)) \tag{5}$$

Conditional on the true quality of the candidate being $\theta_i$, the score for this candidate is centered around $\theta_i + \boldsymbol{H}_j\boldsymbol{\alpha}$, where $\boldsymbol{H}_j$ is the vector of characteristics of the judge (e.g. gender and experience). The precision of this score depends on video characteristics $\boldsymbol{X}_i$.

Suppose we increase the number of human interventions, then in the simple version of the hybrid model, the hybrid estimator is a weighted average of the AI prediction and the average of the human scores $\bar{h}_i$. The weight placed on the AI vs. human scores depends in part on the hyper-parameters in the prior specification, which we are free to choose. We tune these hyper-parameters via a cross-validation procedure.

We note that the AI model also depends on human judgement for initial training. However, the distinction between AI and AI-human hybrid models is in the post-training stage. Once trained, an AI model can be used independently to predict future outcomes without any human intervention. On the other hand, the hybrid will continue to need human scores even after it is constructed.

# 5. Results

The result section has two parts. In the first part, we describe the best AI model and what drives the performance of the model. In the second part, we focus on the hybrid.

## 5.1. Optimal AI model

In Table 7, we first compare the performance of the different AI models we built using the linguistic, audio-visual and interactivity features described in the model section. As described in section 4.4, the most important measure of performance is balanced accuracy as we have a largely imbalanced dataset. We can see from Table 7 that the three machine learning classifiers bring only a small improvement in selection of the top 10/25 percentile compared to a dummy classifier which classifies every observation into the most frequent class. However, in screening bottom candidates, all classifiers and especially SVM does considerably better than random chance.

In Table 8, we further break down the performance of the best ML classifier (SVM) into different scoring dimensions — objective, persuasion, confidence and rapport. *Objective* consists of scores from the *Needs Identification* and *Objection Handling* stages (see Figure 2), which are considered to be more objective as opposed to subjective. *Persuasion*, *Confidence* and *Approach* are considered to be more subjective. *Approach* is the first stage of the sales process and it is based on just the initial 2-3 minutes of the interview. Overall, we find that the AI models do better in objective scoring especially for screening. However, AI is surprisingly good at scoring *confidence* and selecting the top candidates based on this measure. This shows that the notion that AI is bad at subjective scoring is not correct; if we have good quality training data and sufficient agreement among human scorers on how a trait is to be measured (e.g., what constitutes as high confidence), then AI can be equally good at predicting subjective and objective scores.

**Table 7 Performance: AI models**

| | Balanced Accuracy (5 fold cross validation avg) | | | | |
|---|---|---|---|---|---|
| Estimator | Top 10 | Top 25 | Top 50 | Bottom 25 | Bottom 10 |
| Dummy classifier | 50% | 50% | 50% | 50% | 50% |
| Random Forest | 50% | 52% | 60% | 54% | 50% |
| XGBoost | 54% | 53% | 63% | 54% | 52% |
| SVM | 50% | 55% | 66% | 59% | 54% |

**Table 8 SVM Performance for different Scoring Dimensions**

| | Balanced Accuracy (5 fold cross validation avg) | | | | |
|---|---|---|---|---|---|
| Estimator | Top 10 | Top 25 | Top 50 | Bottom 25 | Bottom 10 |
| Objective | 51% | 57% | 55% | 52% | 58% |
| Persuasion | 53% | 48% | 51% | 46% | 53% |
| Confidence | 63% | 57% | 65% | 55% | 49% |
| Approach | 57% | 50% | 54% | 47% | 55% |

**5.1.1. Interpretability: What features drive AI performance** We now look at which features are most important in driving the decisions of the AI model. Since the AI is trained on human judgement, it also throws light on which factors influence the success of candidates in interviews. We use the permutation importance method for Random Forest (Altmann et al. 2010) to unpack the decisions of the random forest estimator. Figure 4 shows which features impact the screening decision the most i.e. identifying the bottom $X$ percentile. We find that the most important feature is an energetic voice and animated body language (audiovis3). More specifically, having wide hand gestures and less frequent hand movements are correlated with higher success. This is followed by turns in the conversation or being *interactive*. Thus, the initial screening decisions seem to place less importance on the *content* or *linguistic style* of the interviewee and are primarily driven by their voice quality and body language. Figure 5 shows which features are driving the selection of the top candidate. In this case, however, we find that the most important features are those around linguistic style (liwcs1). This particular style stands for being precise and quantitative, asking questions and being agreeable. The interactivity parameter that matters the most seems to be the "willingness to listen" as derived from the share of voice. A higher buyer's share of voice leads to a more successful conversation.

**Figure 4 Factors Influencing Screening**

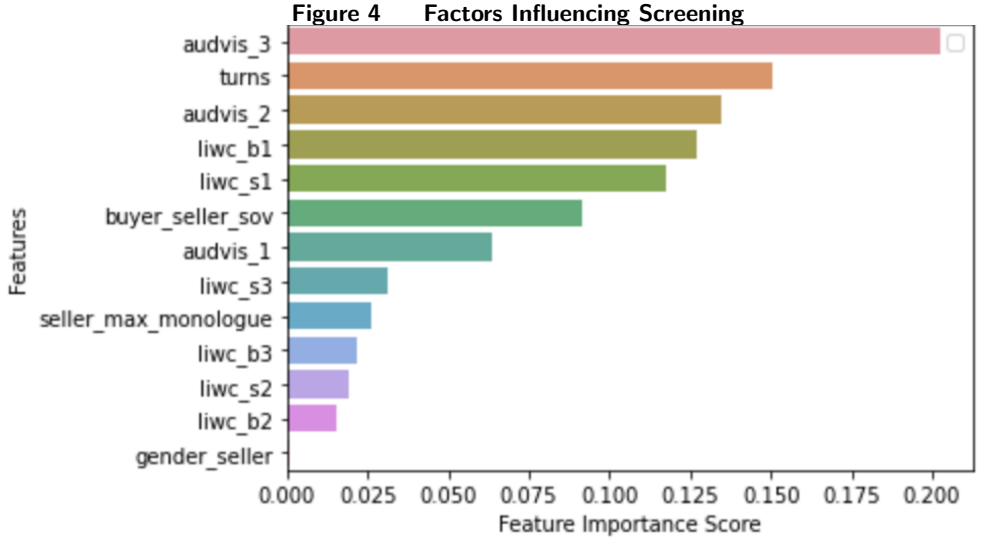
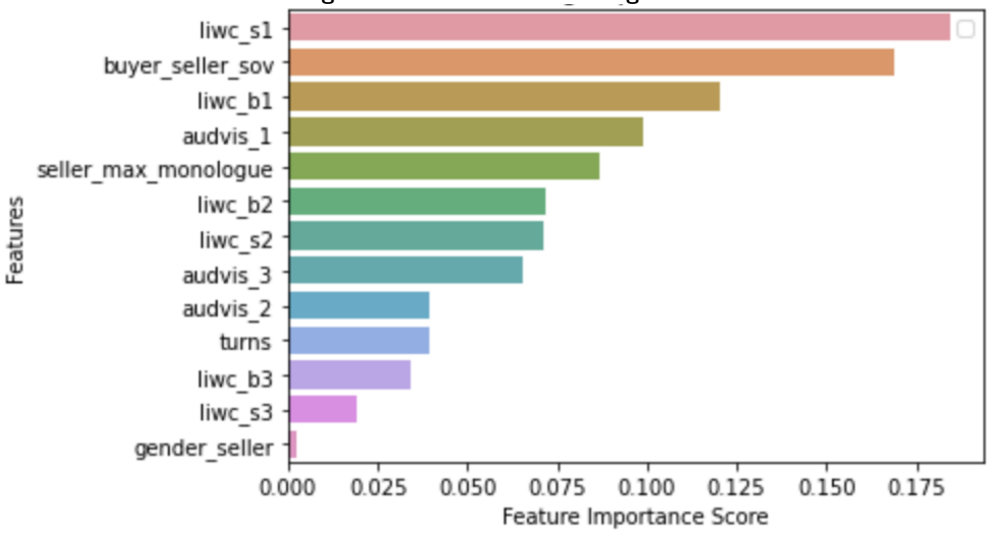



**Figure 5 Factors Influencing Selection**




### 5.2. Optimal AI-human hybrid model

Recall that in the simple version of the hybrid model, the hybrid estimator is a weighted average of the AI prediction $g(\boldsymbol{X}_i)$, and the average of the human scores $\bar{h}_i$. That is, AI-human hybrid estimate $= \left(\frac{g(\boldsymbol{X}_i)}{\sigma_0^2} + \frac{\bar{h}_i}{\sigma_1^2}\right) \Big/ \left(\frac{1}{\sigma_0^2} + \frac{1}{\sigma_1^2}\right)$. The optimal hybrid model consists of choosing the optimal weight to put on AI versus human judgement. We tune these hyper-parameters via a cross-validation procedure.

The optimal weights are reported in Figure 6. In the figure, the blue and green line show the case of screening bottom 25 percentile and selecting top 25 percentile respectively

whereas the orange line denotes choosing the median. The most important insight from Figure 6 is that the relative weight to be put on AI versus human judgement varies by the nature of the task (screening vs. selection); in the case of screening bottom candidates, one must put a high weight on AI, however, to select the best candidates, a higher weight needs to be put on human judgement.

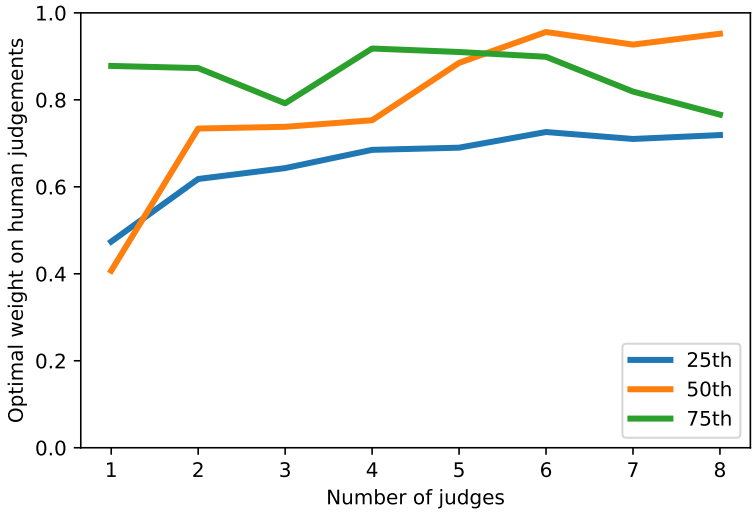



**Figure 6 Optimal weight placed on human judgements**

We vary the degree of human interventions by increasing the number of human judges used in the Hybrid model. Specifically, when the number of judges is $J$, it means that we randomly draw $J$ judges per candidate without replacement (from the testing dataset), and update these $J$ human scores into the AI model according to the Bayesian updating model in Section 4.6. Then for a given updating weight, we can calculate the balanced accuracy score of screening and selection in the testing dataset. The screening and selection process is then repeated 100 times to arrive at an average accuracy (to take into account the randomness in subsampling the judges). The weight is then adjusted to optimize the accuracy in the testing dataset, with 5-folds cross-validation.

| | Balanced Accuracy Score | | Screening Error Rate | |
|---|---|---|---|---|
| Number of judges | Screening (Bottom 25) | Selection (Top 25) | Top 25 | Top 10 |
| 0 (AI) | 58.9% | 54.9% | 11.4% | 5.0% |
| 1 | 73.82% | 70.14% | 3.3% | 0.58% |
| 2 | 77.9% | 74.36% | 1.1% | 0.083% |
| 3 | 79.54% | 76.06% | 0.75% | 0.083% |
| 4 | 80.44% | 78.11% | 0.41% | 0.17% |
| 5 | 81.76% | 78.5% | 0.48% | 0.0% |

**Table 9 Human-AI hybrid model of Section 4.6. 5-fold cross-validation.**

### 5.3. Hybrid model outperforms AI and Human Average Heuristic

Having constructed the hybrid model using the optimal weights as described above, we now compare how well it does with respect to AI and the Human Average Heuristic benchmark (see Section 4.2) [10] for screening and selection. Our metrics of comparison are Balanced Accuracy for both screening and selection and *Screening Error Rate* for screening as described in Section 4.4. Table 9 summarizes the main results. In this table, we consider the screening and selection threshold of Top/Bottom 10 or 25. In the appendix, we show that our results hold for a large range of screening and selection thresholds.

In terms of screening, we can see from Table 9 that the hybrid outperforms AI both in screening accuracy as well as screening error rate. Adding even one human judge in the loop significantly increases screening accuracy by 15% and reduces screening error rate by 8%. However after adding 3 judges, there is hardly any incremental increase in accuracy and beyond 5 judges, the screening error rate drops to almost zero and there is no further gain by adding more judges. In terms of selection, there are also sharp gains in accuracy when

[10] In this benchmark, the hiring company does not need access to a training dataset (previous scores and other relevant judging characteristics) and hence is widely used in practice.

we use a hybrid model over AI. Incorporating just one human intervention can increase accuracy by over 15%.

In Figure 7, is a pictorial depiction of the comparison between the AI-human hybrid and the Human Average Heuristic for the tasks of screening and selection. Here we can see more clearly that the AI-human hybrid model outperforms the Human Average Heuristic in both screening and selection but the gains are particularly higher for selection (top). There is a significant accuracy gain even by incorporating the score of a single human judge however the incremental benefit is small beyond 3 judges.

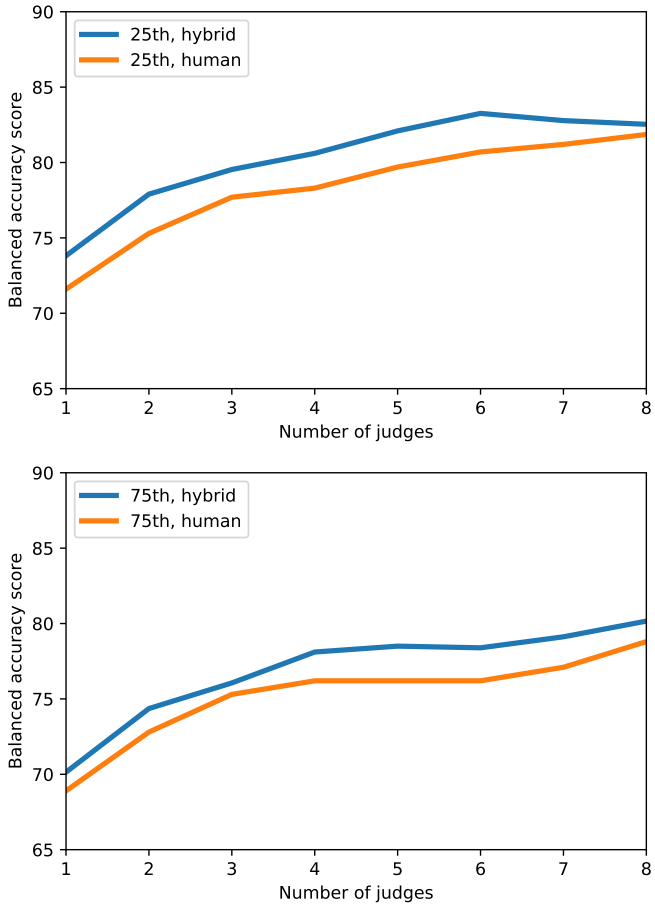



**Figure 7 Comparison of AI-human hybrid and human (Top: Selection, Bottom: Screening)**

### 5.4. Hybrid model is more cost-effective for selection

Previously we showed that a hybrid model improves on screening and selection compared to the AI model. However, this improvement comes about at a cost to the company – the company has to obtain human evaluation. The purpose of using AI for screening and selection is because it is more cost-effective. With hybrid models, there will be additional cost associated with having humans in the loop. In this section, we answer whether the cost of the hybrid model is justified by the increase in accuracy.

We calculate the cost of incremental accuracy (CIA) from augmenting AI model with human judgements in the hybrid model relative to the AI model. From the AI model to the Hybrid model with one human judge per candidate, we incur a cost of $100c$ when there are 100 candidates. We measure the ROI as following:

$$\text{CIA}_{\text{hybrid selection}} = \frac{\text{Additional cost of human judges}}{\Delta\text{Number of candidates correctly selected}}$$

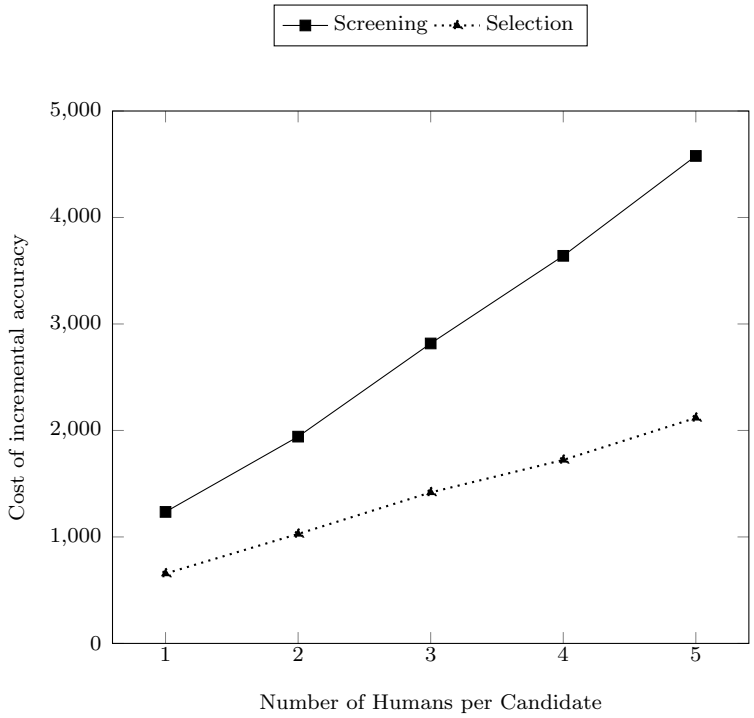
$$\text{CIA}_{\text{hybrid screening}} = \frac{\text{Additional cost of human judges}}{\Delta\text{Number of screened candidates in the top 25th}}$$

We assume $c = 25$, which is equivalent to a cost of \$100 per hour to get an expert human judge to evaluate the videos, each of which is 15 minutes in duration.[11]

In Fig 8, we can see that the use of Hybrid model is more cost-effective for selection rather than screening. Even with a conservative estimate of the cost of human judgement, the hiring company would need to spend over \$1,200 for an additional correct hire, if the company is using the hybrid model for screening. Whereas if the company is using the hybrid model for selection, the cost of an additional correct hire is around \$600. Thus, it

[11] This is a conservative estimate of cost of human judgement. Our judges are experienced industry sales professional whose typical hourly salaries would range from \$100-150

**Figure 8 CIA for using hybrid for screening and selection**




is more cost-effective for firms to use AI models for screening and augment it with human judgement (hybrid) only for selection.

### 5.5. What drives performance of Hybrid

Having demonstrated that the Hybrid model significantly improves accuracy over pure AI and human, we now dig deeper into what drives the improvement in accuracy. A model can improve accuracy in one or both of the two ways: reducing false negatives or false positives. In this scenario, false positive would mean classifying a candidate into one of the buckets (top 25, bottom 25) when the candidate actually belongs to a different bucket. For instance, the model might classify a candidate into top 25 (or in other words, select a candidate) when she/he should not have been selected. Likewise, a false negative in the top 25/bottom 25 buckets would mean not selecting/not rejecting a candidate who should have been selected/rejected. Which of the two is a more serious problem? In the screening stage, which is mostly about choosing "who is the worst", *false positive* is more dangerous

| Number of judges | False Positive Rate | | False Negative Rate | |
|---|---|---|---|---|
| | Bottom 25 | Top 25 | Bottom 25 | Top 25 |
| AI | 0.208 | 0.237 | 0.625 | 0.680 |
| 1 | 0.142 | 0.154 | 0.426 | 0.455 |
| 1 + AI | 0.130 | 0.152 | 0.400 | 0.437 |
| 2 | 0.120 | 0.137 | 0.370 | 0.401 |
| 2 + AI | 0.108 | 0.135 | 0.338 | 0.387 |
| 3 | 0.107 | 0.125 | 0.330 | 0.367 |
| 3 + AI | 0.0996 | 0.124 | 0.305 | 0.355 |
| 4 | 0.108 | 0.119 | 0.330 | 0.355 |
| 4 + AI | 0.0979 | 0.115 | 0.303 | 0.329 |
| 5 | 0.0975 | 0.119 | 0.310 | 0.354 |
| 5 + AI | 0.0961 | 0.116 | 0.294 | 0.323 |

**Table 10 AI-human hybrid: False positive versus False negative rates**

as choosing a bad candidate would lead to higher time cost in the long run (the candidate would be eliminated at a much later stage and result in more loss of managerial time). However when it comes to choosing the top talent, *false negative* is more dangerous. As at that point, the choice set is already small and so the goal is to minimize the chances of losing out on a good candidate.

Now, let us understand how the hybrid leads to improvement in each of these dimensions. Consider the problem of selecting the top candidates, the primary benefit of AI-human hybrid in terms of accuracy improvement stems from lowering the false negative rates as opposed to the false positive rates (see Table 10, Columns 3 and 5). In the table, $k$ + AI denotes a Hybrid model with $k$ human judges, which we compare with the Human Average Heuristic benchmark with $k$ judges. In contrast, for the problem of screening out the bottom candidates, the primary benefit of AI-human hybrid in terms of accuracy

improvement stems from lowering the false negative rates (see Table 10, Columns 2 and 4), and to some extent, a reduction in the false positive rates compared to the Human Average Heuristic benchmark as well. Thus, the Hybrid model improves on both the human and AI benchmarks in the relevant dimensions.

### 5.6. Task-Based Hybrid Models

We see that there is value in adding human judgement to pure AI primarily for selection. We now try to understand at which stage of the interview evaluation does human judgement matter the most? Table 11 shows that most of the benefit of including human in the loop is in the first stage i.e., rapport building and gaining attention. There is not much incremental benefit in selection accuracy as we move to the next stages. This finding is important as it means that we can reduce human involvement time by almost 80% as they need to evaluate only the first 3-4 minutes of an interview instead of the entire conversation.

| Human Judges in Hybrid Model | Stage 1 | Stages 1-2 | Stages 1-3 | Stages 1-4 | Overall |
|---|---|---|---|---|---|
| 1 | 66.8% | 68.8% | 68.9% | 71.1% | 71.5% |
| 2 | 68.2% | 69.9% | 72.2% | 73.0% | 75.0% |
| 3 | 68.3% | 72.6% | 74.1% | 74.3% | 75.7% |

**Table 11 Balanced accuracy score of selection (top-25). Task-based, Sequential Hybrid model where human scores from the first $k$ stages are used.**

## 6. Conclusion and Future Work

Salesforce hiring is costly yet essential for the firm. We propose using AI for screening and selecting candidates based on videos of salesforce interviews. Our first contribution is, we extract persuasion-relevant features from the video data, which consist of 3 modes: text, audio and visual. We develop an AI model based on these features and show that certain

styles of body language, conversational interactivity, voice and linguistic styles, matter for the candidate's success in the sales interview.

Our second contribution is, we construct and test a human-AI hybrid in the domain of salesforce recruitment building on the theoretical literature of human-AI hybrids proposed in past literature. We develop a Bayesian updating framework for incorporating human interventions in the decision-making loop and augmenting AI predictions with human judgements. We assess how the Hybrid model combines the strengths of AI and human judgements, and we find that the Hybrid model outperforms both pure AI and human in terms of accuracy and screening error rate. We show that incorporating even one human in the loop can lead to significant improvements in both selection and screening accuracy though it is more cost efficient for selection. Further, human input is most important for judging the first 2-3 minutes of the interaction. This suggests a cost-effective way to deploy AI in sales hiring—use it exclusively for screening but augment it with human judgement for selection especially for evaluating the early stages.

Our focus has been in the domain of one-one interviews and primarily in salesforce recruitment. It would be interesting to see if the results generalize to a broader set of evaluation scenarios e.g., venture capitalist pitches and student applications.

# 7. Appendix

**Table 12 Most common words for each content class**

| Technology | Business | Greeting | Pricing |
|---|---|---|---|
| laptop | salesforce | come | chargeable |
| security | decision makers | hey | buy |
| mobile | business | morning | prices |
| pipeline | leads | meet | discount |
| data | quoting | weather | monetary |
| storage | customers | sunny | expenditure |
| cloud | sales | week | subscriptions |
| digital | communication | thank you | budgeted |
| transformation | employees | yankee | wallet |
| electronic | problems | holidays | cost |
| screenshot | territories | spring break | expensive |
| licenses | ownership | christmas | pay |
| firewalls | people | business card | dollars |
| emails | agents | traditions | worth |
| video | company | greeted | price |
| licensing | potential client | thanksgiving | money |
| website | insurance | cooking | |
| touchscreen | manager | enjoyed | |
| login | oversee | breakfast | |
| leaderboards | monthly reports | Friday | |
| printout | team | family | |
| technologies | region | | |
| usb | visibility | | |
| kiosk | growth | | |
| android | profitability | | |
| versions | promotion | | |
| charts | revenues | | |
| database | forecasting | | |
| mobility | goal | | |
| phone | supervisors | | |
| app | premium | | |
| tablet | competitive | | |
| dashboard | due diligence | | |
| computer | reps | | |
| tool | trajectory | | |
| | elaborating | | |
| | underperforming | | |
| | policy holder | | |

## 7.1. Sensitivity Analysis

In this section, we show how the results in section XX hold for a range of screening and selection thresholds. Fig 10 shows the impact of

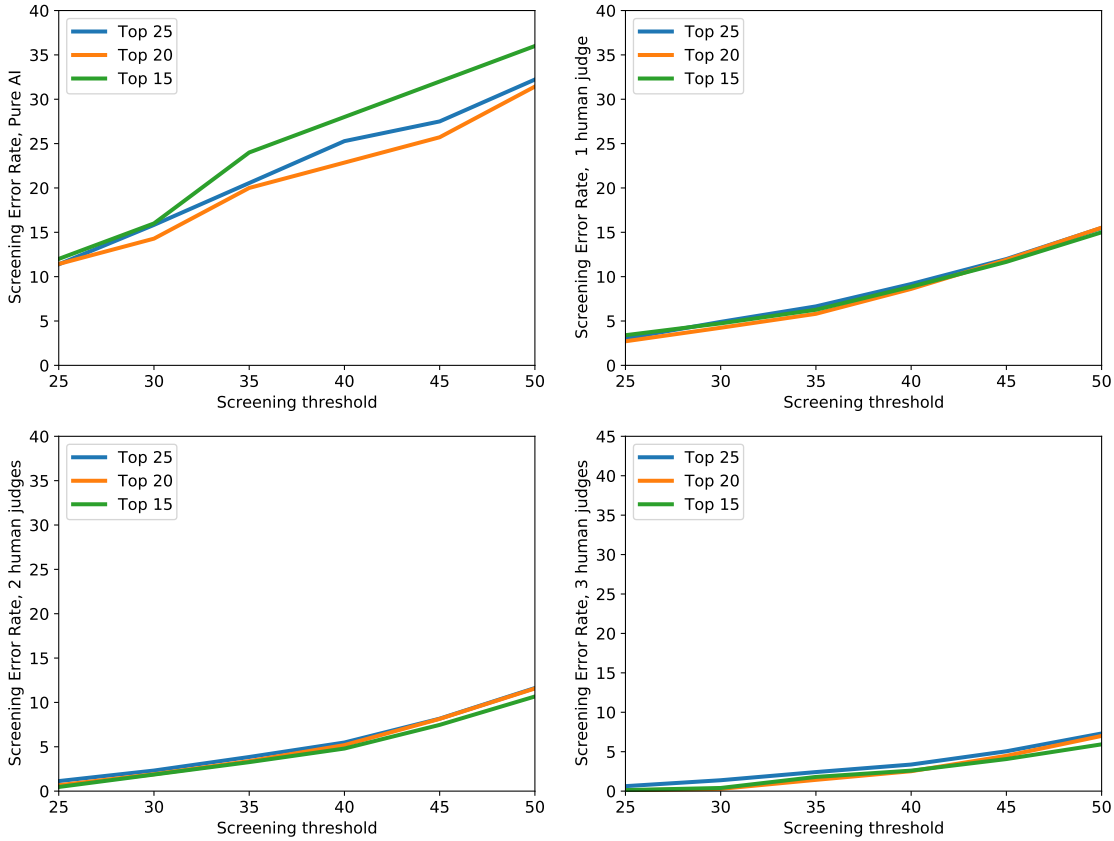



**Figure 9 AI-human Hybrid model significantly reduces the post-screening error rates. However if we are screening out the bottom 25, there is little incremental gain in using more human judges. The more candidates we are screening out, the greater the incremental benefit in using more human judges.**

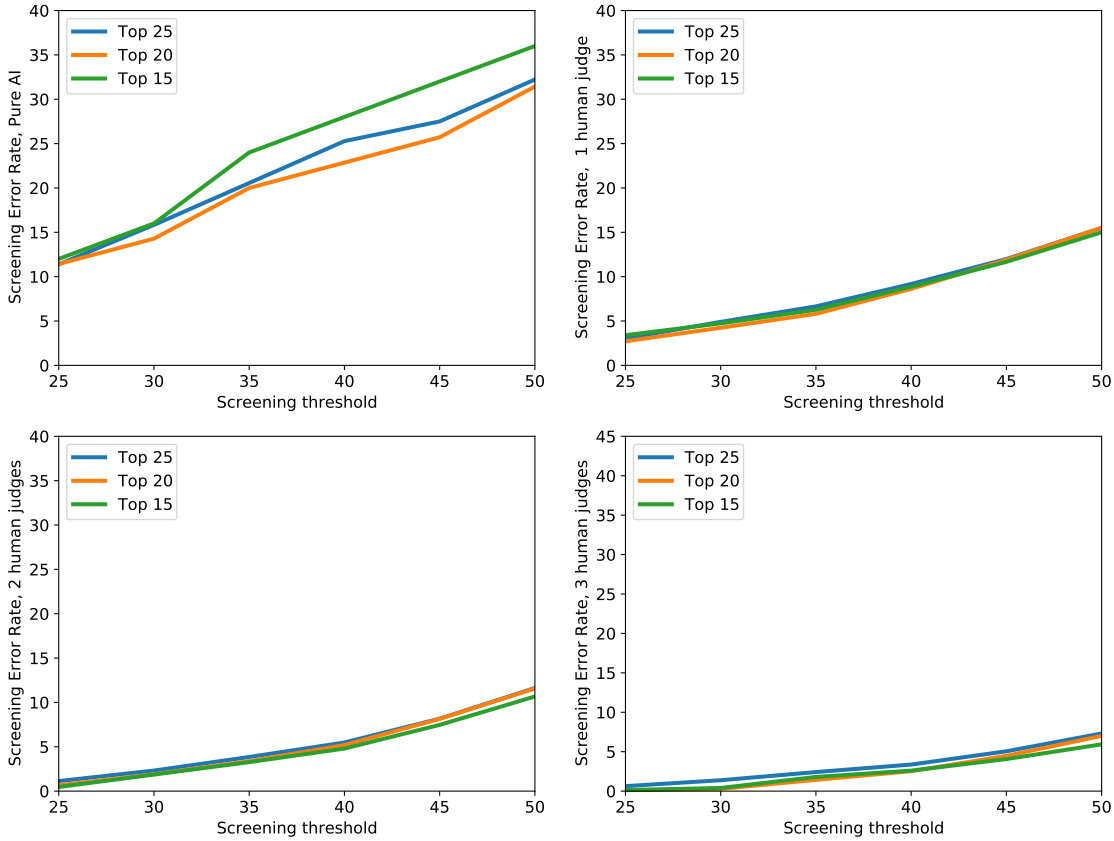



**Figure 10 AI-human Hybrid model significantly reduces the post-screening error rates. However if we are screening out the bottom 25, there is little incremental gain in using more human judges. The more candidates we are screening out, the greater the incremental benefit in using more human judges.**

## References

Agrawal A, Gans JS, Goldfarb A (????) Exploring the Impact of Artificial Intelligence: Prediction versus Judgment 15.

Altmann A, Toloşi L, Sander O, Lengauer T (2010) Permutation importance: a corrected feature importance measure. *Bioinformatics* 26(10):1340–1347.

Autor DH, Scarborough D (2008) Does job testing harm minority workers? evidence from retail establishments. *The Quarterly Journal of Economics* 123(1):219–277.

Bansal G, Nushi B, Kamar E, Weld DS, Lasecki WS, Horvitz E (2019) Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 2429–2437.

Bendick Jr M, Nunes AP (2013) Developing the research basis for controlling bias in hiring. *Journal of Social Issues* 68:238–262.

Bertrand M, Mullainathan S (2004) Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American economic review* 94(4):991–1013.

Campion MA, Palmer DK, Campion JE (1997) A review of structure in the selection interview. *Personnel psychology* 50(3):655–702.

Cao Z, Hidalgo Martinez G, Simon T, Wei S, Sheikh YA (2019) Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* .

Cao Z, Simon T, Wei SE, Sheikh Y (2017) Realtime multi-person 2d pose estimation using part affinity fields. *CVPR*.

Chakraborty I, Chiong K, Sudhir K (2021) Reducing bias in interviewing using balanced panels and structured interviews .

Chapman DS, Webster J (2003) The use of technologies in the recruiting, screening, and selection processes for job candidates. *International journal of selection and assessment* 11(2-3):113–120.

Churchill GA, Collins RH, Strang WA (1975) Should retail salespersons be similar to their customers. *Journal of Retailing* 51(3):29.

Cialdini RB, Cialdini RB (2007) *Influence: The psychology of persuasion*, volume 55 (Collins New York).

Cowgill B (2018) Bias and productivity in humans and algorithms: Theory and evidence from resume screening. *Columbia Business School, Columbia University* 29.

Dane E, Pratt MG (2007) Exploring intuition and its role in managerial decision making. *Academy of management review* 32(1):33–54.

Doshi-Velez F, Kim B (2017) Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* .

Eisenhardt KM (1989) Making fast strategic decisions in high-velocity environments. *Academy of Management journal* 32(3):543–576.

Evans FB (1963) Selling as a dyadic relationship–a new approach. *American Behavioral Scientist* 6(9):76–79.

Farr JL, Tippins NT (2013) *Handbook of employee selection* (Routledge).

Frazier GL, Summers JO (1984) Interfirm influence strategies and their application within distribution channels. *Journal of Marketing* 48(3):43–55.

Hartigan J, Wigdor A (1989) Fairness in employment testing. *Science* 245(4913):14–14.

Hoffman M, Kahn LB, Li D (2018) Discretion in hiring. *The Quarterly Journal of Economics* 133(2):765–800.

Holstein K, Aleven V, Rummel N (2020) A conceptual framework for human–ai hybrid adaptivity in education. *International Conference on Artificial Intelligence in Education*, 240–254 (Springer).

Huang F, Cappelli P (2006) Employee screening: theory and evidence.

Kamar E (2016) Directions in hybrid intelligence: Complementing ai systems with human intelligence. *IJCAI*, 4070–4073.

Longpre L, Durmus E, Cardie C (2019) Persuasion of the undecided: Language vs. the listener. *Proceedings of the 6th Workshop on Argument Mining*, 167–176.

Luo X, Qin MS, Fang Z, Qu Z (2021) Artificial intelligence coaches for sales agents: Caveats and solutions. *Journal of Marketing* 85(2):14–32.

Manzoor E, Chen GH, Lee D, Smith MD (2020) Influence via ethos: On the persuasive power of reputation in deliberation online. *arXiv preprint arXiv:2006.00707* .

Ning H, Yin R, Ullah A, Shi F (2021) A survey on hybrid human-artificial intelligence for autonomous driving. *IEEE Transactions on Intelligent Transportation Systems* .

Pennebaker JW, Francis ME, Booth RJ (2001) Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates* 71(2001):2001.

Perlow LA, Okhuysen GA, Repenning NP (2002) The speed trap: Exploring the relationship between decision making and temporal context. *Academy of Management journal* 45(5):931–955.

Pulakos ED, Schmitt N (1995) Experience-based and situational interview questions: Studies of validity. *Personnel Psychology* 48(2):289–308.

Raghavan M, Barocas S, Kleinberg J, Levy K (2020) Mitigating bias in algorithmic hiring: Evaluating claims and practices. *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 469–481.

Rivera LA (2012) Hiring as cultural matching: The case of elite professional service firms. *American sociological review* 77(6):999–1022.

Sheth J (1976) Buyer-seller interaction: A conceptual framework. *Advances in Consumer Rasearch* 3(3B):382—386.

Shmueli-Scheuer M, Herzig J, Konopnicki D, Sandbank T (2019) Detecting persuasive arguments based on author-reader personality traits and their interaction. *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*, 211–215.

Shrestha YR, Ben-Menahem SM, Von Krogh G (2019) Organizational decision-making structures in the age of artificial intelligence. *California Management Review* 61(4):66–83.

Simon T, Joo H, Matthews I, Sheikh Y (2017) Hand keypoint detection in single images using multiview bootstrapping. *CVPR*.

Spiro RL, Weitz BA (1990) Adaptive selling: Conceptualization, measurement, and nomological validity. *Journal of marketing Research* 27(1):61–69.

Tambe P, Cappelli P, Yakubovich V (2019) Artificial intelligence in human resources management: Challenges and a path forward. *California Management Review* 61(4):15–42.

Vellido A (2020) The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural computing and applications* 32(24):18069–18083.

Vincent VU (2021) Integrating intuition and artificial intelligence in organizational decision-making. *Business Horizons* 64(4):425–438.

Wager S, Hastie T, Efron B (2014) Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *The Journal of Machine Learning Research* 15(1):1625–1651.

Walton D, Reed C, Macagno F (2008) *Argumentation schemes* (Cambridge University Press).

Wei SE, Ramakrishna V, Kanade T, Sheikh Y (2016) Convolutional pose machines. *CVPR*.

Yakubovich V, Lup D (2006) Stages of the recruitment process and the referrer's performance effect. *Organization science* 17(6):710–723.

Yukl G, Seifert CF, Chavez C (2008) Validation of the extended influence behavior questionnaire. *The Leadership Quarterly* 19(5):609–621.