Statistical Methods for Assessing Candidate Employability: A Survey-Informed, Simulation-Based Comparison of Supervised Classifiers for Hiring Decision Support

Tanmay Hinge
Department of Computer Science
California State University, Chico
hingetanmay@gmail.com

Master's capstone research, completed May 2024

Abstract

Hiring teams rank candidates against an implicit model of employability that is rarely written down. This work makes that model explicit and testable. Two survey instruments (candidates, $n = 12$; hiring managers and team leads, $n = 10$) elicited which attributes practitioners weight in practice. From the responses I specified a six-feature representation of a candidate (GPA, years of experience, coding assessment score, Leetcode progress, relevant skills score, and resume ATS score), generated a 1,000-observation synthetic dataset under normality assumptions, and benchmarked five supervised classifiers on it, with logistic regression as the interpretable reference model. Principal component analysis combined with hyperparameter tuning raised SVM accuracy from 89.00% to 96.67% and ROC AUC from 0.9344 to 0.9950. Both survey groups and the fitted models rank GPA as the dominant feature, closing the loop between elicitation and estimation. The central contribution is the elicitation-to-benchmark framework; the accuracy figures matter less. The results support a design argument for high-stakes selection, where interpretability and fairness must be built in from the start because post-hoc additions arrive too late to shape the model.

1 Introduction

Candidate screening is a ranking problem carried out under constraint. There are many applicants, little time, noisy signals, and reviewer judgments that are inconsistent between raters and within the same rater over time. Decisions with material consequences for individuals are therefore made using criteria that are never stated. Three gaps follow.

1. Which attributes do practitioners actually weight, as opposed to those they report valuing in job descriptions?
2. Once employability is operationalised as a binary outcome over measurable features, can it be predicted well enough to be useful, and does the prediction go beyond mere plausibility?
3. What accuracy is given up by insisting on an interpretable model, and is that cost worth paying where a candidate may be owed an explanation?

This paper addresses all three with a pipeline that runs from practitioner elicitation to feature specification, simulation with known ground truth, multi-model benchmarking, and dimensionality reduction with tuning. The synthetic setting is a deliberate choice. Because the data-generating process is known in full, estimated coefficients can be checked against it, and the pipeline can be validated as an instrument before it ever touches real applicant data.

Contributions. The work contributes (i) a reproducible pipeline from two-sided survey elicitation through simulation, benchmarking, and tuning; (ii) an instrument design that surfaces divergence between what candidates emphasise and what reviewers weight, in place of an assumed shared view of merit; and (iii) quantified evidence on the trade-off between interpretability and accuracy in a selection setting, together with the argument that fairness constraints belong in the design phase of such systems.

2 Related Work

Employability prediction has an established empirical literature. A 2023 systematic review of machine learning models for predicting the employability of management students, covering work from 2006 to 2022, reports accuracies up to 95% for logistic regression and support vector machines, with ensemble methods reaching 93% [1]. ElSharkawy et al. benchmark five binary classifiers on survey data from IT graduates and hiring managers in Egypt, with decision trees strongest on their data [2]. Both lines of work fit models to collected records and take the feature set as given. This study differs in construction. The feature set is itself elicited from both sides of the hiring interaction, and the data are simulated from a known generative process, so recovery of the encoded structure can be checked directly.

3 Survey Elicitation and Instrument Design

I ran two instruments so that both sides of the screening interaction were represented. The candidate survey (12 responses) recorded self-reported preparation, credentials, and experience. The employer survey (10 responses, drawn from hiring managers and team leads) recorded the attributes and thresholds reviewers apply in practice.

The design is diagnostic. Divergence between what candidates emphasise and what reviewers weight is itself a result, and it grounds the feature set in elicited evidence, replacing a choice that would otherwise rest on intuition. The responses bear this out. Both groups put GPA first, with 43.5% of the importance weight among managers and 41.7% among candidates, and they part ways below it. Managers assign recommender scores 21.7% of their weight where candidates assign 8.3%, and candidates assign coding assessments 37.5% where managers assign 8.7%. Candidates, in other words, over-weight the signal they can grind and under-weight the one that requires a relationship. The manager pool included hiring managers and senior engineers at large technology firms and startups.

The elicitation fixed a six-feature representation of a candidate, with GPA, years of experience, coding assessment score, Leetcode progress, relevant skills score, and resume ATS score as the features and employability encoded as a binary target.

4 Synthetic Data Generation

The six-feature specification parameterised a synthetic dataset of 1,000 observations with normally distributed features. I chose simulation deliberately, and the small survey samples were only part of the reason. The choice buys three things.

- **Known ground truth.** The data-generating process is fixed by construction, so estimated coefficients can be compared directly against the true generative parameters. A pipeline that fails to recover known structure cannot be trusted on unknown structure.

- **Experimental control.** Class balance and feature correlation become experimental factors that can be varied, so the sensitivity of each classifier to these conditions is measurable.
- **No exposure of applicant data.** No real candidate records are used, which avoids the consent problems that make published hiring datasets scarce.

The generator was implemented in Python with NumPy; each observation carries the six feature values, a composite score computed from them, and the binary label. The data were split 80/20 into training and test sets. The cost of the simulation choice is external validity, discussed in Section 8.

5 Reference Model: Logistic Regression

The logistic model estimates the probability that a candidate is employable ($Y = 1$) given a feature vector $\mathbf{x} = (x_1, \ldots, x_6)$,

$$P(Y = 1 \mid \mathbf{x}) \;=\; \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_6 x_6)}}. \tag{1}$$

Logistic regression serves as the reference model because each term in Equation 1 can be explained to a non-technical audience. The intercept β_0 gives the log-odds of employability when every feature is zero, fixing the baseline against which the remaining terms are read. Each coefficient β_j gives the change in log-odds for a one-unit change in x_j holding the other features constant, and exponentiating it returns an odds ratio, the form a hiring panel can reason about directly. The linear combination $\boldsymbol{\beta}^\top \mathbf{x}$ is unbounded and the exponential maps it onto $(0, \infty)$; the 1 in the denominator compresses that range into the unit interval, so the output is a calibrated probability, and it can be read as one.

Table 1 reports the fitted coefficients. GPA dominates, in agreement with both survey groups, and resume ATS score contributes least. The raw odds ratios must be read against feature scale. A one-unit change in GPA spans half its observed range, so the headline ratio of roughly 1256 reflects the narrowness of that scale as much as the strength of the effect, and cross-feature comparison is safer on the coefficient ranking.

Table 1: Fitted logistic regression coefficients and odds ratios.

Feature	Coefficient	Odds ratio
GPA	7.136	1256.36
Years of experience	0.822	2.28
Coding assessment score	0.332	1.39
Leetcode progress	0.176	1.19
Relevant skills score	0.113	1.12
Resume ATS score	0.062	1.06

6 Experiments and Results

Five classifiers were compared on accuracy and ROC AUC, chosen to span the interpretability spectrum. Logistic regression and decision trees sit at the readable end; random forests, gradient boosting, and a support vector machine (SVM) trade that readability for flexibility.

6.1 Baseline

Table 2 reports the full baseline comparison. Logistic regression achieved the strongest performance of the five, at 98.33% accuracy and 0.9988 ROC AUC, and the decision tree was weakest at 85.33% and 0.8154. The headline result should be stated carefully. It is what would be expected when the data-generating process is close to linear in the log-odds, so it says more about the design of the simulation than about the merits of the estimator. Its value is as a control, confirming that the pipeline recovers known structure.

Table 2: Baseline performance of the five classifiers.

Model	Accuracy	Precision	Recall	F1	ROC AUC
Logistic regression	0.9833	0.9753	0.9634	0.9693	0.9988
Decision tree	0.8533	0.7317	0.7317	0.7317	0.8154
Random forest	0.9400	0.9324	0.8415	0.8846	0.9833
SVM	0.8900	0.8025	0.7927	0.7975	0.9344
Gradient boosting	0.9333	0.9306	0.8171	0.8701	0.9811

6.2 Dimensionality Reduction and Tuning

I applied principal component analysis (PCA) to the SVM and the random forest to address correlation among the six features and to test whether performance survived once the representation was compressed. Combined with hyperparameter tuning, the effect was substantial for both, and largest for the SVM (Table 3). SVM accuracy rose from 89.00% to 96.67% (a gain of 7.67 percentage points) and ROC AUC from 0.9344 to 0.9950; the random forest reached 95% accuracy and 0.9986 AUC.

Table 3: Performance after PCA and hyperparameter tuning.

Model	Accuracy	Precision	Recall	ROC AUC
SVM + PCA	0.9667	0.9459	0.9211	0.9950
Random forest + PCA	0.9500	0.9400	0.8690	0.9986

Of the SVM's two gains, the AUC gain is more informative, since it holds across all decision thresholds and is insensitive to the class balance on which accuracy silently depends. The most defensible interpretation is that the feature space contained redundancy and the margin was fitted partly to noise; decorrelated components gave the kernel a cleaner geometry to separate.

7 Discussion

The framework carries the weight of the contribution. A tuned SVM reaching 96.67% accuracy on favourable synthetic data is unsurprising; a pipeline that elicits practitioner criteria, encodes them as a generative process, and then verifies that standard estimators recover the encoded structure is an instrument that transfers. The same design applies wherever a selection judgement is implicit and contested, and hiring is one instance of a broader class.

The elicitation loop also closes. GPA carries the largest weight in the manager survey, the candidate survey, and the fitted coefficient ranking, so the criterion practitioners state is the criterion the encoded process rewards, and the pipeline detects that agreement. The results also sharpen the interpretability question posed in the introduction. On this data, the interpretable reference model gave up little at baseline, and the flexible model overtook it only after decorrelation and tuning. In a setting where a rejected candidate may be owed an explanation, that ordering matters. Interpretability and fairness function here as design requirements, and treating them as post-hoc additions concedes the model's structure before the audit begins.

8 Limitations

These are stated plainly, because they define what the work does not yet establish.

(i) **Survey power.** At $n = 12$ and $n = 10$, the instruments support qualitative elicitation and the justification of a feature set, and they do not support inference.

(ii) **Synthetic data.** Independent normal features are an idealisation. Real applicant attributes are skewed, censored, correlated, and partly categorical, so the reported accuracy is an upper bound obtained under favourable conditions.

(iii) **Construct validity.** The label "employable" is a binary stand-in for a contested, context-dependent judgement, and it inherits whatever the elicitation encoded.

(iv) **Fairness not audited.** No disparate-impact or subgroup-calibration analysis was carried out, a gap that matters most precisely because the tuned model performs well.

9 Directions for Future Work

Validation on real data. Replicate the pipeline on a consented applicant-tracking dataset, treating the simulation as a pre-registered prior and quantifying the gap between simulated and real performance.

Fairness-constrained learning. Estimate subgroup performance and impose equalised-odds [3] or calibration constraints, so that the accuracy and fairness frontier is characterised empirically and does not have to be asserted.

Causal framing. Screening intervenes in the world, and a passive-prediction framing misses this. Outcomes are observed only for those who are hired, which places the problem in the selective-labels regime [4] and suggests a potential-outcomes reformulation.

Explanation as a deliverable. Evaluate coefficient-based, SHAP-based [5], and counterfactual [6] explanations against a concrete criterion, namely what a rejected candidate can act upon.

References

[1] Crasta, L. and Shailashri, V. T. A systematic review on the employability prediction model for the management students. *International Journal of Case Studies in Business, IT, and Education*, 2023.

[2] ElSharkawy, G., Helmy, Y., and Yehia, E. Employability prediction of information technology graduates using machine learning algorithms. *International Journal of Advanced Computer Science and Applications*, 2022.

[3] Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems 29*, 2016.

[4] Lakkaraju, H., Kleinberg, J., Leskovec, J., Ludwig, J., and Mullainathan, S. The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017.

[5] Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*, 2017.

[6] Wachter, S., Mittelstadt, B., and Russell, C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2), 2017.