Human vs. Machine:

Evaluating AI Decisions in Technical Resume Review

Isabella Castillo, Kaylee Kim, Rithika Cheela

University of Texas at Austin

PSY 371F: Foundations of Psychological Data Science II

Professor Franco Pestilli

April 28th, 2025

Abstract

As AI systems increasingly assist in resume screening for technical roles, concerns about transparency and algorithmic bias have become more prominent. This study investigates how AI evaluates candidate resumes and how its assessments align with human recruiter decisions. Using a dataset of over 1,000 synthetically generated resumes tailored for technology-related positions—each scored by an AI model and reviewed by recruiters—we address three core research questions: (1) What are the most influential resume features in predicting recruiter hiring decisions? (2) To what extent do AI-generated scores align with recruiter outcomes? (3) Are there discrepancies that indicate potential bias in AI evaluations?

We hypothesized that (1) skills and experience would emerge as the most predictive features, (2) AI and recruiter decisions would align in over 50% of cases, and (3) AI scoring may overemphasize quantitative resume attributes. However, our findings show that the AI Score alone was the dominant factor, with experience and education having no measurable impact. At a threshold of 65, the model achieved 100% alignment with recruiter decisions, raising concerns about overfitting to an opaque metric.

Further analysis revealed a statistically significant difference in AI scores between hired and rejected candidates, yet there was no evidence of educational bias in scoring. These insights suggest that the model may replicate the logic of the external AI scoring system rather than evaluating resumes independently. This highlights the need for greater transparency and caution when integrating AI tools into hiring pipelines.

Keywords: AI Resume Review, AI Hiring, Technology, Biases

Introduction

Imagine you are a recent graduate with a degree in the technological field. You will spend countless hours tailoring your resume to enter the competitive job market. You will highlight your skills, refine project descriptions, and carefully submit your resume to stand out among thousands of applications. But you get a rejection email in the middle of the night, within a few hours of your submission. This is even before a human sees your application; just by the artificial intelligence (AI) technology's filter system, you automatically get rejected with zero transparency. The rejection email does not tell you why you got rejected and what the AI saw or overlooked.

As companies increasingly rely on AI to filter resumes for technical roles, there is growing concern that these systems may be optimizing for the wrong traits—prioritizing surface-level indicators like years of experience or keyword frequency over more meaningful but less quantifiable qualities (Raghavan et al., 2020). Prior research has shown that even well-intentioned algorithms can internalize subtle biases from the data they are trained on, resulting in skewed decision-making (O'Neil, 2016; Binns et al., 2018).

This study aims to uncover what characteristics AI screening tools favor when evaluating resumes. Using a dataset of over 1,000 synthetically generated resumes tailored for technical roles, we analyzed how an AI scoring model weights different features such as skills, education, experience, and project count.

We hypothesize that the AI model will emphasize quantifiable traits–particularly years of experience and the number of projects–while undervaluing qualitative or context-rich attributes. By identifying which features are disproportionately favored, this research sheds light on potential biases in the evaluation process.

Data Analysis Plan

This analysis aims to explore three key areas: (1) identifying the two most influential resume features in predicting recruiter hiring decisions, (2) assessing the alignment between AI-generated scores and recruiter decisions, and (3) examining potential biases in AI evaluations.

The dataset used for this study is sourced from Kaggle, a platform offering publicly available datasets for educational purposes. It consists of 1,000 synthetically generated resumes tailored for technical roles. The dataset includes 11 columns featuring information such as candidate skills, years of experience, highest education level, relevant industry certifications, recruiter decisions, number of projects completed, and AI-generated resume ranking scores.

During the data cleaning phase, several preprocessing steps will be taken to ensure consistency, reduce noise, and prepare the data for analysis. First, non-predictive identifiers such as resume ID and candidate name will be removed to mirror AI resume screening practices that aim to reduce bias by anonymizing applicant information. Second, categorical outcome variables will be numerically encoded: "hired" candidates will be labeled as 1, and "rejected" candidates as 0, enabling binary classification. Education levels will also be encoded on an ordinal scale to facilitate analysis—for example, 'b.sc' will be mapped to 1, 'b.tech' or 'BTech' to 2, 'mba' to 3, 'm.tech' to 4, and 'PhD' to 5. Finally, missing values will be replaced with 'None' to improve interpretability and maintain data integrity.

Once cleaned and encoded, the data will be organized into a structured data frame for analysis. To begin exploratory data analysis, the team will employ summary statistics and visualization tools such as correlation pair plots and heatmaps. Pair plots will be used to examine relationships between numeric and encoded variables—including experience, education, project count, and AI score—while heatmaps will help identify multicollinearity among predictors.

Building on these initial insights, the analysis will proceed with statistical and machine-learning methods. Logistic regression will estimate the impact and directionality of features like AI score, education, and years of experience on hiring decisions. To further evaluate feature importance, decision tree classifiers will help determine which variables most strongly influence AI scoring. Additionally, inferential statistical tests—including t-tests and ANOVA—will be conducted to evaluate the following hypotheses: (1) quantitative features (e.g., experience, project count) are more predictive than qualitative ones, and (2) AI scores may be biased based on education level. T-tests will assess differences in AI scores between hired and rejected candidates, while ANOVA will evaluate variance in scores across different education levels.

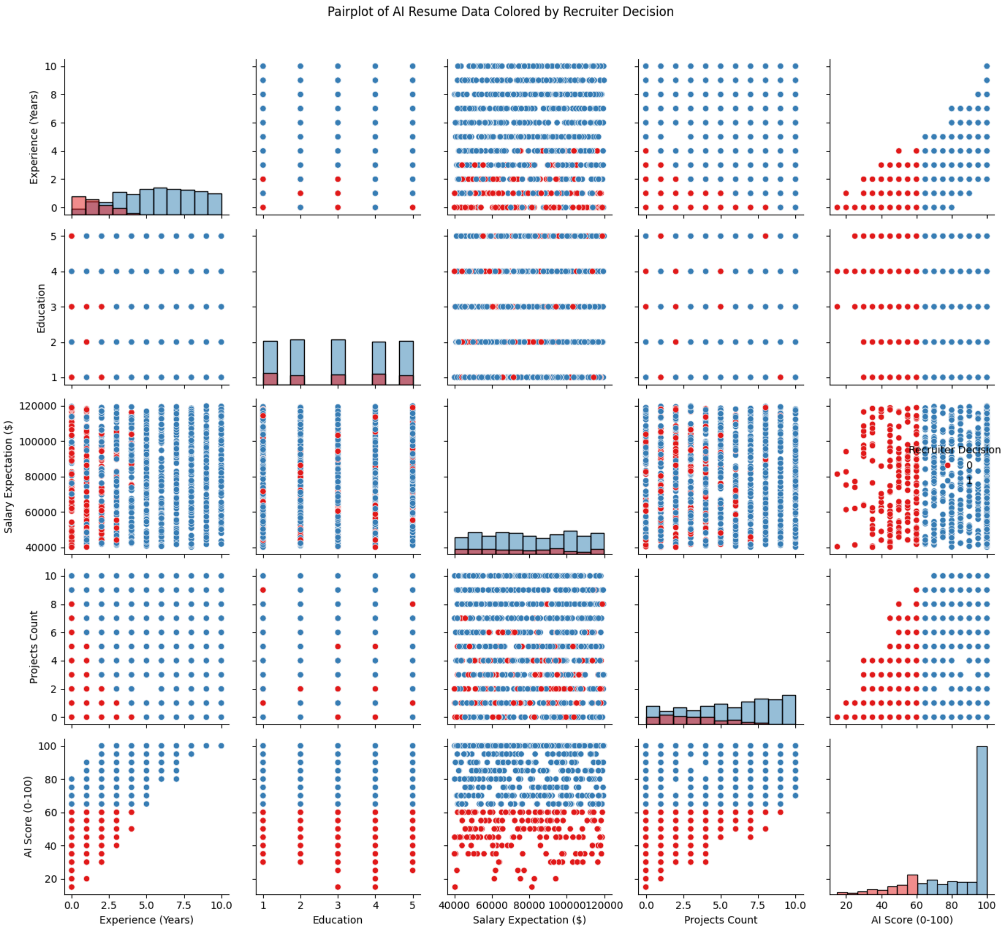
Model performance will be assessed using precision, Cohen's d, and label shuffling metrics. Cohen's d will quantify the effect size of score differences, and label shuffling will test for potential overfitting by examining whether models capture generalizable patterns or memorize data-specific noise. All analyses will be conducted using Python, with the support of libraries including pandas, numpy, seaborn, matplotlib, plotly, sci-kit-learn, stats models, scipy, and graphviz.

In summary, this study leverages a combination of data preprocessing, visualization, statistical testing, and machine learning to systematically evaluate the fairness and accuracy of AI-driven resume screening.

Results

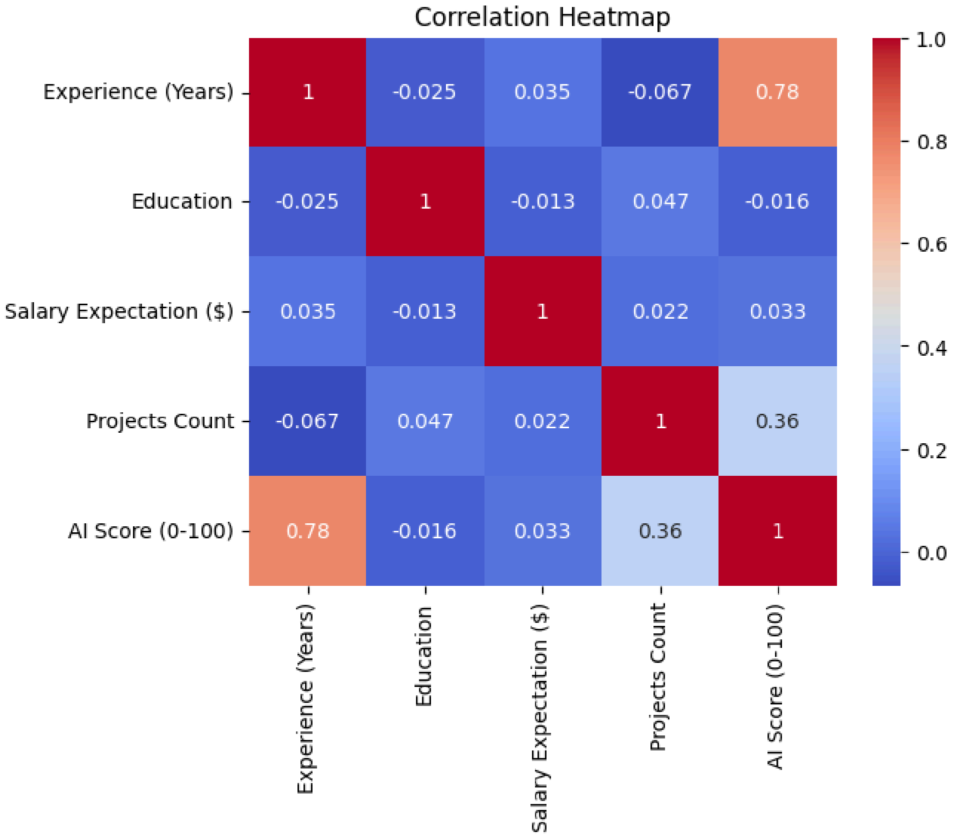
To investigate the top two resume features that predict recruiter hiring decisions, assess whether AI scores align with human recruiter decisions, and oversee potential biases using AI resume screening technologies, we have conducted data analysis as planned in the previous section. The dataset included 1,000 synthetically generated job applications with key features such as project counts, certifications, skills, AI scores, recruiter decisions, etc. Throughout the dataset analysis, the team will present evidence that might or might not support the hypothesis presented.

Figure 1: Pairplot of AI Resume Features Colored by Recruiter Hiring Decision



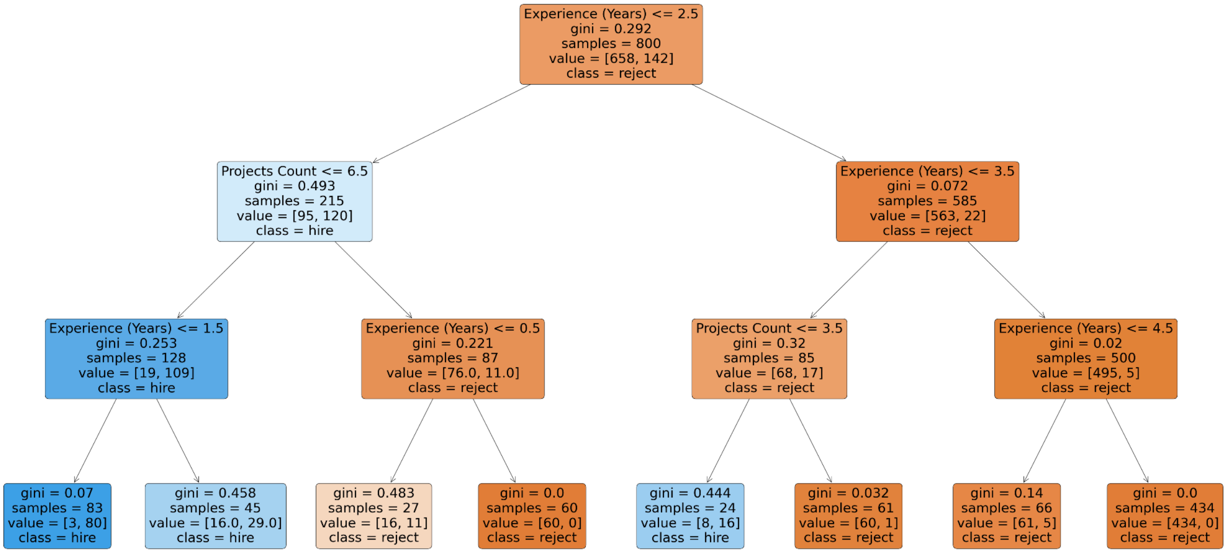
The figure above shows some distinct patterns in color, with red mapped to 0 for the recruiter's decision, which corresponds to being 'rejected,' and blue as 1, corresponding to 'hired.' When taking a closer look at the column with AI scores, there seems to be a distinct color change in scores around 60. Also, it is evident that there are variables that do not get as much distinctiveness, such as education and years of experience.

Figure 2: Correlation Heatmap of Candidate Features for Hiring Analysis



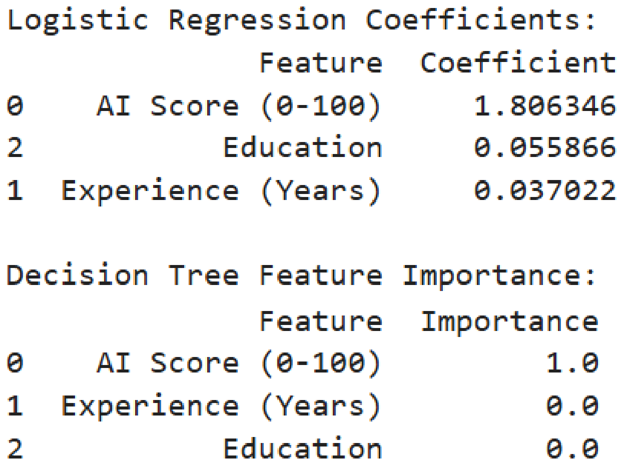
The figure above presents the numeric correlation values of the numeric variables. The variables with the highest correlations are experienced and AI score, at 0.78, and the next highest is between the projects count and AI score, at 0.36. On the other hand, we can tell that education has the lowest correlation when assessed with the AI score variable, which allows further exploration by using different analytic techniques.

Figure 3: Decision Tree Visualization Based on Experience and Project Count



This figure presents a hiring prediction model trained without incorporating the AI Score, enabling a clearer understanding of how project count and years of experience independently influence recruiter decisions. The decision tree reveals that experience is the most impactful variable, as evidenced by the root node's initial split at 2.5 years—candidates with less experience are predominantly classified as "reject." As the tree progresses, project count emerges as another key determinant, appearing in several subsequent splits across both low- and high-experience candidate groups. The model indicates that individuals with limited experience and few projects are typically rejected, while those with moderately higher experience and greater project involvement are more likely to be hired. Overall, the visualization underscores that experience and project count are the strongest predictors of hiring outcomes when AI-based scoring is excluded.

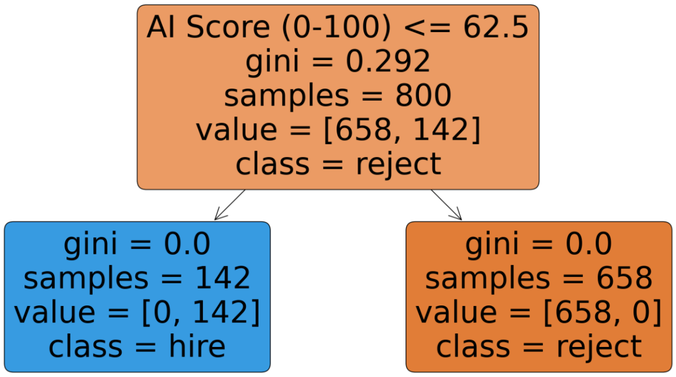
Figure 4. Comparison of Feature Importance in Logistic Regression vs. Decision Tree Models



```
Logistic Regression Coefficients:
              Feature  Coefficient
0    AI Score (0-100)     1.806346
2           Education     0.055866
1  Experience (Years)     0.037022

Decision Tree Feature Importance:
              Feature  Importance
0    AI Score (0-100)         1.0
1  Experience (Years)         0.0
2           Education         0.0
```

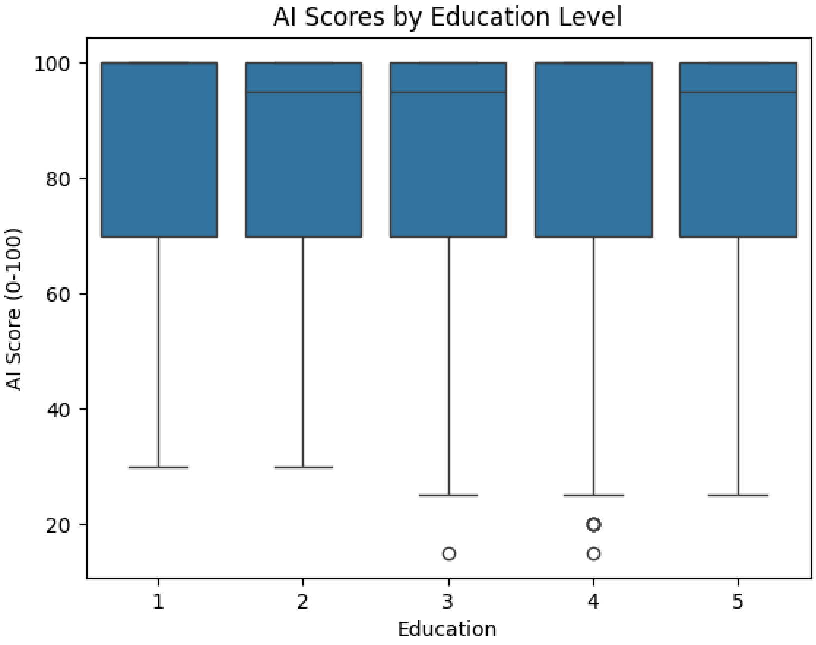
In the figure, the AI Score (ranging from 0 to 100) has an importance value of 1.0, indicating that it is the sole feature the decision tree uses to make hiring decisions. At the same time, experience and education have importance values of 0.0, meaning the model completely ignores them. This suggests that once the AI Score is available, the decision tree relies exclusively on that metric and finds no additional predictive value in traditional resume attributes such as years of experience or educational background. In contrast, the logistic regression model assigns small but non-zero coefficients to education and experience, indicating that it still considers them slightly useful. However, the AI Score remains the dominant factor across both models. Together, these findings emphasize the overwhelming influence of the AI Score when it is included, particularly in tree-based models.

Figure 5: Single-Split Decision Tree Based Solely on AI Score for Hiring Classification



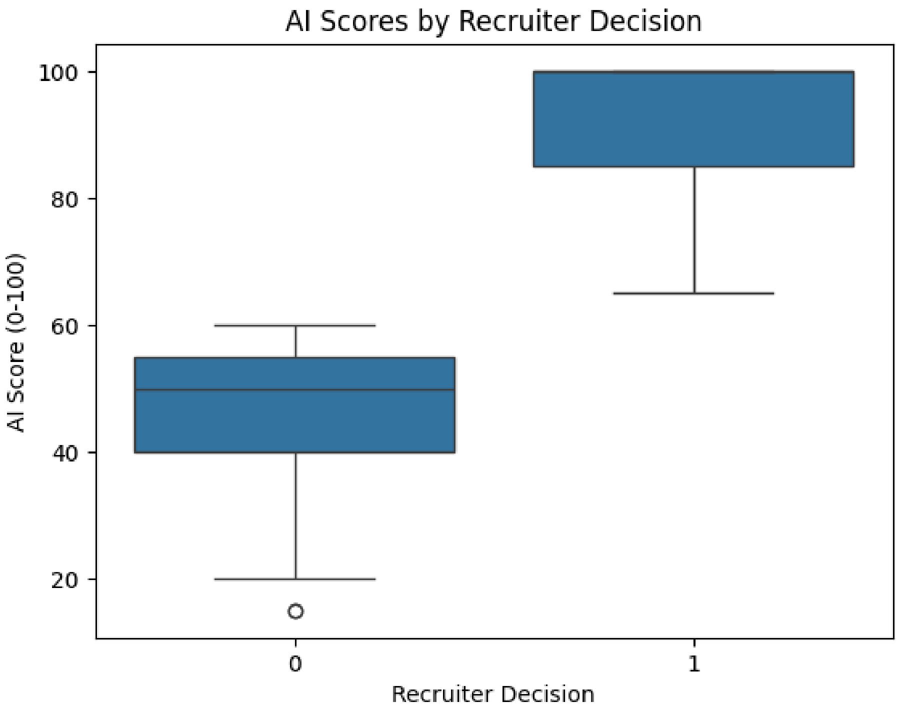
This decision tree shows that AI Score alone fully determines hiring decisions. Candidates with an AI Score of 62.5 or lower are always hired, while those above 62.5 are always rejected, with perfect classification in both cases. The model ignores other features like experience and education, relying solely on this score to make all decisions.

Figure 6: Boxplot of AI Scores by Education Level Demonstrating Score Consistency



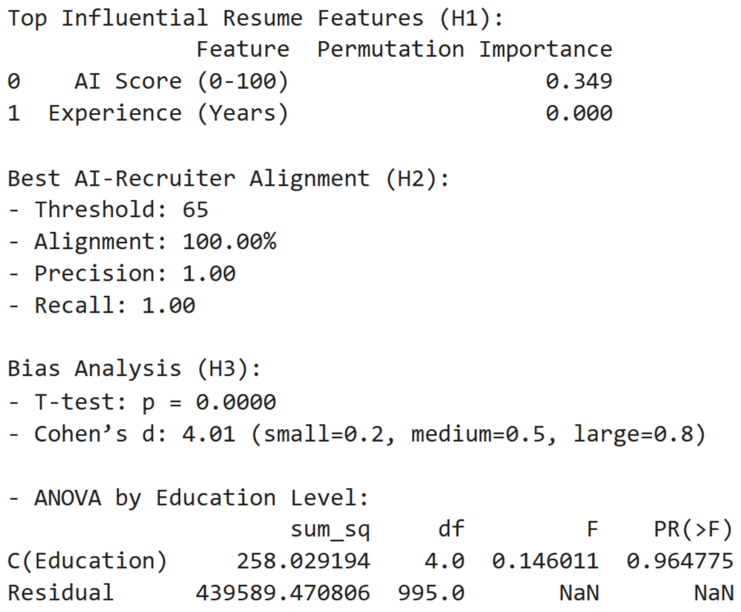
The figure shows that the AI assigned consistently high scores across all education levels (1–5), with median scores close to 100 and similar distributions regardless of degree. There is no clear trend or preference for candidates with higher educational qualifications like PhDs over those with lower degrees such as a BSc. This observation is supported by the ANOVA result (p = 0.96), indicating no statistically significant difference in AI scores between education levels. Overall, this suggests that the AI scoring system is education-neutral, which helps alleviate concerns about educational bias in the algorithm's evaluations.

Figure 7: Boxplot of AI Scores by Recruiter Decision Outcome Demonstrating Strong Alignment



The figure demonstrates a clear separation in AI scores between candidates who were hired (1) and those who were rejected (0), with hired candidates receiving significantly higher scores on average and minimal overlap between the two groups. This pattern is statistically supported by a t-test result of p = 0.0000, indicating a highly significant difference in AI scores based on recruiter decisions. The strong distinction suggests that the AI score is a powerful predictor of recruiter behavior, potentially directly or indirectly influencing hiring outcomes. However, this raises a critical concern: the model may rely almost exclusively on the AI score, potentially neglecting other meaningful candidate attributes such as experience or education.

Figure 8: Summary of AI Feature Importance, Recruiter Alignment, and Bias Analyses

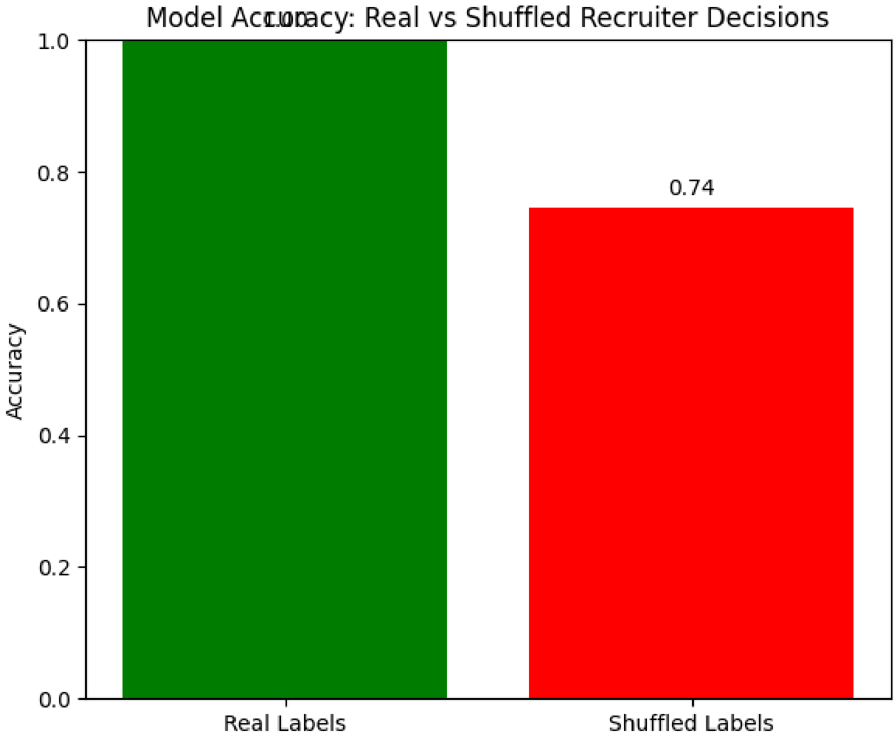


```
Top Influential Resume Features (H1):
              Feature  Permutation Importance
0    AI Score (0-100)                   0.349
1  Experience (Years)                   0.000

Best AI-Recruiter Alignment (H2):
- Threshold: 65
- Alignment: 100.00%
- Precision: 1.00
- Recall: 1.00

Bias Analysis (H3):
- T-test: p = 0.0000
- Cohen’s d: 4.01 (small=0.2, medium=0.5, large=0.8)

- ANOVA by Education Level:
                     sum_sq     df         F    PR(>F)
C(Education)     258.029194    4.0  0.146011  0.964775
Residual      439589.470806  995.0       NaN       NaN
```

The figure above summarizes key findings from the three hypotheses tested in this study. Regarding feature importance, the analysis shows that the AI Score was the only influential predictor of hiring decisions, with a permutation importance of 0.349. At the same time, experience had no measurable impact, receiving a score of 0.000. This indicates that the model relies almost entirely on the AI-generated score when making predictions. For the AI-Recruiter agreement analysis, the optimal threshold for alignment between the model and recruiter decisions was identified as 65, at which the model achieved 100% alignment, with perfect precision and recall—matching recruiter decisions exactly. While this suggests strong agreement, it raises concerns about overfitting, as the model may mimic recruiter behavior based solely on the AI score.

For the bias analysis, a t-test comparing AI scores between hired and rejected candidates produced a p-value of 0.0000, indicating a statistically significant difference, and a Cohen's d value of 4.01, reflecting a considerable effect size. However, an ANOVA across education levels (p = 0.96) found no significant difference in AI scores by educational attainment. This suggests that the AI scoring process is not biased by education level and may be considered education-neutral.

Figure 9: Model Accuracy on Shuffled Labels Suggests Overfitting and Pattern Memorization



The graph above compares model accuracy when trained on real recruiter decisions versus shuffled labels, revealing only a minimal drop in performance—from near-perfect accuracy to 74%. This slight decline suggests that what initially appears to be meaningful insight is, in reality, mere repetition of patterns, highlighting a significant concern. Rather than genuinely learning from candidate features such as experience or education, the model relies almost entirely on a single input: the AI score assigned by an external system. This issue is further compounded by the fact that the AI score's exact algorithm is undisclosed but is likely influenced by factors such as keyword presence, skill mentions, formatting, and education level—criteria commonly used by automated resume screeners. As a result, the model recognizes and replicates the AI score rather than drawing from the underlying, substantive features of the resumes. This behavior risks reinforcing the shallow or biased logic embedded within the original AI system without offering any new or independent judgment.

Discussion

This study provides a focused analysis of AI-influenced hiring practices within the technology industry, but several limitations should be considered when interpreting the findings. First, the domain-specific scope means that any patterns or biases observed are most applicable to tech-related roles and may not generalize to other industries with different hiring norms, job structures, or candidate expectations.

Additionally, the synthetic nature of the dataset and its reliance on simplified numerical and short categorical features introduces significant constraints. Notably, the AI-generated resume scores, which played a dominant role in model behavior, originate from an opaque external system. The lack of transparency regarding how these scores were calculated—including whether full resumes were evaluated or only select features—raises concerns about the interpretability and fairness of the model's decision-making process.

Another critical limitation is the omission of demographic and contextual variables, which are often essential for understanding and mitigating bias in hiring systems. This is especially relevant in the technology sector, which is known to have a pronounced gender imbalance (with men making up 77.4% of the workforce). Without incorporating such variables, the analysis is limited in uncovering or challenging potential sources of systemic bias.

Finally, this study does not simulate the typical use case for AI in hiring, where multiple candidates compete for a single role. Instead, each candidate was modeled as applying for a unique position, which diverges from real-world workflows where AI systems are used to rank or filter candidates within a shared job context. This mismatch limits the realism and applicability of the findings for practical deployment scenarios.

Overall, while the study illuminates how an opaque AI score can dominate hiring decisions, it also underscores the need for greater transparency, representativeness, and contextual alignment in future work to better reflect real-world hiring practices and ensure equitable outcomes.

Work Cited

O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Crown Publishing Group.

Binns, R., Veale, M., Van Kleek, M., & Shadbolt, N. (2018). "'It's reducing a human being to a percentage': Perceptions of justice in algorithmic decisions." In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*.

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*.