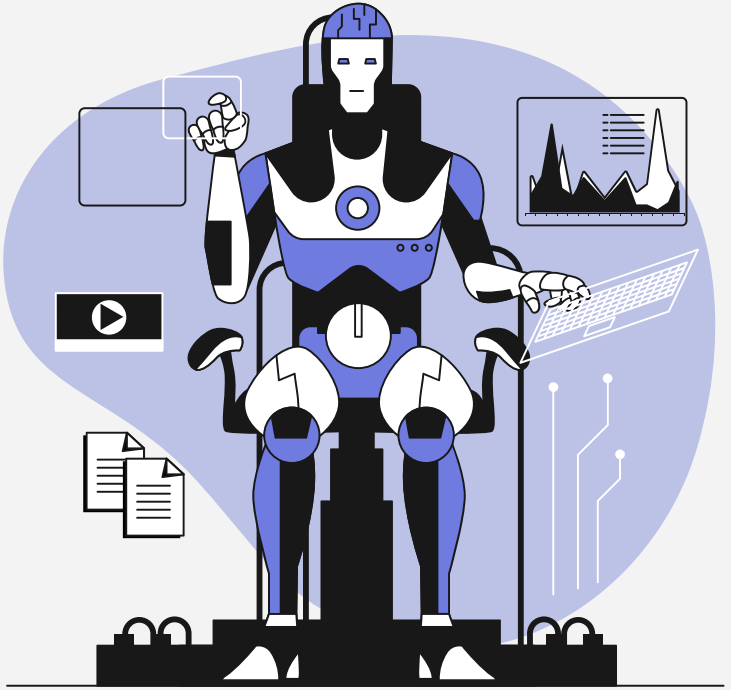
Byte Me – a little salty, a little nerdy, but always efficient with our code

Human vs. Machine: Evaluating AI Decisions in Technical Resume Review

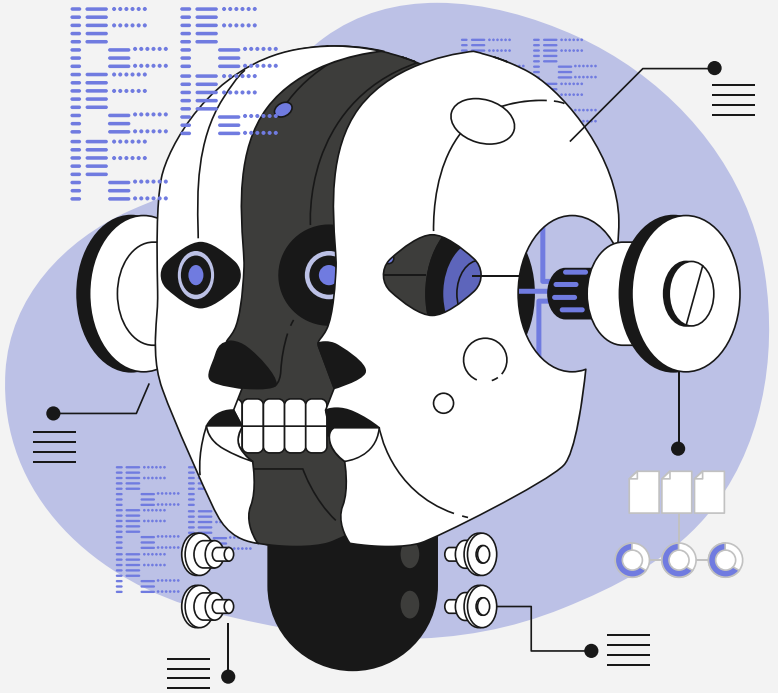


Rithika Cheela, Isabella Castillo, Kaylee Kim

AI Resume Screening

Imagine you're applying for a job in tech. You've polished your resume, added your projects, and listed your skills—but before a recruiter even sees it, it's scanned and scored by an AI.

This lack of transparency in the hiring process is a growing concern, especially as AI becomes more common in hiring. Our project aims to shed light on **what features AI models are really prioritizing in the resumes they evaluate.**



AI Resume Screening Dataset

Our dataset is 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.

Example data

Resume ID	Names	Skills	Experience	Education	Certification	Job Role	Recruiter Decision	Salary Expectation ($)	Projects Count	AI Score (0-100)
1	Ashley Ali	TensorFlow, NLP, Pytorch	10	B.Sc	None	AI Researcher	Hire	104895	8	100

Hypothesis

The AI likely favors **quantifiable traits**—especially years of experience and project count—while undervaluing qualitative, context-rich attributes. Identifying these patterns may reveal hidden biases in the evaluation process.

Data Analysis Plan

#1 Identify the top 2/3 resume features influencing AI decisions

#2 Measure AI–recruiter decision alignment (%)

#3 Investigate signs of potential bias in AI scoring

Methodology Overview

- **Data Cleaning & Preprocessing**
  - Removed non-predictive fields (e.g., resume_id, name)
  - Encoded recruiter decisions: hire = 1, reject = 0
  - Mapped education levels to numerical values:
    - b.sc → 1, b tech/b.tech → 2, mba → 3, m.tech → 4, phd → 5
  - Replaced missing values (NaN) with "None" for interpretability

- **Hypotheses**
  - Quantitative features (e.g., experience, projects) will be most influential
  - Qualitative features (e.g., skills, certifications) will have lower impact

- **Modeling and Evaluation**
  - Decision Tree Classifier – Reveals which features drive predictions
  - Logistic Regression – Estimates the weight and direction of predictors
  - Label Shuffling Test – Assessed whether the model was memorizing patterns (overfitting test)
  - T-Test – Confirmed score differences between hired vs. rejected resumes ($p = 0.0000$)
  - Cohen's d – Measured effect size of AI Score differences ($d = 4.01$)
  - ANOVA – Checked for education-based score bias ($p = 0.96$, no significant difference)

Correlation Pairplot

- Overall view of quantitative columns
- Recruiter Decision marked
  - Rejected = 0, red
  - Hired = 1, blue
- Some distinct trend observed
  - Clear separation of hired/not hired

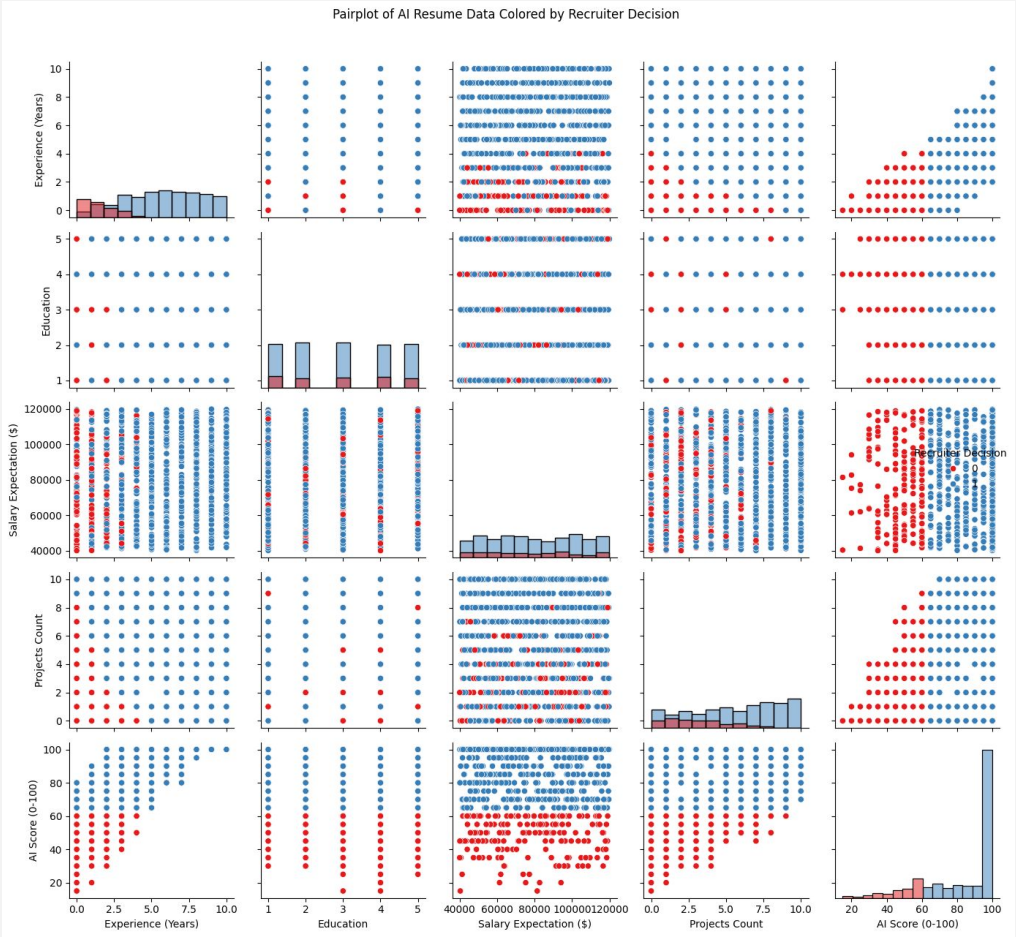


Figure 1.

Correlation Heatmap

- Overall view of quantitative columns' correlations
- Decided to look at other variables more closely using more sophisticated analysis tools
    - Decision trees
    - Logistic regression
    - etc

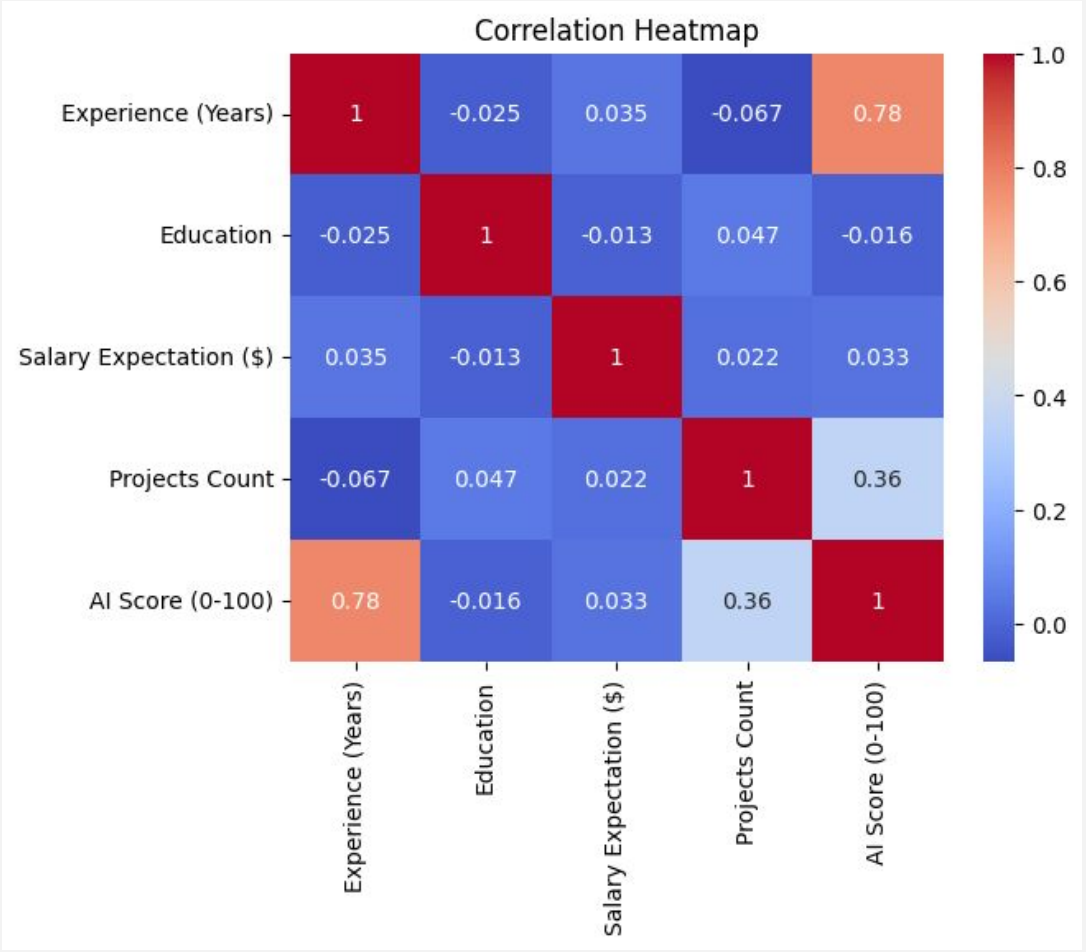


Figure 2.

Decision Tree

- AI Score is major predictor
- Every instance where AI Score is higher than 60 is hired.
    - above 62.5 are always rejected.
    - 62.5 or below are all hired.

Figure 4. The tree only uses AI Score to decide.
- Education and experience don't help at all in the decision — the model just ignores them.

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

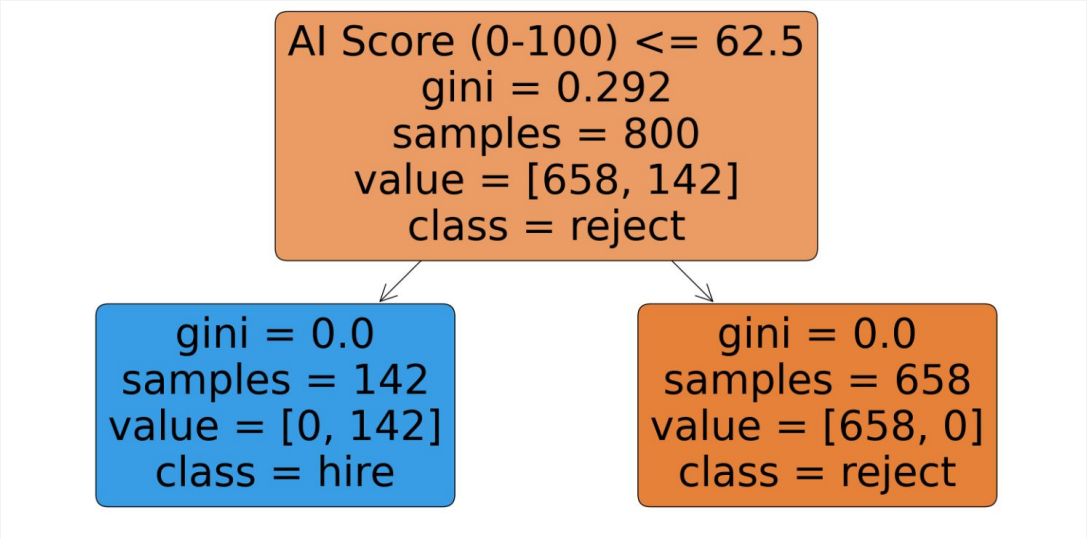


Figure 5.
Blue Gini = 0, Everyone in that group was hired.
Orange Gini = 0, Everyone there was rejected.

142 hires vs. 658 rejections in dataset.

Decision Tree without AI Score

- To see what features have the greatest predictive value.
- A lot of the splitting happens with experience and projects count.

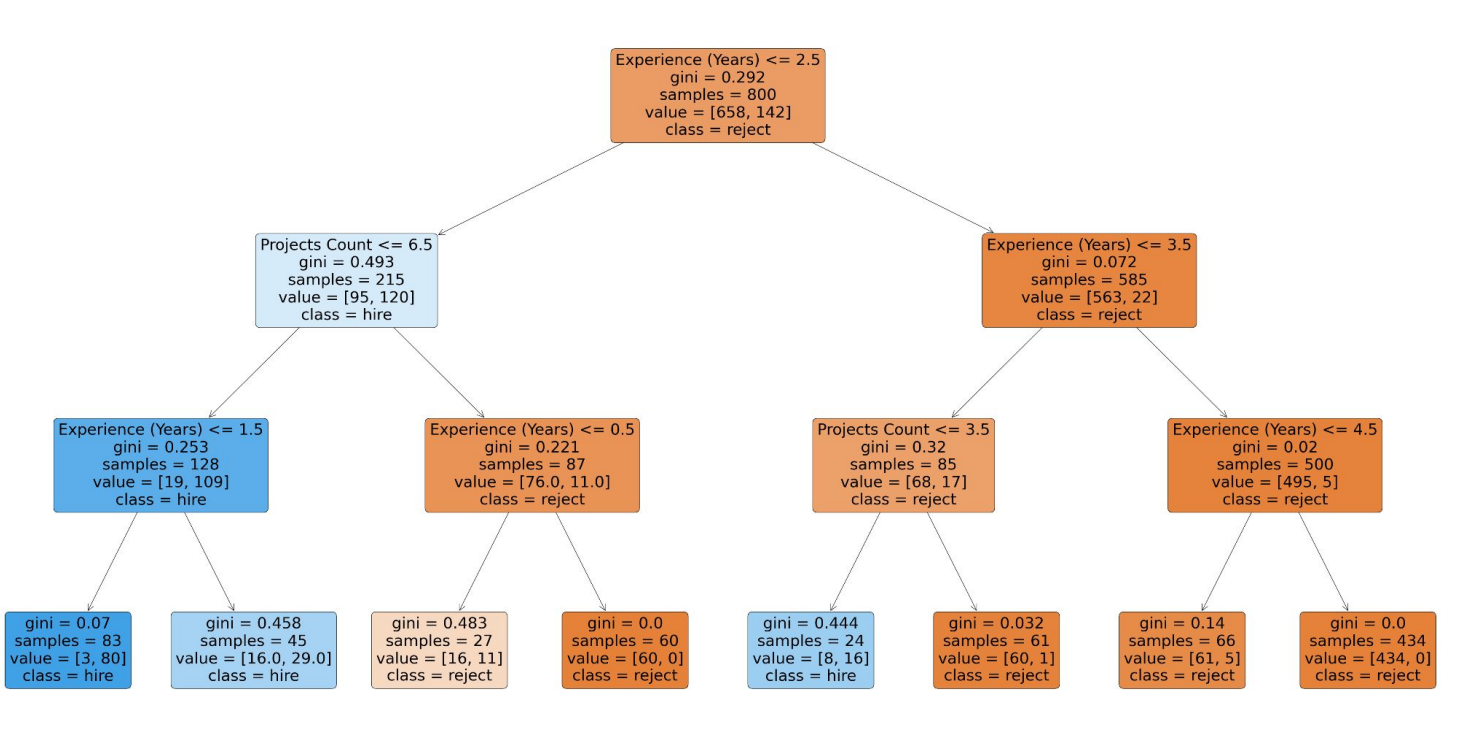


Figure 3.

Education Levels & AI Score Correlation

- The AI assigned consistently high scores across all education levels (1–5)
- No clear trend or preference for higher degrees (e.g., PhD vs. BSc)
- Supported by ANOVA result: p = 0.96 (no significant difference)
- Suggests that the AI score was education-neutral
- This result reduces concern about educational bias in scoring, as education does not influence AI Score.

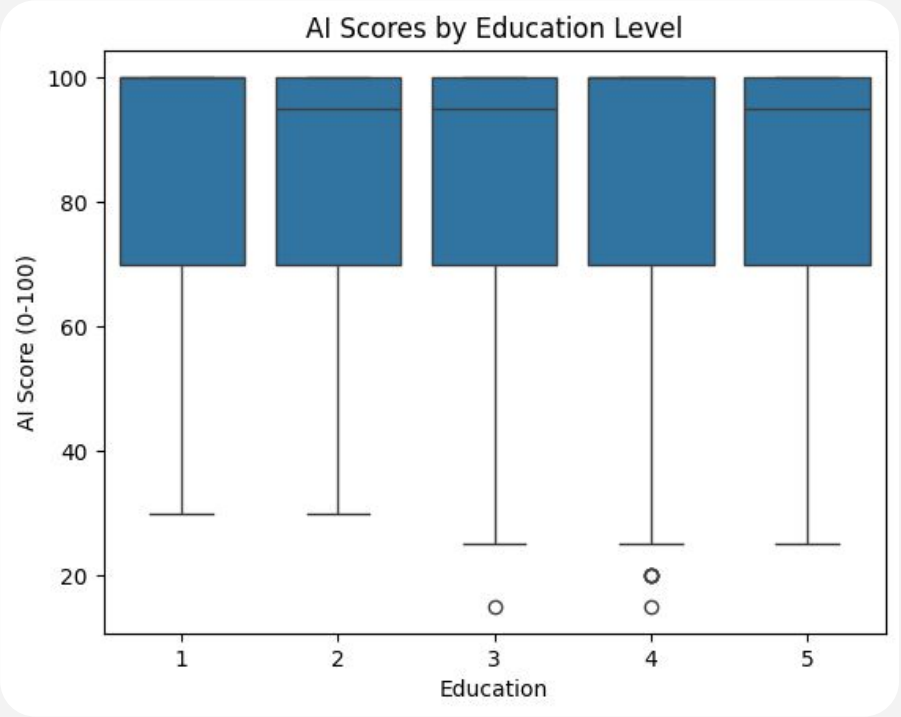


Figure 6. AI scores by education level show no significant variance (ANOVA p = 0.96).

Recruiter Decisions & AI Score Correlation

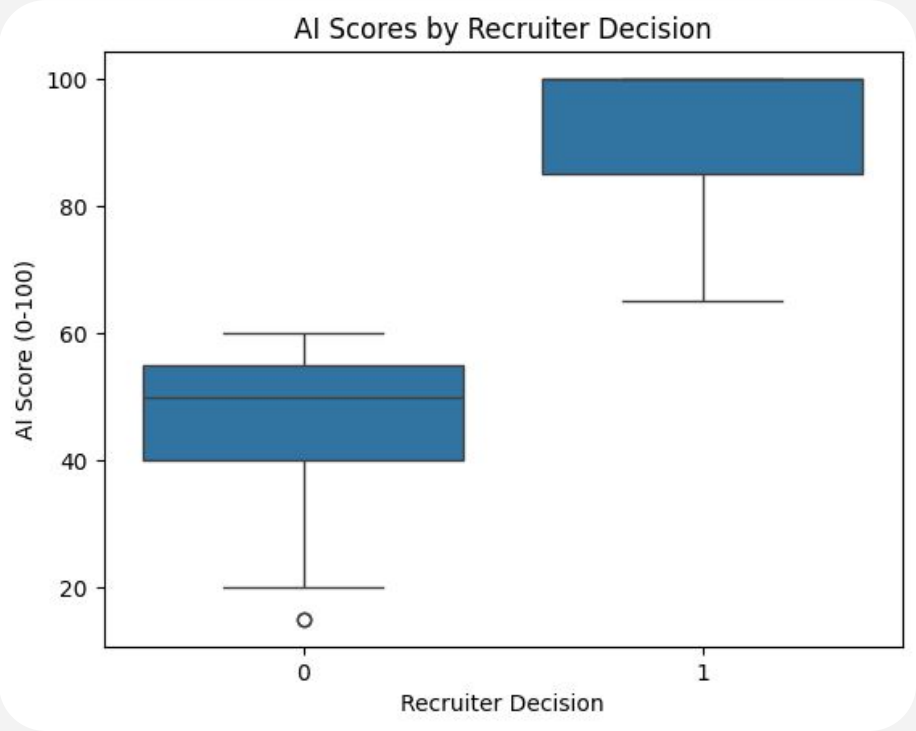


Figure 7. AI scores vary significantly by recruiter decision (t = 49.02, p < 0.0001).

- Recruited candidates received much **higher AI scores** on average
- Clear separation between "hired" (1) and "rejected" (0) resumes
- T-test result: p = 0.0000
  → statistically significant difference
- Suggests that the AI score is a strong predictor of recruiter behavior
- Reinforces concern: model may be learning *ONLY* from the AI score

Key Findings from Model Evaluation

Hypothesis 1 – Feature Importance

- The AI Score (0–100) was the only influential factor in predicting hiring
- Experience had no measurable impact

Hypothesis 2 – AI-Recruiter Agreement

- Best threshold: 65
- Model matched recruiter decisions 100% of the time
- Suggests possible overfitting to the score

Hypothesis 3 – Bias & Fairness

- T-test p = 0.0000 → Strong difference between hired vs. rejected AI scores
- Cohen's d = 4.01 → Extremely large effect size
- ANOVA p = 0.96 → No significant difference in AI score by education level

```
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
```

Figure 8.

Is the Model Really Learning?

It relies almost entirely on an AI score assigned by an external system.

the exact algorithm is not disclosed it is assumed to rely on factors such as,

- keyword presence
- skill mentions
- Formatting
- education level

— common features in AI-driven resume screening tools.

Our model has learned to recognize the score, but not the individual features

This may reinforce biased/shallow logic already in the AI score.

What looks like insight is just repetition.

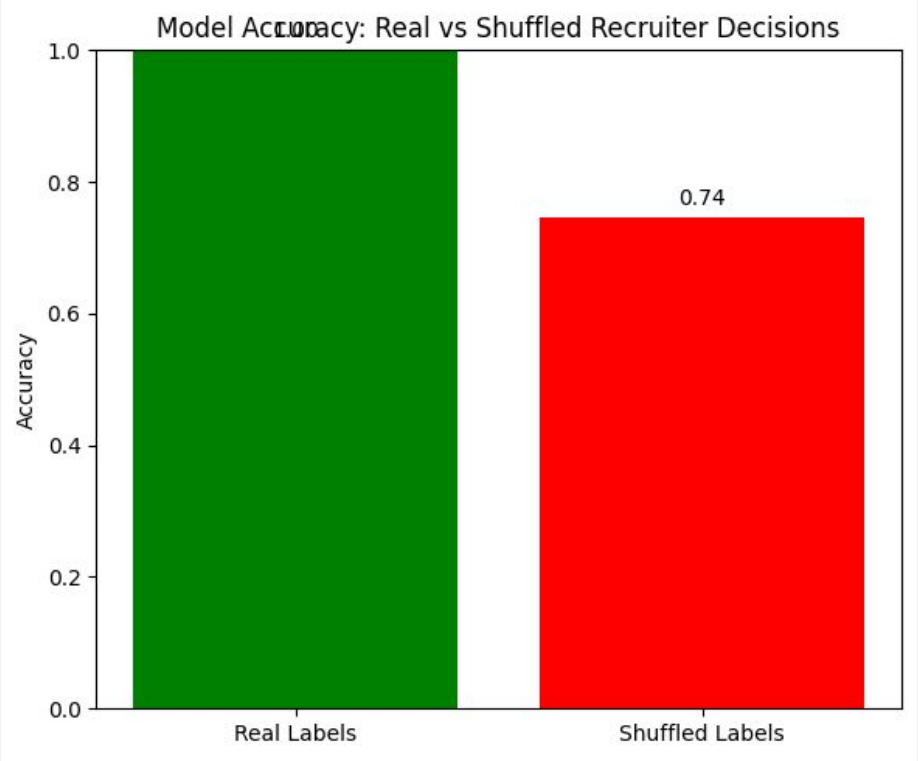
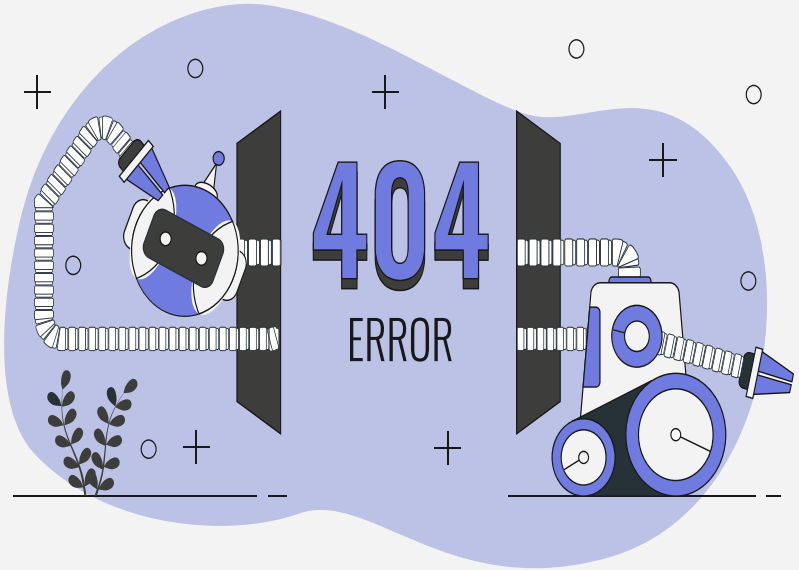


Figure 9. Minimal accuracy drop after randomizing outcomes indicates pattern memorization.

4085959751

In conclusion, our analysis shows that hiring outcomes were determined solely by AI score. This highlights how automated systems may overlook human qualifications and reinforce oversimplified decision-making in real-world hiring.



Limitations and Considerations

- **Domain-Specific Scope on Tech Related Jobs**
  - Any insights found will be most applicable for technology industry
  - May not generalize to other industries

- **Dataset Composition and Feature Transparency**
  - The dataset was synthetically generated with only numerical and short categorical features.
  - Lacks transparency in the following areas:
    - The external resume review model used to generate the AI scores.
    - The depth and logic behind the AI's evaluation of candidate resumes.
    - Whether full resumes were assessed or only select features were considered.

- **Omission of Demographic and Contextual Variables**
  - Often critical to identifying and assessing bias in hiring models
  - Especially given that the tech field has a known gender imbalance (Men 77.4%)

- **Candidate and Role Matching Context**
  - Each candidates are applying for different positions → deviates from typical AI use
    - In real world, AI would be used to see the pool for one job with multiple applicants and find the candidate with the best fit
  - *Limits the realism and applicability of the findings* in simulating actual hiring workflows