D-BIAS Analysis Report

cached_dataset.csv

Generated: 18/11/2025, 03:00:16

Fairness Score: 31/100 Fairness Label: Critical Bias Risk: Critical

Detected Biases

- Categorical Imbalance (col: job_simp) — Moderate
- Categorical Imbalance (col: seniority) — High
- Numeric Correlation Bias (col: min_salary max_salary) — High
- Numeric Correlation Bias (col: min_salary avg_salary) — High
- Numeric Correlation Bias (col: max_salary avg_salary) — High
- Numeric Correlation Bias (col: hadoop spark) — Moderate
- Outlier Bias (col: rating) — Moderate
- Outlier Bias (col: avg_salary) — Moderate
- Outlier Bias (col: company_age) — Moderate

Recommendations

- **Resolve Multicollinearity (Highest Priority)**: Immediately drop `min_salary` and `max_salary` from the feature set, retaining only `avg_salary` to represent compensation. This is essential for model stability.
- **Address Missing `seniority` Data**: Due to the extreme level of missingness, the `seniority` column should be removed from any modeling efforts unless the missing values can be reliably filled from an external source.
- **Counteract `job_simp` Imbalance**: Implement class weighting during model training to force the model to pay more attention to underrepresented job titles. This is crucial for improving fairness and utility across different roles.
- **Mitigate Outlier Effects**: Apply a log transformation to skewed numeric features like `avg_salary` and `company_age` to reduce the disproportionate influence of extreme values and create more stable models.
- **Strategic Data Augmentation**: To build a truly representative model, the dataset must be supplemented with more examples of non-'data scientist' roles and job listings that include seniority information.

Conclusion

:
The dataset's overall "fairness health score" is **Needs Significant Improvement**. The combination of high-severity data quality issues, strong imbalances, and statistical redundancies means that using this data as-is would likely produce a biased, unreliable, and potentially unfair model. Foundational issues must be resolved before it can be trusted for any serious application.

*