Analyzing & Predicting Recruiter Decisions Using Synthetic Resume Data

Team Member: Advait Patil [asp292]

Research Question: What factors significantly correlate with a positive hiring decision, and how can we build & train a model to successfully predict positive hiring outcomes given a list of tangible details?

Context/Motivation: This question is important because many companies use AI-powered tools to parse resumes in order to pick up various keywords & phrases. If a candidate doesn't have enough of these keywords, then he/she will not progress through the hiring process. Oftentimes, there is a lack of transparency or fairness in these automated decisions. Therefore, it is imperative that we explore the relationship between the importance of various details in candidate resumes and a positive hiring outcome to improve hiring practices & enhance AI fairness. By diving deeper into this aspect, we can potentially improve AI-based resume evaluation.

Data Sources:

- To build & train a model to predict positive/negative hiring outcomes, I will be using the AI-Powered Resume Screening Dataset (2025) from Kaggle. The dataset can be accessed using this link: Dataset Link
    - This dataset contains 1000+ rows of synthetic resume details, from skills, experience, education to the AI score, salary expectations and recruiter decisions

- **Preprocessing of Data:**
    - In order to effectively clean & preprocess the dataset, I will be using NumPy and Pandas to remove duplicates or inconsistent rows of data, encode categorical data, & potentially normalize numerical data.

- **Challenges with preprocessing data:**
    - Since the dataset is synthetic, the data may not perfectly mimic real-world scenarios, leading to potential limitations in generalizability.
    - There could be biases in how certain skills, job titles, and degree levels may be represented, leading to mixed results when training models.
    - A decent amount of feature engineering might be needed to preprocess categorical variables.
    - There might be a class imbalance between those who were hired and those who weren't. In that case, I might need to resample the rows and potentially use class weights in modeling to give more importance to the underrepresented class.

Methodology:

- Initially, the dataset will be cleaned & preprocessed using NumPy & Pandas as mentioned in earlier sections. I will perform Exploratory Data Analysis (EDA) to analyze data distributions across multiple variables along with feature importance. I also plan on implementing libraries such as Matplotlib & Seaborn to generate boxplots, histograms, heatmaps, etc. to visualize the correlation between different variables

- When it comes to building & testing an ML model, I will be splitting the dataset into a testing and training dataset. As of now, I plan on utilizing linear regression, logistic regression, decision trees, random forests (possibly neural networks) to predict recruiter decisions and eventually find a model that works best for the data. When building models for each of these methods, I will be focusing on choosing certain features that are correlated with a positive hiring outcome.
- In terms of evaluating each ML model, the following metrics will be used:
    - F1 Score
    - Root Mean Squared Error (RMSE)
    - Confusion Matrices
    - Overall Accuracy (Classification)

- **Tools & Libraries:**
    - NumPy: Data manipulation
    - Pandas: Data preprocessing
    - Matplotlib: Visualization of data distribution
    - Seaborn: Visualization of data correlation & distribution
    - Scikit-Learn: Implementation of basic ML models
    - TensorFlow: Implementation of Neural networks
    - Jupyter Notebook: Main IDE to handle the above tasks

Expected Outcomes:

- While the dataset contains various details regarding the outcome of a certain applicant, I expect certain details, such as the applicant's skills, AI screening score, & experience, to have a significant impact on the recruiter's decision. Ultimately, there might be other variables that influence the decision-making of a recruiter. Regardless, building various models will allow us to identify which one aligns best with the decision of the recruiter or if there are biases surrounding the variables that limit the accuracy of our model.
- As mentioned earlier, the following metrics will be used to evaluate every ML model:
    - F1 Score
    - Root Mean Squared Error (RMSE)
    - Confusion Matrices
    - Overall Accuracy (Classification)
- **Implications of this project:** By undergoing the various steps of this project, we will be able to determine if there are certain skills, job roles, or a certain salary range that contribute to biased results or fairness in certain areas. If some bias is present, then our model can help AI-driven recruitment in the future by identifying certain areas where bias is most prevalent. Likewise, this process would also help potential applicants improve their chances of moving forward in the hiring process by giving them an idea of what areas of their resume hold the most weight. Moreover, companies themselves can leverage these insights to promote more reliable AI hiring tools that align with recruiter decisions while mitigating discrimination in all areas.
- **Extensions of this project:** One way to extend this project is by analyzing & predicting recruiter decisions in various fields, such as Finance & Healthcare, to determine if certain industries weigh hiring factors differently. Another idea that is particularly relevant to us as college students is to perform the same analysis & model training on a dataset with resume data from internship applicants of various industries. This could help determine whether there are certain aspects of a resume that internships & full time roles value the most, along with those that are negligible. Moreover, using NLP techniques to analyze resumes can test whether certain resume formats, phrasing techniques, & layouts affect the overall outcome of a jobseeker.