Proposal: Resume Analysis and Classification System

Jesse DiMarzo
Statistics Major, Mathematics Minor
University of Connecticut

October 28, 2024

Objective Statement

The primary objective of this project is to build a resume analysis and classification system that categorizes resumes into different job roles and provides insights into skill gaps between applicants. The goal of this project is to develop a method for recruiters to quickly and efficiently identify suitable candidates for their job openings, as well as to help applicants improve their resumes to meet employer standards.

Data Science Pipeline

Data Collection and Preparation

The first step is to find suitable data to accomplish the project's goals. I have found a dataset on Kaggle that contains over 2400 resumes sourced from livecareer.com, stored in text and HTML formats, along with their respective job categories.

It is essential to first load the dataset, inspect it, and begin cleaning and preprocessing it using tools such as Pandas and NumPy. It may be useful to convert all text to lowercase to allow for ease of identification in machine learning. Additionally, removing special characters and punctuation to reduce noise in the data will be critical parts of processing the data.

Exploratory Data Analysis (EDA) and Visualization

The next step is to begin Exploratory Data Analysis (EDA) to better understand the dataset and identify patterns, underlying trends, correlations, and statistically significant variables. It will be useful to employ the following visualizations during EDA:

- Bar charts to investigate the distribution of resumes across job categories, and address issues such as class imbalance.

- Heatmaps to identify correlations between skills and job categories.

Data Processing for Modeling

At this stage of the pipeline, I will focus on transforming the cleaned dataset into a format suitable for modeling. The job categories will be label encoded, and any other relevant categorical variables will be one-hot encoded if needed. Any numerical features will be standardized, and missing (NaN) values will be handled appropriately to ensure consistency in the data.

Machine Learning Techniques

To classify resumes into different job categories, I will use learning models available in Scikit-Learn, such as Logistic Regression, Decision Trees, and RandomForest. Decision Trees are especially useful due to their ease of interpretation and effective visualization of the decision-making process. Evaluation metrics such as accuracy, precision, and recall will be used to assess model performance.

Discussion

The most challenging parts of this task will likely be dealing with the variability in resume formats, i.e., different section titles, bullet points, etc. This, as well as potential imbalances in the dataset, where certain categories and job roles are more prominent than others, could cause issues if they arise.

The difficulty in obtaining diverse data may be a limitation in this project, as the current dataset only contains approximately 2400 observations, which may not be representative of the true job market. To address this, it will be crucial to either search for supplementary datasets or conduct web scraping to expand the current dataset.

GitHub Repository