**Automated Resume Categorization Using Machine Learning**

**1.Introduction**

This work investigates the creation and use of a machine learning model to automatically classify resumes (CVs) into different field of employment, including healthcare, IT, finance, and others. Employing Natural Language Processing (NLP) and a Random Forest classifier, the model attained a high level of precision in categorizing resumes according to their content. This work examines the methodologies of data preparation, model training, assessment metrics, and deployment strategy to illustrate the efficacy of automation in optimizing HR operations. Categorizing resumes according to their content is crucial in recruiting as it enables HR departments to effectively organize and screen candidates. Conventional hand sorting is very labor-intensive and susceptible to mistakes, thereby requiring an automated method. This project targeted the development of a machine learning model that use natural language processing (NLP) methods to automatically categorize resumes into predetermined job areas.

**2. Data Acquisition and Preprocessing**

**2.1 Data Acquisition**

The information was obtained via a Drive search link and consisted of resumes from different job categories saved in PDF format. This dataset provides the fundamental basis for both training and assessing the model. There seems to have been a problem with the direct presentation of the dataframes. Nevertheless, I am able to succinctly outline the main discoveries and procedures derived from the statistical analysis and exploratory data analysis (EDA).

**2.2 Statistical Analysis**

**Data Types and Non-Null Counts:** All columns (ID, Resume\_str, Resume\_html, Category) in the dataset include 2484 non-null items, indicating the absence of any missing values.

**ID Column Statistics:** The ID column is numeric, with a range between approximately 3.5 million and 36 million.

* **Summary of Statistics:**
* Count: 2,484
* Mean: 31,826,160
* Standard Deviation (std): 21,457,350
* Minimum (min): 3,547,447
* 25th Percentile (25%): 17,544,300
* Median (50%): 25,210,310
* 75th Percentile (75%): 36,114,440
* Maximum (max): 99,806,120

**2.3 Exploratory Data Analysis (EDA)**

* **Category Distribution:**

The bar chart shows the distribution of resumes across different categories. This helps in understanding which categories are most or least represented in the dataset.

A chart of different colors

Description automatically generated

**Figure 1: Category Distribution**

* **Word Count in Resumes:**

To facilitate the analysis of resume length, the word\_count column was included into the dataset.   
The analysis of word counts in a histogram offers valuable information on the average length of resumes, therefore facilitating feature engineering and text preprocessing procedures.

A close up of words

Description automatically generated

**Figure 2: Word Count in Resumes**

**2.4 Data Preprocessing**

Preparation of the data for model training included the following procedures:

* **Text Extraction:** Text extraction was performed using the pdfplumber library to extract text from PDF files, specifically targeting parts such as Summary, Experience, Education, and Skills.
* **Cleaning and Formatting:** Data cleaning and formatting included removing non-alphabetical letters, filtering out stopwords, and lemmatizing the text to reduce words to their basic forms.
* **Data Structuring:** The processed data was saved in a CSV file, and prepared for extracting features and training machine learning models.

**3. Feature Extraction**

TF-IDF The textual input was converted into numerical characteristics using vectorization, a crucial methodology for training machine learning models. By quantifying the significance of each word in a text in relation to the whole corpus, this approach enables the model to differentiate between various job domains.

**4. Model Training**

The selection of a Random Forest classifier was based on its resilience and capability to effectively process high-dimensional data. To guarantee equitable representation of all employment categories, the model was trained using a balanced dataset generated using SMOTE (Synthetic Minority Over-sampling Technique).   
Training Procedure: The model underwent training using the training dataset, which included 80% of the whole data source. The remaining 20% was divided into validation and test sets to assess the performance of the model and optimize its parameters.

**5. Model Evaluation**

The efficacy of the approach was assessed utilizing essential performance indicators:

* **Accuracy:** Evaluation of the model's overall accuracy in categorizing resumes.
* **Precision:** Accuracy rate of optimistic forecasts.
* **Recall:** The model's capacity to accurately detect all relevant instances.
* **F1-Score:** A harmonic mean of accuracy and recall is a balanced metric that evaluates the performance of a mathematical model.

A screenshot of a computer screen

Description automatically generated

**Figure 3: Classification report**

**6. Prediction and Automation Script**

A resume classification automation script was created to optimize the resume categorizing procedure. The following script:

* **Processes Batches of Resumes:** Utilises the trained model to categorise a directory of PDF resumes into job categories.
* **Output Results:** The classified resumes are arranged into their corresponding folders, and the outcomes are stored in a CSV file.

**7. Discussion**

The model's exceptional precision and equitable performance across different job categories underscore its efficacy in automating resume categorization. By using a Random Forest classifier in conjunction with TF-IDF vectorization, we have devised a resilient method that demonstrates exceptional accuracy in precisely classifying resumes. Incorporating data balancing methods, such as SMOTE, guaranteed that the model maintained its robust performance even in the presence of skewed datasets. This degree of automation has the potential to greatly reduce the need for human labor in HR departments, hence improving both productivity and precision. The model's performance was optimized by using many data pretreatment methodologies, such as text cleaning, normalization, and vectorization. A meticulous selection of these preprocessing procedures was made to ensure the optimal conversion of the unprocessed text input into a format that is well-suited for machine learning algorithms. Various computational models were evaluated, including Logistic Regression, Support Vector Machine (SVM), XGBoost, and Gradient Boosting Machine (GBM). Nevertheless, the Random Forest classifier consistently achieved the greatest level of accuracy. Furthermore, thorough hyperparameter tweaking was performed on all models, however the present configurations were shown to be the most optimum for attaining the greatest outcomes.

**8. Conclusion**

The initiative effectively created and implemented a machine learning model that can accurately classify resumes into various job categories. The automation script notably improves the practical use of the model in actual human resources operations. Potential future enhancements may include enlarging the dataset, integrating more advanced NLP methodologies, and optimizing model parameters to achieve even superior performance.

**9. Future Work**

* **Expand Dataset:** To enhance the resilience and extrapolability of the model to a wider range of employment categories.
* **Advanced NLP Techniques:** Leverage transformer-based models such as BERT to enhance contextual comprehension.
* **Model Refinement:** Utilise ensemble techniques or deep learning models to further optimise performance.