Milestone Report 2

Understanding The Preprocessed Dataset and Feature Engineering

Group I

Beema Sajeev

Jaisy Joy

ParminderKaur

Harsh Kundal

Diksha

Capstone-project/milestone-2.ipynb at main · Beema20/Capstone-project (github.com)

Table of Contents

Introduction

During this milestone, we focused on crucial data preprocessing and feature engineering for our project. These steps are essential for preparing our datasets for advanced analysis and modeling. Data preprocessing involves cleaning and transforming raw data, while feature engineering involves creating new features or modifying existing ones to improve predictive models. We completed various tasks such as text preprocessing and numerical feature scaling to enhance the quality and usability of our datasets. Our goal is to extract insights into job market trends, salary predictions, and remote work dynamics. We integrated datasets, encoded categorical variables, and engineered new features to lay a solid foundation for exploratory data analysis and model development in the next phases.

Key Tasks Completed

1. **Text Preprocessing**
    - Tokenized text
    - Removed stopwords and non-alphanumeric tokens
    - Lemmatized tokens
2. **Data Integration**
    - Merged 'company_industries' and 'company_specialities' with 'companies'.
    - Merged 'employee_counts' with 'companies'.
    - Merged 'job_industries', 'job_skills', 'salaries', and 'benefits' with 'postings'.
3. **Categorical Encoding**
    - Used LabelEncoder from sklearn to encode categorical columns across datasets ('companies', 'company_industries', 'skills', 'industries', 'postings').
4. **Feature Engineering**
    - Extracted a new feature 'location' from 'state' and 'city' columns in the 'companies' dataset.
    - Combined 'state' and 'city' data into a unified format using string concatenation.
    - Encoded the newly created 'location' feature using LabelEncoder to transform it into numerical values.

- Stored the LabelEncoder instance used for 'location' in the 'label_encoders' dictionary for future reference and consistency in data transformation.

5. **Numerical Feature Scaling**

   - Applied StandardScaler from sklearn to scale numerical columns ('company_size', 'max_salary', 'min_salary', 'views', 'applies') in 'companies' and 'postings' datasets.

Output Overview

1. **Data Quality Issues Identified:**

- Most datasets are fairly complete, with some notable exceptions in postings and salaries where many entries have missing values, especially for salary-related columns.
- The companies dataset also has a significant amount of missing data in several columns, including description, company_size, and state.
- salaries dataset has missing values in max_salary, med_salary, min_salary, pay_period, currency, and compensation_type columns.
- Other datasets like company_industries, company_specialities, employee_counts, benefits, job_industries, job_skills, and skills are complete and have no missing values.
- The industries dataset has a few missing values in the industry_name column.

2. **Remediations:**

- Data Cleaning: Addressed the missing values in critical columns to ensure the integrity of the analysis. This included deletion, or further investigation to find the missing data.
- Data Integration: Ensured that the company_id and job_id fields are properly linked across datasets for coherent analysis.
- Feature Engineering: Create new features that could enhance the predictive power of any models or analyses that will be conducted.

Next Steps

1. **Exploratory Data Analysis (EDA)**

- Conduct a thorough EDA to uncover patterns, trends, and insights into job market trends, salary predictions, and remote work dynamics.
- Visualize data distributions, correlations, and other relevant aspects.
- Identify correlations and patterns within the datasets.

2. **Feature Selection**
    - Identify additional features from existing data, such as industry-specific metrics, regional trends, etc.
    - Evaluate feature importance and potential impact on model performance.
3. **Model Development**
    - Develop and train predictive models for salary prediction, job trend analysis, or classify companies based on various attributes etc.
    - Optimize model performance through hyperparameter tuning.
4. **Model Evaluation**
    - Assess model performance using appropriate metrics.
    - Iterate on model development based on evaluation results.
5. **Deployment**
    - Deploy the final models into production for real-world applications.

Challenges Faced

- **Complex Data Structures:** Handling complex data structures and inconsistencies across datasets.
- **Time Management:** Balancing time between data preprocessing, feature engineering, and analysis.
- **Data Integrity:** Ensuring data accuracy and reliability throughout preprocessing stages.

Conclusion

Milestone 2 completes the extensive data preprocessing and initial feature engineering phase. We have meticulously cleaned, integrated, and transformed our datasets to address data quality issues and prepare them for advanced analysis and modeling. The datasets are now coherent, enriched with meaningful features, and free from significant inconsistencies. This sets the stage for exploratory data analysis, feature selection, and model development. Through these steps, we will uncover patterns, select impactful features, develop predictive models, and evaluate their performance. The challenges we faced have provided valuable learning experiences. With the datasets primed for analysis, we are confident in our ability to derive actionable insights that will benefit stakeholders in understanding job market dynamics. Milestone 2 has prepared our datasets for in-depth analysis and modeling, positioning us to advance to the next phases of the project and derive valuable insights for job seekers, employers, and policymakers.