Requirement 8: Final Project Scripts

Meng

### Project Title

Job Hunting for Data Crew: Salaries Prediction based on Job Postings Data

### Description of data & Link

Descriptions: This dataset contains a nearly comprehensive record of 124,000+ job postings listed in 2023 and 2024. Each individual posting contains dozens of valuable attributes for both postings and companies, including the title, job description, salary, location, application URL, and work-types (remote, contract, etc), in addition to separate files containing the benefits, skills, and industries associated with each posting. Most jobs are also linked to a company, which are all listed in another csv file containing attributes such as the company description, headquarters location, and number of employees, and follower count.

Kaggle link: <https://www.kaggle.com/datasets/arshkon/linkedin-job-postings/data>

Download files: Please see data folder “FINAL PROJECT/ Final Project Data Files/ Kaggle Data Archive”

Cleaned-up: “FINAL PROJECT/ Final Project Data Files/ ML\_processed\_job\_postings\_finale\_v2.csv”

### Attributes

Original Data: Please refer to Kaggle Data Card

CLEAN-UP DATA: ‘ML\_processed\_job\_postings\_finale\_v2.csv”

Shape: 3902 rows x 50 columns

Data types:

* String (29 cols)
  1. Categorical: 6 cols

sector, region, degree level, work type, formatted experience level, job classification

* 1. Geographical: 2 cols
  2. Text: 6 cols

title, descriptions, company name, company desc, industry, specialty

* 1. ID number: 2 cols
* Bool (1 col)
* int & float (21 cols)

### General Statistics

Please see file in folder: ‘’ /FINAL PROJECT/ df\_describe.xlsx’’

### Tools/ Methods

LR& Logistic Reg: predicting salary with influential predictors

SVM: Predicting jobs with salary over 140K

Decision Tree: Predicting jobs with salary over 140K

Clustering: what jobs clustering looks like.

### Detail description of what problems/questions your team plans to predict / study

* What predictors are the key influencers for data job salary?
* Find predictors for salary, and based on the predictors and results, it could help job hunting for data people to identify where to relocate, what skills they should study, and what companies/industry they should focus on.

### Scripts

|  |  |  |
| --- | --- | --- |
| Slide Number | Slide Image | Key points |
| 1 | A screenshot of a computer  Description automatically generated | 1. Topic: Job analysis based on LinkedIn posting dataset. 2. Reason for this topic: 3. Relative to our majors 4. Large size dataset allowing more possible models. 5. Expectations for this presentation: what the job market looks |
| 2 |  | 1. Goal: Explore the relationships between salary and predictors. 2. Is there a model that can predict salary based on skills, industries and locations and etc.? 3. If not, what are the trends in current job market? Examples. |
| 3 |  | 1. Steps for data filtering and cleaning, and dataset size changes. 2. **Imputation, missing values handling, collinearity, skewness.** 3. Lists of transformations for data for each step. 4. Final dataset size for our models. |
| 4 |  | 1. How to define “data jobs” 2. How hard skill scores are extracted. - from ‘job description’ 3. Classification of industries into 11 sectors 4. Classification of locations into 8 regions 5. Failed extracting: ’moving score’, ’business/engineering classification’ 6. Calculating economic scores by adding GDP and avg income data. |
| 5 |  | 1. LR model & Predictors 2. Results: no linear relation |
| 6 |  | 1. Lasso results 2. Identify two predictors (did you standardize it ?????) |
| 7 |  | Conclusion for LR models |
| 8 |  | 1. Logistic Analysis: predicting 60K salary with all predictors 2. Results: not good |
| 9 |  | 1. Logistic Analysis model 2: improved model 2. Experience level and headquarter locations are striking predictors. |
| 10 |  | 1. SVM model 1: 140k prediction, PCA features. 2. Steps:   Predictors trying 🡪 3D plotting🡪 partially manual PCA 🡪 auto PCA🡪 Gamma, C and kernel testing 🡪 Classification reports/ROC/Visualization   1. Cons: class imbalance (75% vs. 25% ) because 140k is 3rd quantile. |
| 11 |  | 1. SVM model 1: 140k prediction, 3 PCA features. 2. What those 3 features are. 3. Results: improved class 1 prediction. |
| 12 |  | 1. Lasso to pick top predictors 2. SVM models with 5 predictors 3. Results: acceptable but not super good. |
| 13 |  | 1. Decision tree for predicting 110kor more 2. Root nodes |
| 14 |  | Root nodes analysis: interpretations |
| 15 |  | Model quality |
| 16 |  | 1. Lasso for top predictors 2. Run DT model to predict 140k 3. Similar results: Company size and Information sector is the key predictors for 140k jobs |
| 17 |  | 1. Model 2 continued: other methods ( GridSearch, Random Forest, xgbBoost) 2. Results: not ideal, bad performance |
| 18 |  | 1. K-means clustering results 2. Interpretations of the clustering 3. Model quality: silhouette score |
| 19 |  | Conclusions for Clustering analysis |
| 20 |  | 1. PCA for all features, seems having 2 clusters 2. Comparisons with other methods with silhouette score =0.552 3. Interpretation |
| 21 |  | 1. Summary 2. Disadvantages of this projects 3. Next steps for this project |

**Script Showing Our Regression Analysis:**

[Final\_Project\_Script\_Regression\_Analysis.ipynb](https://uofstthomasmn.sharepoint.com/:u:/r/sites/MachineLearningGroupProject577/Shared%20Documents/General/Final_Project_Script_Regression_Analysis.ipynb?csf=1&web=1&e=cKs3QM)