Project Title: Skill-Based Job Role Predictor

1. Introduction

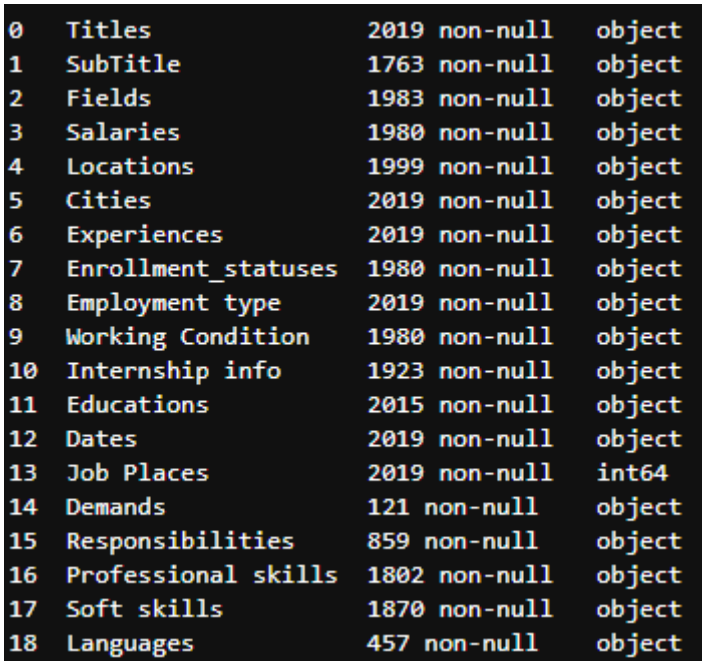
In the evolving landscape of employment, skill-based hiring has emerged as a practical and effective approach for matching talent to jobs. Traditional qualifications like degrees or job titles no longer fully reflect a candidate's capabilities or suitability for a role. This project, **Skill-Based Job Role Predictor**, leverages machine learning to identify appropriate job roles for individuals based on their skill sets and other attributes such as education, experience, and work preferences.

This system can support job seekers by recommending roles that align with their unique profiles and assist employers or platforms in automating and improving candidate-job fit.

2. Dataset Overview

The dataset used in this project was scraped from the Kazakhstani employment portal enbek.kz. After cleaning, it consists of **5,412** job listings with **53**(20) **columns** detailing various job attributes. Some key columns include:

- **job_title, subtitle :** Target variable, representing the job role.
- **education_requirement, work_experience, languages_required, soft_skills, professional_skills:** Candidate requirements.
- **job_skills, field, working_conditions, employment_type, enrollment_status, internship_info,demands**: Job characteristics.
- **salary, location, schedule, city**: Work environment indicators.



0	Titles	2019 non-null	object
1	SubTitle	1763 non-null	object
2	Fields	1983 non-null	object
3	Salaries	1980 non-null	object
4	Locations	1999 non-null	object
5	Cities	2019 non-null	object
6	Experiences	2019 non-null	object
7	Enrollment_statuses	1980 non-null	object
8	Employment type	2019 non-null	object
9	Working Condition	1980 non-null	object
10	Internship info	1923 non-null	object
11	Educations	2015 non-null	object
12	Dates	2019 non-null	object
13	Job Places	2019 non-null	int64
14	Demands	121 non-null	object
15	Responsibilities	859 non-null	object
16	Professional skills	1802 non-null	object
17	Soft skills	1870 non-null	object
18	Languages	457 non-null	object

Preprocessing Steps:

- Missing values were filled or dropped based on relevance.
- Skills, education, and other categorical variables were one-hot encoded and label encoded.
- Standardization, robust scaling, log1p, cutting, boxcox for the data to get suitable.
- The dataset was balanced and filtered to keep the most frequent job roles.

0	Titles	5412 non-null	object
1	SubTitle	5412 non-null	object
2	Fields	5412 non-null	object
3	Salaries	5412 non-null	int64
4	Locations	5412 non-null	object
5	Experiences	5412 non-null	int64
6	Working Condition	5412 non-null	object
7	Educations	5408 non-null	object
8	Dates	5412 non-null	object
9	Job Places	5412 non-null	int64
10	Demands	5412 non-null	object
11	Responsibilities	5411 non-null	object
12	Professional skills	5412 non-null	object
13	Soft skills	5412 non-null	object
14	Kazakh	5412 non-null	int64
15	Russian	5412 non-null	int64
16	English	5412 non-null	int64
17	Chinese	5412 non-null	int64
18	Turkish	5412 non-null	int64
19	seasonal	5412 non-null	float64
20	permanent	5412 non-null	float64
21	temporary	5412 non-null	float64
22	remote	5412 non-null	float64
23	unpaid_intern	5412 non-null	float64
24	paid_intern	5412 non-null	float64
25	no_intern	5412 non-null	float64
26	Abay_reg	5412 non-null	float64
27	Almaty	5412 non-null	float64
28	Almaty_reg	5412 non-null	float64
29	Astana	5412 non-null	float64
30	Atyrau_reg	5412 non-null	float64
31	Aqmola_reg	5412 non-null	float64
32	Aqtobe_reg	5412 non-null	float64
33	West_kz_reg	5412 non-null	float64
34	Zhambyl_reg	5412 non-null	float64
35	Zhetisu_reg	5412 non-null	float64
36	Mangistau_reg	5412 non-null	float64
37	Pavlodar_reg	5412 non-null	float64
38	North_kz_reg	5412 non-null	float64
39	Turkistan_reg	5412 non-null	float64
40	Shymkent	5412 non-null	float64
41	East_kz_reg	5412 non-null	float64
42	Qaragandy_reg	5412 non-null	float64
43	Qostanay_reg	5412 non-null	float64
44	Qyzylorda_reg	5412 non-null	float64
45	Ulytau_reg	5412 non-null	float64
46	rotational	5412 non-null	float64
47	other_specific_jobs	5412 non-null	float64
48	shift_based	5412 non-null	float64
49	part_time_work_week	5412 non-null	float64
50	part_time	5412 non-null	float64
51	full_day	5412 non-null	float64

3. Methodology

3.1 Feature Engineering

- **Preprocessing** data types, replaced, removed unnecessary data

- **Text-based features** all of the text data were parsed and removed stopwords in kazakh(**self-made**) and russian(**nltk**) then transformed into binary vectors using **tf-idf** then all of the text columns were inserted into one, also gave the language prefix as: **kz, ru, unknown**
- **Categorical features** such as city and education columns were label encoded and one hot encoded.
- **Numerical features** like salary and experience were standardized, robusted.

3.2 Model Selection

The following classification models were trained and compared:

- **ML(Linear Regression, Random Forest, Gradient Boosting) hyperparameters: (GridSearchCV, Ridge, Lasso, cross_val)**

- **NLP(Random Forest Regressor, PipeLine with Tf-idf)** other models did not give good results even **Gradient Boosting**

- **Combined NLP, ML with Pipeline(Random Forest, XGBoost, CatBoost)**

- **Neural Networks(MLPRegressor)** with SimpleImputer, TfidfVectorizer

Models were evaluated using **MSE**, **RMSE, R2 squared, Cross_val_score.** Also **GridSearchCV** for finding best parameters

3.3 Tools Used

- Python
- Libraries: **pandas, numpy, scikit-learn, matplotlib, seaborn, BeautifulSoup, api (for scraping), sklearn, xgboost, nltk, re, scipy**

4. Results and Discussion

After experimentation, the **Random Forest Regression,MLPRegressor** achieved the highest performance with an **accuracy of 65.1%**. **Catboost, XGBoost** followed closely, offering fast training but slightly lower generalization. **Linear Regression, Lasso, Ridge** provided interpretability and insight into feature importance.

```
Catboost MSE: 5368270814.47
Catboost R2: 0.6062922457977351
Catboost RMSE: 73268.48
```

Top 10 Predicted Job Titles:

- Sales Manager
- Driver
- Cook

- Accountant
- Engineer
- Storekeeper
- Nurse
- Security Guard
- Cashier
- Teacher and so on

These reflect the most common roles in the dataset and those most easily predicted based on skills.

5. Conclusion

This project demonstrated how machine learning can effectively predict suitable job roles based on skill and profile data.However, as you can see, the machine did not perform that good, because of the unbalanced salary in kz work industry :) Whatsoever, the tool could be integrated into job search platforms to enhance recommendations and reduce mismatches in applications.

Salaries	1.000000
Working Condition	0.205842
Educations	0.205281
Experiences	0.192676
Qostanay_reg	0.186117
Astana	0.144262
English	0.138314
Job Places	0.115548
full_day	0.090174
shift_based	0.085645
permanent	0.072084
Almaty	0.058944
Qaragandy_reg	0.058462
paid_intern	0.048397
Chinese	0.024398
unpaid_intern	0.017720
Russian	0.017588
no_intern	0.006111
remote	0.004518
other_specific_jobs	-0.000169
Abay_reg	-0.001691
Kazakh	-0.001939
rotational	-0.011234
Mangistau_reg	-0.013953
Atyrau_reg	-0.015309
Turkish	-0.015330
North_kz_reg	-0.018566
part_time_work_week	-0.025020
seasonal	-0.026525
Qyzylorda_reg	-0.028679
Turkistan_reg	-0.033230
Almaty_reg	-0.034710
Ulytau_reg	-0.036291
Aqmola_reg	-0.037448
Shymkent	-0.037603
East_kz_reg	-0.050556
Pavlodar_reg	-0.050587
Aqtobe_reg	-0.053880
temporary	-0.066926
Zhambyl_reg	-0.074291
Zhetisu_reg	-0.079921
Experience_Job_Interaction	-0.083393
West_kz_reg	-0.086392
part_time	-0.150934

Correlation between the columns above

Limitations:

- Bias in the dataset toward certain roles
- Reliance on structured input; real-world resumes may require NLP parsing

Future Work:

- Add complex NLP models like (BERT, GPT) for better result
- Extend the model with job similarity scores
- Improve recommendations using collaborative filtering or hybrid models

Prepared by: Amanekov Farkhat, Daulet Serikkazy,