Technical Assessment Activity - Data Scientist position

1- Problem Statement:

Mohammad, a recruiter at KABI based in KSA, has a list of jobs that need to be published. He is seeking assistance in establishing the appropriate salary range for each position based on job descriptions, taking into consideration factors such as skills, work experience, education, and more. Mohammad is struggling to determine the optimal salary range. As a Data Scientist, you will assist him in determining the most suitable salary range for each position.

2- Request:

You have been tasked with creating a solution to accurately predict salary ranges for positions in the Kingdom of Saudi Arabia (KSA), taking into account factors such as skills, work experience, education, and more.

Part I: Research:

Conduct a review of current and recent research in the field of position salary prediction
and provide a concise summary of the prevalent findings and recent development:

I did my research to check the best practices to finish this task I found that we need firstly to scrap the data or get it using some resources such as Glassdoor then process the data to be cleaned to use it in machine learning, after all of that we need to load it in warehouse or database, by doing all of that we will be did the ETL pipeline, by extracting the data from the needed sources then processing it and finally load it to be used in ML. For this task we can solve it using many ways, the easiest solution to use LLM such as GPT-4 to learn the model about the features by zero shot learning then test it to check the quality, we can also use open source models to do that, from my side I will do this by using free LLM model such as mistral:7b-instruct-v0.2-q4_K_S and the second option to do this task by using classical machine learning which means after scarping the data we will

process it and extract the best features for predictions the salary by using classical machine learning models, the main issue here that the data is textual data which means that we can use it directly to the classical machine learning models but this need many processing steps, these steps may also have many other options which I will explain them in details. After processing the data to be used in classical machine learning for a regression task, I searched for the best features that I can extract I found that: **Skills, Experience, Education Level, Sector, and Geographical location, I used also word embedding techniques with RNN-LSTM also BERT model, all of that failed unfortunately. I used another dataset with deep learning and got acceptable results.**

Objective:

Review current trends and developments in salary prediction research.

Summary of Findings:

1. **Features Impacting Salaries:**
    - I searched for the best features that I can extract I found that: **Skills, Experience, Education Level, Sector, and Geographical location,**
2. **Modeling Approaches:**
    - **I used Linear Regression, Tree-Based Models like** Random Forests and Boosted Trees, which was bad quality.
    - **Deep Learning:** I used RNN with LSTM for word embedding, also BERT, I used MLP with another dataset which I gained good results.
3. **Recent Trends:**
    - I created Streamlit app to scrap the data.
    - I used transfer learning to learn BERT about my data but the results was very bad.

Final model:
After trying many approaches I decided tat using classical machine learning better than some approaches like BERT, Word Embdeeings with RNN and LSTM, also another method may bay good which is using zero shot learning with llama,

so my final decision to use the new dtaset from the internet or just using the zero-shot learning with LLM like llama

---

Regression Metrics: Training Data

- MAE = 310.230
- MSE = 511,260.350
- RMSE = 715.025
- R^2 = 0.973

Regression Metrics: Test Data

- MAE = 458.173
- MSE = 1,030,904.107
- RMSE = 1,015.334
- R^2 = 0.947

Theres no overfit, the mean for the data 8659.778207 which means its acceptable to have some errors with 1,015.334 on Testing data

How to use this model?

1- Extract the data using LLAMA:

```
OLLAMA_ENDPOINT = "http://localhost:11434/api/generate"
OLLAMA_CONFIG = {
    "model": "mistral:7b-instruct-v0.2-q4_K_S",
    "keep_alive": "5m",
    "stream": False,
}

PROMPT_TEMPLATE = Template(
"""Extract the following details from the job description below,
- Years of Experience: Extract the required experience as a sing
```

```
- Education Level: Extract the education level choose one(choose
- Job Level: Extract the level of the role (choose from: Senior,
- Job Sector: Provide the job sector (choose one: Technology, He
- Gender: Extract gender (choose one: Male, Female, Other).
- Age: Extract age (choose a number from 18 to 100, or Other if

Job Description: $text

Respond in JSON format with the keys: "Job Title", "Years of Exp
As
- "Years of Experience"
- "Education Level"
- "Job Level"
- "Job Sector"
- "Gender"
- "Age"
For missing fields, use the value "Other".
"""
)

def extract_job_details(description):
    print("*"*10)
    prompt = PROMPT_TEMPLATE.substitute(text=description)
    response = httpx.post(
        OLLAMA_ENDPOINT,
        json={"prompt": prompt, **OLLAMA_CONFIG},
        headers={"Content-Type": "application/json"},
        timeout=240,
    )
    if response.status_code != 200:
        print(f"Error {response.status_code}: {response.text}")
        return None

    try:
        result = response.json()["response"].strip()
        return eval(result)
```

```
    except Exception as e:
        print("Error parsing response:", e)
        return {
            "Years of Experience": 0,
            "Education Level": "Other",
            "Job Level": "Other",
            "Job Sector": "Other",
            "Gender": "Other",
            "Age": 25,
        }
```

2- Process the data:

Convert the categories into: Education_Level = ['Other', 'High School', 'Diploma', 'Bachelor', 'Bachelor/Master', 'Master', 'Doctorate'] Level = ['Other', 'Senior', 'Manager', 'Junior', 'Mid'] Industry = ['Technology & Engineering', 'Miscellaneous', 'Sales & Marketing', 'Finance & Business Services', 'Creative & Entertainment','Healthcare & Pharmaceuticals', 'Retail & Consumer Goods','Hospitality & Tourism', 'Industrial & Manufacturing', 'Education & Services', 'Energy & Resources'] Anything out of this make it 'Other'.

For Numerical data use column.median for missing values.

3- Let the model to predict the results using rfmodel.predict(data)

Note: I used Level and Education Level as ordinal because here the order is important

We can deploy this inside basic Streamlit App to enter the text, then process it using LLAMA to extract data then process the results and predict it.

Job Description Analysis and Prediction

Enter the Job Description:

ALGOS LABS project for AASOS Academy, using React TS, JAVA SPRING, Mongo, and Postgres databases. Additionally, I have trained as a front-end React JS developer at EXALT Technologies and worked on a project in Oracle PL/SQL for one year, mainly in AI and Database for a web app. I aim to leverage my diverse and versatile skills to create value and positively impact the data science and machine learning community.

Analyze and Predict


```
{
  "Job Title" : "AI Engineer and Instructor"
  "Years of Experience" : 3
  "Expected Salary Range" : "Other"
  "Expected Salary Currency" : "USD"
  "Job Type" : "Other"
  "Location" : "Palestine"
  "Education Level" : "Master's Degree"
  "Major" : "Computer Engineering"
  "Job Level" : "Other"
  "Job Function" :
  "AI/Machine Learning Specialist, Instructor, Full-Stack Developer"
  "Job Sector" : "Engineering & Technology"
  "Gender" : "Other"
  "Age" : 25
}
```


Extracted Details:


```
{
  "Job Title" : "AI Engineer and Instructor"
  "Years of Experience" : 3
  "Expected Salary Range" : "Other"
  "Expected Salary Currency" : "USD"
  "Job Type" : "Other"
  "Location" : "Palestine"
  "Education Level" : "Master's Degree"
  "Major" : "Computer Engineering"
  "Job Level" : "Other"
  "Job Function" :
  "AI/Machine Learning Specialist, Instructor, Full-Stack Developer"
  "Job Sector" : "Engineering & Technology"
  "Gender" : "Other"
  "Age" : 25
}
```

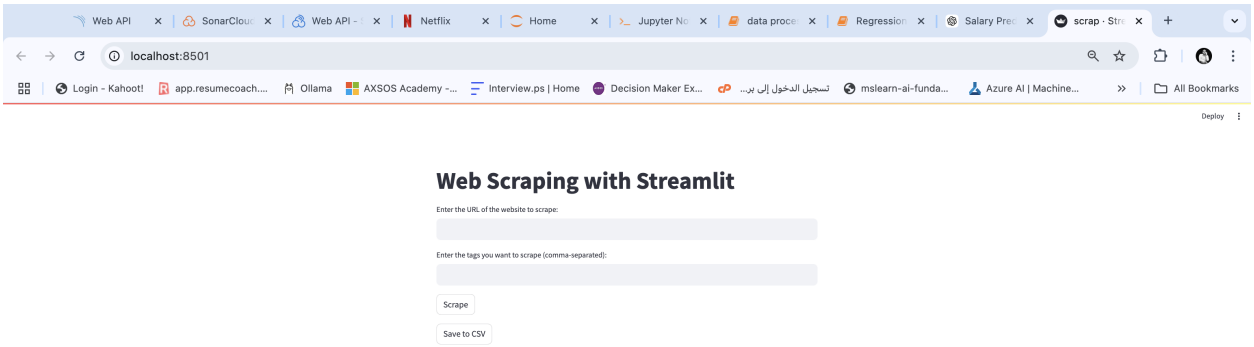

Preprocessed Features for Model:


```
{
  "Job Title" : "AI Engineer and Instructor"
  "Years of Experience" : 3
  "Expected Salary Range" : "Other"
  "Expected Salary Currency" : "USD"
  "Job Type" : "Other"
  "Location" : "Palestine"
  "Education Level" : "Master's Degree"
  "Major" : "Computer Engineering"
  "Job Level" : "Other"
  "Job Function" :
  "AI/Machine Learning Specialist, Instructor, Full-Stack Developer"
  "Job Sector" : "Engineering & Technology"
  "Gender" : "Other"
  "Age" : 25
}
```


Predicted Job Level: 6271.69521434271

For data scraping:

You can use this basic app I created to scrap the data.



Using zero-shot learning with LLAMA:

Job Description Analysis and Prediction

Enter the Job Description:

ALGOS LABS project for AXSOS Academy, using React TS, JAVA SPRING, Mongo, and Postgres databases. Additionally, I have trained as a front-end React JS developer at EXALT Technologies and worked on a project in Oracle PL/SQL for one year, mainly in AI and Database for a web app. I aim to leverage my diverse and versatile skills to create value and positively impact the data science and machine learning community.

Analyze and Predict

Predict Salary from Ollama Model

Based on your extensive experience as a data scientist, machine learning engineer, front-end developer, and Oracle PL/SQL project specialist, as well as your expertise in various AI frameworks, full-stack development, and teaching, I would estimate your salary to be around $120,000—150,000$ per year. However, this is just an estimation and the actual salary may vary depending on several factors such as location, industry, company size, and specific job responsibilities.