IE 7374 MLOps - Project Scoping

February 2024

1 Team Members:

1. Niloufer Syed
2. Susheel Kumar Yadav Konda
3. Vamshi Krishna Konyala
4. Veda Upasan Pedeagadi
5. Venkateshwaran Sundar
6. Zizheng Wang

2 Introduction

Our project will leverage the Data Science Job Postings & Skills (2024) dataset available on Kaggle to build a model that uncovers insights into the skills required, preferred companies, and salary ranges for various roles within the field. Despite the dataset containing approximately 12,000 data points—a modest size—it presents a valuable opportunity to apply the concepts taught in our Machine Learning Operations course. Our aim is to construct an efficient pipeline that adheres to best practices in data and code versioning, incorporates effective logging, and leverages data visualization techniques.

This pipeline will be designed to be adaptive, accommodating updates to the training data, supporting continuous training, and adjusting for data drift over time. We also plan to explore feature engineering opportunities within job descriptions and consider integrating data from the Jobs and Salaries in Data Science dataset to enhance our salary insights analysis.

In summary, by synthesizing these two valuable data sources through a robust modeling approach, we aim to generate actionable insights into the data science job market—a topic of significant relevance and interest to all students enrolled in the course. We'd like demonstrate effective use of concepts of data versioning, logging, data visualization and continuous learning/deployment in this project.

3 Data card

3.1 Dataset introduction

This dataset provides a raw dump of data science-related job postings collected from LinkedIn. It includes information about job titles, companies, locations, search parameters, and other relevant details. The main objective of this dataset is not only to provide insights into the data science job market and the skills required by professionals in this field but also to offer users an opportunity to practice their data-cleaning skills. By working with this dataset, users can gain hands-on experience in cleaning and preprocessing raw data, a critical skill for aspiring data scientists.

3.2 Data Card

There are four files in this dataset

1. job_postings.csv
2. job_skills.csv
3. job_summary.csv
4. jobs_in_data.csv

3.3 Dataset Overview

This section provides an overview of the datasets used in the project, detailing their content and potential applications.

3.3.1 job_postings.csv

Records: 12,218 **Columns:** 15
This dataset encompasses a comprehensive array of features including job titles, company names, locations, and search parameters among others. It serves as a primary source for understanding the job market landscape.

3.3.2 job_skills.csv

Records: 12,218 **Columns:** 15
This file is pivotal for analyzing the skill sets required for various job postings. It links directly to specific jobs and outlines the associated skills, offering a lens into the qualifications demanded by employers.

3.3.3 job_summary.csv

Records: 12,218 **Columns:** 15
Containing links to job postings along with concise summaries, this dataset presents an opportunity for Natural Language Processing (NLP) applications to distill and categorize job-related information effectively.

3.3.4 jobs_in_data.csv

Records: 9,356 **Columns:** 12
Rich in critical details such as job descriptions, experience levels, salaries, locations, and company sizes, this dataset is slightly smaller yet highly relevant. It is instrumental for correlating job postings with salary data through exploratory data engineering.

3.4 Data Sources

Both datasets have been sourced from Kaggle, ensuring the absence of Personally Identifiable Information (PII) and adherence to data privacy norms. The compiled data originates from publicly accessible LinkedIn profiles, thereby simplifying its aggregation for model training purposes.

3.4.1 Data Rights and Privacy

Given the public availability of the information contained within these datasets, there are minimal concerns regarding data rights and privacy. However, it is crucial to acknowledge that despite the data's public nature, ethical considerations in its use and the potential impact on individuals described by the data should be carefully evaluated.

4 Data Planning and Splits

Our dataset encompasses 12,218 data points, distributed across three files that provide comprehensive details on job postings, skills required, and job descriptions. A separate dataset delivers insights into roles, experience levels, and salary information. Our initial step involves devising a methodology for leveraging both datasets effectively, with a particular focus on identifying and engineering relevant features.

In terms of dataset partitioning for model development:

- The **training set** will consist of approximately 8,000 records. This subset is designated for the initial training phase of our model, enabling it to learn and adapt to the patterns within our data.
- The **testing set** will comprise around 4,000 records. This portion is reserved for evaluating the model's performance, ensuring its ability to generalize and make predictions on unseen data.

Our strategy is aimed at maximizing the efficiency and accuracy of the model by carefully balancing the training and testing phases, thereby facilitating a robust assessment of our model's predictive capabilities.

5 GitHub Repository

IE 7374 Machine Learning Operations - Project Scoping :
https://github.com/VenkyGitRep/MLOps_Project_Scoping

6 Project Scope

6.1 Problems

6.1.1 Integrating Datasets

The challenge lies in correlating job postings—detailing skills, job summaries, companies, and locations—with salary and experience level data from a secondary dataset. Our objective is to seamlessly merge these datasets to provide comprehensive insights.

6.1.2 Identifying Target Variables

Post data cleansing, exploration, and Principal Component Analysis (PCA), our goal is to understand the data's structure and pinpoint viable target variables for model prediction.

6.1.3 Developing Outputs

Aim to implement a user interface that accepts input, such as a student's profile or specific details, to recommend suitable roles, potential salaries, or job opportunities.

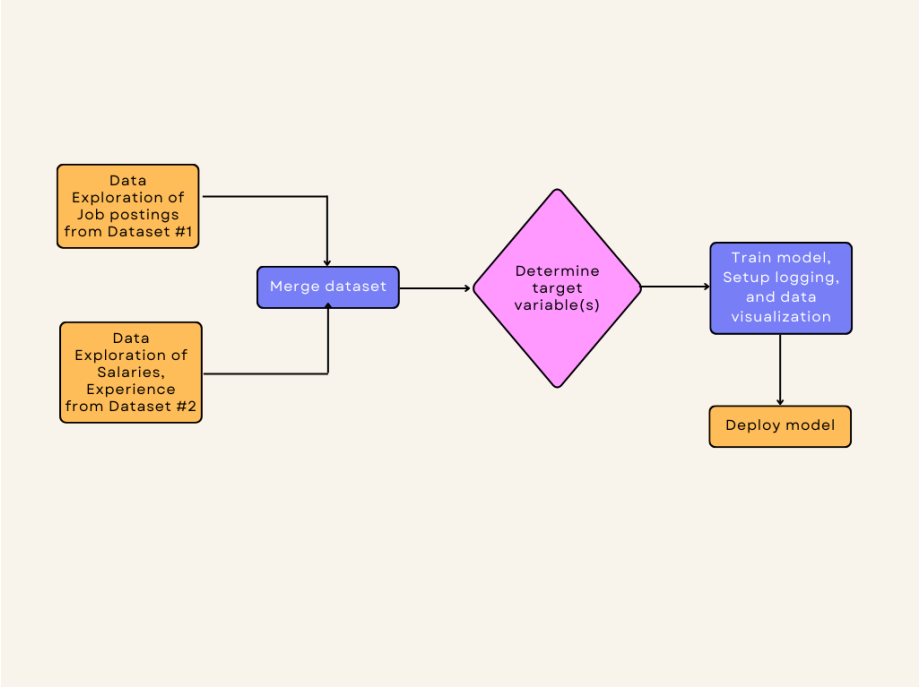
6.2 Current Solutions

Platforms like LinkedIn and Indeed already offer similar functionalities. Our project aspires to replicate an embryonic version of such features, providing foundational insight into potential career paths.

6.3 Proposed Solutions

The project intends to develop a prototype offering rudimentary functionality akin to that found on established job search platforms.

7 Current Approach Flow Chart and Bottleneck Detection



Initial efforts are focused on extracting actionable insights from the integrated datasets, as highlighted in the problems section.

8 Metrics, Objectives, and Business Goals

8.0.1 Objective 1

Establish a robust pipeline that supports version logging of datasets, efficient retraining, and the creation of training parameter snapshots to facilitate collaboration.

8.0.2 Objective 2

Determine valuable target variables through comprehensive data analysis, enhancing the model's learning potential and applicability.

8.0.3 Objective 3

Deploy a model complemented by a user interface capable of predicting suitable job roles for candidates based on their profiles.

9 Failure Analysis

1. If dataset integration proves ineffective, the focus may shift to utilizing a single dataset or exploring a new objective with a more substantial dataset.
2. Should the model fail to predict useful outcomes, further feature engineering may be necessary to align candidate profiles with appropriate job opportunities.

10 Deployment Infrastructure

The Google Cloud Platform, chosen for its provision of course-related free credits, alongside MLflow, Kibana, and Logstash, will support the project's infrastructure needs. Specifics regarding the infrastructure for model training and deployment remain under development.

11 Monitoring Plan

To ensure the effectiveness and efficiency of our model, we have devised a comprehensive monitoring plan:

- Continuous monitoring of each inference made by the model, coupled with real-time visualization of these inferences, to track the model's performance and adjust as necessary.
- Regular oversight of model performance metrics and log data, allowing for prompt identification and resolution of any issues.

12 Success and Acceptance Criteria

The project will be deemed successful upon the establishment of a fully operational pipeline. The hallmark of success will be the pipeline's ability to incorporate new data, execute preprocessing, and facilitate learning processes with minimal manual intervention—an essential benchmark for efficiency and automation.

13 Timeline Planning

Our project timeline is strategically organized to optimize progress and achieve key milestones within specific time frames:

- **Spring Break**: Dedicated to data exploration and clarifying the model's objectives to ensure a solid foundation for subsequent phases.
- **20 March**: Completion of the pipeline construction and activation of a baseline model to initiate the learning process.

- **31 March**: Implementation of logging mechanisms and deployment functionalities through the pipeline, enhancing operational transparency and efficiency.
- **15 April**: Evaluation of three distinct models to ascertain the most effective approach based on performance metrics.
- **20 April**: Development of a user interface (UI) that enables query processing and model inference in a production environment, marking the project's transition into the operational phase.

This timeline represents a structured approach to achieving the project's goals, with each phase designed to build upon the previous one, culminating in a comprehensive and functional solution.