Job Transition Pathway: Navigating Career Transitions with Precision

Deekshita Prakash Savanur, Gouri Benni, Uzair Riyaz Pachhapure

Department of Applied Data Science, San Jose State University

DATA 240 - Sec 12: Data Mining/Analytics

Dr. Shayan Shams

December 07, 2023

Motivation

The modern workforce is dynamic, and individuals must constantly modify their knowledge and abilities to meet the evolving needs of their professional settings. Effective tools for skill development and career planning are essential in this dynamic environment, especially in light of the growing significance of career flexibility and lifelong learning. The project was motivated by the acute knowledge of the difficulties professionals face when attempting to navigate their career trajectories. It is becoming more difficult to establish clear and strategic avenues for career advancement and personal development due to the expansion of varied job categories and the quick evolution of essential skill sets. By building the ML model that can forecast career advancement trajectories and recommend pertinent educational courses, this research attempts to close this gap. The main goal of this project is to build a data-driven, practical solution that will meet the demand for individualized career counseling in the quickly evolving professional landscape.

Background Information

A prominent trend in the job market today is career switching, which is a reflection of the changing nature of work, the need for certain skills, and individual goals. This pattern is noticeable in a range of age groups and is impacted by things like personal development and financial security. The frequency of job changes decreases with age; individuals between the ages of 25 and 34 change jobs roughly 2.4 times, while those between the ages of 35 and 44 change employment roughly 2.9 times. A significant percentage of younger people, especially those under 25, have a propensity to reevaluate their present employment circumstances. Of this group, almost 87% are thinking about switching careers. (Kurtuy, 2023). These figures highlight the need for flexibility in job routes and the expanding trend of labor mobility which is addressed in our project. They draw attention to the need for tools and resources that can help people through these changes and customize their professional pathways to meet their goals, the needs of the market, and their changing skill set.

Literature Review

Baldwin et al. (2022) investigate the enhancement of job-skill representation through the use of a transformer-based model that was trained on online job advertising. The model outperforms previous models in the transfer of job taxonomy. Our project, on the other hand, stands out since it focuses on creating a seq2seq model with a bidirectional LSTM and attention mechanism for predicting educational courses and career advancement.

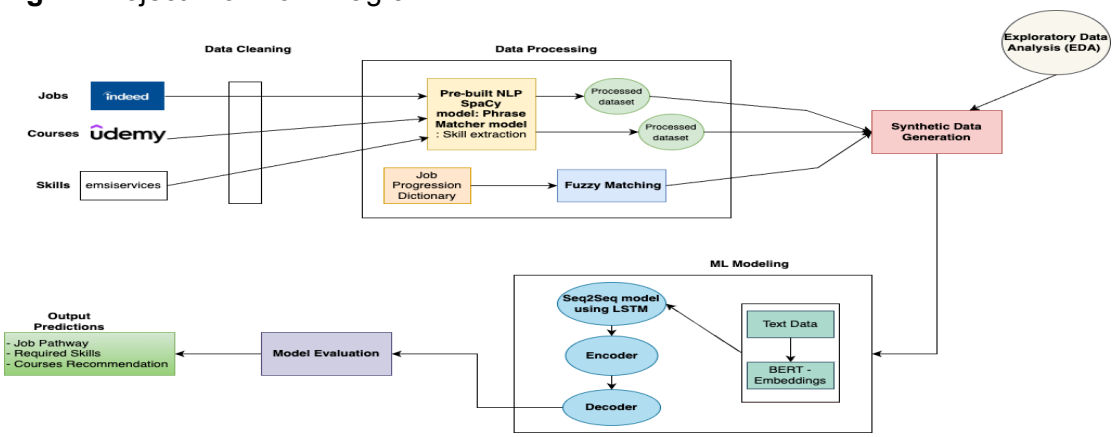
Using work transition patterns for job recommendations, the study by Lee et al. (2017) addresses the importance of job ordering in career paths and user-item matrix sparsity. It provides a more comprehensive approach to professional development that goes beyond job transitions, in contrast to our project, which concentrates on forecasting career progression trajectories and recommending educational courses.

The goal of the study by Paparrizos et al. (2011) is to utilize machine learning, which has been trained on online employee profiles, to forecast job transitions. It places a strong emphasis on precisely predicting a person's next career change, showing a notable advancement over baseline models. In comparison, our project uses a seq2seq model to forecast career trajectories while also integrating recommendations for educational courses and skill gap analysis. This takes a more comprehensive approach to career development.

The study by Feng and Zhu (2016) offers a thorough analysis of trajectory data mining methods and applications, with a focus on technical approaches and their application in a range of contexts, including movement behaviour analysis and location prediction. It provides an organized synopsis of trajectory mining techniques and data collection methodologies. This work focuses mostly on the technical aspects of trajectory data mining, without a specific focus on individual skill development or educational assistance. This is in contrast to our project, which blends skill gap analysis and educational suggestions in career path prediction. Obstacles like maintaining data representativeness and striking a balance between interpretability and model complexity offer important lessons for future study.

Methodology

Fig 1 - Project Workflow Diagram



Data Collection

The datasets were collected for 3 sections namely Jobs, Skills, and Courses. For the jobs section, the data was web-scraped from the Indeed website, an employment website that contains job listings. Several job titles were selected and their page was webscraped and all job title data was further aggregated. It contains features such as titles, job descriptions, urgently hiring, and others. For the Courses section, the data was web-scraped from Udemy, an educational tech company that serves online learning & teaching platforms. Its features include ID, title, URL, price details, and others. For the skills dataset, the data was webscraped from the Emsiservices website and its features include ID, name, category, and others.

Data Cleaning & Exploratory Data Analysis (EDA)

The data cleaning process was applied to three different datasets namely, Skills, Jobs, and Courses. The Skills Dataset underwent lemmatization and stopword removal, conversion of skills data into lists, and manual addition of specific skills including “Bard ai” and “Machine Learning”. The Jobs Dataset was similarly preprocessed with lemmatization, conversion of salary data from dictionary format to columns, job description conversions to lists, and treating of missing values (NaN). For the Course Dataset, HTML tags were removed from the course descriptions, ensuring that the textual data is clean and uniform across all datasets for subsequent analysis. We then performed EDA on the data.

Data Processing (Mining)

The data processing was structured to extract, clean up, and correlate data from employment, skill, and course datasets. To accurately identify the skill sets, the Phrase Matcher Model was applied to extract skills from job descriptions and course descriptions. A job progression dictionary was created from the jobs dataset to generate job titles and pathways, guaranteeing consistency and coherence within the dataset. Fuzzy matching was applied, which normalizes job title variations against a set of well-defined titles (e.g.,'sr' abbreviations to 'Senior') further refines the cleaning of job titles. In parallel, the abilities found in job descriptions were compared to the skills taken from course descriptions to link them. Advanced methods including cosine similarity measurements and TF-IDF for text vectorization were included in the data processing, which aided in suggesting courses that closely matched the necessary skill set. The simulation of career trajectories and focused skill development benefited greatly from this methodical processing, which laid the groundwork for the latter phases of modeling and analysis.

Synthetic Data Generation

The data generation technique describes a methodical process for developing synthetic career trajectories and related learning recommendations. The procedure begins

with a job role being selected at random and its respective required skills from the dataset. Subsequently, it predicts the next job role, either by randomly selecting (if no path exists) or by following a predetermined career advancement path. The skills needed for this next job are then evaluated and compared to the current skill set to find any gaps. The method uses cosine similarity to make sure the courses closely match the required abilities and TF-IDF for text vectorization of course descriptions is utilized to fill in these gaps by searching a course dataset for courses that offer the requisite skills. It ranks every course according to how relevant it is to the skill gap, giving priority to the most relevant courses. To create a multi-step professional development route with targeted course suggestions, this iterative procedure repeats, replicating sequential job changes and skill requirements.

ML Modeling

Model: Sequence-to-Sequence Model Using LSTM Layers

Seq2Seq Model: The architecture translates sequences from one domain to another and is crucial to NLP. It's a perfect fit for the project's requirements because it works especially well for activities where the output length fluctuates with regard to the input. Our project makes use of NLP pioneering BERT embeddings in addition to the Seq2Seq models using LSTM layers to provide a profound comprehension of language context and semantics.
LSTM Networks: LSTM layers are incorporated into the Seq2Seq model and were selected due to their ability to capture long-term dependencies in sequential data. To effectively model career paths and skill progression, the capacity for long-term memory is essential.
BERT Functionality: Unlike conventional models that read text sequentially, BERT processes words in relation to all other words in a phrase & is perfect for comprehending intricate job responsibilities and skill descriptions as it allows it to take subtle meanings and contexts.

Utilization:
Modeling Employment Transitions: Based on a series of previous roles and experiences, the model is used to forecast possible future employment roles.
Skill Progression Prediction: In a similar vein, the model forecasts how skill requirements specific to each career path will change by examining the order in which skills are acquired.
Course Recommendation Engine: The model efficiently suggests courses and learning routes that correspond with anticipated career improvements by comprehending the sequential relationship between employment job titles and skills.

The model necessitates a structured dataset with sequential job roles, skills, and courses, which undergoes tokenization, normalization, and padding post-synthetic data generation to align with LSTM input specifications.
Architecture Model: The encoder and decoder of the Seq2Seq paradigm are both outfitted with LSTM layers. The input sequence, such as past job roles, is processed by the encoder, while the output sequence, such as upcoming job roles or talents, is produced by the decoder.
Training and Validation: To avoid overfitting, the number of LSTM units, learning rate, and dropout rate are fine-tuned when the model is trained using historical data. A different dataset is used for validation to guarantee the generalizability of the model.

Challenges and Resolutions:
Handling Sparse Data: Individualistic career paths often result in trajectories with sparse data. To lessen this, strategies like transfer learning and data augmentation are used.
Dynamic Industry Trends: The model must be flexible due to the quickly evolving nature of skill requirements and job marketplaces. To keep the model relevant, it must be updated frequently and retrained with new data.

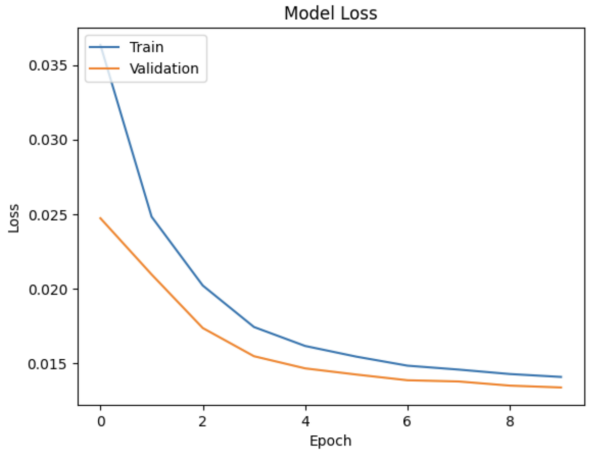
Experiments
Having a contextual representation of text, BERT embeddings, as opposed to one-hot encoding improved the input data for our model. BERT encodes semantic meaning and links between words based on the context, in contrast to one-hot encoding, which produces sparse and high-dimensional vectors. This was essential for correctly analyzing skills data and job descriptions, which enhanced the model's predictive power and facilitated the recommendation of pertinent courses.

Initially Bidirectional LSTM model was explored and tried but the accuracy results were not up to the mark. Therefore, the Seq2Seq model with LSTM was used as the accuracy was better on the model prediction for job title progression that was not included in the dataset (Generative capability was higher).

Fig 2- Model Summary of Seq2Seq Model with LSTM

```
Model: "model"
__________________________________________________________________________________________________
 Layer (type)                Output Shape              Param #     Connected to
==================================================================================================
 input_2 (InputLayer)        [(None, None)]            0           []

 input_1 (InputLayer)        [(None, 2, 768)]          0           []

 embedding (Embedding)       (None, None, 256)         9984        ['input_2[0][0]']

 lstm (LSTM)                 [(None, 256),             1049600     ['input_1[0][0]']
                              (None, 256),
                              (None, 256)]

 lstm_1 (LSTM)               [(None, None, 256),       525312      ['embedding[0][0]',
                              (None, 256),                         'lstm[0][1]',
                              (None, 256)]                         'lstm[0][2]']

 dense (Dense)               (None, None, 39)          10023       ['lstm_1[0][0]']

==================================================================================================
Total params: 1,594,919
Trainable params: 1,594,919
Non-trainable params: 0
```

Fig 3 - Model Loss
The model's accuracy in predicting outcomes improves as it learns from the data, as seen by the model loss graph, which shows the decline in training and validation loss over epochs. We fine-tuned the model with various epoch counts but it showed the best result with 20 epochs.



Model Evaluation

The accuracy of the model in forecasting the next work role and the necessary abilities for that role are measured as part of the evaluation process. In skill prediction, the model's ability is gauged to anticipate the skill gap between the current role and the next, and for job prediction, the advised job transitions are compared by the model against a collection of established career progressions. The model's accuracy of 51.43% in predicting the next job position and 68.84% in predicting the next set of abilities needed for that job respectively was used to evaluate the performance. The BLEU score for next job predictions is 51.43%, while the prediction of next skills achieved a higher score of 68.84%. These metrics give an objective assessment of the model's predictive ability. The accuracies show how effectively the model has assimilated the training set.

Project Results

Fig 4 - The resultant view of the model generated output of the job transition pathway with the skills required and the course recommendation.

Data Engineer was given as input to the model and the predicted progression pathway includes :
Data Engineer -> Senior Data Engineer -> Data Architect -> Data Manager -> Director of Analytics -> Chief Data Officer. The corresponding skills required for their jobs and the top 10 predicted course categories with predicted course titles are displayed for each row.

	Predicted Job Title	Predicted Skills	top_10_predicted_course_category	all_skills_learnt	top_10_predicted_course_title
0	data engineer	python, sql, computer science, database design, data quality, data analysis, data collection, data management, data mining	{data science machine learning data analytics, data mining, improving data quality data analytics machine learning, intro data data science, introduction data science, getting started data management, python data science learn data science scratch, datascience machine learning nlp python r bigdata pyspark, data collection beginners, data science python complete guide}	{data quality management, data mining, database management, data science}	{data quality management, data mining, database management, data science}
1	senior data engineer	python, sql, computer science, azure blob storage, data pipelines, data integration, analytics, apache spark, azure data factory, databricks, data engineering	{azure cloud azure databricks apache spark machine learning, azure data engineer workshop weekend, microsoft azure databricks data engineering, machine learning, real world azure data engineer project end end, python sdk azure bootcamp, azure databricks build data engineering ai ml pipeline, azure data factory data engineering cloud, azure databricks spark sql python, apache spark master big data pyspark databricks}	{azure synapse analytics, azure data factory, undefined, apache spark, databricks, microsoft azure}	{azure synapse analytics, azure data factory, undefined, apache spark, databricks, microsoft azure}
2	data architect	ci cd, data modeling, sql, data strategy, data architecture, business analysis, business analysis, data quality, data governance, analytics, data management, pl sql	{data science machine learning data analytics, data scientist sql tableau ml dl, python data analytics beginner advanced, python sql, introduction data science, getting started data management, sql, complete sql bootcamp data analysis level, sql data visualization complete bootcamp}	{oracle sql, data science, database management, big data, data analysis, sql}	{oracle sql, data science, database management, big data, data analysis, sql}
3	data manager	quantitative analysis, sql, data access, computer science, statistical analysis, data driven decision making, business requirements, analytical techniques	{statistics data analytics, data science machine learning python, data science machine learning data analytics, data analyst python beginners, python sql, introduction data science, data science interview questions answers, r data analysis statistics data science, master data analysis pandas, data science r}	{sql, undefined, data science, data analysis}	{sql, undefined, data science, data analysis}
4	director of analytics	analytics, project management, data science	{complete data science project management course, data analytics python projects hindi, statistics data science business analytics python, introduction data science, getting started data management, data analytics python, practical guide alteryx data science analytics, introduction data science python module, data science complete beginners, cdmp metadata specialist exam questions practitioner master}	{data mining, data science, database management, data analysis, automation, undefined}	{data mining, data science, database management, data analysis, automation, undefined}
5	chief data officer	predictive analytics, data modeling, data strategy, data strategy, problem solving, data quality, data governance, data governance, analytics, business requirements, data management, business intelligence, cloud storage, data literacy, data engineering, big data, business performance management	{x data management governance security ethics masterclass, python data analytics beginner advanced, data science fundamentals, probability statistics data science, intro data data science, introduction data science, getting started data management, big data, cdmp metadata specialist exam questions practitioner master, data analytics crash course}	{data science, data quality management, big data, database management, data analysis, undefined}	{data science, data quality management, big data, database management, data analysis, undefined}

Discussion

The research examines critically its all-encompassing approach to career development. Our approach utilizes Indeed, Udemy, and Emsi data to anticipate job paths and required skill sets, and then suggests courses that are relevant to those job pathways. By using Spacy for reliable NLP processing, the model's comprehension was improved and parsing of intricate job descriptions. Nuanced textual data was captured by integrating seq2seq LSTM models with BERT embeddings, which is essential for the precision of the predictions.

Future Improvement

Future project improvements will include a range of learning tools and opportunities for real-world experience to diversify recommendations. It will expand the dataset to include a greater range of employment roles and industry representation and make use of real-time labor market data for dynamic trend response. While a sophisticated recommendation engine personalizes ideas based on user data, advanced natural language processing algorithms increase knowledge of job advancement. User feedback loops will improve the precision and applicability of recommendations, and continuous learning mechanisms will eliminate the need for periodic retraining.

References

Baldwin, T., Clarke, W., Macedo, M. M. G., De Paula, R., & Das, S. (2022). Better skill-based job representations, assessed via job transition data. *2022 IEEE International Conference on Big Data (Big Data)*. https://doi.org/10.1109/bigdata55660.2022.10021087

Lee, Y., Lee, Y., Hong, J., & Kim, S. (2017). Exploiting job transition patterns for effective job recommendation. *IEEE Conference Publication | IEEE Xplore*. https://doi.org/10.1109/smc.2017.8122984

Paparrizos, I., Cambazoğlu, B. B., & Gionis, A. (2011). Machine learned job recommendation. *Fifth ACM Conference on Recommender Systems (RecSys '11)*. https://doi.org/10.1145/2043932.2043994

Feng, Z., & Zhu, Y. (2016). A survey on Trajectory Data Mining: Techniques and Applications. *IEEE Access*, *4*, 2056–2067. https://doi.org/10.1109/access.2016.2553681

Kurtuy, A. (2023, January 4). *60+ career change statistics for 2022 [That you didn't know!]*. Novorésumé. https://novoresume.com/career-blog/career-change-statistics#:~:text=Nearly%209%20out%20of%2010,age%20of%2039%20years%20old