Capstone Project Proposal

Team members

Asia Paige
Thomas James
Chauncey Raggie

asiam@umich.edu, tjamesum@umich.edu, craggie@umich.edu

Team Name

"Team ACTion"

Motivation

With so many paid and free courses across various platforms on the internet, it can be difficult to ensure you're making the best investment of your time from a relevance perspective and from a return perspective. The proposed project is a course recommender system that takes in job attributes (title, location, company etc.), predicts the salary and recommends relevant courses. The tool informs job seekers about the salary expectations for a given position and provides relevant courses to acquire the skills for the position in one location. Currently, this process would require multiple searches across websites. The system mitigates the information and cognitive overload problem by suggesting related and relevant content to the users.

Dataset(s) we plan to use

Dataset name: Udemy Courses
Source URL:
https://www.kaggle.com/datasets/andrewmvd/udemy-courses?resource=download
Access method: Download from Kaggle.com
Key features: 'course_title': name of the course, 'url': link to course, 'price': cost of course, 'level': difficulty of course.
Restrictions: None

Dataset name: US Jobs on Monster.com
Source URL: https://www.kaggle.com/datasets/PromptCloudHQ/us-jobs-on-monstercom
Access method: Download from Kaggle

Key Features: ‘job_description’ ,‘job_title’,‘job_type’,‘location’,‘organization’,‘page_url’, ‘salary’, ‘sector’
Restrictions: None

Dataset name: Glassdoor-Analyze Gender Pay Gap
Source URL:
https://www.kaggle.com/datasets/nilimajauhari/glassdoor-analyze-gender-pay-gap
Access method: Download from Kaggle
Key features: ‘JobTitle’: Name of job, ‘BasePay’: salary of job.
Restrictions: None

Dataset name: Coursera
Source URL: https://www.coursera.org
Access method: Download from Coursera
Key features: ‘specialization title’: Course title, ‘project courses’: ‘online degrees’.
Restrictions: None

Dataset name: May 2021 National Occupational Employment and Wage Estimates United States
Source URL: https://www.bls.gov/oes/current/oes_nat.htm
Access method: Download from Occupational Employment and Wage Estimates website
Key features: ‘Occupation title’: name of job, ‘Annual mean wage’: annual mean wage for job, ‘Employment’: Number of U.S. Citizens employed in this role overall.
Restrictions: None

Dataset name: EdX Courses Dataset 2021
Source URL: https://www.kaggle.com/datasets/khusheekapoor/edx-courses-dataset-2021
Access method: Download from Kaggle
Key Features: ‘Name: name of course, ‘Difficulty’: course difficulty level, ‘Link’: url link to course, ‘About’: short description of course.
Restrictions: None

Should more data be required, platform APIs will be utilized to acquire additional data.

There are no restrictions pertaining to the datasets we wish to use. Should we come across any datasets that prohibit redistribution for educational purposes we will drop this dataset.

Minimum viable product (MVP):

Our MVP is a model that predicts salary with an R squared value of above 0.75 and a recommender system that passes our ground truth and smoke test.*

*Since we will be evaluating our recommender systems in absence of labeled data, we won't know what recommendations are actually optimal. We may try to do ground truth tests, i.e. providing a small number of job titles we believe are similar to see if the recommender system produces the same results. This could also be combined with "smoke tests" where we measure the overlap between the input and output samples in certain key features and determine if we have passed a specific threshold, say average feature overlap of 50%.

Ethical challenges or concerns

There are no currently known ethical concerns for this particular project.

Anticipated technical challenges

Our evaluation step of the recommendation system component may prove challenging given that we will not have ground truth labels.

Evaluation outcomes

A working predictive model performing up to the standard stated above and working recommendation model performing up to the standard stated above.

Planned contributions of each team member

EDA/Dataset Creation: Asia
Project Manager: Asia
Lead Visualizer: Chauncey
Lead Report/Blog Writer: Chauncey
Lead Model Evaluator: Thomas
Lead Model Developer: Thomas

Tentative schedule

Activity /Milestone	Start Date	End date
Data acquisition and set up of github repository	Feb 12, 2023	Feb 21, 2023

Data preparation and cleaning	Feb 21, 2023	Feb 28, 2023
Data exploration	Feb 28, 2023	March 7, 2023
Model Development and Evaluation	March 7, 2023	March 14, 2023
Model Deployment (front end app creation if time permits)	March 14, 2023	March 21, 2023
Report/Blog writing	March 21, 2023	March 28, 2023
Video Preparation	March 28, 2023	April 4, 2023
Project submission - GitHub repository reviewed and finalized, blog report and 3-5 min summary video prepared	April 4th, 2023	April 11th, 2023