Predicting likelihood of a `successful job match’

Sprint 1 - Nivedita Prasad

Introduction

What's the problem area?

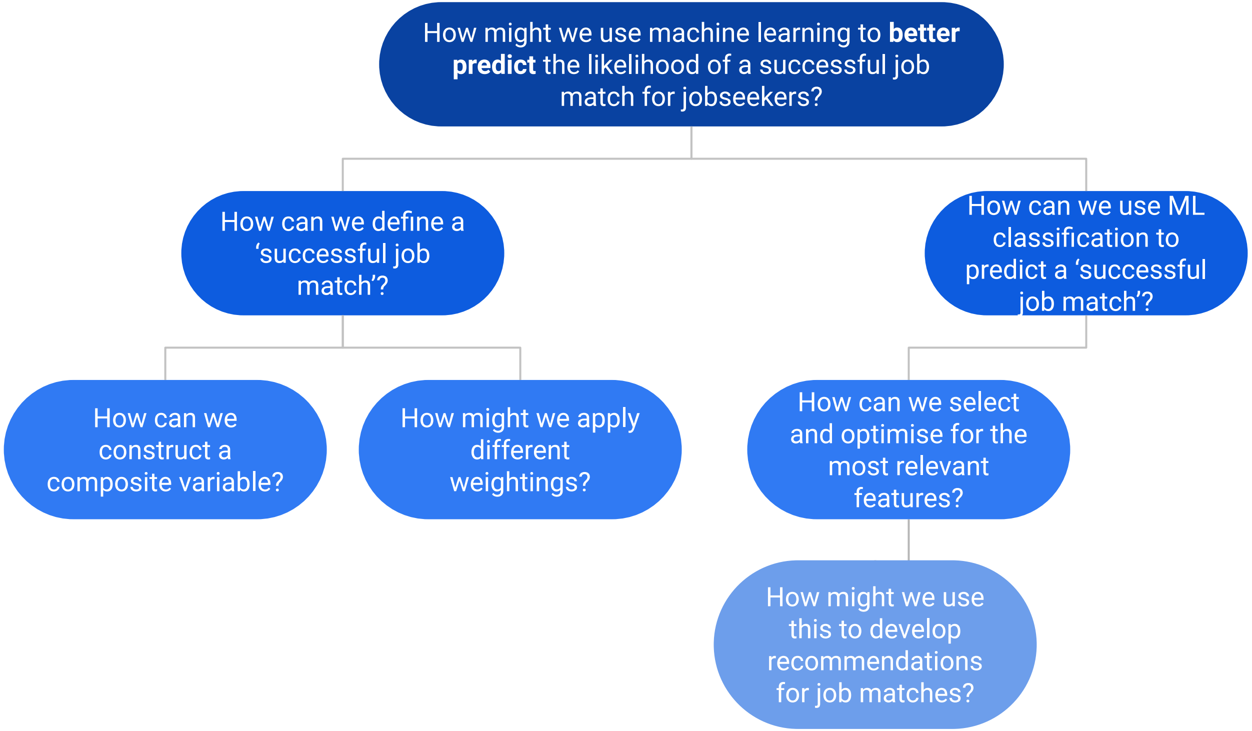
- Job matching continues to be a critical issue in today's labour market (and something many of us can relate to, more personally!).
- We talk a lot about the skills we need to be the right fit for a job, our work-life balance, and levels of financial wellbeing we aspire to.
- But it continues to be challenge to get the 'right fit', as we know from stark unemployment figures*

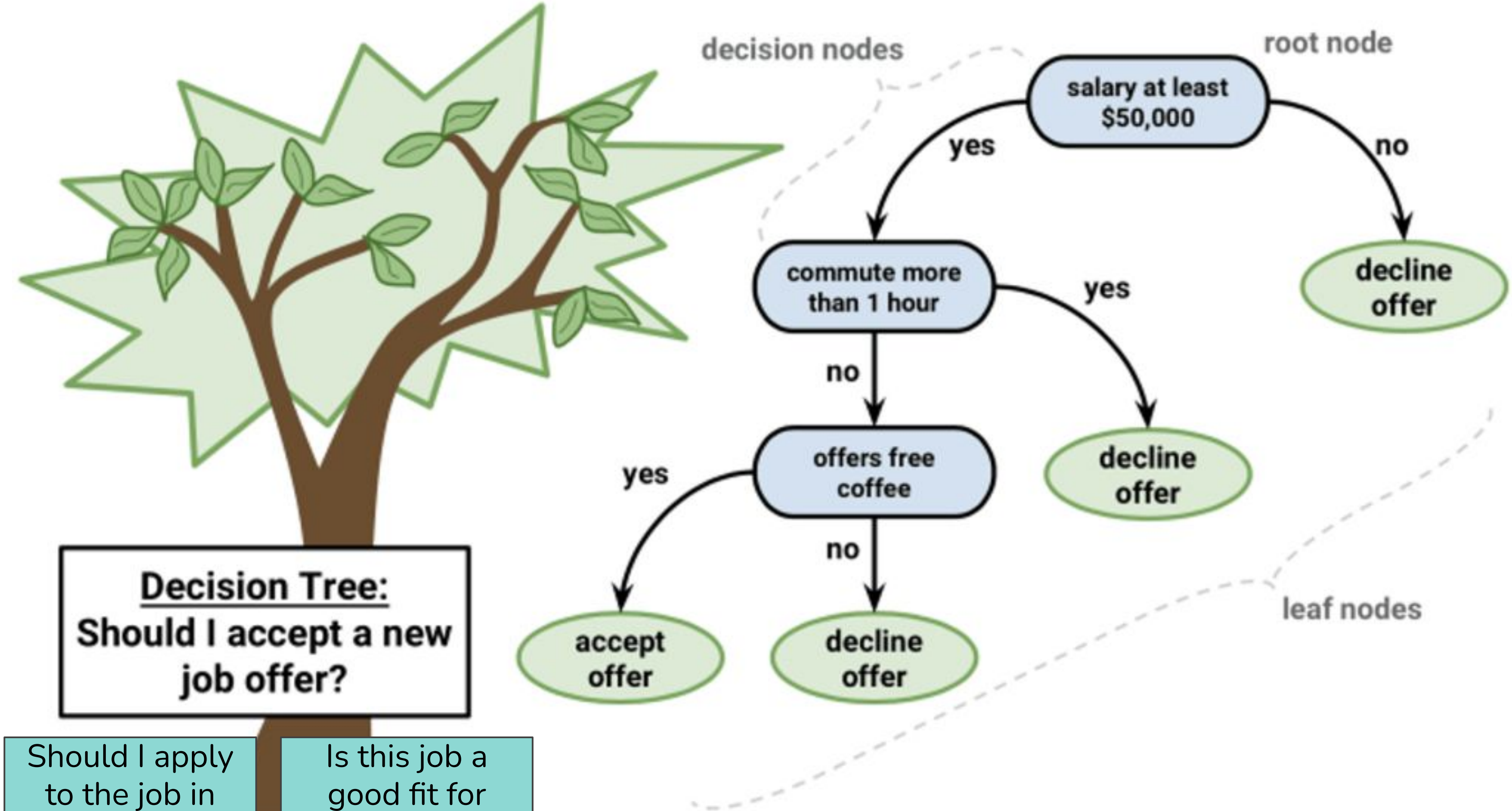
How can we go beyond usual notions of 'job matching'?

- Often studies and articles focus on salary as the key outcome for a successful job match
- **We can go beyond this** to consider aspects of job satisfaction, career growth and skills utilisation
- This leads us to our problem statement...

** Illustrated by these stark labour market figures relating to youth unemployment, and long-term unemployment - e.g., over a year of youth unemployment can lead to a wage reduction of around 30% for men and 15-20% for women. And data on elements of job satisfaction, showing lower satisfaction with 'opportunities for promotion'*

Breaking down our problem statement





Should I apply to the job in the first place?

Is this job a good fit for me?

The National Survey of College Graduates (U.S)

What is it?

- Recurring survey (every 2-3 years), conducted by the National Science Foundation
- Each dataset is a snapshot of the U.S. college graduate population, at a specific point in time
- Given the richness of the data - this focuses our scope **on college graduates**
- Raw data is usually provided as SAS data files; so main task was in importing and renaming variables

Why is it useful for this project?

- It's used to assess trends in the labor market
- We get insights into the experiences and outcomes of college graduates, like **salary, job satisfaction, demographics**. Often unavailable in other datasets, or too aggregated (as found in the UK data)
- We'll focus on three timepoints of survey data from **2017, 2019 and 2021**

For Sprint 1 - we focus on 2021 data

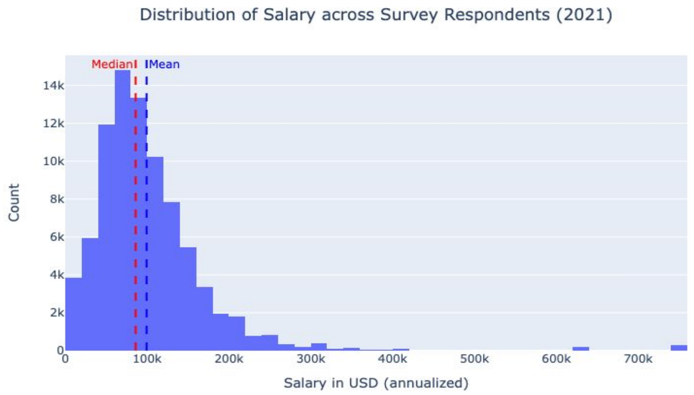
(~84,000 rows, 25 columns)*

*Whittled down from a whopping 530+ variables - selection detailed in notebook, and is based on broadly relevant variables (with the potential to expand or reduce as we conduct feature engineering in depth).

Emerging insights for 2021 data

What do we know so far?

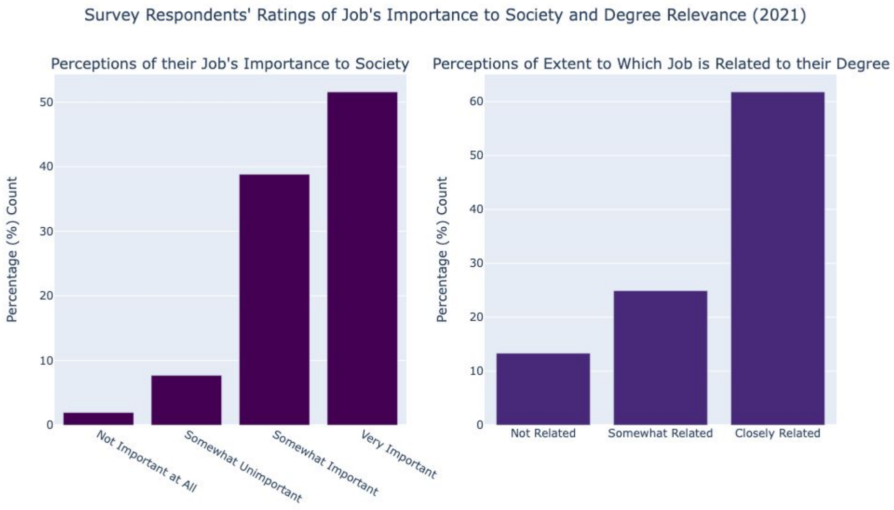
- Salary distributions amongst graduate survey respondents
- Overall salary distribution is heavily right-skewed (longer tail corresponding to higher values)
- Distribution by gender highlighting pay gaps; lower median salary as well as interquartile range



Emerging insights for 2021 data

What do we know so far?

- Exploration of a previously unexplored variable - usually unavailable in other datasets
- Self-reported perceptions of an individual's job's importance to society - over 50% of respondents rate this as 'Very Important'
- Perceptions of extent to which job is related to their degree - some variability, most most find it's 'closely related'
- This is a helpful proxy of 'skills matching - and we'll want explore this distribution more by different demographics, and by degree subject

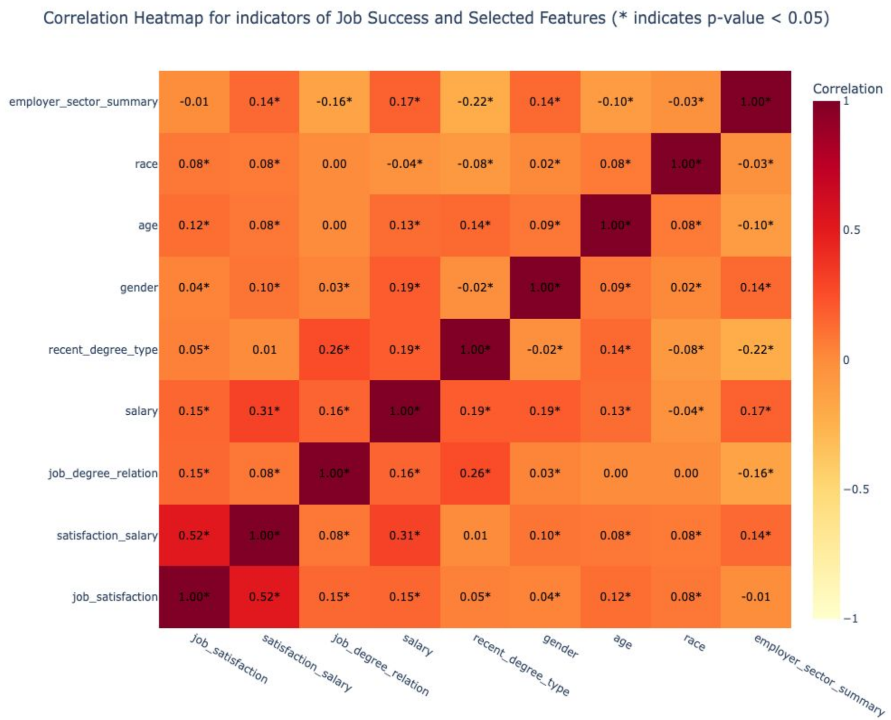


Emerging insights for 2021 data

What are our hypotheses?

- Job satisfaction elements will be central to our composite y-variable
- Thinking about correlation - **honing in on selected features**
- Exploring significant positive correlations **(darker shades)**
    - Salary and job satisfaction
    - **Satisfaction** with salary, job satisfaction
    - Job-degree skills matching and job satisfaction
    - Degree type (i.e., BSc, MSc), salary and job-degree skills matching
- Exploring significant negative correlations **(lighter shades)**
    - Age and employer sector
    - Salary and race

Note: We need to interpret these with caution given the coding of categorical columns as numeric - and explore further as we build out models.



*Please note the darkest shades on the heatmap can indicate high collinearity - to investigate further during the modelling process...

Immediate next steps

1. **Investigating the time element**
   a. Merge with datasets beyond 2021 to model trends in employment over time, and clean appropriately
   b. It may be interesting to explore outcomes pre-Covid vs post-Covid
2. **Build on correlation heatmap to understand relationships**
   a. Finish plotting relationships across other variables - e.g., can we dig deeper into the relationship between popular degree subjects (e.g., Engineering) and job satisfaction?
3. **Inspect and continue pre-processing of variables**
   a. Get data ready (i.e., inspecting data-types) for baseline modelling
   b. Research and finalise proposed modelling approach - to consider decision trees?
   c. Identify appropriate feature engineering strategies
4. **Experiment with our first composite y/target-variable**
   a. Consider salary, job satisfaction and degree-job alignment as core variables
   b. Experiment with different weightings to improve model accuracy - could strength of correlations help make a start on this?
5. **Think about the 'stretch goal' use case for my final models - a recommendation tool**
   a. Is there additional data required for this? E.g., online job postings, basic user inputs