Who's likely to love their job?

A Predictive Tool for Job Satisfaction, and Key Drivers

Agenda

01

Intro

02

Data exploration

03

Machine Learning Roadmap

04

Model evaluation

05

'So what?'

06

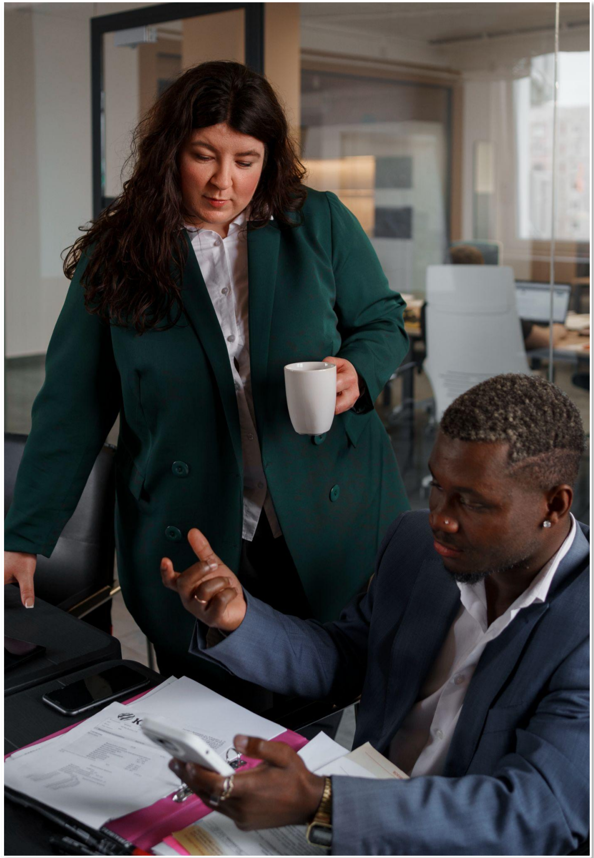
Next steps

01

Intro: Problem space & opportunity

Finding a job we 'love' is hard...

- ..and a critical issue in today's labour market, to support **productivity and wellbeing**
- But it continues to be a challenge to get the 'right fit', as we know from stark unemployment figures



** 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'*

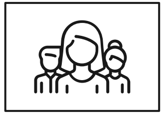
It can be tricky to measure & predict job satisfaction

- ..and even trickier to determine **key drivers of 'high satisfaction'**

- First, understanding **who** in the labour force self-reports as 'satisfied' is invaluable

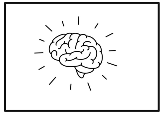
- It can help us dig deeper into the **'why'**...

Data science can help us tackle this



Predict the 'who'

Develop a 'classification' to **predict** with more accuracy



Dig deeper into the 'why'

Analyse and recommend priority areas for improving job satisfaction

Why? Developing a predictive tool for policymakers and jobseekers can provide actionable insights, and demystify drivers of job satisfaction

02

Data exploration

We analysed survey data from college graduates in the US

What?

1. Recurring survey, conducted by the **National Science Foundation**

2. Rich data on **job satisfaction** and **aspirations**

3. Focus on those **‘currently employed’ at time of survey**

How?

1. Loading, cleaning, preprocessing survey data - **2015, 2017, 2019, 2021**

2. Extensive variable-readable name mapping

295, 323 rows and 31 columns

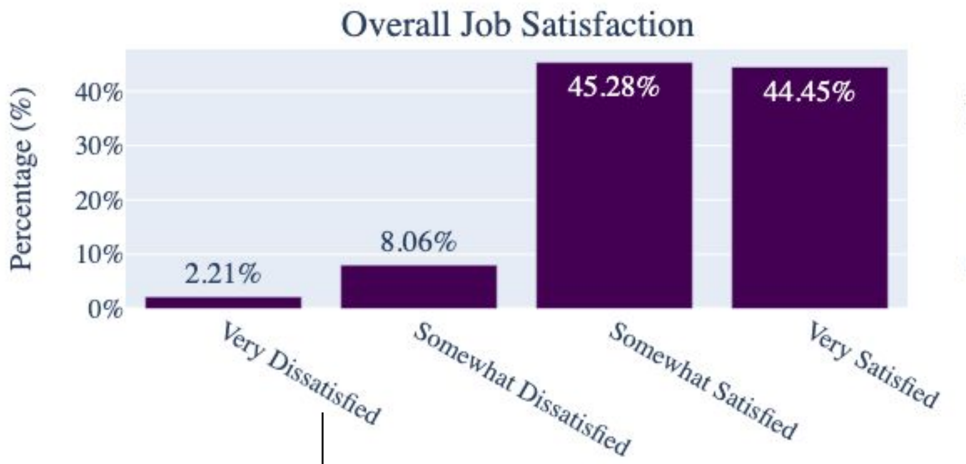
Challenges?

1. Inconsistent variable naming across years

2. Feature engineering and selection; reducing from 200+ survey Qs to 31

3. Lengthy documentation of surveys; **required careful interpretation**

Respondents generally like their jobs

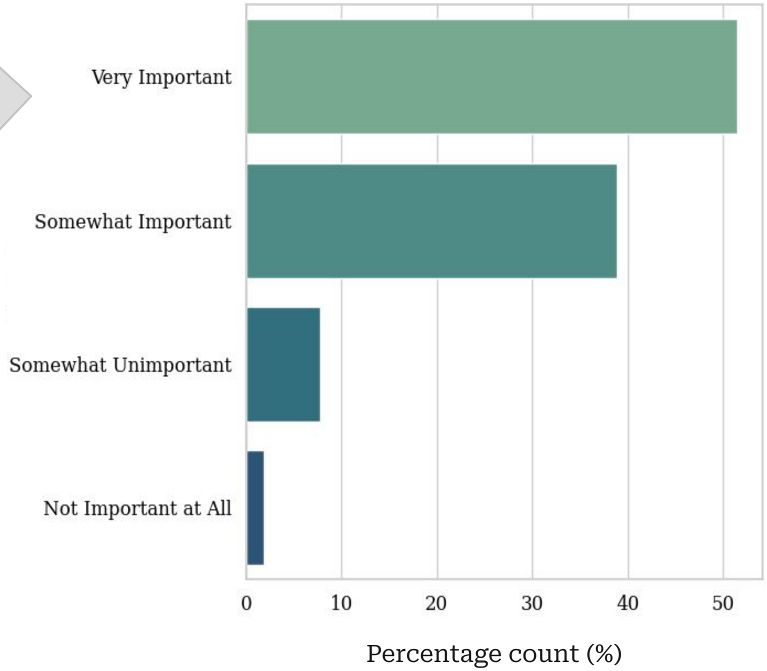


Nearly 90% of respondents were 'somewhat' or 'very' satisfied

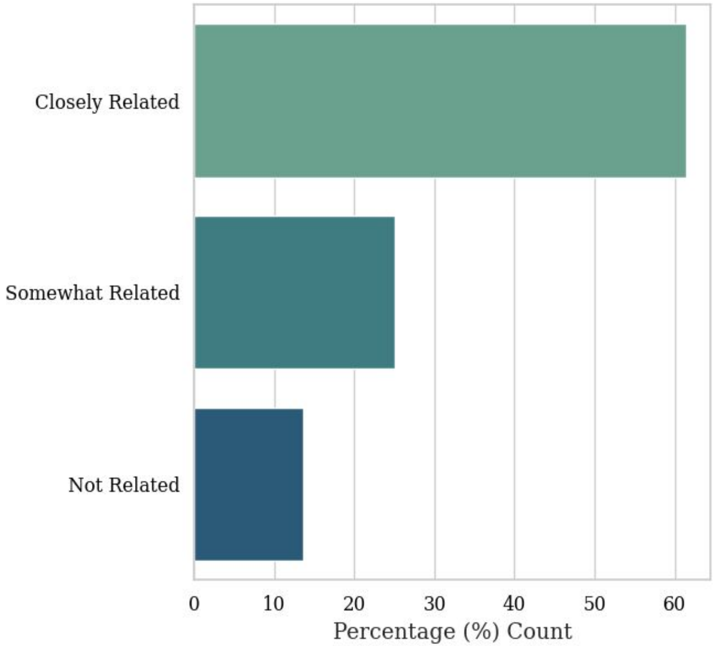
~10% of respondents had low satisfaction - an imbalance dataset

Many feel their job is important to society

Over 50% report their job as 'very important' to society



Respondents thought their jobs are closely related with their degree subject



Over 60% report their job as '**closely related**' to their academic background.

Another potential indicator of 'highly satisfied' employees

Running a logistic regression, we found top predictors of 'high job satisfaction'

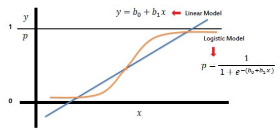
Individuals who are happier with career advancement opportunities, are 4.06x more likely to be satisfied

Job tenure, salary & marital status are strong predictors of satisfaction

03

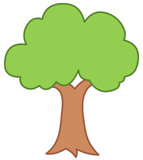
Machine Learning Roadmap

Experimenting with 4 different models, of varying complexity



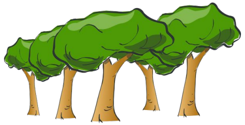
Baseline Model 1:
Logit Regression*

- Variance Inflation Factor
- Backwards selection



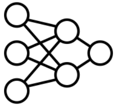
Model 2: Decision Tree

- Maximum depth, maximum features, minimum samples leaf…
- Coarse Analysis, 5-fold CV
- RandomizedSearchCV



Model 3: Random Forest

- Maximum depth, maximum features, minimum samples leaf…
- Number of estimators
- Coarse Analysis, 5-fold CV



Model 4: Neural Network

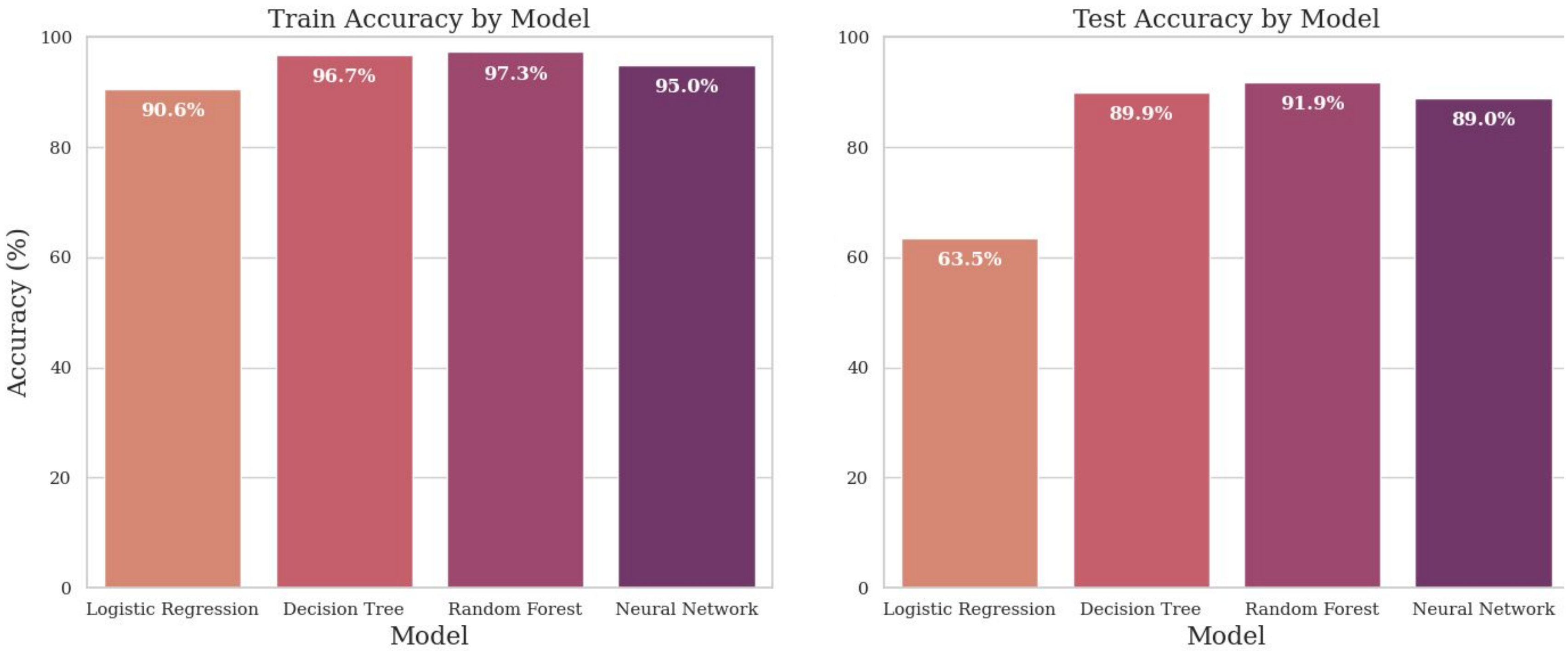
- Learning rate, regularisation
- Epochs
- Coarse Analysis, 5-fold CV

*Including exploration of Principal Component Analysis, followed by further scaling and preprocessing of numerical and categorical variables as required

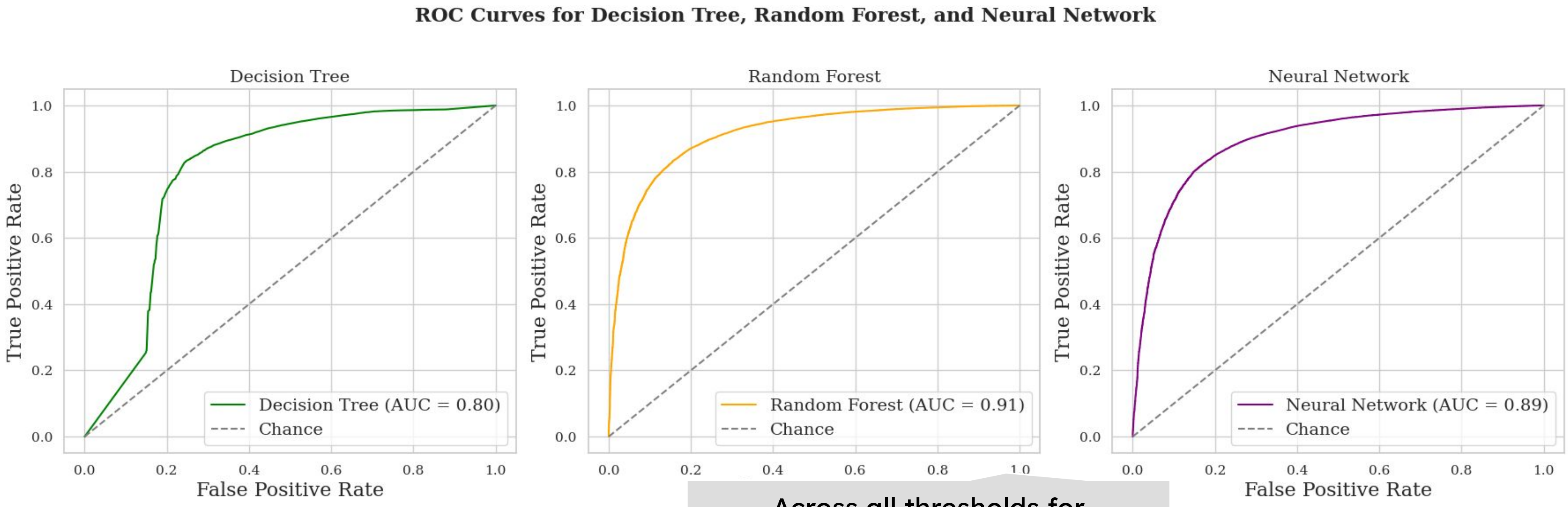
04

Model evaluation

Random Forest came out on top, for both train...and test data accuracy



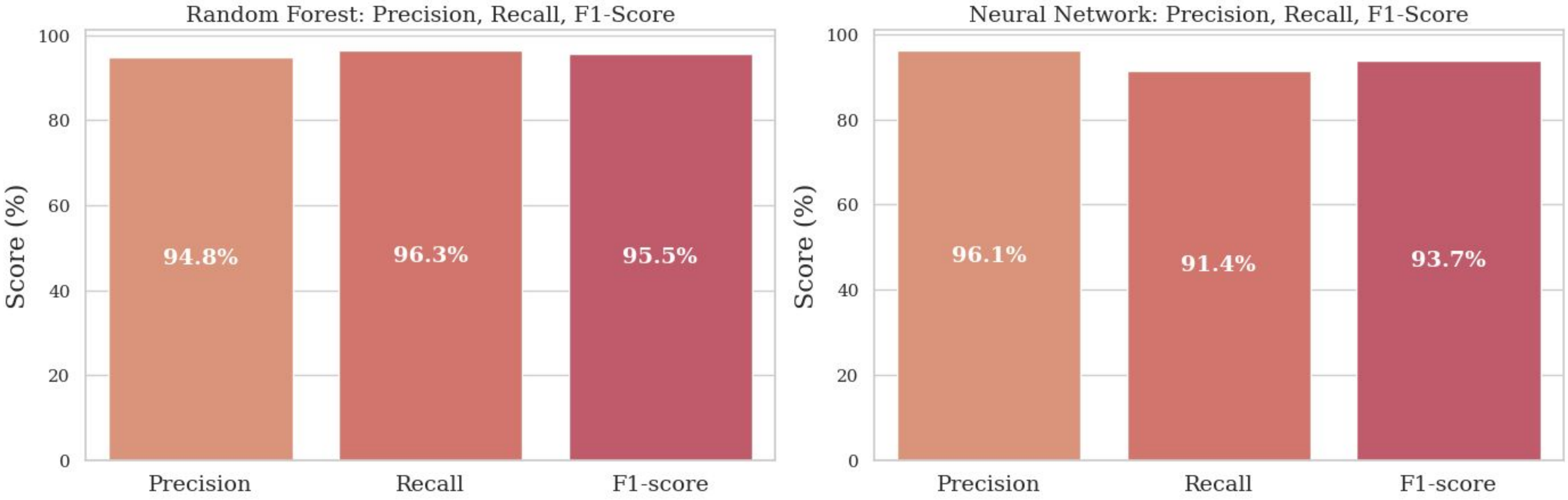
All models did better than 'random guessing', with the most complex models doing best



Across all thresholds for classification, Random Forest had the highest overall performance

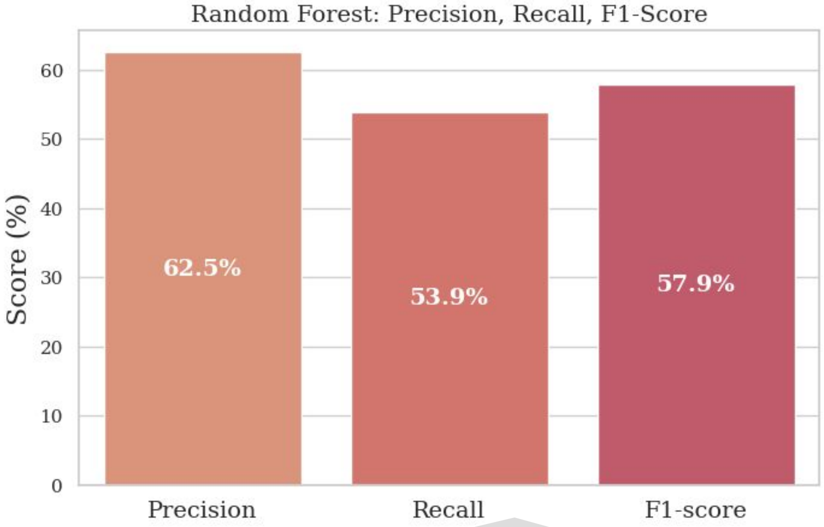
We saw good performance across classification accuracy for 'High Satisfaction' (Class 1)

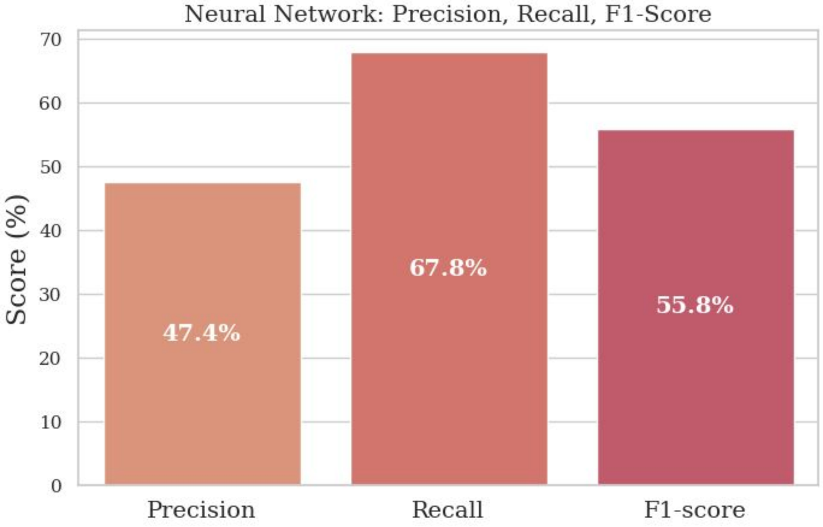
Whilst our neural network wasn't the best overall, it's precise in predicting 'highly satisfied' individuals



We saw <u>reasonable</u> performance across classification accuracy for 'Low Satisfaction' (Class 0)

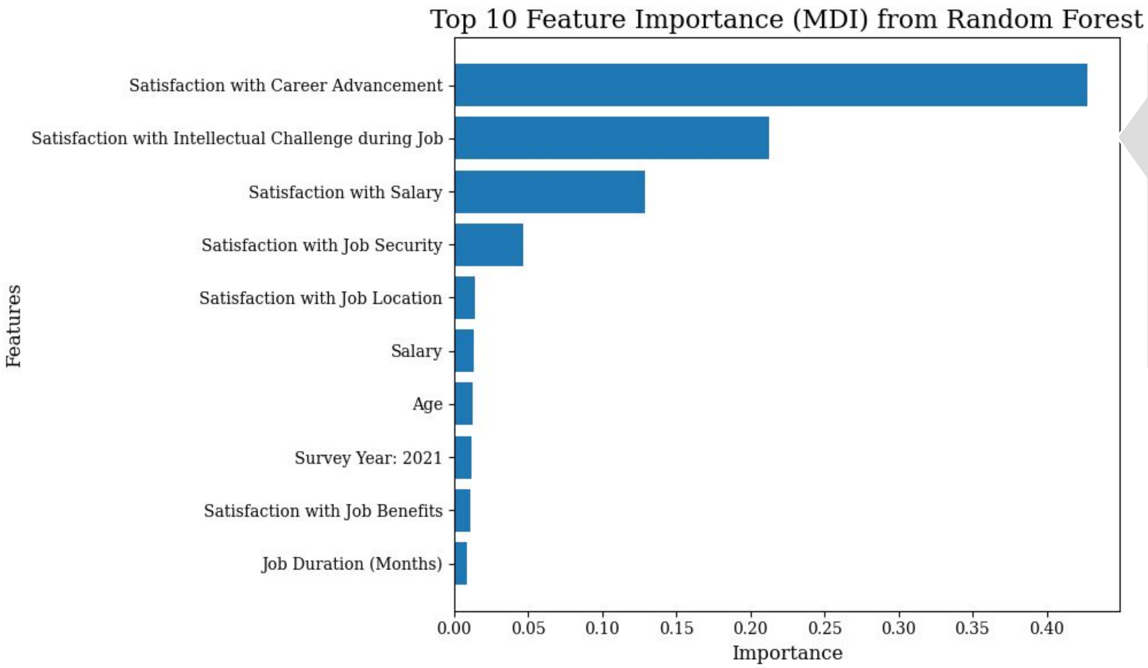
Whilst our neural network wasn't the best overall, it has better recall: identifying unsatisfied individuals





Random Forest doesn't do as well with spotting all unsatisfied respondents

Our best performing model shows that scores for job satisfaction matter



Components of **job satisfaction** such as **salary, security** and **location** are important features

05

‘So what?’

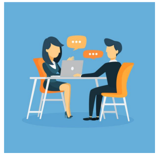
Using our well-tuned model can generate crucial insights for...

Policymakers

1. Explore and address wider disparities in job satisfaction

2. Use insights to formulate target labour market interventions

3. Revise and update the predictive model against a changing landscape



Jobseekers

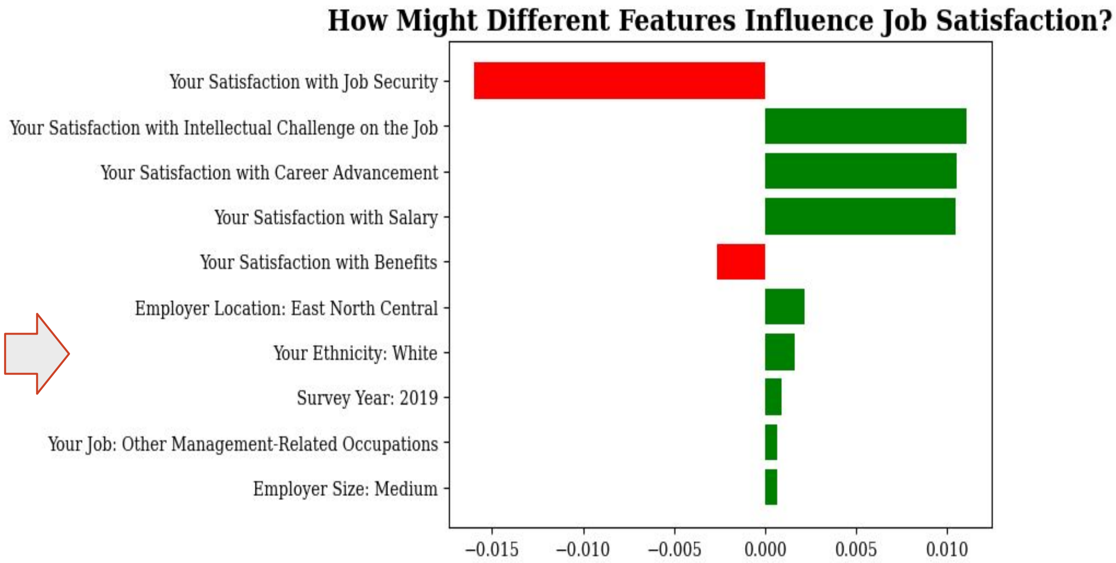
1. Understand drivers of job satisfaction, individualised

2. Keep up-to-date with labour market trends, such as most popular degrees

3. Hone in their job search on maximising job satisfaction in specific areas

What interactive insights can this prediction tool provide? (e.g., Streamlit)

1	A **user inputs** information about themselves (degree, occupation, region)
2	**Model outputs a probability score** of ‘High Job Satisfaction’
3	App generates insights about **what factors influence this probability** (like this plot)
4	User learns about what factors **could positively or negatively impact job satisfaction**



This graph was generated using LIME explainer and best Random Forest model on one test instance- e.g., ***dissatisfaction with salary could significantly reduce job satisfaction***

Users learn and make more data-driven decisions about the job market

06

Next steps

Next steps

To build on learnings and insights…

- **Building out proof-of-concept Streamlit app** (using LIME and Random Forest, as seen earlier)
- **Neural networks** - further tuning, and potential to hone in on drivers of ‘low satisfaction’
- Continue to **explore benefits of interpretable ML** - using SHAP values