Predicting ‘high job satisfaction’

EDA, establishing a baseline, and modelling roadmap

Sprint 2 - 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*

Refining our focus since Sprint 1

- Further EDA highlighted overall job satisfaction as our key target variable of interest.
- Job satisfaction can serve as a proxy for self-reported job fit, with datasets explored offering deeper insight into what might predict high job satisfaction
- These findings can guide jobseekers in evaluating potential roles and inform policymakers aiming to improve labor market outcomes.

** 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'*

Refining our problem statement and purpose

- How might we use machine learning to **better predict** the likelihood of ‘high job satisfaction’, as a proxy for job fit?
  - How can we define ‘high job satisfaction’?
    - Harness survey data on self-reported satisfaction scores
    - Define the threshold for ‘high job satisfaction’
  - How can we use ML classification to produce actionable insights?
    - Evaluate and improve classification accuracy
    - Translate into insights for jobseekers and policymakers

Bottom Line: A well-constructed ML classification model for predicting 'high job satisfaction' has the potential to help **jobseekers** to make informed career decisions and provides **policymakers** with actionable insights to enhance labour market policies, driving better job fit and satisfaction across the workforce.

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

How did we load, clean and process it?

For Sprint 2 - I merged survey data from **2015, 2017, 2019 and 2021:**

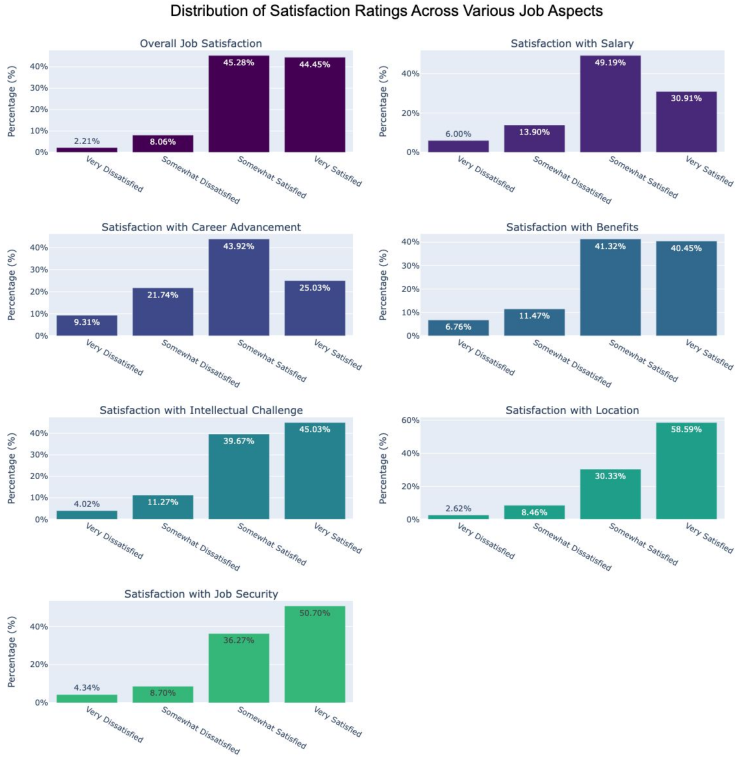
- Read in the data, inspecting yearly user guides to ensure correct mapping of variables (e.g., degree code of '089' could mean engineering)
- Cleaned and checked missing
- Inspected distribution, and aggregated categories for interpretability

Ready for EDA: **295,323 rows** and **31 columns**

Highlights from EDA (1)

Distribution of job satisfaction ratings

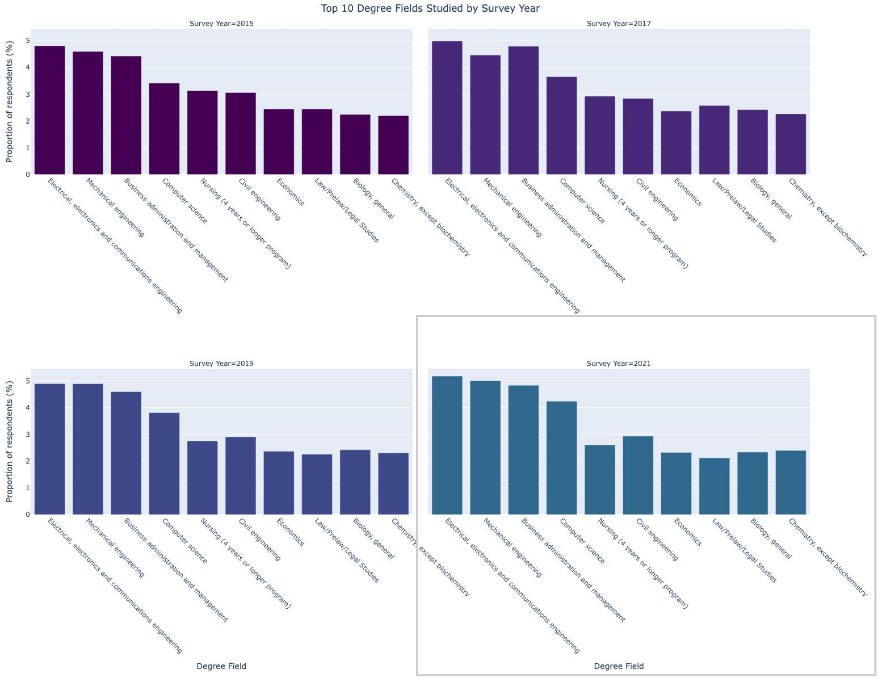
- Exploration of a previously unexplored variable - usually unavailable in other datasets
- Observed similar distributions across components of satisfaction
- Most respondents where either ‘somewhat’ or ‘very satisfied’ overall
- Used insights to refine my **target variable as binary:**
- 1: ‘Highly satisfied’
- 0: ‘Not satisfied’



Highlights from EDA (2)

Popular degrees over time

- The top 10 degree fields remain stable over the 4 survey years.
- We can notice that as we move towards 2021, popularity in electronic and mechanical engineering increases.
- Some degrees, such as nursing reduce in popularity with 3.1% of respondents doing a nursing degree in 2015, compared to 2.6% of respondents in 2021. T
- Interesting to note, given the likely impact of the pandemic on the popularity of some degrees.



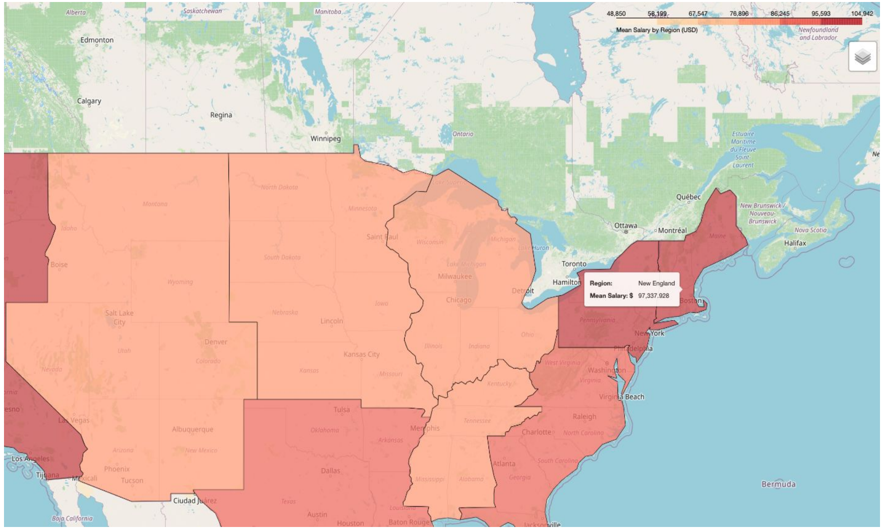
Post-Covid

Highlights from EDA (3)

Geographical patterns

- We see higher mean salaries over in the Pacific and New England (darker colours indicate higher salaries)
- This seems sensible, given the type of occupations we know our respondents hold in tech, and the higher salaries in these larger cities or tech hubs.
- Given that we see salary varies by the respondent's location...
- We might hypothesise that so would the likelihood of an individual being highly satisfied in their job.

(Note: due to data availability, mean salaries are calculated for higher-level regions, rather than by-state)



Download as HTML

Constructing a baseline model: Logit

Train/test split	Dummy encoding	Collinearity and multicollinearity checks	Backwards selection & Statsmodel	Evaluation

Train/test split

Keeping in mind issues of overfitting and data leakage

Stratifying to preserve proportion of our binary target var - which has a high class imbalance (90% positive)

Scaling features and seeing distributions before and after

Dummy encoding

As part of pre-processing, inspecting categorical variables and using pd.get_dummies()

Collinearity and multicollinearity checks

Generating correlation coefficients - dropping highly correlated features, if they illustrated lower variance

Exploring VIF to determine further features to drop.

Backwards selection & Statsmodel

Inspecting features with highest p-values, and conducting iterations of logistic regression, using statsmodel.

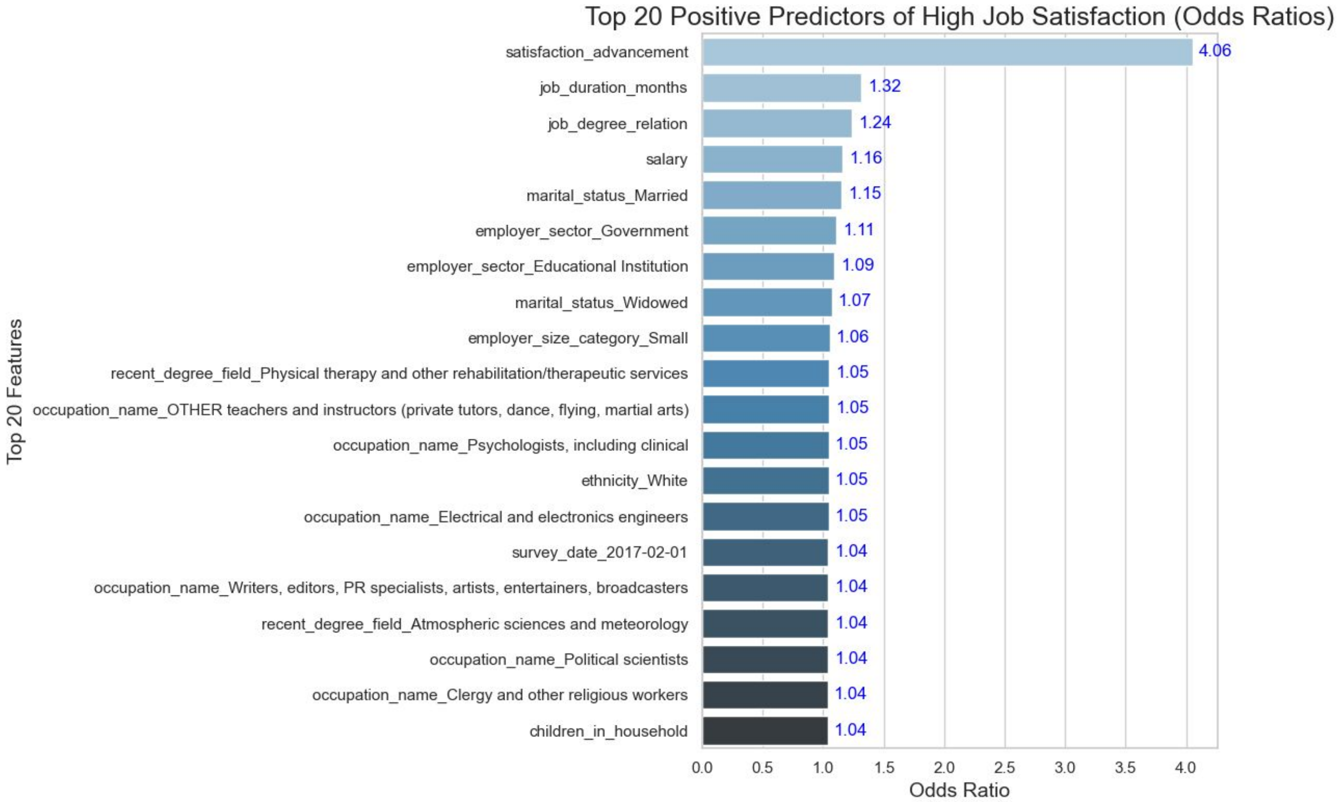
Starting with a 'baseline', and conducting 4 more iterations.

Evaluation

Observed a marginal improvement in train accuracy following backwards selection

However, clear signs of overfitting when fitting to unseen data

Excerpt: Insights from baseline modelling



Evaluation framework

Class Imbalance Consideration

- Over 90% of respondents belong to the "High Job Satisfaction" (class 1) category.
- This imbalance impacts the model's ability to generalise well to both classes, so accuracy alone may not reflect true performance.

Going beyond accuracy score

- ROC curve and AUC score
- Confusion matrix
- Classification report

In addition to the above, we also care about interpretability -

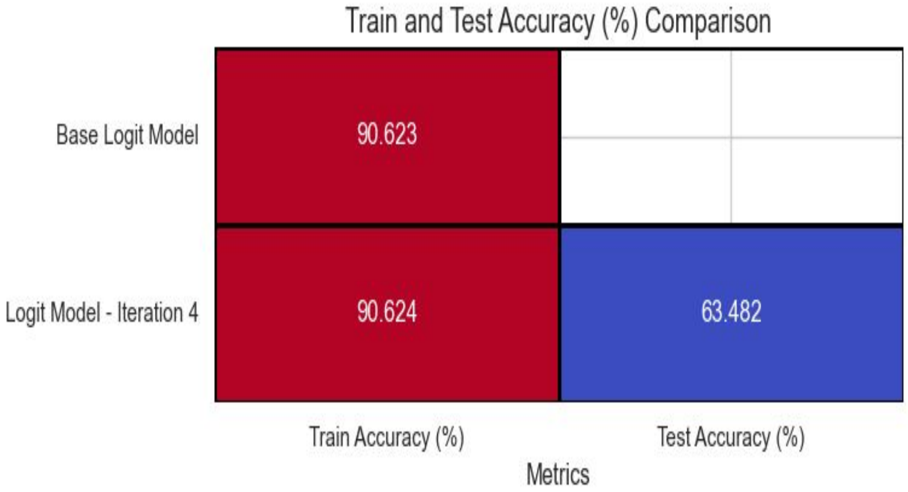
- Our end users may be individuals in or seeking employment...
- Or policymakers seeking to understand inequities in the labour market when it comes to job satisfaction.

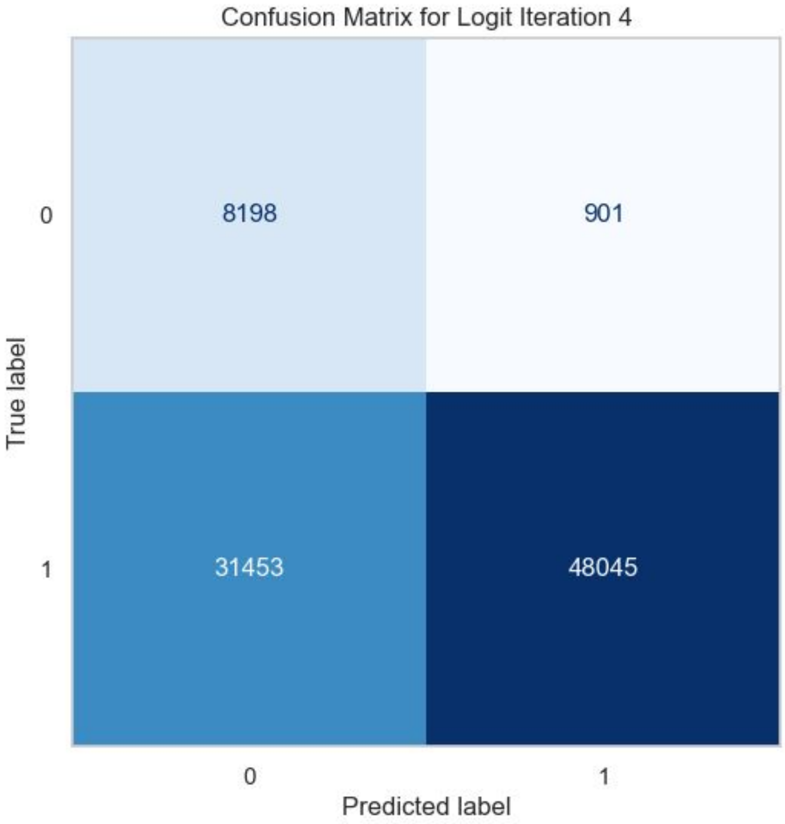
Evaluation: Applying to Baseline Logit Model

- Over 90% train accuracy; only 60% test accuracy
- The model is good at identifying those with high job satisfaction (48,045 True Positives).
- It also correctly identifies those without high satisfaction in 8198 True Negative cases.

However:

- The model misses a significant portion of individuals who are actually highly satisfied.
- Precision will be high since the model has a low number of False Positives compared to True Positives.
- Recall will likely be lower due to the high number of False Negatives—it misses many people who actually have high job satisfaction.


Train and Test Accuracy (%) Comparison
Base Logit Model
Logit Model - Iteration 4
90.623
90.624
63.482
Train Accuracy (%)
Test Accuracy (%)
Metrics


Confusion Matrix for Logit Iteration 4
True label
0
1
8198
901
31453
48045
0
1
Predicted label

Next steps: Modelling Roadmap

01	**Construct baseline model**	• Baseline Logistic Regression using statsmodel • Baseline evaluation
02	**Further explore PCA**	• First iteration completed in notebook • Further explore impact of adjusting n_components
03	**Model 1: Decision Tree**	• Build baseline DT model • Perform coarse analysis of hyperparameters • Use RandomizedSearchCV prior to GridSearch • Evaluate model & classification accuracy
04	**Model 2: Ensemble learning**	• Explore Random Forest, building up in complexity • Leading up to neural network

05: Comprehensive model evaluation and write-up