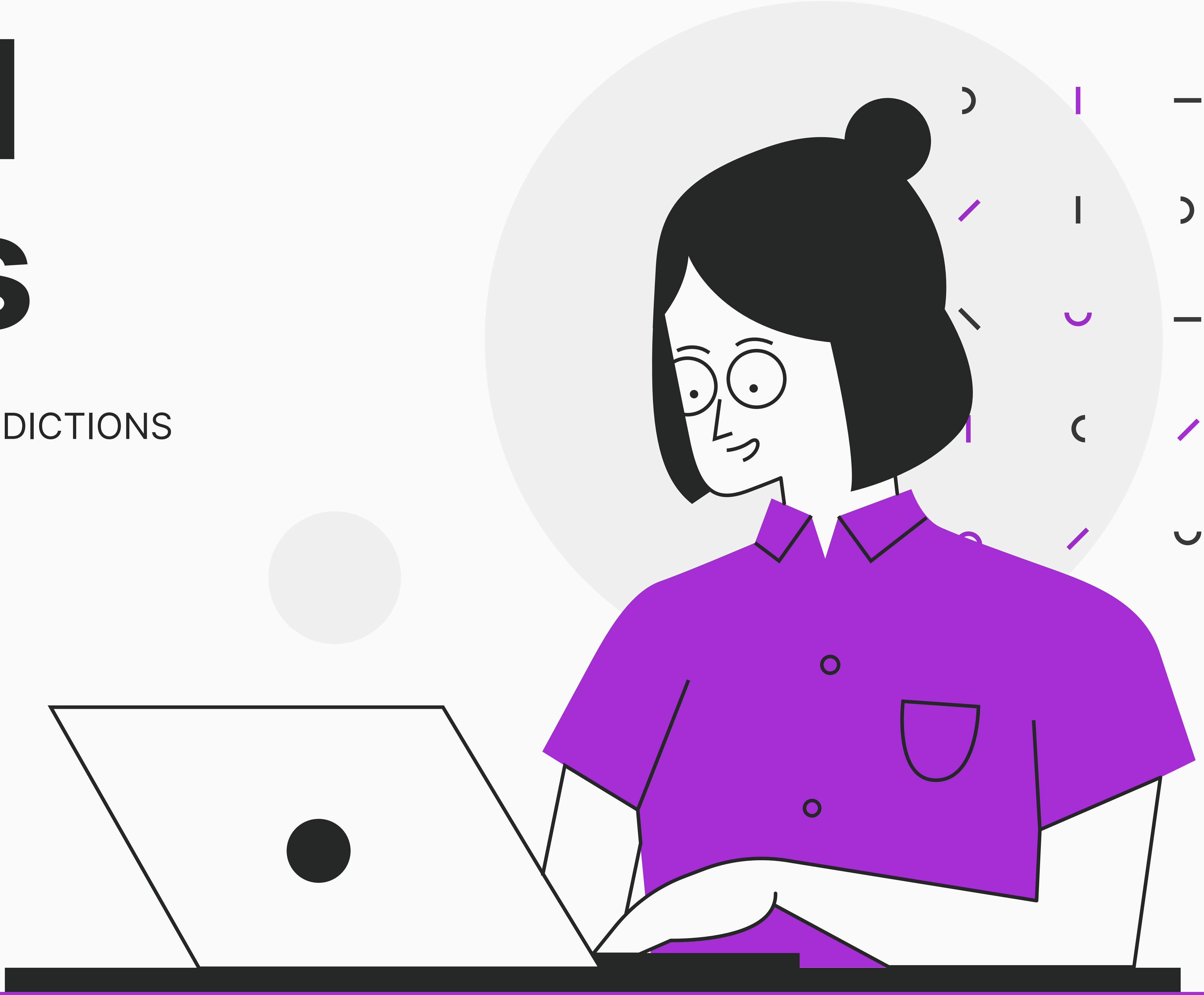
# Bias and Fairness

IN DATA SCIENTIST JOB CHANGE PREDICTIONS

01

Claire Saint-Donat, Xiangyue Wang



# Is this data scientist seeking a new job?

is the driving question behind the classification model we analyze.

Using a set of features about a given data scientist, the model categorizes them as either a **“job seeker” or “non-job seeker”.**

The model was trained on features corresponding to each candidate’s **current credentials**, **demographics** as well as **work experience**. Many of the features are categorical, some with high cardinality.



The model, titled "Awesome HR Data Visualization & Prediction" is created by the Kaggle user Josh.

# The Data

03

The data were published by a data science company looking to hire data scientists who successfully passed some certification courses conducted by the company. The data consist of features on those data scientists, including but not limited to:

### GENDER

Gender of the candidate.

### EDUCATION LEVEL

Ranging from "Primary School" to "PhD".

### EXPERIENCE

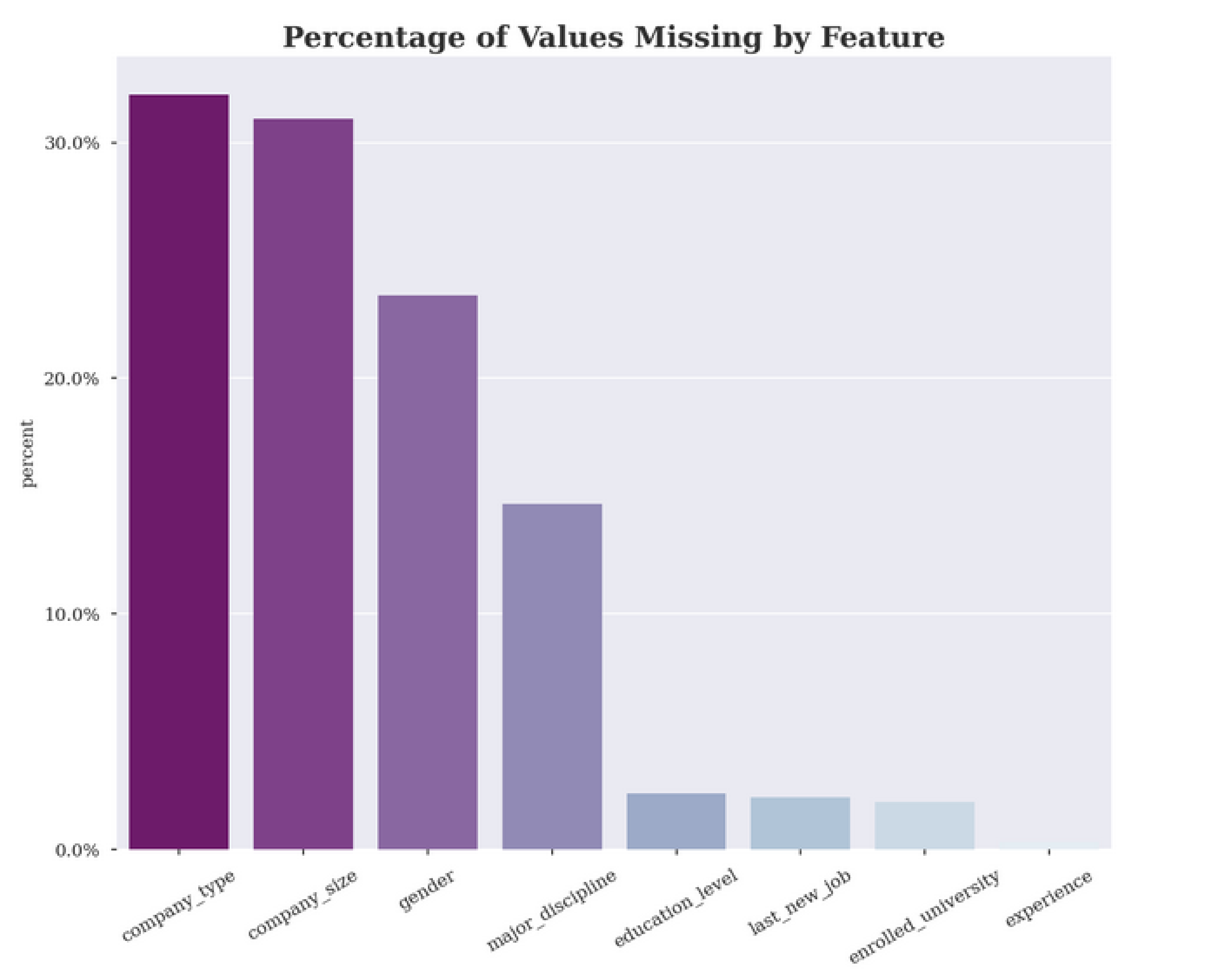
Total experience in years.

### CITY DEVELOPMENT INDEX

A numeric measure how developed is the city the person resides.

### COMPANY SIZE

Size of the company the person currently works at.



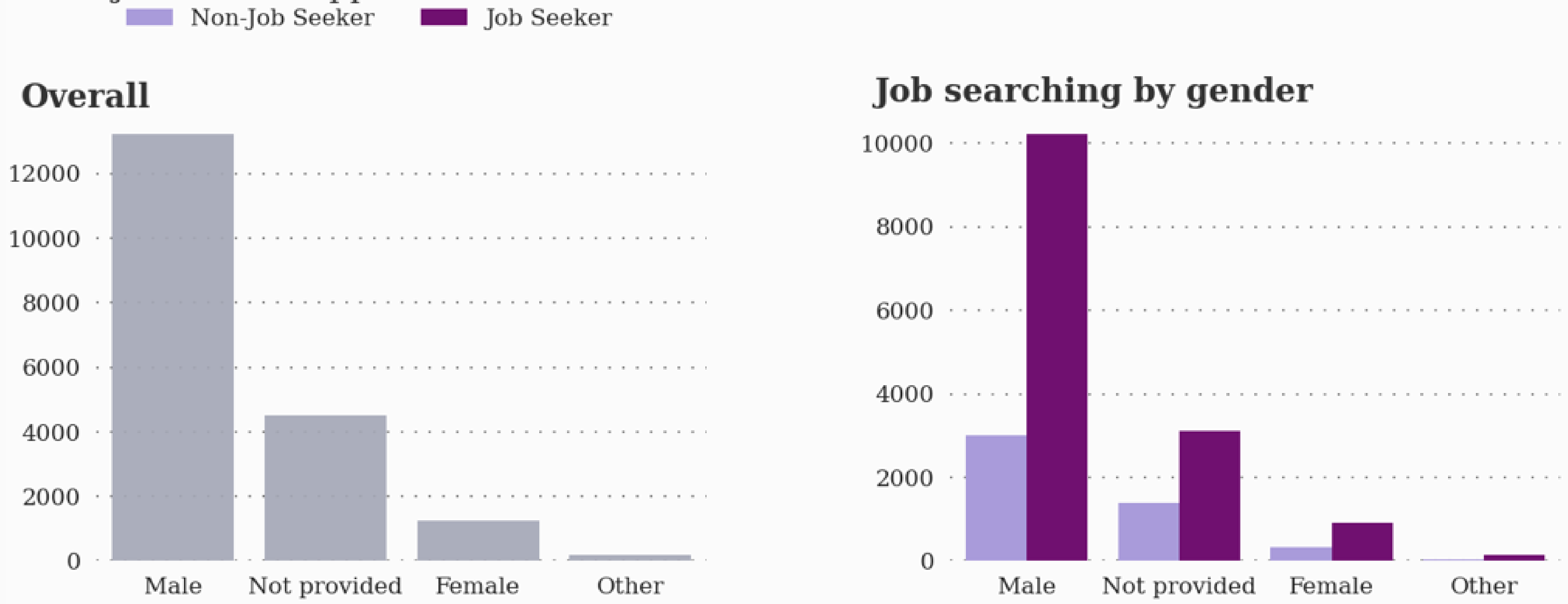

What are the missing data?

## How many are looking for a new role?

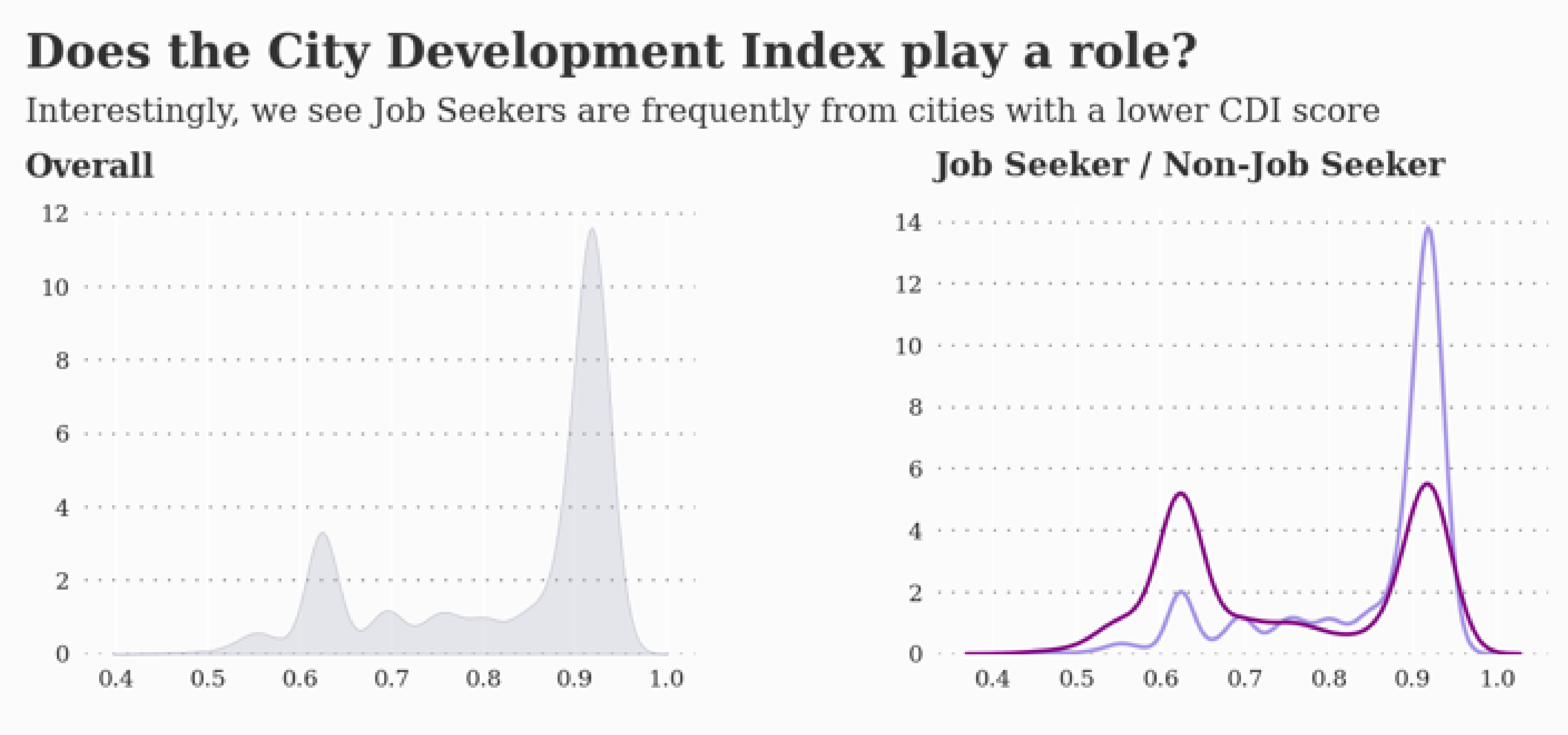
We see an imbalanced dataset;
most trainees are not job-seeking

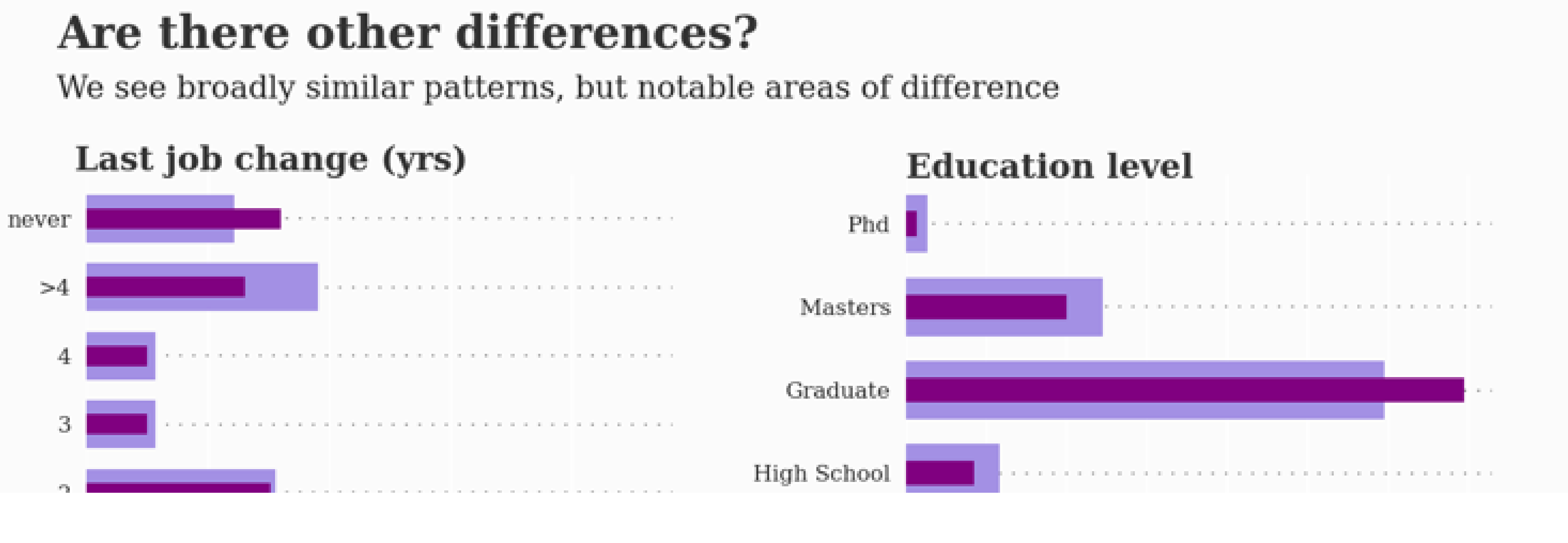

## Who is looking for a new job?

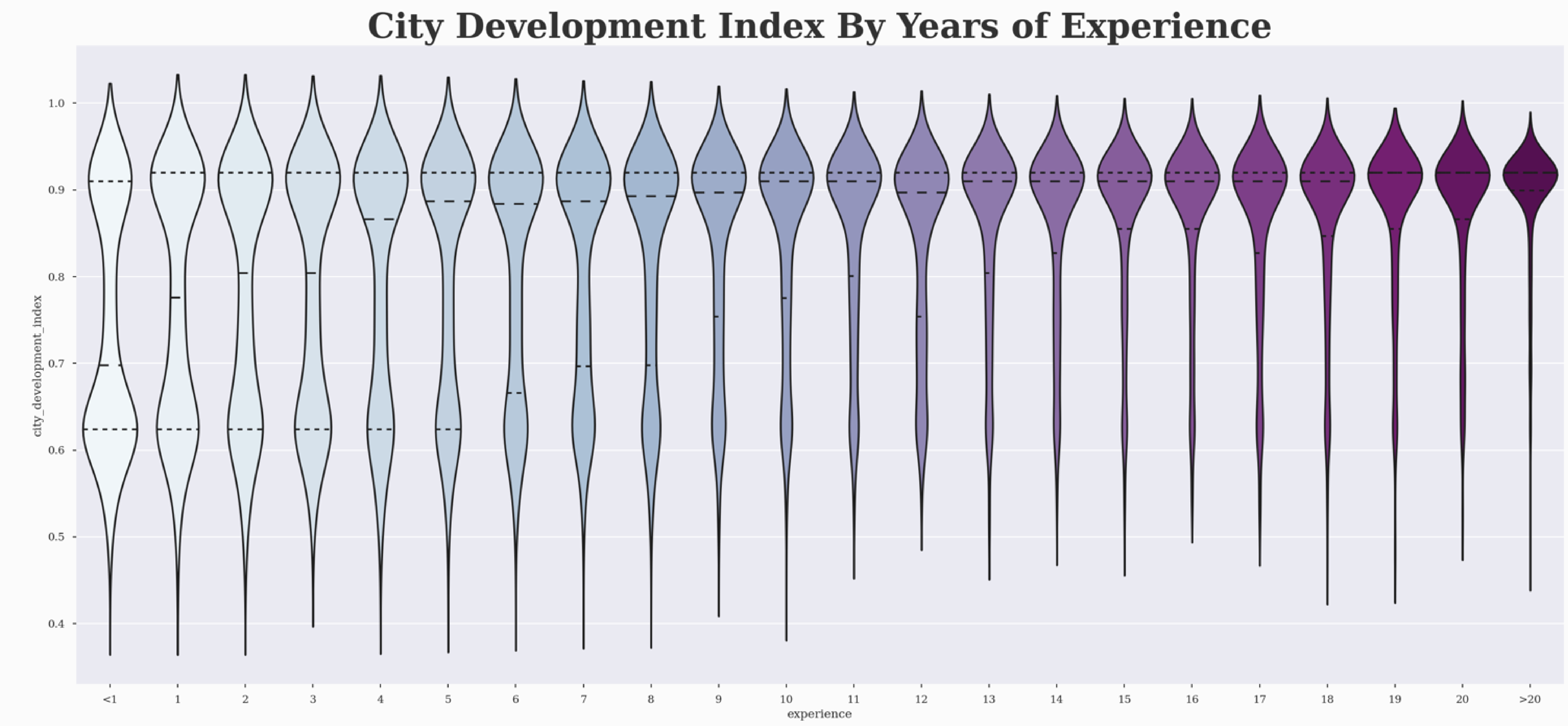
Most job-seekers appear to be male



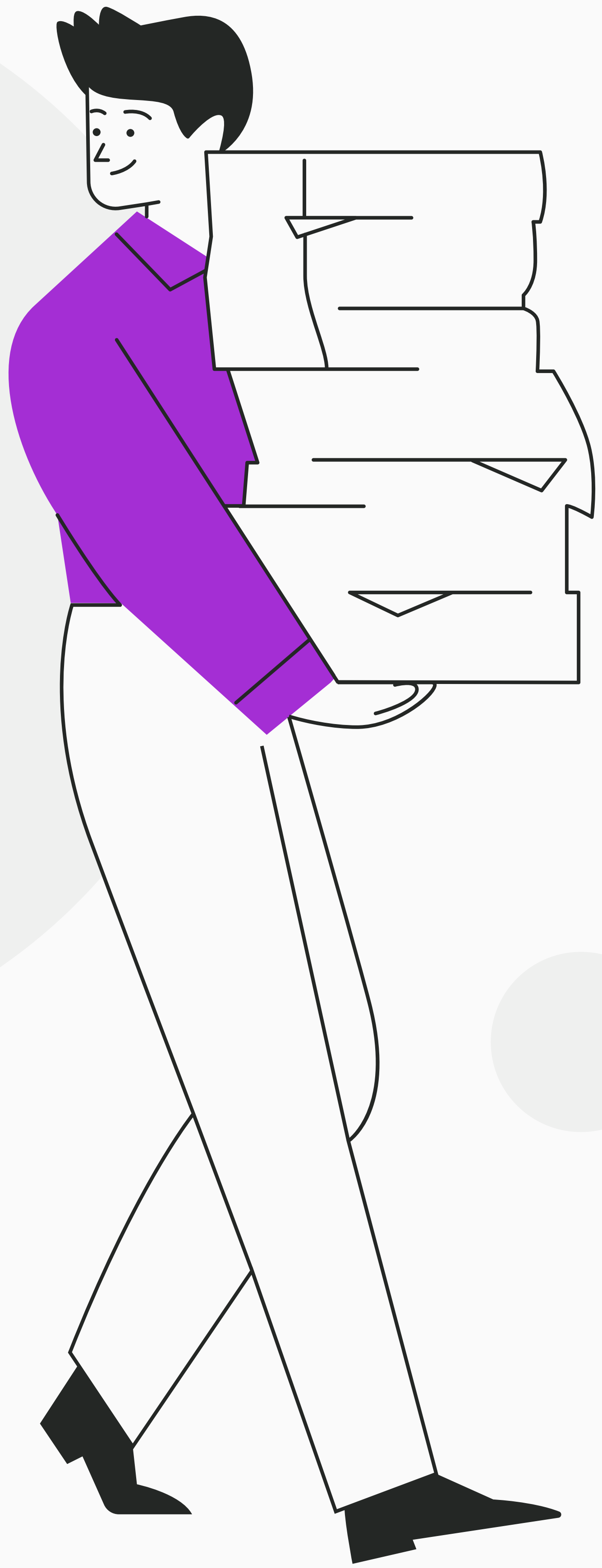
# Training and City Development





City Development Index By Years of Experience
city_development_index
1.0
0.9
0.8
0.7
0.6
0.5
0.4
<1
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
>20
experience



# Implementation

### DATASET PREPARATION

Imputing of missing values, test/train split

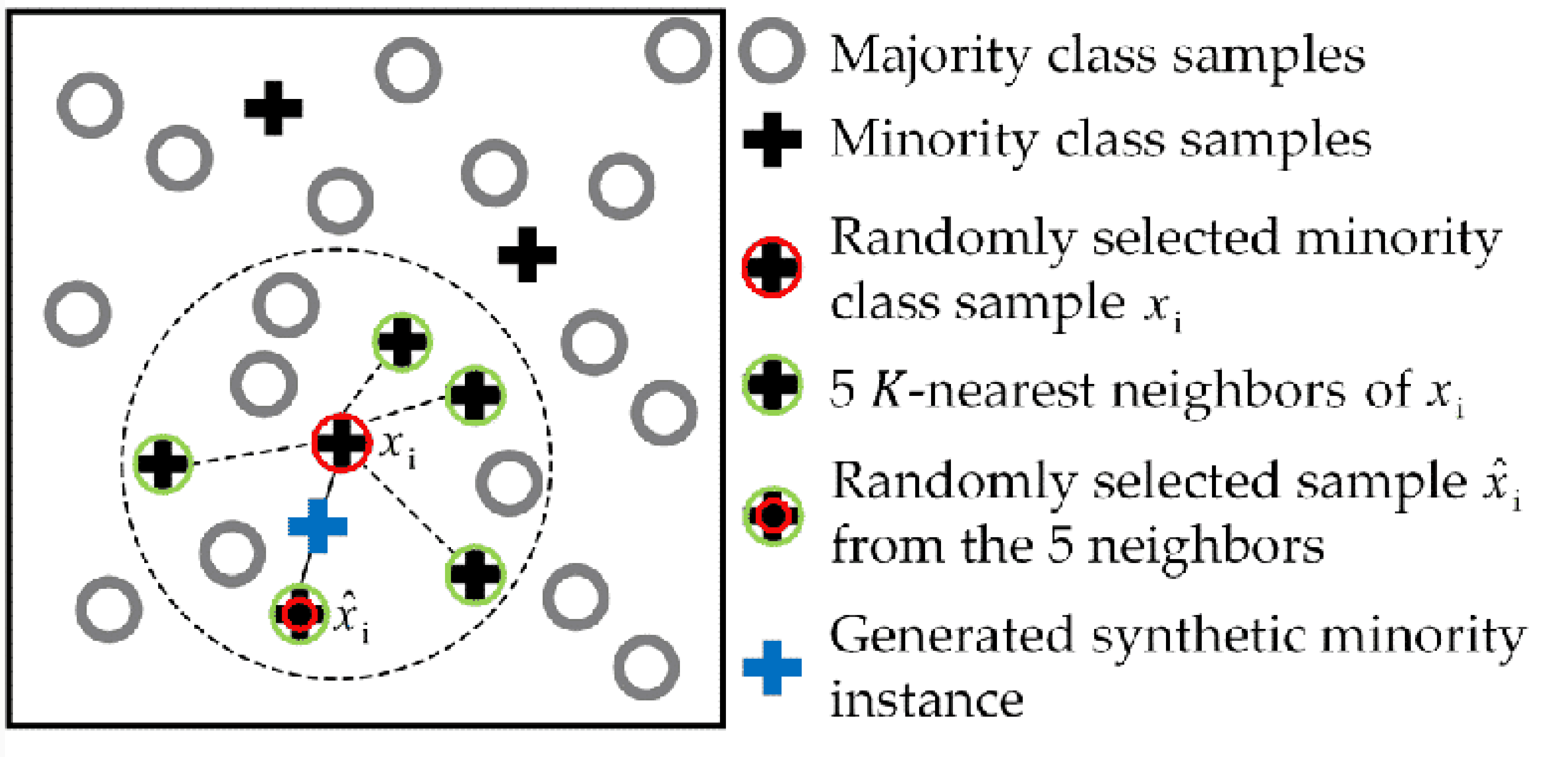
### MODEL TRAINING

SMOTE (Class Imbalance upsampling), Classification Models, Hypterparameter tuning

### EVALUATION

Measuring accuracy, precision, recall and ROC AUC for each of the models trained and making final selection

# SMOTE for Target Class Imbalance



# Evaluation

| | Accuracy | Recall | Precision | ROC AUC Score |
|---|---|---|---|---|
| SVC Score | 78.4% | 39.4% | 57.0% | 65.0% |
| Decision Tree Score | 72.5% | 46.6% | 43.1% | 63.6% |
| Random Forest Score | 79.0% | 48.8% | 57.2% | 68.6% |
| Tuned Random Forest Score | 76.9% | 10.7% | 60.4% | 54.3% |
| SMOTE Random Forest Score | 78.6% | 49.5% | 56.1% | 68.6% |
| Logisitc Regression Score | 78.1% | 34.2% | 56.9% | 63.0% |
| **SMOTE Logistic Regression Score** | 77.1% | 74.4% | 51.4% | 76.2% |
| KNN Score | 77.8% | 36.1% | 55.5% | 63.5% |

**Support Vector Machine (SVM)**

| | Predicted Non-Job Seeker | Predicted Job Seeker |
|---|---|---|
| Actual Non-Job Seeker | 3726 | 385 |
| Actual Job Seeker | 784 | 510 |

**Decision Tree**

| | Predicted Non-Job Seeker | Predicted Job Seeker |
|---|---|---|
| Actual Non-Job Seeker | 3316 | 795 |
| Actual Job Seeker | 691 | 603 |

**Random Forest**

| | Predicted Non-Job Seeker | Predicted Job Seeker |
|---|---|---|
| Actual Non-Job Seeker | 3639 | 472 |
| Actual Job Seeker | 663 | 631 |

**Random Forest (w/Adjustments)**

| | Predicted Non-Job Seeker | Predicted Job Seeker |
|---|---|---|
| Actual Non-Job Seeker | 4020 | 91 |
| Actual Job Seeker | 1155 | 139 |

**Logistic Regression**

| | Predicted Non-Job Seeker | Predicted Job Seeker |
|---|---|---|
| Actual Non-Job Seeker | 3756 | 355 |
| Actual Job Seeker | 828 | 466 |

**K-Nearest Neighbours (KNN)**

| | Predicted Non-Job Seeker | Predicted Job Seeker |
|---|---|---|
| Actual Non-Job Seeker | 3737 | 374 |
| Actual Job Seeker | 827 | 467 |

**SMOTE Random Forest**

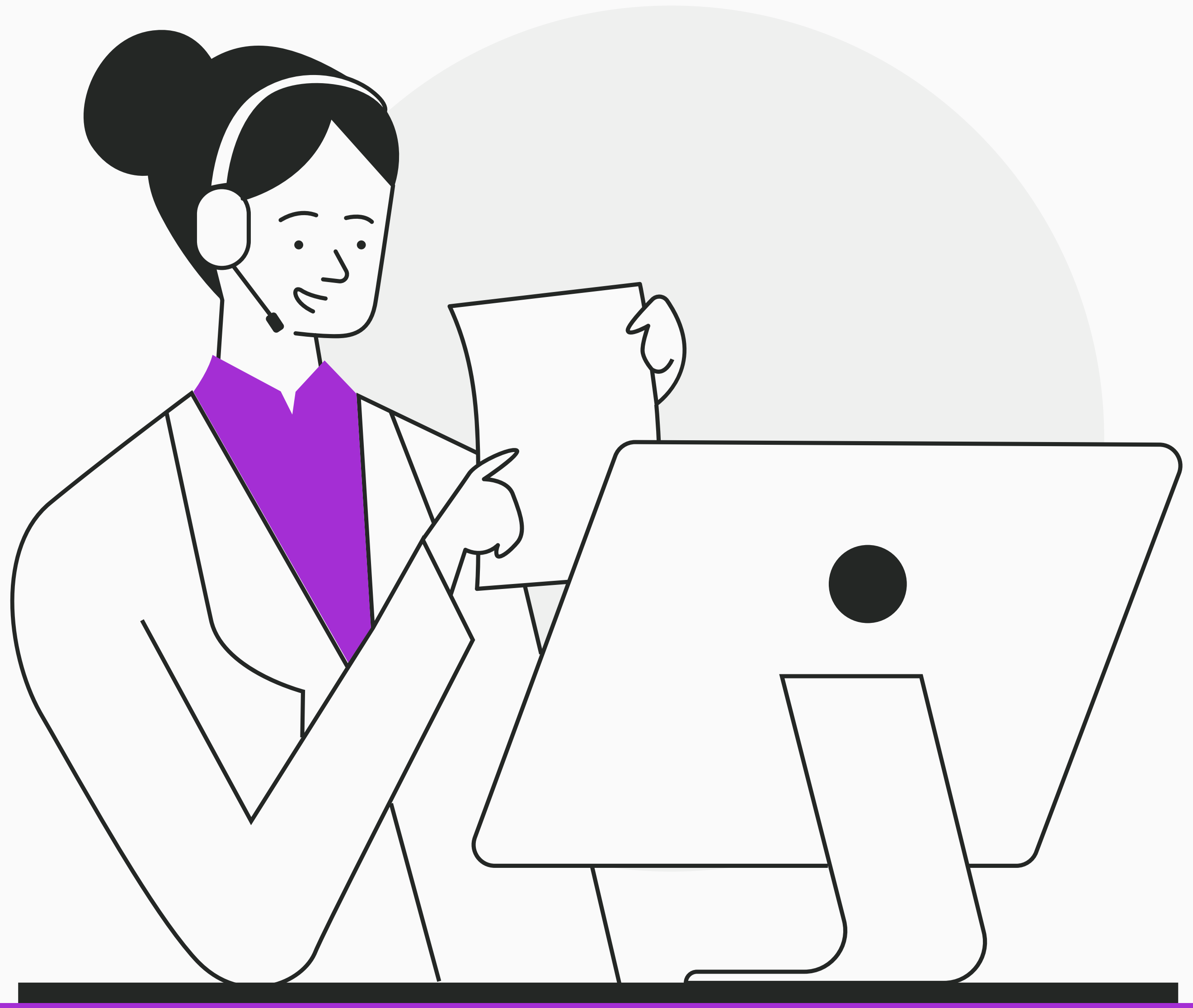
| | Predicted Non-Job Seeker | Predicted Job Seeker |
|---|---|---|
| Actual Non-Job Seeker | 3610 | 501 |
| Actual Job Seeker | 654 | 640 |

**SMOTE Logistic Regression**

| | Predicted Non-Job Seeker | Predicted Job Seeker |
|---|---|---|
| Actual Non-Job Seeker | 3202 | 909 |
| Actual Job Seeker | 331 | 963 |

# Outcome Summary

08



### GENDER

All measures demonstrate reasonable parity across gender groups, indicating no bias.

### RELEVANT EXPERIENCE

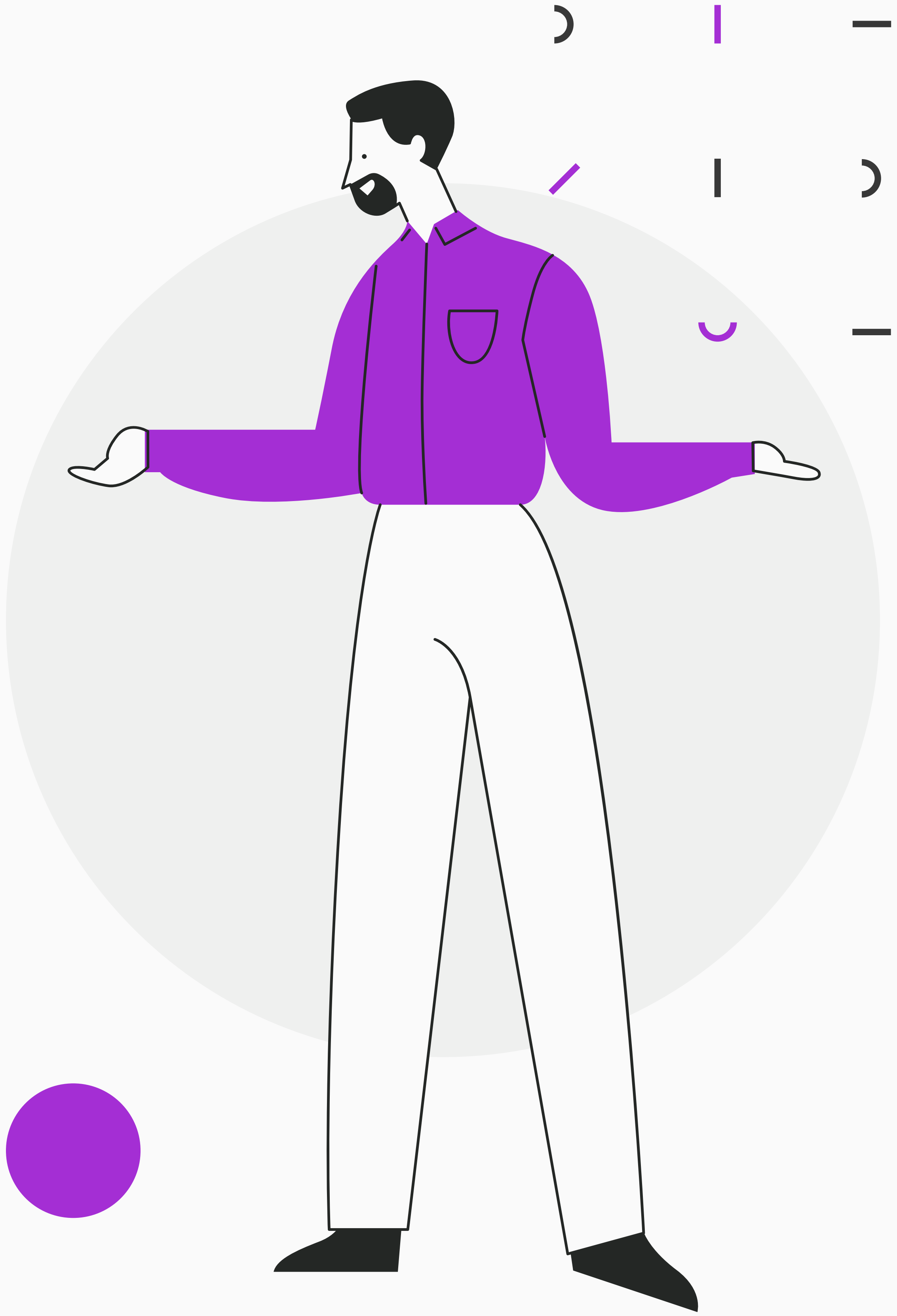
The ADS has the tendency to label those with no relevant job experience as “job seekers” and those with relevant experience as "non-job_seeker."

### EDUCATION LEVEL

The ADS is highly likely to label those with an undergraduate degree as "job-seekers", and those with high school diplomas as non-job-seekers, potentially taking opportunities away from the latter group.

## EXPERIENCE

The more experience a person has, the less likely the ADS will think they are job-seeking. This puts older people at a disadvantage. Since they are not viewed as job-seeking, they will miss out on job opportunities and potential promotion.



## CITY DEVELOPMENT

Being in a relatively less-developed city automatically makes the ADS think of a data scientist as a ”job-seeker”. Such a bias will bring benefits to those living in those relatively less developed cities, but companies will waste resources as a result and potentially lose employees to competing firms in more developed cities .

# Key Points

**THE ADS IS ACCURATE, YET BIASED**

While we believe that the ADS has high accuracy, it discriminates candidates based on experience (which correlates with age), location (city-development), and education.

Make sure you do enough research to support your points. It’s also a good idea to pair data with visual aids like charts, graphs, or images.

**MORE BALANCED DATA WILL IMPROVE THE MODEL**

We recommend collecting or synthesizing more data, especially from those with low education.

# References

Chakraborty, J., Majumder, S. & Menzies, T. Bias in Machine Learning Software: Why? How? What to Do?. Proceedings Of The 29th ACM Joint Meeting On European Software Engineering Conference And Symposium On The Foundations Of Software Engineering. pp. 429-440 (2021),

Chawla, N., Bowyer, K., Hall, L. & Kegelmeyer, W. SMOTE: Synthetic Minority Over-sampling Technique. Journal Of Artificial Intelligence Research.

Joshuaswords HR data visualization & prediction. Kaggle. (April 2021)

Yan, S., Kao, H. & Ferrara, E. Fair Class Balancing: Enhancing Model Fairness without Observing Sensitive Attributes. Proceedings Of The 29th ACM International Conference On Information & Knowledge Management. pp. 1715-1724 (2020)