Context

A company wants to identify true job seekers from the resource pool it maintains. The project is designed to understand signs that suggest a person is seeking new jobs and help HR identify true job seekers.

Data

The data is from kaggle. The data includes the candidate's demographics, education, experience and location data. A sample row from the data has a candidate's relevant experience, enrolled university, education level, major discipline, experience, current company and size, city, training hours and gender. There are a number of categorical features, I convert them into numbers according to the natural order.

Algorithms

Feature Engineering

1. Map eight categorical data to numeric values
2. Converting major into binary flag
3. Fill in missing feature data with zero
4. Data scale to fit logistic regression model
5. Drop less relevant features
6. Implement class weight to handle imbalanced target

Models

Logistic regression was used as a baseline model. The model doesn't fit well. A lot of false job seekers are miscategorized as positive.

- Precision: 0.55
- Recall: 0.27
- F1: 0.36

Gradient boosting classifiers, k-nearest neighbors, random forest classifiers and ensembling (voting classifier and stacking classifier) were used before settling on random forest. Random forest was chosen for the performance and interpretability.

Model Evaluation and Selection

The metric for evaluation is F1 score. The target has over 76% zero class, which makes accuracy a not good measure. The ideal model will have a balanced precision and recall score. This means the model should not miss too many true job seekers and do not miscategorize non-job seekers as a good candidate. This way the HR department can focus on the true job seekers, reducing time and cost of chasing people who do not plan to change jobs.

Final random forest (tuned with class weights) score

- Precision 0.54
- Recall 0.71
- F1 0.61

Tools

- Numpy and Pandas for data manipulation and exploration
- Scikit-learn for modeling
- Matplotlib and Seaborn for data visualization