HR Analytics:

Likelihood a Data Scientist is Looking for a Job Change

I. CONTEXT:

Company X offers a few training courses related to data science and big data. Many people enroll in these courses. Of those who successfully complete and pass the course, only some are actively looking for a job change. Company X has collected demographic, educational, and career information on all course graduates in order to better understand which individuals are more likely to be looking for a job change.

As resources are limited, Company X wants to reduce the cost and time of reaching out to candidates who are less likely to be interested in working for them. The aim of this project was to build a predictive model that determines which candidates are most likely to be looking for a job change, prioritizing reducing *Type II errors*. Additionally, the results from this project might add more insight into which types of workers are more likely to be open to a job change in general.

II. DATA

a. Raw Data

The data, retrieved from Möbius on kaggle, was pre-split into separate training (19,158 rows) and testing (2,129) datasets, with each row representing a potential candidate for the company. There were 14 columns, representing:

a) Demographic Information:
- Gender
- City (Coded)
- City Development Index

b) Educational Background:
- College Major
- Training Hours at Company X
- Education Level
- College Enrollment (Part-Time, Full-Time, Not Enrolled)

c) Work Experience
- Has Relevant Experience (Yes or No)
- Years of Experience
- Years Since Last New Job
- Current Company Type
- Current Company Size

Many of the features were categorical, some with high cardinality. The **target feature** was binary, with 1 representing that the candidate was looking for a job change, and 0 representing that they were not. This was an imbalanced class, as only about 25% of candidates in our training set were actively looking for a job change.

b. Data Cleaning

Some modifications were made to the original dataset:

- **Experience**, originally listed by year, was binned into six possible ranges. The binned groups were not entirely equal in size, with 5,904 candidates in the 2-5 years range, and only 1,201 candidates in the ≤ 1 year range.
    - ≤ 1 year
    - 2 - 5 years
    - 6 - 10 years
    - 11 - 15 years
    - 16 - 20 years
    - > 20 years
- **City**, a column with 123 different possible values, was limited to cities that represented at least 1% of total candidates. All others were binned into an 'Other' category.
- Rows with 4 or more missing values (out of 14 total columns) were dropped. This totalled to 878 rows.

c. Data Transformation

Prior to model training, data was transformed through the following means:

- Missing values of **Gender** (approximately 20% of the column) were imputed as 'Unknown'.
- Both **Company Size** and **Company Type** were missing about 30% of data. For candidates who had never held a previous job, this was imputed as 'N/A'. For all others, this was imputed as 'Unknown', as it was not certain whether this information was missing due to the candidate not being currently employed, or due to other reasons.
- Missing values of **Major** were filled in with 'N/A' for candidates who did not have a college-level or higher education (which explained ~95% of the missing values for this column).
- **Education level**, **company size**, **last new job**, and **experience range** were encoded using an ordinal encoder.

- **Gender**, **city group**, **enrollment status**, **major**, and **company type** were encoded using a One-Hot Encoder.
- The remaining null values were imputed using a KNN-imputer.
- Data was scaled using a standard scaler.
- For some models, the minority class (1 - looking for a job change) was oversampled using SMOTE.

III. MODELING:

Four general types of models were trained and tested:

1) Logistic Regression
2) Random Forest Classifier
3) Support Vector Machine (Linear and RBF Kernels)
4) Gradient Boost Classifier (XGBoost).

Due to the imbalanced nature of the target feature (only ~25% positive case), models were tested using either class weights parameters or oversampling of the minority class using the **imbalanced-learn** library. As the goal was to reduce Type II errors, each model was also tuned to optimize F0.5 using Grid Search CV. Mean model metrics from cross-validation are recorded in the table below.

Model	Precision	Recall	Accuracy	F 0.5	AUC
Baseline: Dummy Classifier	0.00	0.00	75.29	0.00	.50
Logistic Regression (Class Weights 0.4 : 0.6)	59.16	55.67	79.54	58.42	.797
Random Forest (Class Weights 0.4 : 0.6)	58.95	62.90	80.23	59.89	.800
SVC - Linear (Class Weights 0.4 : 0.6)	59.16	54.10	79.42	58.07	.797
SVC - RBF (Class Weights 0.4 : 0.6)	57.91	59.45	79.29	58.20	.790
XGBoost	**60.10**	**59.92**	**80.24**	**60.04**	**.796**

The **XGBoost model** trained without oversampling of the minority class was found to have the highest F 0.5, precision, and accuracy, although performance did not differ significantly between most models. The training and fitting time of this model was noticeably faster than

the Random Forest and SVC (RBF) model, offering it an advantage in this regard. It was selected as our best model, and saved in order to be deployed by Company X.

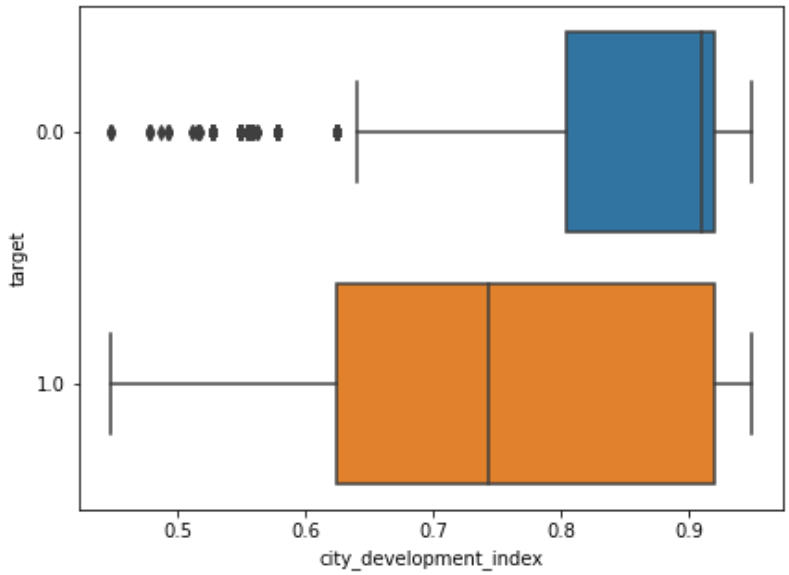
IV. KEY FINDINGS:

Through exploratory data analysis and modeling, we identified these attributes to have the most impact in predicting openness to a job change:

- City development index
- Years of experience
- Current company information missing
- Lack of relevant experience
- Current full-time university student

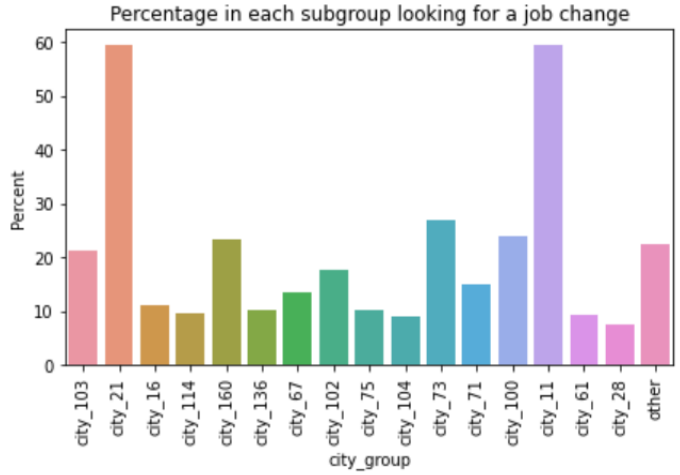
a. City and CDI

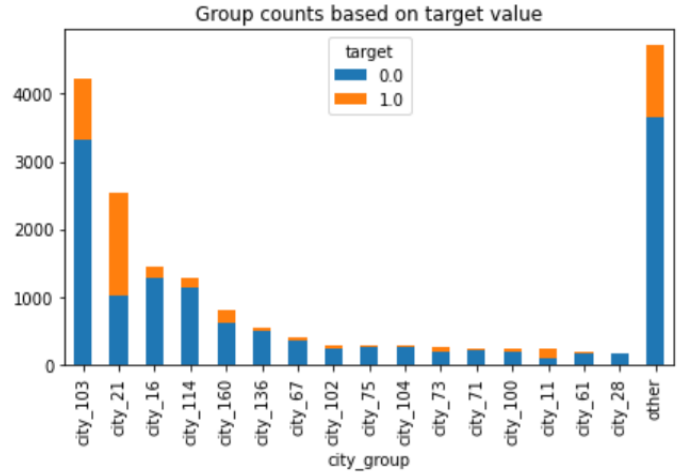
City-based information seems to be the most influential predictor in our model, with candidates from cities with higher CDIs being much less likely to be interested in working for Company X. This may be because these candidates are older and more established in their careers, as CDI showed a somewhat positive correlation with years of experience. As seen in the figure below, the median CDI of candidates uninterested in a job change is 0.91, while candidates interested in a job change come from a wider range of CDIs (median = 0.74).



Further trends based on city type can be seen by plotting each city against the target. In the left figure, it is clear that City_21 (CDI = 0.62) and City_11 (CDI = 0.55) contain the highest

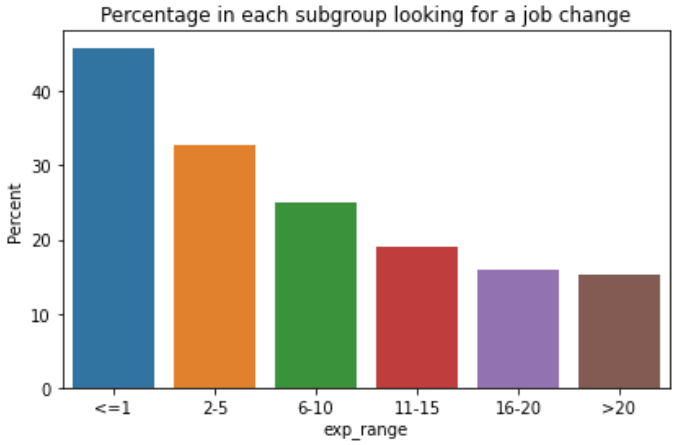
percentage of candidates interested in working for Company X, and are also the two cities with the lowest CDI. City_21 also represents the second largest group of candidates

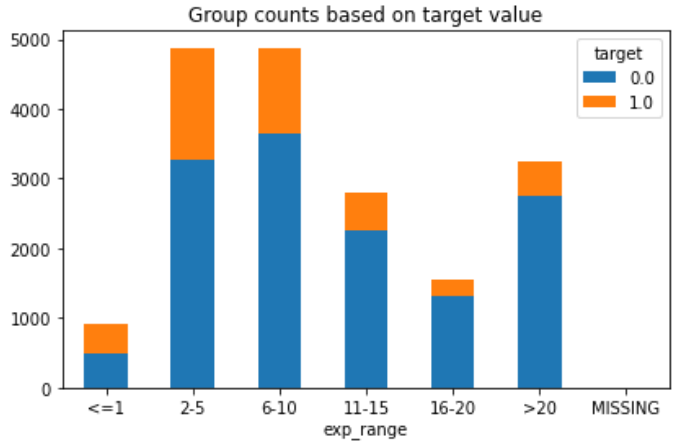




b. Experience Range

Experience range was another strong predictor of the target. The majority of candidates in the dataset possessed 2+ years of experience, with most having between 2-10 years. In the left figure, there is a steady trend in which the more years of experience a candidate has, the more unlikely they are to be looking for a job change. About 46% of candidates with 1 or fewer years of experience were interested in working for Company X, contrasted with 15% of those with more than 20 years of experience.

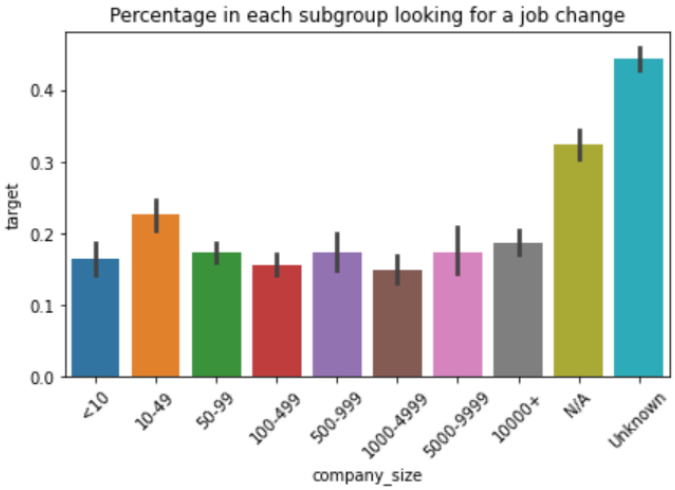


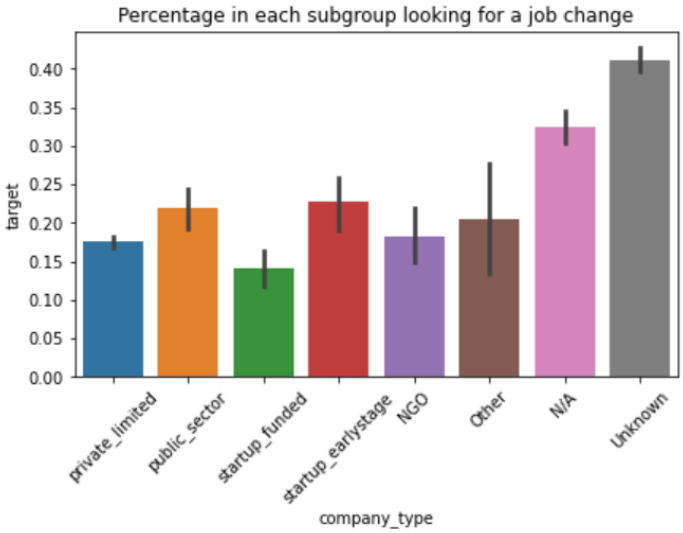


c. No Current Company

Additionally, candidates with missing current company information were more likely to be open to a job change. Candidates labeled as 'N/A', meaning no previous work experience, were more likely to be open to a job change than currently employed candidates, but were still

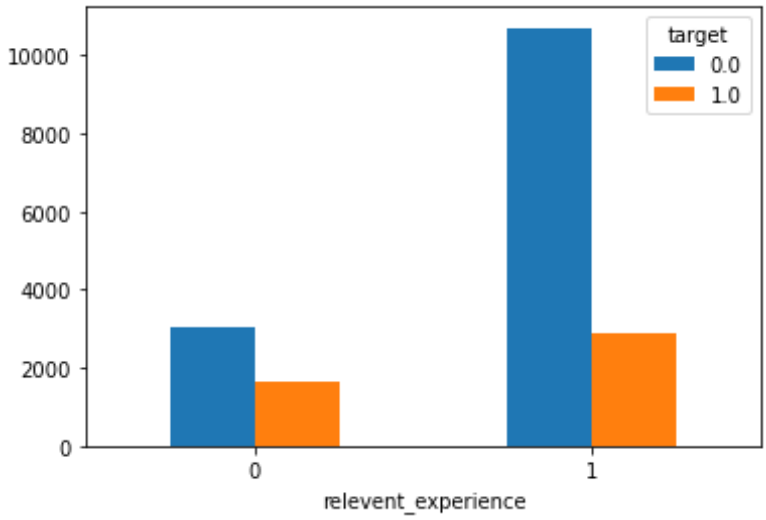
outnumbered by candidates with 'Unknown' work information. This may be because candidates with no former work experience are more likely to be college students, and thus less urgently in need of immediate full-time employment.





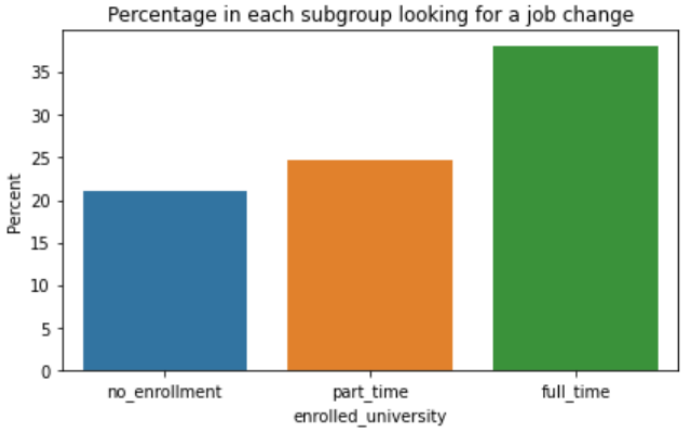
d. Lack of Relevant Experience

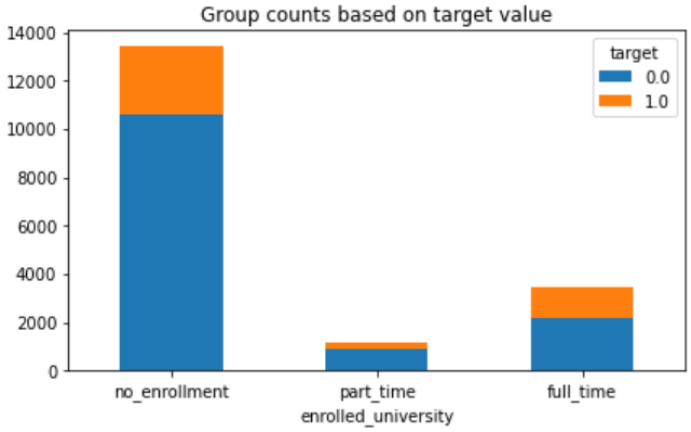
Most candidates do have relevant experience within the field, and unsurprisingly, candidates who do not have relevant experience were more willing to work for Company X. The majority of candidates in both groups, however, were not looking for a job change.



e. Full-time University Student

The vast majority of candidates are not currently enrolled in a university (73.2%). Full-time students were about twice as likely as non-students to be open to working at Company X, supporting the trend found in section **c. No Current Company**.





V. CONCLUSION:

In summary, our best model was able to predict a candidate's willingness to work for Company X with reasonably high accuracy, despite the imbalance nature of the target. Some key features were identified to be a) city development index, b) experience range, c) no current employment listed, which are all fairly logical conclusions.

A primary limitation of this project is the unexplained missing values in Company Size and Company Type. Although it seems likely that candidates missing this information are not currently employed (in some cases, due to being a full-time student), it would be beneficial to have more certainty on this factor. Or, preferably, have more data on the type and length of unemployment.

Additionally, this dataset does not specify whether a candidate's years of experience pertain primarily to data science or another field. Relevant experience is recorded only as a binary value (yes or no), and therefore limits our ability to analyze candidate background, although this information would likely have a strong impact on whether or not a candidate is looking for a job change.

Among the previous suggestions, it would also be useful to collect more data on each city, as city background was one of our main predictors of the target variable. This could include standard socioeconomic and demographic information, as well as distance from Company X, proximity to nearby universities, etc.