Job Change Prediction for Data Scientists

Man Shi

1. Business Problem

Data scientist has been a popular occupation in recent years, a company which is active in Big Data and Data Science wants to hire data scientists and many people signed up for their training. Company wants to know which of these candidates really wants to work for the company after training or looking for a new employment because it helps to reduce the cost and time as well as the quality of training or planning the courses and categorization of candidates. Information related to demographics, education, experience are in hands from candidates signup and enrollment.

In this project, we will explore the factors that lead a person to leave their current occupation through Exploratory Data Analysis and predictive models. We would like to forecast whether a candidate will look for a new job or work for the company as well as interpreting affected factors on employees' decisions.

2. Dataset

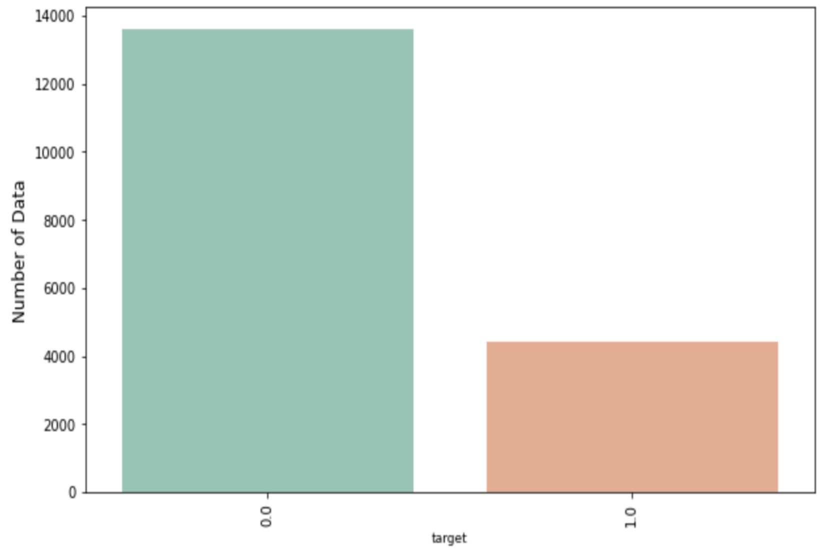
The dataset was from Kaggle: https://www.kaggle.com/arashnic/hr-analytics-job-change-of-data-scientists

The dataset consists of the records of candidates who signed up for the training of the company. The entire dataset was classified into train data and test data. The dataset we used contains 19185 observations with 'target' label (0: not looking for job change; 1: looking for a job change). Besides, there are 13 features correlated to the information of candidates such as their gender, whether they have relevant experience or not, their education level etc. The reason we chose this dataset is that it gathered the comprehensive information associated with the candidates' occupation, which would be very useful for finding patterns of decision making for a job. The information of variables is shown below.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 19158 entries, 0 to 19157
Data columns (total 14 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   enrollee_id             19158 non-null  int64
 1   city                    19158 non-null  object
 2   city_development_index  19158 non-null  float64
 3   gender                  14650 non-null  object
 4   relevent_experience     19158 non-null  object
 5   enrolled_university     18772 non-null  object
 6   education_level         18698 non-null  object
 7   major_discipline        16345 non-null  object
 8   experience              19093 non-null  object
 9   company_size            13220 non-null  object
 10  company_type            13018 non-null  object
 11  last_new_job            18735 non-null  object
 12  training_hours          19158 non-null  int64
 13  target                  19158 non-null  float64
dtypes: float64(2), int64(2), object(10)
memory usage: 2.0+ MB
```

3. Exploratory Data Analysis

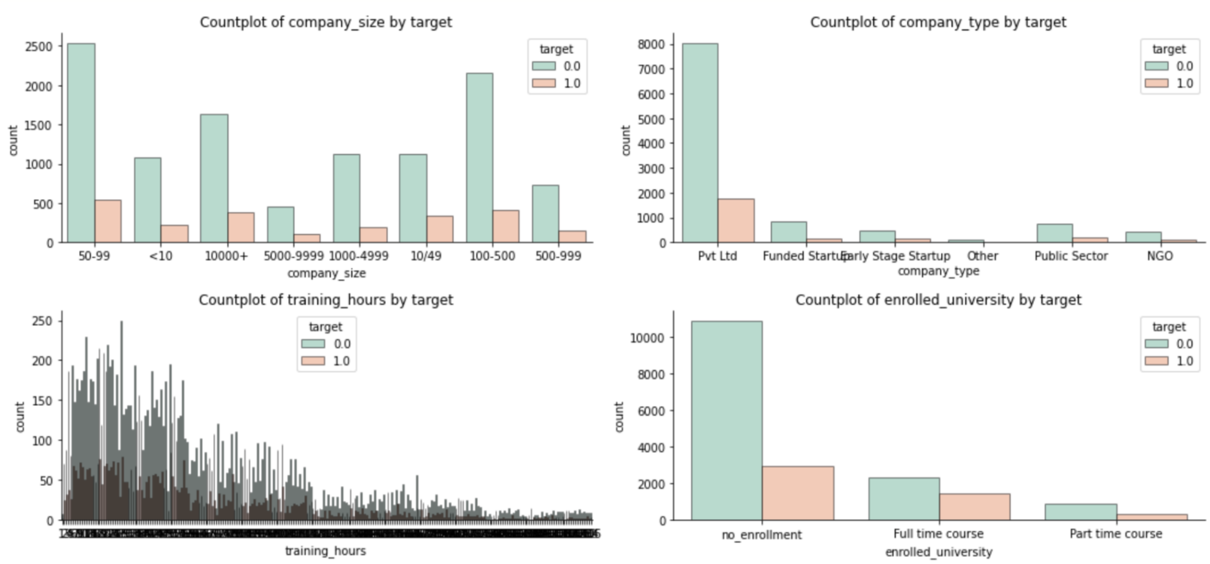
1) The count of target data ('target')

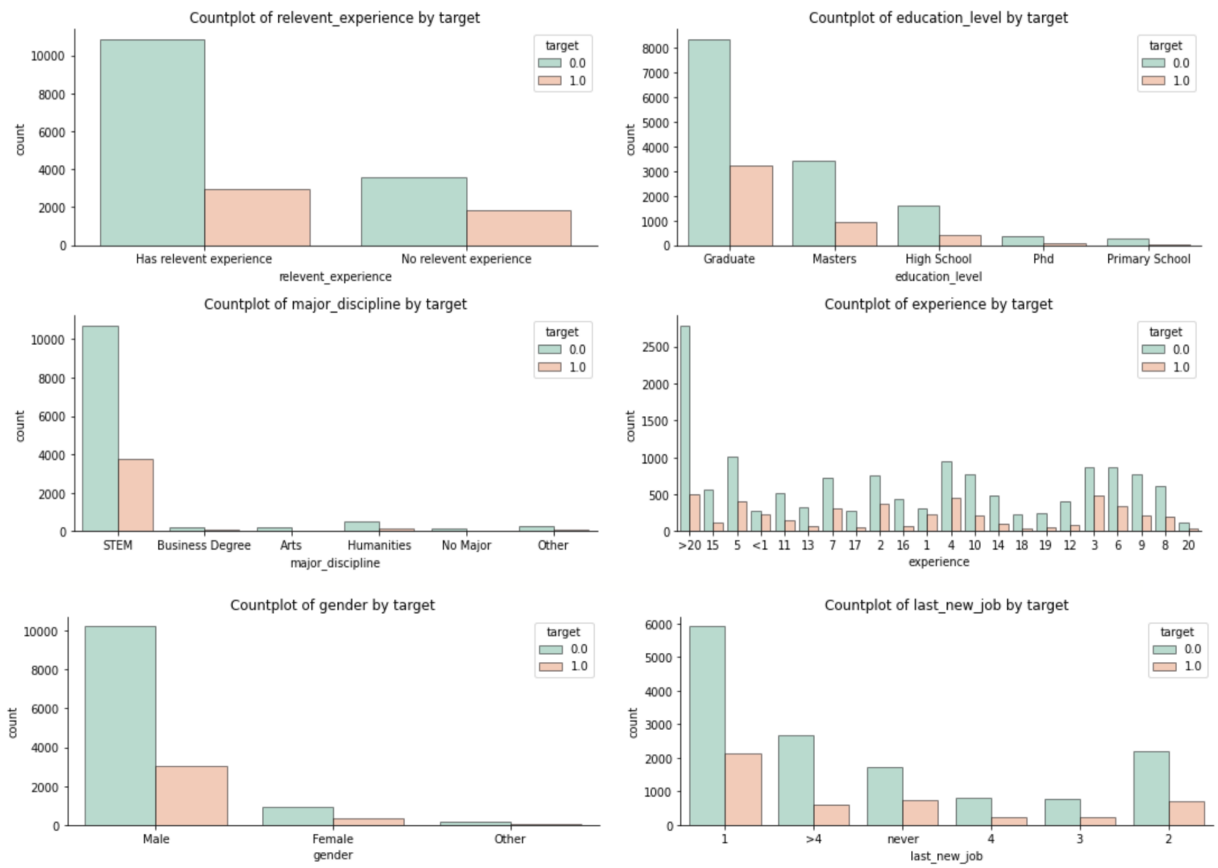



```
0.0    13593
1.0     4421
Name: target, dtype: int64
```

From the bar plot above, we can see that the number of 1 (Looking for a job change) < 0 (Not looking for a job change).

2) The frequency of each category separated by label ('target')



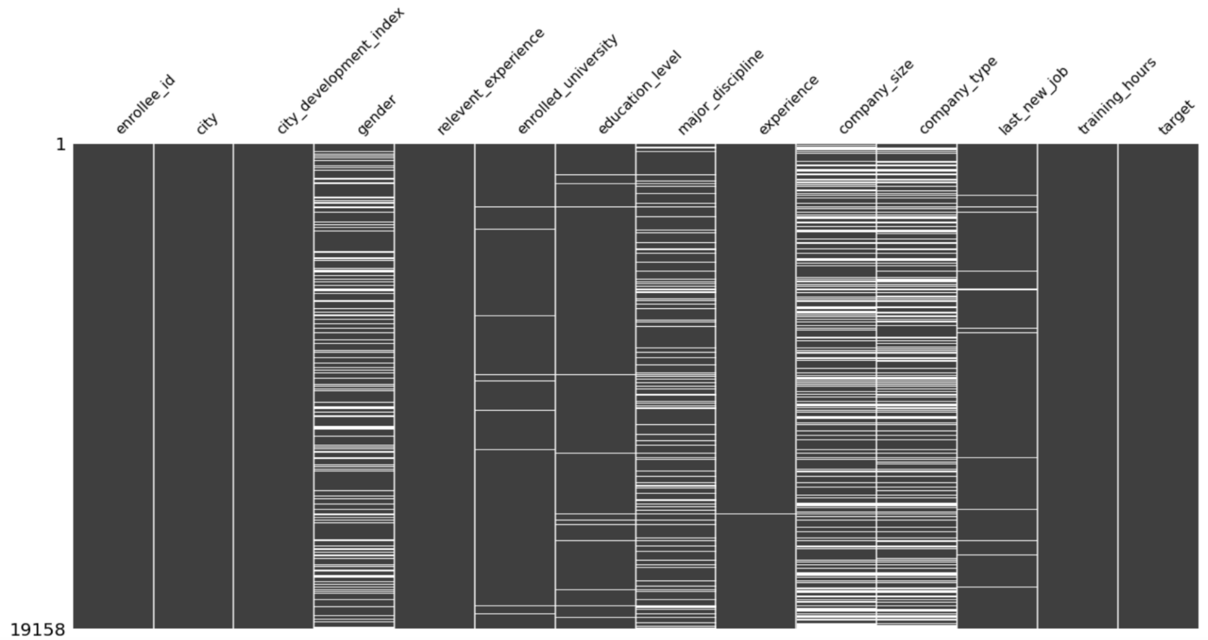


From the graphs shown above, we can see the distribution of employees who look for a job change and who do not look for a job change in different sub-group categories. Based on the results, we notice that people who work for a small or middle company, who equip relevant experience, who only have one year between previous job and current job would be more likely to change their jobs. According to these meaningful findings through data visualization, we could know more about the dataset and it's helpful to further operate the modeling part.

4. Data Preprocessing

1) Check and deal with missing values

```
enrollee_id               0.000000
city                      0.000000
city_development_index    0.000000
gender                    0.235306
relevent_experience       0.000000
enrolled_university       0.020148
education_level           0.024011
major_discipline          0.146832
experience                0.003393
company_size              0.309949
company_type              0.320493
last_new_job              0.022080
training_hours            0.000000
target                    0.000000
dtype: float64
```



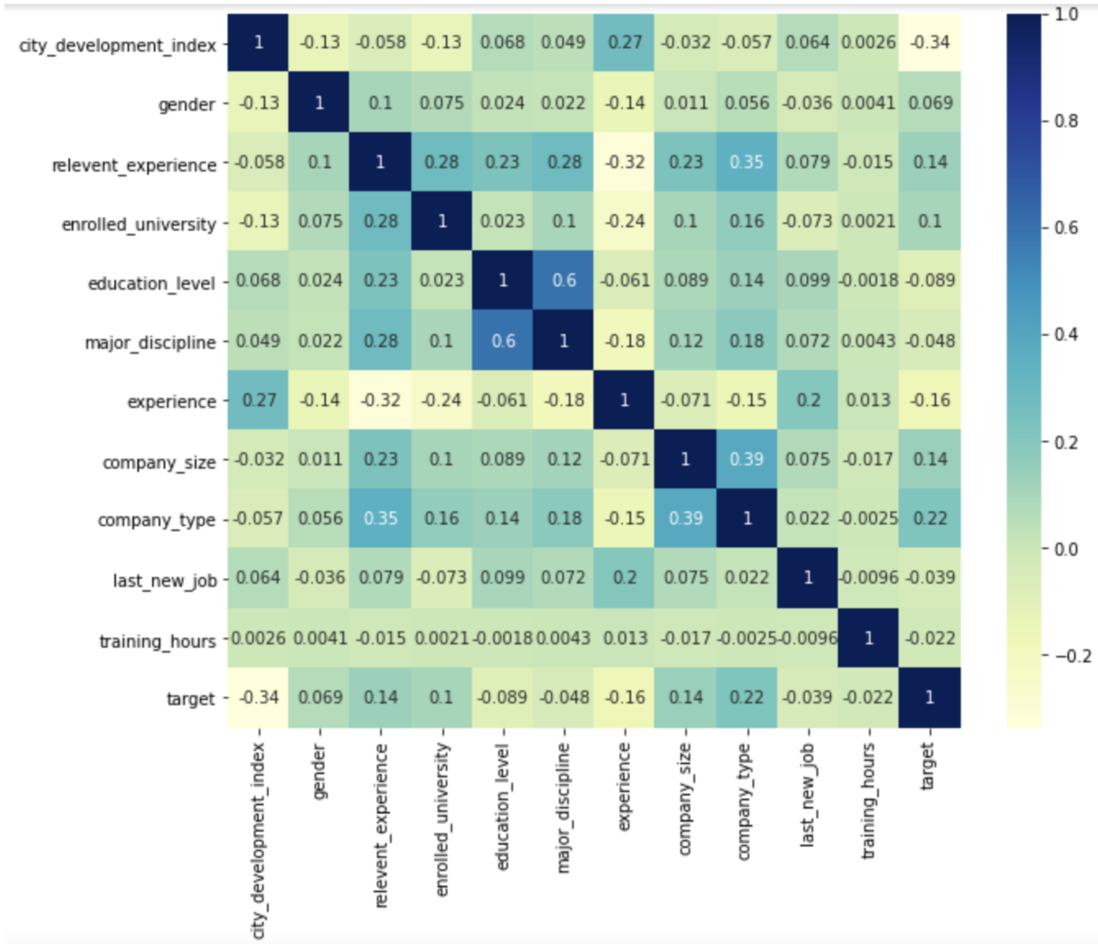
From the above proportion of NAs in each column and the null value graph, we can notice that the variables 'experience', 'enrolled_university', 'last_new_job' and 'education_level' contain relatively few missing values, we can just drop null from our dataset. Besides, we would like to fill the null with 'Unknown' for the variables ' major_discipline', 'company_size', 'company_type' and 'gender' as they have more than 10% null values.

2) Handling categorical variables
Since there exist several categorical variables in the dataset, before constructing the machine learning models we need to handle them first. Here we transfer the type of categories from 'object' into 'int' with different levels.

5. Analytical Findings - Predictive models

1) Prepare the dataset that used in models
First we should check the correlation among variables, the heatmap of the correlation matrix is shown below. From the graph below we can see that the 'relevent_experience', 'enrolled_university', 'company_size' and 'company_type' slightly correlated with the 'target'. On the other hand, the feature 'education_level' shows moderate correlation with 'major_discipline', meaning that we need to remove one of them to delineate multicollinearity. Here we drop the 'education_level' column from the dataset.

Then we randomly split the dataset into 80% train set and 20% test set for fitting the predictive models and forecasting the probability of employees looking for a job change.



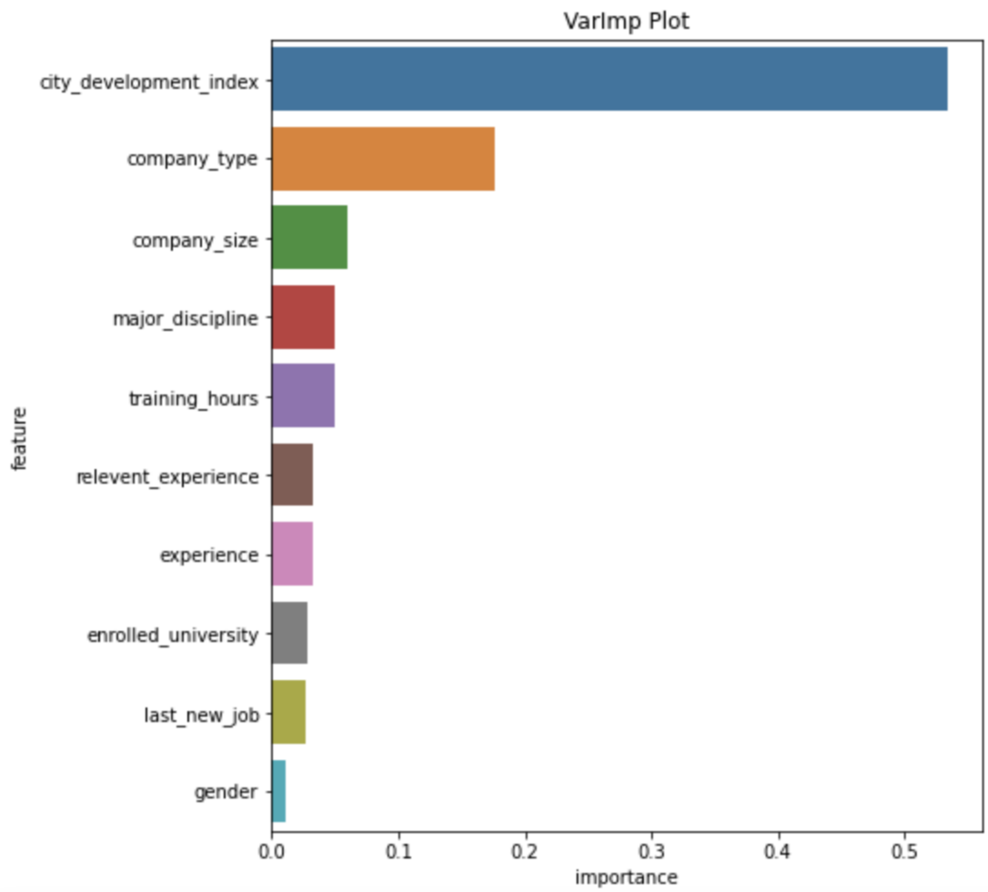
2) Building the machine learning models
We construct 4 predictive models based on scikit-learn to fit with the train set and forecast the response with the test set, the models include Logistic Regression, Random Forest, K-Nearest-Neighbors and Support Vector Machine models.

3) Model performance evaluation
The below table shows the accuracy score of each predictive model, we can notice that the Random Forest model achieves the highest accuracy which exceeds 0.78.

	Model Result	Models
0	0.729947	Logistic Regression
1	0.784901	Random Forest
2	0.750763	KNN
3	0.747710	SVM

4) Important Features

From the feature importance score graph for Random Forest model shown below, we notice that the ‘city_development_index’ and ‘company_type’ are two variables which have the most significant effect on determining whether people would change their job or not.



6. Summary

According to the analysis we did, we can conclude that the Random Forest model has the highest accuracy score for forecasting the possibility of people looking for a job change (0.78), and the most two important factors that affect people whether changing jobs or not are the development status of the cities where they worked and the type of the company they worked for.

The predictive models could also be improved by tuning parameters and deeper feature engineering for future research, methods for resolving imbalance dataset would also be applied to increase the predictive accuracy and model performance.