Sawyer Jacobson

University of Denver

Ritchie School of Engineering

Data Science Tools II

3/15/2021

HR Analytics: Looking for a New Job?

**Introduction**

Recruiting data science talent has become a key priority for some companies due to the recent rise in popularity of the data science profession. There is no set path for one to start a career in data science. One could take bootcamps, courses, or a formal university education program. For a particular Big Data and Data Analytics company, offering their own data science courses and recruiting those that successfully complete them is an efficient recruitment model. However, not everyone that takes or completes their courses is, in the end, looking for a new job. Utilizing collected professional and demographic data from the candidates, the HR department of this company would like to use data science and machine learning methods to determine which candidates are likely to be looking for a new job using the collected data. The ability to accurately predict which candidates are seeking new employment would allow the company to better evaluate their courses and how and where to invest in them to optimize the course content and the overall benefit of offering the courses.

The dataset used for this project is an HR analytics job change dataset obtained from Kaggle.com[[1]](#footnote-1). The research question is how accurately can a candidate be classified as looking for a new job or not given the demographic and profession data available? Various classification machine learning methods will be utilized in our attempt to answer this question.

**Data and EDA**

The original dataset obtained from Kaggle.com was already split into both training and test datasets. For this project, these data were combined, and new training and test data were separated at a 70/30 train/test split. The demographic variables include a unique ID for the candidate, an anonymized city code, a scaled development index for said city, the gender of the candidate, whether or not the candidate had relevant experience, the type of university course the candidate was enrolled in if any, their education major discipline, and the total number years of experience. The candidate and city IDs will be dropped from all analyses as they do not convey useful information. The professional variables include information about current employment such as current company size, company type (i.e. startup, corporation, etc.), the difference in years between their previous and current job if relevant. The number of training hours of course time is also included. The target variable is binary of whether or not the candidate is looking for a job change. A total of 11 features will be used in modeling the binary target variable.

The exploratory data analysis was performed only using the training data so as not to expose any information about the test data. The variables are mostly categorical in nature with both ordinal and nominal variables included. These categorical variables are explored with barplots colored by the target variable, and the continuous variables are displayed with density plots also colored by the target.

In the plots below, we see that there is an imbalance in the target variable with only about a quarter of candidates looking for a new job. The scaled city development index tends to be lower for those looking for a new job potentially indicating they are looking to move to a better city. Training hours are roughly the same between the 2 target groups. More graduate or master’s students were seeking new employment than all other education levels, and more STEM majors and males were seeking new employment than others in their respective categories. Lastly, more candidates were looking for jobs were either not employed at a company or were working for a private company, and many were in their first year since changing jobs likely seeking to gain a boost in their new field.

Chart, bar chart

Description automatically generatedChart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart

Description automatically generated

**Data Preprocessing**

The majority of the variables in this dataset are categorical. For use in most machine learning models, these variables will have to be encoded in one way or another. Nominal categorical variables will simply be encoded using One-Hot Encoding, creating a binary variable for each variable level. Ordinal variables will be encoding using an ordinal encoder with the correct order specified. The numeric variables tend to show significant skew and different scales and will be transformed using a standard scaler as models that use distance metrics require similar scaling.

Years of experience and difference in years between previous and current job are inherently numeric but are presented as categorical due to a ceiling category. Experience will be transformed to numeric with 21 replacing those with >20 years of experience, and missing values were converted to 0s. We note that there are numerous values of >20 because of the experience ceiling. Years since last new job will also be converted to numeric with a cap at 5 as there is not information on those >4 with Never and missing values converted to 0s. There are certainly limitations to converting variables in this fashion, but given the data, we believe this is the best method to move forward.

More than half of the variables in this dataset contained missing values. However, in nearly all cases, the absence of data is still useful data. The data are not missing completely at random. Imputation methods were considered, but the missing data will be assumed to be either its own category or reassigned to the lowest variable level where appropriate. The methods for handling these missing values follow below.

Missing values for company size, type, and time since the last new job will be assumed to mean that the enrollee doesn't work for a company and hasn't had a job to fill out. It is also reasonable to assume missing values for Major Discipline mean the enrollee did not attend or finish college. Missing values for gender will considered a separate value for gender since it is possible the enrollee does not align with a binary gender or the answers to the question did not represent them properly. Missing values for university enrollment, level of education, experience, major discipline, company size, company type, etc. will be considered as either the lowest ordinal level, its own separate value, a value that could reasonably represent missing data, or 0 appropriately dependent on the type of variable. Based on the variables and their meaning in this dataset, it is reasonable to assume that missing data should not be thrown out.

**Model Selection and Evaluation**

As stated above, the data were split into training and test sets based on a 70/30 split leaving 15934 training observations and 5353 test observations. The preprocessing described above using the training data as reference was also applied to the test data. The models were further optimized using 5-fold Cross Validation (CV) on the training set while using the test set to finally evaluate the optimized models.

For this classification problem, the machine learning algorithms implemented included the robust Random Forest, the classic Logistic Regression, and the non-parametric K-nearest neighbors (KNN). Each model was first fit on only the training set and evaluated on the test set to gain a baseline performance for the out of the box hyperparameter settings. Afterwards, each model was optimized using 5-fold CV on a grid of prespecified hyperparameters. A snippet of the model fitting process can be seen in the figure below.

Text

Description automatically generated

The data are first preprocessed using the given transformer pipeline. Then each model is looped through performing a CV on the given hyperparameters to find the optimal model. Afterwards, the best model is evaluated and returned along with the data and predictions.

After these models were optimized, an ensemble model in the form of a Voting Classifier using hard majority rule settings was fit using the models to return new predictions. Ensemble models can be great for prediction under the idea that the combined models will assist in evening out the bias-variance tradeoff. However, the ensemble model is only as good as the underlying models that comprise it so the performance will be diminished if simple, inaccurate predictors are used.

Chart, bar chart

Description automatically generated

The test accuracy for each model can be seen in the figure above. While not displayed, each model performed better with optimized hyperparameters than baseline. The bars are sorted in descending order. Random Forest was the most accurate model on our test data with the ensemble Voting Classifier coming in close second. The dotted line represents the baseline prediction accuracy of predicting the majority class: not looking for a job. Each model was able to outperform that baseline although not by a large margin. Model evaluation metrics for classification problems can be dependent on what error is desired to be minimized. Below, a plot of the confusion matrix for each model predictions can be seen to attain a better understanding of where each model made their prediction errors.

Graphical user interface, application, Teams

Description automatically generated

In the labels above, 1.0 is a candidate looking for a job, and 0.0 is not looking for a job. A key takeaway from this figure is that Random Forest was largely more accurate in correctly predicting candidates looking for jobs than each other model, but Random Forest also falsely predicted job seekers at a higher rate than all of the other models. Optimizing a different evaluation metric may produce better results.

Additionally, using feature importance of model such as Random Forest or parameter estimates for Logistic Regression would allow HR to determine which candidate attributes have the most pull in making predictions.

Chart

Description automatically generated

In the above plot, we see that the city development index, company size, training hours, and experience were the most important prediction variables based on the Gini index criterion.

**Conclusion**

Overall, we were able to build a pipeline that correctly predicted a large number of candidates on their future job seeking aspirations. Being able to predict above the majority baseline is the minimum expected from an effective model, and the models fitted were able to beat this baseline accuracy by several percentage points thus would provide at least a minimal benefit to the companies HR department.

Further work could be done by working with the HR department directly in determining whether it is more important to minimize either false positives or false negatives in candidate predictions. The models could then be optimized on this new metric to potentially see better performance. The Receiver Operating Characteristic (ROC) and Area Under the Curve (AUC) could also be used in evaluation metrics to optimize the models. Optimizing the probability cutoff point could also potentially improve performance.

1. <https://www.kaggle.com/arashnic/hr-analytics-job-change-of-data-scientists> [↑](#footnote-ref-1)