#Enkhchimeg Tsendnyam

#HR Analytics: Job Change of Data Scientists

* ***Introduction to the problem***

The dataset I picked is on “*HR Analytics: Job Change of Data Scientists*” from Kaggle (<https://www.kaggle.com/arashnic/hr-analytics-job-change-of-data-scientists>). Company X is looking to hire data scientists who successfully pass their data science courses and trainings they offer. Company X’s HR department wanted to see which of these candidates really want to work for the company after completing the training and which of these candidates are taking advantage of their opportunity and in fact, they are not looking for employment with Company X. Some of the questions that I’m trying to answer through this dataset are :

* What do the variables look like? Are the numerical variables correlated?
* Are the distributions of categorical variables the same of different among looking for a job and not looking for a job? For example, do participants that have fewer degrees are looking for job in high number compared to the participants that have higher degrees?
* What are some of the factors that lead a person to leave their current job?

This is an important topic and problem to solve for company X’s HR department because they want to know whether their training offerings are helping with their recruitment. Also, since they have candidates’ information such as current credentials, demographics and experience, they want to analyze what are some of the current factors might lead a person to leave their current job.

* ***Organized and detailed summary of the three parts of your case study (include any important visuals in here)***

Some of the basic data summary findings are:

* There are 19158 candidates that participated in company X’s training offerings.
* Out of 19158 participants, 14650 are identified their gender on the survey. Clear majority of 90% that or 13221 are male.
* 71% of the participants or 13792 people have relevant data science experiences.
* 75% of the participants or 14492 people have majors in the STEM area.
* Over 60% of the participants or 11598 people have Graduate degrees.
* Only 27% of the participant or 5341 people are enrolled in part-time or full-time courses.
* Average number of training hours these participants completed were 65.
* Pearson correlation matrix did not find any correlation amongst the quantitative variables
* Some of the observations that are made in regard to what factors lead a person to leave their current job are:
  + Looking at the relevant experience variable, larger percentage of people who do not have the data science experience are looking to change their current jobs
  + Looking at the relevant education level variable, larger percentage of people who have graduate degrees are looking to change their current jobs. On the other hand, larger percentage of people who have have Phd or primary school degrees are not looking to change their current jobs
  + Looking at the relevant enrolled university variable, larger percentage of people who have full-time courses are looking to change their current jobs
  + Looking at the last new job variable, larger percentage of people who never had a before are looking for a job. Perhaps more recent graduates are looking to take this training opportunity and looking to get a job with company X.
  + Looking at the major discipline variable, larger percentage of people who have STEM degrees are looking to change their current jobs compared to the people who have non-STEM degrees.

**Supporting Graph Visualizations:**

A picture containing text, receipt

Description automatically generated

A picture containing chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart

Description automatically generated

Chart, treemap chart

Description automatically generated

Table

Description automatically generated

Chart, line chart

Description automatically generated

* ***Conclusion: What does the analysis/model building tell you? Do you have a recommendation based on these findings?***

When it comes to evaluating performance of the model, I used several different evaluation metrics such as Confusion matrix - Accuracy, Precision, Recall and F1 score, ROC Curve and multiple machine learning algorithms such as logistic regression, decision tree, random forest to try to predict whether someone would be a job seeker or not.

* First when looking at the confusion matrix to evaluate the performance of a classification model, I looked at the accuracy, precision, recall % F1 scores. Below are my TP, TN, FP and FN numbers:

1. True Positive (TP) =131; meaning 131 is an outcome where the model correctly predicts the positive(Looking to change a job) class
2. True Negative (TN) = 4212; meaning 4212 is a outcome where the model correctly predicts the negative(Not Looking to change a job) class
3. False Positive (FP) or Type I error= 105; meaning 105 is an outcome where the model incorrectly predicts the positive(Looking to change a job) class
4. False Negative (FN) = 1300; meaning 1300 is an outcome where the model incorrectly predicts the negative(Not looking to change a job) class

* Based on this, my Accuracy became = (4112+131)/(4212+131+105+1300)= 4343/5748=75%. I got **75%** of Accuracy for our model. Accuracy represent the model’s ability to correctly predict both the positives and negatives out of all the predictions. The model is saying I can predict the candidates who are looking for a job 75% of the time. It is also predicting that the candidates that are not looking for a job with 75% accuracy as well. Accuracy is one of the useful metrics to get the general idea but it’s not the most insightful metric when evaluating a performance of a model. Especially with 75% accuracy, it’s not impressive. So, we look at precision, recall and F1 score.
* Precision tells us the proportion of every observation predicted to be positive that is actually positive. In other words, how many of these candidates that said are looking for a job are actually are actually looking for are actually looking for a job? Precision = TP/ (TP +FP) = 131/(131+105) = **55%.** So, the classifier revealed only 55% of the people who said are looking to change jobs are actually looking for a job change.
* Recall is the proportion of every positive observation that is truly positive. In other words, of all these candidates that are looking for a job, how many are correctly classified as actually looking for a job?
* Recall = TP/(TP +FN) = 131/131+1300 = **9%.** So, when the classifier predicts that person is looking for a job change, only 9% of the time, the classifier is correct.
* As Albon pointed out, better metrics often involve using some balance of precision and recall – that is a trade-off between optimism and pessimism of our model. So, I further looked into F1 score. F1 score is a measure of correction achieved in positive direction – that is, of observations labeled as positive, how many actually positive. F1 score = 2x (Precision x Recall)/(Precision + Recall) = 2 x (0.55 x 0.09/ (0.55 +0.09)) = 2 x 0.0495/0.65 = **1.5%.** So in other words, 1.5% of the “Not looking for job change” candidates were classified incorrectly as “Looking for a job change” candidates.

Out of all the model, random forest had the highest precision, recall & f1 scores overall compared to the logistic regression and decision tree models:

Precision = 48%

Recall = 32%

F1 = 38%

* Based on these findings, I would recommend Random Forest algorithm for this data. When looking at the Receiver Operator Characteristics(ROC) metric, area under the curve(AUC) was 0.68. ROC Curves summarize the trade-off between the true positive rate and false positive rate for a predictive model using different probability thresholds. AUC being 0.68 detonates a poor classifier.