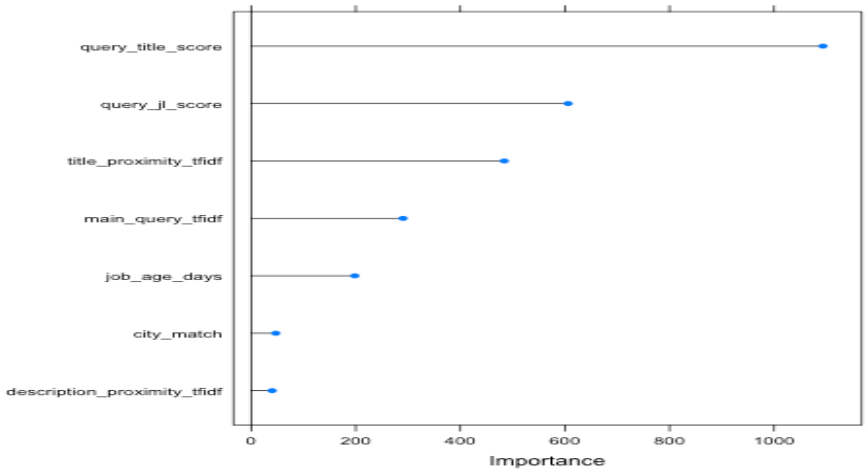
Prediction of apply rate

Xiangyu Liu

03/24/2018

For part 1: I run logistic regression model and gradient boosting model to predict whether the user will apply or not. For the logistic regression model, the accuracy rate is 90.91%; the AUC is 0.5869. For gradient boosting model, the accuracy rate is 90.91%; the AUC is 0.6043.

According to these two models, we can conclude the importance and impact of the variables. The top 5 important variables are query_title_score, query_jl_score, title_proximity_tfidf, main_query_tfidf, and job_age_days.



- The ***most important*** one is query_title_score, which measures the popularity of query and job title pair. If the pair of query and job title is more popular, then the user is more likely to apply. This may be because the trend of job market will guide people's career paths. For instance, data scientists and machine learning engineers are very popular positions due to high salary and large demand from different industries. As a result, new graduates would, to some extent, follow others's career choices - applying hot job positions.
- The ***second important*** factor is query_jl_score, which is similar to the first one. The popularity of query and job listing pair, including title names, job description, and technical requirement, reflects the trend of job market. The popular the job position (or techniques that this position requires) is, the more likely people apply.

- The ***third important*** variable is title_proximity_tfidf, which measures the closeness of query and job title. According to the logistic regression results, the more accurate the query is, the more likely the user will apply. The reason is that when our website can answer people's query accurately, people will have more job matches based on their preference and interest. Of course, they are more likely to apply.
- The ***fifth important*** variable is job_age_days. The negative relationship with the dependent variable, according to logistic regression results, is because the job listing may expire or no longer available. As a result, people will less likely to apply such positions.

Overall, we can **make people more likely to apply** on our website in these ways:
1) provide more accurate query results;
2) list keywords that are as popular as possible in the job listing (e.g. supply chain analyst is not a very popular position, but its job requirement includes analyzing data and writing metric reports, which is similar to data analyst. Thus, the title of this position can be changed to data analyst. This can attract more attention)
3) delete obsolete and expired job postings to reduce distraction

For part 2: I segment users based on their interests and redo the model training and testing on the observations where the class ID of job title is 10148. For the new logistic regression model, the accuracy rate is 90.91%; the AUC is 0.5518136. For gradient boosting model, the accuracy rate is 90.86%; the AUC is 0.5339. Therefore, we cannot achieve a better classification performance on this new data set.

Improvement: This result is somewhat dependent on the split of the data that I made earlier, therefore for a more precise score, I would be better off running some kind of cross validation such as *k-fold cross validation* for the next time.