On the Prediction of Apply Rate

O. Ozan Koyluoglu
ozan.koyluoglu@glassdoor.com

I. Introduction

The problem of interest is the prediction of apply rate. Imagine a user visiting Glassdoor, and performing a job search. From the set of displayed results, user clicks on certain ones that she is interested in, and after checking job descriptions, she further clicks on apply button therein to land in to an application page. The apply rate is defined as the fraction of applies (after visiting job description pages), and the goal is to predict this metric using the dataset described in the following section.

II. Dataset

Dataset is located at the URL http://bit.ly/applyratemarch2018. Each row in the dataset corresponds to a user's view of a job listing. It has 11 columns as described below.

1) $title\_proximity\_tfidf$: Measures the closeness of query and job title.
2) $description\_proximity\_tfidf$: Measures the closeness of query and job description.
3) $main\_query\_tfidf$: A score related to user query closeness to job title and job description.
4) $query\_jl\_score$: Measures the popularity of query and job listing pair.
5) $query\_title\_score$: Measures the popularity of query and job title pair.
6) $city\_match$: Indicates if the job listing matches to user (or, user-specified) location.
7) job_age_days: Indicates the age of job listing posted.
8) $apply$: Indicates if the user has applied for this job listing.
9) $search\_date\_pacific$: Date of the activity.
10) u_id: ID of user (for privacy reasons ID is anonymized).
11) $mgoc\_id$: Class ID of the job title clicked.

III. Analysis

Please use the "$search\_date\_pacific$" column (9-th column) to split the dataset into training and test dataset. Train your model(s) using the data between 01/21/2018-01/26/2018, and test your model on 01/27/2018.

Split the analysis into two parts:

- Focus on the first 7 columns. Use these as features to predict the 8-th column, "$apply$". Discuss the model you choose. Primarily focus on AUC as the metric of interest for your binary classifier. You can also investigate/discuss other metrics.
- Consider now adding the remaining two columns to your feature set. Is it possible to segment users ("u_id") based on their interests ("$mgoc\_id$"), and achieve a better classification performance?

IV. Deliverable

Please summarize your assumptions and findings using **at most two pages** and **submit your .pdf and (preferably python) code to the email address above with subject "ML Intern Summer 2018"**. Be reasonable with your presentation (i.e., font size, spacing, etc.), and focus on your major findings. Make sure to add a final section in your report that describes what could be done if you have more time/data to approach this problem.

Approximate time that this analysis should take is 3 hours.

Deadline is March 27, 2018 by 11:59pm PST. (Due to the amount of the applications we received, we will not be able to grant any extension on the deadline.)