

The data set includes 1200+ google jobs description. The variables include the title of the job, category, location, responsibility of the job, minimum and preferred qualifications.

The goal is to identify the important attributes or skills of the candidates for a particular job category. In general, this may help the candidates to match with the fitted jobs. Please use the machine learning methods to address this problem and justify your model/method. For instance, you may provide an estimate of the test error (or test MSE) of your method.

Since the data contains the full job description, you may have to define and extract your own features. You may apply the methods not covered in the class if you prefer. The grade is mostly based on whether the question is addressed and the methods are applied appropriately or not.

Format: Report should be submitted in a pdf file by December 1 at 6pm to blackboard. Report that is late receives 20% off within 24 hours late and 0 after that. The report has no more than 8 pages in normal format with font size 12. All the statistical analysis should be done in R (you may use other tools or text mining softwares to extract features). The R programs used for data analysis should be turned in together with the report (programs not counted into the 8 page).

Ideally, everybody should contribute equally in the project. In general, I will give the same grade to all group members. When submitting the report, please upload the file only through one student's account. This makes me easier to know which is the right version for grading. (Last year, the different versions are submitted by the same group.) Please make sure the names of the all group members and the NetID are printed clearly on the first page of the report.