Weibin's Warriors
April 8th, 2018

Datafest 2018 -- Indeed Analysis Write Up

We decided to use clustering, an unsupervised learning technique. Our goal is to group jobs by common characteristics shared by these postings. Given the Indeed dataset, we randomly drew a sample of 500,000 observations and collectively decided to use 12 out of 29 variables: jobId, Word Count, Character Length, numReviews, Date, Salary, Country, normTitleCategory, clicksPerDay. We performed NA reassignment, outlier reassignment aggregation by jobID, variable selection, variable transformation, and variable standardization. Interestingly, the clicks variable was heavily skewed right and adjusted it by using fourth root.

We decided to apply the k-means clustering technique and discovered k=5 to be the optimal number of clusters. We took our standardized variables and clustered by industry and location separately. For Cluster 1, Character Length and Word Count seemed to be the most important. However, there is nothing significant. For Cluster 2, *Salary* seemed to be the most important. For Cluster 3, we had *Low Character Length and Word Count.* For Cluster 4, it was relatively *average.* For Cluster 5, *high Clicks Per Day* seemed to be the most important. We can use this information to see what kind of listings can be catered to certain groups of people. Surprisingly, the technology and medical industries had a fewer amount of words for the job description.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Most common industries	Sales, retail, service	Accounting, architecture, engineering, finance, meddr	Veterinary, sanitation, personal, science,	Admin, agriculture, hospitality, service	Accounting, childcare, service, tech software
Least common industries	Techsoftware, meddental, meddr	Food, personal, retail, warehouse	Finance, marketing, military, socialscience	Childcare, engchem, engid, techinfo	aviation, agriculture, engchem, military

Github Repository: https://github.com/akan72/Datafest-2018