General Assembly
Data Science Immersive - 5
Baurjan Safi

July 28, 2015

EXECUTIVE SUMMARY

Introduction

I have requested to provide analytical report on the job market in 15 major cities of the United States for positions of Data Scientist or related positions. The source of the data was web resource www.indeed.com. The task is to build a model for predictions and make it as reliable as possible.

Data Description

The search was based on web scraping of about nine thousand job announcement placed on indeed.com across 25 major cities in the US. The criteria for selection of the cities were their size, density of population. The other source of data was the list of publicly available statistical data on median household income for each city as well as its population, density and location expressed by latitude and longitude.

Of those nine thousand announcements, only 477 contain salary information. Some numbers were yearly, some were dayly, and some were hourly. All rates were extrapolated into annual salaries on the assumption that there are about 200 working days, 12 months, 1600 hours a year.

Models

Three models were applied for the data set:
Random Forest with limited number of variables
Random Forest with variables including words
Stratified K-folds
Bagging Classifier

Through sifting the variables that eventually would be sensibly applicable I identified that the inverse of the City Median Household Income is almost completely linearly describes the average salary per city and decided to use it for modeling. Introduction of this variable highly predicts the salary expectations.

Gridsearching through the number of categories I identified that the split into 0/1 provides the best results for prediction.

Unfortunately for the HR, the word importance did not produce any visible result.