Indeed Salary Project

Avinash TAMBY

Introduction

- Scrape Indeed for Data Scientist positions
  - Included related positions

- 30 most populous cities + tech cities + big-ish cities (Palo Alto, Atlanta, Boston)

- Determine if a job will pay above or below median Data Science salary

Scraping

- Built a scraper to pull ~100 job listings for each city I wanted to look at

- Took job title, company name, location (encoded city name), snippet of job description and salary (if applicable)

- Most listings did not come with salary information (about 200 of 2500 total listings)

What Industry Factors Influence Pay ?

- Location?
  - Data scientists in NY and SF are probably paid more than data scientists in OKC or ATL.
  -
- Title?
  - If the job has a ‘senior-level’ title
  - Includes the words ‘Senior’, ‘Director’, ‘Principal’, ‘Chief’, ‘Manager’
  -
- Company?
  - Big companies (e.g. Google) probably pay more than small startups
  -
- Job Description?
  - Do certain words in a job description correlate with higher pay?

What I looked at

- Location
- Seniority
- TF-IDF
  - measures importance of certains words in a document
  - Weight increases if it appears a lot in a document
  - Weight decreases if it appears a lot across many documents
  - This algorithm helps emphasize a certain word’s importance

Classifiers

- Random Forests
  - Use a random subset of features to build decision trees and makes a classification based on the agglomeration of the trees (a forest)

- Gradient Boosting Classifier
  - Builds decision trees and tries to iteratively explain the error to create a stronger overall model

- These are generally considered to be the industry favorites because they usually perform better than other classifiers

Models

- (1): RF with location
- (2): RF with location and seniority
- (3): RF with TF-IDF on job summaries
- (4): RF with location, senority and TF-IDF on job summaries
- (5): GBT with location, seniority, and TF-IDF on job summaries

Results

- Best performing were the last 3 models:
  - (3): RF with TF-IDF on summaries (76% accuracy)
  - (4): RF with location, seniority, and TF-IDF on summaries (77% accuracy)
  - (5); GBT with location, seniority, and TF-IDF on summaries (77% accuracy)

- Most important features for (4) were words: ‘looking’, ‘scientist’, ‘experience’, ‘team’ and feature ‘location_code’
- Most important features for (5) were words: ‘looking’, ‘role’, ‘science’, ‘knowledgable’ and feature ‘senior’

Interpretation

- ‘Looking’ seemed to be important, though I thought this word would only come up if a job description said ‘We are looking for someone who can …’. I think it’s odd that it comes up as an important feature in 2 different models

- Job summaries were only snippets, not the full job summaries
  - Lots of missing qualification information
  - Most important qualification requirements comes later in the description, and I did not scrape that information

Questions