Predicting Salaries for Data Scientist Jobs

Khatereh Mohajery

Jan. 06, 2017

Summary

The main goal of this project was to determine what industry factors influence the salary of data scientist jobs in the market.

- Includes:
    - The source of the data
    - Steps in processing the dataset
    - The methods and models used for predicting salaries
    - The results of the models and important factors in determining the salary for data scientist jobs

The Data

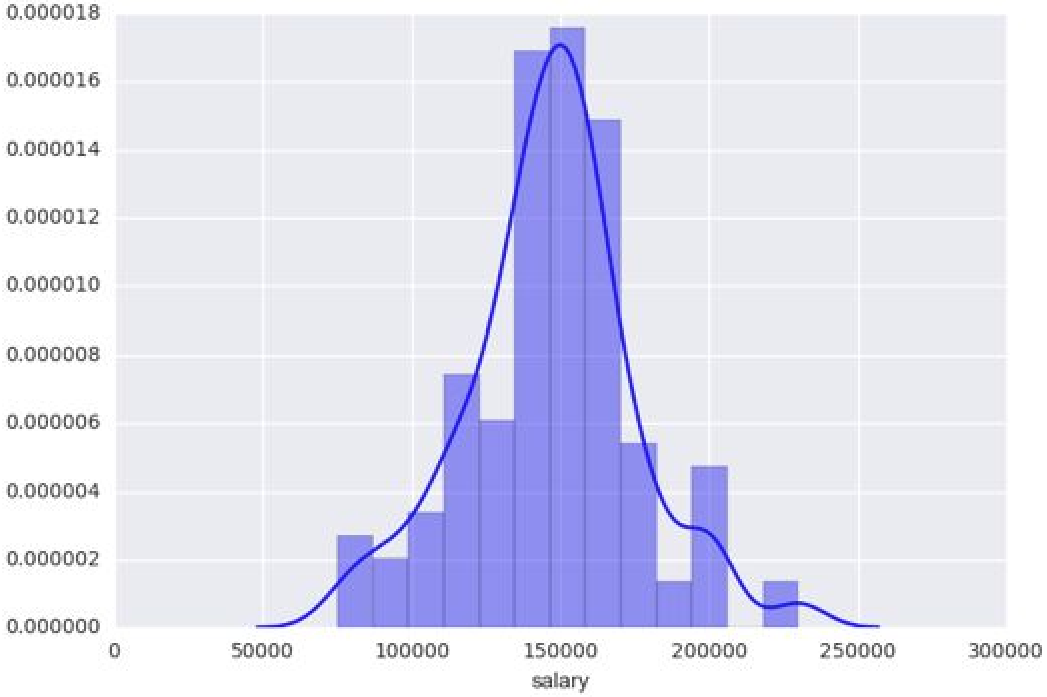
- The data was scraped from indeed.com
- The data was collected searching "data scientist" jobs for the following cities across US for estimated salary ranges of $75,000 - $90,000 - $105,000 - $120,000 - $135,000 and over $150,000

New York	Chicago	San Francisco	Seattle	Los Angeles	Austin
Philadelphia	Atlanta	Dallas	Pittsburgh	WashingtonDC	Portland
Phoenix	Denver	Houston	Miami	San Jose	Palo Alto

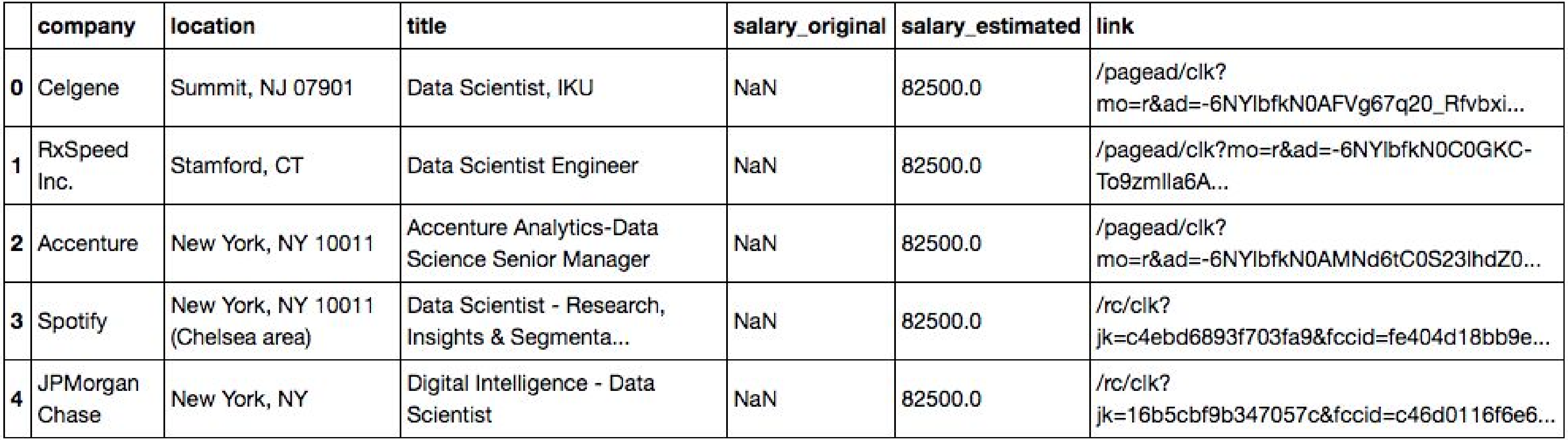
The Data (continued)

- For each job posting :
  - Title
  - Company
  - Location
  - Summary for job postings in indeed format pages
  - Original posted salary
  - Indeed estimated salary
- Dropping the duplicate entries (15113 ----> 10900)
- Finding the median for posted salaries = $140,000

Histogram of Data Scientist Salaries

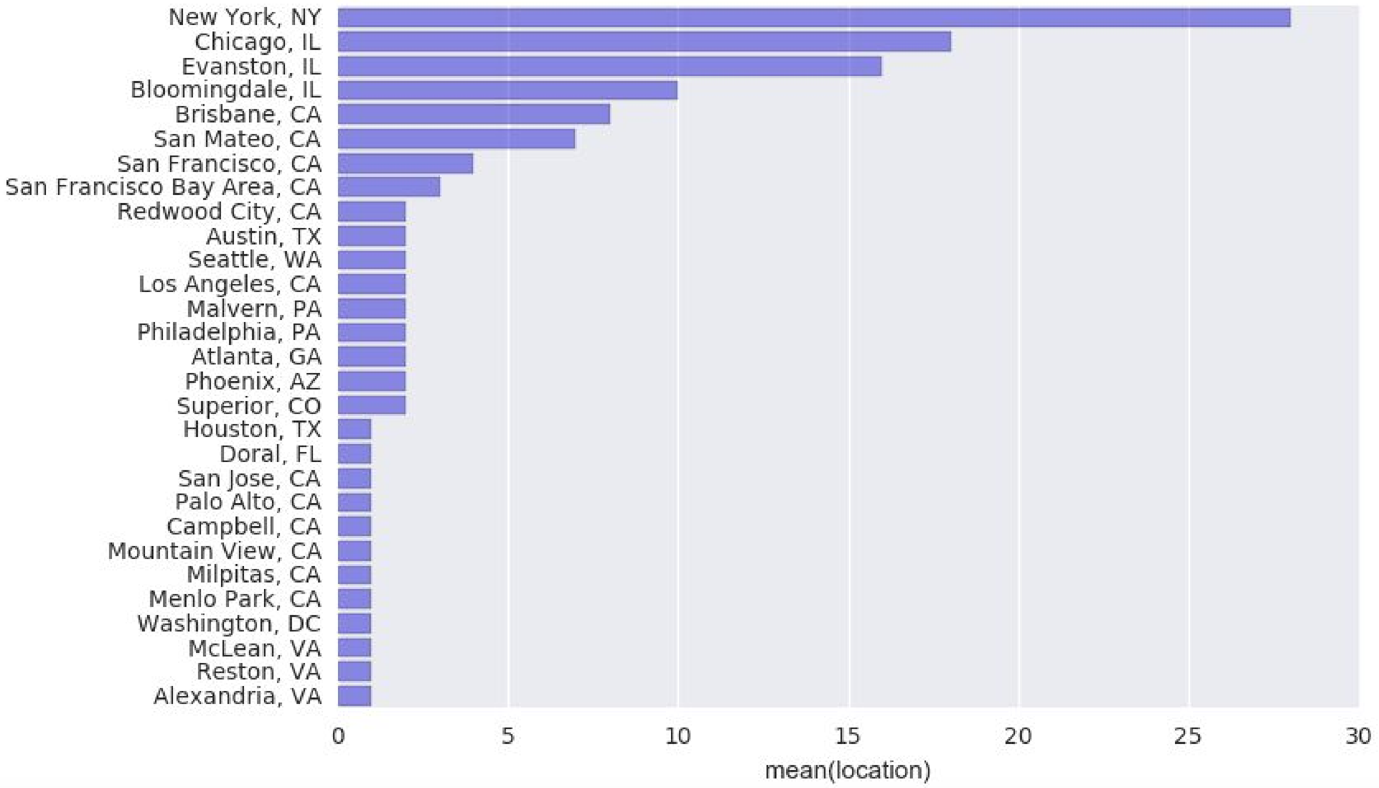


The Data Set

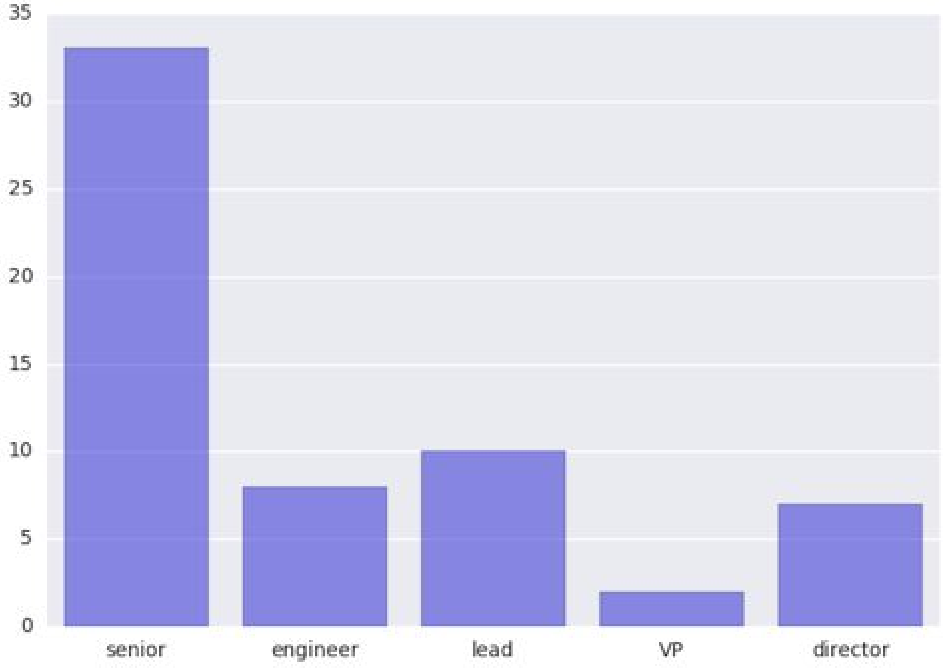


	company	location	title	salary_original	salary_estimated	link
0	Celgene	Summit, NJ 07901	Data Scientist, IKU	NaN	82500.0	/pagead/clk?mo=r&ad=-6NYIbfkN0AFVg67q20_Rfvbxi...
1	RxSpeed Inc.	Stamford, CT	Data Scientist Engineer	NaN	82500.0	/pagead/clk?mo=r&ad=-6NYIbfkN0C0GKC-To9zmlla6A...
2	Accenture	New York, NY 10011	Accenture Analytics-Data Science Senior Manager	NaN	82500.0	/pagead/clk?mo=r&ad=-6NYIbfkN0AMNd6tC0S23lhdZ0...
3	Spotify	New York, NY 10011 (Chelsea area)	Data Scientist - Research, Insights & Segmenta...	NaN	82500.0	/rc/clk?jk=c4ebd6893f703fa9&fccid=fe404d18bb9e...
4	JPMorgan Chase	New York, NY	Digital Intelligence - Data Scientist	NaN	82500.0	/rc/clk?jk=16b5cbf9b347057c&fccid=c46d0116f6e6...

Frequency of each city appearing in the data set



Frequency of each keyword appearing in the data set



Defining the classification problem:

Approached the problem as a classification problem trying to find out if a job posting has above median salary or below median salary

- Binary value for salaries : 1 and 0
- Predictors : locations and keywords “senior” ,”lead”, “engineer”, “director” and “VP”) in the job title
- Random forest, knn and Logistic Regression models
- Optimizing parameters of the models using grid search

How good are the model predictions?

- Baseline accuracy of 0.5

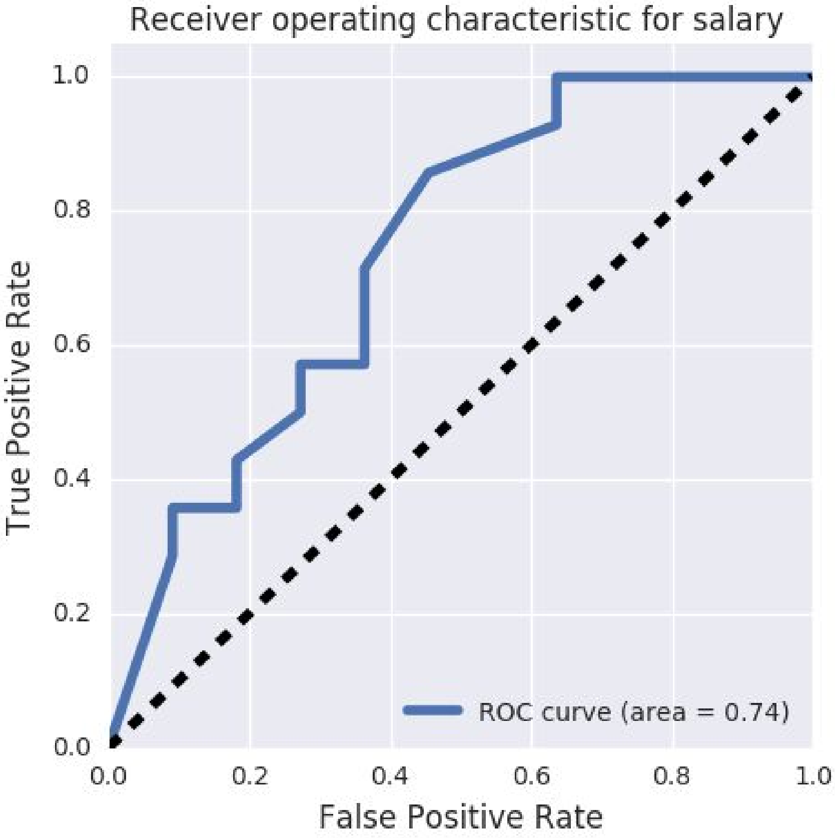
Model	Accuracy Score	Precision	Recall	f1-score
Random Forest (max_features =0.5)	0.64	0.66	0.64	0.64
Knn (n=12)	0.68	0.75	0.68	0.67
Logistic Regression	0.62	0.62	0.62	0.62

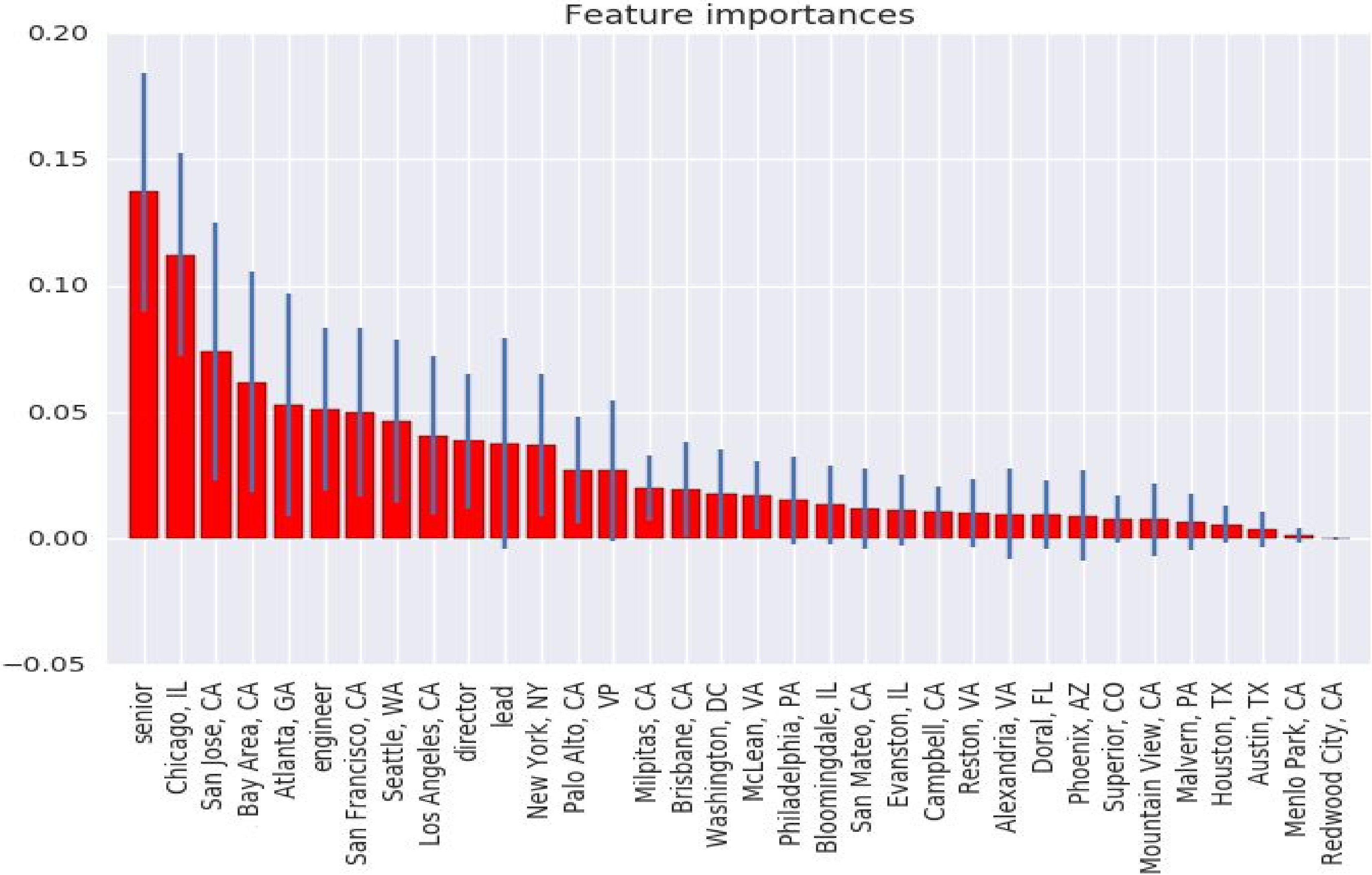
Improving the Precision of the Random Forest Model

- Why to improve precision: to be sure to tell someone that a job pays above median
- Increasing threshold probability from 0.5 to 0.75
- Overall accuracy drops to 0.60 from 0.64

Predicted by RF	Precision
Below median	0.53
Above median	0.83
Overall	0.70

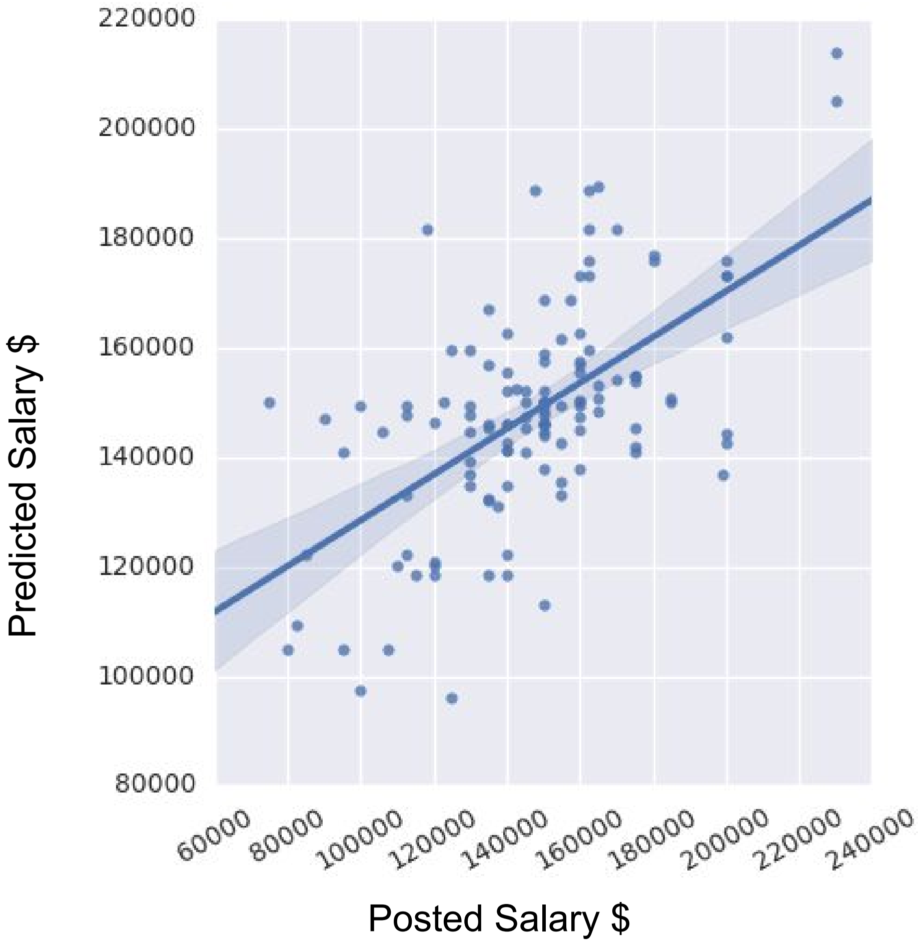
ROC-AUC for Random Forest model




Feature importances
0.20
0.15
0.10
0.05
0.00
−0.05
senior
Chicago, IL
San Jose, CA
Bay Area, CA
Atlanta, GA
engineer
San Francisco, CA
Seattle, WA
Los Angeles, CA
director
lead
New York, NY
Palo Alto, CA
VP
Milpitas, CA
Brisbane, CA
Washington, DC
McLean, VA
Philadelphia, PA
Bloomingdale, IL
San Mateo, CA
Evanston, IL
Campbell, CA
Reston, VA
Alexandria, VA
Doral, FL
Phoenix, AZ
Superior, CO
Mountain View, CA
Malvern, PA
Houston, TX
Austin, TX
Menlo Park, CA
Redwood City, CA

Regression Problem?

- Using linear Regression to estimate salary
- Using locations and keyword in the job title
- Cross_validation Score of 0.20



Using Text in Summary as

- Estimated salaries from indeed.com
- Tf_idf method to produce features from summary (2645 words)
- Random Forest as the model
- Probability threshold from 0.5 to 0.75 (RF1 and RF2)
- Most important words: expertise, winner, financial!!

	Accuracy Score	Precision	Recall	f1-score	AUC score
RF1	0.59	0.54	0.59	0.55	0.59
RF2	0.62	0.68	0.62	0.63	0.59

Most important words in the job summary for determining job salaries:

	importance
expertise	0.051971
winner	0.050840
financial	0.038376
good	0.034072
houston	0.033172
production	0.026776

Steps to Improve the Model

- Scraping more data using other job titles such as data analyst

- Using other sources to get salaries such as glassdoor