Hiring marketplace for Data Scientists

LinkedIn

Peter Gao

GA - SF DAT 22
June 2, 2016

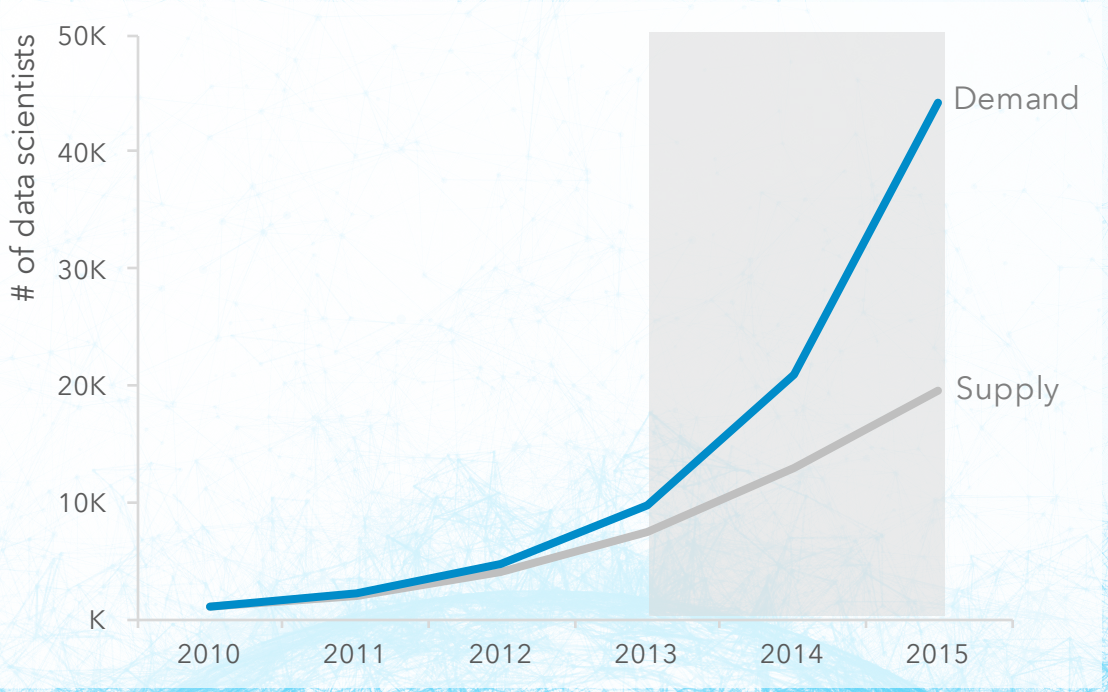

When I grow up, I want to be a... Data Scientist
Ready or not, skills gap, here I come!

McKinsey projects that by 2018, demand in U.S. for data scientists may be as much as

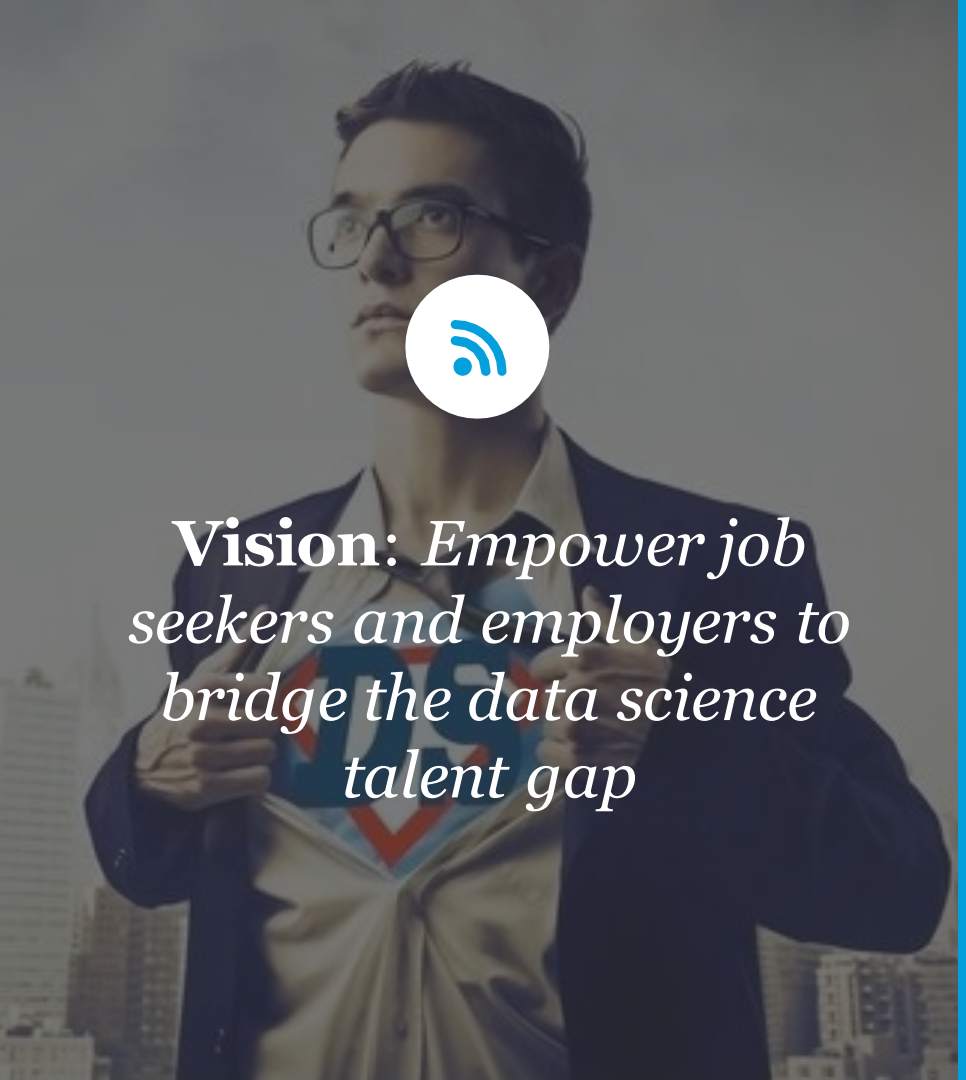
60%

greater than the supply.

LinkedIn data suggests that the talent gap in U.S. already **exceeds 60%** and has **doubled past 2 yrs**



Mission: *Predict hiring outcomes in data science talent marketplace*



Vision: *Empower job seekers and employers to bridge the data science talent gap*

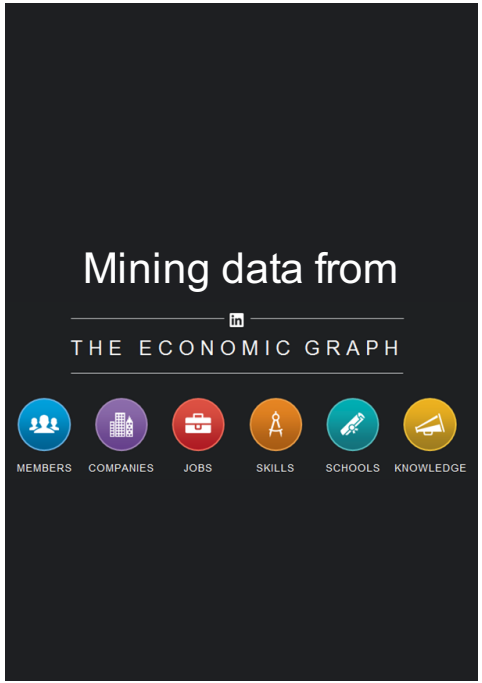
Mission: *Predict hiring outcomes in data science talent marketplace*

Methodology Overview

For predicting hiring outcomes in data science talent marketplace

Data Exploration

Prepare: Obtain the data



What's in training dataset: Record for each member in U.S. who viewed 10+ data scientist jobs and changed jobs within last 2 years

Data Shape: (27519, 82)

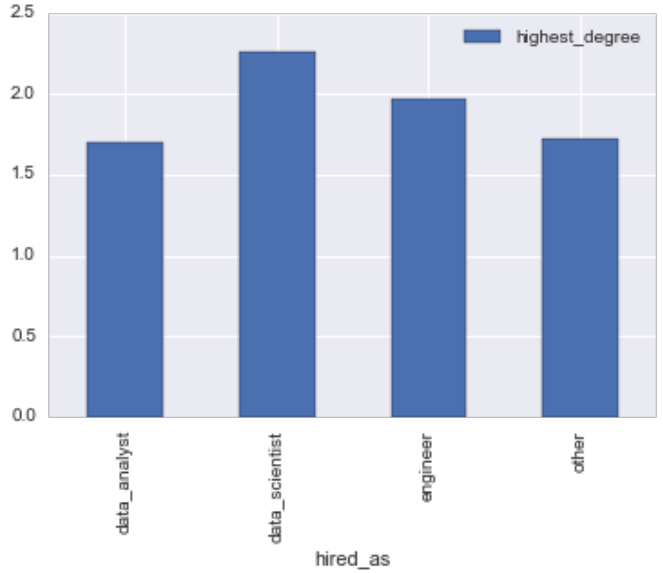
- **Response Variable (multinomial classification):** hired_as (data scientist, data analyst, engineer, other)
- **Example columns**
  - **Metrics:** connections, skills, endorsements, etc.
  - **Standard Dimensions:** gender, region, school, etc.
  - **Custom Variables**: # of bootcamps, ds_connections, etc.

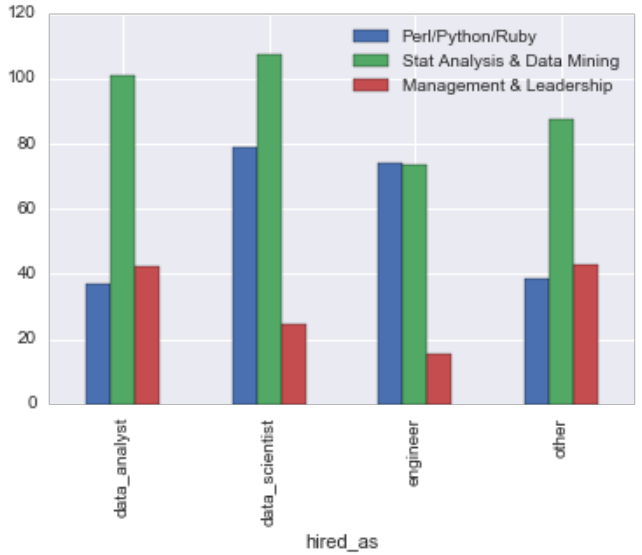
Prepare: Clean and pre-process the dataset

50% of time was spent cleaning and pre-processing

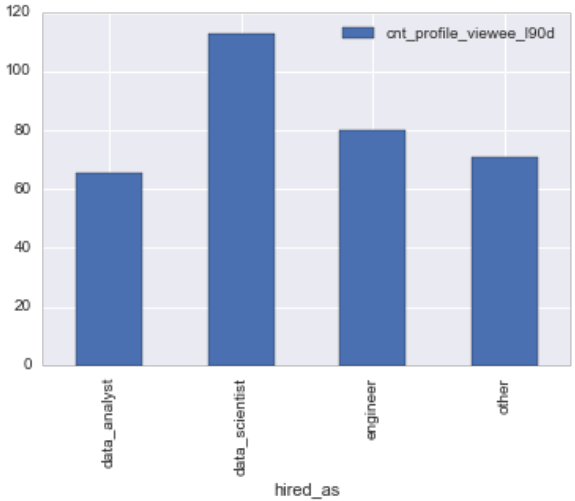
1. Identify and fill in **missing values**
2. Convert categorical data to **numeric** or **boolean** (e.g. highest_degree)
3. **Impute** some missing values (e.g. gender, university acceptance rate)

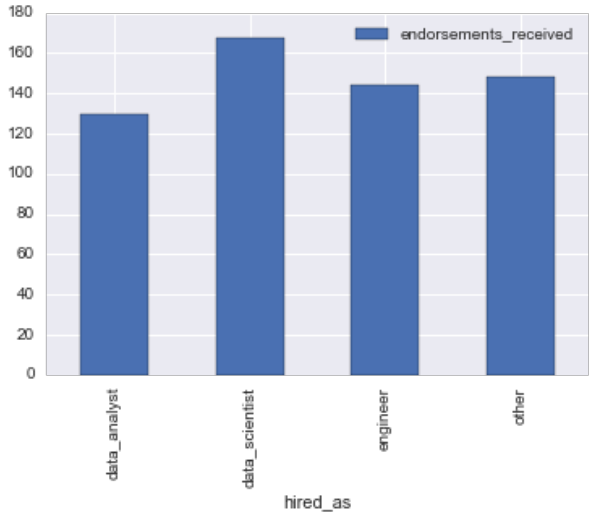
Explore: Uncover qualification patterns

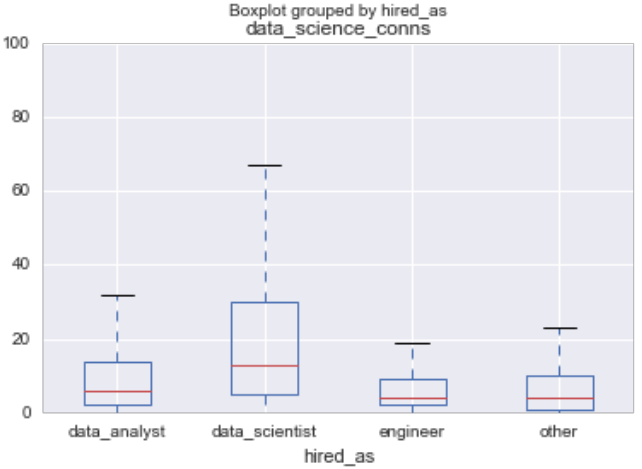




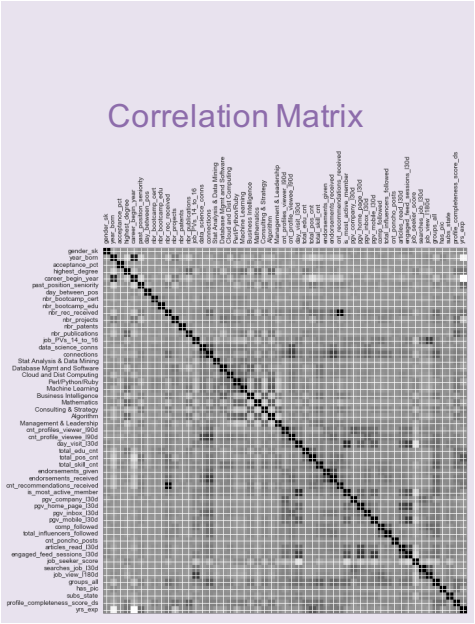
Explore: Uncover behavioral patterns







Explore: Choose model features



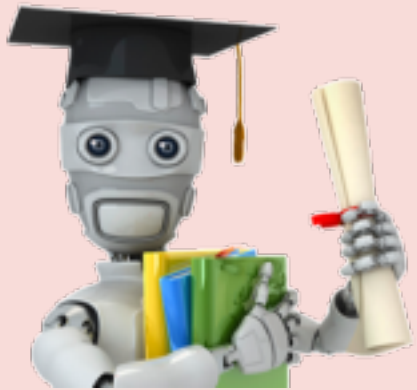
1. **Multicollinearity:** Remove features such as year_born which highly correlates with career_begin_year

2. **Add dummy variables** (industry, region, field of study)

3. **Choose predictor variables**
   - **Metrics:** connections, skills, endorsements, # recs, PVs, engagement, job searches, profile completion, etc.
   - **Standard Dimensions:** gender, region, school, degree, field of study, age, industry, company, profile summary
   - **Custom Variables**: # of bootcamps, # online courses, data science connections, university acceptance rate

Modeling

Model: Select models



Response Variable

Observed from dataset

y_i = {1 = data scientist; 2 = data analyst; 3 = engineer; 0 = other}

Selected Classification Models

K-Nearest Neighbors

Logistic Regression

Random Forest classifier

Model: Apply and Validate models

Null Accuracy Rate (to beat) - 0.36

Model	Metric
Logistic Regression	Accuracy – 0.35
K-Nearest Neighbors	Accuracy – 0.41
Random Forest Classifier (determined best parameters)	Accuracy – **0.54** OOB Score – 0.52

Random Forest Wins!

Most Important Features

rank	feature	importance
1	Stat Analysis & Data Mining	3.259956e-02
2	data_science_conns	3.145870e-02
3	cnt_profile_viewee_l90d	2.882765e-02
4	job_PVs_14_to_16	2.860416e-02
5	Perl/Python/Ruby	2.774114e-02
6	connections	2.519730e-02
7	job_view_l180d	2.511343e-02
8	day_between_pos	2.461743e-02
9	endorsements_received	2.335754e-02
10	cnt_profiles_viewer_l90d	2.257608e-02
11	comp_followed	2.246757e-02
12	endorsements_given	2.245722e-02
13	Database Mgmt and Software	2.160170e-02
14	total_skill_cnt	2.143237e-02
15	Machine Learning	2.112286e-02

Most important

- Endorsements for key skills (statistics & scripting)
- Connections within data community

Lower importance than expected

- Degree, field of study

Low importance

- # Patents
- # Bootcamps attended (sorry GA)
- Recommendations
- Industry


What do you think are the avg. class probabilities for sfdat22 students?

predict_proba(GA_class)

Recommendations

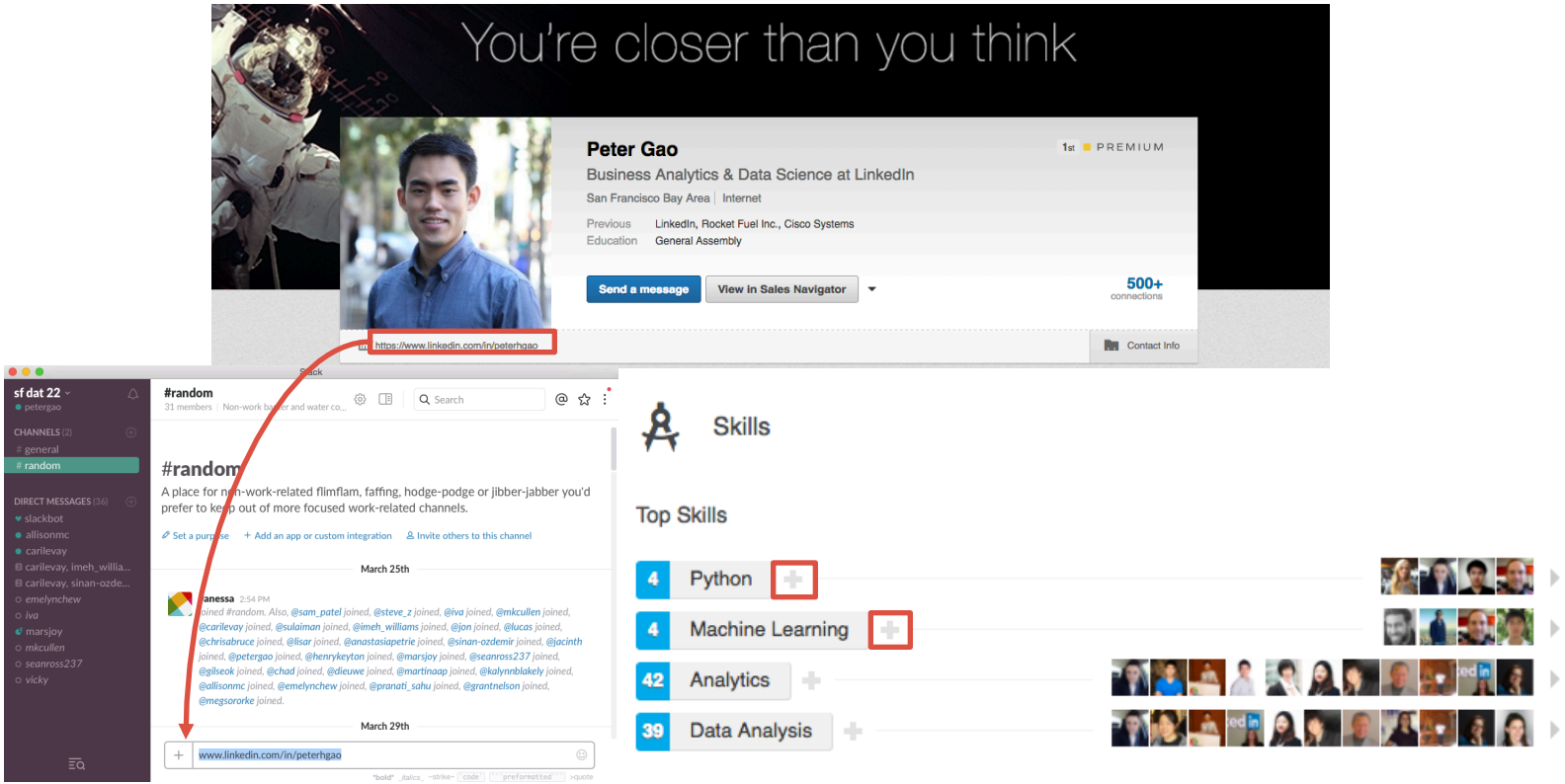
Key Learnings about getting hired

Connect	Connect with others in Data Science and Analytics community
Attain Skills	Develop skills in Python, Machine Learning, and Statistics
Get noticed	Leverage LinkedIn to build your brand and get noticed

Possible Extensions of this Project

Web Development	Integrate with LinkedIn sign in API so users get **personalized probability** of getting hired
Skill Recommendation Engine	**Recommend skills** to learn based on member's existing skillset
New Economic Research	Identify **economic opportunities** for hiring marketplace to bridge the data science talent gap

Let’s connect with and endorse each other



Thank You!

Questions?

| Peter Gao | Endorse my skills @ linkedin.com/in/peterhgao