

Analysis of H1B Visa Petitions (2011-2016)



**Siyuan Liu
Mar. 12, 2018**



Overview

01

Introduction

02

Current Job Market for Foreigners

03

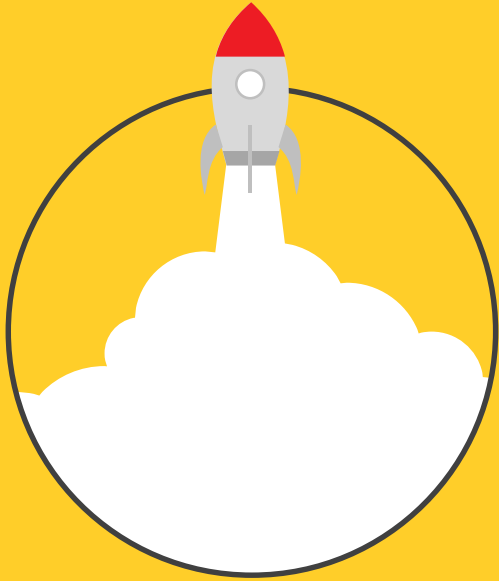
Predictive Analysis of Petition Status

04

Further Analysis

What is the process of H1B visa application?





Current Job Market for Foreigners

Data

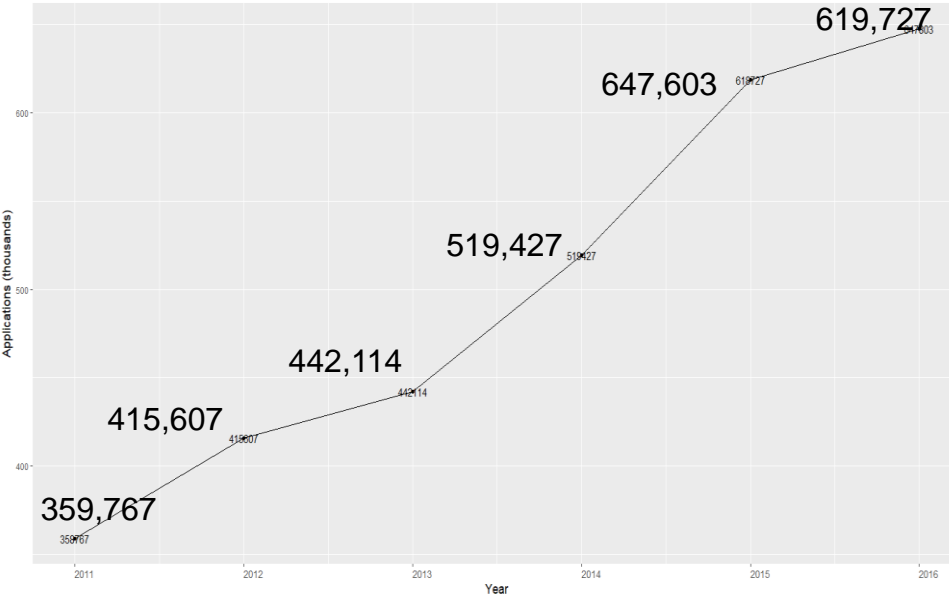
- **Data Source:** <https://www.kaggle.com/nsharan/h-1b-visa/data>
- **10 Variables:**

Case Status	Employer Name
SOC Name	Job Title
Full Time Position	Year
Worksite	Prevailing Wage
Longitude	Latitude
- **Data Size:** 3,002,458 records

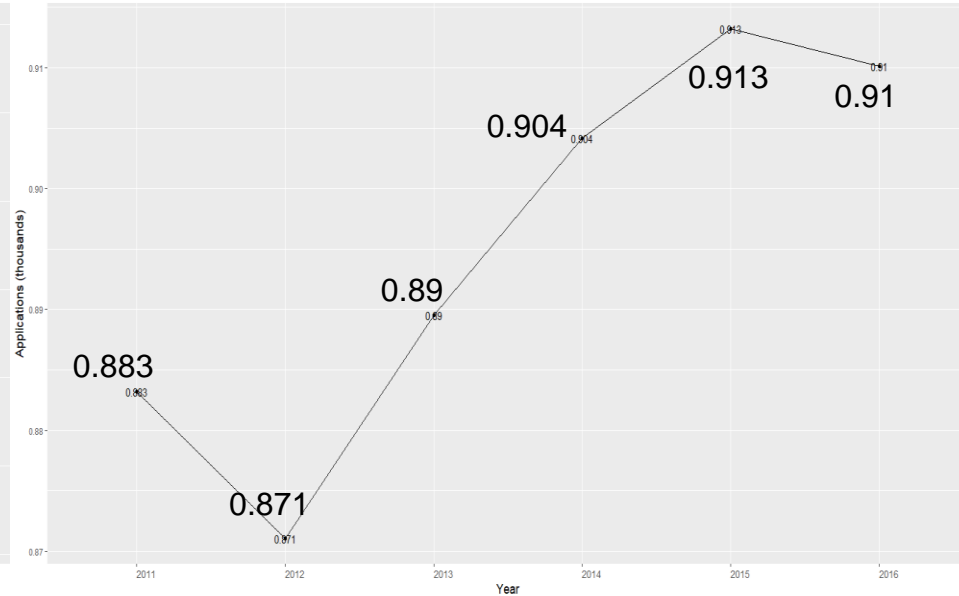


General Trend

Amount of petitions trend (2011-2016)

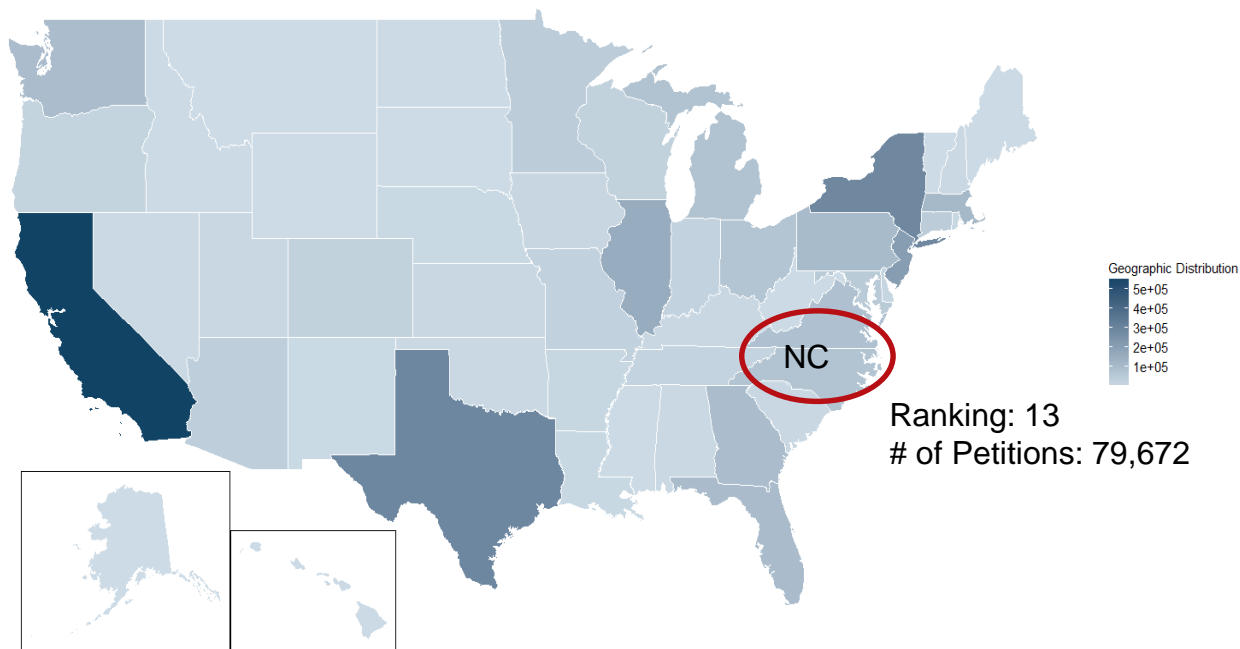


Certified Rate (2011-2016)



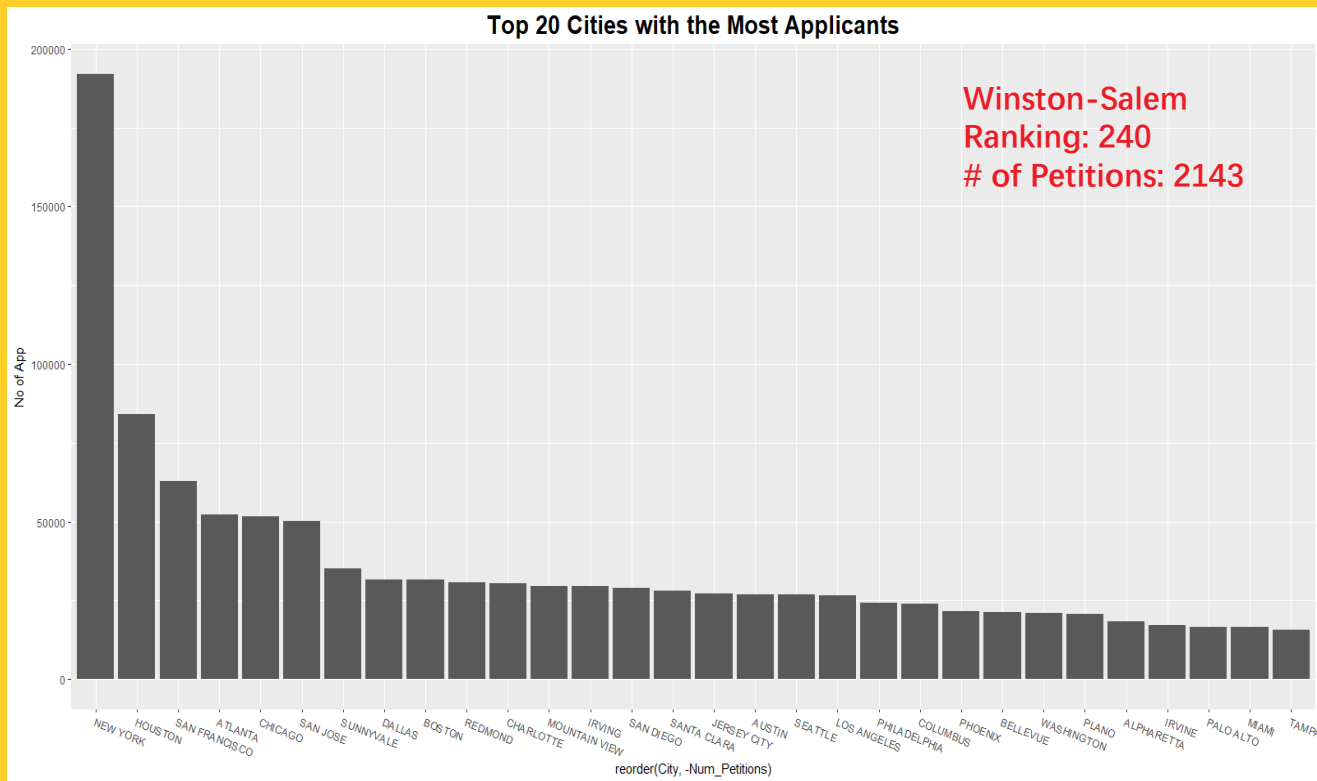
States with the Most Applications

Geographic Features for # of Petitions(2011-2016)



State	Count
CA	559,942
TX	294,454
NY	291,411
NJ	214,076
IL	160,814
MA	116,464
PA	109,960
FL	105,722
GA	103,168
WA	102,777
VA	90,012

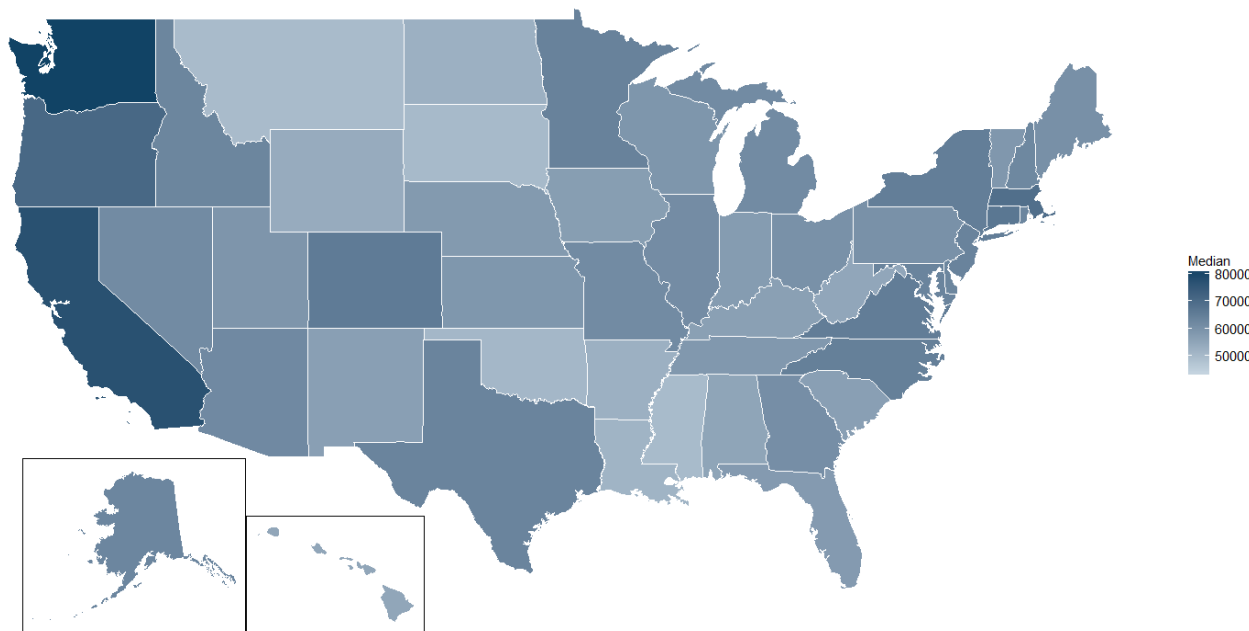
Cities with the Most Applications



City	Count
New York	191,980
Houston	84,070
San Francisco	62,854
Atlanta	52,306
Chicago	51,575
San Jose	50,145
Sunnyvale	35,203
Dallas	31,733
Boston	31,545
Redmond	30,724
Charlotte	30,387

States with the Highest Salary*

Geographic Feature of Salary(2011-2016)

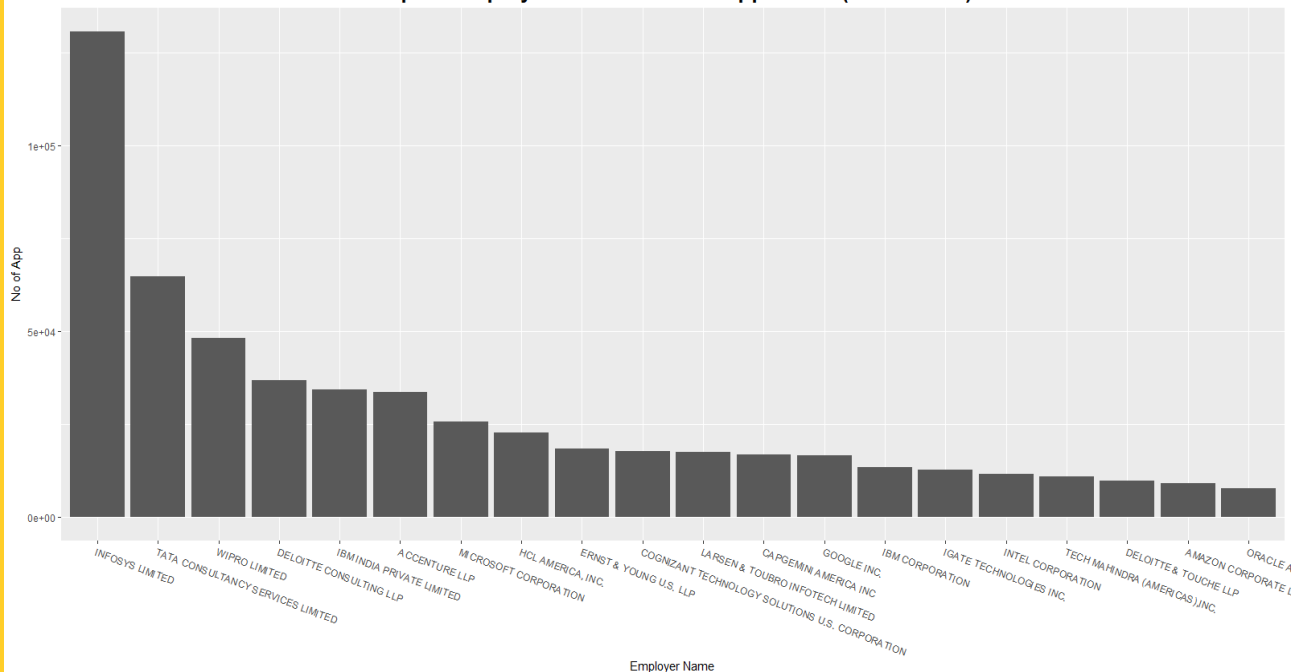


State	Wage*
Washington	81,432
California	77,522
Oregon	71,242
Massachusetts	69,243
Connecticut	67,059
DC	66810
Colorado	65,957
Virginia	65,519
New York	65,187
North Carolina	64,792

*Wage: Median Prevailing Wage

Popular Companies

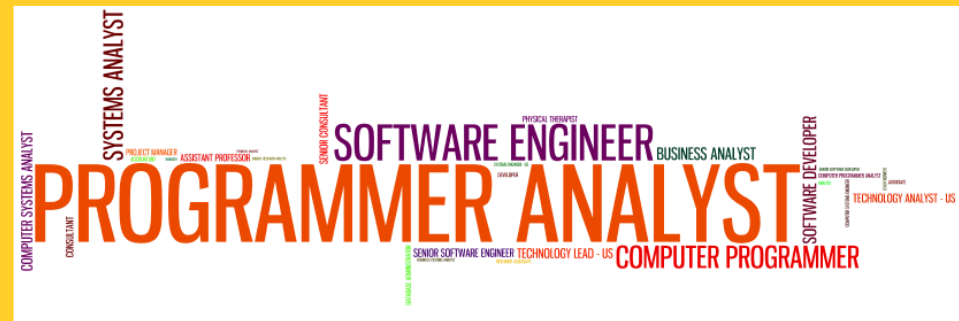
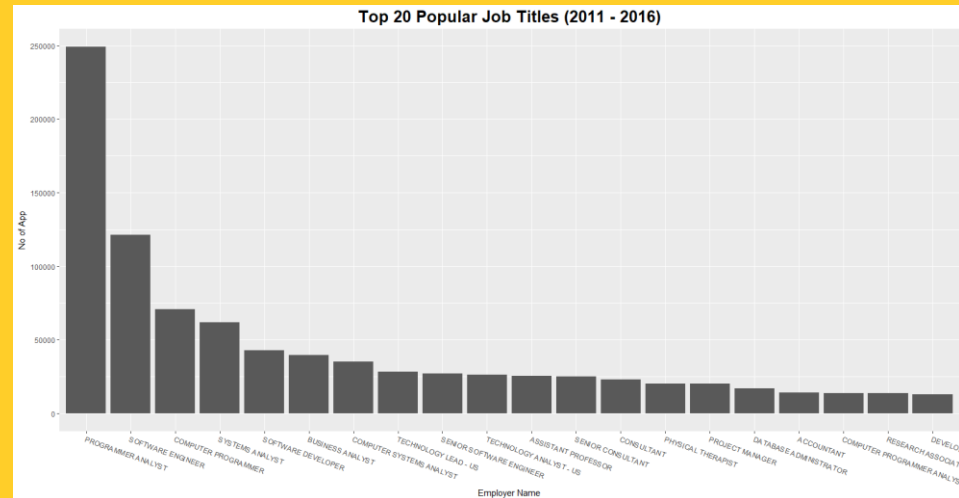
Top 20 Employers with the most Applicants (2011 - 2016)



Company Name	Petitions
Infosys Limited	130,592
Tata Consultancy	64,729
Wipro Limited	48,117
Deloitte Consulting	36,742
IBM India	34,219
Accenture LLP	33,447
Microsoft Corp.	25,576
HCL America. INC	22,678
Ernst & Young U.S.	18,232
Cognizant Tech. U.S.	17,528

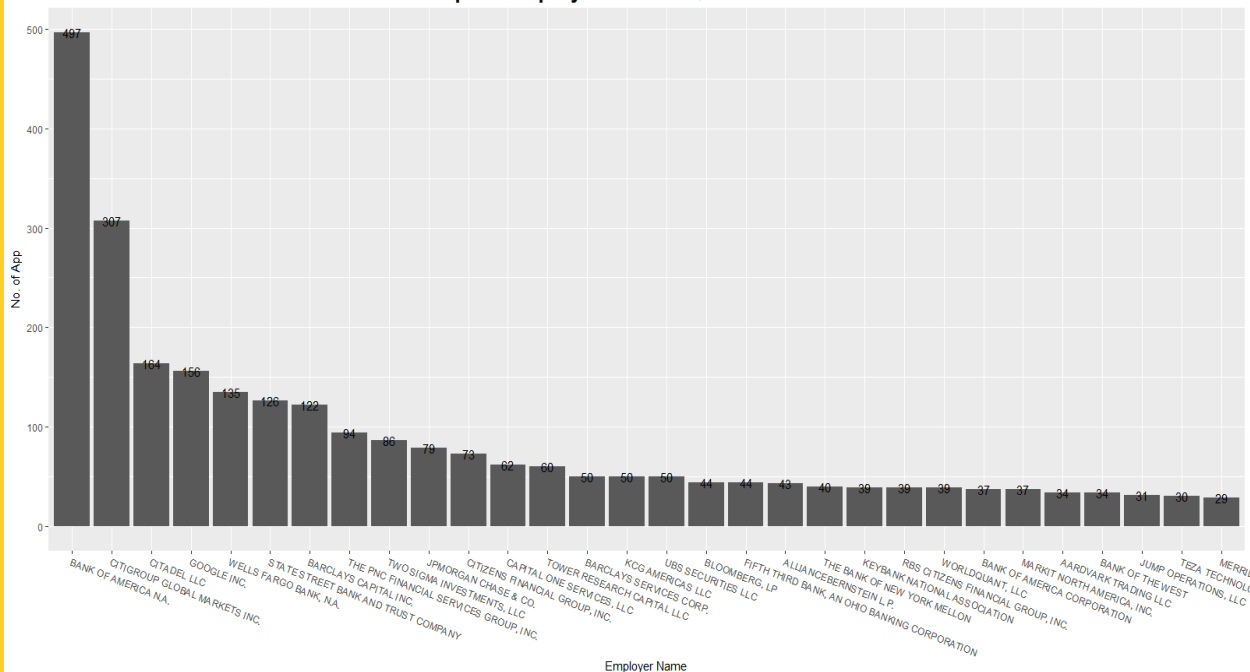
Popular Job Titles

Ranking	Job Titles
1	Programmer Analyst
2	Software Engineer
...	
6	Business Analyst
...	
12	Senior Consultant
...	
14	Physical Therapist
...	
17	Accountant



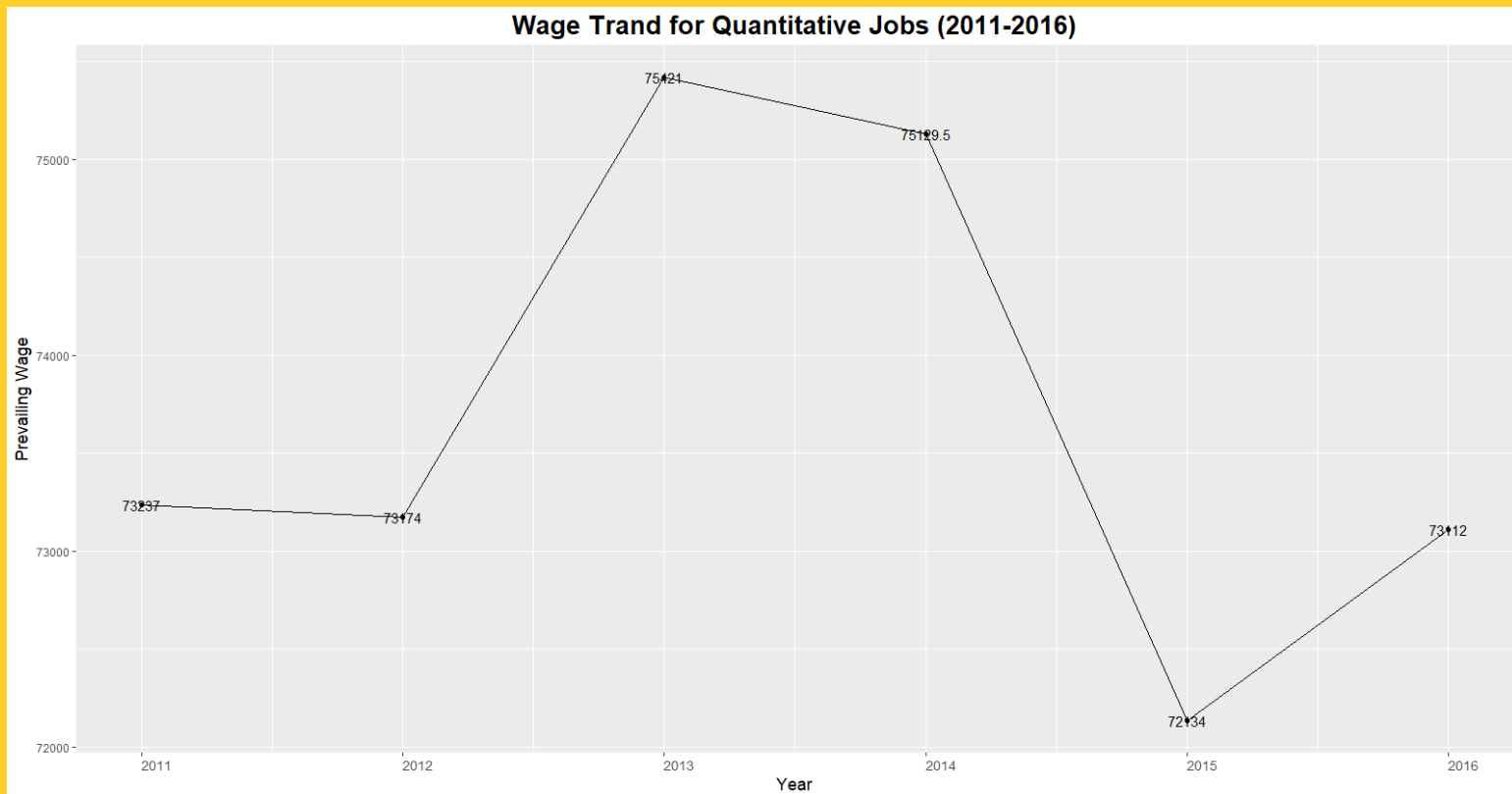
Employers for Quantitative Jobs

Top 30 Employers about Quantitative Jobs



Employer Name	Count
Bank of America	497
Citi Group Global.	307
Citadel LLC	164
Google Inc.	156
Wells Fargo Bank	135
States Street	126
Barclays Capital Inc.	122
The PNC Financial	94
...	
BB&T Corporation	14 (67/1415)

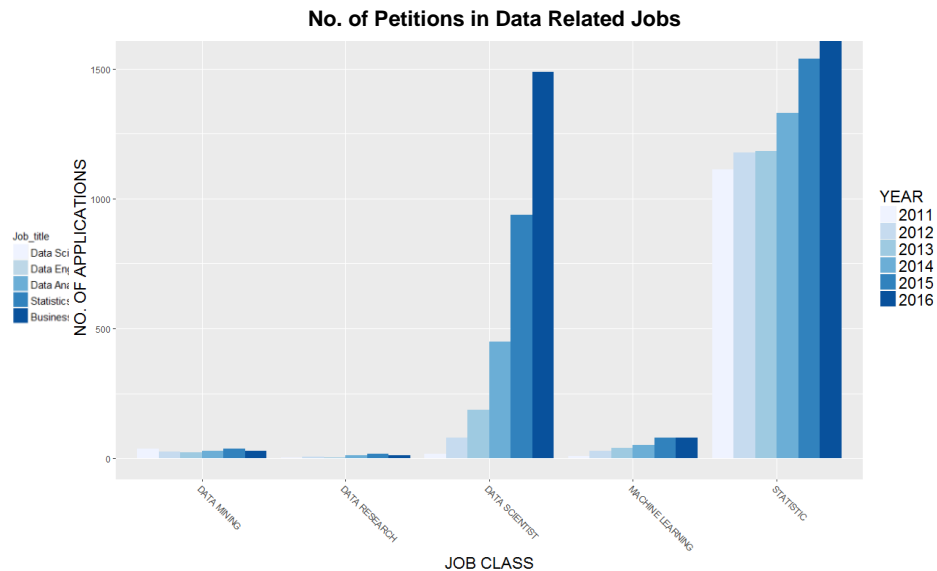
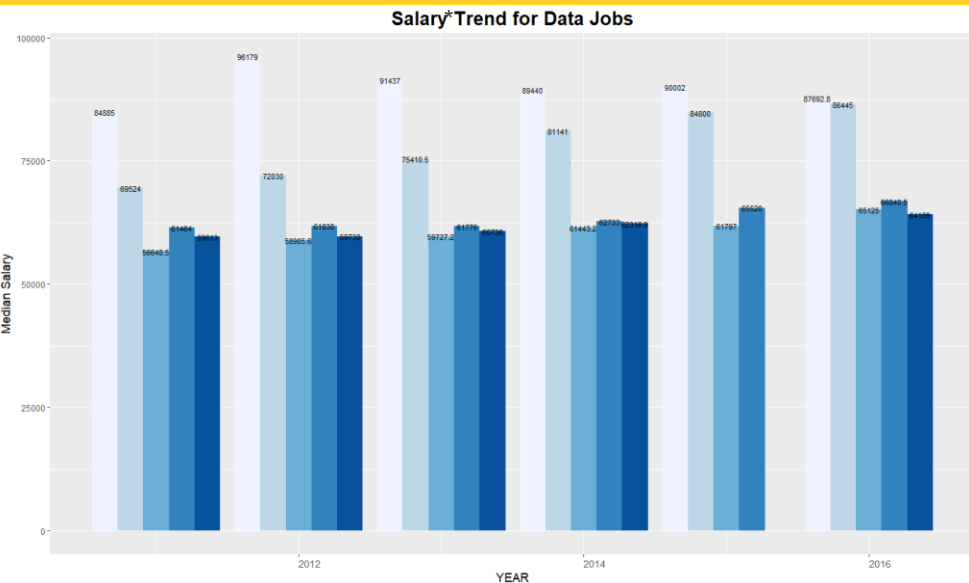
How much does a QA earn*?



*Median Prevailing Wage



Some Insights for Data Related Jobs



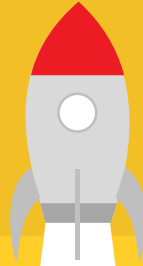
*Median Prevailing Wage

Conclusions

Some tips for
foreigner job seekers



- The chance of winning H1B lottery is **decreasing**.
- **Relocate** for better opportunities. East coast and west coast are both ideal worksite. Especially **CALIFORNIA, TEXAS** and **NEW YORK**.
- The demand for data scientist is **exploding**.
- Prepare yourself with more **programming skills**.
- **Quantitative Analyst** is a good option for seekers with preference to **financial services**.



Predictive Analysis of Case Status (demo)

Data Preparation

*

01

Missing Values

- Latitude and Longitude: deleted
- SOC_NAME: imputation

02

Oversampling

- Balancing the data
(10% of cases are denied)

03

Feature Engineering

- Encoding
- Dimension Reduction



Data Preparation

Oversampling

- Original: Training dataset – 2500 records (**9.8%** negative cases)
Testing dataset – 500 records
- After-Oversampling: Training dataset - 4838 records (**52.3%** negative cases)

Feature Engineering

SOC_NAME

- Number of Petitions
- Certified Rate

WORKSITE

- State
- City

EMPLOYER_NAME

- Number of Petitions
- Certified Rate

CASE_STATUS

- Certified
- Denied
- Certified - Withdrawn



Methodologies

Model Evaluation Methods

Logistic Regression

Random Forest

Adaboost

Cross-Validation

ROC and AUC

Modeling Methods



Model Selection



Logistic Regression:

Mean_AUC = 0.794

Mean_Best_Cutoff = 0.510



Random Forest

Max_AUC = 0.880

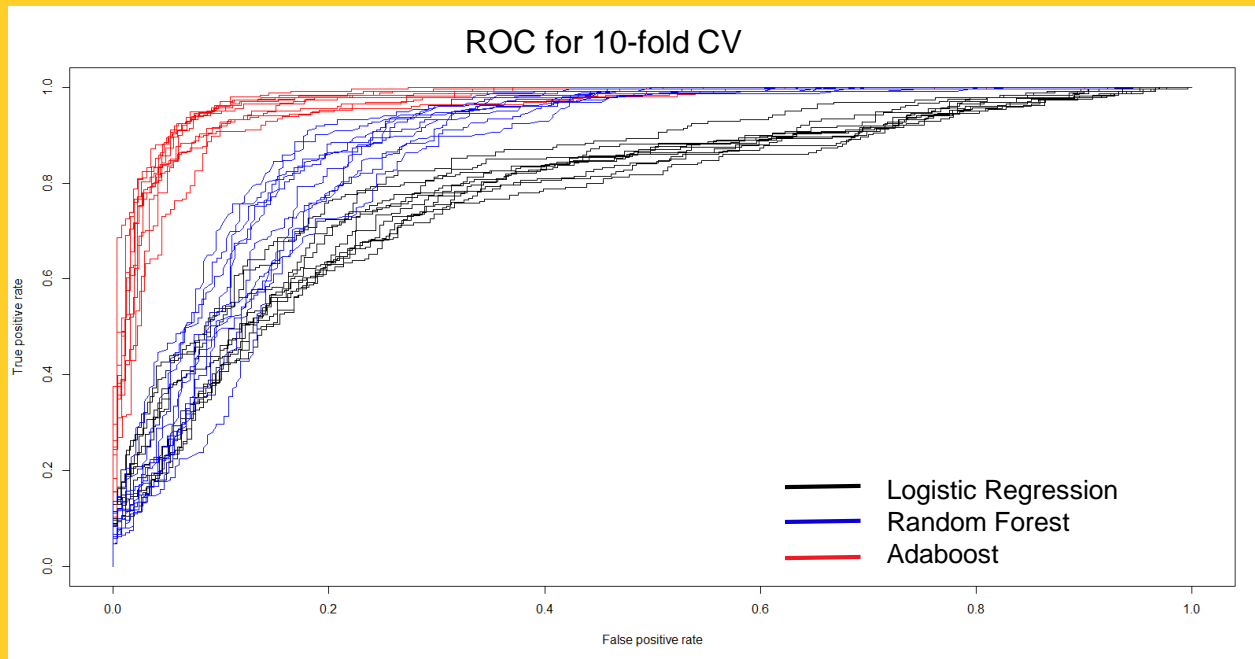
Mean_Best_Cutoff = 0.210



Adaboost

Mean_AUC = 0.968

Mean_Best_Cutoff: **0.499**



Prediction Results Analysis

Adaboost

	Denied	Certified
Pre_Denied	31	53
Pre_Certified	17	399

Accuracy Rate: 0.860

64.6% negative cases are detected

88.3% positive cases are detected

Random Forest

	Denied	Certified
Pre_Denied	3	18
Pre_Certified	45	434

Accuracy Rate: 0.874

6.25% negative cases are detected

96.0% positive cases are detected

Important Features:

- Employer Certified Rate
- Full_Time_Position
- Year
- Prevailing Wage

Logistic Regression

	Denied	Certified
Pre_Denied	17	115
Pre_Certified	31	337

Accuracy Rate: 0.708

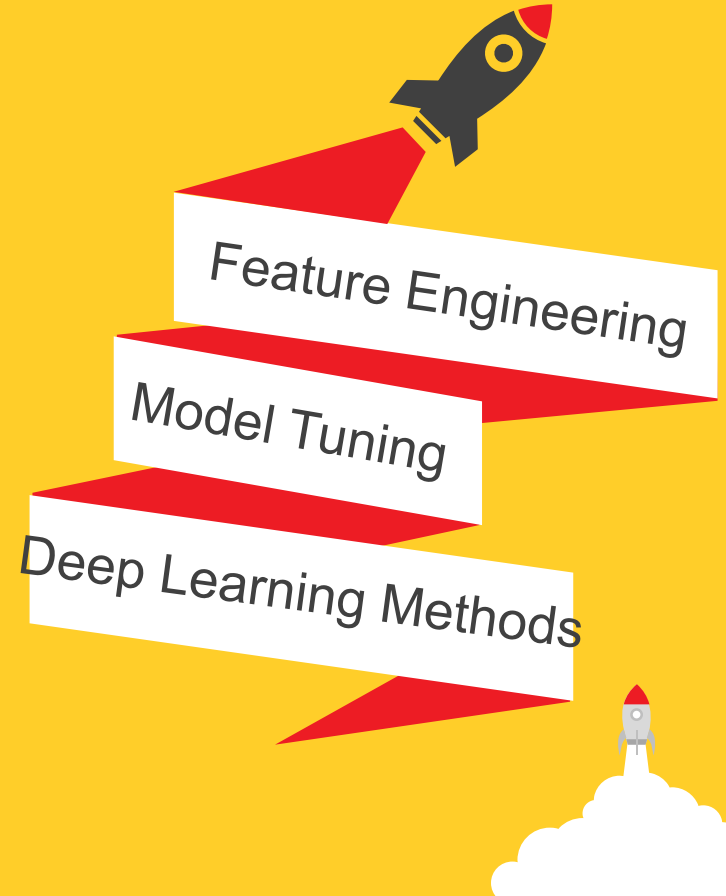
35.4% negative cases are detected

74.6% positive cases are detected



Further Analysis

- **Extract more information from text by word embedding**
(City, Job title, SOC name, Employer Name)
- **Use grid search method to optimize existing models**
- **Implement more advanced classifier to the whole dataset**





Thank you for listening!