# H1B Data Exploration

Barry Ke, Angela Qiao

Oct 29, 2018

# Introduction

- Motivation
- Exploratory data analysis
- Data analysis: wage
  - LASSO regression: optimization problem
  - Coordinate descent
  - Result
- Data analysis: status prediction
  - Logistic regression with L1 panelty: optimization problem
  - Quadratic approximation
  - Alternative: Newton's method
  - Result

# References


- Jerome Friedman, Trevor Hastie and Robert Tibshirani. *Regularization Paths for Generalized Linear Models via Coordinate Descent*. Journal of Statistical Software Vol 33. P1-22. 2010
- Andreas C. Müller and Sarah Guido. *Introduction to Machine Learning with Python*. O'Reilly Media. Inc. 2016.
- Trevor Hastie, Robert Tibshirani, Jerome Friedman *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag New York. 2009. 9781449369897
- Bureau of Labor Statistics, Department of Labor *OFLC Performance Data*. 2018.
- Anna Maria Mayda, Francesc Ortega, Giovanni Peri, Kevin Shih and Chad Sparber. *The Effect of the H-1b Quota on Employment and Selection of Foreign-Born Labor*. NBER Working Paper No. w23902. 2017.
- Statistical package: scikit-learn

# Motivation

- 19% of students at Columbia are international students
- 88.9% of Columbia College and Columbia Engineering - Undergrad degree earners are employed or in grad school
- $70,000 median starting salary for working graduates of Columbia College and Columbia engineering

# Data Set used

- https://www.foreignlaborcert.doleta.gov/performancedata.cfm#dis
- Labor Condition Application ("LCA") disclosure data from UNITED STATES DEPARTMENT OF LABOR

# H1B Process Introduction

What is H1B?

- H-1B is a temporary (nonimmigrant) visa category that allows employers to petition for highly educated foreign professionals to work in specialty occupations that require at least a bachelor's degree or the equivalent.
- Jobs in fields such as mathematics, engineering, and technology often qualify
- Duration: Three years

# H1B Process Introduction

Employer Qualification

- Employers must attest, on a labor condition application (LCA) certified by the Department of Labor (DOL), that employment of the H-1B worker will not adversely affect the wages and working conditions of similarly employed U.S. workers.

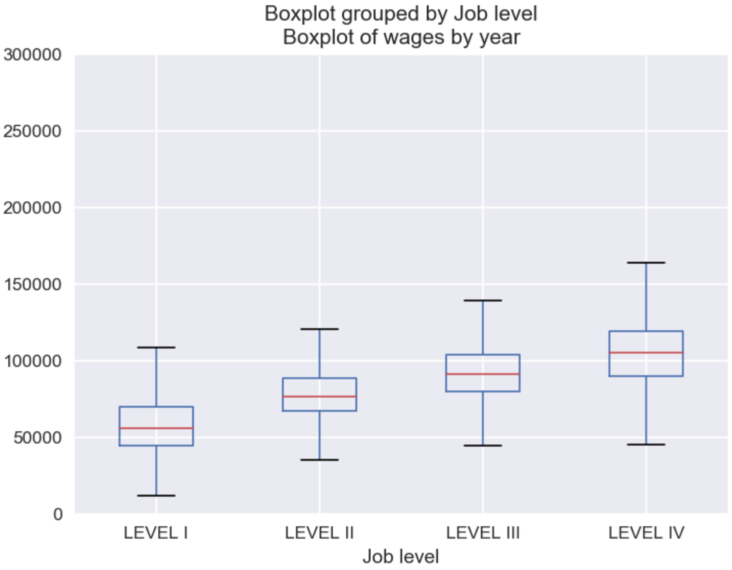
# Data cleaning

| Data before 2015 | Additional data after 2015 |
|---|---|
| Date | Total number of employee |
| Employer Name | Firm's founding year |
| Location | Education level |
| Economic sector | University |
| Job title | Major |
| Wage | Prior working experience (months) |
| Citizenship | |

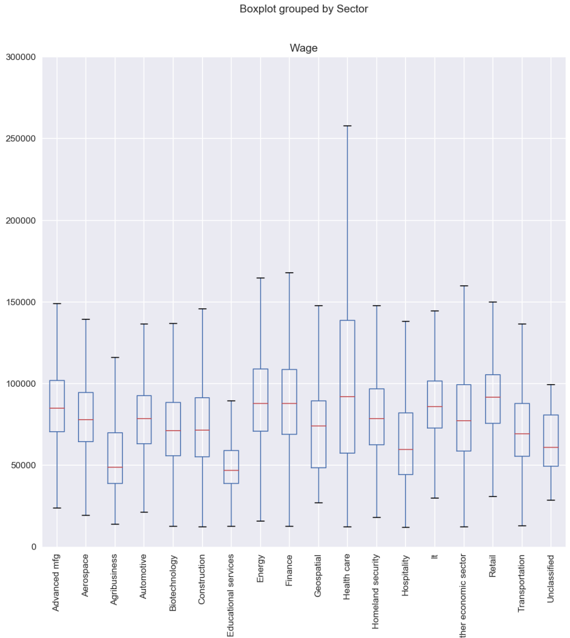
Table 1: Variables included

ltab:my$_l$*abel*

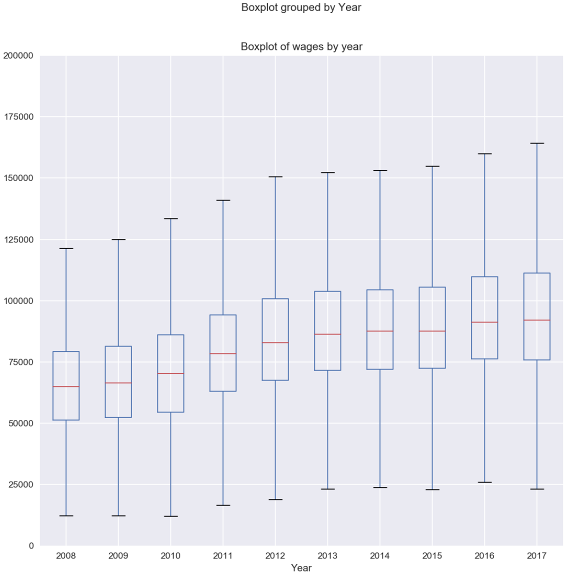
# Box plot of wages by job level



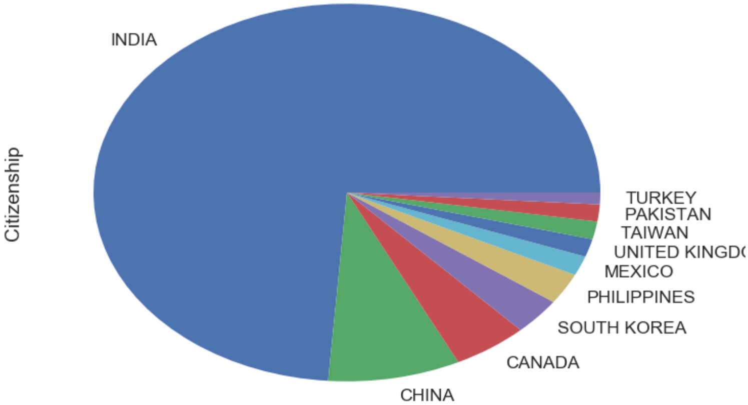
# Box plot of wages by sector



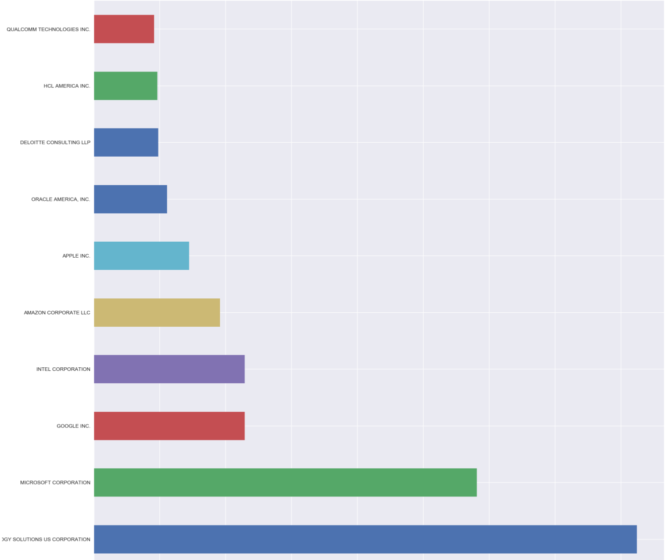
# Box plot of wages by year



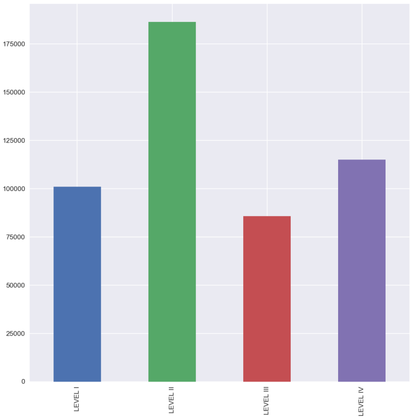
# Number of applications by citizenship



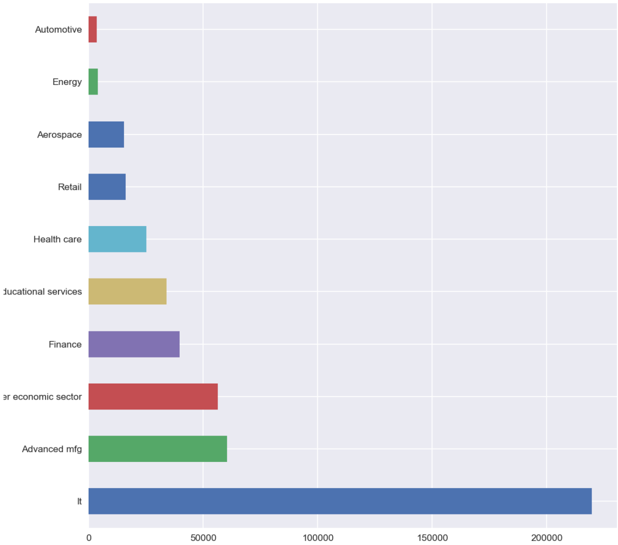
# Number of applications by firm



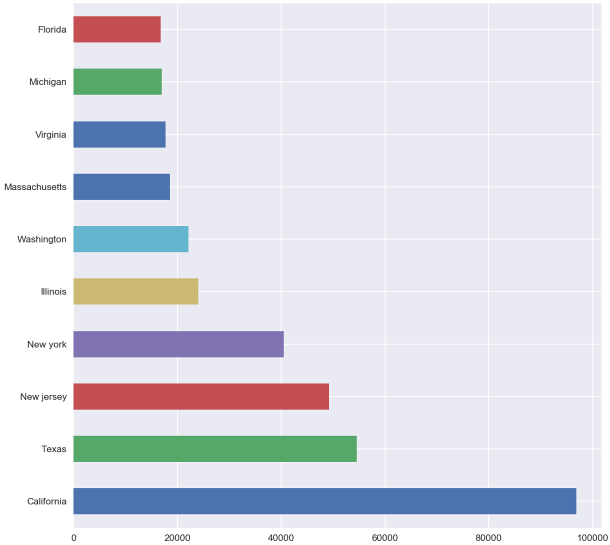
# Number of applications by job level



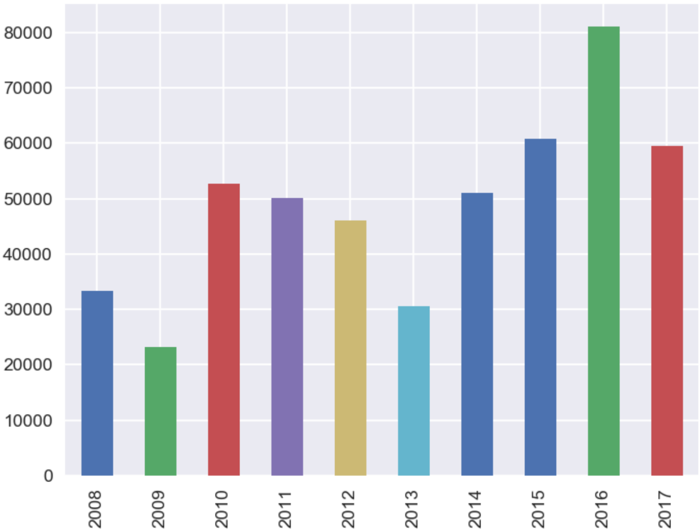
# Number of applications by sector



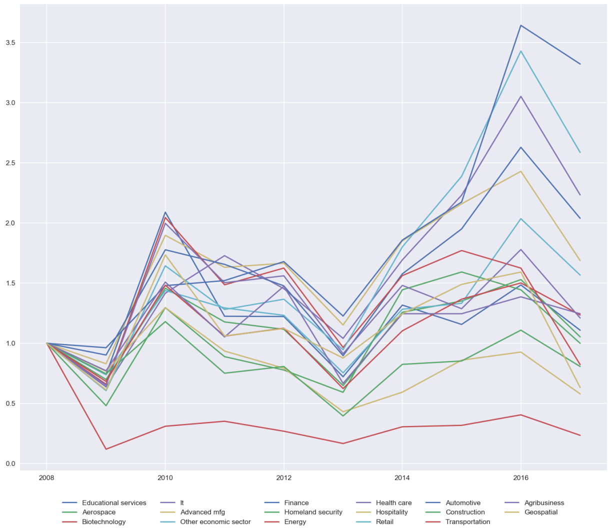
# Number of applications by states



# Number of applications by year



# Percentage change relative to 2008



# Linear model

Main objective:

- Understand how wages are related to each features
- Predict whether an H1-B case will be denied

We use linear models for both regression and classification problem

- Feature selection with L1 penalty
- Regression: LASSO
- Classification: Logistic regression with L1 penalty

# LASSO: optimization problem

- Response variable $Y \in R$ and predictor vector $X \in R^p$
- N observation pairs $(x_i, y_i)$
- For simplicity assume variables are standardized $\sum_{i=1}^{N} x_{ij} = 0$, and $\frac{1}{N}\sum_{i=1}^{N} x_{ij}^2 = 1$

LASSO solves the following problem

$$\min_{(\beta_0,\beta)\in\mathbb{R}^{p+1}} R_\lambda(\beta_0,\beta) = \min_{(\beta_0,\beta)\in\mathbb{R}^{p+1}} [\frac{1}{2N}\sum_{i=1}^{N}(y_i - \beta_0 - x_i{}^T\beta)^2 + \lambda P(\beta)]$$

where

$$P(\beta) = ||\beta||_{\mathbb{L}_1} = \sum_{j=1}^{p} |\beta_j|$$

# LASSO: coordinate descent

Coordinate descent: partially optimize with respect to one coordinate, assuming other coefficients are known.

- Suppose we have estimates $\tilde{\beta}_0$ and $\tilde{\beta}_l$ for $l \neq j$, and we wish to partially optimize with respect to $\beta_j$.
- We want the gradient at $\beta_j = \tilde{\beta}_j$. Because of L1 penalty, it only exists if $\tilde{\beta}_j \neq 0$

# LASSO: coordinate descent

If $\tilde{\beta}_j > 0$,

$$\frac{\partial R_\lambda}{\partial \beta_j}|_{\beta=\tilde{\beta}} = -\frac{1}{N}\sum_{i=1}^{N} x_{ij}(y_i - \tilde{\beta}_0 - x_i^T\tilde{\beta}) + \lambda$$

If $\tilde{\beta}_j < 0$,

$$\frac{\partial R_\lambda}{\partial \beta_j}|_{\beta=\tilde{\beta}} = -\frac{1}{N}\sum_{i=1}^{N} x_{ij}(y_i - \tilde{\beta}_0 - x_i^T\tilde{\beta}) - \lambda$$

# LASSO: CD naive update scheme

Because we assume standardization $\frac{1}{N}\sum x_{ij}^2 = 1$

$$\tilde{\beta}_j \leftarrow S(\frac{1}{N}\sum_{i=1}^{N} x_{ij}(y_i - \tilde{y}_i^{(j)}), \lambda)$$

where

- $\tilde{y}_i{}^{(j)} = \tilde{\beta}_0 + \sum_{l \neq j} x_{il}\tilde{\beta}_l$ is the fitted value excluding $x_{ij}$, and thus $y_i - \tilde{y}_i{}^{(j)}$ is the partial residual of fitting $\beta_j$.
- $S(z, \gamma)$ is the soft-thresholding operator with value

$$sign(z)(|z| - \gamma)_+$$

Many coefficients will remain 0 in updating, thus no need to change (feature selection).

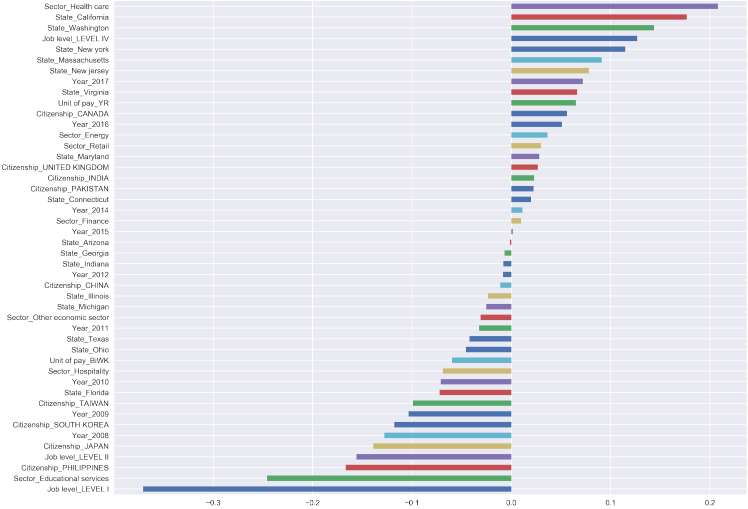
# LASSO Application: Wage Regression

Two models:

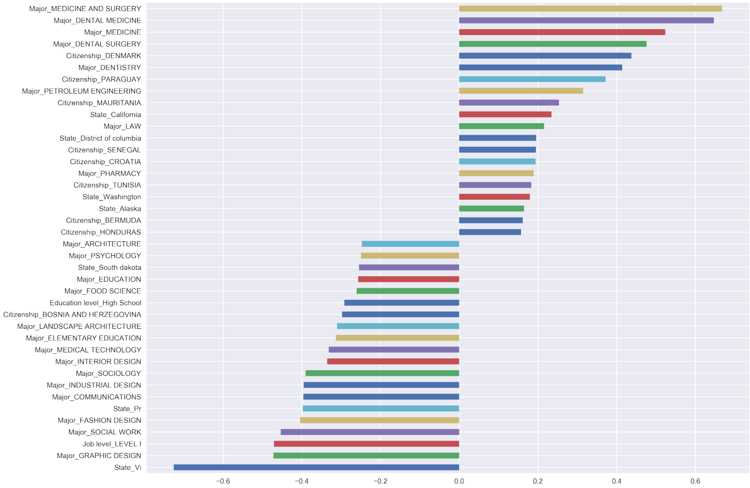
- Data from 2008 to 2017
  - Features: sector, state, citizenship, job level, pay unit, year
- Data from 2015 to 2017
  - Features: sector, state, citizenship, job level, pay unit, year, major, education level, ownership interest, prior job experience, employer's founding year, employer's total employee number

Use 10-fold cross validation and grid search for best $\lambda$

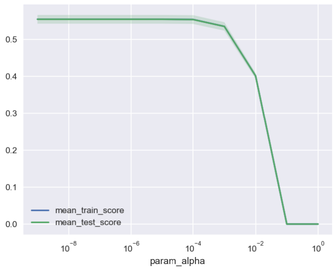
# Feature selection from wage: pre2015

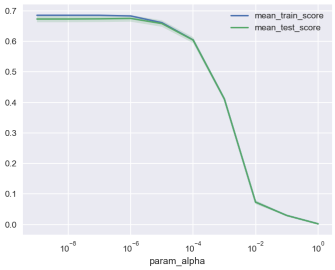


# Feature selection from wage: post2015



# Regularization path



# Wage regression: conclusion

- Pre-2015 most important features: Healthcare, California, Washington, Job level IV, New York, Massachusetts
- Pre-2015 negative contributors: Job level I and II, Educational services, Philippines, Japan, Year 2008
- Post-2015 added features: Medicine and surgery major, petroleum engineering major, Law major, pharmacy major
- Post-2015 negative contributors: Graphic Design major, Social work major, Fashion Design Major, Communications major

# Status prediction

Classification with 2 states, use logistic regression with L1 penalty.
Let $y = 0, 1$ be the response variable, logistic regression model:

$$Pr(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + x^T\beta)}}$$

$$Pr(y = 0|x) = 1 - Pr(y = 1|x) = \frac{1}{1 + e^{(\beta_0 + x^T\beta)}}$$

# Logistic regression: Maximum likelihood

Let $p(x_i) = Pr(y = 1|x_i)$, the maximum log likelihood problem for logistic regression with L1 penalty is given as

$$\max_{(\beta_0,\beta)\in\mathbb{R}^{p+1}} [\frac{1}{N}\sum_{i=1}^{N}\{y_i \log p(x_i) + (1-y_i)\log(1-p(x_i))\} - \lambda P_{\mathbb{L}_1}(\beta)]$$

Let's call the un-penalized likelihood function $\ell(\beta_0, \beta)$

$$\ell(\beta_0,\beta) = \frac{1}{N}\sum_{i=1}^{N} y_i(\beta_0 + x_i^T\beta) - \log(1 + e^{(\beta_0 + x_i^T\beta)})$$

# Logistic regression: quadratic approximation

Let $\tilde{\beta}_0$ and $\tilde{\beta}$ be current estimators, we do Taylor expansion of the likelihood function about current estimate

$$\ell_Q(\beta_0, \beta) := -\frac{1}{2N} \sum_{i=1}^{N} w_i (z_i - \beta_0 - x_i^T \beta)^2 + \mathcal{O}(||\beta - \tilde{\beta}||^2)$$

where

$$z_i = \tilde{\beta}_0 + x_i^T \tilde{\beta} + \frac{y_i - \tilde{p}(x_i)}{\tilde{p}(x_i)(1 - \tilde{p}(x_i))}$$

$$w_i = \tilde{p}(x_i)(1 - \tilde{p}(x_i))$$

# Logistic regression: Coordinate descent

Now the problem becomes

$$\min_{(\beta_0,\beta)\in\mathbb{R}^{p+1}}[-\ell_Q(\beta_0,\beta)+\lambda P_{\mathbb{L}}(\beta)]$$

It is similar to the optimization problem in LASSO:

$$\min_{(\beta_0,\beta)\in\mathbb{R}^{p+1}}[\frac{1}{2N}\sum_{i=1}^{N}(y_i-\beta_0-{x_i}^T\beta)^2+\lambda P_{\mathbb{L}}(\beta)]$$

and can be solved using the coordinate descent update scheme

# Alternative: Newton's method

For single variable convex function $f(x)$, if we want to find its minimum, the update is given by

$$x^i = x^{i-1} - \frac{f'(x^{i-1})}{f''(x^{i-1})}$$

If the function is multivariate,

$$\mathbf{x}^i = x^{i-1} - (\nabla^2 f(\mathbf{x}^{i-1}))^{-1} \nabla f(\mathbf{x}^{i-1})$$

It it optimizing using the knowledge of second order derivative. Better solution but computationally costly.

# Newton's Method: Logistic regression with L1 penalty

$$\min_{(\beta_0,\beta)\in\mathbb{R}^{p+1}}[-\ell_Q(\beta_0,\beta)+\lambda P_{\mathbb{L}}(\beta)]$$

$$=\min_{(\beta_0,\beta)\in\mathbb{R}^{p+1}}[\frac{1}{2N}\sum_{i+1}^{N} w_i(z_i-\beta_0-x_i^T\beta)^2+\lambda P_{\mathbb{L}}(\beta)$$

In matrix form,

$$f=\frac{1}{2N}(\mathbf{Z}-\mathbf{X}\beta)^T\mathbf{W}(\mathbf{Z}-\mathbf{X}\beta)+\lambda P_{\mathbb{L}}(\beta)$$

The gradient of $f$:

$$\nabla f=-\frac{1}{N}\mathbf{X}^T\mathbf{W}\mathbf{Z}+\frac{1}{N}\mathbf{X}^T\mathbf{W}\mathbf{X}\beta+\lambda\frac{\partial P_{\mathbb{L}}(\beta)}{\partial\beta}$$

The Hessian is given by

$$\nabla^2 f=\frac{1}{N}\mathbf{X}^T\mathbf{W}\mathbf{X}$$

# Logistic regression application: Status prediction

We use the data from 2015 to 2017, excluding data points with empty features (in total 191693 observations).

- Features: wage, sector, job level, pay unit, state, education level, job experience, major, employer's founding year, ownership interest, employee number
- Two status: certified (184308) and denied (7385)

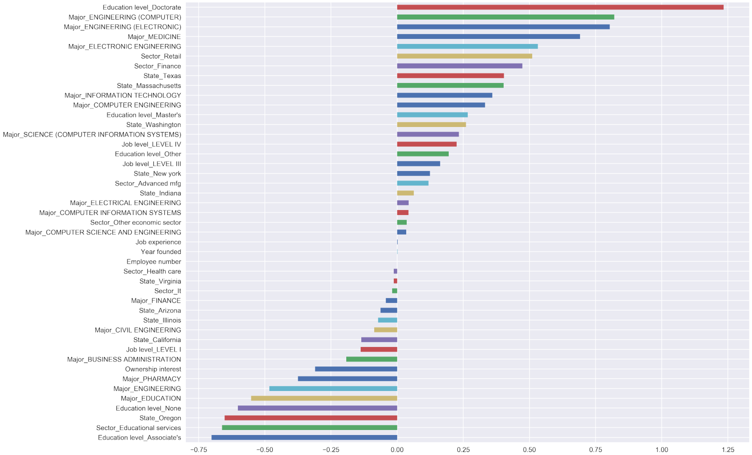
Since the dataset is very unbalanced, we need to do random sampling.

# Status prediction: random sampling

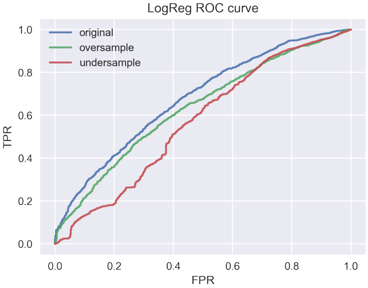
Model evaluation: AUC (Area under ROC curve)
We want to find the parameter $\lambda$ such that it maximize the AUC of the model

- Generate a list of $\{\lambda_1, \lambda_2, \cdots, \lambda_n\}$
- For each $\lambda_i$:
  - Oversampling from denied class
  - Undersampling from certified class
  - For each sample:
    - Use coordinate descent method to compute MLE of penalized LogReg with quadratic approximation of each sample
    - Calculate AUC for each model
  - Calculate average AUC and call it the score of $\lambda_i$
- Find the $\lambda_i$ with the highest average AUC

# Status prediction: result



# Status prediction: result




# Status prediction: conclusion

- Status prediction most important factors: education level PHD, Major Computer science, Major electronic, Major medicine, Major electronic engineering, Retail Sector, Finance Sector
- Status prediction negative contributors: Educational level associate, educational services sector, State Oregon, Educational level non, Major in education
- Original sample gives the best performance

# Future

- For the field
- For us