What's in a Name?

Gabrielle Salamanca

1.
The Dataset

Rnames

Are Emily and Greg More Employable than Lakisha and Jamal?

About the Data

- Vincent Arel-Bundock's Github projects
  - R datasets
- 4870 rows x 23 columns
  - Applicants' resumes
  - Job requirements
  - Type of job

The Origins

- The National Bureau of Economic Research
  - Working Papers
- *Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination*
  - Marianne Bertrand
  - Sendhil Mullainathan

The Experiment

- Send 4,870 fictitious resumes to job adverts
  - Randomly assigned a name
    - Caucasian or African-American
  - Which ones got a call back

Was the name the main reason for not receiving a call back, or were there other factors affecting it?

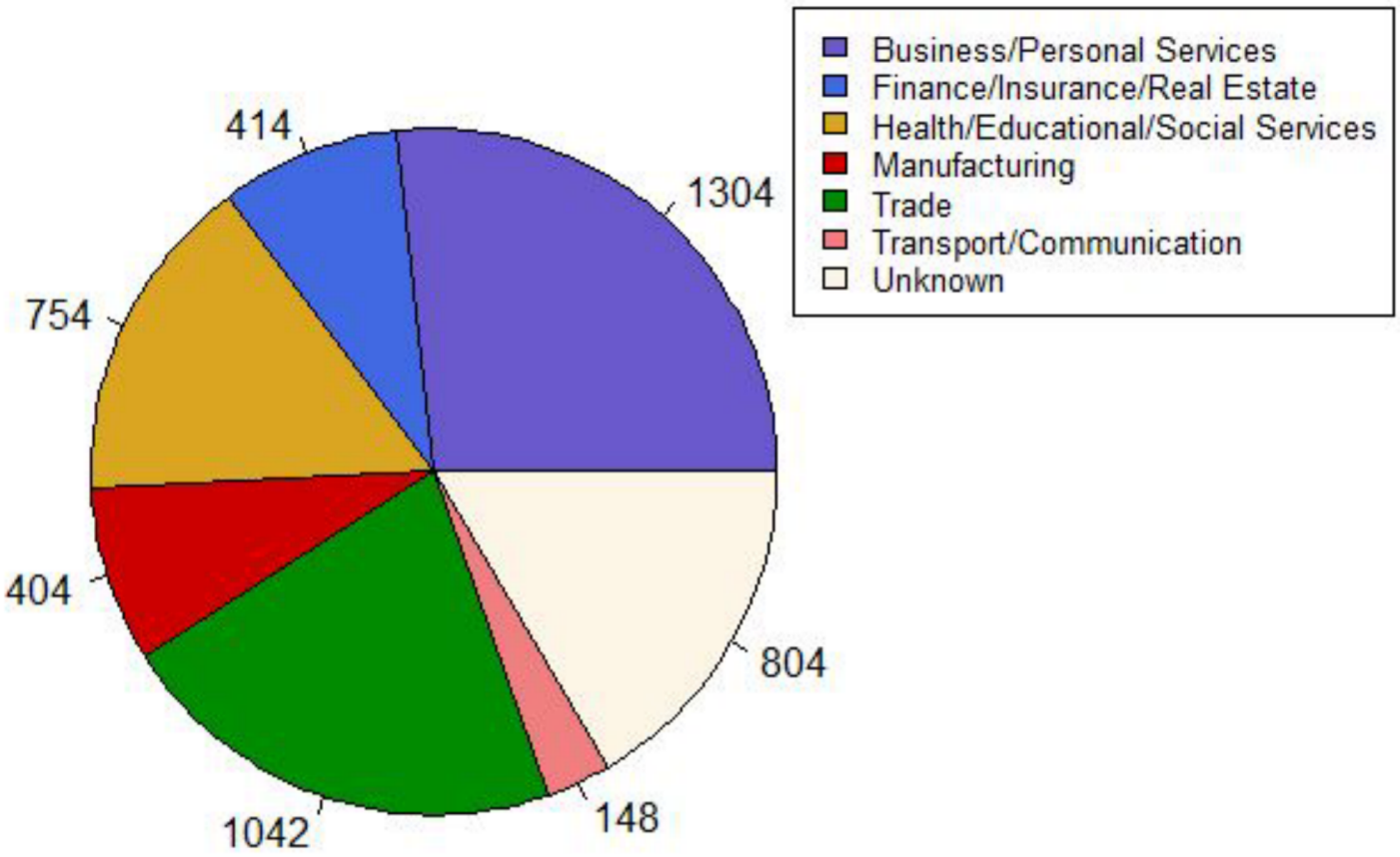
Goal: Classification

- Find a model that can provide the most accurate prediction
- Response variable
  - Call back

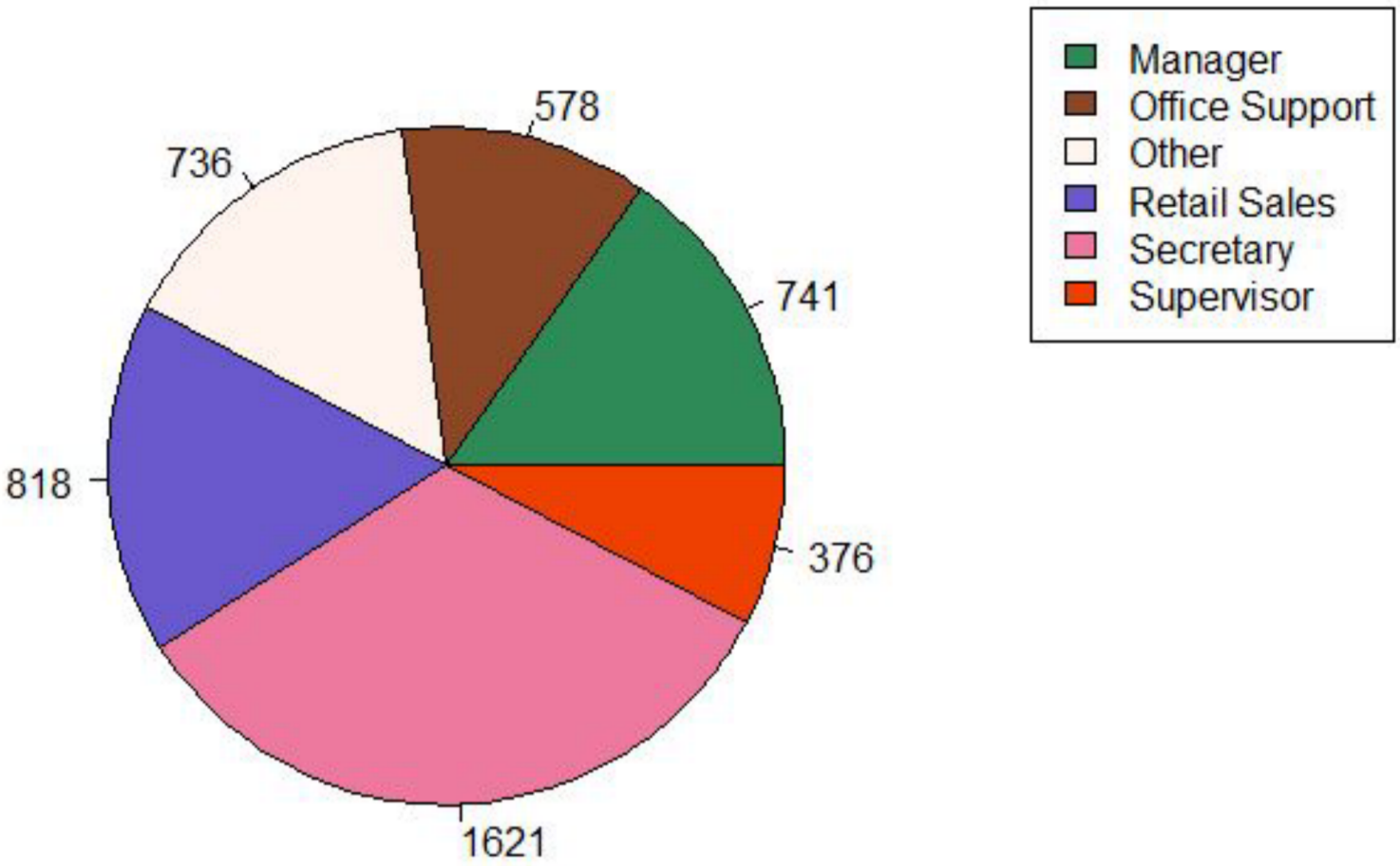
Data Visualization

Resume Names

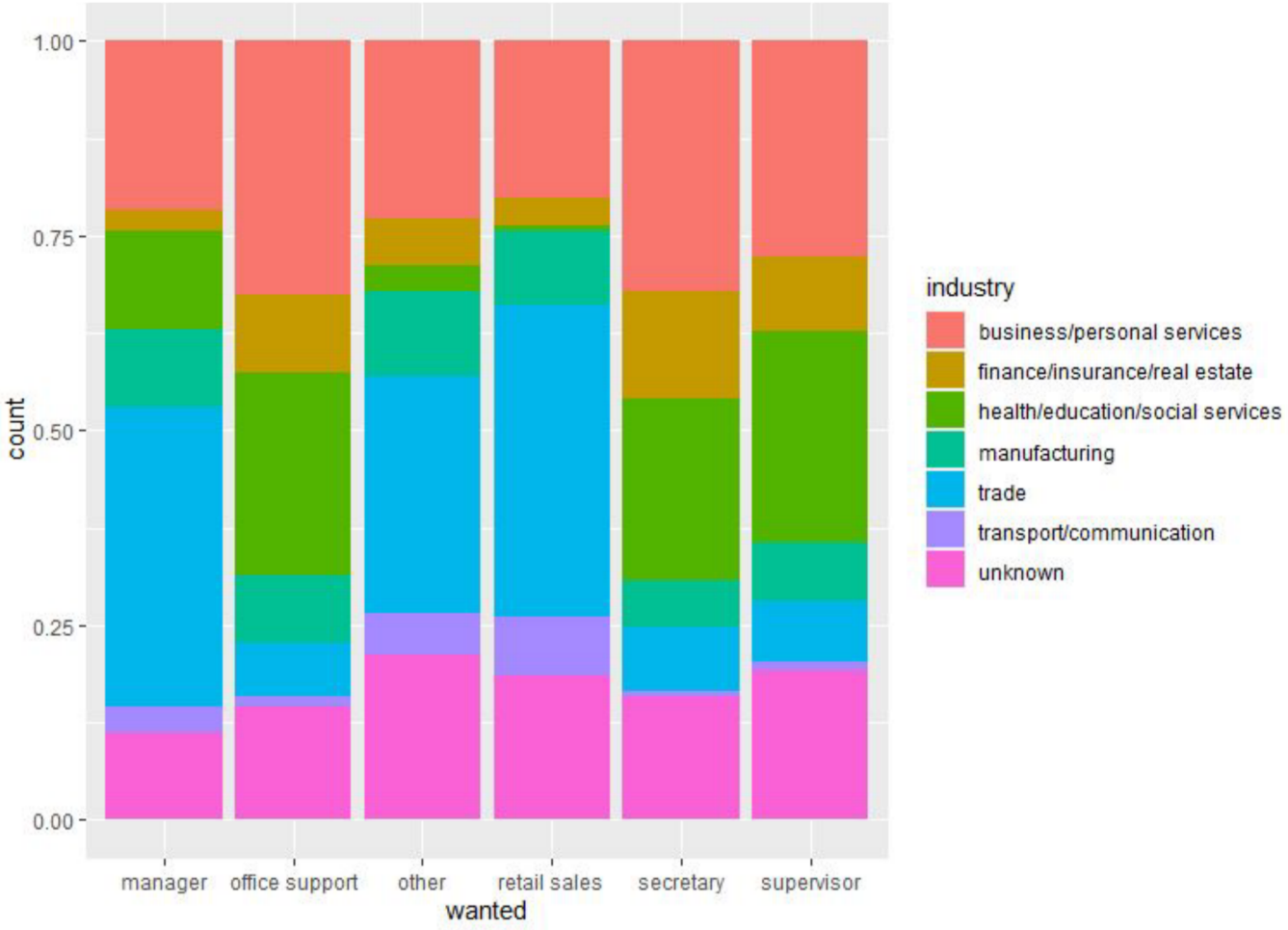
X	*name*	*gender*	*ethnicity*	*quality*	*call*	*city*	. . .
1	Allison	female	cauc	low	no	chicago	. . .
2	Kristen	female	cauc	high	no	chicago	. . .
3	Lakisha	female	afam	low	no	chicago	. . .
4	Latonya	female	afam	high	no	chicago	. . .
5	Carrie	female	cauc	high	no	chicago	. . .
6	Jay	male	cauc	low	no	chicago	. . .
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .



What industries were the jobs offered in?



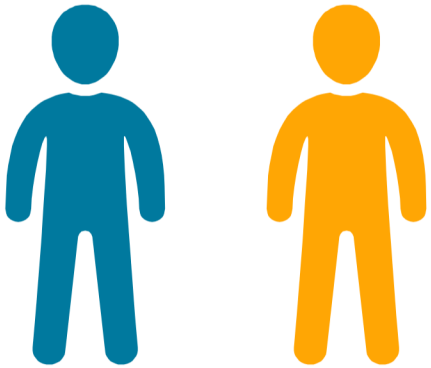
What were the offered positions?

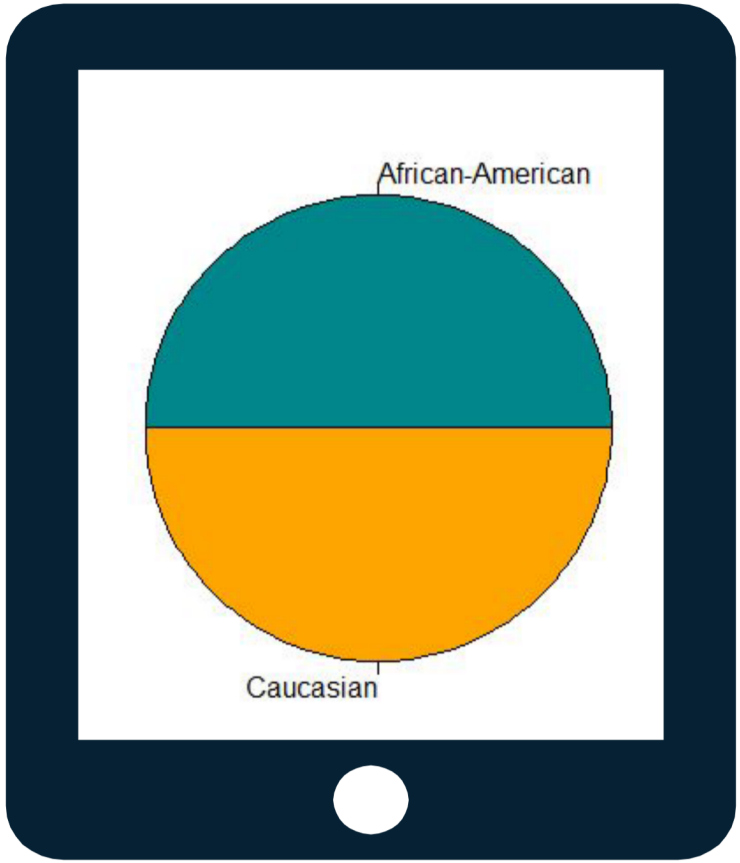


Positions in terms of Industry

Ethnicity: Caucasian vs African-American

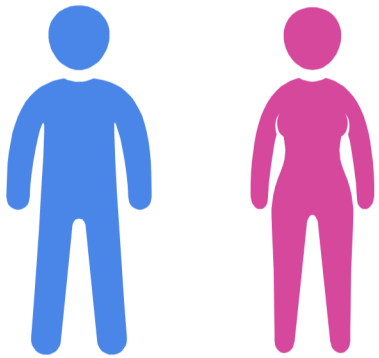
- 4,780 applicants
  - 2,435 Caucasian
  - 2,435 African-American

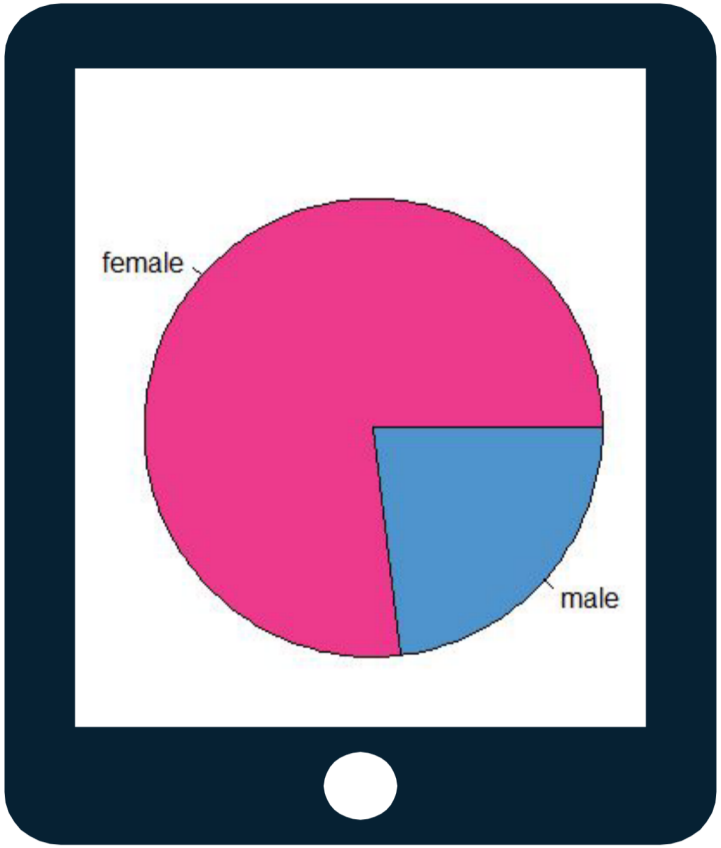




Gender: ♀ vs ♂

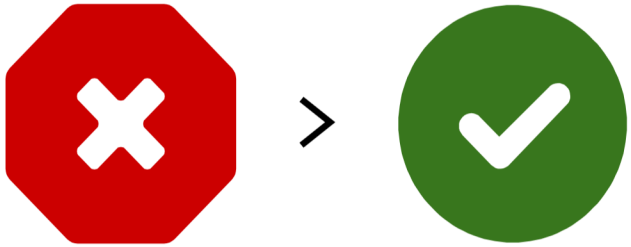
- 4,870 applicants
  - 3,746 female
  - 1,124 male

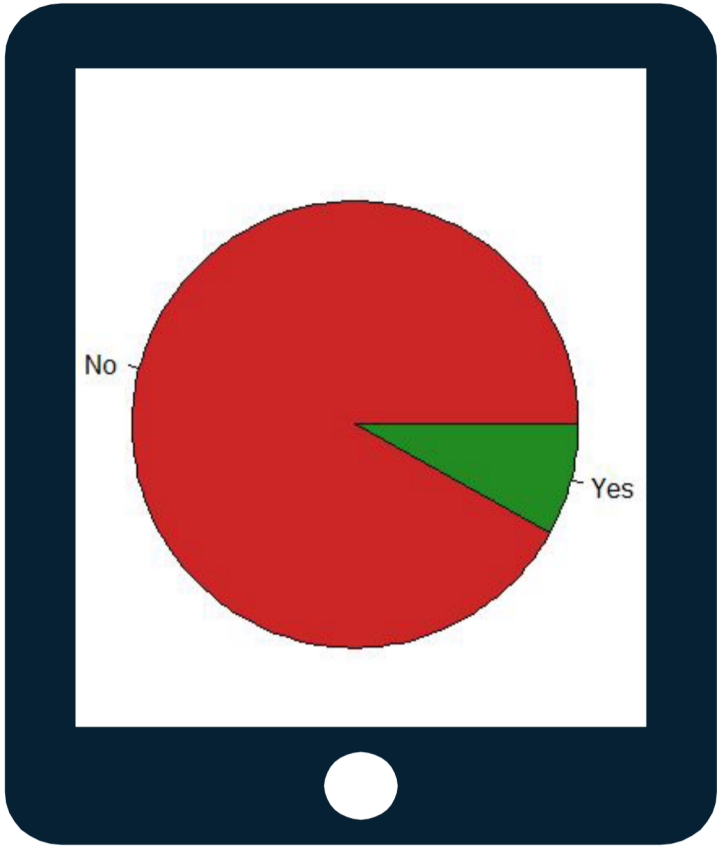




Call back: Yes vs No

- 4,870 applicants
  - 4478 'no's
  - 392 'yes's



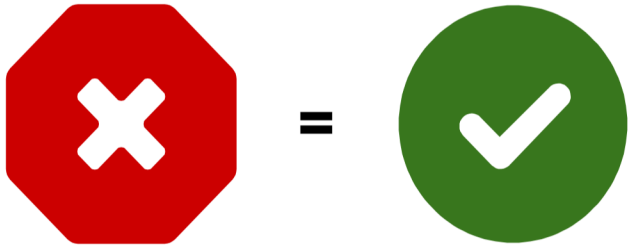


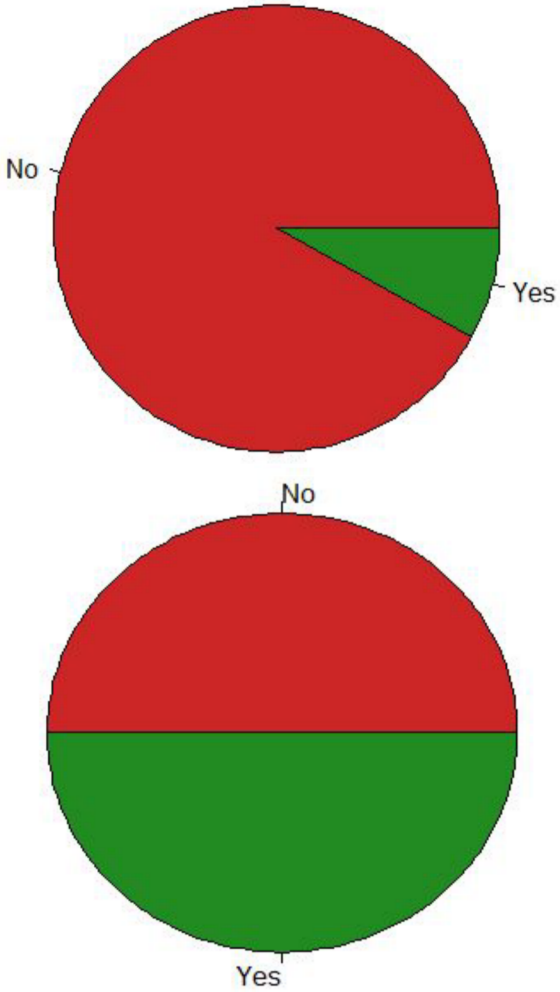
2. Cleaning and Reorganizing

Rnames → Names

Downsampling

- Most important variable
  - Call back
- Before vs After
  - 4,478 vs 392
  - 392 vs 392





Updating Dataset

- Newnames
  - downSample(Rnames, Rnames$call)
- Names
  - Removing variables
    - subset()
- X
  - Cell number
- Name
  - Descriptor of Ethnicity
- City
  - Boston vs Chicago
- Class
  - Added by downSample()
  - Classifying each case

Industry - Wanted Removal?

- Originally excluded due to "missing values"
  - unknown/other

During further testing

- Inclusion of Industry - Wanted
  - Accuracy boosted (in some cases)

3. Methods

What have I tried?

GLM Fitting

Splitting the Data

- 80% training
- 20% test

summary()

- Finding the most significant variables
- Use a variation of those variables to find the best fit

Without Industry-Wanted Vars

Accuracy

- 55% < y < 65%

Best fit

- 63.69427%
  - glm.fit1

Variables

- Ethnicity
- Experience
- Special
- Reqeduc

With Industry-Wanted Vars

Accuracy

- 50% < y < 65%

Best fit

- 63.69427%
  - glm.fit5
  - glm.fit7

Variables

- Ethnicity (5, 7)
- Experience (5, 7)
- Special (5, 7)
- Reqeduc (5, 7)
- Honors (5, 7)
- Reqcomp (5, 7)
- Requirements (7)
- Industry (5, 7)

The Methods

- Best Subset Selection
  - Cross-Validation
- LDA
- QDA
- KNN
- Ridge Regression
- LASSO

The Fits Being Used

glm.fit1

- No Industry var
- 58.59873%

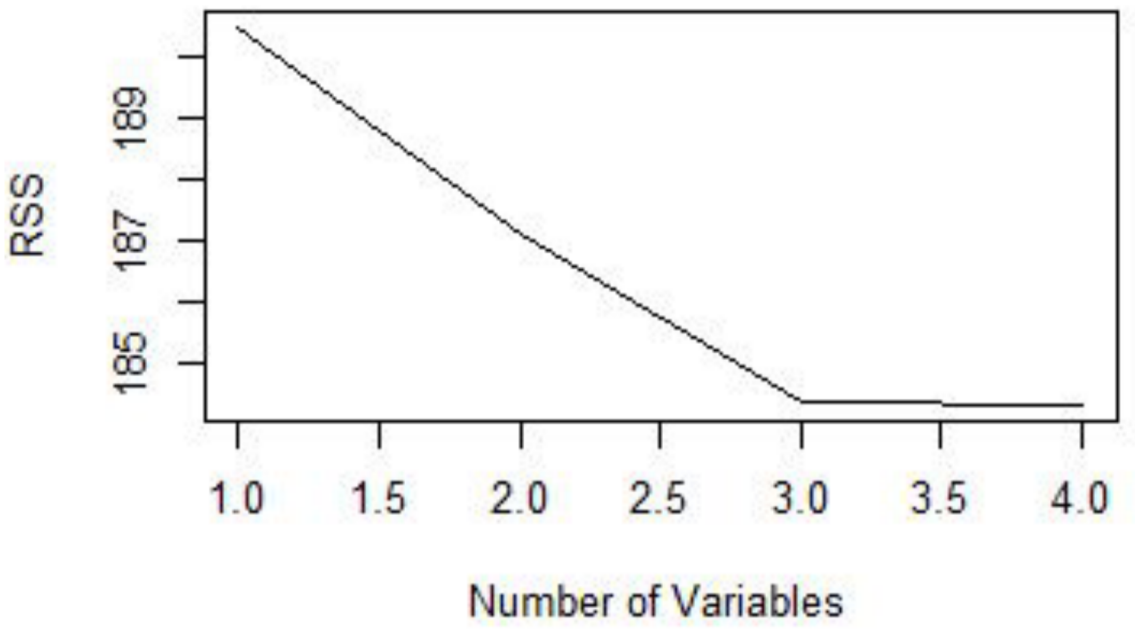
	No	Yes
No	48	30
Yes	35	44

glm.fit5

- Industry var
- 62.42038%

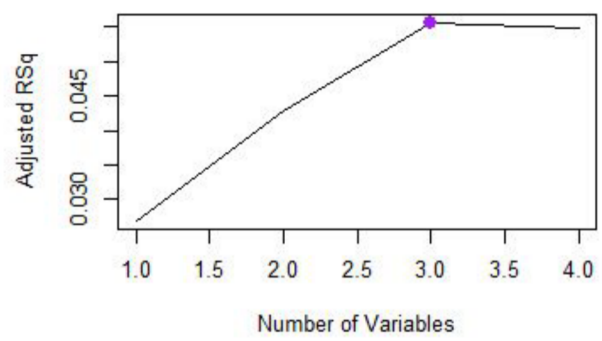
	No	Yes
No	54	31
Yes	28	44

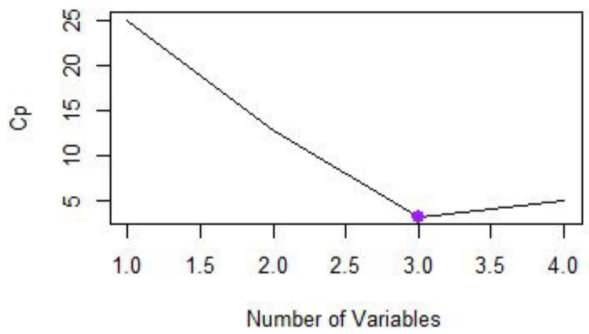
Best Subset Selection: No Industry

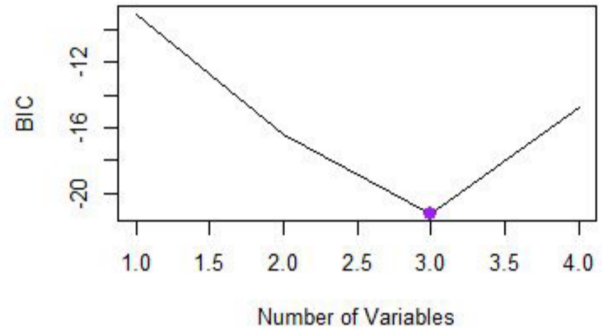


When n > 3, RSS is small

Cross-Validation: No Industry



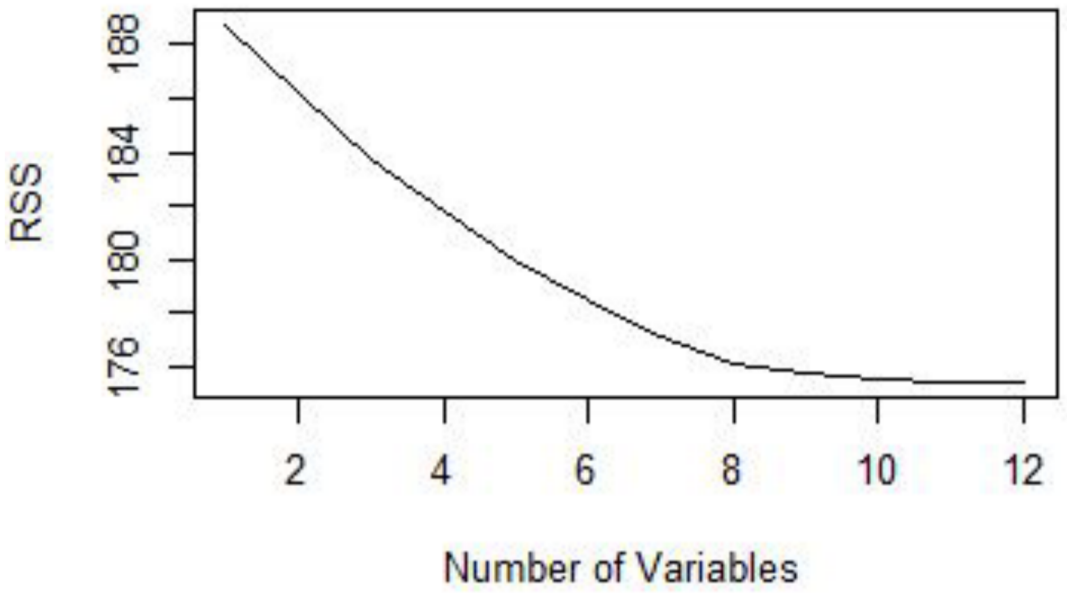




N = 3

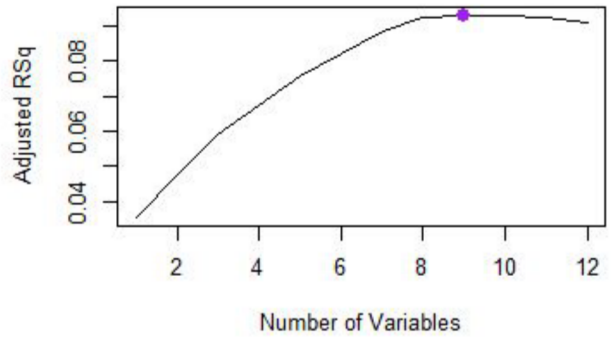
- Adjusted R^2
- BIC
- C_p

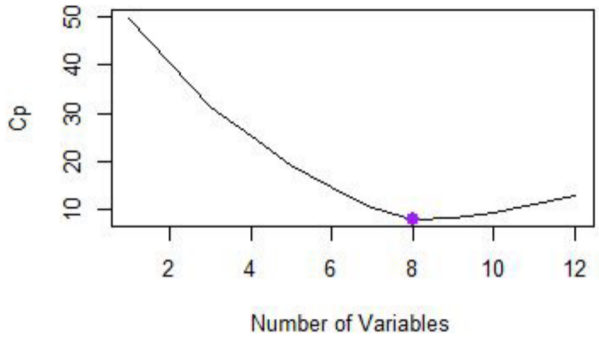
Best Subset Selection: Industry

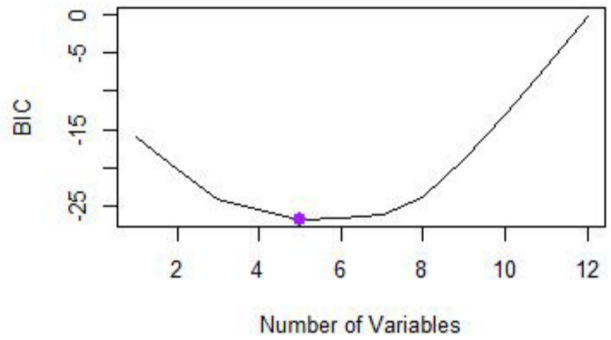


When n > 10, RSS is small

Cross-Validation: Industry







Adjusted R^2

- N = 9

BIC

- N = 5

C_p

- N = 8

Classification Methods: No Industry

LDA

- 58.59873%

QDA

- 54.77707%

KNN

- K = 1
  - 60.50955%
- K = 5
  - 64.96815%

LDA		
	No	Yes
No	48	30
Yes	35	44

QDA		
	No	Yes
No	51	39
Yes	32	35

KNN		
K = 1	No	Yes
No	55	34
Yes	28	40

KNN		
K = 5	No	Yes
No	54	26
Yes	29	48

Classification Methods: Industry

LDA

- 62.42038%

QDA

- 61.78344%

KNN

- K = 1
  - 60.50955%
- K = 5
  - 64.33121%

LDA			QDA		
	No	Yes		No	Yes
No	54	31	No	60	38
Yes	28	44	Yes	22	37

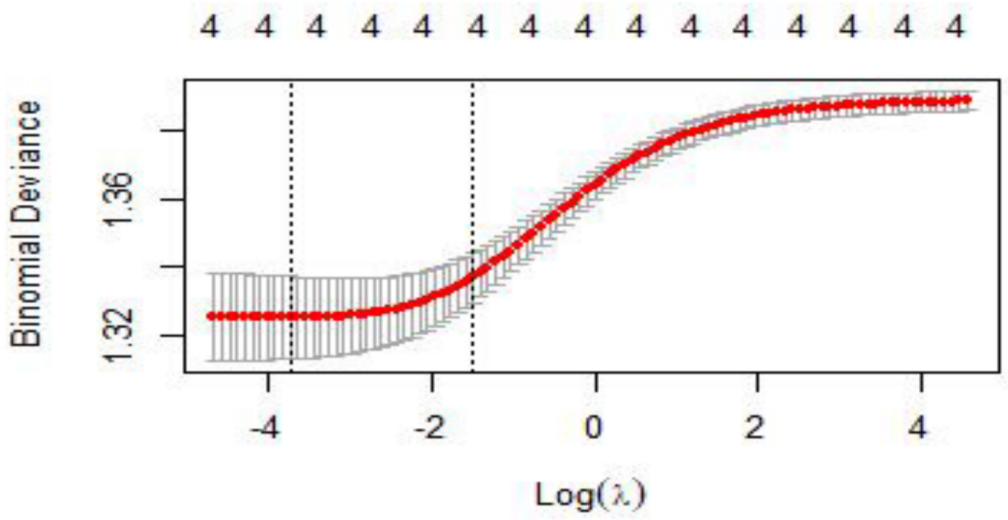
KNN			KNN		
K = 1	No	Yes	K = 5	No	Yes
No	53	33	No	52	26
Yes	29	42	Yes	30	49

Ridge Regression: No Industry

Accuracy

- 59.87261%

	No	Yes
No	51	31
Yes	32	43



Ridge Regression: No Industry

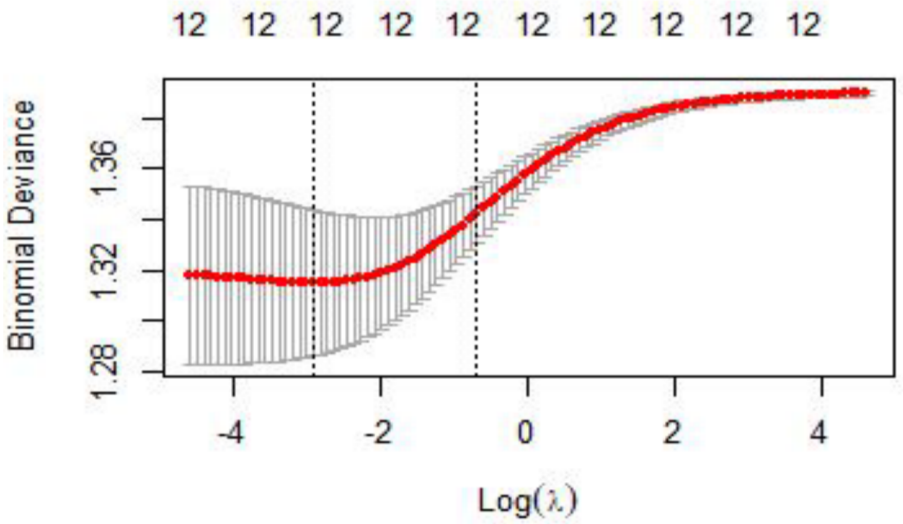
Ridge	Coefficients
Intercept	— 0.8972270
Ethnicity (cauc)	0.4511977
Experience	0.0492280
Special (yes)	0.6192874
Reqeduc (yes)	— 0.1183023

Ridge Regression: Industry

Accuracy

- 63.69427%

	No	Yes
No	55	30
Yes	27	45



Ridge Regression: Industry

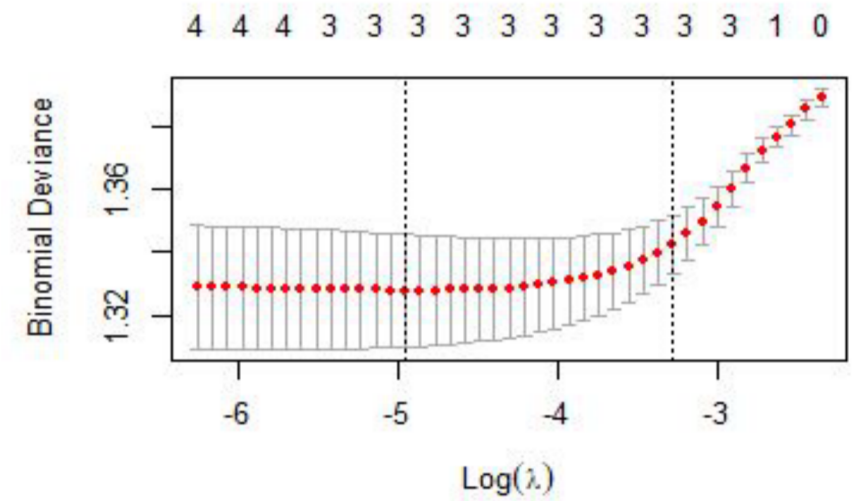
Ridge	Coefficients	Ridge	Coefficients
Intercept	— 0.46743944	Reqcomp (yes)	— 0.35486429
Ethnicity (cauc)	0.30914913	Industry • Finance • Insurance • Real Estate	— 0.24239488
Experience	0.03065452		
Special (yes)	0.56394817		
Reqeduc (yes)	— 0.53606360	Industry • Health • Education • Social Services	0.33985341
Honors (yes)	0.58931603		

LASSO: No Industry

Accuracy

- 59.87261%

	No	Yes
No	51	31
Yes	32	43



LASSO: No Industry

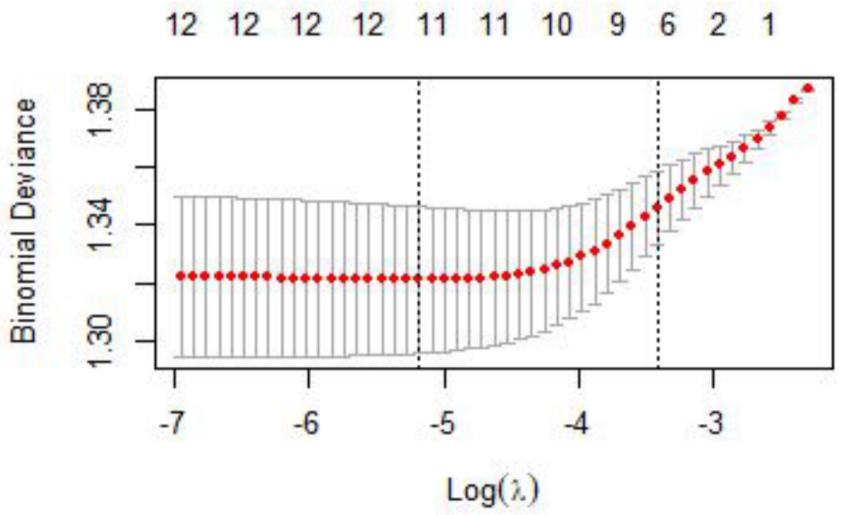
Ridge	Coefficients
Intercept	— 0.84673040
Ethnicity (cauc)	0.40982207
Experience	0.04588278
Special (yes)	0.59580626
Reqeduc (yes)	0

LASSO: Industry

Accuracy

- 64.33121%

	No	Yes
No	55	29
Yes	27	46



LASSO: Industry

LASSO	Coefficients	LASSO	Coefficients
Intercept	— 0.59813958	Reqcomp (yes)	— 0.40395699
Ethnicity (cauc)	0.34333831	Industry • Finance • Insurance • Real Estate	— 0.12613050
Experience	0.03385181		
Special (yes)	0.69530054		
Reqeduc (yes)	— 0.60125455	Industry • Health • Education • Social Services	0.39792262
Honors (yes)	0.62044904		

Best Classification Accuracy

Without Industry

KNN, K = 5

- 64.33121%

64.33121%

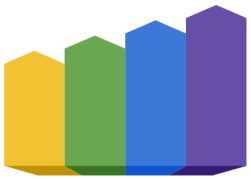
- KNN, K = 5
- LASSO

With Industry

Same percentage

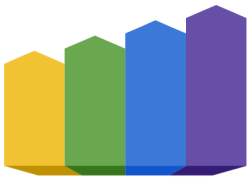
- Industry not as important as expected

The Rankings: No Industry



- KNN, K = 5
  - 64.96815%
- KNN, K = 1
  - 60.50955%
- Ridge Regression
  - 59.87261%
- LASSO
  - 59.87261%
- LDA
  - 58.59873%
- QDA
  - 54.77707%

The Rankings: Industry



- KNN, K = 5
  - 64.33121%
- LASSO
  - 64.33121%
- Ridge Regression
  - 63.69427%
- LDA
  - 62.42038%
- QDA
  - 61.78344%
- KNN, K = 1
  - 60.50955%

Conclusion

Best Accuracy Rate

- KNN, K = 5
  - With or without industry

Best RSS

- With Industry

Improvements?

- Minimum variable
  - Problems in code
  - Too many lvls
- Further explore Best Subset
  - Find best fits with the limitations

Credits

Special thanks to all the people who made and released these resources for free:

- Presentation template by SlidesCarnival
- Photographs by Unsplash

Bertrand, M. and Mullainathan, S. (2003). *Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination* (Working Paper No. 9873). National Bureau of Economic Research. http://www.nber.org/papers/w9873