# Data Science Salaries Analysis

Cheng Ji

# Contents

- The Data
- EDA
- Mine
- Refine
- Model
- Conclusion

# The Data

- Data Science salary data scraped from indeed.com
- Over 50,000 raw data, most of which are shit.
- After dropping data without salary info and duplicates, only 294 were used.

# EDA & Mine

- Create binary variable ‘high salary’ as Y
- Uniform location data to city level, and categorize it.
- Create ‘high_position’ feature from ‘title’

# Refine

- Utilize NLP on summary data
- Feature selection using Random Forest

# Refine

| | importance |
|---|---:|
| scientists | 0.029172 |
| scientist | 0.029123 |
| data | 0.028931 |
| high_position | 0.027156 |
| team | 0.026821 |
| location_num | 0.025325 |
| big | 0.024341 |
| analysis | 0.013789 |
| analytics | 0.012868 |
| large | 0.012140 |
| looking | 0.011791 |
| responsible | 0.011668 |
| company | 0.011122 |
| python | 0.010743 |
| experience | 0.010456 |

# Models

- Random Forest with grid search and cross validation
- Mean cross validation score: 0.74

# Models

- Confusion Matrix

|  | pred high | pred low |
|---|---|---|
| **high** | 102 | 45 |
| **low** | 31 | 115 |

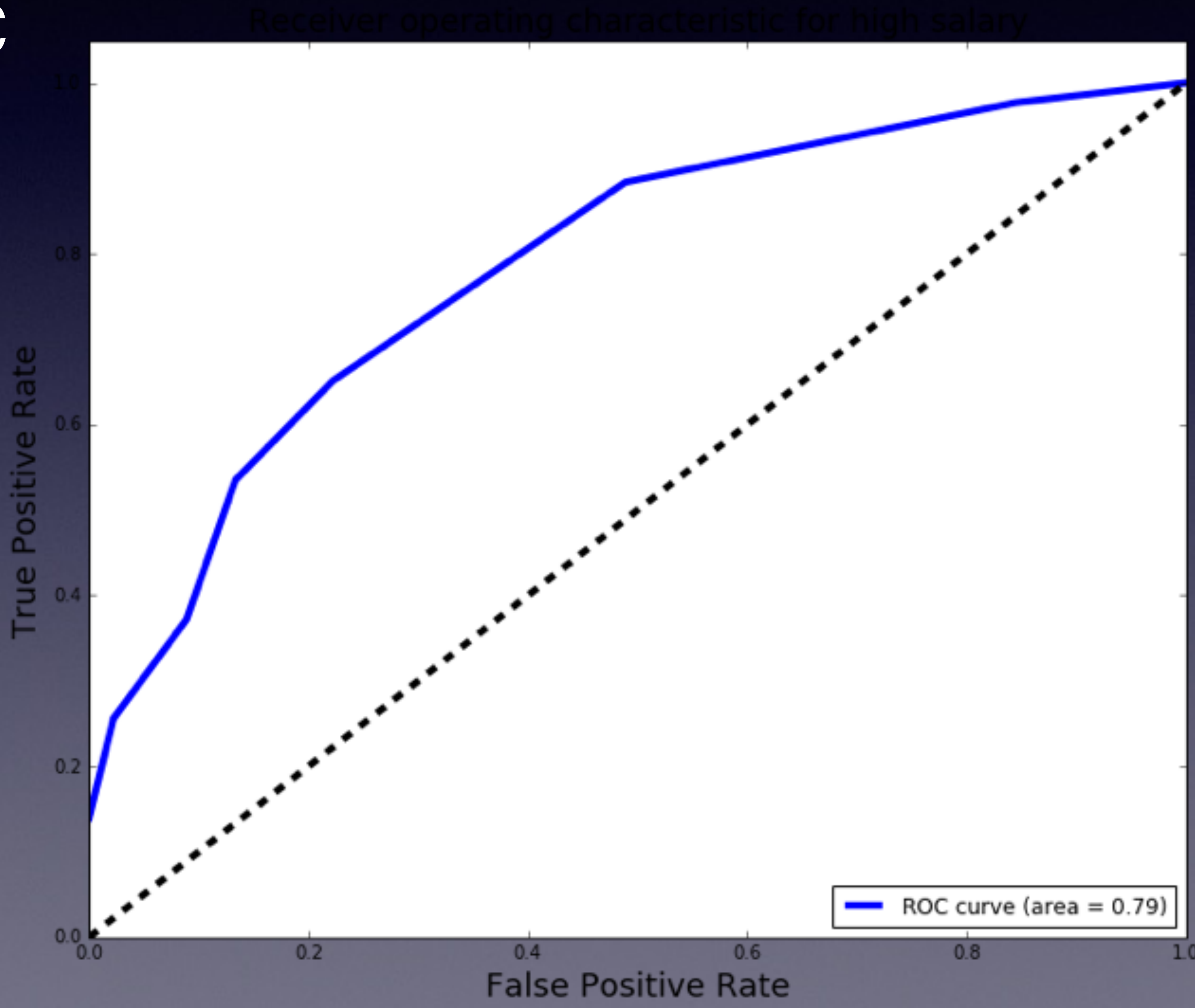
# Models

- Classification report

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.72 | 0.79 | 0.75 | 146 |
| 1 | 0.77 | 0.69 | 0.73 | 147 |
| avg / total | 0.74 | 0.74 | 0.74 | 293 |

# Models

- ROC


Receiver operating characteristic for high salary
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.79)

# Models

- SVM with linear kernel and cross validation
- Mean cross validation score: 0.75

# Models

- Confusion Matrix

| | pred high | pred low |
|---|---|---|
| **high** | 104 | 43 |
| **low** | 30 | 116 |

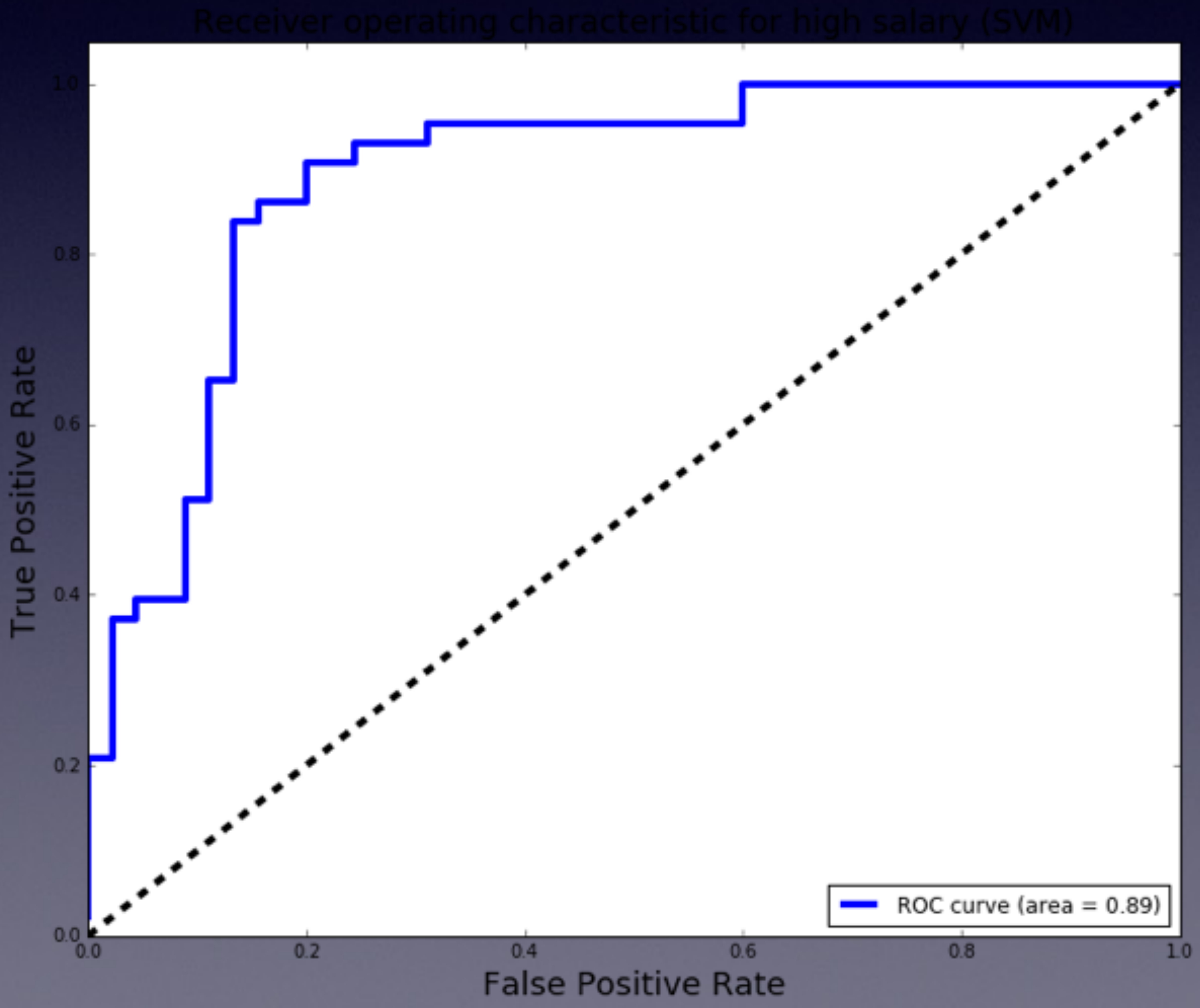
# Models

- Classification report

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.73 | 0.79 | 0.76 | 146 |
| 1 | 0.78 | 0.71 | 0.74 | 147 |
| avg / total | 0.75 | 0.75 | 0.75 | 293 |

# Models

- ROC


Receiver operating characteristic for high salary (SVM)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.89)

# Conclusion

- Current median level of data science salaries is about $105,000

- If we are focusing on managing the chances that incorrectly tell a client that he or she would get a high salary, then we recommend to set our false positive rate at 0.2, which gives us a true positive rate over 0.8

- Limitation: small dataset makes our prediction less reliable.