# 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
- 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
- Due to small dataset, no train test split. K folds cross validation is used.
- Feature selection using Random Forest

# Refine

| | importance |
|---|---|
| **data** | 0.069475 |
| **analytics** | 0.046889 |
| **team** | 0.032315 |
| **big** | 0.026158 |
| **location_num** | 0.024852 |
| **scientists** | 0.023610 |
| **scientist** | 0.022645 |
| **python** | 0.015266 |
| **high_position** | 0.013530 |
| **learning** | 0.012604 |
| **derivative** | 0.010777 |
| **client** | 0.010720 |

# Models

- Random Forest with grid search
- Mean cross validation score: 0.70

# Models

- Confusion Matrix

| | pred high | pred low |
|---|---|---|
| **high** | 89 | 58 |
| **low** | 30 | 116 |

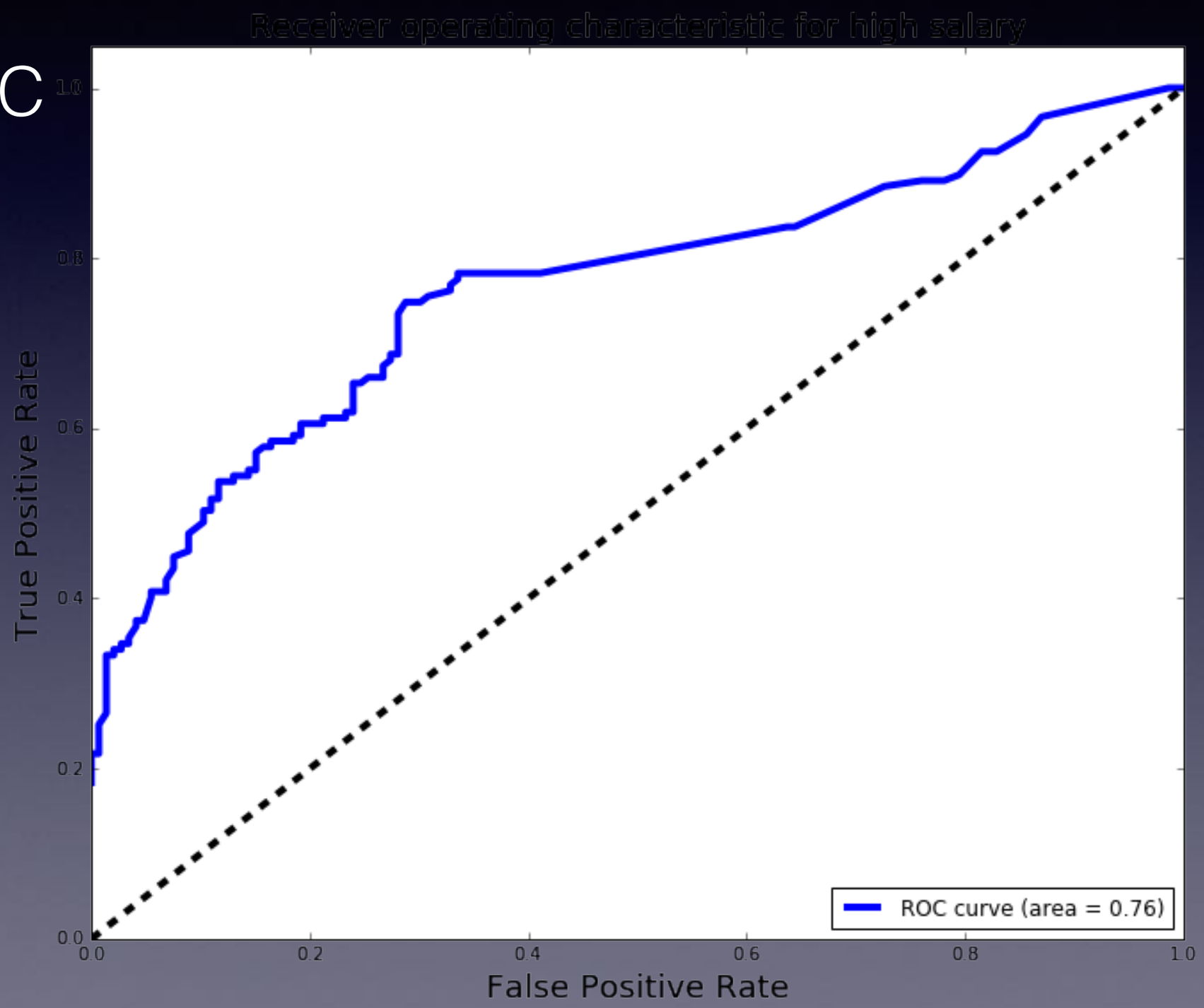
# Models

- Classification report

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.67 | 0.79 | 0.72 | 146 |
| 1 | 0.75 | 0.61 | 0.67 | 147 |
| avg / total | 0.71 | 0.70 | 0.70 | 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
ROC curve (area = 0.76)

# Models

- Gradient Boosting with grid search
- Mean cross validation score: 0.72

# Models

- Confusion Matrix

| | pred high | pred low |
|---|---|---|
| **high** | 95 | 52 |
| **low** | 30 | 116 |

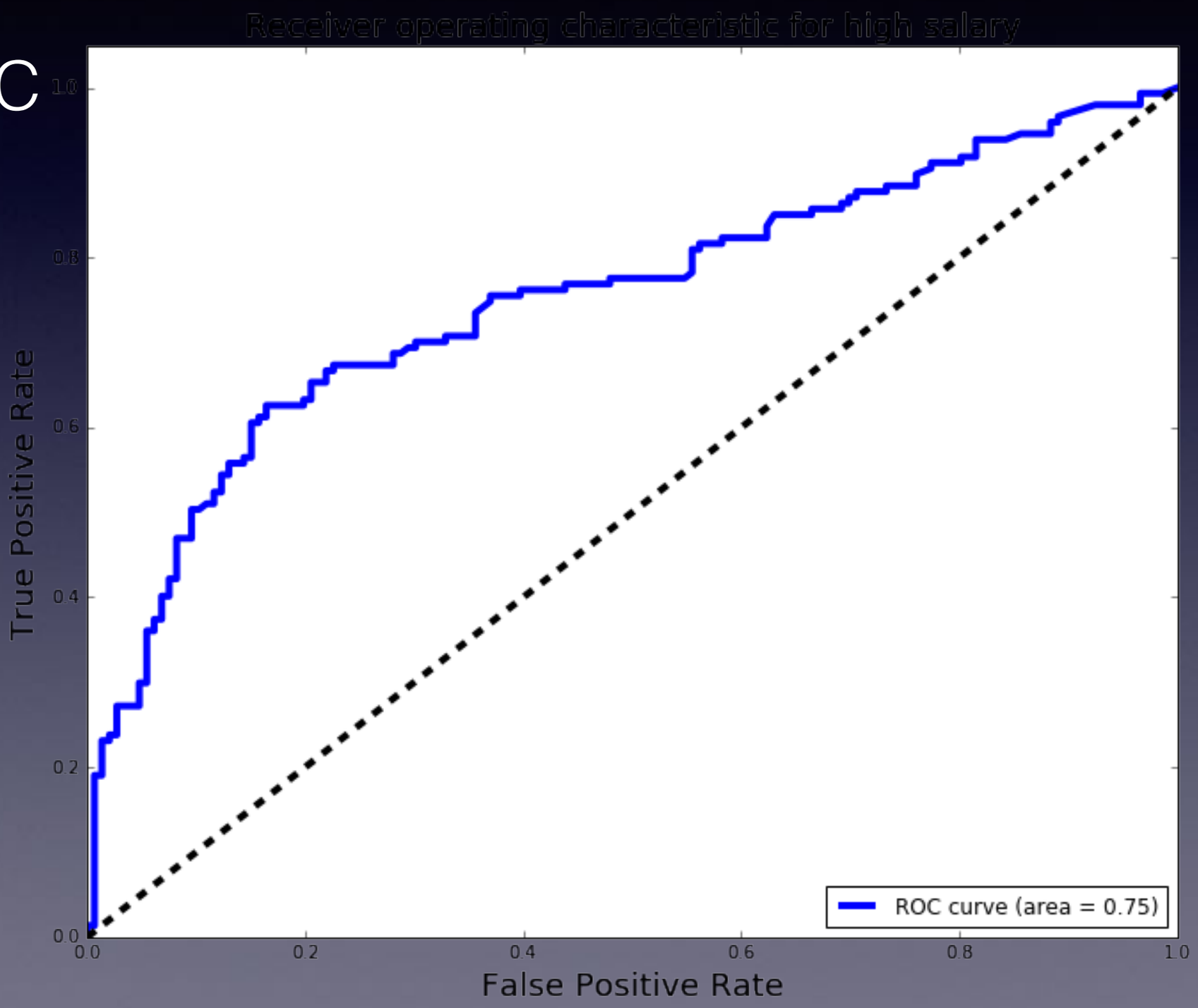
# Models

- Classification report

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.69 | 0.79 | 0.74 | 146 |
| 1 | 0.76 | 0.65 | 0.70 | 147 |
| avg / total | 0.73 | 0.72 | 0.72 | 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
ROC curve (area = 0.75)

# Models

- SVM with ref kernel
- Mean cross validation score: 0.71

# Models

- Confusion Matrix

| | pred high | pred low |
|---|---|---|
| **high** | 108 | 39 |
| **low** | 47 | 99 |

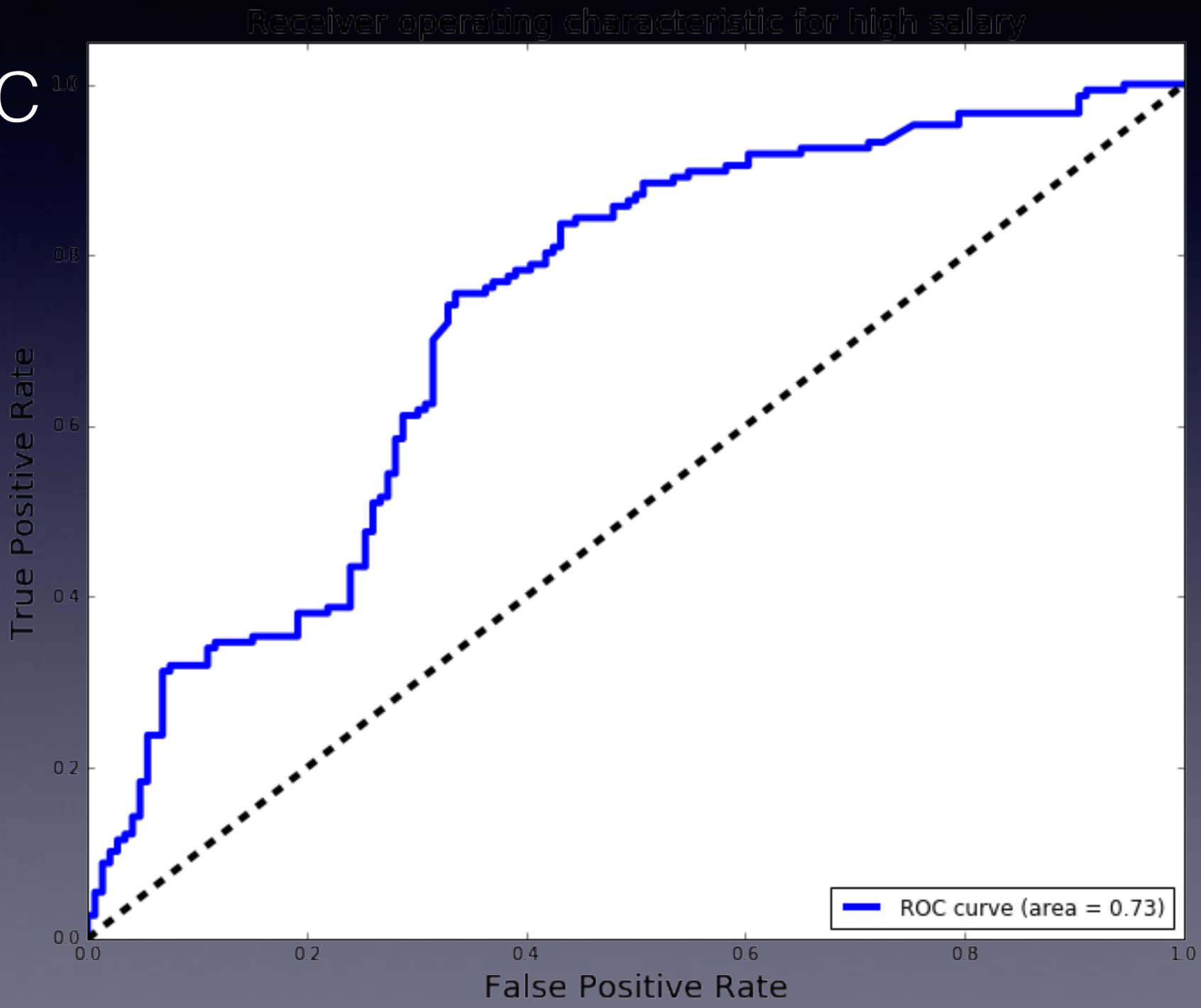
# Models

- Classification report

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.72 | 0.68 | 0.70 | 146 |
| 1 | 0.70 | 0.73 | 0.72 | 147 |
| avg / total | 0.71 | 0.71 | 0.71 | 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
ROC curve (area = 0.73)

# Conclusion

- Current median level of data science salaries is about $105,000
- Judging from the ROC curve, if we focus on controlling false positive rate, random forest (if control fpr<0.4) or gradient boosting (if control fpr<0.2) should be used.
- Limitation: small dataset limits model performance.