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
- Train test split
- Feature selection using Random Forest

Refine

	importance
big	0.044259
data	0.035693
company	0.034145
team	0.025159
location_num	0.022828
scientist	0.022280
high_position	0.018302
looking	0.017426
science	0.014183
lead	0.013826
research	0.013508
analyze	0.011852
join	0.011520
algorithms	0.011329
modeling	0.010705
derivative	0.010500

Models

- Random Forest with grid search
- Accuracy score on test set: 0.60

Models

- Confusion Matrix

	pred high	pred low
high	22	26
low	9	31

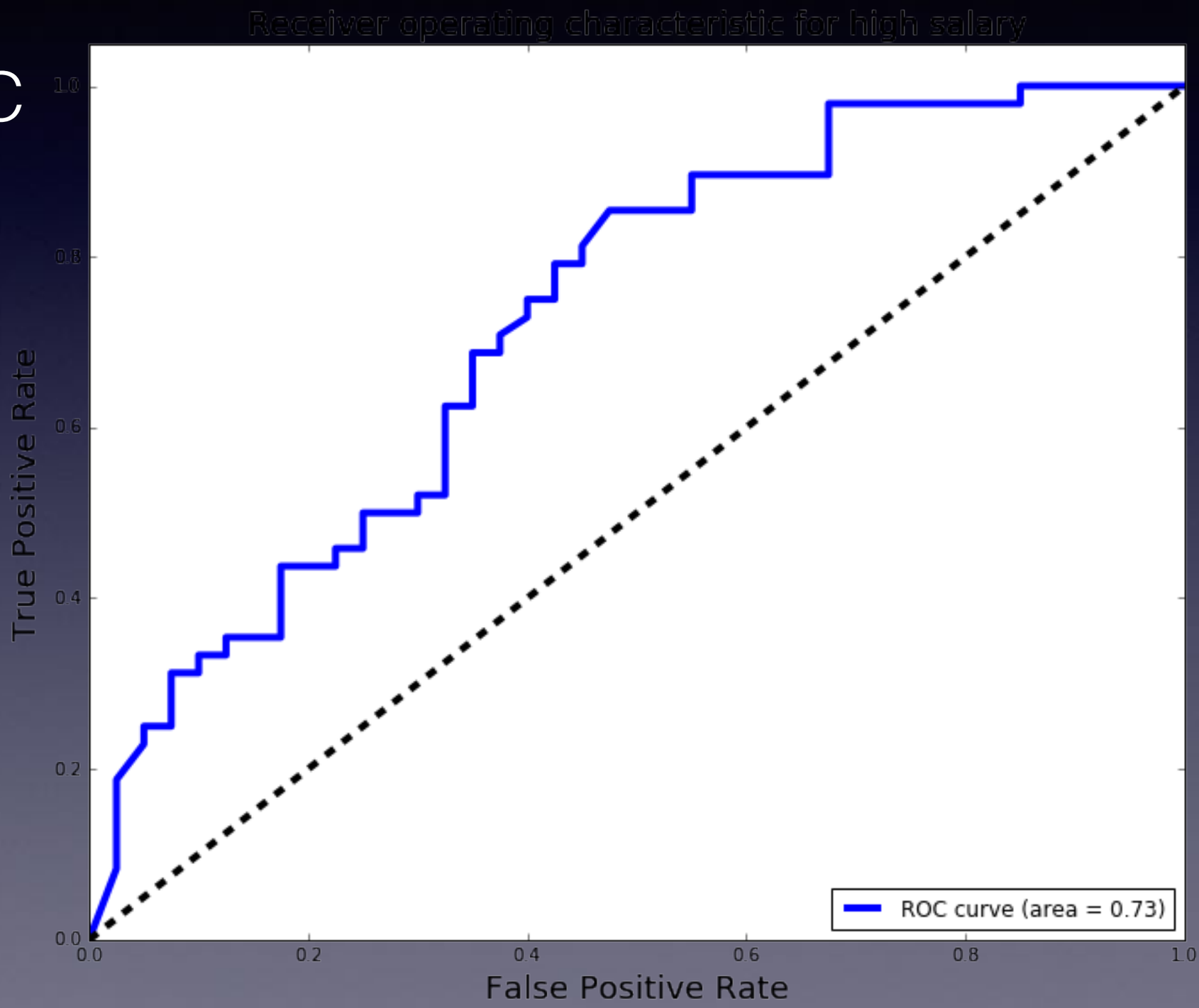
Models

- Classification report

	precision	recall	f1-score	support
0	0.54	0.78	0.64	40
1	0.71	0.46	0.56	48
avg / total	0.63	0.60	0.59	88

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)

Models

- Gradient Boosting with grid search
- Accuracy score on test set: 0.67

Models

- Confusion Matrix

	pred high	pred low
high	32	16
low	13	27

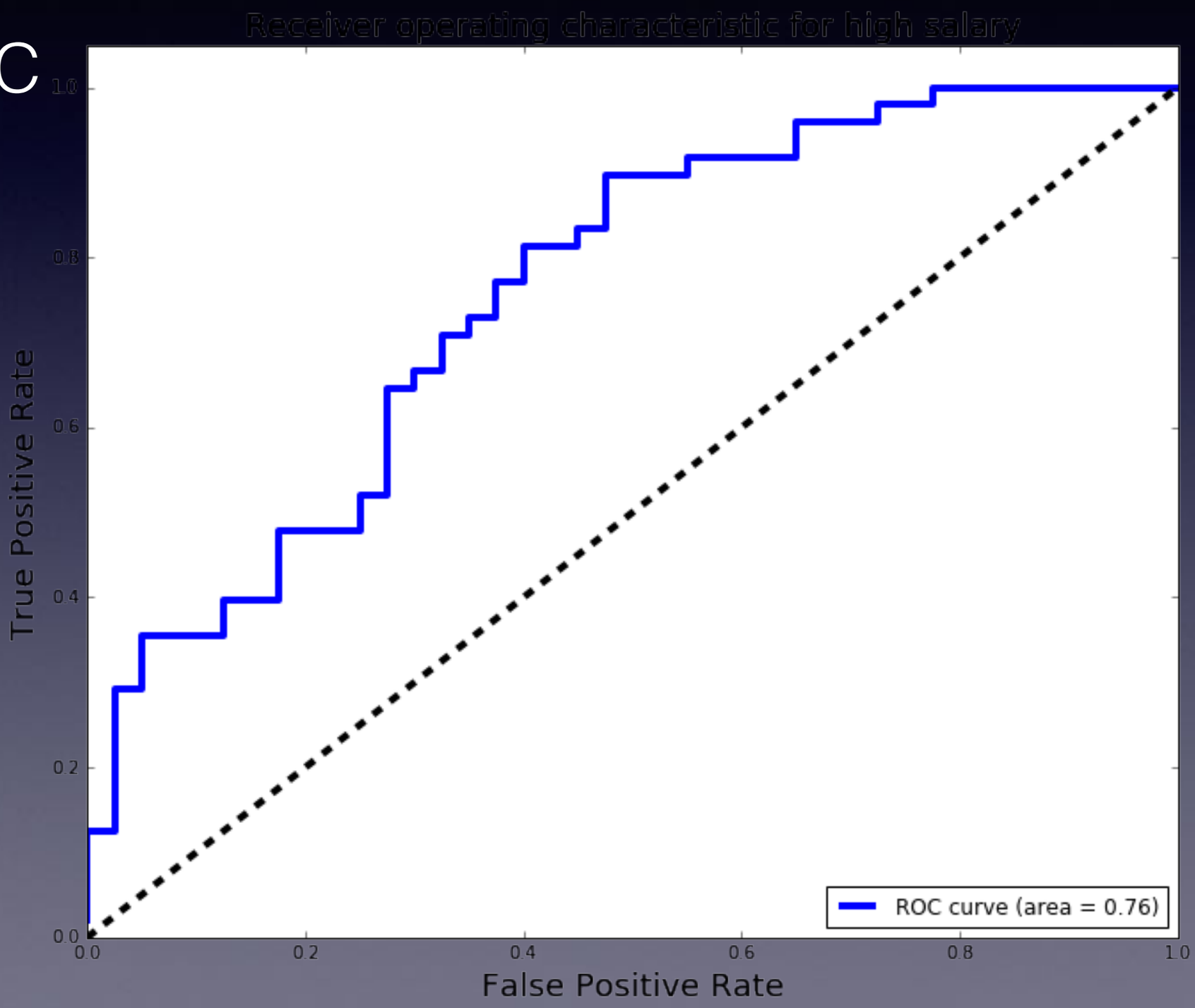
Models

- Classification report

	precision	recall	f1-score	support
0	0.63	0.68	0.65	40
1	0.71	0.67	0.69	48
avg / total	0.67	0.67	0.67	88

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

- SVM with linear kernel
- Accuracy score on test set: 0.66

Models

- Confusion Matrix

	pred high	pred low
high	27	21
low	9	31

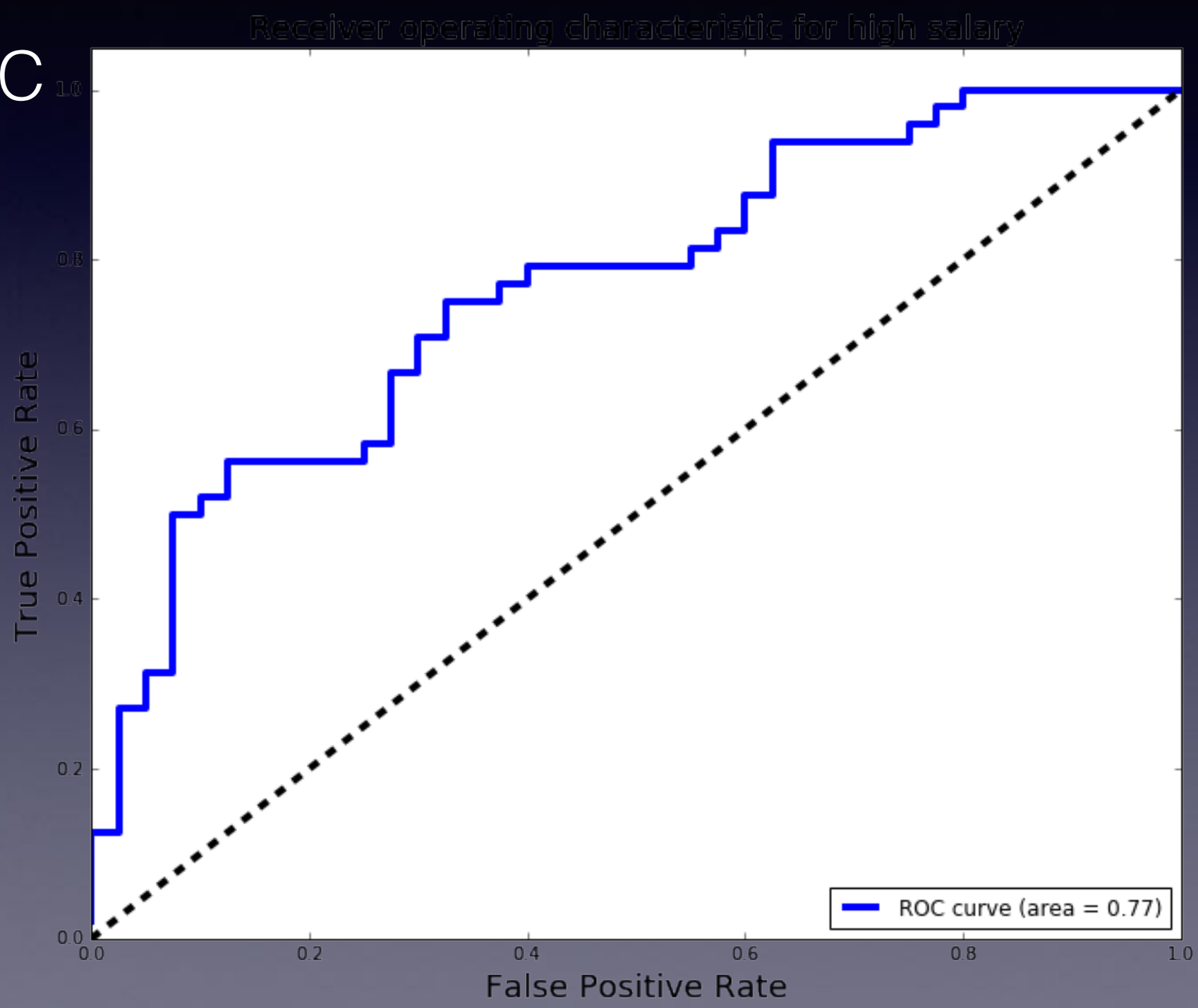
Models

- Classification report

	precision	recall	f1-score	support
0	0.60	0.78	0.67	40
1	0.75	0.56	0.64	48
avg / total	0.68	0.66	0.66	88

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 = 0.77)

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 0.3, which gives us a true positive rate around 0.7

- Limitation: small dataset limits model performance.