Online Dating: Who Wants Children?

October 31, 2018

Spencer Tollefson

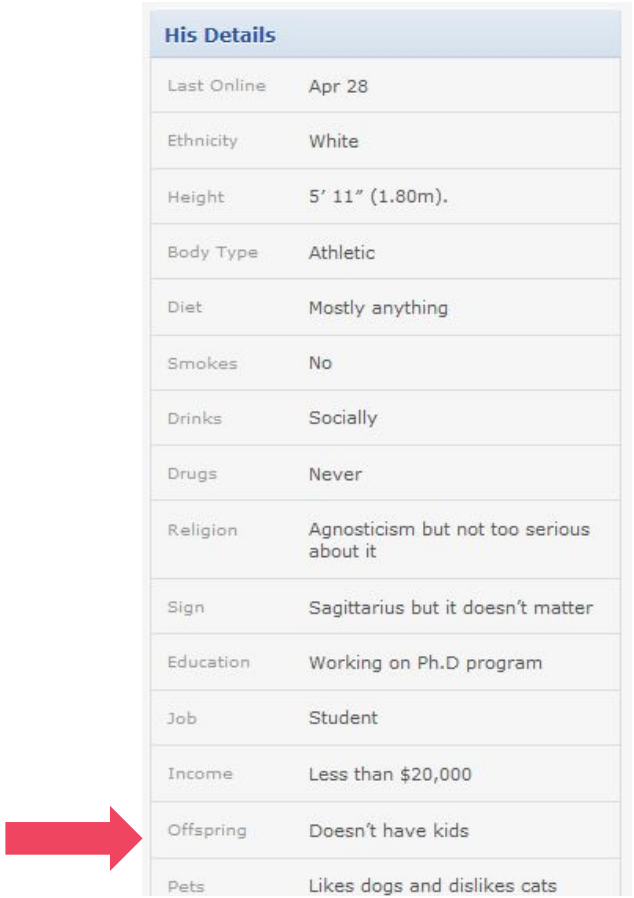
Introduction

- Problem
  - Many online profiles do not reveal children preference
- Objective
  - Create model to predict child-having preferences
- Goal
  - Provide prediction information to interested parties

Data

- OKCupid.com profile details
  - 59,000 unique profiles
  - San Francisco, June 2011

- Filters for project:
  - Relationship status: “Available”
  - Desire children?: “Yes/No”



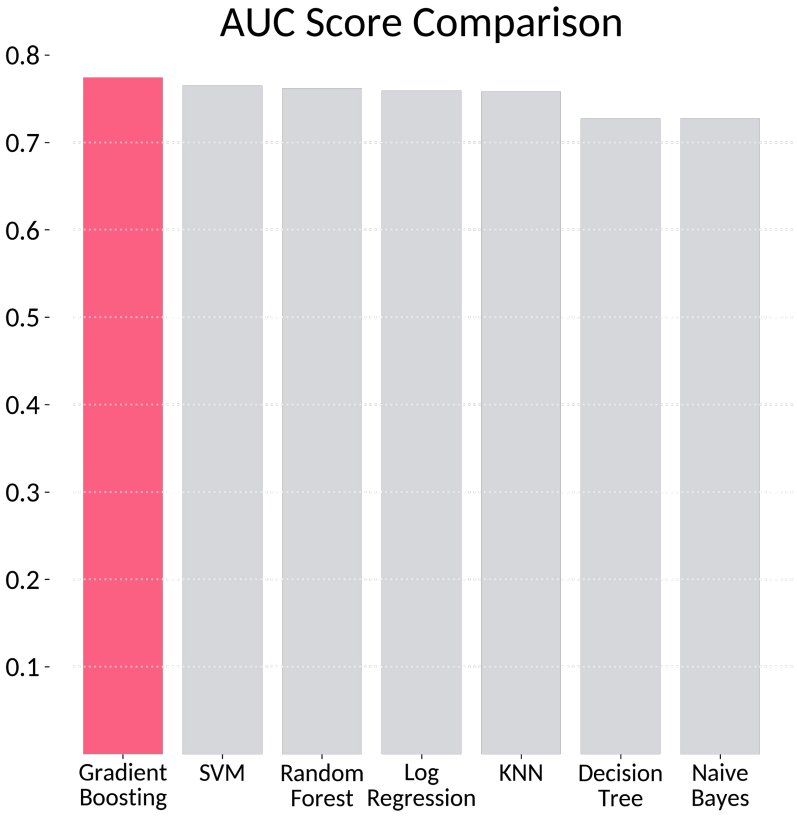
His Details	
Last Online	Apr 28
Ethnicity	White
Height	5’ 11” (1.80m).
Body Type	Athletic
Diet	Mostly anything
Smokes	No
Drinks	Socially
Drugs	Never
Religion	Agnosticism but not too serious about it
Sign	Sagittarius but it doesn’t matter
Education	Working on Ph.D program
Job	Student
Income	Less than $20,000
Offspring	Doesn’t have kids
Pets	Likes dogs and dislikes cats

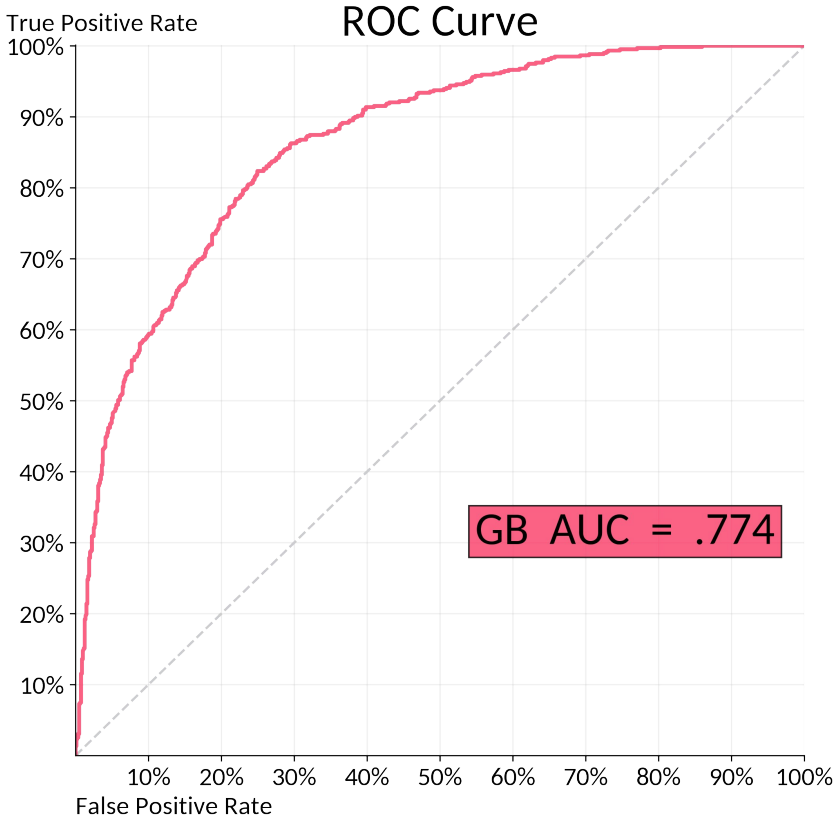
Methodology

- Measure by ROC AUC
  - Binary classifier: 40/60 split
  - Equally valued outcomes
  - SMOTE upsampling

- Classifier Models
  - K Nearest Neighbors
  - Support Vector Machine
  - Gradient Boosting
  - Decision Trees
  - Random Forest
  - Naive Bayes

Results


AUC Score Comparison
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
Gradient Boosting
SVM
Random Forest
Log Regression
KNN
Decision Tree
Naive Bayes


ROC Curve
True Positive Rate
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
False Positive Rate
GB AUC = .774

Conclusions

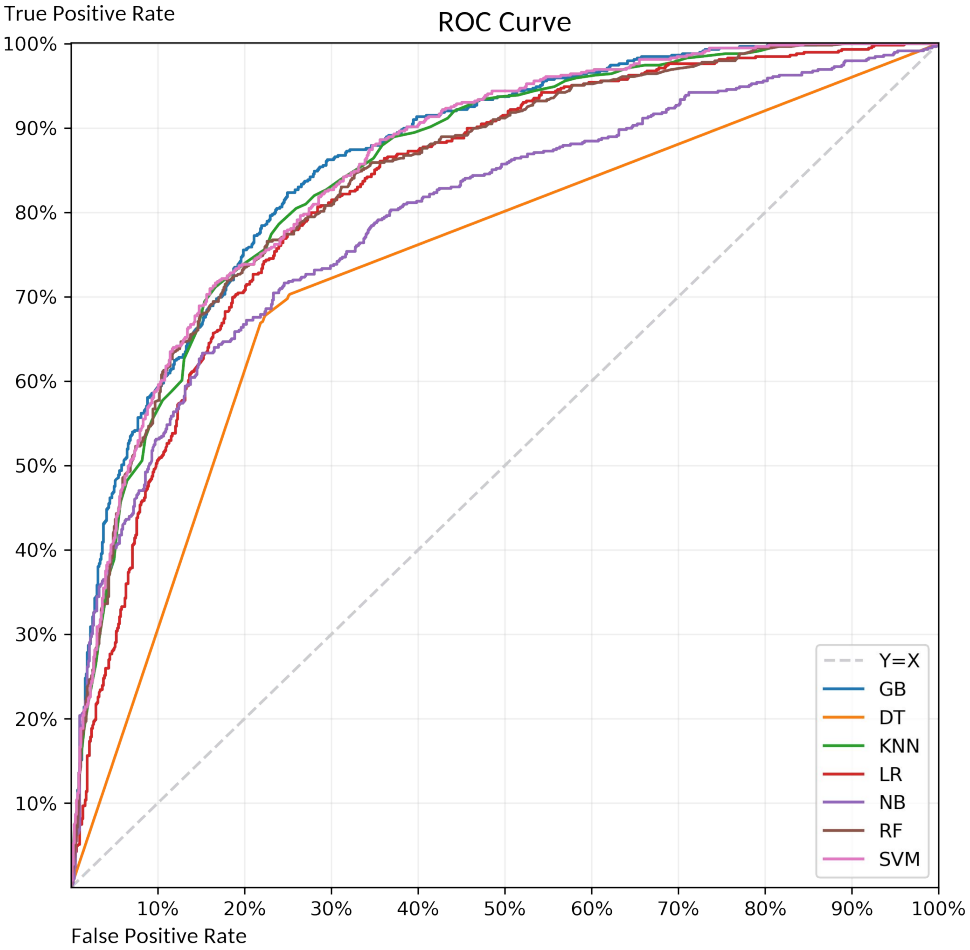
- Value-add: provides real predictive info
- Potential uses:
  - Stand alone web app
  - Online dating sites
  - Health insurance companies
  - Corporate HR

Future Work

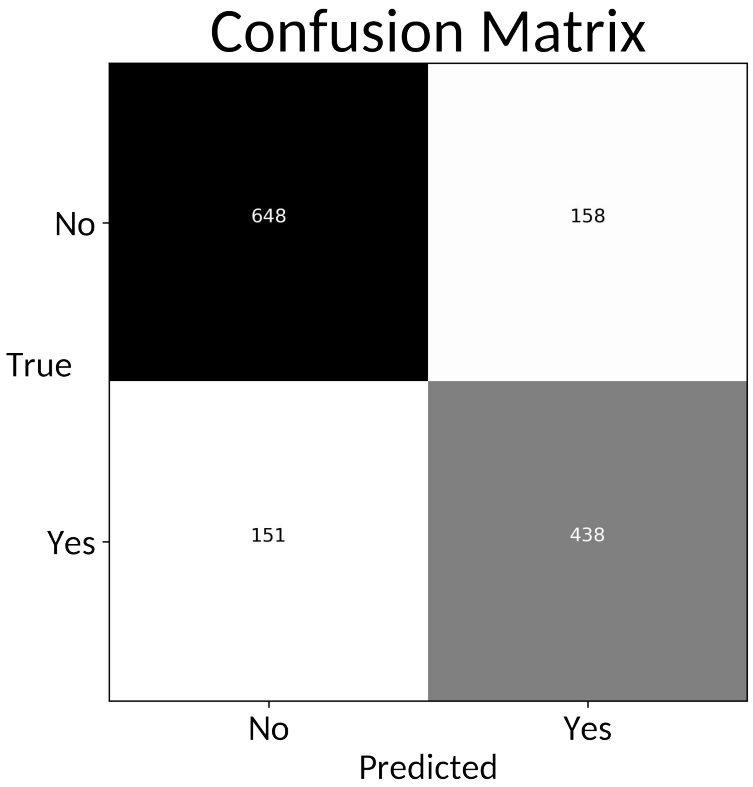
- Compare model to different populations
  - Other dating platforms
  - Locations
  - Time Periods
- Expand classifier from binary to multi-class

Thank you

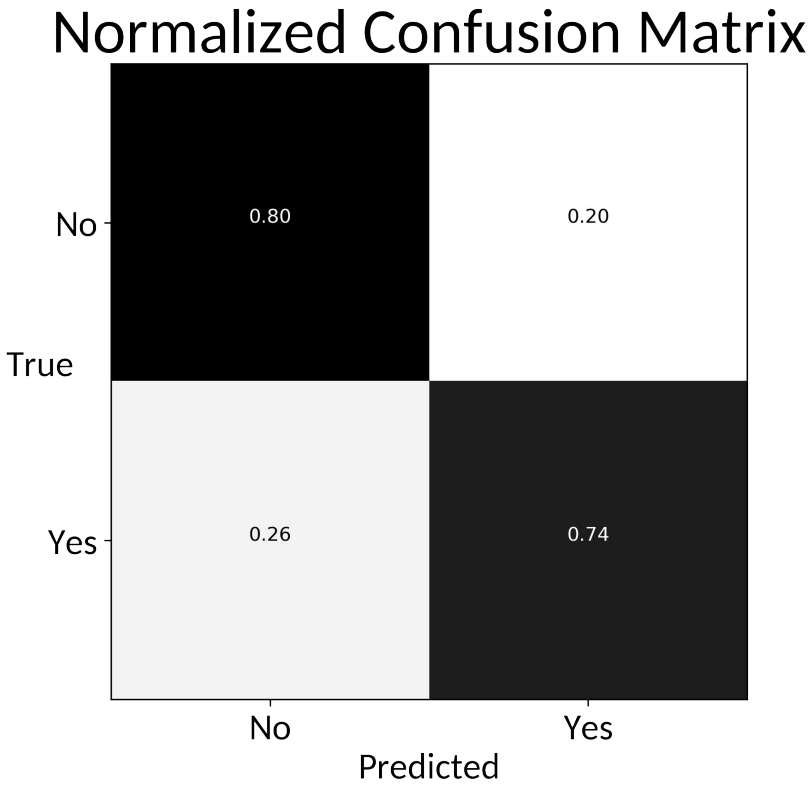
Appendix: ROC Curve



Appendix: Confusion Matrix



Appendix: Confusion Matrix (Normalized)



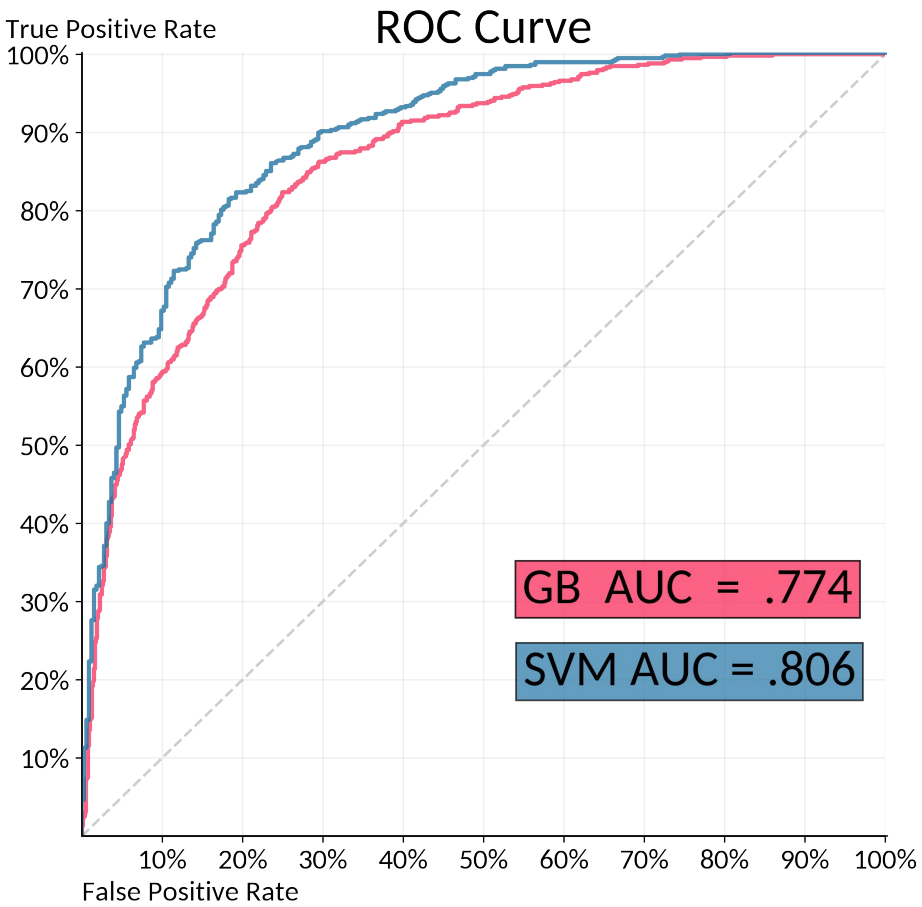
Appendix: SK-Learn Classification Report

Note: these tests done at P > 0.5 threshold

```
Accuracy on holdout set: 0.778494623656
```

	precision	recall	f1-score	support
0	0.81	0.80	0.81	806
1	0.73	0.74	0.74	589
micro avg	0.78	0.78	0.78	1395
macro avg	0.77	0.77	0.77	1395
weighted avg	0.78	0.78	0.78	1395

Appendix: ROC Curve with “current children”



Appendix: Feature Correlation