Escaping the Ivory Tower

Ravi Charan
//Flatiron School (NYC Data Science)

Dec 13, 2019

Predicting Career Intent

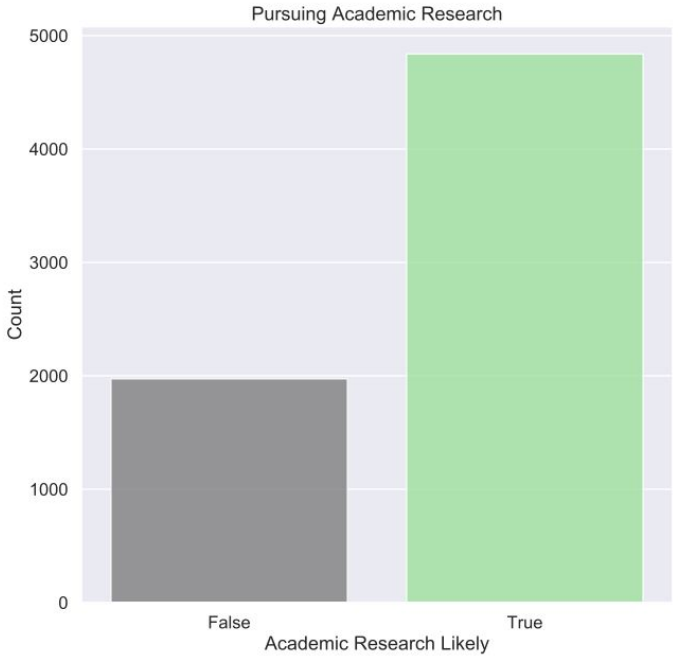
- 6,812 PhD students worldwide took a survey distributed by Nature, the top academic journal (alongside Science)
- Topics covered include:
    - Experiences in the program (academic, personal, financial, and workplace issues)
    - Career plans
- We used the data to try to predict whether students say they are unlikely to pursue academic research

Who Cares?

- Employers now use the data they collect to predict employee career moves and treat people accordingly (sometimes positively, sometimes negatively)
    - This has been going on at a personal in professional environments since there were any but computers bring an impersonal angel
- How much information leaks from unrelated questions?

Data

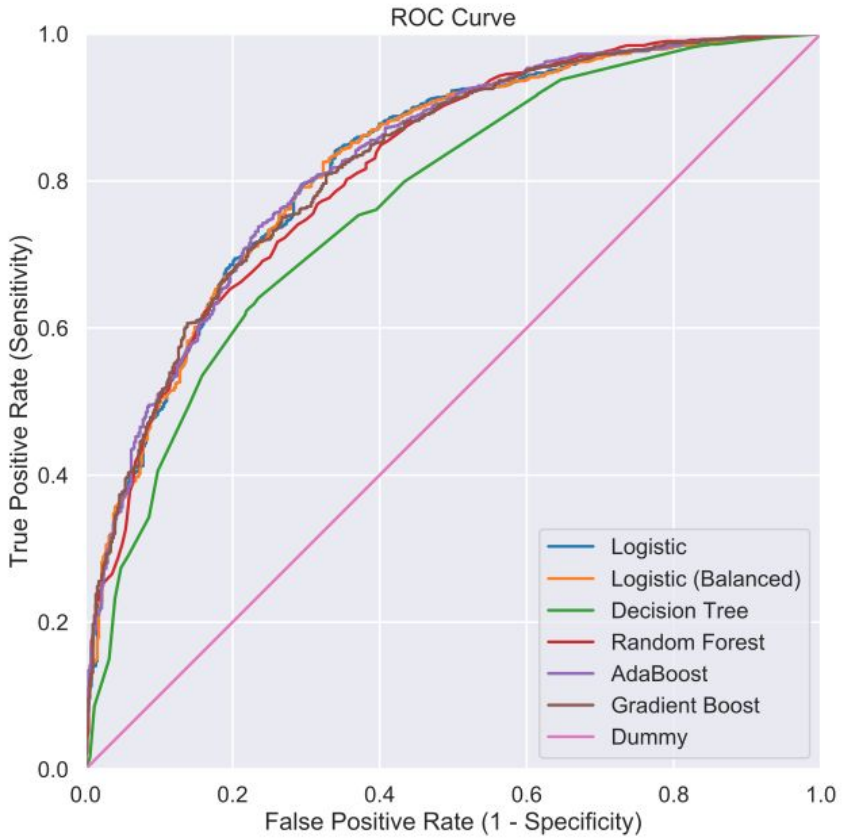
- 6812 Rows
- 65 Questions + subparts (not all asked to all)
    - Geography
    - Motivation
    - Satisfaction and Engagement
    - Workplace issues (bullying, harassment, discrimination)
    - Mental health
    - Career plans
    - Demographics, family, other jobs
- Categorical, Binary, and Ordinal (1–5 or 7; or rankings)
    - Median imputation (ordinal 1–5 or 7)
    - Bottom imputation for rankings
    - Unanswered category for categorical
- Dependent Variable: Pursuing Academic Research
    - “Please use the scale below to indicate how likely you are to pursue one of these career paths upon completion of your programme: **Research in Academia**”
    - 1–5 scale; 4 or 5 coded as “pursuing academic research”



Baseline 71%

Modelling (1/2)

- Top AUCs were 81–83%
- Sensitivity: Percent of those pursuing academia identified as such
- Specificity: Percent of those leaving identified as such



Modelling (2/2)

Preferred Metrics

Model	Precision	Recall	F1	AUC	Cross-Entropy (Train)	Cross-Entropy (Test)	Fit Time (min)	Time per Fold* (sec)
Baseline	71%	100%	83%	50%			0	
Logistic	83%	90%	86%	82%	1.8	1.9	30	10.15
Logistic Weighted	87%	76%	81%	82%	2.0	2.2	6	7.75
Decision Tree	78%	92%	84%	77%	1.9	2.5	12	0.05
Random Forest	80%	94%	87%	82%	0.6	2.0	14	1.17
AdaBoost	82%	90%	86%	83%	2.9	3.0	33	0.68
Gradient Boost	82%	91%	86%	82%	1.6	2.0	101	0.63

Hardware: 2.7 GHz Intel Core i5 (dual core)

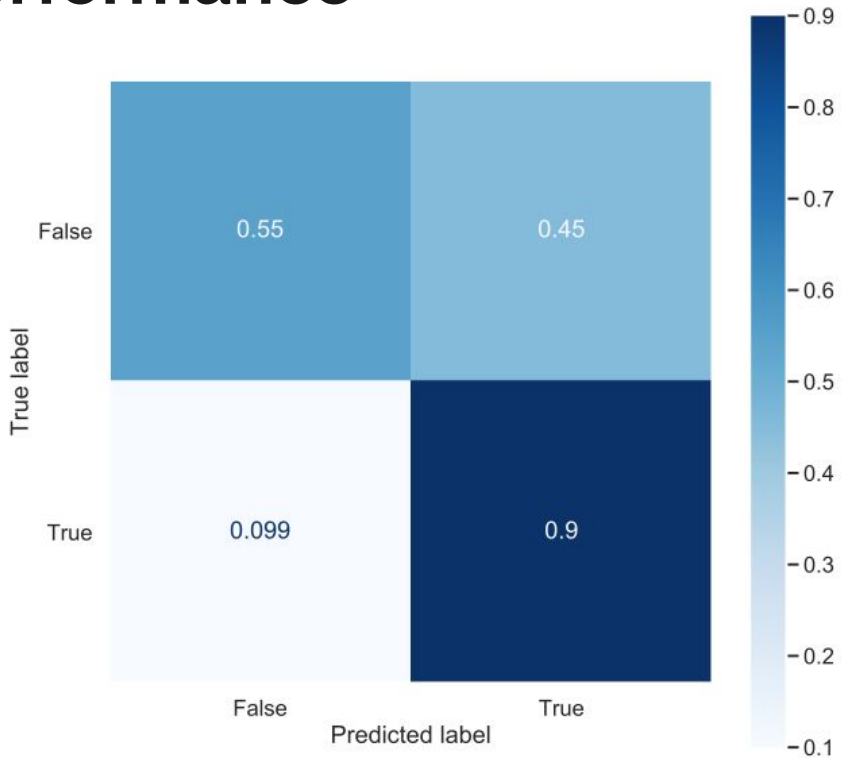
- 5-fold cross-validation on training data (80/20 split except:

  *Random Forest not cross-validated (OOB)
- Preferred metrics: AUC, Cross-Entropy Loss
- Cross Entropy is in decibels (not nats) and is per sample. 4.3 decibels = 1 nat
  - Nats are the default output of software packages
- Winner: **Logistic Regression**
  - Prefer for cross-entropy, interpretability

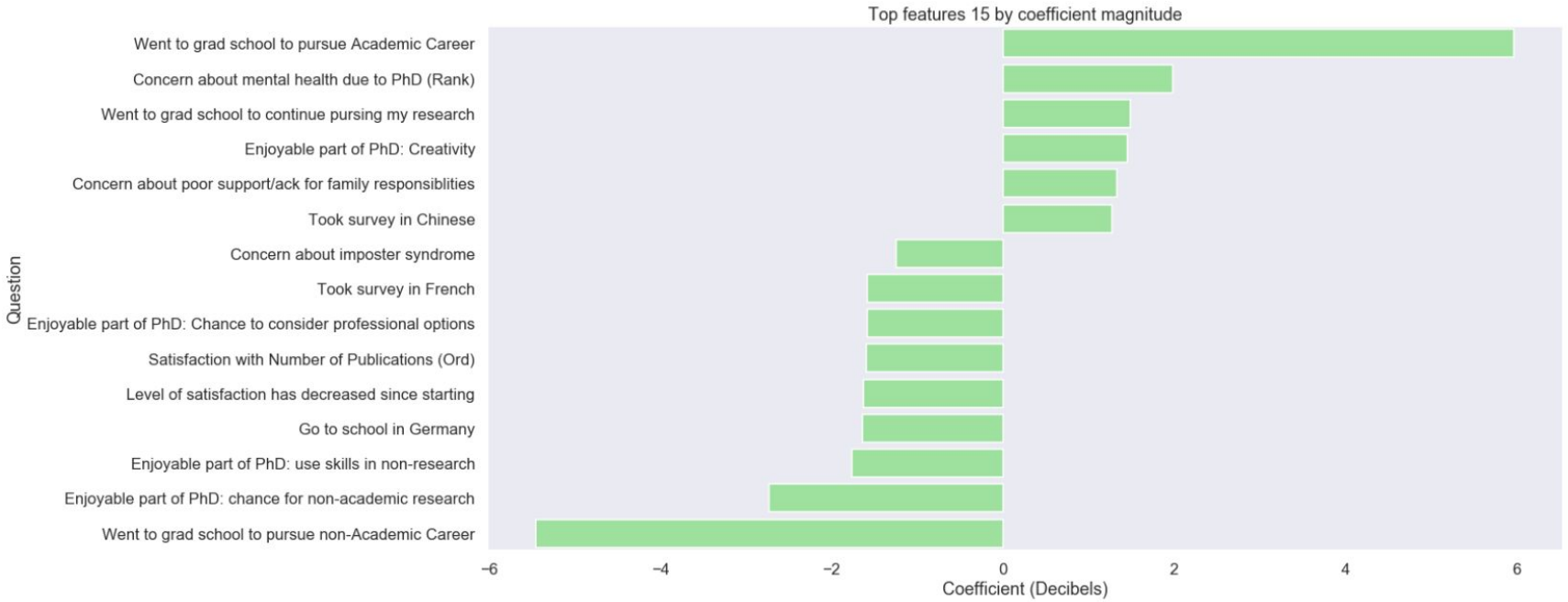
Logistic Regression: Performance

	Precision	Recall	F1	Count
True	83%	90%	86%	1205
False	70%	55%	61%	498

- Accuracy: 80%
- Class imbalance issues
  - Balanced weightings did not fix (results not shown)
  - ROC curve suggests easy tunability (AUC = 82%)
- Fitting
  - 5 fold cross-validation on training data to select hyperparameters (Grid Search)
  - ElasticNet regularization (p = 1.4)
  - Regularization strength λ = 10
  - 30 minutes at 10sec/fold



Feature Importance



- Potentially some data leakage despite removing many career-related questions.
- Some surprising results

Reflections / Future Work

- Model fitting occured over a period of about 45 chronological hours but I only spent about 3 hours in usable computation. (fits that were thrown out/redone not included)
    - Could be more efficient
- Grid Search is not magic
- Need to work on class balance issues
- Need to investigate data leakage