Detecting Fake Job Postings

Patrick Norman

Our Problem
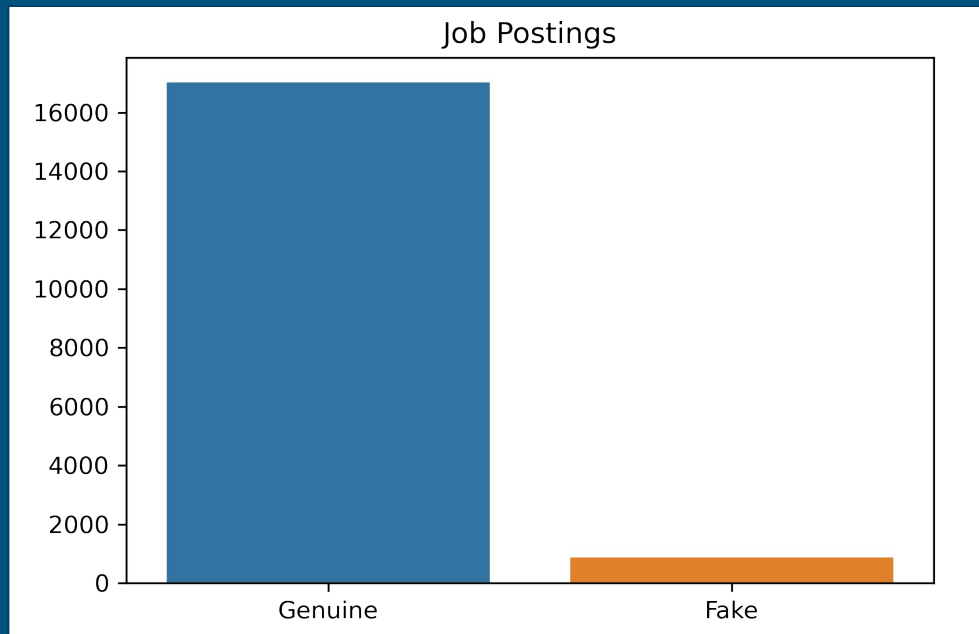
- Identity theft
- Email / resume harvesting

Model Applications
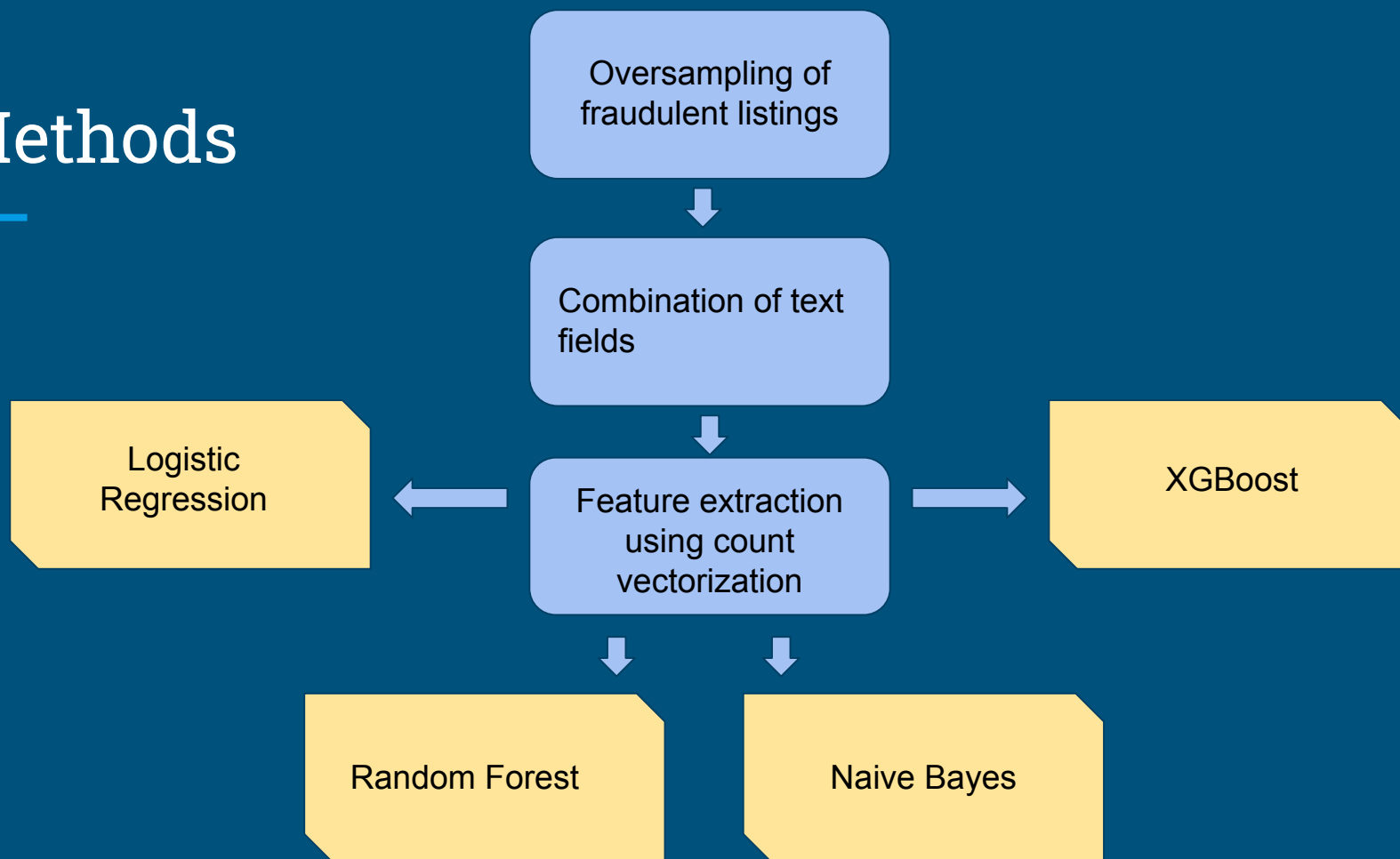
- Job sites
- Job seekers

The Data

- 17,880 job listings
- Unbalanced (4.8% fraudulent)
- Features include...
  - Job title
  - Job description
  - Company description
  - Telecommuting status
  - Presence vs absence of company logo
- Lots of missing data!

Methods

Oversampling of fraudulent listings → Combination of text fields → Feature extraction using count vectorization → Logistic Regression, XGBoost, Random Forest, Naive Bayes
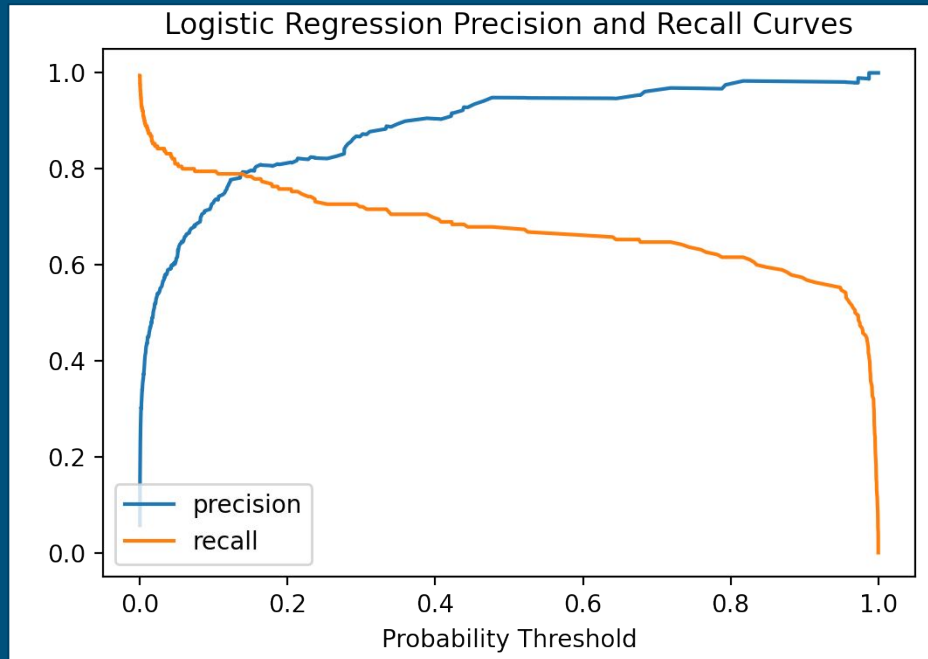
Choosing a Metric

- Model selection
- Model tweaking
- Recall or precision?

Final Model Performance

- Logistic regression
- Detected 78% of fraudulent job postings
- False positive rate was around 20%

Logistic Regression Precision and Recall Curves

Indications of Fraud

- Language as a proxy for region
- No dead giveaways
- Intentional deception prevents obvious clues

Possible Improvements

- Deeper language analysis
- More features
  - Company history on the site
  - User reviews of the company

Conclusions

- 78% of fraudulent listings detected
- 20% false positive rate
- Valuable applications for job listing sites

Appendix

https://www.thebalance.com/how-to-borrow-from-social-security-interest-free-2894595

https://www.roberthalf.com/blog/job-market/10-best-job-search-websites

https://www.traveltradecaribbean.com/its-official-more-cyber-related-criminal-activity-in-the-caribbean/