How to Get Your Questions Answered Quickly

•••

Paul Lim

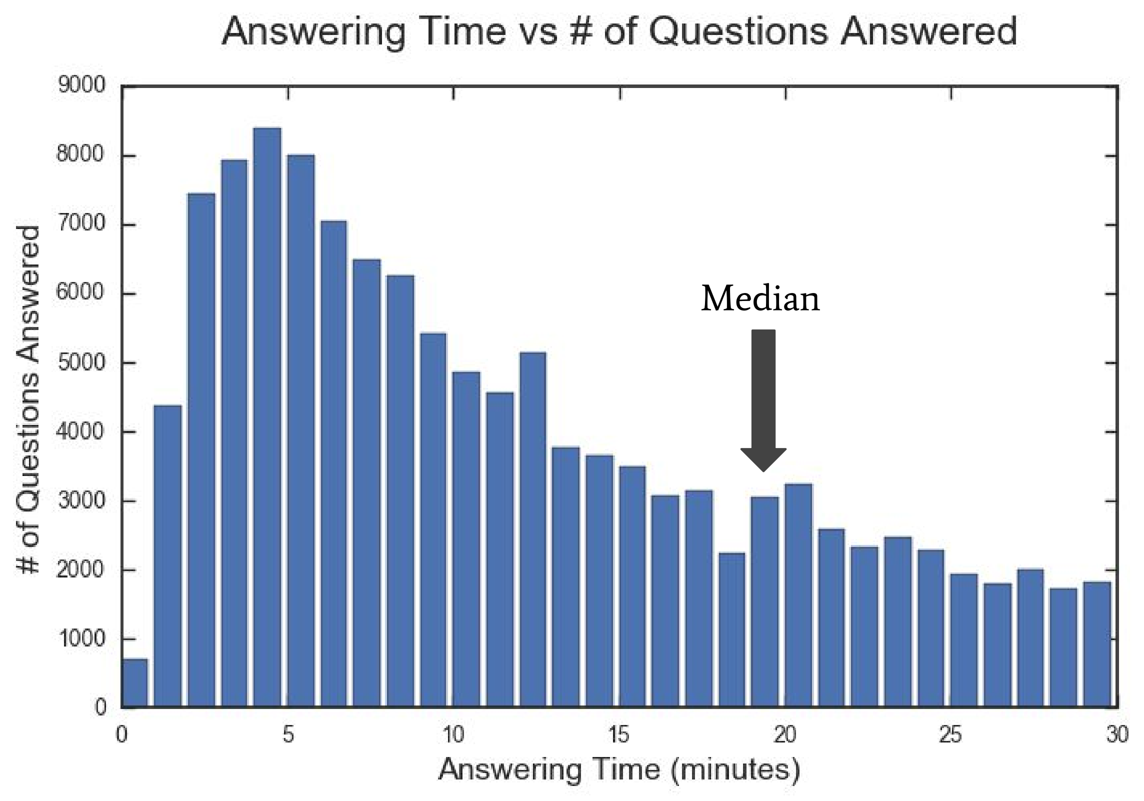
05/17/2017

Objective

We all have questions...

1. What features are important in getting answers?
2. Optimize the complexity of models and prediction time of new observations.

Target Selection

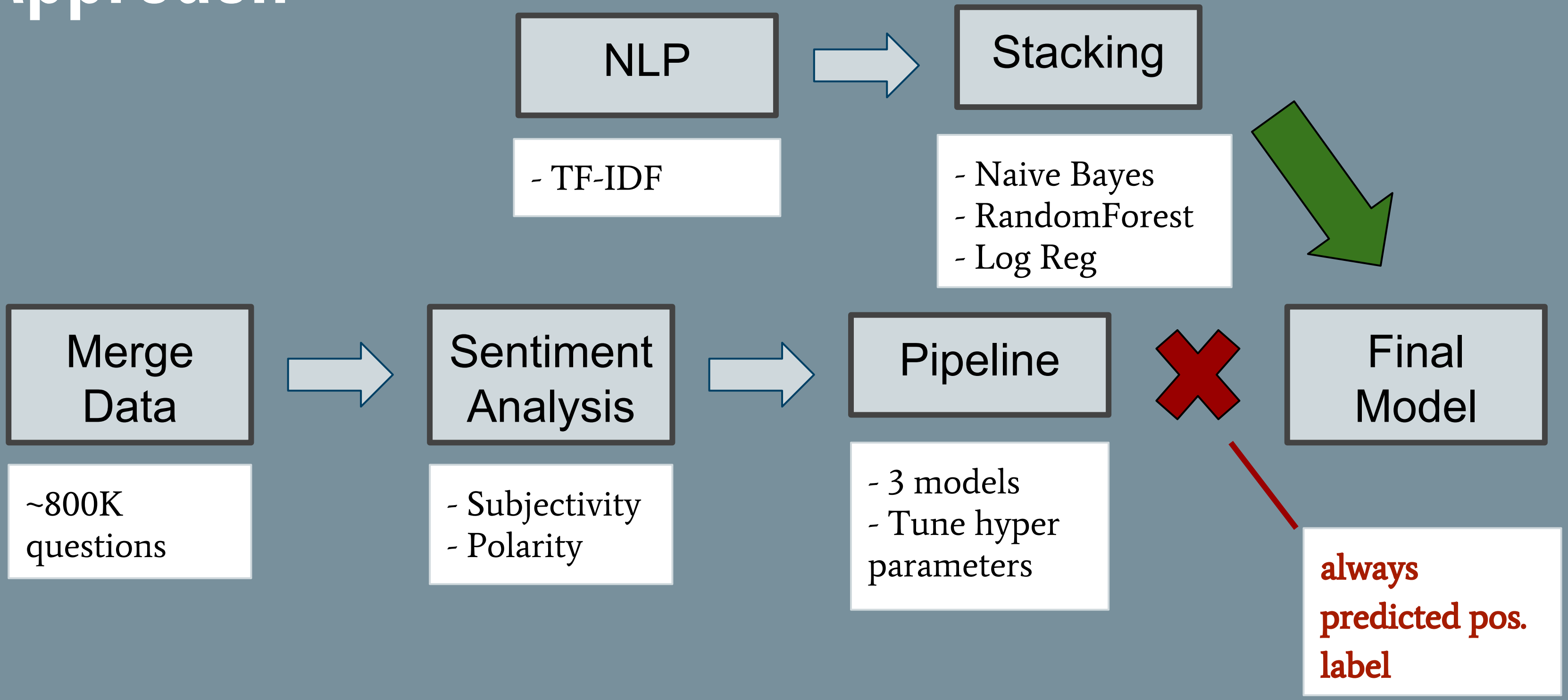


- Data limited to questions answered within 24 hours

Label:	**Criteria:**
Quick Answer	< 30 minutes
Slow Answer	>= 30 minutes

Modeling

Approach



NLP

- TF-IDF

Stacking

- Naive Bayes
- RandomForest
- Log Reg

Merge Data

~800K questions

Sentiment Analysis

- Subjectivity
- Polarity

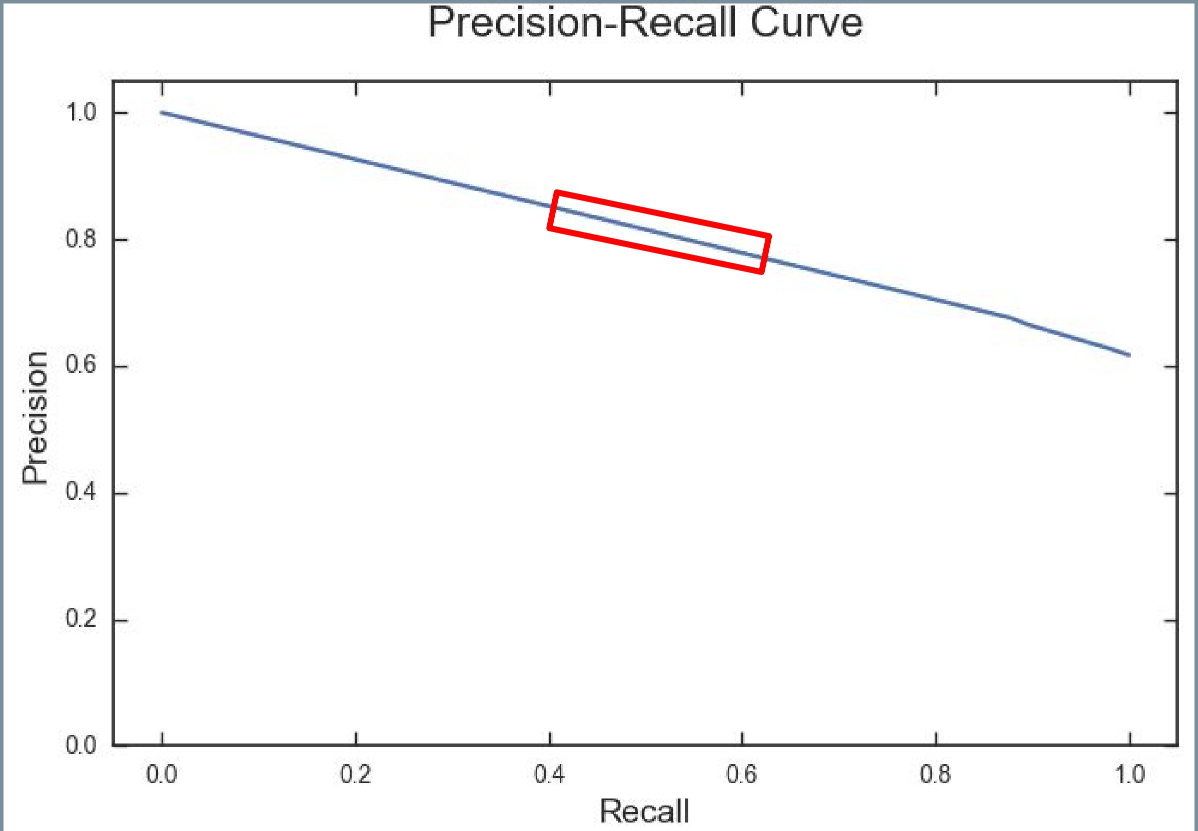
Pipeline

- 3 models
- Tune hyper parameters

Final Model

always predicted pos. label

NLP → Multinomial NB → Stacking



- Find the balance between precision and recall

- No more guessing only the positive label

- Limited to ~10,000 obs.

Scores

- FBeta with a beta of 0.5 places a higher weight on precision.

- **Stacking**:

Multinomial NB + Random Forest → Logistic Regression

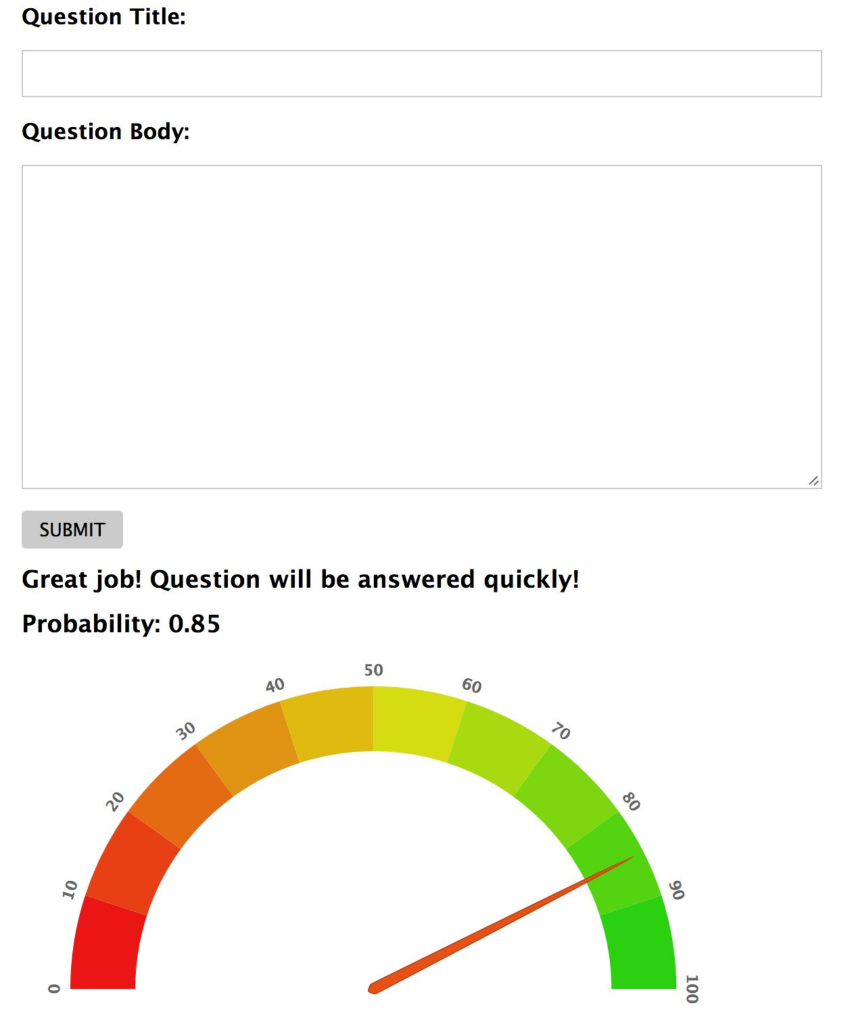
Model:	**FBeta:**
Multinomial NB	0.68
Random Forest	0.70
Logistic Regression	0.487
Stacking	0.70

Fastest model

Small gains for higher complexity

Visualization

Flask App + D3



Question Title:

Question Body:

Great job! Question will be answered quickly!

Probability: 0.85

Conclusions

Takeaways

- According to the model, ~70% FBeta is possible.
- When something goes wrong.. try again!

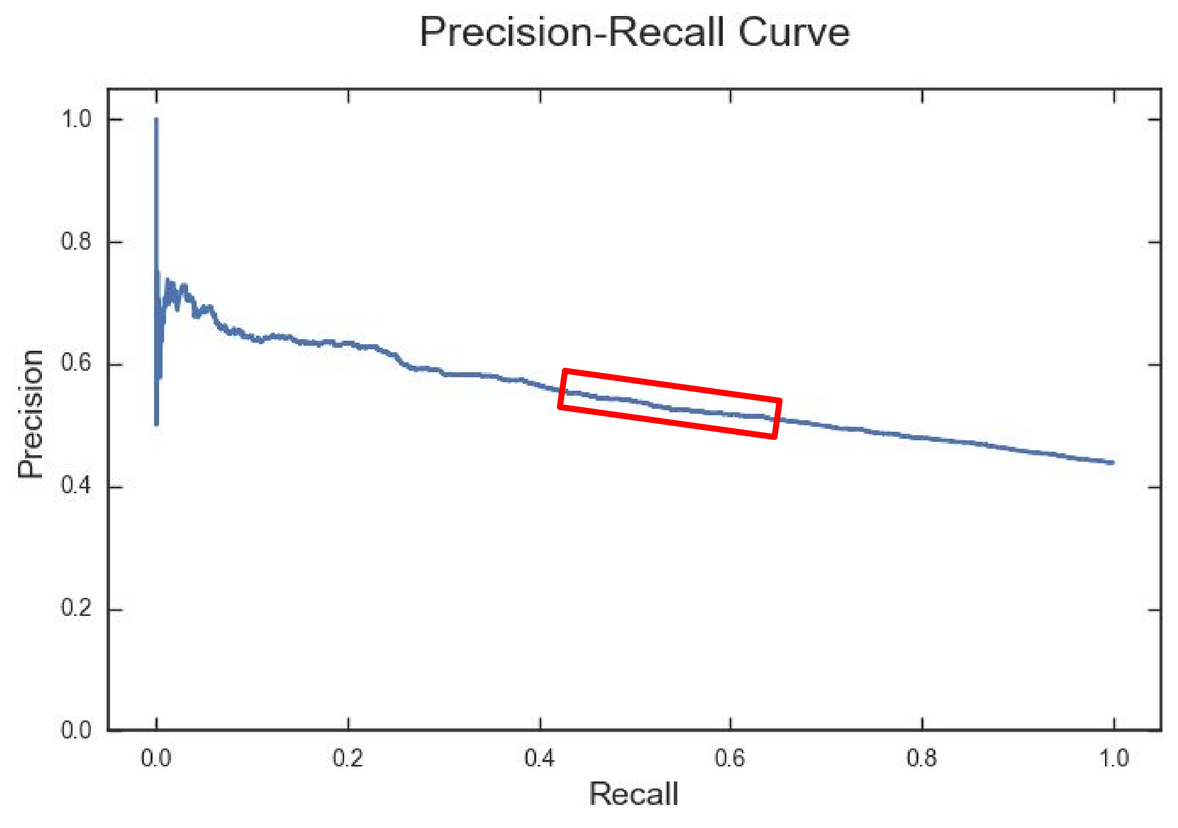
Future Works

Next steps

- Different combinations of **stacking** or *boosting* for better scores
- Find a way to use all of the available data rather than a subset of it.

Appendix

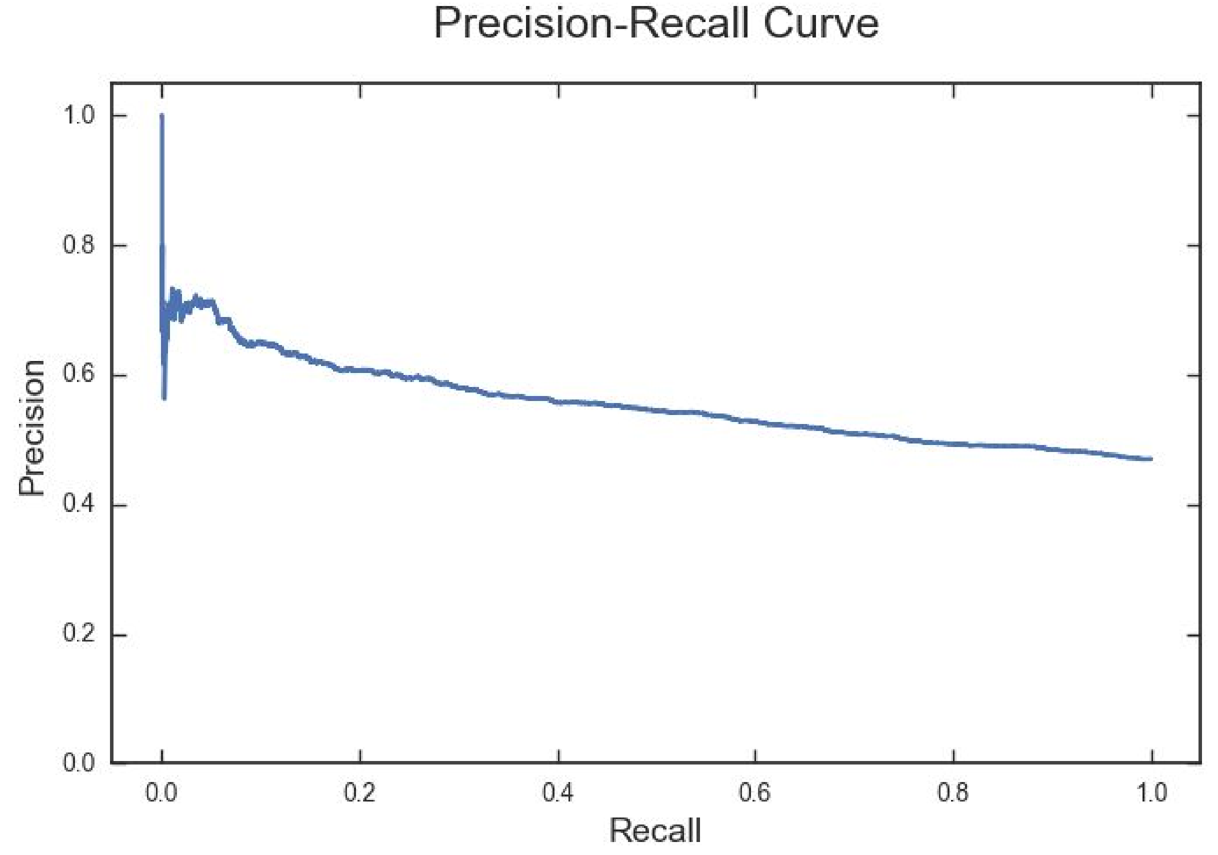
What went wrong?



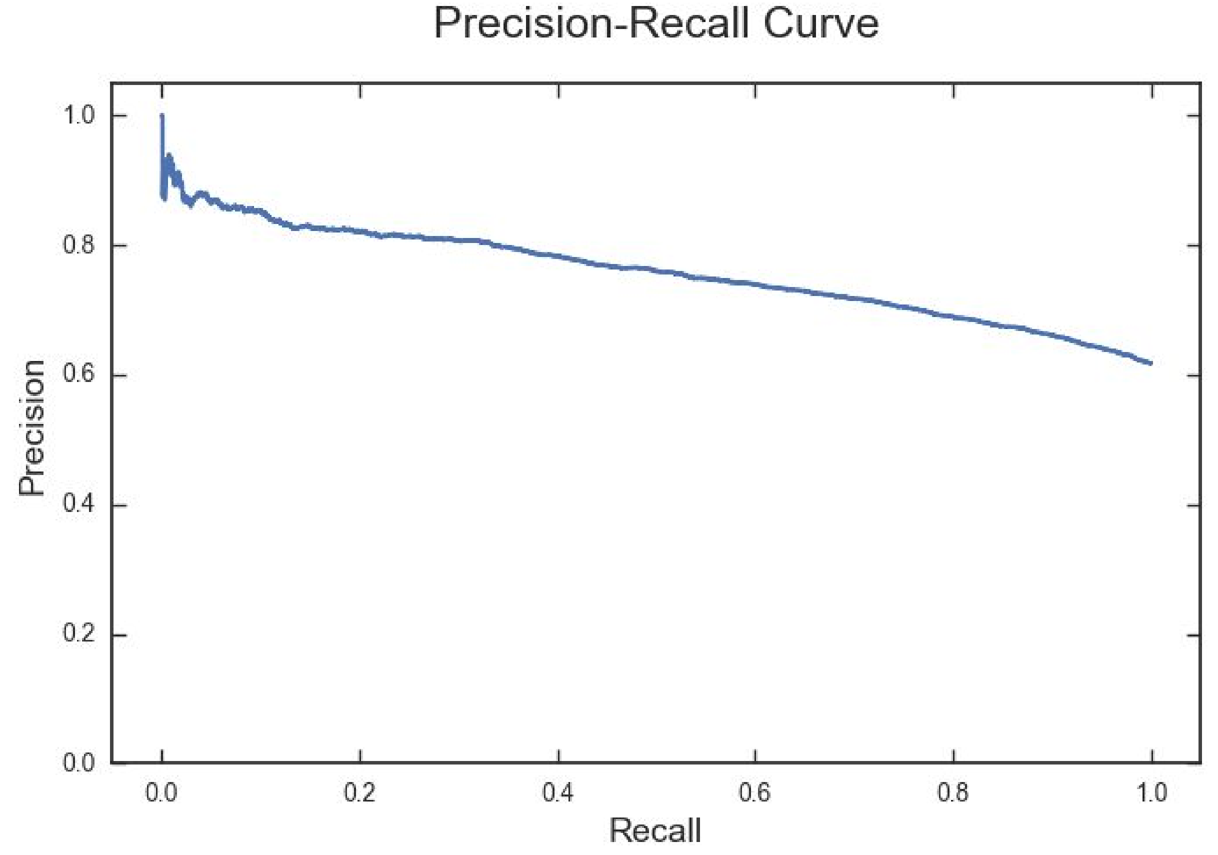
- Small precision increases

- Logistic Regression model was always predicting the positive label

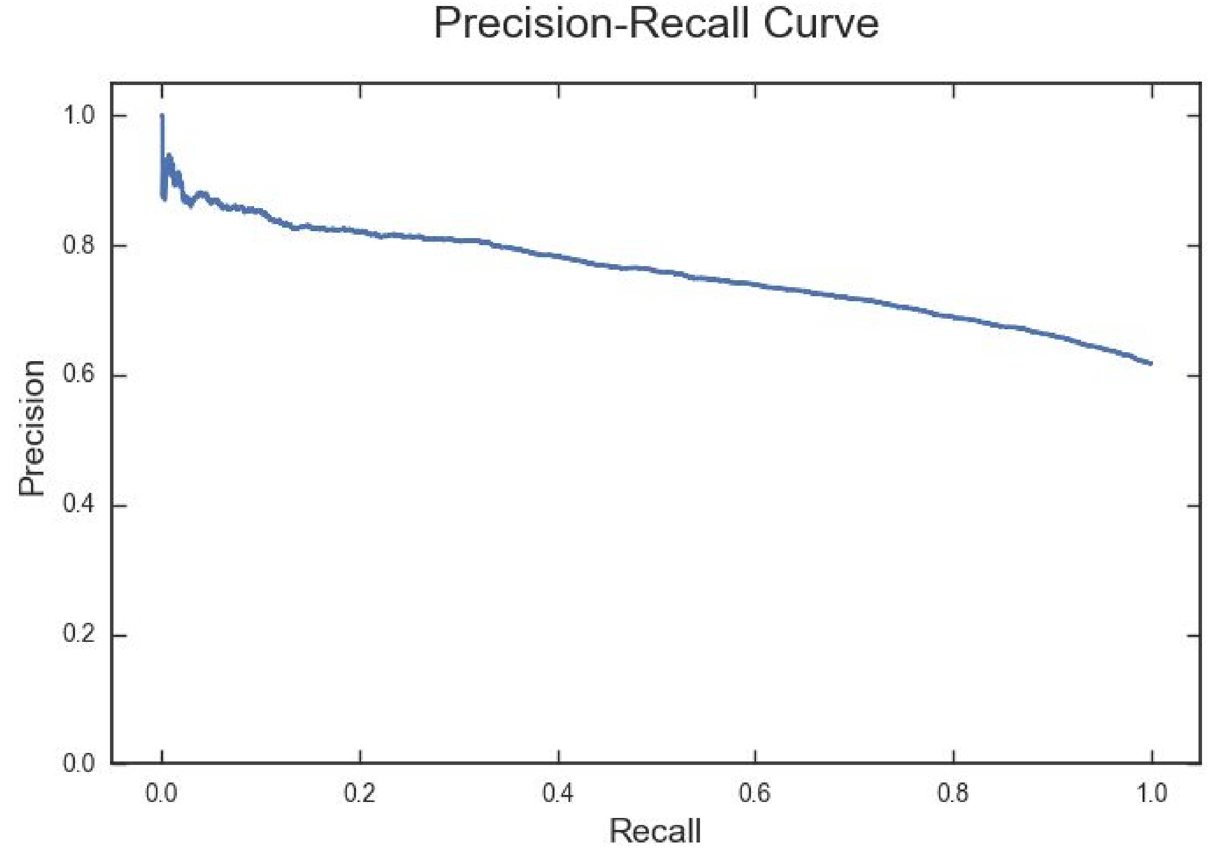
Logistic Regression



Multinomial NB



Multinomial NB


Precision-Recall Curve
Precision
Recall