

Ad-Hoc Classification of Python 2 vs 3 Questions

Katharina Huang
2017/05/17

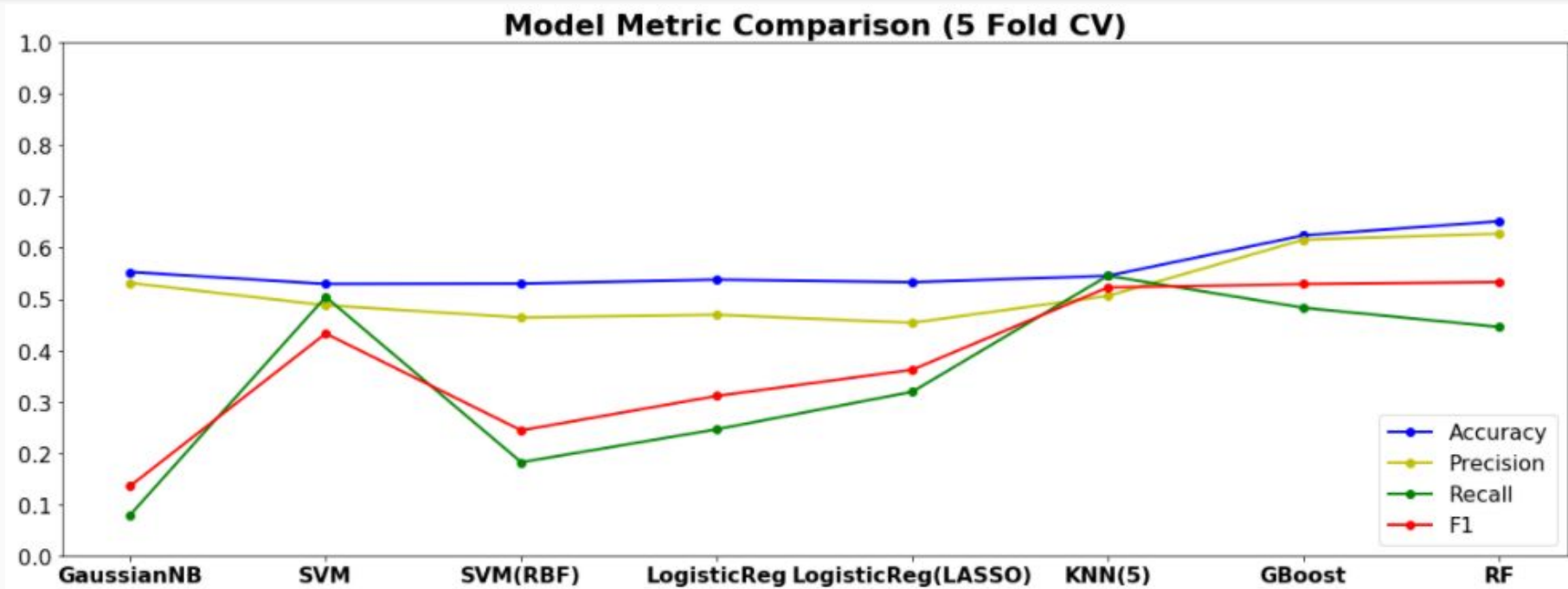
Data

- Year 2014
- Questions with at least one answer

python-2.7	python-3.x	
	No	Yes
No	67839	6809
Yes	7577	655

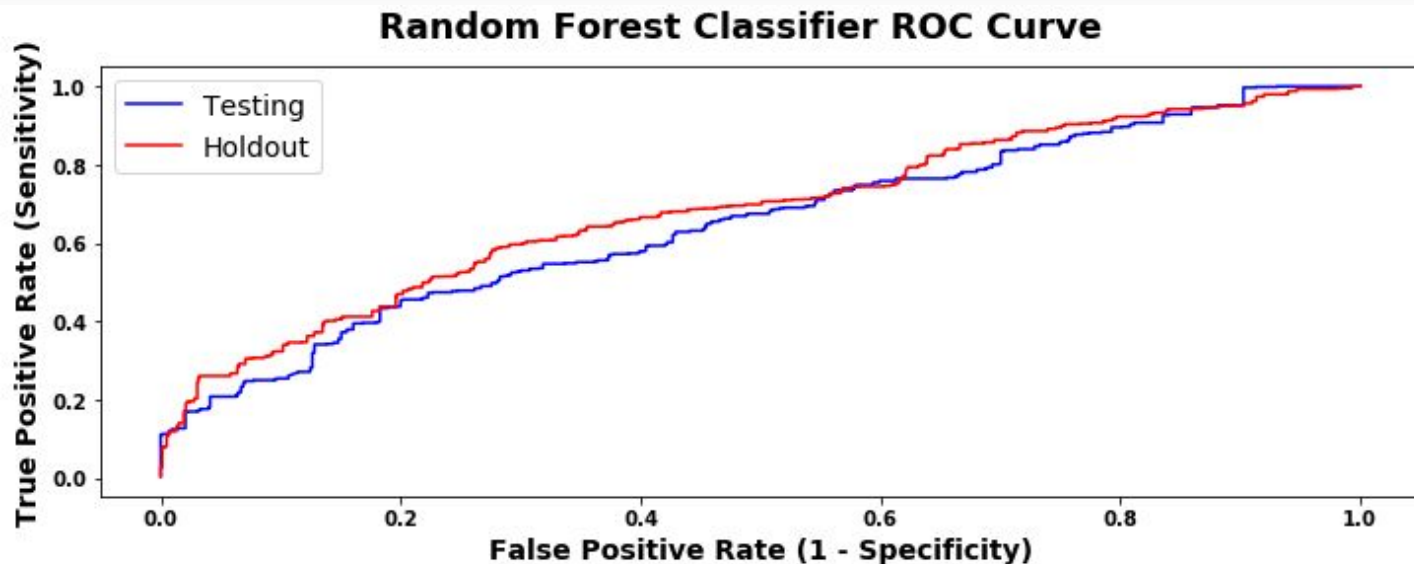
The screenshot shows a Stack Overflow question page. The question title is "How do I convert a currency string to a floating point number in Python?". The question body contains the following text: "I have some strings representing numbers with specific currency format, for example: money='\$6,150,593.22'. I want to convert this string into the number 6150593.22. What is the best way to achieve this?". The question has 11 votes and is tagged with "python" and "localization". The question was asked 5 years, 5 months ago and has been viewed 11527 times. The question was edited by Dave Mooney on Dec 7 '11 at 20:31. The question has 4 answers. The top answer is by Dave Mooney, with 1,730 votes, 6 answers, 25 comments, and 54 votes. The answer body contains the following code: "from re import sub; from decimal import Decimal; money = '\$6,150,593.22'". The question is linked to a Google code page: "code.google.com/p/python-money ? - Stephen Gross Dec 7 '11 at 20:18".

Method



Result

*Treating Python3 as the “event”



Random Forest 5 fold CV accuracy 0.6514 precision 0.6272 recall 0.4456 f1 0.5330 auc -

Random Forest testing set accuracy 0.6229 precision 0.5936 recall 0.4783 f1 0.5297 auc 0.6494

Random Forest holdout set accuracy 0.6415 precision 0.6972 recall 0.5135 f1 0.5914 auc 0.6842

Result

Holdout Set		
True Value	Prediction	
	Python 2.7	Python 3.x
Python 2.7	1848	545
Python 3.x	1189	1255

Importancy Ranking	Feature Name
1	Asker Reputation
2	Log Seconds between Question and Answer
3	Log Seconds between Answer and Present Day
4	Question Views
5	Asker Bronze Badges
..	..
38	Flask Tag
39	Beautifulsoup Tag
40	Matplotlib Tag
41	Datetime Tag
42	Python Tag

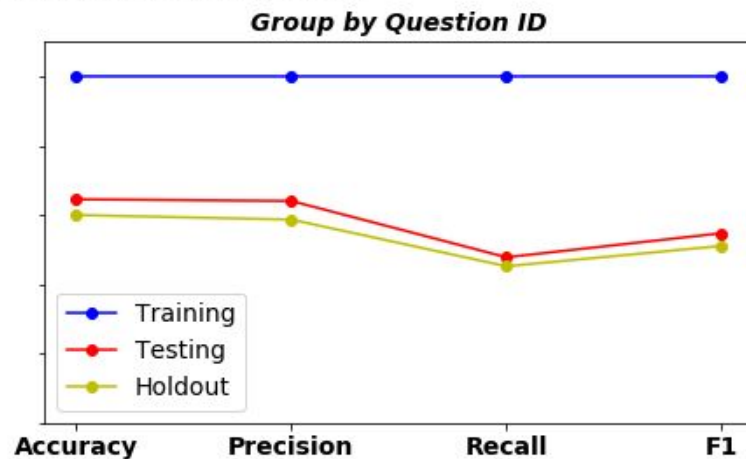
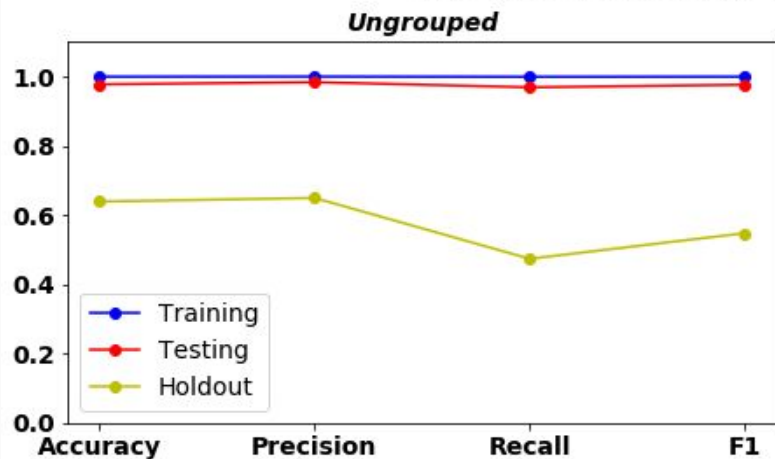
API Demo

https://youtu.be/D8_yesxONsM

Future Work

- More natural language processing (NLP)
- Grouping by question and answer aggregation

Two Different Ways to Split Train/Testing Data



Q & A

Appendix:

A better looking feature chart

Importancy Ranking	Feature
1	Asker Reputation
2	Asker Bronze Badges
3	Length of Question
4	Asker Silver Badges
5	Asker Gold Badges
6	Question Score
7	Day of Answer
8	Log Seconds between Question and Present Day
9	Question Views
10	Year of Question
..	..
39	Unit-testing Tag
40	Dataframe Tag
41	Flask Tag
42	Datetime Tag
43	Python Tag