Project 3: Reddit Classification

Chang Chu Hua

Context and Problem Statement

- We wish to create a monetizable automated homework helper
- The first step in this process will be to create a smart classifier that can classify questions into the subjects they come from.

Problem Statement:

- Can we automate the classification of math and physics questions?

Data

- Posts were scraped from Reddit's 'AskPhysics' and 'askmath' subreddits as a proxy for a homework help forum.
- Scraped data format:

```
phys.head()
```

l_by	archived	author	author_flair_background_color	author_flair_css_class	author_flair_richtext	author_flair_template_id	author_flair_text	...	title	total_awards_received	ups	url	user_reports	view_count	visited	whitelist_status	wls	author_ca
NaN	True	[deleted]	NaN	NaN	NaN	NaN	NaN	...	[META] - Yes, homework questions are OK. Here'...	0	68	https://www.reddit.com/r/AskPhysics/comments/5...	[]	NaN	False	all_ads	6	
NaN	False	GregwiseNoah	NaN	misc	[]	983e1d48-c6b2-11e4-8b2d-22000b39cdde	High school	...	How to excel in a physics undergraduate degree	0	45	https://www.reddit.com/r/AskPhysics/comments/b...	[]	NaN	False	all_ads	6	
NaN	False	Annyunatom	NaN	NaN	[]	NaN	NaN	...	Why do shorter wavelength form better images?	0	1	https://www.reddit.com/r/AskPhysics/comments/b...	[]	NaN	False	all_ads	6	
NaN	False	bl00dinyourhead	NaN	NaN	[]	NaN	NaN	...	How to solve for maximum compression in a spring?	0	4	https://www.reddit.com/r/AskPhysics/comments/b...	[]	NaN	False	all_ads	6	
NaN	False	RedeemedDeus	NaN	NaN	[]	NaN	NaN	...	Practice Exam Question	0	1	https://www.reddit.com/r/AskPhysics/comments/b...	[]	NaN	False	all_ads	6	

Cleaning

I have this question
http://imgur.com/nJAMkV9\...

Issues:

- Html links
- Punctuation
- Numerical values

```
#check original text
cleaned.text[5]
```

```
'does pressure in a sealed container rise as it ascends in altitude?the source of this discussion is talking about
inflating an inflatable stand up paddleboard at lower altitude and then driving it up to a mountain lake. these pad
dle boards have a stiff strong structure that holds their shape. they are designed to hold aprox. 15psi. if someone
was to fill the paddleboard to 15psi say at 4,000ft altitude, then drive it to a mountain lake at say 8,000ft altit
ude, will the pressure in the paddleboard change from the change in altitude or is 15psi in a container, 15psi rega
rdless of ambient pressure?'
```

```
#check cleaned text
cleaned.stoptext[5]
```

```
'pressure sealed container rise ascends altitude the source discussion talking inflating inflatable stand paddleboa
rd lower altitude driving mountain lake  paddle boards stiff strong structure holds shape  designed hold aprox   ps
i  someone fill paddleboard  psi say    ft altitude  drive mountain lake say    ft altitude  pressure paddleboard c
hange change altitude  psi container   psi regardless ambient pressure '
```

We observe that our cleaning has been performed successfully.
Next, we lemmatize our words to prevent repetition of different word variations.

Baseline Model

For our baseline model, we will utilize the reliable Logistic Regression classifier with default settings.

```
logreg=LogisticRegression()
logreg.fit(X_train,y_train)
logreg.score(X_test,y_test)
```

```
C:\Users\chang\Anaconda3\lib\site-packages\sklearn\linear_model\logistic.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.22. Specify a solver to silence this warning.
  FutureWarning)
0.893574297188755
```

We obtain a mean accuracy of 89.4% from our baseline model, which is quite impressive as a start.

```
pred=logreg.predict(X_test)
f1_score(y_test,pred)
```

```
0.8893528183716075
```

We calculate our f1 score that indicates the balance between our predicted and true positive rates.

Investigation of Variables

Variable Changed	Effect	Comments
Punctuation and Numeric	Negligible	Once stopwords are applied, further cleaning results in miniscule differences.
Models	Noticeable	All models perform reasonable well except k-Nearest Neighbours
Tf-Idf	Negligible	
Max Features	Some	With reduced dimensionality, models are still highly serviceable
Bigrams	Negligible	f1 score actually **reduced**

Summary of models

Section	Classifier	Parameter	f1 score
Baseline	Logistic Regression	default	0.889
Models	MultinomialNB	alpha=1.15	0.912
Tf-Idf	MultinomialNB	alpha=1.60	0.916
Max Features	MultinomialNB	alpha=1.35	0.865
2-gram Count Vectorization	MultinomialNB	alpha=0.15	0.887

Evaluation

Confusion matrix for the baseline model:

	predicted askmath	predicted AskPhysics
from askmath	231	19
from AskPhysics	45	203

Classification report for the baseline mode:

	askmath	AskPhysics	micro avg	macro avg	weighted avg
f1-score	0.878327	0.863830	0.871486	0.871078	0.871108
precision	0.836957	0.914414	0.871486	0.875685	0.875530
recall	0.924000	0.818548	0.871486	0.871274	0.871486
support	250.000000	248.000000	498.000000	498.000000	498.000000

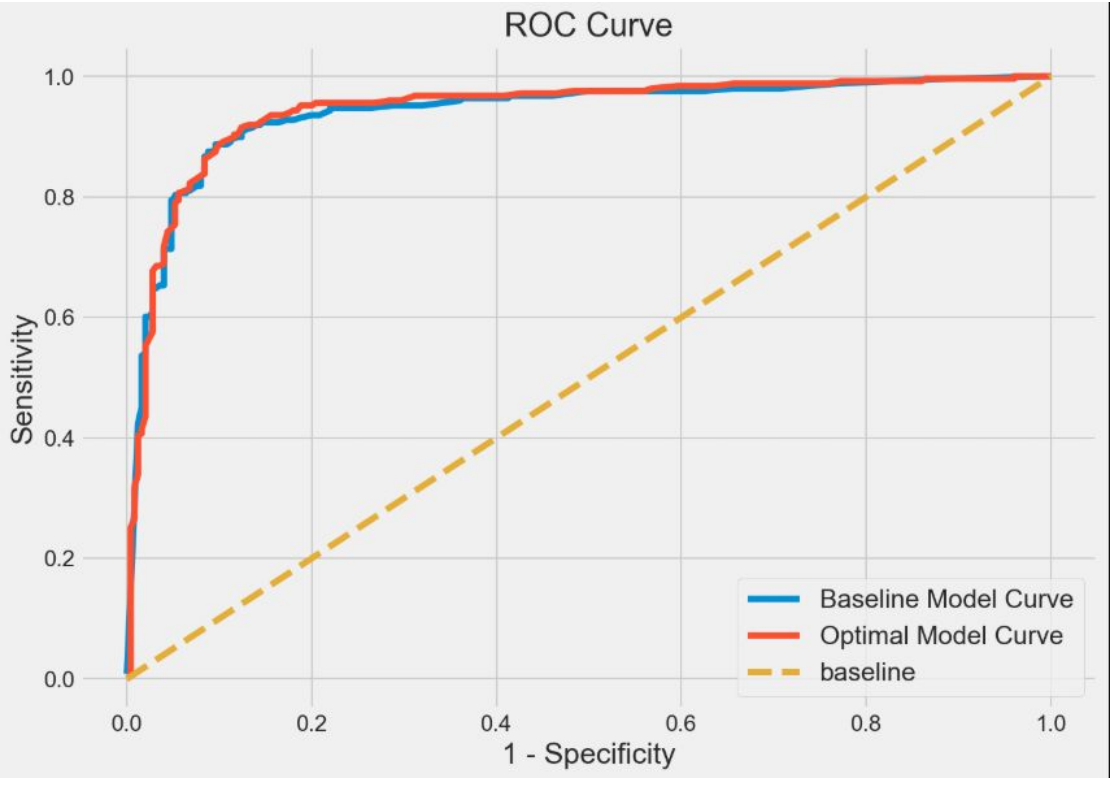
Confusion matrix for the optimal model:

	predicted askmath	predicted AskPhysics
from askmath	214	36
from AskPhysics	18	230

Classification report for the optimal mode:

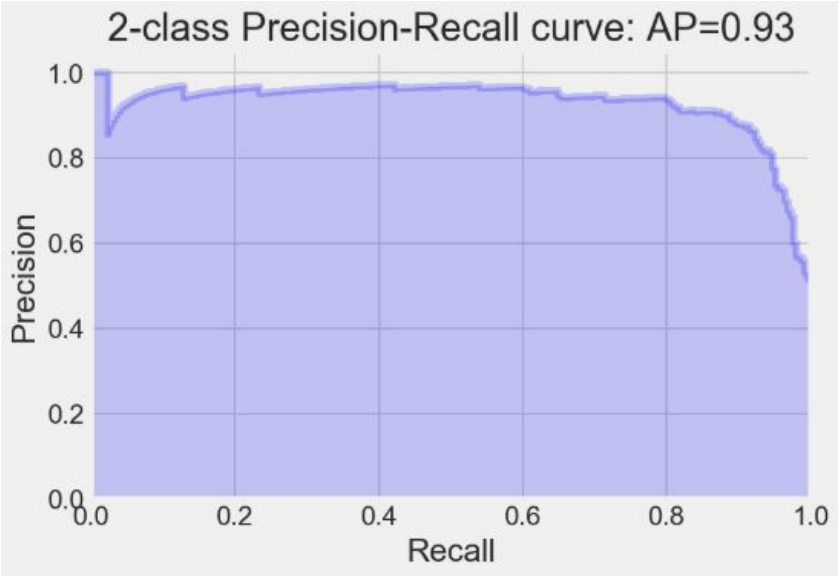
	askmath	AskPhysics	micro avg	macro avg	weighted avg
f1-score	0.887967	0.894942	0.891566	0.891454	0.891440
precision	0.922414	0.864662	0.891566	0.893538	0.893654
recall	0.856000	0.927419	0.891566	0.891710	0.891566
support	250.000000	248.000000	498.000000	498.000000	498.000000

Evaluation

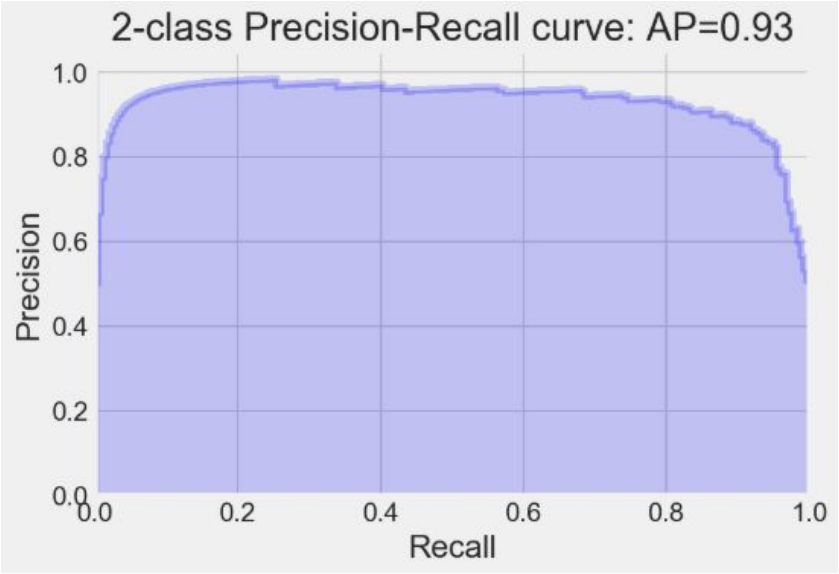

ROC Curve
Sensitivity
1 - Specificity
0.0
0.2
0.4
0.6
0.8
1.0
Baseline Model Curve
Optimal Model Curve
baseline

Evaluation

Baseline Model



Optimal Model



Conclusion

Our automated classifier is serviceable for use in our automated homework helper system.