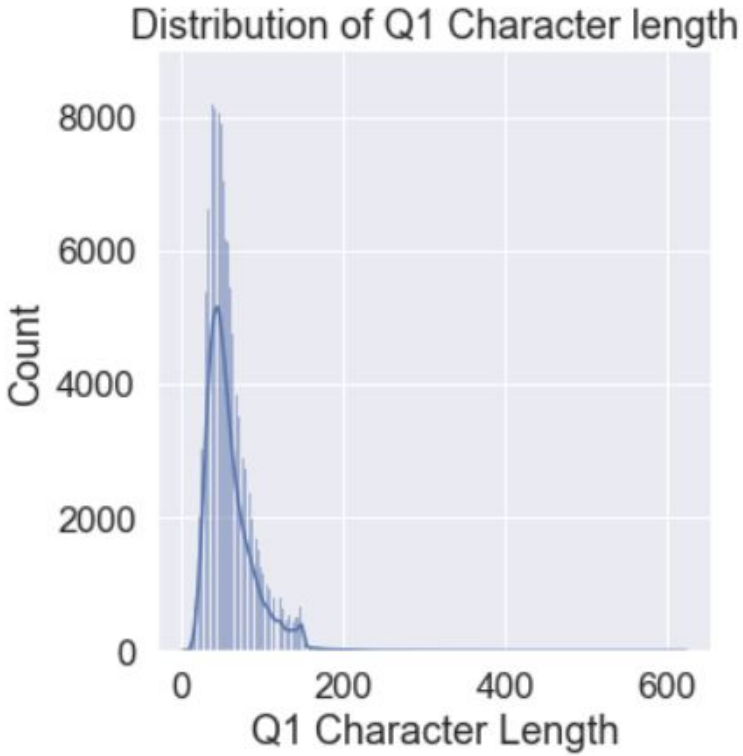
Lighthouse Labs - Mini-Project V

Quora Question Duplicates

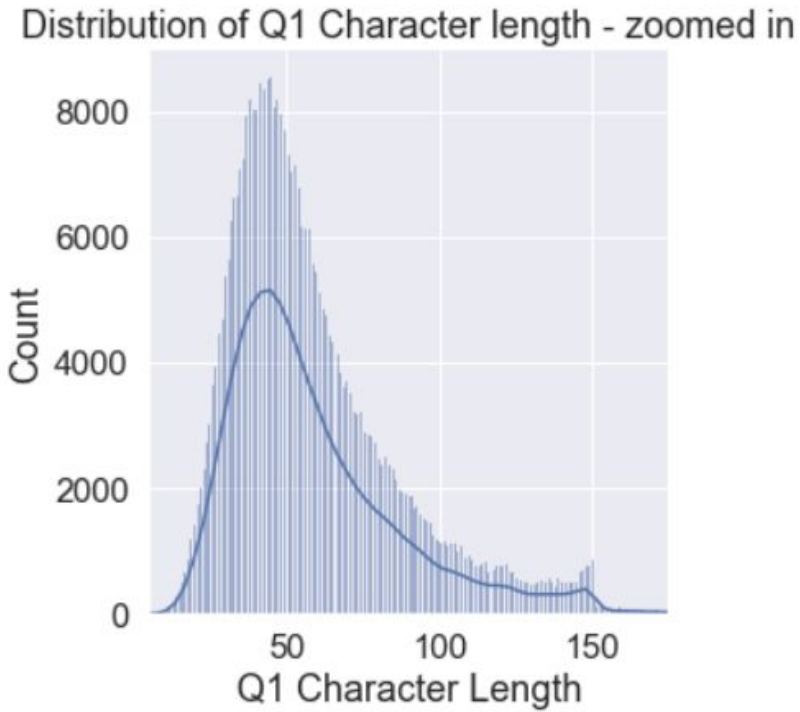
Agenda

- Dataset
- Approach
- Results
- Next Steps

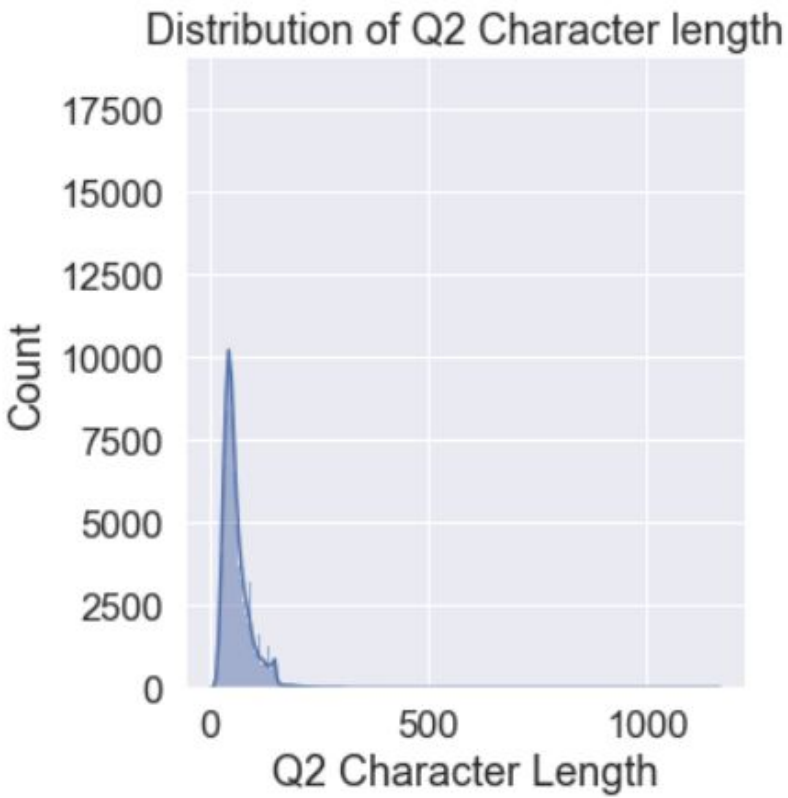
Character Count Distribution - Question 1



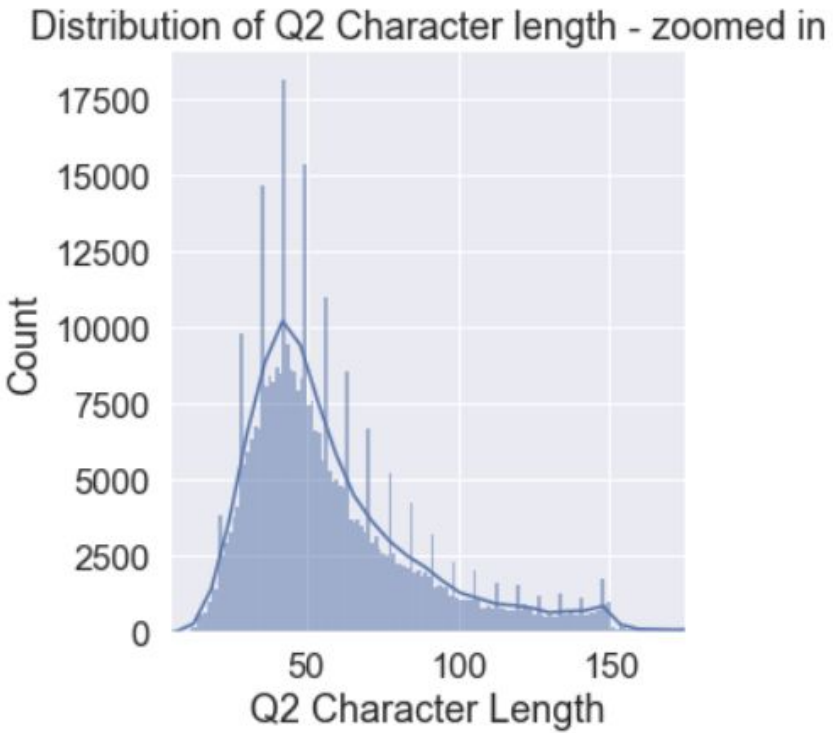

```
(5.0, 175.0)
```



Character Count Distribution - Question 2




```
(5.0, 175.0)
```



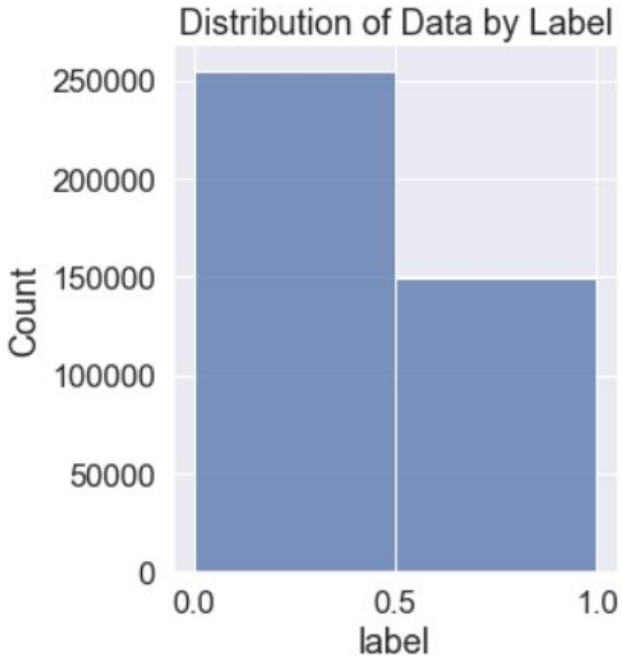
< 10 Characters

```
['Big data?',
 '……….',
 'What?',
 '?',
 'What?',
 'Hh',
 'Does?',
 'null ',
 'Delete',
 'Deleted.',
 'HH',
 'Why',
 '[removed]',
 'o',
 'Null ',
 'Edit',
 '????',
 'deleted',
 'Spam',
 'Hh ',
 'What?',
 'Deleted.',
 'lol',
 'Spam',
 '……….',
 'What is']
```

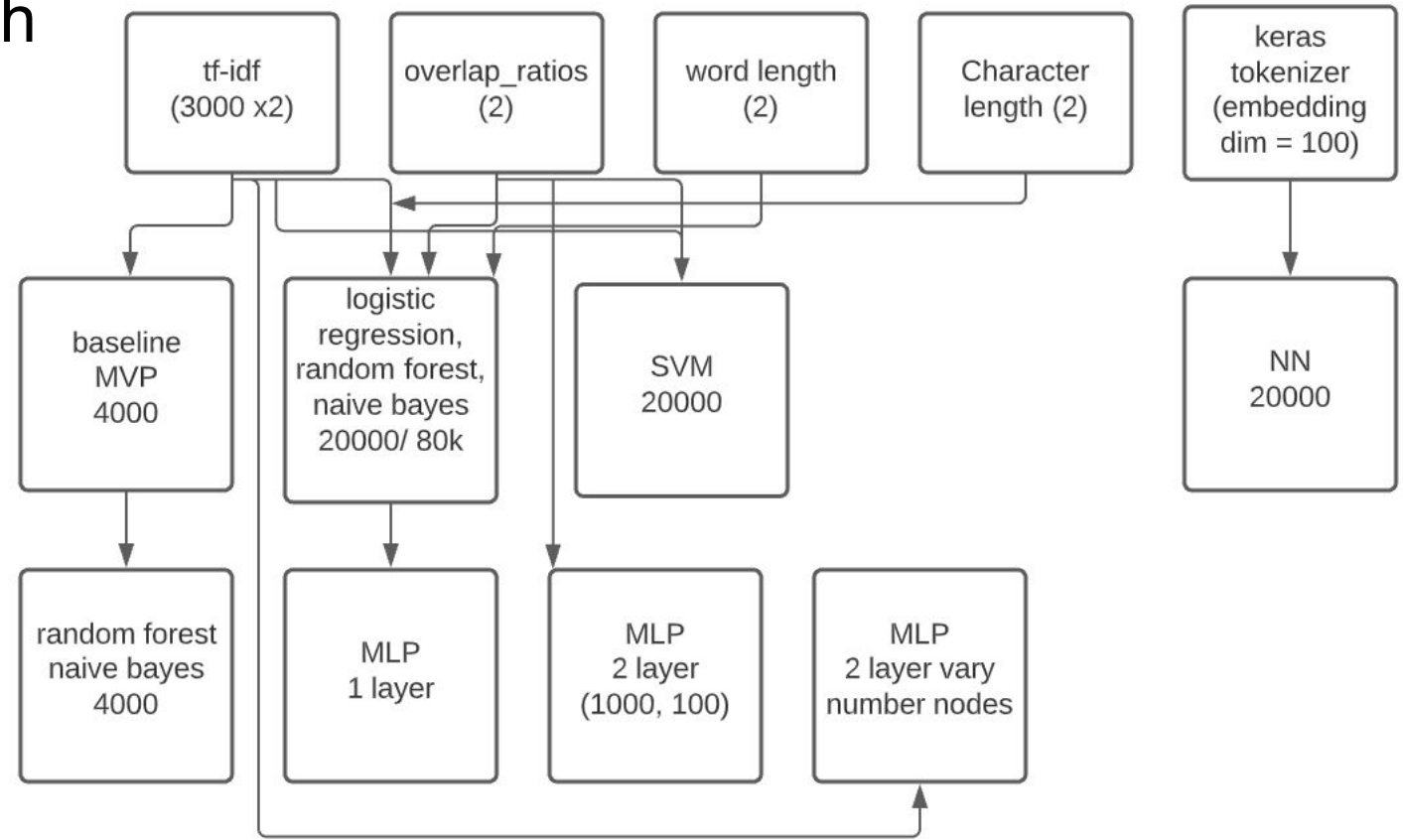
```
['.',
 '?',
 'deleted',
 '?',
 'deleted',
 'HH',
 'What?',
 'deleted',
 'deleted',
 'Na',
"I'm ",
 'grammar',
 'How long?',
 'What?',
 'lol ?',
 'Is?',
 'Deleted.',
 'How I am?',
 'Who Am I?',
 '?',
 'ok ?',
 '?',
 'What?',
 'i',
 'What',
 'o',
 '?',
 'deleted',
 'Deleted.',
```

Feature Exploration and Engineering

- Dropped null values
- Dropped <10, > 500 Characters
- Stripped leading and trailing spaces
- Lowercase
- Removed all punctuation
- Removed any word < 2 letters in length
- Removed stopwords
- Lemmatize

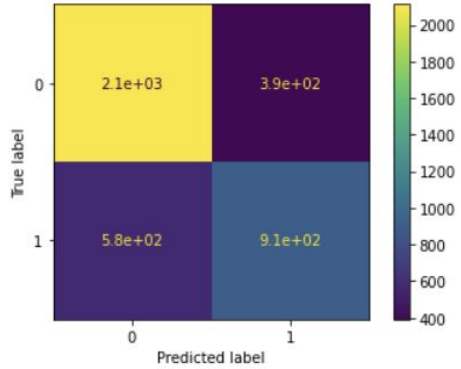


Approach


tf-idf
(3000 x2)
overlap_ratios
(2)
word length
(2)
Character
length (2)
keras
tokenizer
(embedding
dim = 100)
baseline
MVP
4000
logistic
regression,
random forest,
naive bayes
20000/ 80k
SVM
20000
NN
20000
random forest
naive bayes
4000
MLP
1 layer
MLP
2 layer
(1000, 100)
MLP
2 layer vary
number nodes

Results

```
Test set accuracy: 0.7561890472618155
Test set recall: 0.7266071943752872
Precision: 0.7561890472618155
```



	data points	features	accuarcy	recall	precision		
log reg	20000	tfidf, and overlap_ratio	0.74	0.71	0.74		
log reg	4000	tfidf	0.67	0.6	0.67		
log reg	20000	tfidf					
	80000	tfidf, and overlap_ratio	0.76	0.73	0.76		
Naive Bayes							
SVM	20000	tfidf, and overlap_ratio	0.69	0.64	0.68		
random forest	20000	tfidf, and overlap_ration and word count	0.75	0.71	0.75		
random forest	20000	tfidf, and overlap_ratio	0.76	0.73	0.76		
					train	test	
MLP	20000	tfidf, and overlap_ratio	1000, 100	cross entropy, ac	0.79	0.74	batchsize 10, epochs 2
	20000	tfidf, and overlap_ratio	1000, 1000, 100		0.81	0.75	batchsize 10, epochs 2
keras word embedding , max word len 100			embedding, flatten, dense 10		0.997	0.62	batchsize 10, epochs 2=3

Next Steps

- Reduce number of features significantly / add more data
- Add different features Ngrams
- Incorporate word2vec (to account for number of words) or doc2vec
- Tuning hyperparameters (at each step - really improve ensemble models with tweaking and tuning- felt like I was hypertuning ram)
- Pipeline and deploy model in cloud
- Additional models xgboost, LSTM