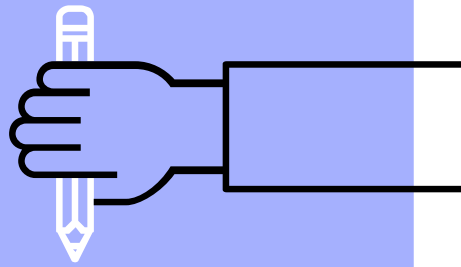
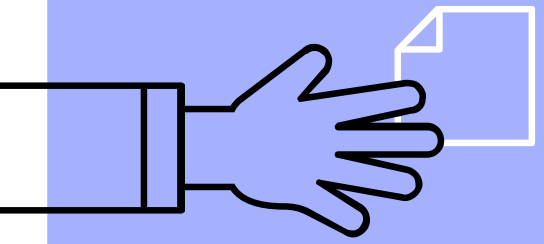


Using NLP on Crossword and Sudoku Subreddit Content to Interpret Keywords

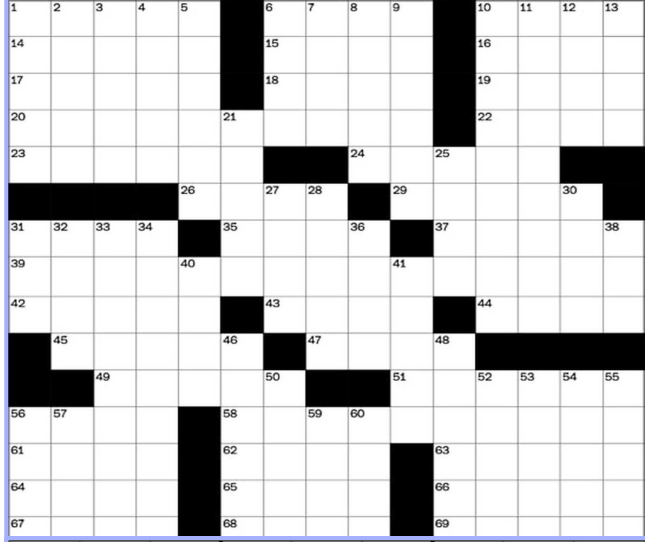


Gabe Cano

December 3, 2020

Agenda

- Problem Statement
- Preprocessing
- Modeling Process
- Interpretation
- Areas for Improvement
- Conclusion



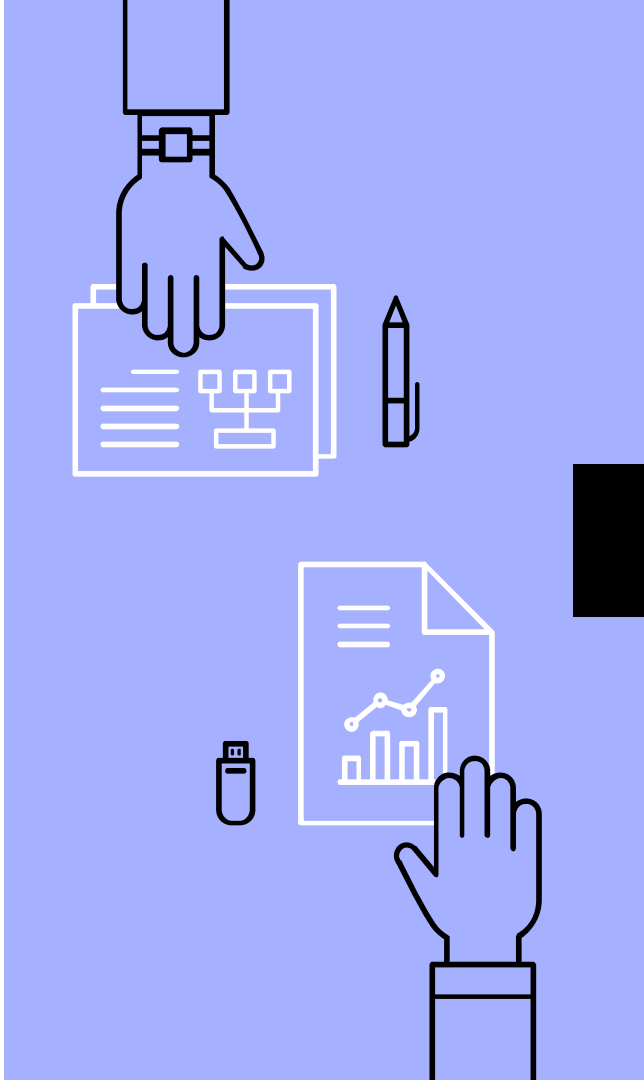
5	3			7				
6			1	9	5			
	9	8					6	
8				6				3
4			8		3			1
7				2				6
	6					2	8	
			4	1	9			5
				8			7	9

Problem Statement

The Puzzle Society hired me to help increase membership size. By modeling on crossword and sudoku subreddit content, the goal is to understand what aspects of those puzzles generate most discussion, so that the Puzzle Society can use that information to better know its audience and improve marketing efforts

Preprocessing

- Combined all text into one DataFrame
- Dropped unnecessary characters
- Lemmatized data



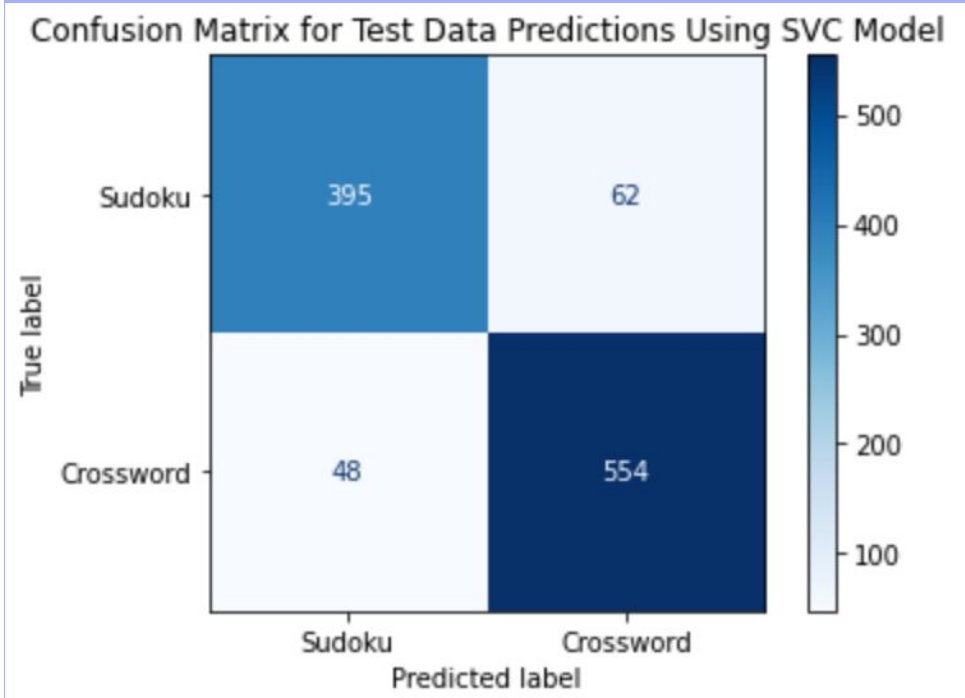
Modeling Process

- **Models ran:**
  - Logistic Regression
  - Naive Bayes
  - Support Vector Classifier
  - Baseline Accuracy**:** 56.85%
- **Methodology**:
  - Modeled data including and excluding 'crossword' and 'sudoku'
  - Pipeline/GridSearch to streamline process and optimize parameters

	Train Score	Train Score (Crossword and Sudoku Omitted)	Test Score	Test Score (Crossword and Sudoku Omitted)
svc_model	0.995138	0.985818	0.933900	0.897073
nb_model	0.968801	0.961507	0.928234	0.891407
logreg_model	0.995138	0.991086	0.935788	0.893296

Metrics Interpretation

- Plotted a confusion matrix using the Standard Vector Classifier model

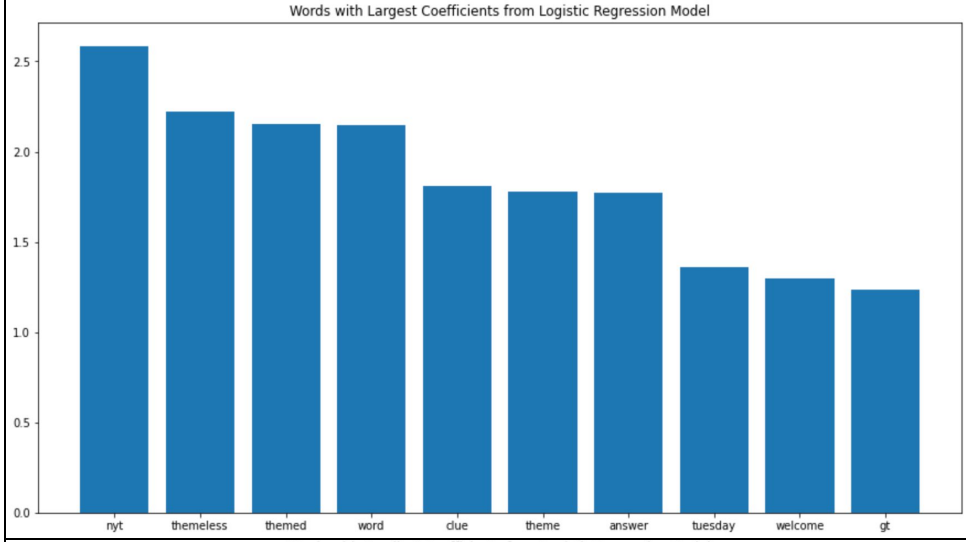


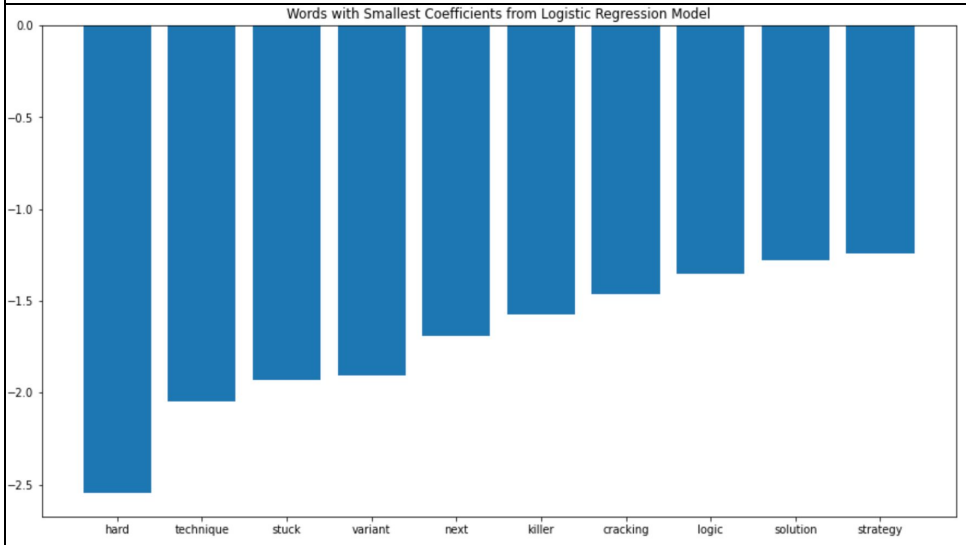

```
Accuracy: 0.8961284230406044
Misclassification Rate: 0.10387157695939564
Sensitivity: 0.920265780730897
Specificity: 0.8643326039387309
```

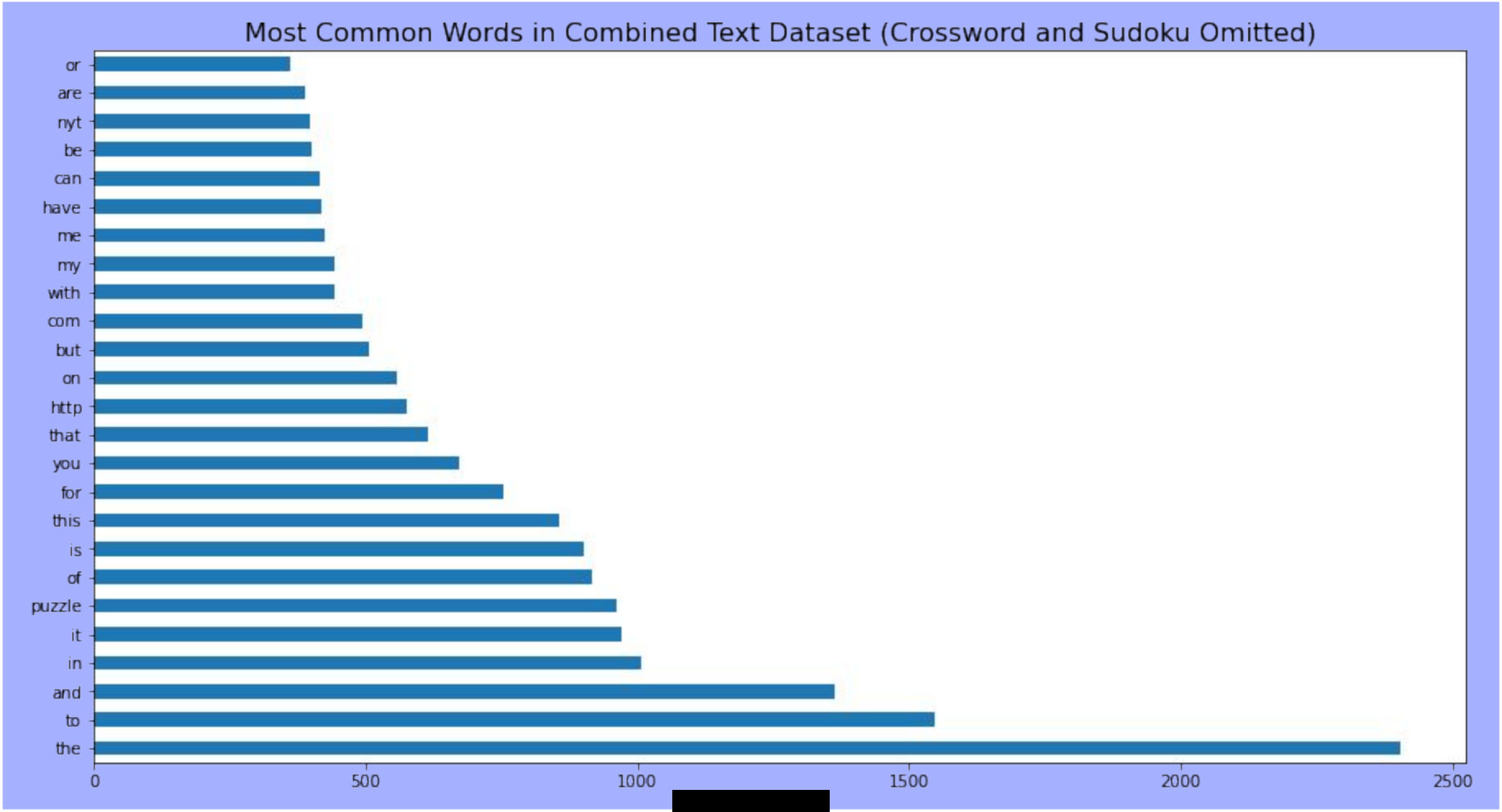
Coefficient Interpretation

- Logistic Regression model ran on text data to interpret coefficients for each word

Smallest Coefficients	Largest Coefficients
hard	nyt
technique	themeless
stuck	themed
variant	word
next	clue
killer	theme
cracking	answer
logic	tuesday
solution	welcome
strategy	gt






Most Common Words in Combined Text Dataset (Crossword and Sudoku Omitted)
or
are
nyt
be
can
have
me
my
with
com
but
on
http
that
you
for
this
is
of
puzzle
it
in
and
to
the
0
500
1000
1500
2000
2500

Conclusion

- Adopt a targeted approach to membership promotion
- Incorporate statistically significant keywords into advertisements directed at crossword and sudoku fans to attract a wider audience.

Areas for Future Improvement

- Research stop words more intensely
- Identify more insightful patterns during exploratory data analysis
- More models!