

# Space Wars: a 'literary' exercise in Natural Language Processing

Matt Paterson | [hello@hireMattPaterson.com](mailto:hello@hireMattPaterson.com)



Photo Credit: Wikipedia

# The Data Science Problem

- Virgin Galactic wants to charge customers \$250K per voyage to bring customers into outer space on a pleasure cruise in null G
- The potential customers range from more traditional HNWIs who have more conservative values, to the Nouveau Riche, and various levels of tech millionaires in between
- Marketing Analysts and Marketing Managers are expensive
- As headcount grows, overall ROI shrinks (VG HC ~ 200 ppl)


# The Solution

- Create a machine learning model to identify what type of interests each user has based on their social media and reddit posts
- Narrowcast to each smaller cohort with the language, tone, and vocabulary that will push each to purchase the quarter-million dollar flight






(Source: Virgin Galactic IR) | <https://seekingalpha.com/article/4359851-virgin-galactic-dont-buy-richard-bransons-lagging-space-startup>



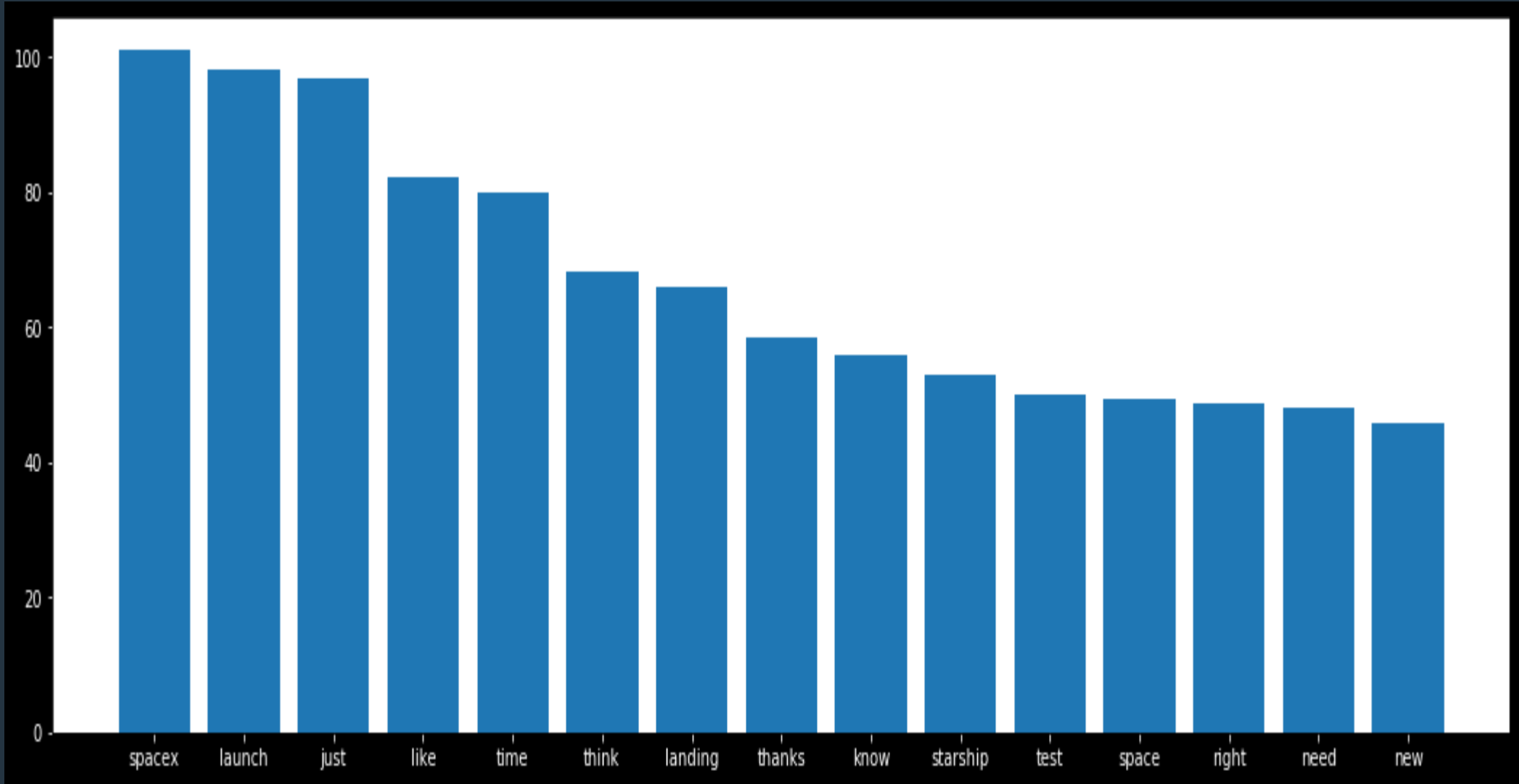
HireMattPaterson.com has been contracted by Virgin Galactic's marketing team to build a Natural Language Processing Model that will efficiently predict if reddit posts are being made for the SpaceX subreddit or the Boeing subreddit as a proof of concept to segmenting the targeted markets.

We've created a model that predicts the silo of the post with nearly 80% accuracy (with a top score of 79.9%). To get there we tried over 2,000 different iterations on a total of 5 different classification modeling algorithms including two versions of Multinomial Naïve Bayes, Random Cut Forest, Extra Trees, and a simple Logistic Regression Classifier. We'd like to use Support Vector Machines as well as Gradient Boosting and a K-Nearest Neighbors model in our follow-up to this presentation.

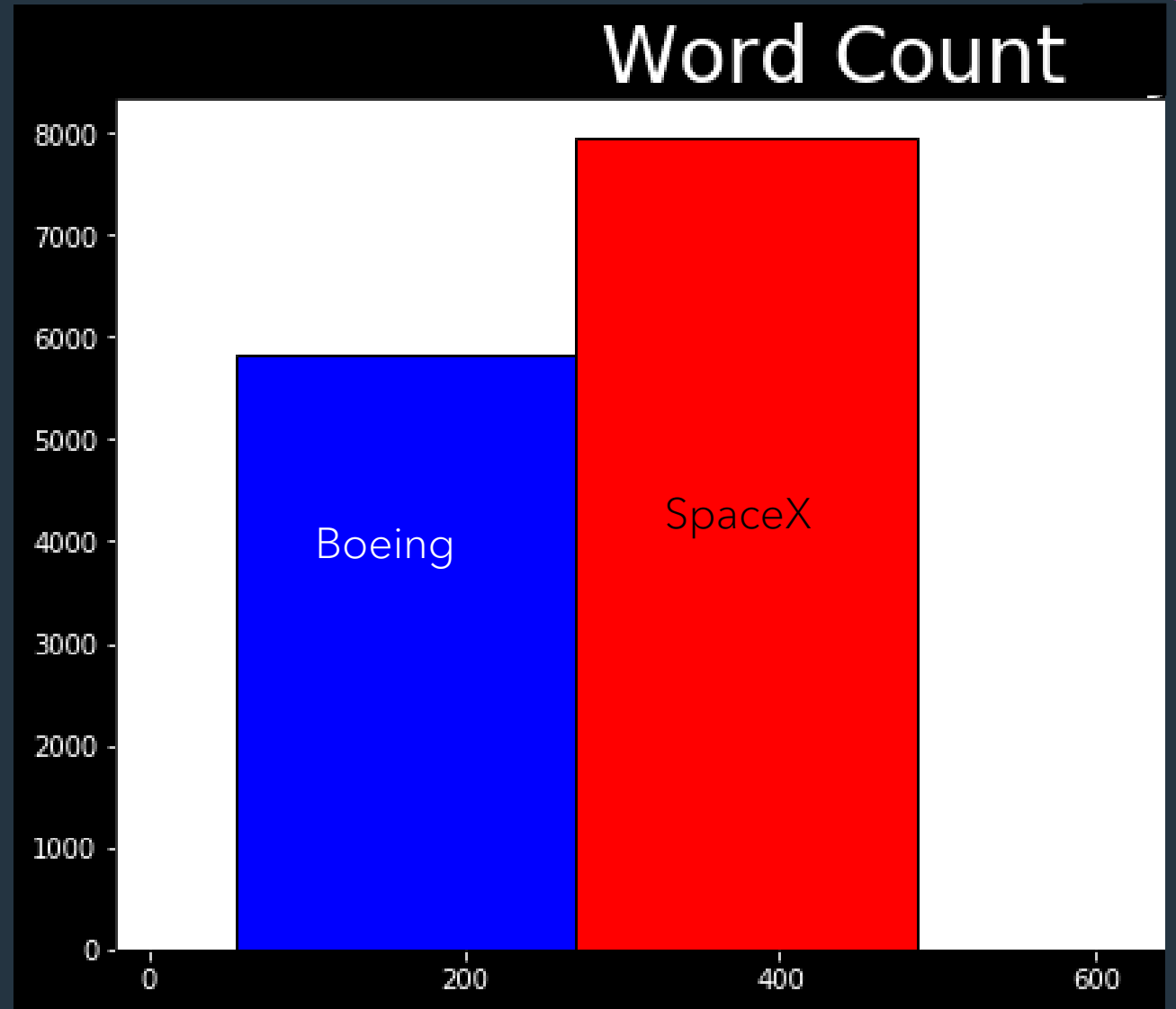
If you like our proof of concept, the next iteration of our model will take in to account the trend or frequency in the comments of each user; what other subreddits these users can be found to post to (are they commenting on the Rolex and Gulfstream and Maserati or are they part of the Venture Capital and AI crowd?); and if their comments appear to be professional in nature (are they looking to someday work in aerospace or maybe they already do). These trends will help the marketing team tune their tone, choose words that are trending, and speak directly to each cohort in a narrow-cast fashion thus allowing VG to spend less money on ads and on people over time.



# Some words are popular



# Size Matters



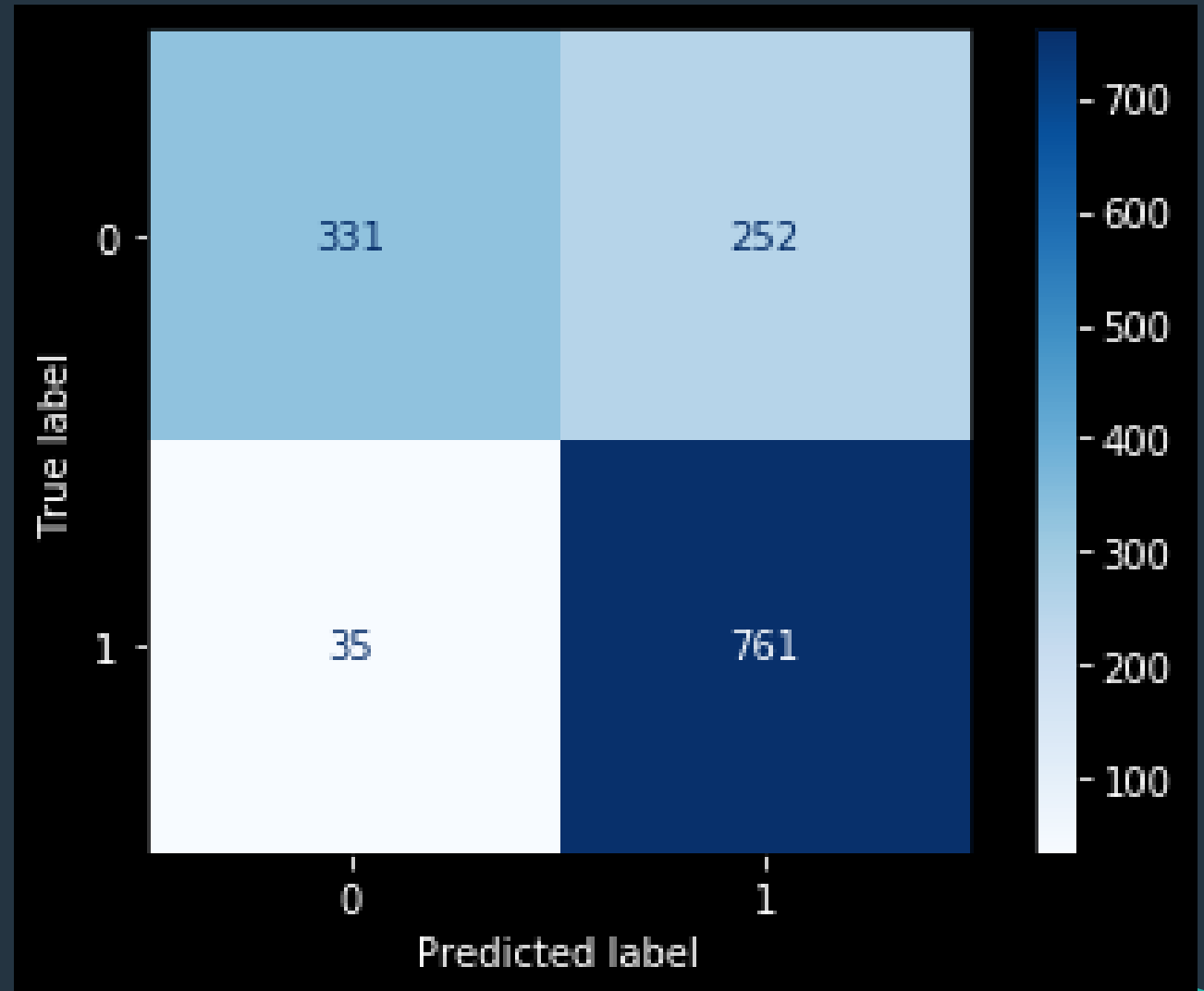
# Confusion Matrix of Naïve Bayes w/C Vec

Shown here, the 0's are Boeing  
and the 1's are SpaceX.

Naïve Bayes predicts Boeing  
better than SpaceX in this model.

False Negatives: 35 of 366  
Specificity: 90%

False Positives: 252 of 1013  
Sensitivity: 75%





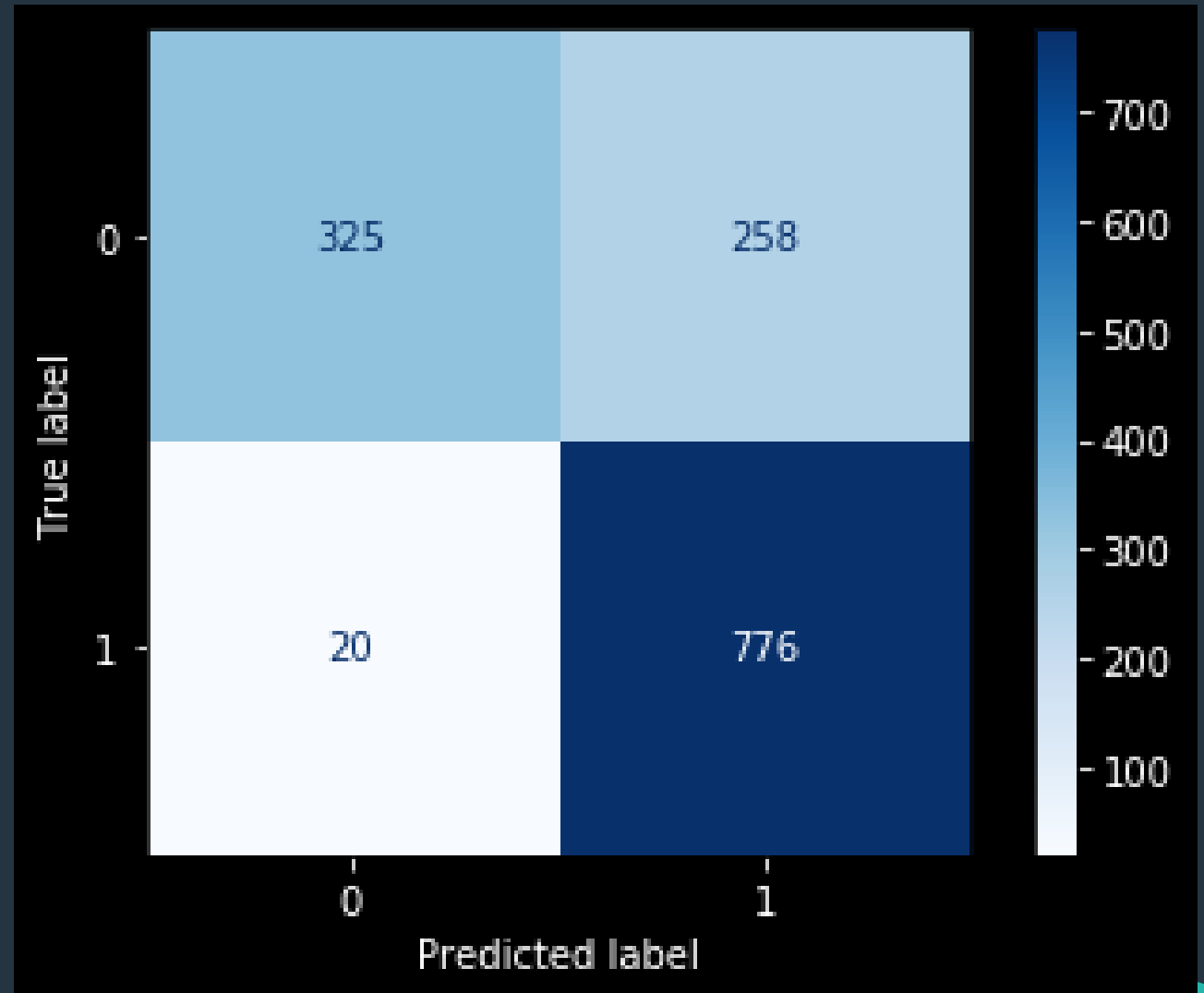
# Confusion Matrix of Naïve Bayes w/TFIDF

Shown here, the 0's are Boeing  
and the 1's are SpaceX.

Naïve Bayes predicts Boeing  
better than SpaceX in this model.

False Negatives: 20 of 345  
Specificity: 94%

False Positives: 258 of 1034  
Sensitivity: 75%



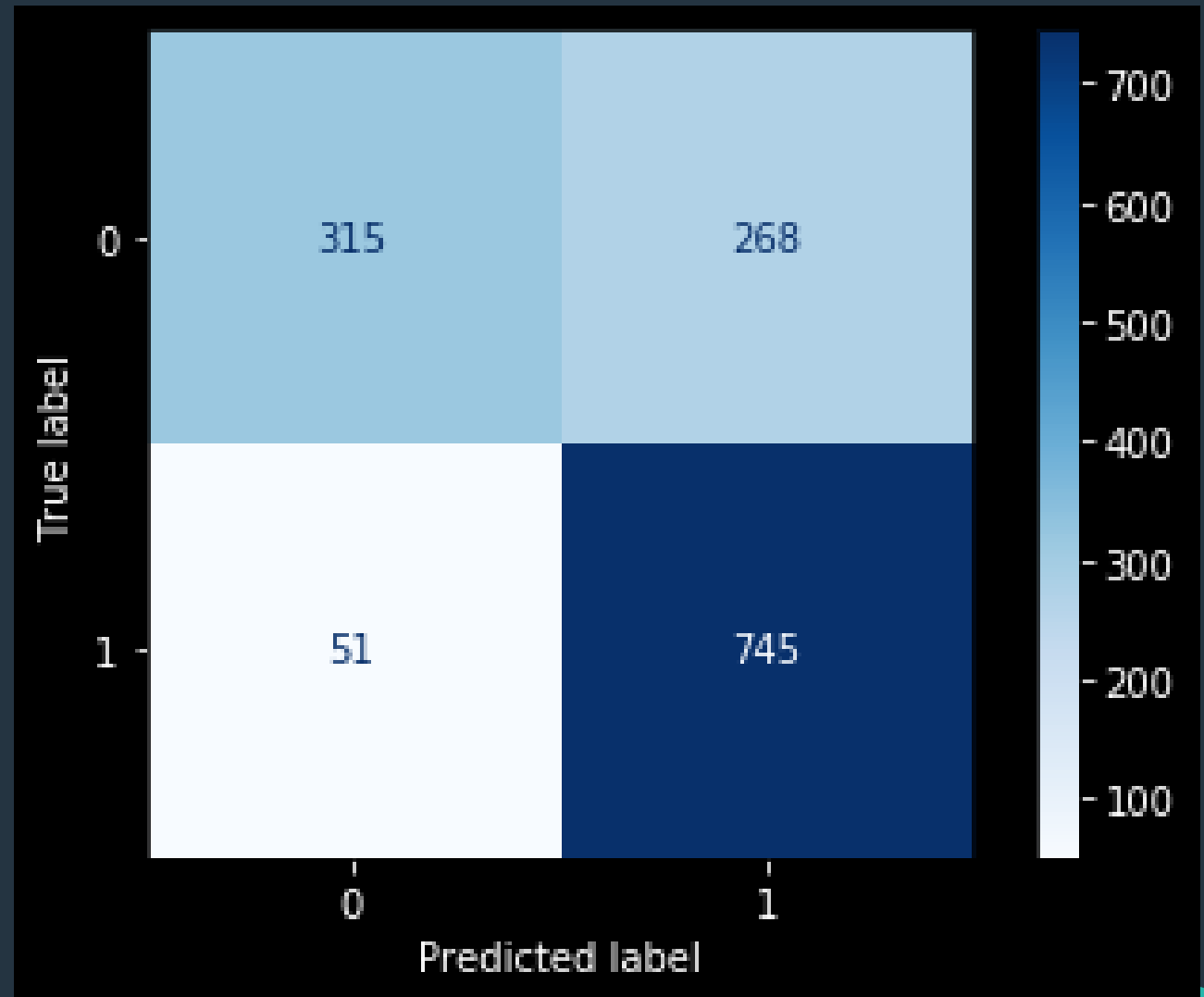
# Confusion Matrix of Random Cut Forest

Shown here, the 0's are Boeing  
and the 1's are SpaceX.

Random Forest predicts Boeing  
better than SpaceX in this model.

False Negatives: 51 of 366  
Specificity: 86%

False Positives: 268 of 1013  
Sensitivity: 74%



# Scoreboard of Algorithms

baseline	Naive Bayes	Naive Bayes TFID	Logistic Regression	Random Cut Forest
0.577573	0.796744	0.795857	0.79913	0.773695

We can see that while all of our models improved the baseline score by at least 20 percentage points, the simple Logistic Regression model was the winner with 79.9% accuracy score.

# Conclusion and Recommendations

- Hire us to run these same models on the users, and then create categories to target market to the various segments
- We can predict with 80% accuracy which posts are meant for which thread, and when we isolate the focus to the users it will allow more efficient and cost-effective target marketing
- Narrowcasting yields far greater ROI in your marketing
- Our ML is cheaper than hiring a whole new team of Marketing Analysts

# Questions? Comments? Concerns?



Virgin Galactic's SpaceShipTwo "Unity" launches toward the edge of space on December 13, 2018.  
Virgin Galactic; MarsScientific.com/Trumbull Studios : Credit 'Business Insider', July 15, 2020