

The background features a complex, abstract visualization of data. It includes a bar chart with bars in shades of orange, yellow, and blue. Overlaid on this are several line graphs: a solid blue line, a dashed white line, and a solid white line. A numerical value '+11,000.00' is displayed in white at the top center. The overall color palette is dominated by dark blues, oranges, and yellows, creating a high-tech, analytical feel.

Project 3: Distinguishing Between Data Science & Analytics Subreddits


Jamie Squires

Background:

“Data Science” vs. “Analytics”


There's little consensus on the specific differences between data science and analytics

Data Scientist
also known as Data Managers, statisticians.



A data scientist will be able to take data science projects from end to end. They can help store large amounts of data, create predictive modelling processes and present the findings.

Skills: Mathematics, Programming, Communication



Will use programmes such as:
SQL, Python, R

Data Analysts
also known as business Analysts.



They typically help people from across the company understand specific queries with charts.

Skills: Statistics, Communication, Business knowledge



Will use programmes such as:
Excel, Tableau, SQL

“Data analytics focuses more on viewing the historical data while data science focuses more on machine learning and predictive modeling”₁



“Whereas data analytics is understanding datasets and gleaning insights that can be turned into actions, data science is centered on building, cleaning, and organizing datasets.”₂

Problem Statement

Given the amount of overlap between the two data-centric specialties of 'Data Science' vs 'Analytics', let's see if there's a tangible difference in the subject matter amongst topics in the Data Science and Analytics subreddit.

Can we accurately categorize reddit posts from the Data Science and Analytics subreddit?



Action Plan

- We will perform a variety of classification methods to accurately predict which subreddit titles came from which subreddit (Data Science vs Analytics)

Success Metrics:

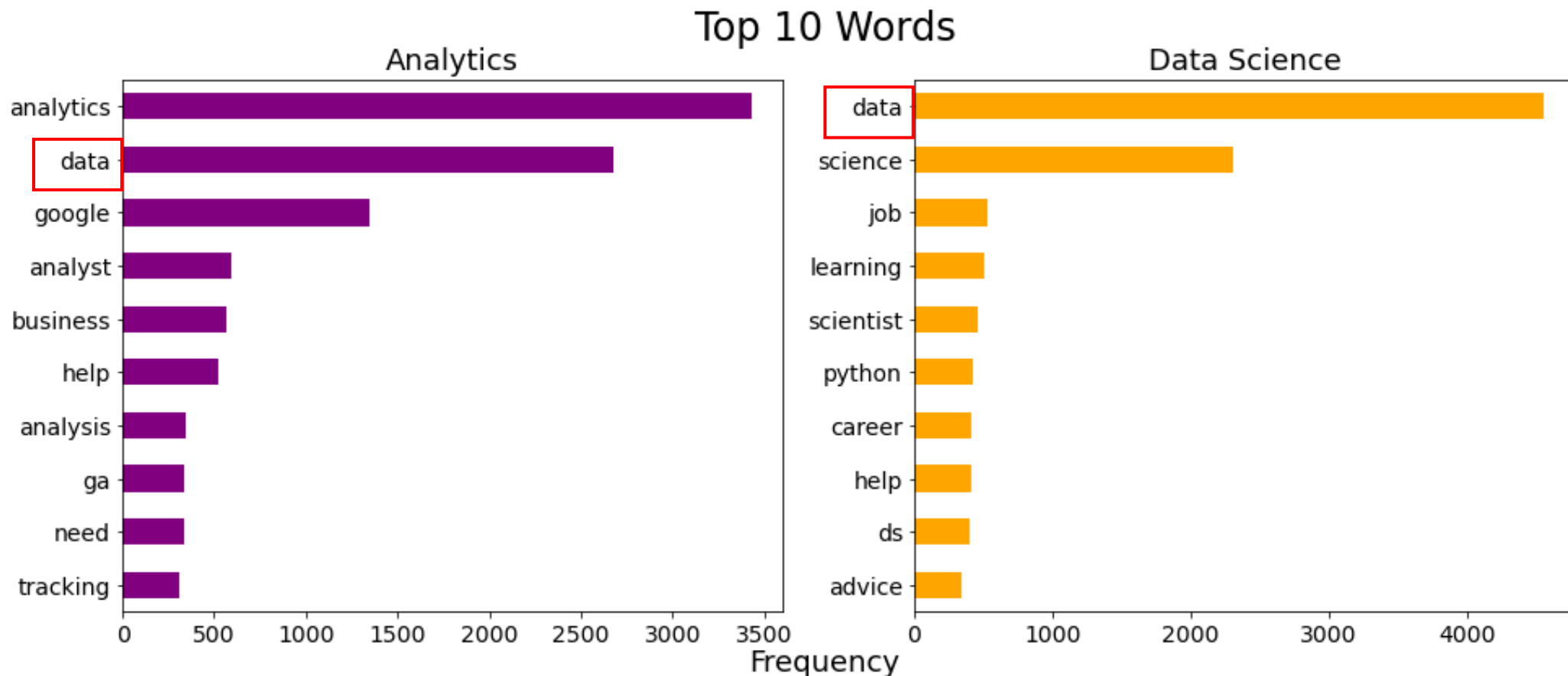
- ✓ Optimize for the highest **accuracy** score in classifying posts

Methodology

1. Use Pushshift API to collect posts (last 10K posts per subreddit)
2. Clean & explore text data through Count Vectorizer and TF-IDF Vectorizer
3. Compare various classification models to find what works well
4. For the models yielding the best results, fine-tune parameters via GridSearch
5. Evaluate best fit model & compare to baseline accuracy score

EDA: Raw Frequency of Words

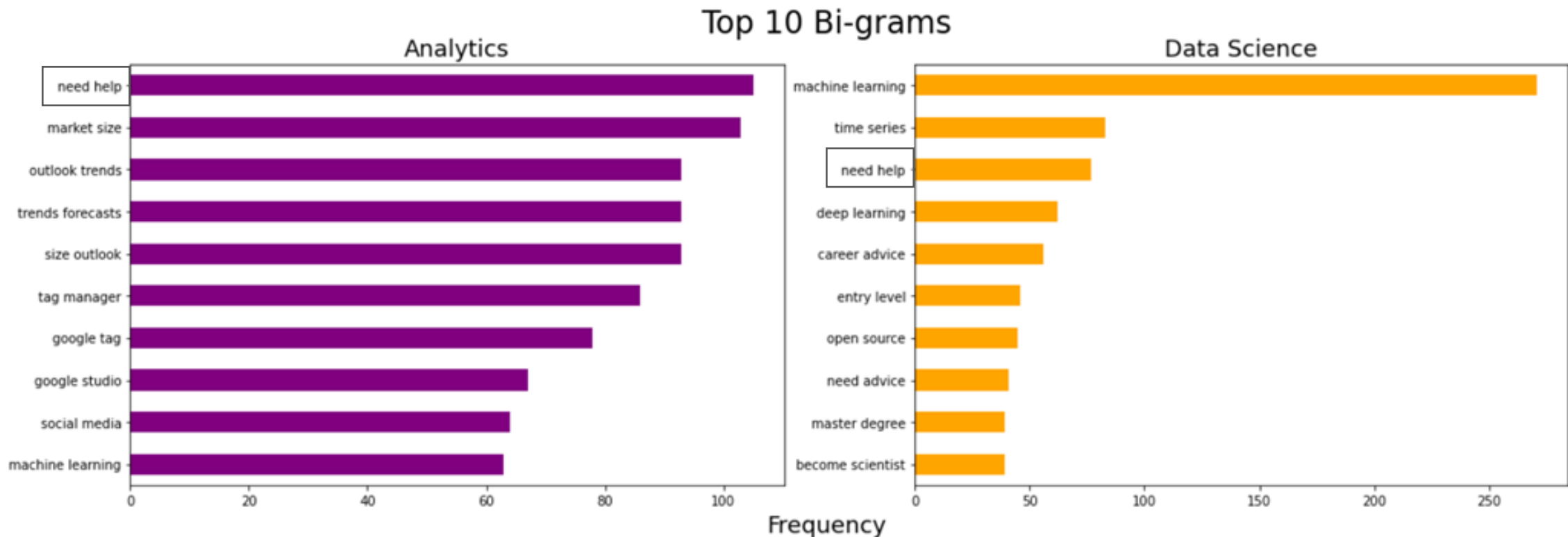
- Unsurprisingly, **data** is a top term in both subreddits



EDA: Bigram Frequency with Stop Words

Insights:

- Both subreddits are a place for users to get help & seek advice
- Analytics subreddit focuses on trends/ forecasts, “google analytics”, business problem oriented
- Data Science subreddit shows that machine learning is very popular, posts are career trajectory oriented



Modeling & Evaluation

Setup & Methodology

1. 1st Pass of Model Selection
 - Compare various classification models to assess which models perform better than others
2. 2nd Pass of Model Selection
 - For the models yielding the best results, fine-tune parameters via GridSearch
3. Evaluate best fit model & compare to baseline accuracy score

Modeling & Evaluation

1st Pass of Model Selection

Vectorizer	Classifier	Train Accuracy	Test Accuracy
TF-IDF	LogisticRegression	0.8460	0.8026
CountVectorizer	LogisticRegression	0.8723	0.8014
TF-IDF	Random Forests	0.9936	0.7998
CountVectorizer	Random Forests	0.9920	0.7954
CountVectorizer	Gradient Boost	0.8272	0.7932
TF-IDF	Gradient Boost	0.8366	0.7890
CountVectorizer	Naïve Bayes	0.8501	0.7872
TF-IDF	Bagging Classifier	0.9935	0.7870
TF-IDF	SVM	0.9236	0.7866
CountVectorizer	Bagging Classifier	0.9935	0.7852
CountVectorizer	AdaBoost Classifier	0.8256	0.7822
TF-IDF	AdaBoost Classifier	0.8352	0.7766
TF-IDF	Naïve Bayes	0.8514	0.7700
CountVectorizer	SVM	0.9399	0.7670
CountVectorizer	Decision Trees	0.9936	0.7602
TF-IDF	Decision Trees	0.9936	0.7473
TF-IDF	KNN	0.6192	0.5471

Findings:

- **Logistic Regression, Random Forests, and Gradient Boost** performed the best
- TF-IDF tended to perform slightly better than CountVectorizer

Next Steps:

- Let's tune the top performing models using GridSearch to improve performance

Modeling & Evaluation

2nd Pass of Model Selection

Vectorizer	Classifier	Cross Val Score	Training Accuracy	Test Accuracy
CountVectorizer	LogisticRegression	0.8009	0.9094	0.8070
TF-IDF	LogisticRegression	0.8022	0.8898	0.8064
CountVectorizer	Gradient Boost	0.7898	0.8167	0.7946
TF-IDF	Naïve Bayes	0.7861	0.9382	0.7936
TF-IDF	RandomForests	0.7844	0.8822	0.7886
TF-IDF	Gradient Boost	0.7860	0.8260	0.7886
TF-IDF	Decision Trees	0.7590	0.8530	0.7538
TF-IDF	KNN	0.7500	0.7596	0.7451

- **Logistic Regression is the winner!**

- Test Accuracy: 80.7%



- **Outperformed our baseline:**

- Analytics: 50%
- Data Science: 50%

Conclusion

Recommendation:

- Use the **Logistic Regression** x CountVectorizer model to predict the analytics vs data science subreddit classification
- This model got **80.7%** of our predictions correct

Takeaways:

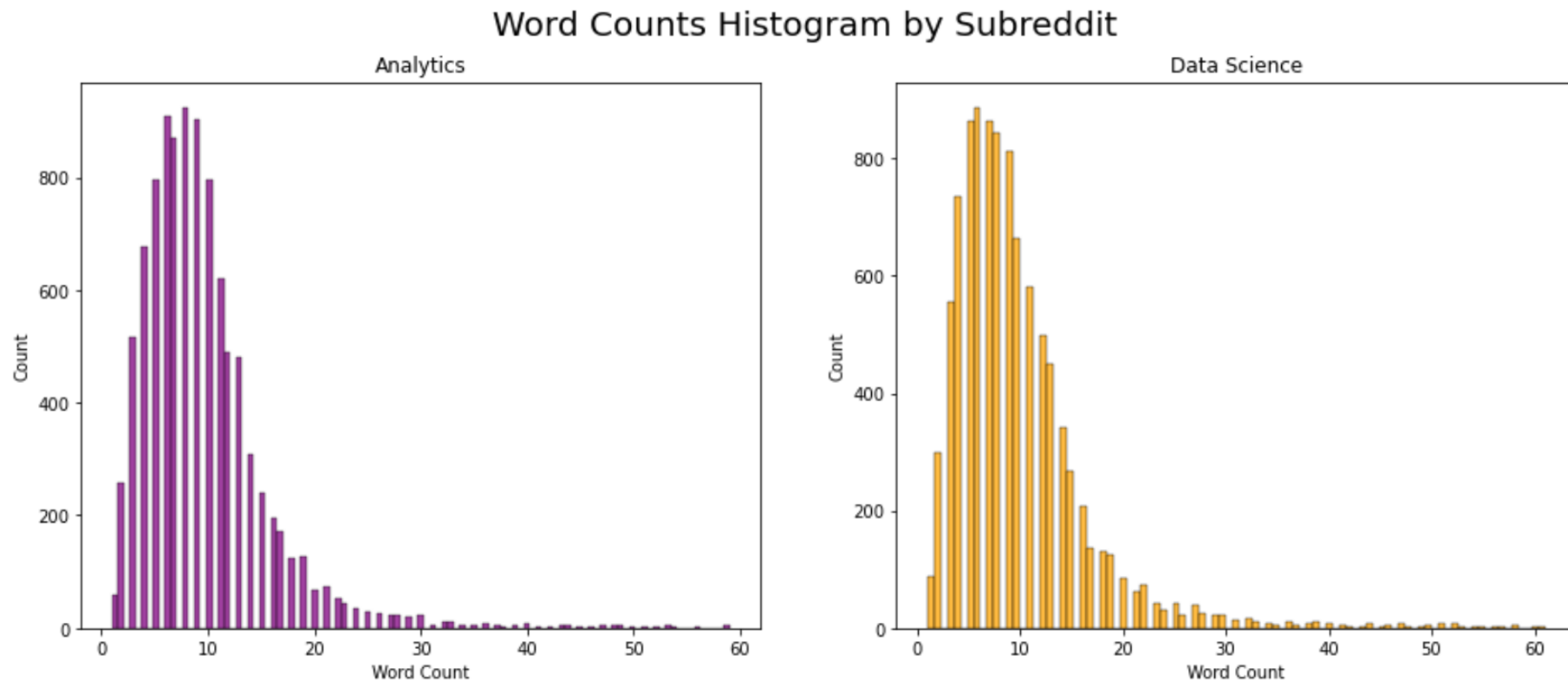
- It's evident that there is a lot of overlap between the topics on Data Science & Analytics subreddits, as shown by our accuracy scores

Thank you!

Appendix

EDA: Length of Post & Word Count

- Both subreddits follow a similar distribution in word count per post



Sources

Image on page 1:

- https://www.nyit.edu/files/degrees/Degree_CoECS_DataScienceMS_Hero.jpg

Sources/image on page 2:

- <https://i.ytimg.com/vi/yFmkK94rnc0/maxresdefault.jpg>
- Source 1: https://medium.com/@springboard_ind/data-science-vs-data-analytics-how-to-decide-which-one-is-right-for-you-41e7bdec080e
- Source 2 : <https://online.hbs.edu/blog/post/data-analytics-vs-data-science>

Image on page 3:

- https://play-lh.googleusercontent.com/MDRjKWEIHO9cGiWt-tlvOGpAP3x14_89jwAT-nQTS6Fra-gxfakizwJ3NHBTCINGYK4