Phase 1

EDA & Basic Modeling Results

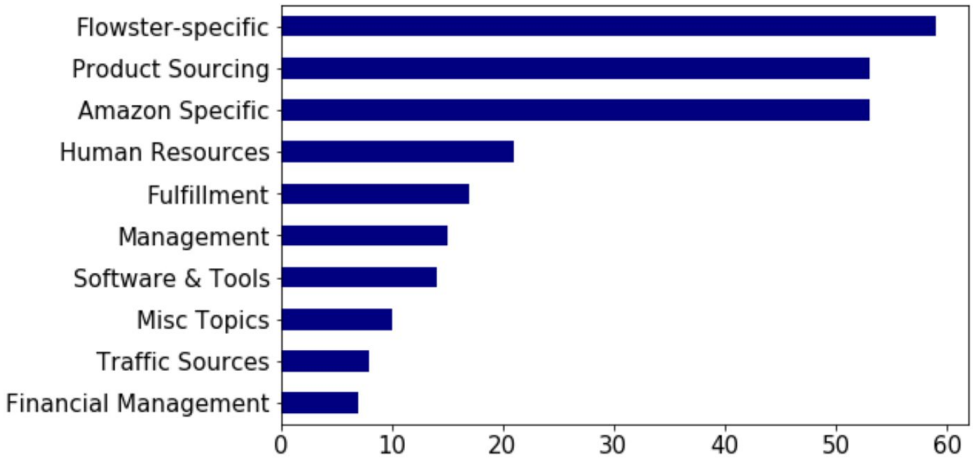
Team Flowster

Agenda

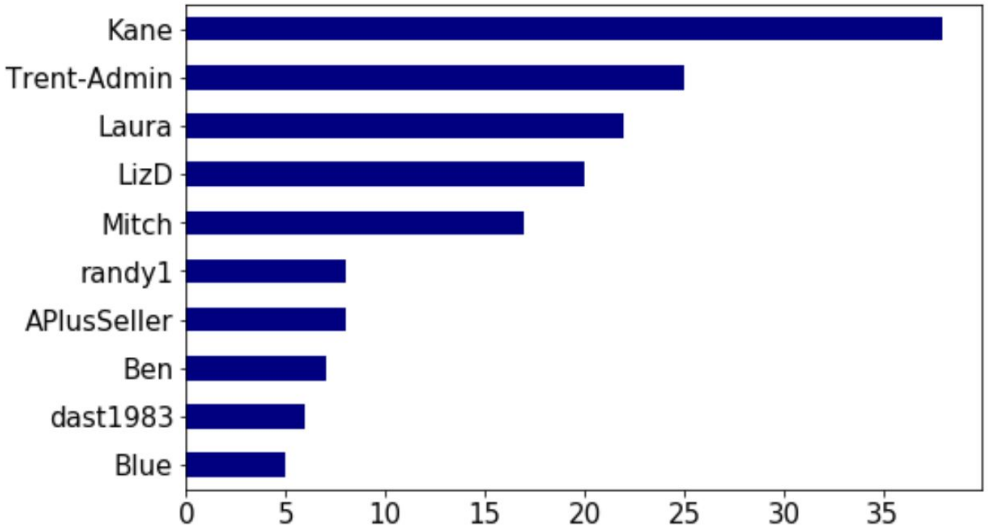
- Exploratory Data Analysis (EDA)
- Data Cleaning Techniques
- Evaluation Metrics
- Results
- Notes and Conclusion

EDA - Class Imbalance

```
Flowster-specific             59
Product Sourcing              53
Amazon Specific               53
Human Resources               21
Fulfillment                   17
Management                    15
Software & Tools              14
Misc Topics                   10
Traffic Sources                8
Financial Management           7
eCommerce Marketplaces         2
Store & Website Management     1
Name: Category, dtype: int64
```



EDA - Author Imbalance and Lack of Page Statistics



	count	mean	std	min	25%	50%	75%	max
Likes	260.0	1.319231	2.346857	0.0	0.00	1.0	2.00	19.0
Replies	260.0	2.780769	4.038666	0.0	1.00	2.0	3.00	51.0
Views	260.0	167.807692	126.395414	0.0	106.75	151.5	216.25	736.0

EDA - Stopwords

	Leading Comment	word_count
0	Have questions about sourcing products? This i...	23
1	Hi! We are new to the forum and are going thro...	63
2	As I am working in Amazon as a seller from las...	81
3	Does anyone have a VA they recommend, have use...	16
4	Can you sell branded products on Amazon Uk or ...	15

	Leading Comment	stopwords
0	Have questions about sourcing products? This i...	10
1	Hi! We are new to the forum and are going thro...	28
2	As I am working in Amazon as a seller from las...	23
3	Does anyone have a VA they recommend, have use...	7
4	Can you sell branded products on Amazon Uk or ...	6

EDA - Numerics and Special/ASCII Characters

	Leading Comment	hastags
0	Have questions about sourcing products? This i...	0
1	Hi! We are new to the forum and are going thro...	0
2	As I am working in Amazon as a seller from las...	0
3	Does anyone have a VA they recommend, have use...	0
4	Can you sell branded products on Amazon Uk or ...	0

	Leading Comment	numerics
0	Have questions about sourcing products? This i...	0
1	Hi! We are new to the forum and are going thro...	0
2	As I am working in Amazon as a seller from las...	2
3	Does anyone have a VA they recommend, have use...	0
4	Can you sell branded products on Amazon Uk or ...	0

EDA - Rare Words

- Spelling mistakes
- Hyperlinks
- Costs (eg. $1.25 or $25,000)
- Slashes (eg. buy/sell)
- Date short forms (eg. Jun-26)
- Acronyms
- Lack of space between punctuation (eg. “Sunday.I”, “this)now”, “for,he”)

Data Cleaning Techniques Used

- *Data augmentation by turning additional comments into “new” topics with the same category
- Lowercase normalization
- Punctuation removal
- Stopword removal
- Rare word removal
- Non-ASCII character removal
- Numerics removal
- Numeric and cost replacements using unique identifiers “########” and “$$$$$$$$”

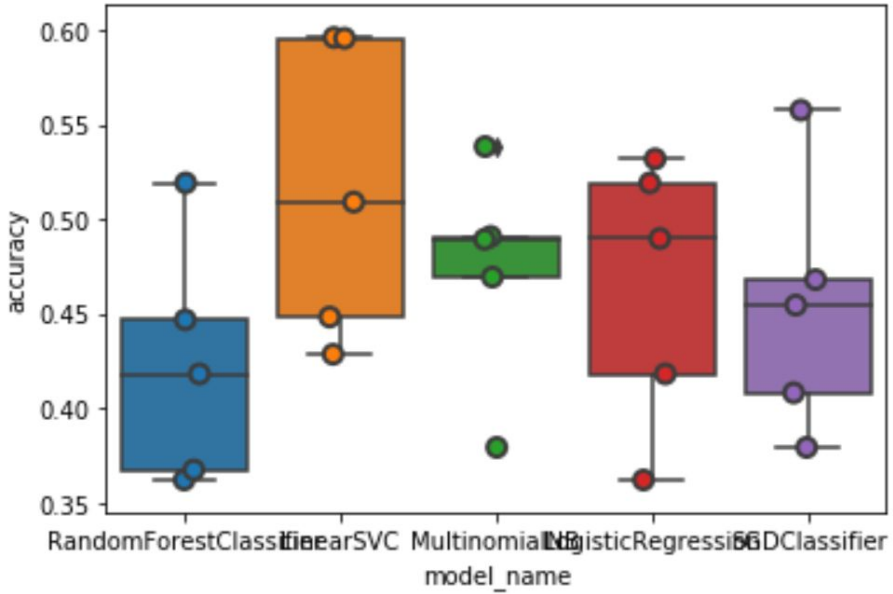
Evaluation Metrics

1. 5-Fold Cross Validation
    a. Used to measure the consistency of accuracy scores across different shuffles of the dataset
2. Scikit-learn Classification Report
    a. Accuracy is main value of interest
    b. Precision, recall and f-1 scores are also reported, which are different ratios involving TPs, TNs, FPs and FNs
    c. Provides accuracy scores for each category

Embedding Techniques vs ML Models Results

	Bag of Words	TF-IDF	Word2Vec	Doc2Vec
Linear SVM		0.55		0.6
Logistic Classifier	0.85	0.51, 0.83	0.19	0.68, 0.55
Naive Bayes	0.59	0.59, 0.48	0.11	0.29

Cross-Validation Box Plot Examples



Notes

- The dataset and evaluation metrics were standardized across tests, but the data pre-processing was left to be flexible and changed based on which combination was being run at the time
- In general, it was found that pre-processing makes less of a difference than simply having more data
- Only the leading comments (and in some cases the topic titles) were used in the construction of topic feature vectors

Conclusion

Based on the results, it appears that most models have the potential to be a decent classifier given the correct embedding technique, proper tuning, and an abundance of data.

Personally, I recommend the logistic classifier. It was shown to have one of the highest scores based on our results and it is more intuitive to train.