AirBnB vs ShoeString

Agenda

Problem statement

Data Analysis

Modelling

Conclusion and Recommendations

Problem Statement

What Airbnb offer isn’t a cheap place to sleep when you’re on holiday, it’s the opportunity to experience your destination as a local would.
It’s the chance to meet the locals, experience the markets, and find the non-touristy places.

AirBnB vs ShoeString

Data Analysis

r/AirBnB		r/Shoestring
889	Posts	935
15	Empty Posts	53
55	Ave Title length	51
895	Ave Post length	613
95	Ave Unique word	70
77	Ave Root word	55

Analysis of these textual statistics, didn't yield clear means to classify the 2 subreddits

Data Analysis (Wordclouds)

r/AirBnB

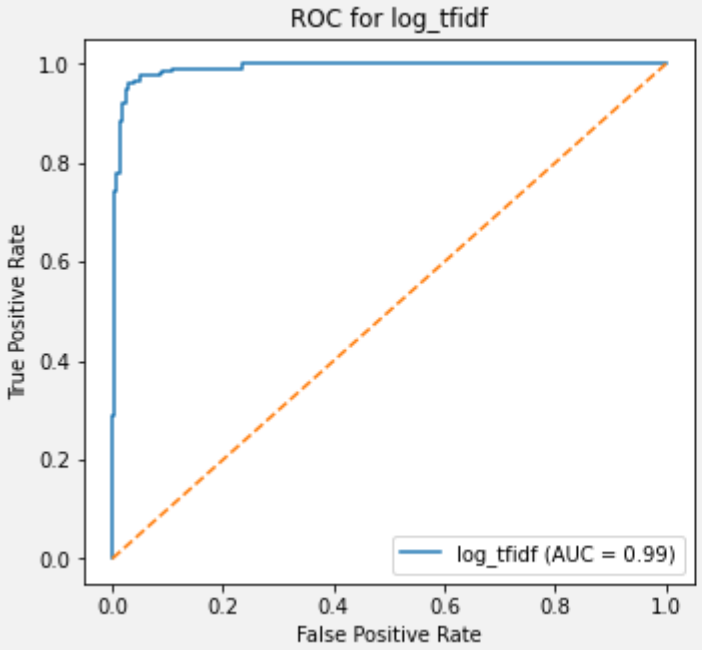
r/Shoestring

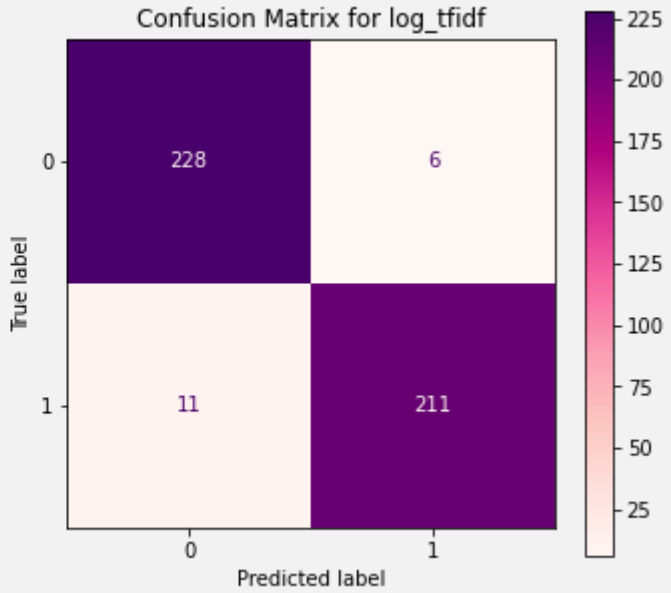
Stop words that were identified: Day, Month, back, book and AirBnB

Modelling

	mnb_hash	log_tfidf
Train score	0.9890	0.9868
Test score	0.9627	0.9627
Score diff	0.0263	0.0241
No. of features	131,072	1000
Precision	0.9680	0.9724
Specificity	0.9701	0.9744
Sensitivity	0.9550	0.9505
True Negatives	227	228
False Positives	7	6
False Negatives	10	11
True Positives	212	211

ROC plot and confusion matrix for Test data

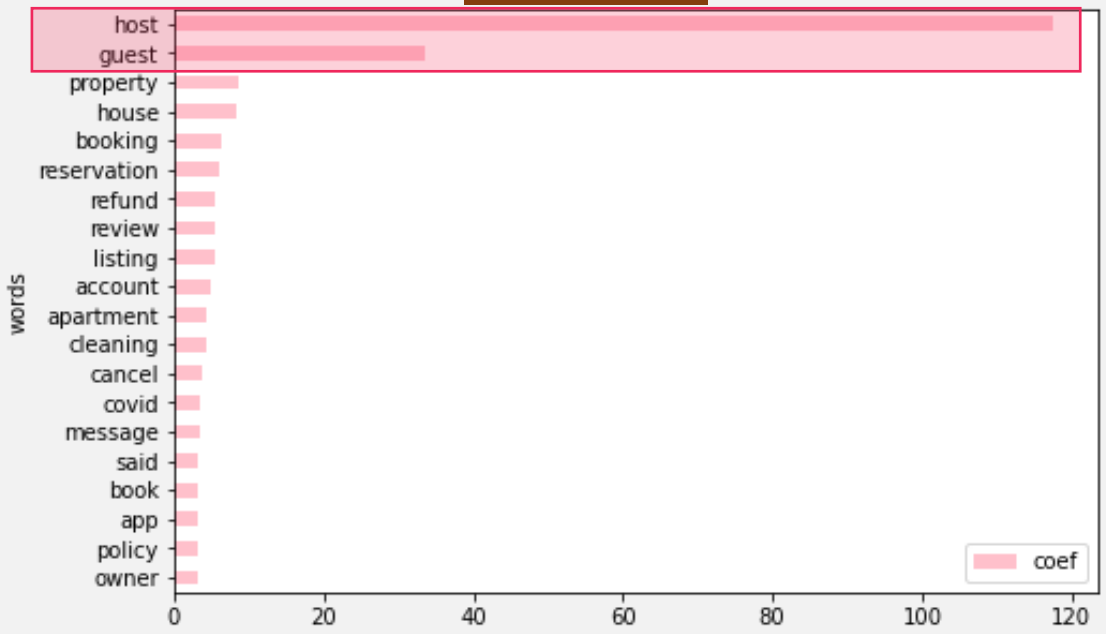




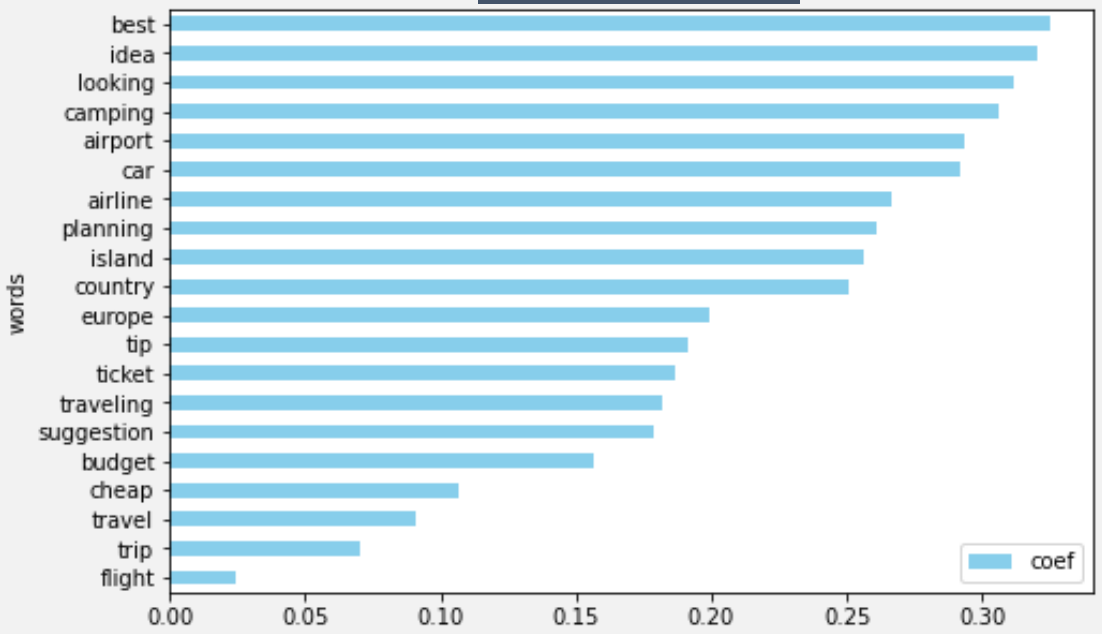
Term Frequency Inverse Document Frequency Vectorizer and Logistic Regression combination

Modelling (features)

r/AirBnB



r/Shoestring



2 Top words that affects the classifications of AirBnB posts are **host** and **guest**

Conclusion

Type 1 Error	Type 2 Error
13 False Positives	22 False Negatives
Posts classified as AirBnB that are actually ShoeString posts	**Posts classified as ShoeString that are actually AirBnB posts**

Overall
35 misclassifications
98% of posts classified correctly

Still within expectation of 95% accuracy as set out in the beginning of the project

Recommendations

For the future phase of the project:

- Dive into analysis of the comments
- Analysis of the sentence structure and POS tagging
- Sentiment analysis on the posts and comments
- Use of Support Vector Machine model

Thank You

AirBnB vs ShoeString