Sci-fi V.S. Fantasy

The Story of a Model

Agenda

1. Data Wrangling
2. Exploratory Data Analysis
3. Model Mania
4. Word Frequency
5. Conclusions and Future Paths

Data Wrangling - Preparing the Ingredients

- PushShift API - Web Scraping Tool
- SKLearn - Analysis Tool Library
- Subreddits - Which to choose?
    - Sci-fi
    - Fantasy
- 6000 Total Posts
    - 100 Posts Every Minute
    - 30 Iterations

EDA - What makes up our data?

- Are there more media posts on one subreddit versus the other?
- Is the existence of self-text enough to determine origin?
- Minimum Viable Product
    - Title
    - Subreddit
    - Selftext

Menagerie of Models

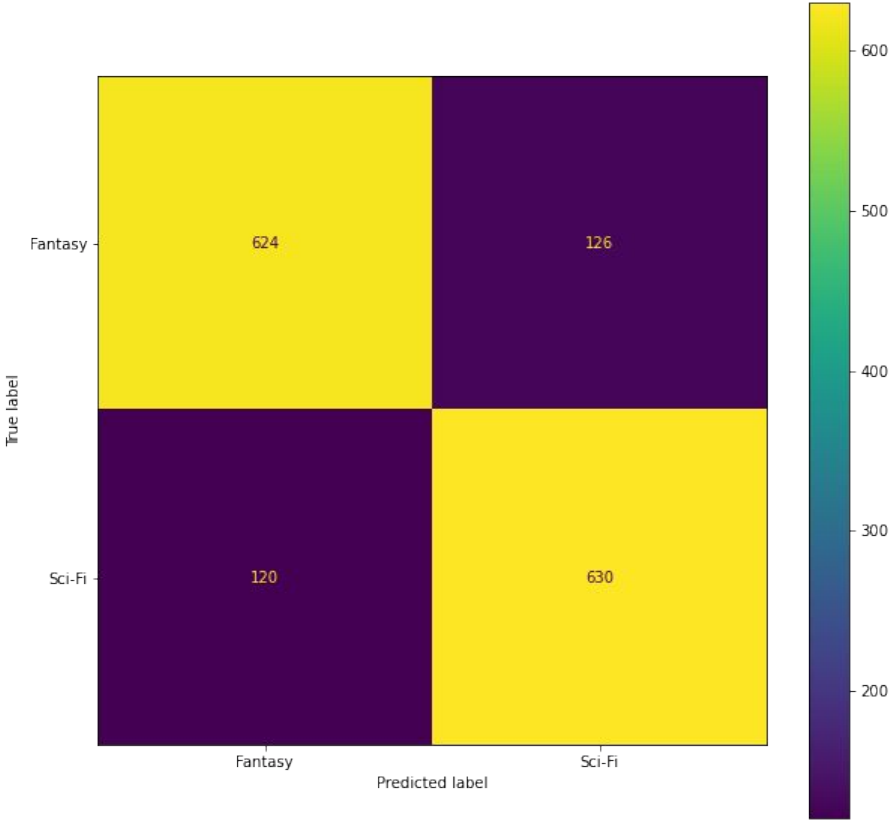
Models Used

- Logistic Regression
- Naive-Bayes: Bernoulli
- K-Nearest Neighbors
- Support Vector Classifier

Baseline Accuracy: 50%

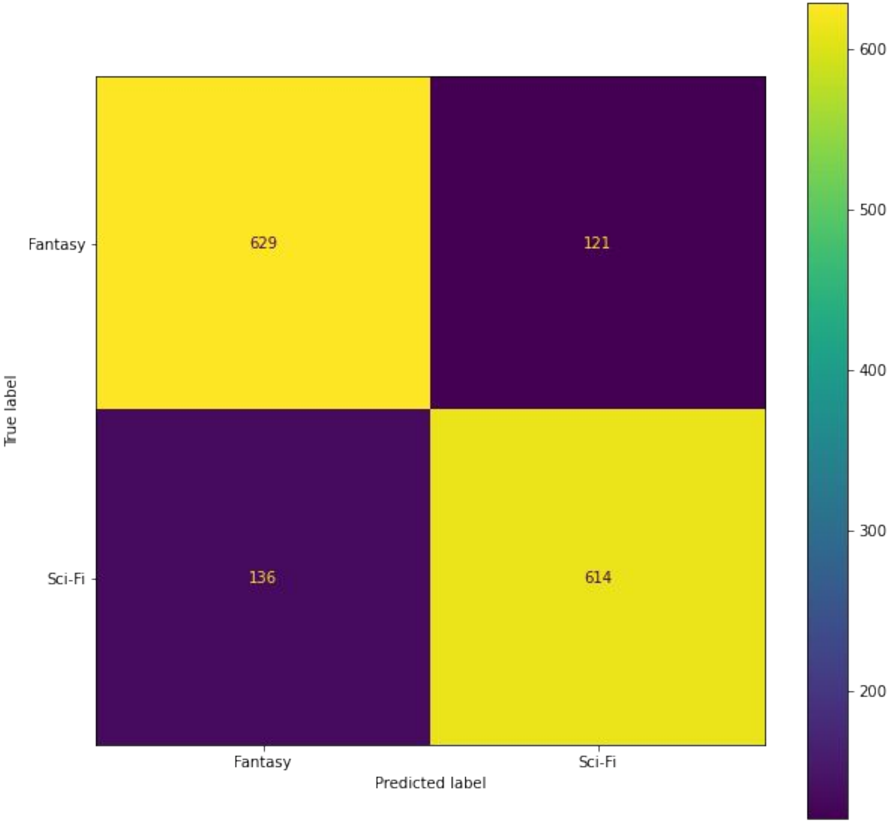
Modeling - MVP TF-IDF Logistic Regression

- Accuracy
  - 0.836
- Precision
  - Fantasy - 0.839
  - Sci-fi - 0.833
- Recall
  - Fantasy - 0.832
  - Sci-fi - 0.840
- F1
  - Fantasy - 0.835
  - Sci-fi - 0.837



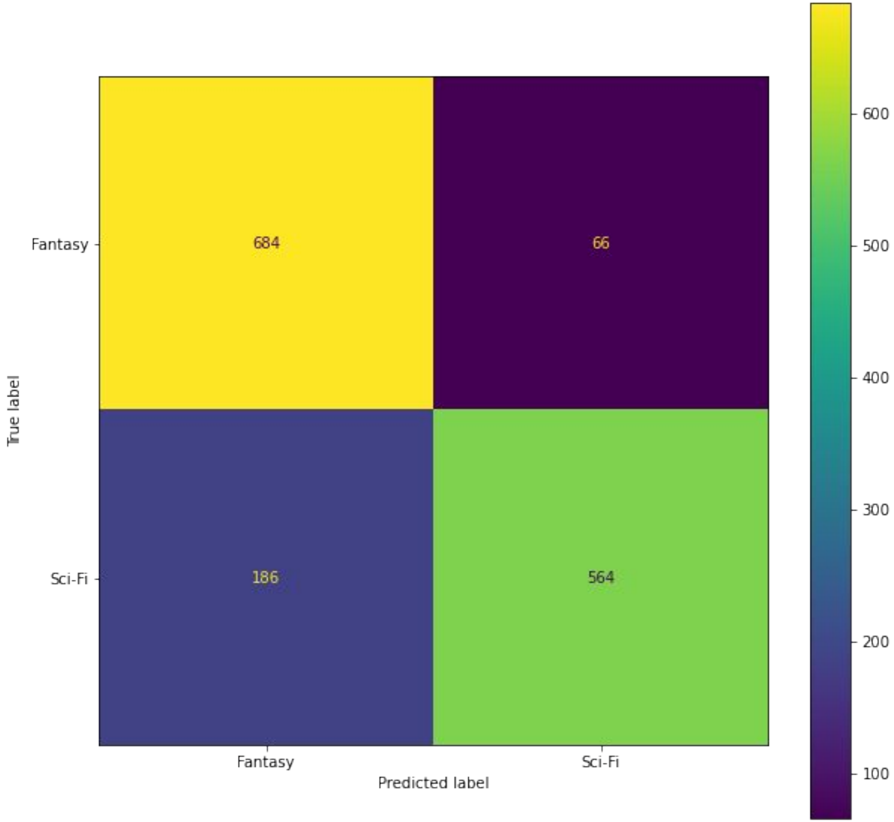
Modeling - Hypertune Count Vectorizer Logistic Regression

- Accuracy
  - 0.829
- Precision
  - Fantasy - 0.822
  - Sci-fi - 0.835
- Recall
  - Fantasy - 0.839
  - Sci-fi - 0.819
- F1
  - Fantasy - 0.830
  - Sci-fi - 0.827



Modeling - Grid Search TF-IDF Logistic Regression

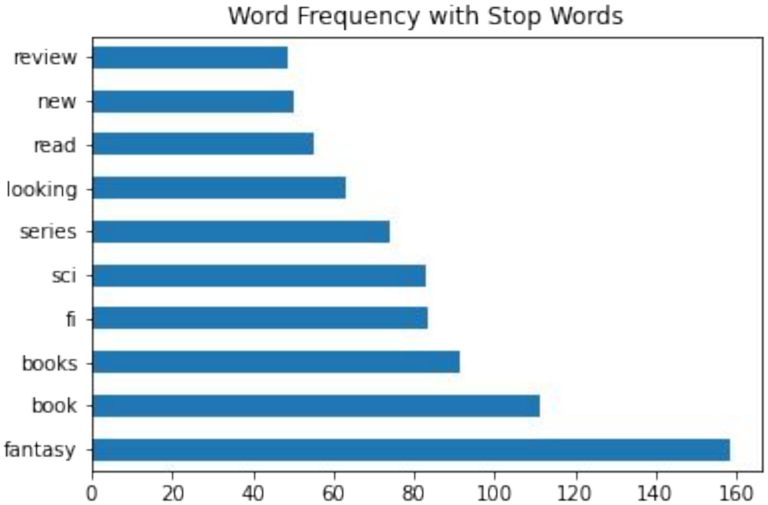
- Accuracy
  - 0.832
- Precision
  - Fantasy - 0.786
  - Sci-fi - 0.895
- Recall
  - Fantasy - 0.912
  - Sci-fi - 0.752
- F1
  - Fantasy - 0.844
  - Sci-fi - 0.817

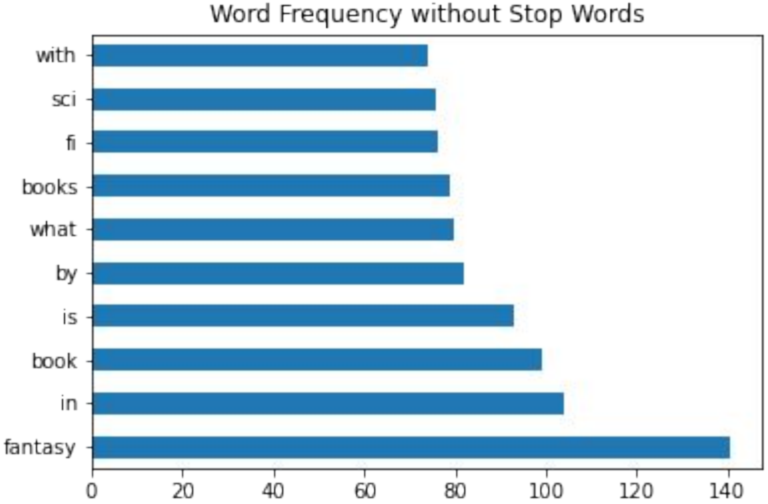


Examining Word Frequency

- Most Frequent Words
  - Fantasy
  - Book
  - Sci-fi

- The Matthew Effect
  - Voracious Readers Utilize Robust Lexicons


Word Frequency with Stop Words


Word Frequency without Stop Words

Conclusions

- Both subreddits focus primarily on books
  - Emphasis on writing

Future Paths

- Features to Engineer
  - Spelling Mistake? Part of Speech?
- Additional Subreddits
  - Subgenres of Fantasy and Sci-fi? Non-fiction?
- Analyzing vocabulary used in comments