USING GOODREADS DATA TO GET ON THE NYT BEST SELLERS LIST

SANCIA YANG

INTRODUCTION

- How can a new publishing company get their book onto the New York Times Best Sellers list?

METHODOLOGY

- Data:
    - Goodreads dataset - ratings and info on ~2 million books
    - NYT Best Sellers dataset - info of books on the list from Jan 2010 through Dec 2019

METHODOLOGY

- Tools
    - Pandas
    - Sklearn
    - Matplotlib

METHODOLOGY

- Feature Engineering
    - Number of ratings and rating distributions were converted into numerical features
- Models
    - Logistic Regression
    - K-nearest Neighbors
    - Random Forest

METHODOLOGY

- Accounting for class imbalance
    - Random Oversampling
    - Random Undersampling
    - SMOTE

METHODOLOGY

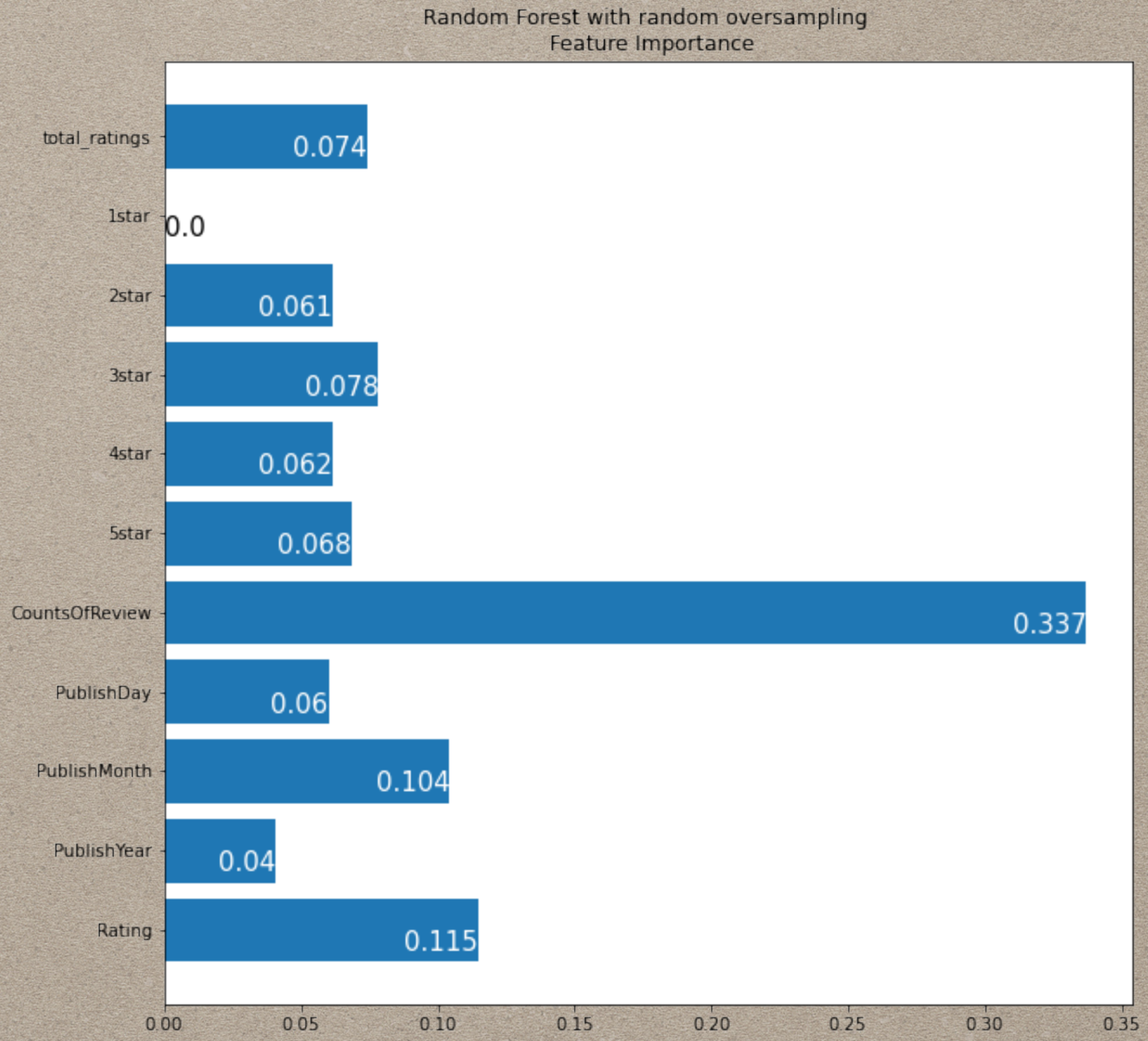
- Model evaluation
    - Dataset split into 80/20 (train, test)
    - Cross-validation
    - Measured on F1

RESULTS

Model	F1 score
Baseline logistic regression	0.0
Logistic regression with random oversampling	0.415
KNN with random oversampling	0.823
Random Forest with random oversampling	0.987
Simplified Random Forest with random oversampling	0.963

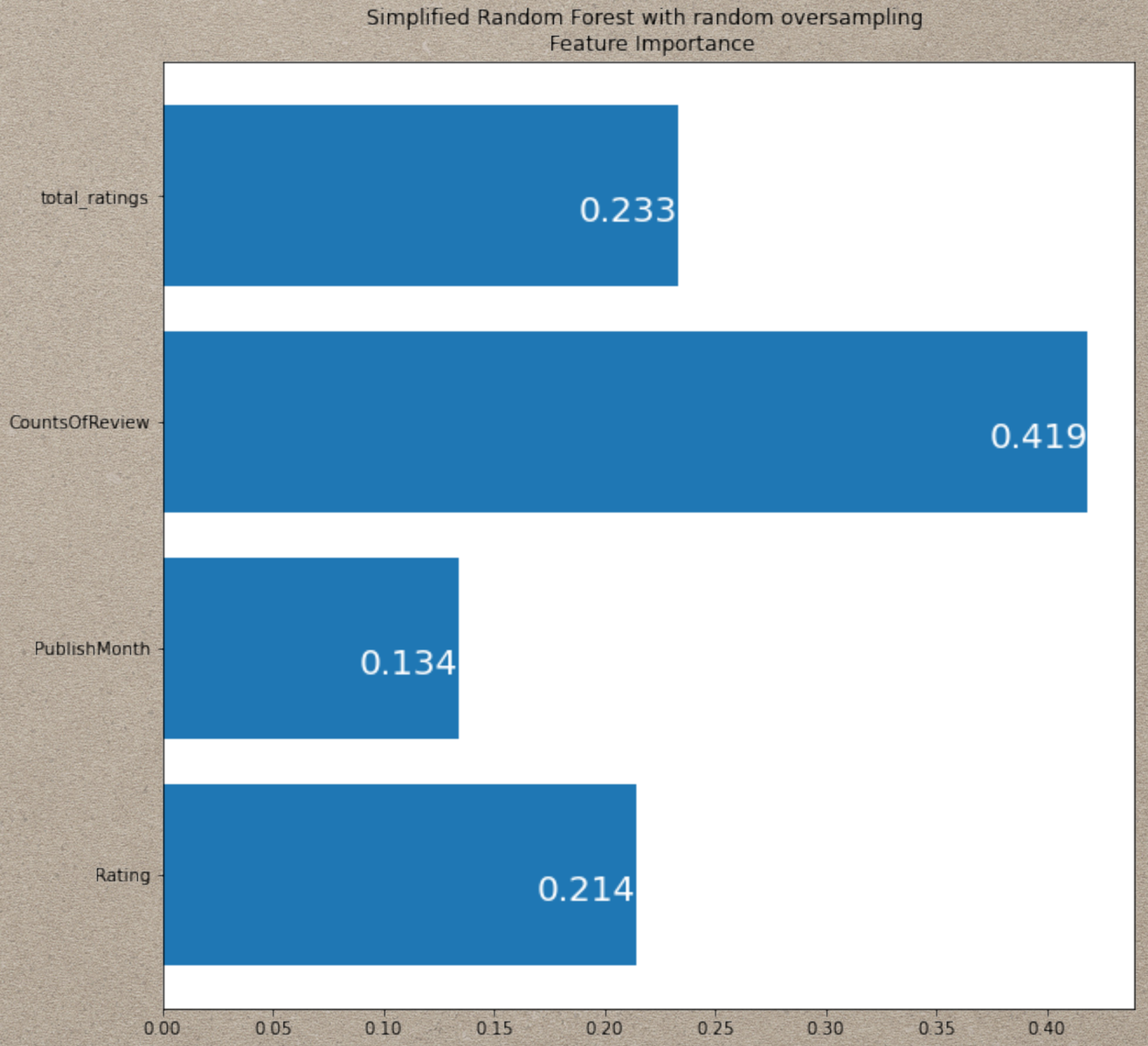
RESULTS

- Final model:
- Random Forest with random oversampling
  - F1: 0.987



RESULTS

- Final model:
- Simplified Random Forest with random oversampling
    - F1: 0.963



CONCLUSION

- Send advanced copies to (1) the top reviewers and (2) the most popular reviewers
- Aim for as many reviews as feasible leading up to publish date

FUTURE WORK

- Perform in-depth analysis on genres and subjects to determine possible imprints and manuscripts to pursue