Predicting Whether A Publisher will Publish A Book

PROPOSAL

This project was aimed at predicting whether a publisher will publish a book or not based on studying several factors. This is a **Multi-Class Classification** problem. We restrict ourselves to the top 4 (based on number of books they published) publishers in our data set.

BACKGROUND AND MOTIVATION

Motivation - Books help us to escape to a different world of imagination and increase our sense of the various things in life. Being said that, we were really eager to explore the world of books and came across the Book-Crossing Dataset and the Goodreads API. After exploring both the dataset, we decided to proceed with the above specified problem.
Background - To learn more about the data, we explored the Goodreads API and Book Crossing
To get to know about what exactly publishers look for, we read various articles [1], [2], [3]. Below are some of the main points that publishers look that are explicit (Abstract Ideas):
Idea - You must have a good and unique idea
Author Platform - Author should have proven ability to write
Manuscript - Writing must prove a quality product
Complicated Book - Too many characters/Too many pages or words
Marketability - Market analysis should indicate potential reader interest
Books Competition - Similar books in category

DATA DESCRIPTION

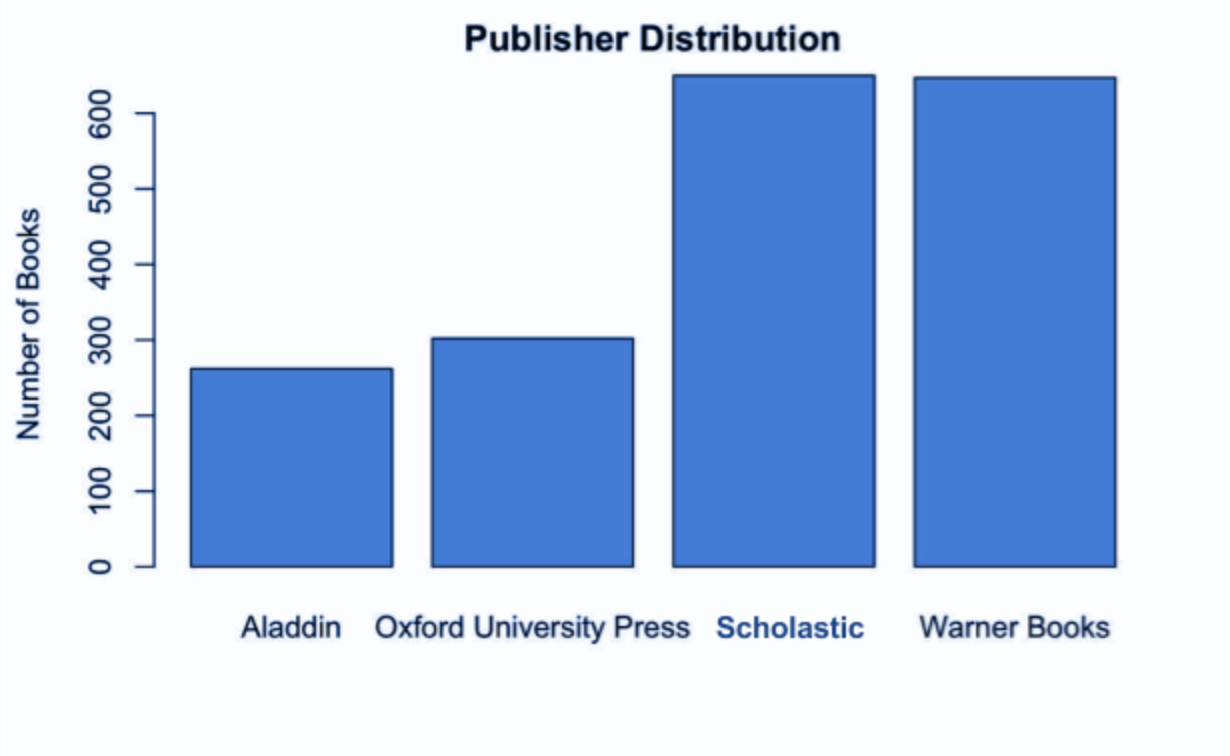
Data Sources: *Book-Crossing Dataset* and *Goodreads API*.

- Book-Crossing is an initiative of bookcrossing.com, a free online book club which was founded to encourage the practice, aiming to "make the whole world a library" (*CSV*)
- Goodreads API allows developers to freely search Goodreads' extensive user-populated database of books, annotations, and reviews (*XML Responses*)

Attribute Name	***Description***	***Type***
ISBN	International Standard Book Number	TEXT
Title	Title of the book	TEXT
Description	Description of the Book	TEXT

Publication Year	Year of Publication	NUMERIC
Publisher	Publisher of the Book	TEXT
Number of Pages	Total Pages in the Book	NUMERIC
Average Rating	Ratings given by GoodReads Users	NUMERIC
Authors	Authors of the Book	LIST
Author Followers	Total Author Followers on GoodReads	NUMERIC

Total number of books considered: **1861**
Aladdin - **262**
Oxford University Press - **302**
Scholastic - **650**
Warner Books - **647**



The Bar chart above shows the total books published by each publisher (after all preprocessing). *Aladdin* and *Oxford University Press* has the least books published (data) among 4 publishers

Data Cleaning on original dataset

- Merged Bookcrossing and Goodreads API
- Removed all books with no Publisher and Description
- Corrected Discrepancies in Publisher and Author Names

Possible Confounders to the problem are mentioned below. :

1. Author rating
2. Author works count
3. Sentiment Score

The Feature Importance graph in Model Section shows how these affect the dependent variable.

MODEL DESCRIPTION (What we Did)

Predictors: The features considered

1. **Author Rating** : Author rating is the average rating of an Author which we calculated by getting the ratings of previous books, which we got from the goodreads API, of the author and taking their average.
2. **Number of Pages**: This is a valuable feature as the industry standards should be met. We got this from the Goodreads API.
3. **Author works count**: Work count gives the total number of books that have been previously published of the particular author.
4. **Author fans count**: The author following is specified by this feature. It tells us how popular the author is.
5. **Sentiment score**: The score feature was calculated by taking the sentiment score of description of each book. This helps us in identifying whether a book is negative or positive.
6. **Number of Characters**: This feature was obtained by named entity recognition technique. It helps in deciding whether a book is too complicated to understand with too many characters in it.
7. **Novelty**: To incorporate this as a feature, we calculated TF-IDF Scores of all terms in the Book Description. Then, cosine similarity was taken from each book to every other book. Books with similarity > 0.90 were considered to have similar idea, and thus not novel.
8. **N-grams**: Uni, Bi and Tri Grams on Description of the book

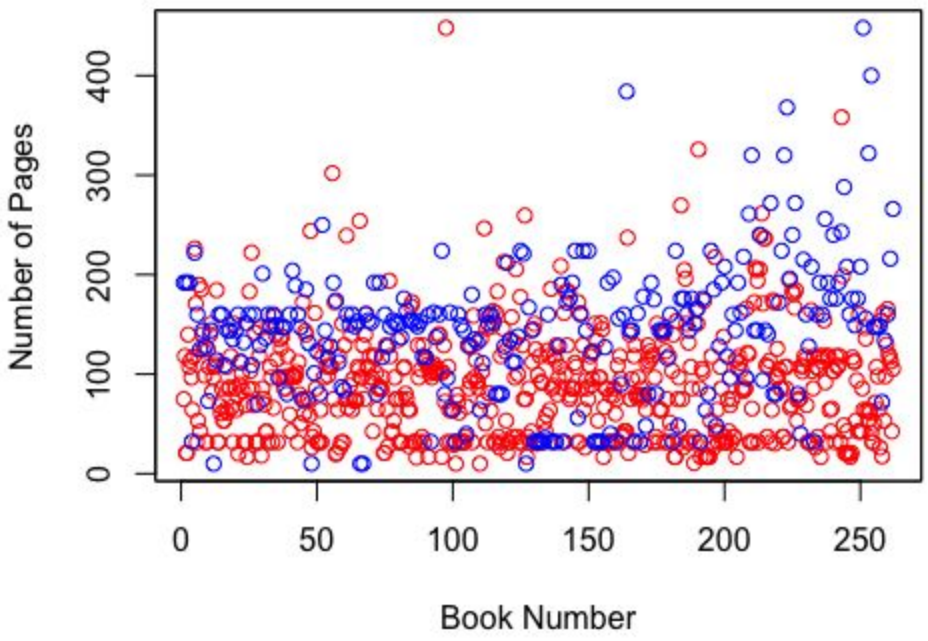
Outcome: One of the 4 Publishers who is likely to publish the book

Models tried:

1. Random Forest (models non-linear relationship well)
2. SVM (nonlinear relationship can be modelled well using Kernels)
3. One Class SVM (We need this since there is no negative data available)

Below curve depicts the following:

- Blue Points are for publisher *Oxford University Press*
- Red Points are for publisher *Aladdin*
- y-axis depicts the Number of Pages feature for both publishers
- x-axis denotes the Book Number (Not Important)
- We see that there is no Linear Relationship between positive and negative data



Assumption - For the first 3 models, we took the assumption that if a publisher P1 publishes a book B then we assume that all other publishers have rejected B.
We get negative training data with this. This assumption might not be valid in all scenarios

Random Forests

Random forests can be used to rank the importance of variables in a classification problem in a natural way. Random forests are a way of averaging multiple deep decision trees, trained on different parts of the same training set, with the goal of reducing the variance.

Support Vector Machines

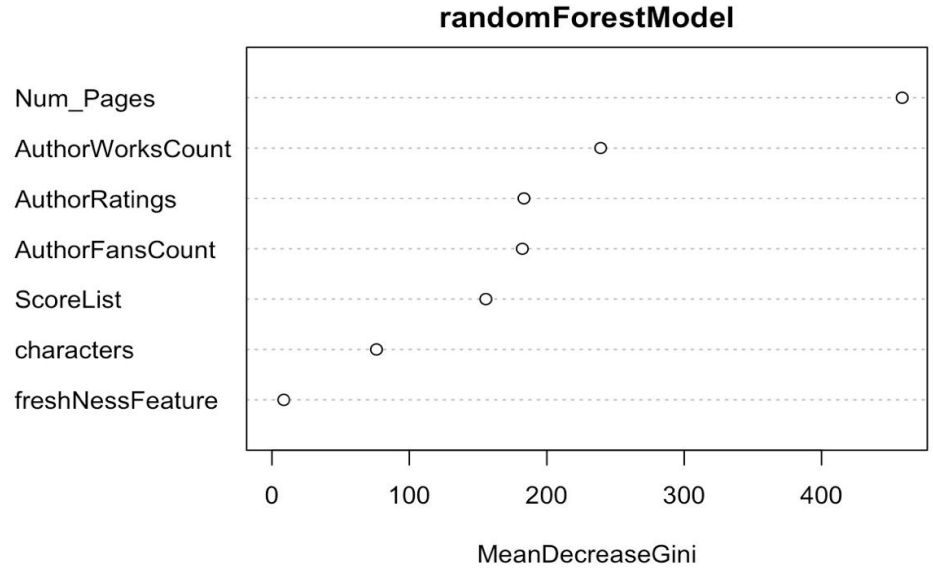
SVM is a robust model for the classification problem. The Kernel features in SVM model Nonlinear Relationships really well. We used this model because of its Kernel Feature

Support Vector Machines with One class

Since for all the models above their was an underlying assumption that we have negative training data. This assumption might not be always true. So we experimented with this model.

EVALUATION & PERFORMANCE

Feature Importance Plot (Using Random Forests)



From the above graph, we can clearly see that Author Ratings and Author Fans count are **confounding variables** as they are not related to the problem directly but still affect the result.

Performance of Models

We used **10-fold Cross-Validation Technique** to evaluate our model. Accuracy is the average accuracy given when a 10-fold cross validation is ran on a dataset.

MODEL	ACCURACY
SVM with Linear Kernel	65.66362
SVM with Radial Kernel	69.58624
SVM with Polynomial Kernel	67.32939

SVM with Sigmoid Kernel	51.10156
SVM with Radial Kernel (N-Grams as Features)	61.213434
Random Forests	62.32332

Accuracy with **One class SVM** for each of the following publishers:

Warner Books	49.45904
Scholastic	50.30769
Oxford University Press	50.66225
Aladdin	48.47328

CONCLUSION

- Number of Pages is one of the most distinguishing feature for this problem
- Author Credentials (His Writing Capability, His Followers) influence publishers too.
- SVM With Radial Kernel works the best among other models. This confirms our observation that their is a nonlinear relationship among features
- One Class SVM works quite well. This is important since assumption of negative training data might not be true always.

FUTURE WORK

- Getting Genre Information into data. This is important since many publishers are inclined towards some genre.
- Converting the idea of Marketability as a feature. All the publishers look for profits to which marketability of a book is very important.
- Also, Novelty feature which we tried didn't work that well. Further improvements to this feature can be very useful

REFERENCES

Articles explored:

- http://www.theadventurouswriter.com/blogwriting/17-reasons-book-manuscripts-are-rejected/
- http://www.writersdigest.com/online-editor/do-you-have-what-publishers-really-want
- http://www.writetosellyourbook.com/blog/2013/09/16/what-do-publishers-want/