CS 418 - Final Project
Rating Analysis and Prediction For Books

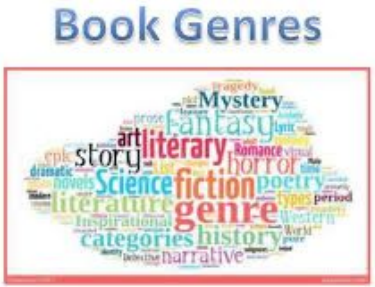
Team:
Anusha Voloju
Maleeha Ahmed
Priyan Sureshkumar

Problem Specification

The goal of this project is to analyse the important factors that make a book more popular and most read compared to others and to use this analysis to predict the average rating of a book and to categorize the books into corresponding genres.

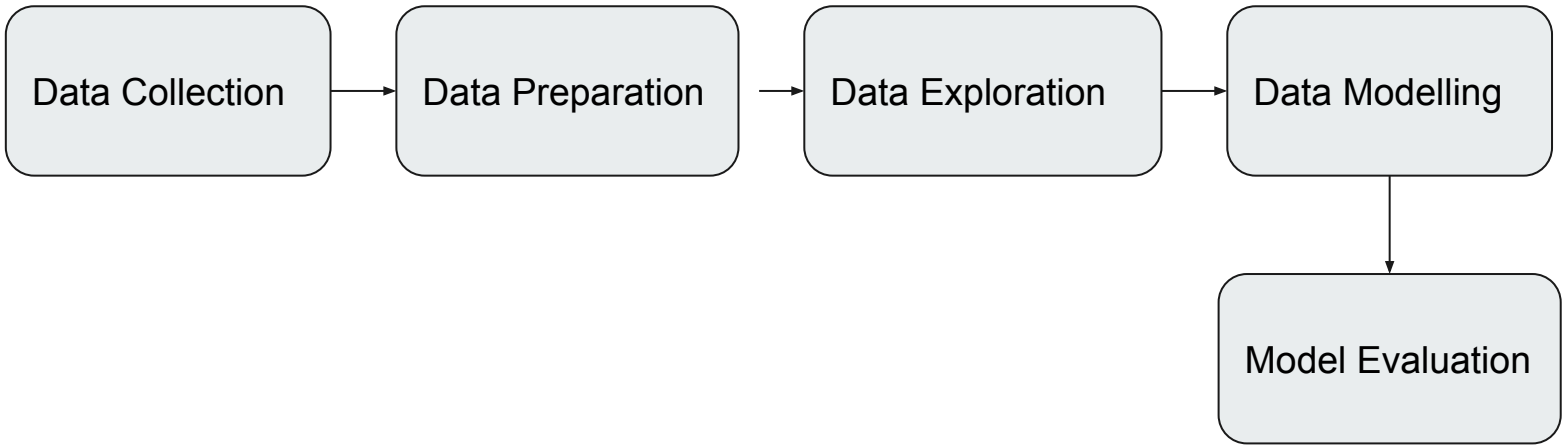
Proposed Solution

The important features are extracted using descriptive statistics and are used to build a regression model for the prediction of average rating of the books.



Using classification or clustering techniques, the books are categorized into specific genres.

Data Science Pipeline



Data Collection

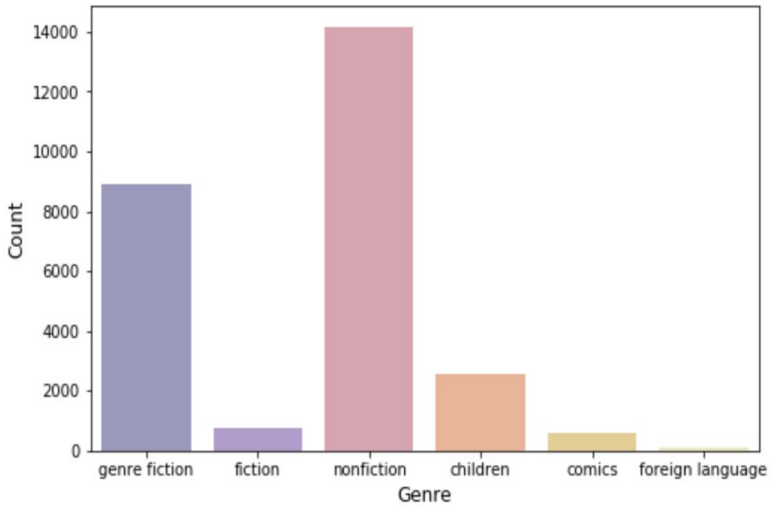
We are using Publishers dataset from CORGIS datasets.
The variables in the dataset include:
genre, sold_by, daily_average_amazon_revenue, daily_average_author_revenue, daily_average_gross_sales, daily_average_publisher_revenue, daily_average_units_sold, publisher_name, publisher_type, average_rating, sale_price, sales_rank, total_reviews

	genre	sold by	daily average.amazon revenue	daily average.author revenue	daily average.gross sales	daily average.publisher revenue	daily average.units sold	publisher.name	publisher.type	statistics.average rating	s
0	genre fiction	HarperCollins Publishers	6832.000	6832.000	34160.00	20496.000	7000	Katherine Tegen Books	big five	4.57	
1	genre fiction	HarperCollins Publishers	2487.500	2487.500	12437.50	7462.500	6250	HarperCollins e-books	big five	4.47	
2	genre fiction	Amazon Digital Services, Inc.	9559.000	9559.000	47795.00	28677.000	5500	(Small or Medium Publisher)	small/medium	4.16	
3	fiction	Hachette Book Group	8250.000	8250.000	41250.00	24750.000	5500	Little, Brown and Company	big five	3.84	
4	genre fiction	Penguin Group (USA) LLC	7590.500	7590.500	37952.50	22771.500	4750	Dutton Children's	big five	4.75	

Data Preparation

1. No Missing Values
2. Valid Zero Values (average rating and total reviews)
3. Genres distribution

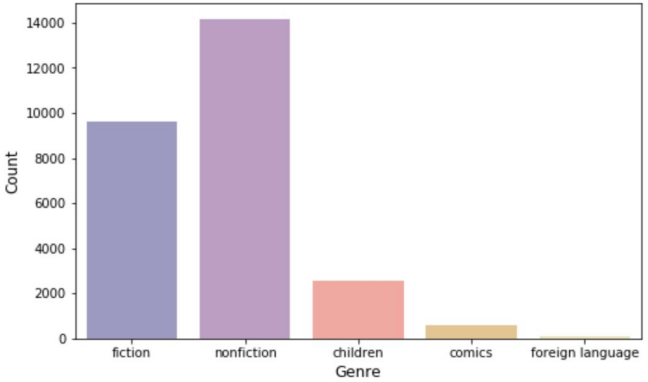
genre	
children	9.401709
comics	2.101602
fiction	2.712103
foreign language	0.447700
genre fiction	32.941133
nonfiction	52.395752



Data Preparation

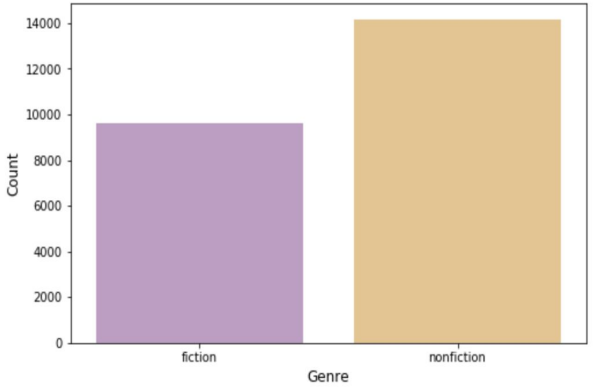
Combine Genres

genre	
children	9.401709
comics	2.101602
fiction	35.653236
foreign language	0.447700
nonfiction	52.395752



Filter Genres

genre	
fiction	40.492499
nonfiction	59.507501



Data Preparation

4. Categorical Attributes

   sold_by and publisher_name are dropped

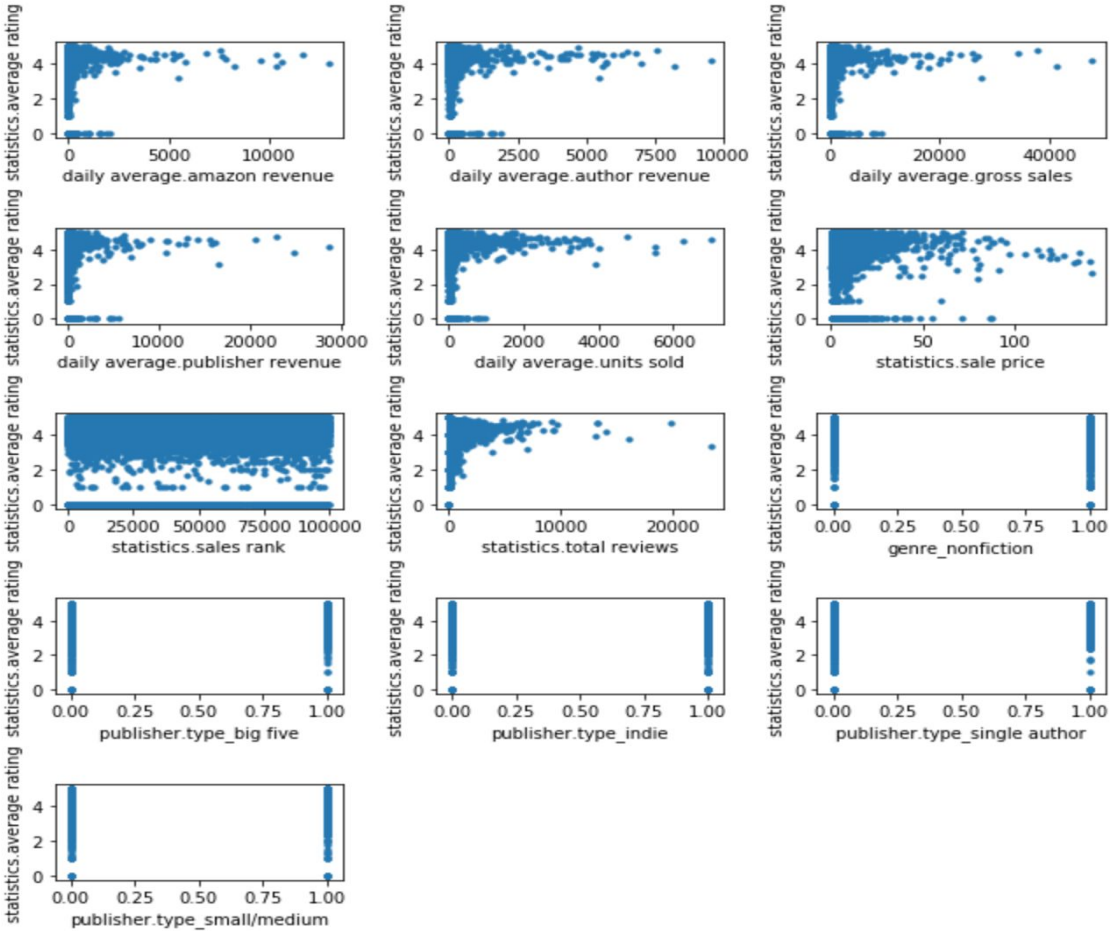
   Dummy variables created for publisher_type

amazon
big five
indie
single author
small/medium

publisher.type_big five	publisher.type_indie	publisher.type_single author	publisher.type_small/medium
1	0	0	0
1	0	0	0
0	0	0	1
1	0	0	0
1	0	0	0

Data Exploration

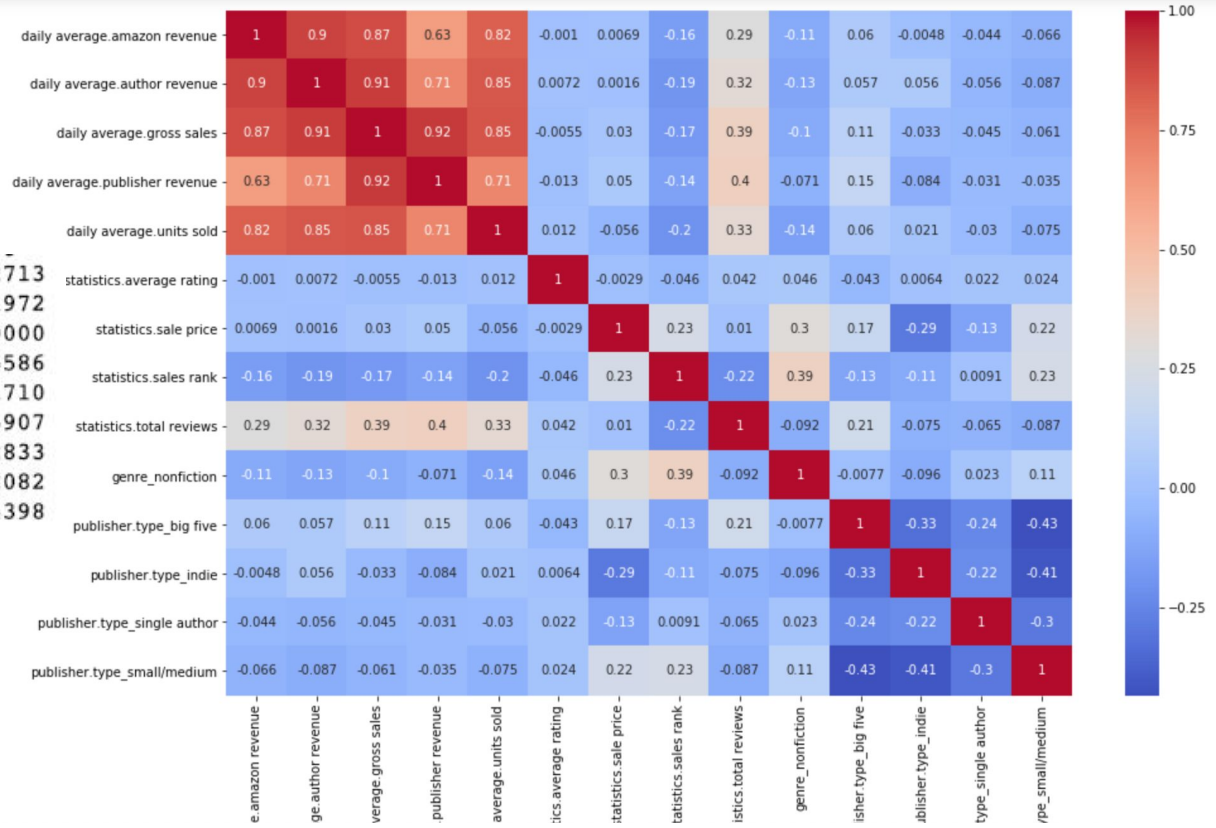
1. Scatter plots



Data Exploration

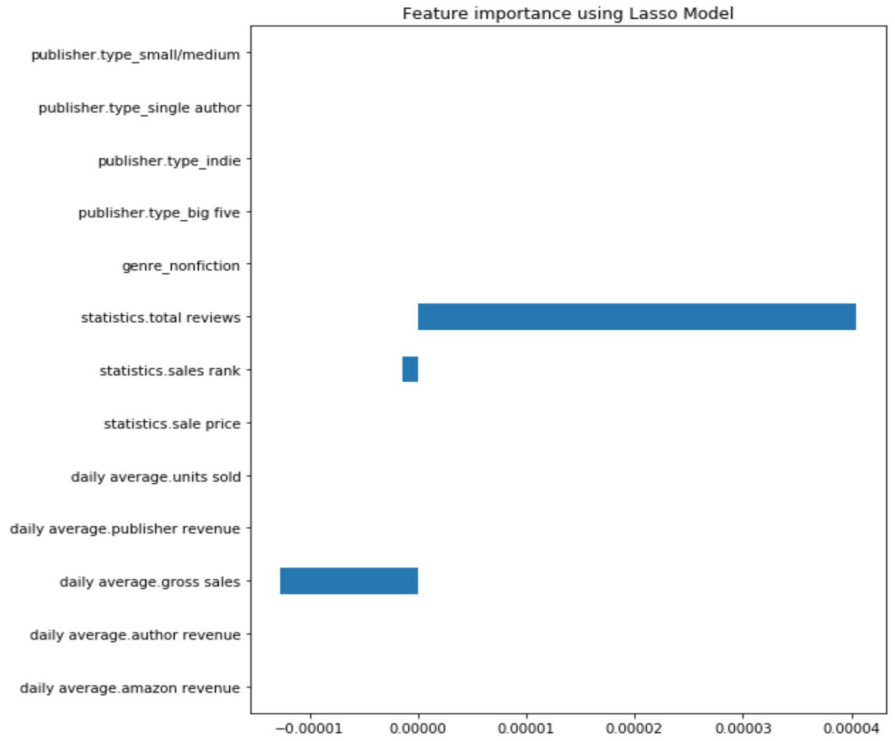
2. Correlation Matrix

```
daily average.publisher revenue    0.012713
daily average.units sold           0.011972
statistics.average rating          1.000000
statistics.sales rank              0.045586
statistics.total reviews           0.041710
genre_nonfiction                   0.045907
publisher.type_big five            0.042833
publisher.type_single author       0.022082
publisher.type_small/medium        0.024398
```



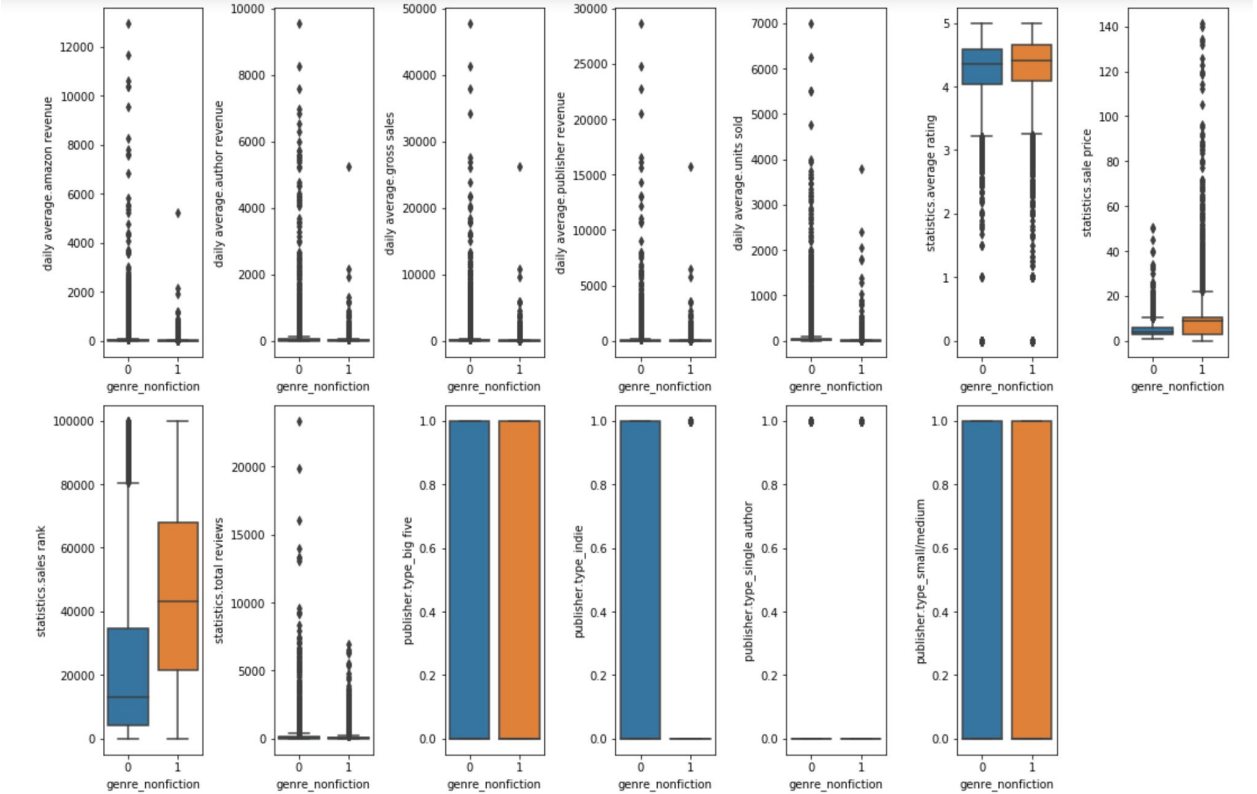
Data Exploration

3. Feature Importance using LassoCV



Data Exploration

4. Box Plots (genres)



Data Modelling

Regression

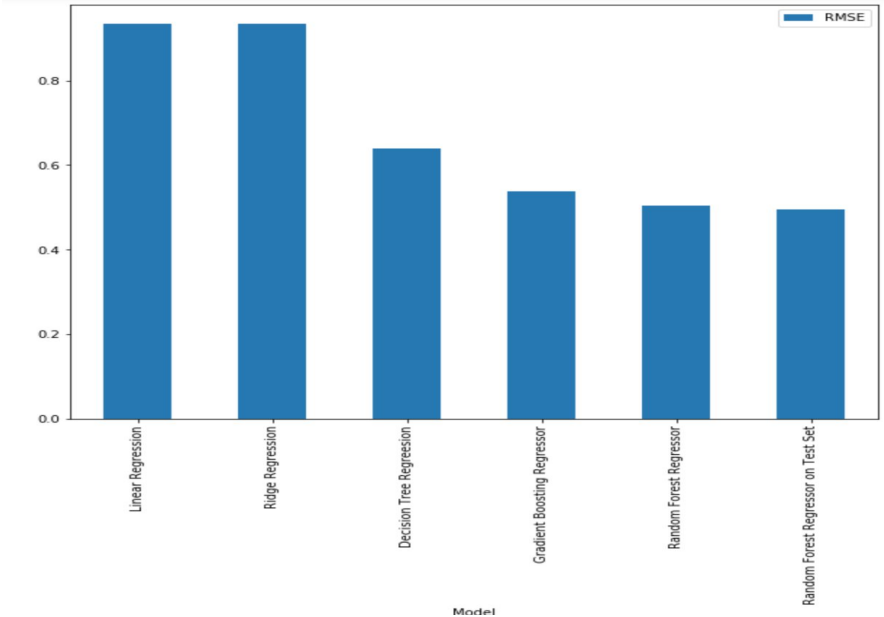
Best Model :
Random Forest Regressor

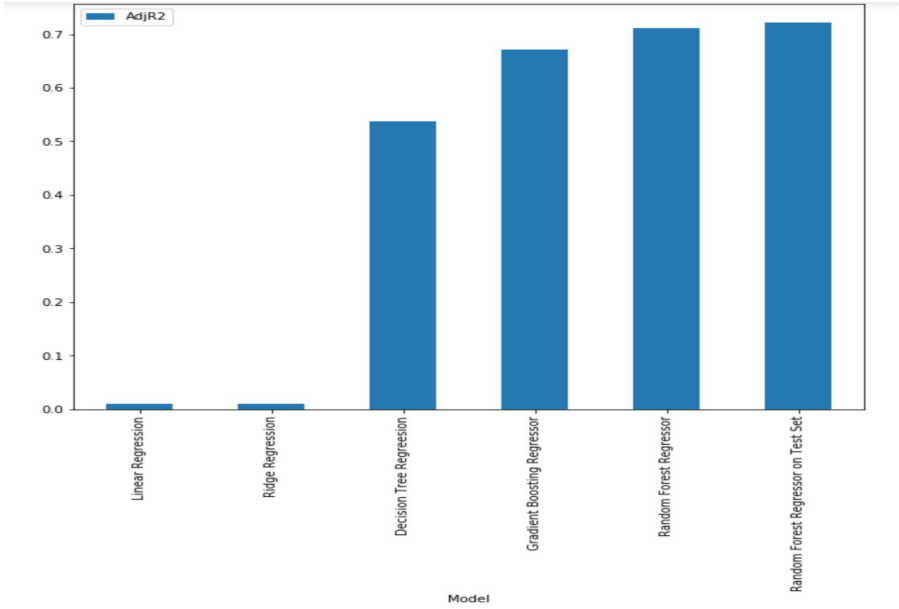
Best Features:
Sales rank,
Total reviews,
Genre,
publisher.type_big five,
publisher.type_single author,
publisher.type_small/medium

Model	R Square	Adjusted R Square	Root Mean Square Error
Linear Regression Model	0.012	0.010	0.939
Ridge Regression Model	0.012	0.010	0.939
Decision Tree Regressor Model	0.536	0.536	0.639
Gradient Boosting Regressor Model	0.672	0.672	0.537
Random Forest Regressor Model	0.711	0.711	0.504
Random Forest Regressor Model On Test Set	**0.721**	**0.721**	**0.495**

Data Modelling

Performance of Regression Models





Classification

Models -

- Basic Decision Tree using the entire feature set
- Basic Decision Tree using a subset of features
- **Bagging Ensemble Method with Decision Tree as a base estimator [Best Model]**
- Adaboost Ensemble Method with Decision Tree as a base estimator
- Random Forest Ensemble Method with Decision Tree as a base estimator

Data Preparation - Used Scaled Features. Used a subset of features for one model.

Results - Most models came within the 70% bracket for Accuracy , Precision and F1 Score.

Analysis - All models were able to predict genre to around a 70% effectiveness even given our data set’s features.

Clustering

Models - Hierarchical Single, Hierarchical Complete, Hierarchical Average, K-Means, DBSCAN

Distance Metrics - Euclidean, Manhattan, Minkowski, Cosine

Data Preparation - Scaling on Features used. Testing on different subsets of data.

Results -

- Low silhouette scores and lower Adjusted Rand. Indices.
- Many singular large clusters.
- DBSCAN - 8-10 clusters

Analysis - Values of features common throughout many different types of books - price range, units sold etc. Clusters not ideal for our purposes.


Thank
you!