Amazon Bestsellers Analysis

Authors: Eric Luong, David Chien, Siva Sai Chandra Annepu, Sri Krishna Chaitanya Kommalapati

Overview

Amazon started out as a bookstore and rapidly expanded into one of the largest ecommerce monopolies in the world.

Situation: Explore the relationship between book attributes and their popularity, as measured by the number of reviews and user ratings. From the publisher's perspective, this could be useful in minimizing the losses from picking up non-performing books/authors.

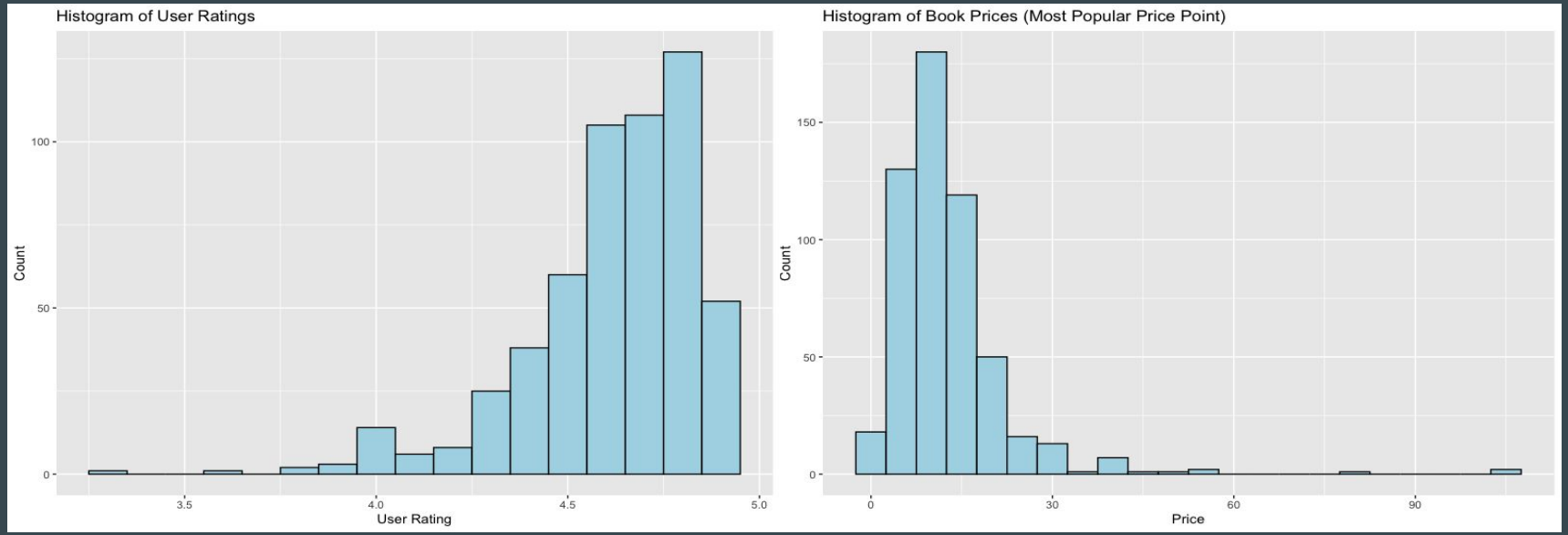
Data

- Dataset has 550 rows with 7 columns.
- No duplicated rows
- No null values
- Contains numerical and categorical data
- Variables:- Name, Author, User Ratings, Reviews, Price, Year, Genre

	Name	Author	User.Rating	Reviews	Price	Year	Genre
1	10-Day Green Smoothie Cleanse	JJ Smith	4.7	17350	8	2016	Non Fiction
2	11/22/63: A Novel	Stephen King	4.6	2052	22	2011	Fiction
3	12 Rules for Life: An Antidote to Chaos	Jordan B. Peterson	4.7	18979	15	2018	Non Fiction
4	1984 (Signet Classics)	George Orwell	4.7	21424	6	2017	Fiction
5	5,000 Awesome Facts (About Everything!) (National Geographic Kids)	National Geographic Kids	4.8	7665	12	2019	Non Fiction
6	A Dance with Dragons (A Song of Ice and Fire)	George R. R. Martin	4.4	12643	11	2011	Fiction

Exploratory Data Analysis

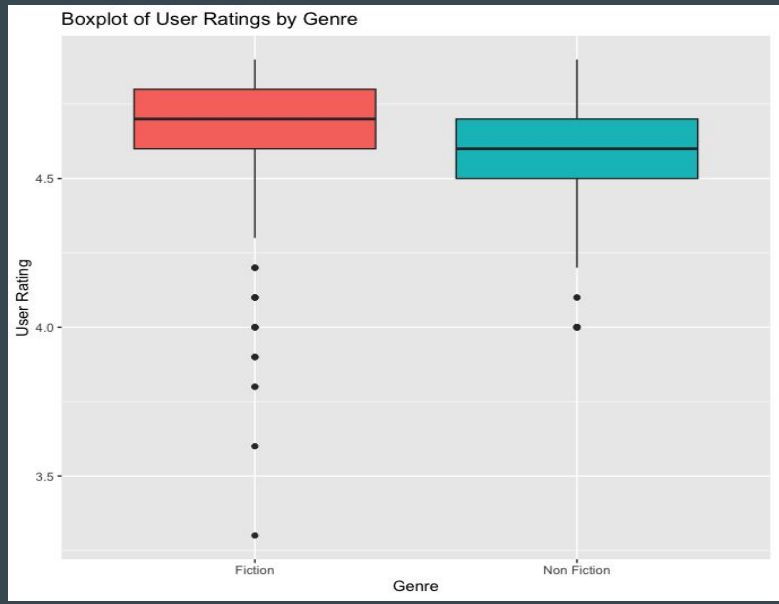
Histograms Of User Ratings And Book Prices

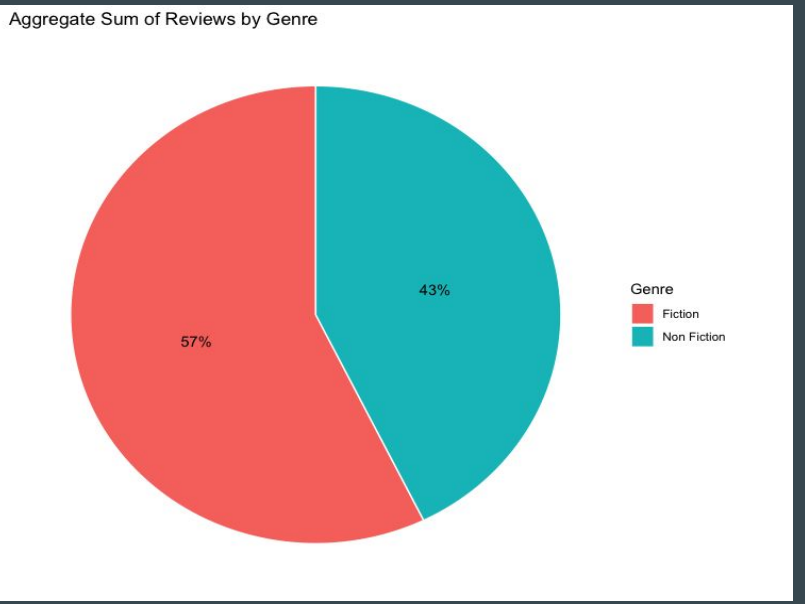


- Left skewed
- Right Skewed

- Use this information to benchmark their books against others in the market and to set realistic expectations for the ratings of their books
- Identify areas for improvement in their books to increase their appeal to readers

User Ratings and Reviews By Genre

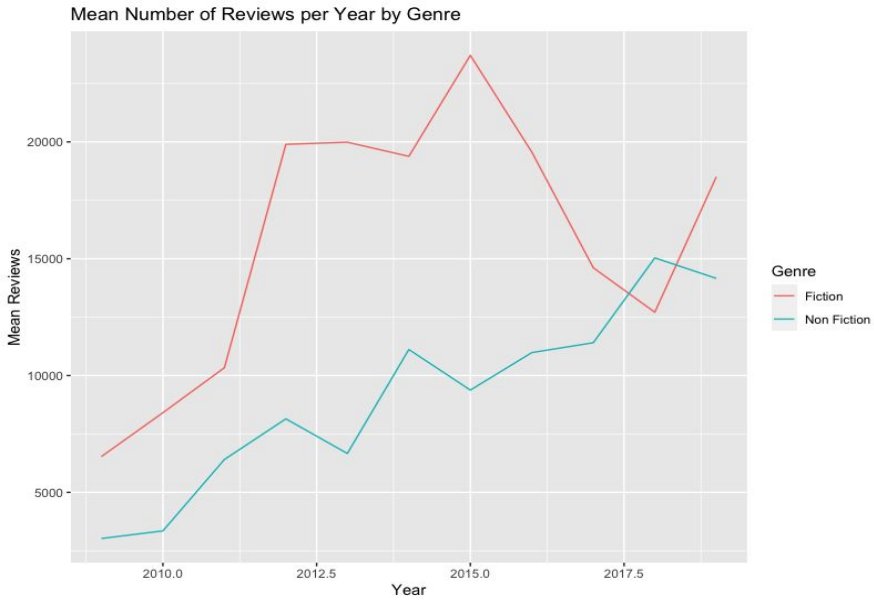


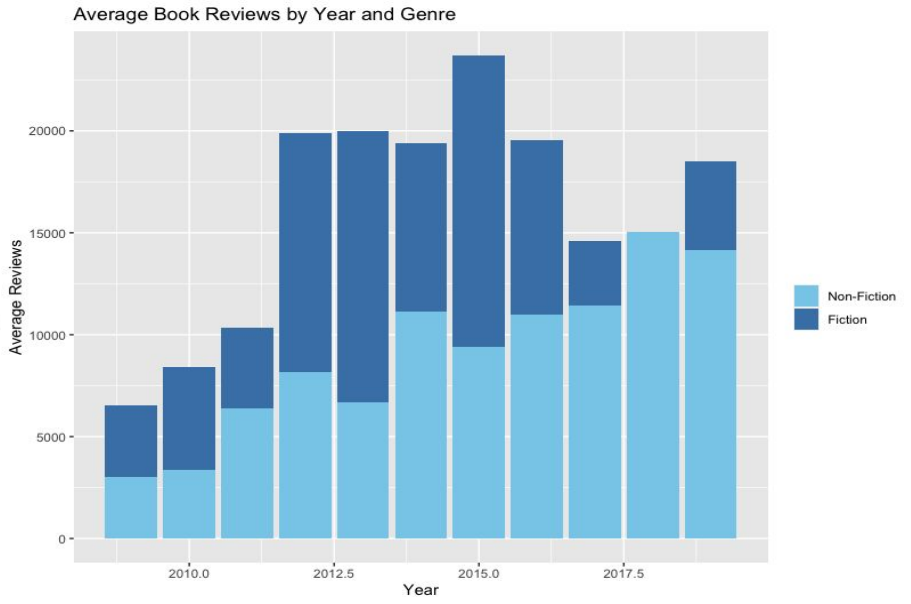


- Spread of user ratings for "Non-Fiction" books is wider than "Fiction" .
- Publisher might focus on "Fiction" genre.

- Publisher may want to consider focusing on publishing more fiction books as they tend to have higher ratings and number of reviews on average
- However, the spread of ratings is wider for non-fiction books, which suggests that there may be a larger market for these types of books as well.

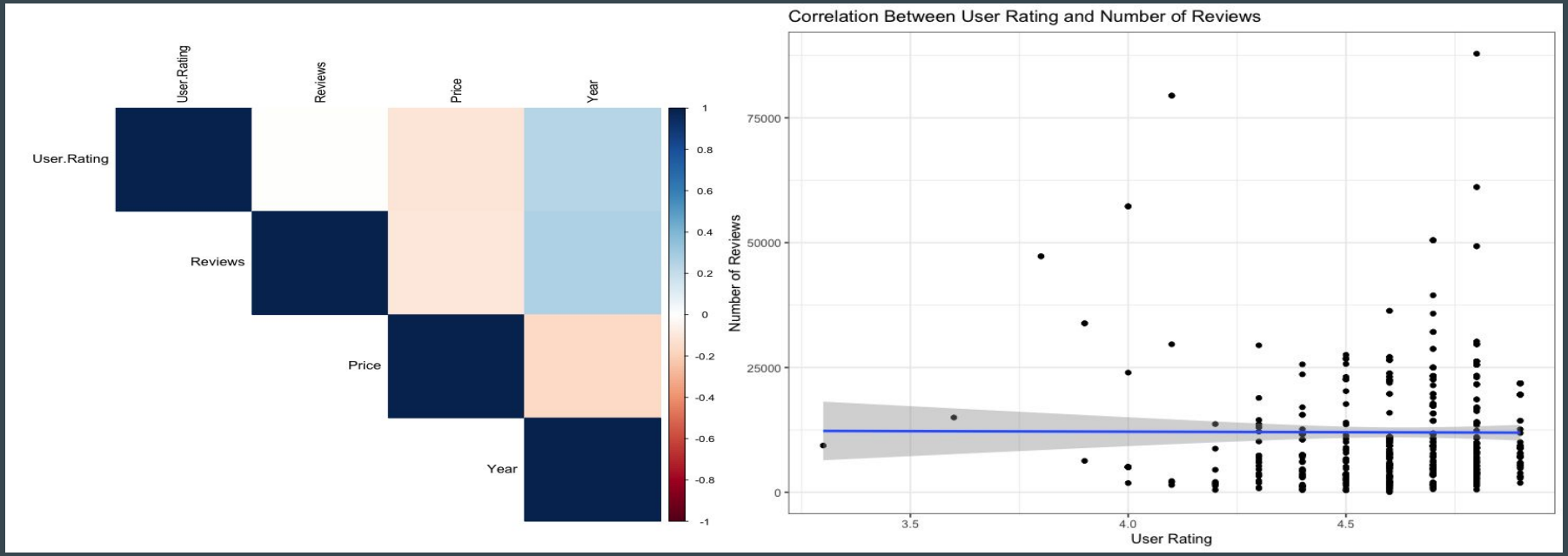
Average Reviews Over The Year By Genre





- Mean of the reviews peaked after 2014 for both genre
- On average the aggregate reviews of fiction genre is greater than that of non-fiction genre

Correlation Among Features



- Weak positive correlation between user ratings and year
- Weak negative correlation between price of the book and year
- Confirms weak correlation between variables number of reviews and user ratings

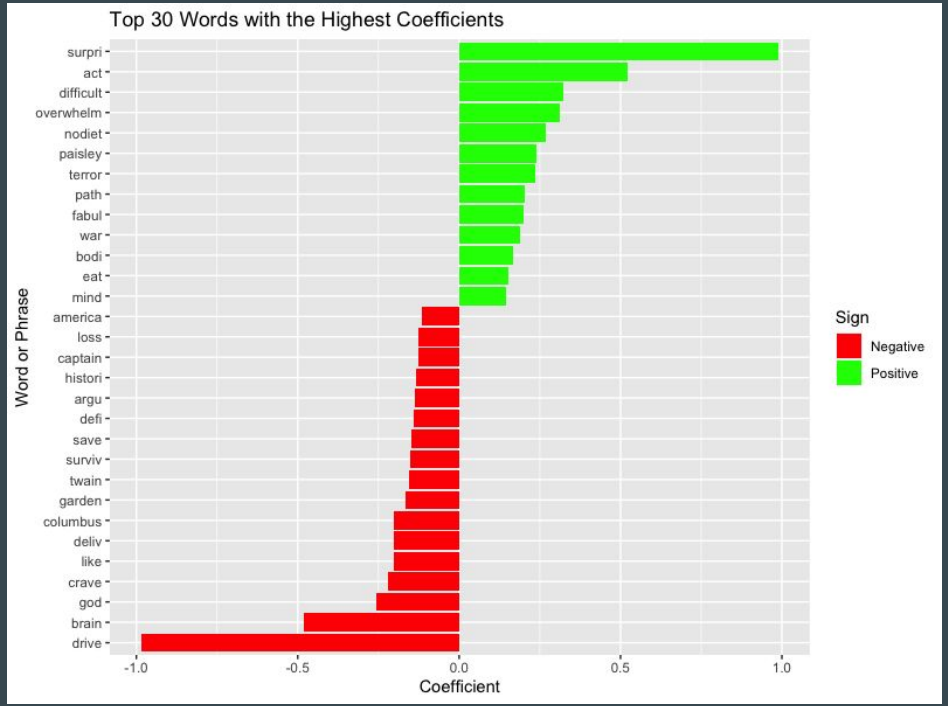
Overview of Machine Learning Methods

3 Models

- NLP & Regression Analysis of Titles
- K-Means Clustering
- Generalized Linear Model (GLM)

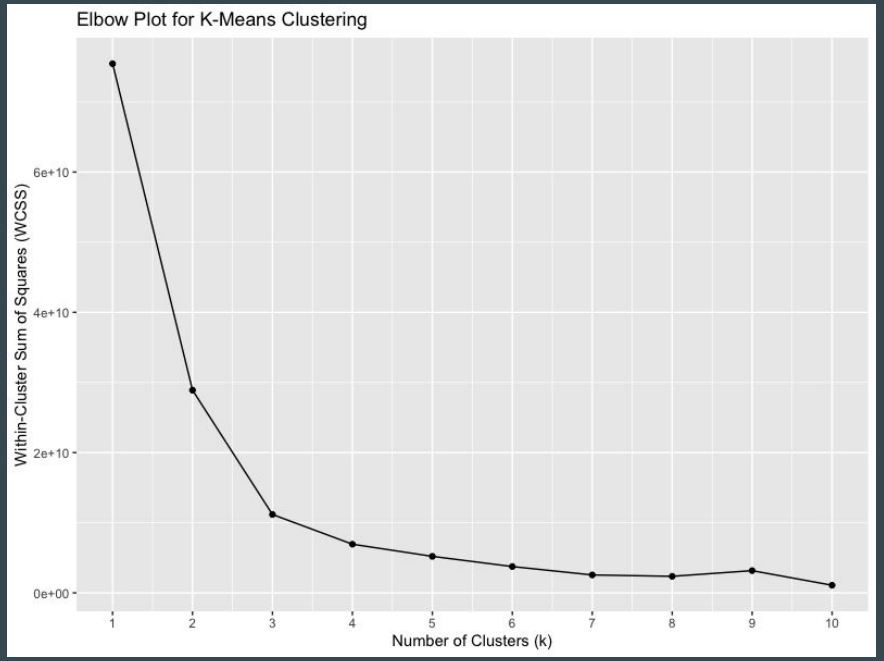
NLP & Regression

- Preprocess Titles (Remove stopwords, punctuation, etc., Tokenize/Stem, Create document term matrix for bag-of-words)
- Train linear model on bag-of-words features with number of reviews
- Analyze coefficients for best and worst words to use in titles



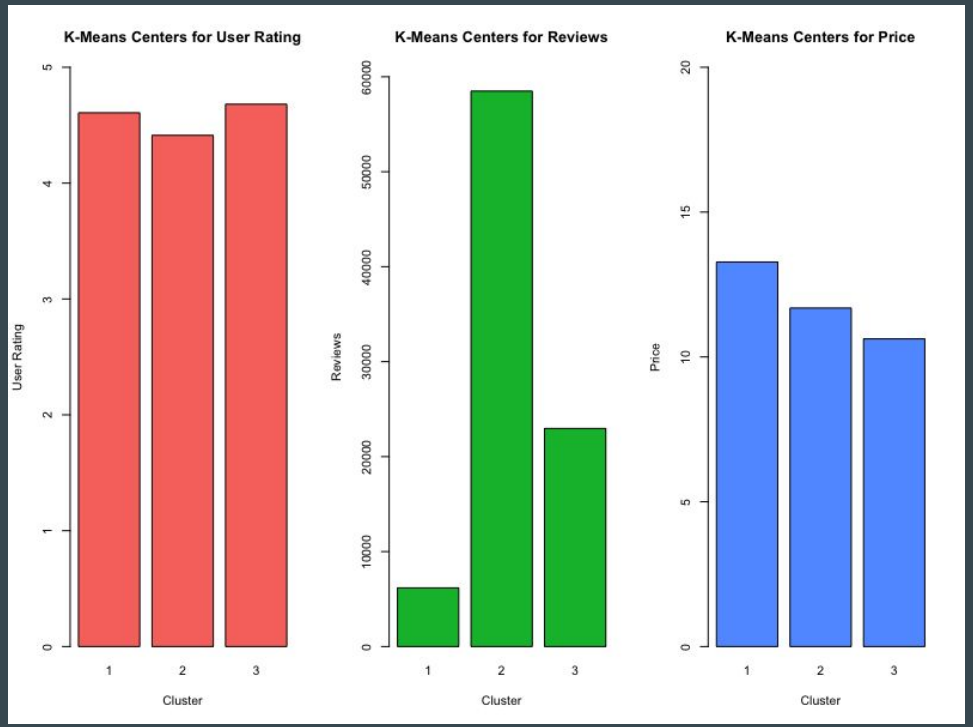
K-Means Clusters

- Elbow Plot of Within-Cluster Sum of Squares vs # of Clusters
- K-Means Clustering for User Rating, Reviews, Price
- Visualize Cluster Centers and Distributions



K-Means Results/Findings

- Cluster 1 represents least popular books, predominantly non-fiction
- Cluster 2 represents the bestselling books with medium user ratings
- Cluster 3 represents niche books that sell well but are critic favorites



Generalized Linear Model (GLM)

- Allows you to Fit a wide variety of regression models

- Split and train the dataset with the Bestseller variable as the response and the User.Rating, Reviews, Price, Year, and Fiction variables as predictors.

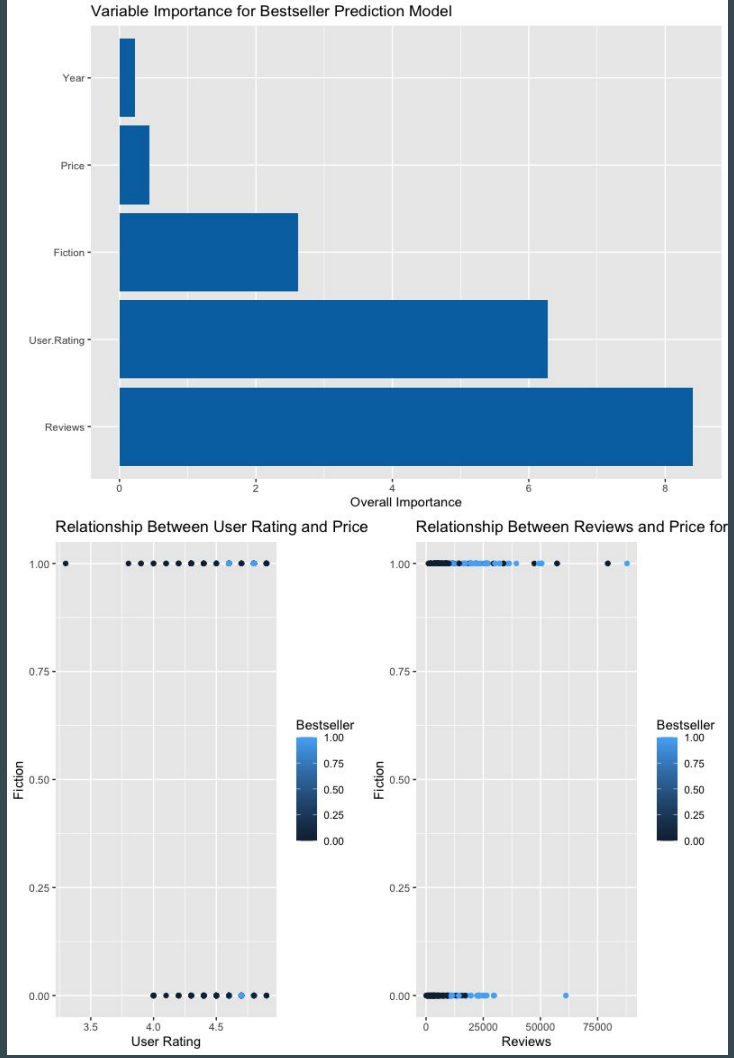
- Make predictions on the test set and use confusionMatrix() to show Accuracy.

```
Confusion Matrix and Statistics


predicted_classes   0   1
                0 100  11
                1   5  47

               Accuracy : 0.9018
                 95% CI : (0.8455, 0.9428)
    No Information Rate : 0.6442
    P-Value [Acc > NIR] : 3.802e-14

                  Kappa : 0.7808

 Mcnemar's Test P-Value : 0.2113

            Sensitivity : 0.9524
            Specificity : 0.8103
         Pos Pred Value : 0.9009
         Neg Pred Value : 0.9038
             Prevalence : 0.6442
         Detection Rate : 0.6135
   Detection Prevalence : 0.6810
      Balanced Accuracy : 0.8814

       'Positive' Class : 0
```

Generalized Linear Model (GLM)

- Use variable importance plot to show the relative importance of each feature in the model

- Visualize the relationship between Fiction and Rating/Reviews



Conclusions & Recommendations

- Publisher Size: Small publishers should focus on niche authors that have high user ratings/medium book sales. Larger publishers should focus on the authors with the highest average book sales

- Genre: Fiction books/authors generally perform better than non-fiction

- Eye Catching Titles: Publisher should title books with words positively correlated with high reviews. Avoid words that are negatively correlated with reviews.