Marketing Analytics Final Project

Amazon Bestseller Books (2009-2019) Analysis

Group 8

Chia-Ling Ni, Mihir Kungulwar,

Tejaswini Edupuganti, Yi-Hsuan Tseng

Leavey School of Business, Santa Clara University

Marketing Analytics (MKTG2505)

Dr. Sujata Ramnarayan

March, 7, 2022

Table of Contents

Introduction to Company Overview: Amazon

Amazon.com, Inc. is an American multinational technology company which focuses on e-commerce, cloud computing, digital streaming, and artificial intelligence. It has been referred to as "one of the most influential economic and cultural forces in the world", and is one of the world's most valuable brands. It is one of the Big Five American information technology companies, alongside Alphabet, Apple, Meta, and Microsoft.

Amazon was founded by Jeff Bezos from his garage in Bellevue, Washington, on July 5, 1994. Initially an online marketplace for books, it has expanded into a multitude of product categories: a strategy that has earned it the moniker The Everything Store. It has multiple subsidiaries including Amazon Web Services (cloud computing), Zoox (autonomous vehicles), Kuiper Systems (satellite Internet), Amazon Lab126 (computer hardware R&D). Other subsidiaries include Ring, Twitch, IMDb, and Whole Foods Market. Its acquisition of Whole Foods in August 2017 for US$13.4 billion substantially increased its footprint as a physical retailer.

Amazon is guided by four principles: customer obsession rather than competitor focus, passion for invention, commitment to operational excellence and long-term thinking. Amazon strives to be Earth’s most customer-centric company, Earth’s best employer, and Earth’s safest place to work. Customer reviews, 1-Click shopping, personalized recommendations, Prime, Fulfillment by Amazon, AWS, Kindle Direct Publishing, Kindle, Career Choice, Fire tablets, Fire TV, Amazon Echo, Alexa, Just Walk Out technology, Amazon Studios, and The Climate Pledge are some of the things pioneered by Amazon. (*Who we are | Amazon*, March 6, 2022)

Situation Analysis

Exploratory Data Analysis - Part I

The following insights have been arrived at while exploring the data initially:

- Non-fiction was a more popular Genre compared to Fiction. Out of the 351 unique books, 54.42 percent were Non-Fiction and 45.58 percent were Fiction as shown in Fig. 1-1.
- On further observation across years shown in Fig. 1-2, the same trend can be noticed with the highest fraction of bestsellers for Non-Fiction books being in 2015, which was coincidentally the lowest for Fiction books.
- User Rating ranges from 0 to 5 with the least rating for the books being 3.3 and the highest being 4.9. Fig. 1-3 shows that the most frequent user rating is 4.8. It may be inferred that a better rating is attracting more customers and in turn more reviews, whereas a book with lower rating is failing to attract more customers which is reflected by the number of reviews for that book.
- Further observation on review numbers by genre across years shown in Fig. 1-4, it can be seen that overall people prefer to leave reviews on Fiction over Non-Fiction books.
- On taking a look at the average prices across genres from 2009 to 2019 from Fig. 1-5, it can be seen that Non-Fiction books are more expensive when compared to Fiction except for the year 2009 where the difference in price is not too prominent.
- Also, when we look at the popularity of authors based on the reviews numbers of their books and the genre that they belong to in Fig. 1-6, among the top 10 authors, it can be seen that Michelle Obama is the only Non-Fiction author. It indicates that Fiction is a category that is most popular among Best Sellers.

Exploratory Data Analysis - Part II

Text Mining - Most Popular Words in the Names of Top Selling Books

- Here, text mining has been used to analyze the most common words in order to provide recommendations on how to name a book for authors to attract more sales. Since the books have already been classified as fiction and nonfiction categories, there is no need to conduct topic modeling analysis on book names for this dataset.

  As seen from Fig. 2-1, “book” is the most frequently used word, coming next with“edition”. This is essentially a dataset on bestselling books, so it makes sense that book is the most common word followed by edition which is also a common term used to represent different releases of a book. On further analysis by ignoring the word “book”, “series” and “edition” from the data as shown in Fig. 2-2, “life” becomes the most common word followed by “coloring”, “kid”, “diary”, “guide” and “love”.

- A way to view word connections is to treat them as a network, similar to a social network. In this network graph, the dots represent individual terms, while the lines connecting the dots represent the possible connections between each of these terms. The overall correlation between words can be seen in Fig. 2-3. Take a closer look at the words with high correlation in Fig. 2-4. There is a cluster that shows “kid”, “wimpy”, “dog”, “diary”, “captain”, and a few others that are more likely to appear together. By utilizing text mining analysis on the top selling book names on Amazon, it can give authors an overview of the trend in most popular keywords that appear in book names and what kind of words usually appear together. These observations are of great use to authors and publishers while creating the storyline of the books, and deciding on book names.

Identification of Problem

The scope of this analysis is to identify and analyze the impact of the key factors that influence a book in becoming a bestseller. This analysis also enables us to examine the data from an Author's and Publisher's perspective on what type of books they should be scouting for and what kind of titles generate more traction, thus maximizing reach, ratings and sales.

Approach and Techniques for Data Analysis

Before building an analytical model, the following steps are followed to ensure that the data is clean and easy to comprehend for analysis :

- Firstly, the dataset needs to be checked for missing values for further analysis conduction.
- In order to do text mining on book names, first the names are tokenized and then stop words which have no impact on semantic meaning are removed.
- To avoid duplicate values, words in plural form are converted into singular as they are basically the same word.
- The texts that do not entirely consist of alphabets are filtered out. For example, the word "6th" will be removed. The distinct() function is used to set all the words to be counted as a single value in each book.
- Explicit binary encoding is done for 'Genre' and implicit one hot encoding for 'Author'.

Classification Analysis : Logistic Regression on Genre

Conditional Model

To predict whether a bestseller belongs to a fiction or nonfiction genre, the maximum likelihood model is being used. It helps identify the parameters and probability distribution that best explain the observed data. The sigmoid function as shown in Fig.3-1a and formula in Fig.3-1b is used to take any theoretical real value and map it to a probability between 0 and 1 shaped like a squiggle.

After plotting the probabilities between 0 and 1, logistic regression is being used to find an optimal value which is used to generate a linear classification line in the feature space that categorizes the books as fictional genre for probabilities above the optimal threshold value and as nonfiction for those below it. For this analysis, the optimal value of the threshold is **0.48**.

Model Selection

Initially, the parameters 'User Rating', 'Reviews', 'Price' have been used as the control variables for the classification resulting in an accuracy of **64.72%**. To improve the accuracy of the model, another parameter 'Author' has been added. This has boosted the accuracy to **80.54%** implying that 'Author' is a significant parameter while analyzing the factors defining a bestseller. This accuracy comparison with and without Author is reflected in Fig.3-2, Fig.3-3, and Fig.3-4.

Feature Importance

The 'Author' parameter is automatically encoded in a binary format by '**glm()**' and is prioritized more than the features used in the previous model and this is reflected in Fig.3-5, Fig.3-6.

Optimization

The ROC Curve in Fig 3.7 depicts the trade off between the true positive rate and the false positive rate. The closer the curve is to the top left corner and the higher the area covered by it, the better the results. Here, 77% of the area is covered by the ROC curve and the area closest to the top left corner is the specific point of interest as shown in Fig.3-7. Interpretation of the confusion matrix (Table.3-1) is as follows:

- Precision = True Positive / (True Positive + False Positive) = 67 / (67 + 19) = 0.77
- Recall = True Positive / (True Positive + False Negative) = 67 + (67 + 18) = 0.788
- Accuracy = (True Positive + True Negative) / Total = (67 + 34) / (67 + 19 + 18 + 34) = 101 / 138 = 0.731

Conclusions and Recommendations

- At an initial glance even though there are more books in NonFiction as compared to Fiction, from the analysis, it can be seen that Authors of Fiction books are more popular than Non Fiction. So, Amazon can take cues from this insight and add more books in the Fiction genre to boost sales of books on their website.
- Also, Amazon needs to invest in advertising the NonFiction books that are being sold to attract more customers, so that it leads to an overall increase in sales of books.
- From our research it can also be seen that NonFiction books are priced more than Fiction books. This is a major contributing factor in the decline of sales of NonFiction books. Lowering the price would bring in more sales as it improves the purchasing capacity of the customers.
- To improve the sales, irrespective of genres, Amazon needs to send their customers personalized emails with recommendations based on their search history, previous purchase of books and the reviews and ratings that they have given for each of these books.
- For the books that have low sales and reviews, Amazon could launch special promotions, or provide a sample copy with a few pages of the book to generate interest.
- Interesting book names lure customers. By taking into account the most popular keywords from the analysis, Amazon can in turn provide recommendations to the Authors regarding naming their books to increase the chances of it being noticed by a wider customer base.
- It would also be a good idea to encourage customers to leave more reviews by offering some perks for leaving reviews such as bonus points that can be used while making a purchase on the website as it would help the company analyze customer preferences, opinions on the books and their purchasing behavior and launch marketing strategies based on the analysis.

References

Amazon (company). (2022, March 4). In *Wikipedia*. https://en.wikipedia.org/wiki/Amazon_(company)

Who we are | Amazon. (n.d.). US About Amazon. Retrieved March 6, 2022, from https://www.aboutamazon.com/about-us

R. Hafen and T. Critchlow, "*EDA and ML -- A Perfect Pair for Large-Scale Data Analysis*," 2013 IEEE International Symposium on Parallel & Distributed Processing, Workshops and Phd Forum, 2013, pp. 1894-1898, doi: 10.1109/IPDPSW.2013.118.

Silge, Julia, and David Robinson. 2017. "*Text Mining with R: A Tidy Approach.* " O'Reilly Media, Inc.".

Wiedemann, Gregor, and Andreas Niekler. 2017. "*Hands-on: A Five Day Text Mining Course for Humanists and Social Scientists in R*. " In Proceedings of the Workshop on Teaching NLP for Digital Humanities (Teach4DH2017), Berlin, Germany, September 12, 2017, 57–65.

Basta, N. (2020, April 1). *The Differences between Sigmoid and Softmax Activation Functions*. Nikola. Retrieved March 6, 2022, from https://medium.com/arteos-ai/the-differences-between-sigmoid-and-softmax-activation-function-12adee8cf322

Appendix

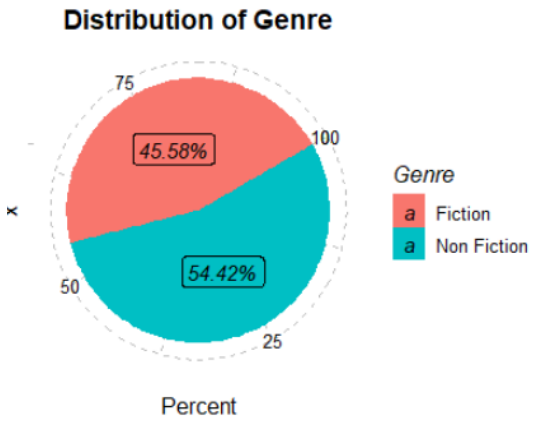

Fig. 1-1: Distribution of Genre

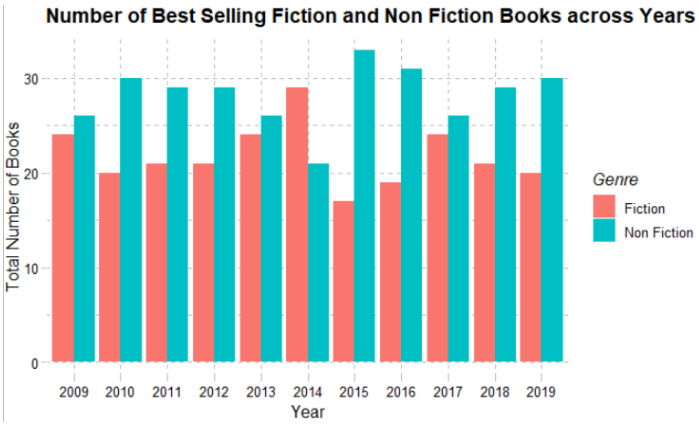

Fig. 1-2: Number of Books by Genre across Years

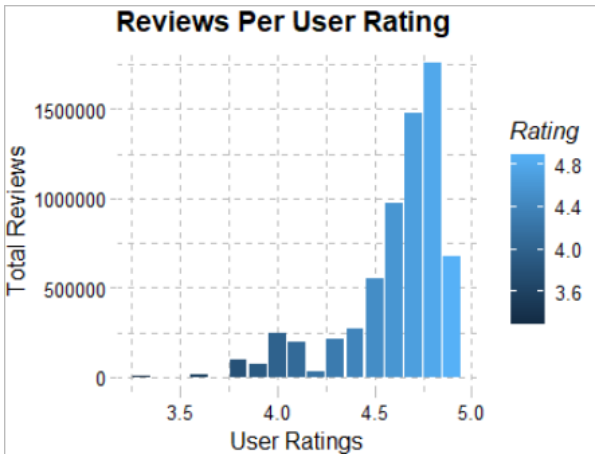

Fig. 1-3: Reviews per User Rating

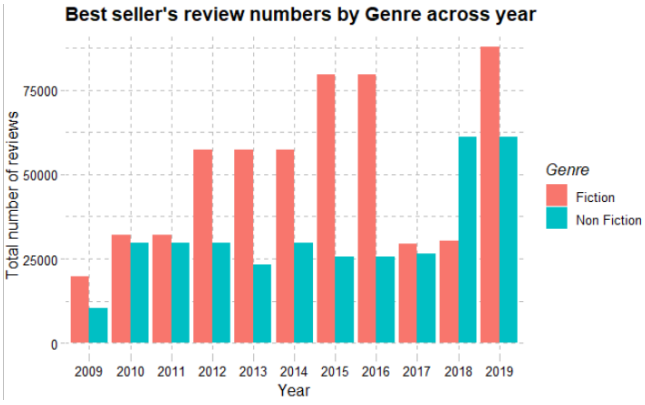

Fig. 1-4: Review numbers of Books by Genre across Years

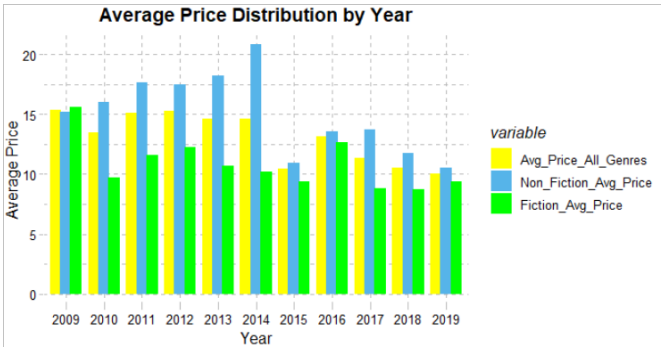

Fig. 1-5: Average Price across Years

Fig. 1-6: Top 10 Authors by Genre and Review numbers

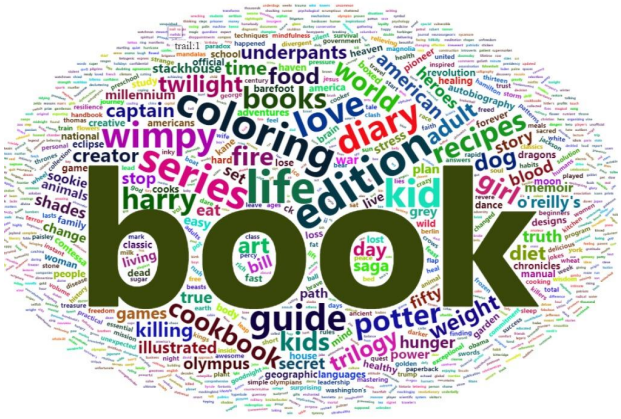

Fig. 2-1: Most Popular Words

Fig. 2-2: Most Popular Words (excluding book, edition, series)

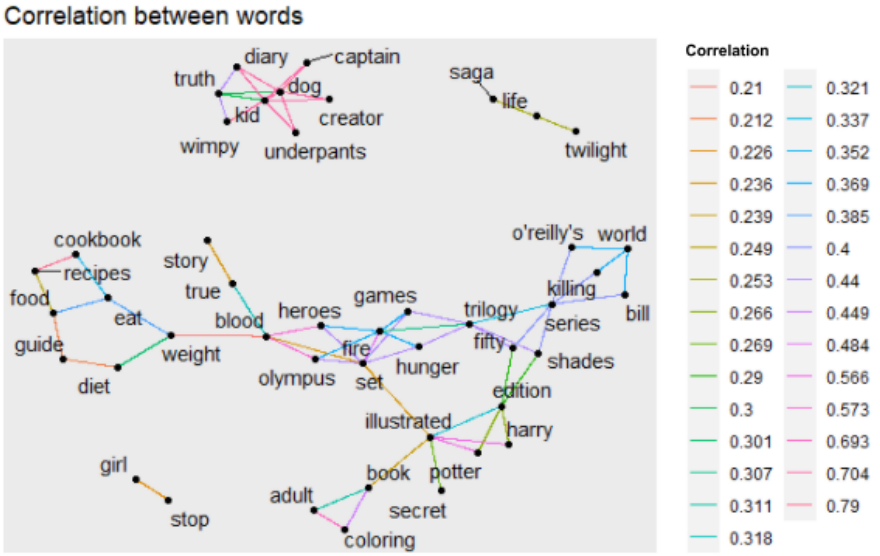

Fig. 2-3: Correlation Between Words

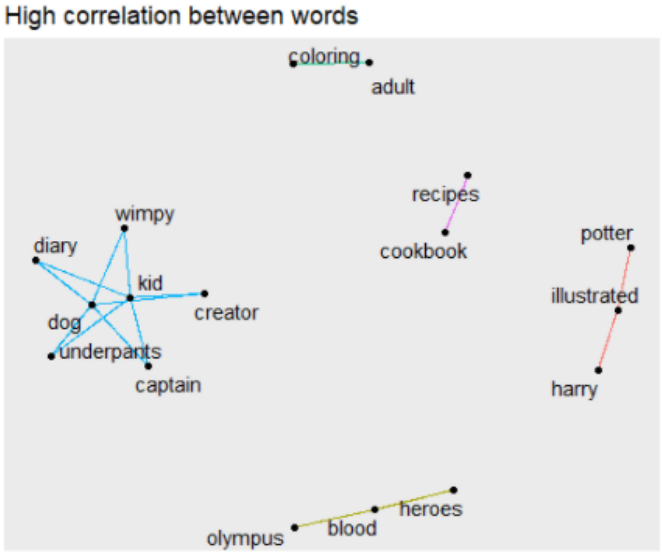

Fig. 2-4: Highly Correlated Words

$$p(y|x;\beta) = \frac{1}{1 + exp - \left\{\beta_0 + \sum_{j=1}^{d} \beta_j x_j\right\}}$$

Fig. 3-1 a: Sigmoid Formula

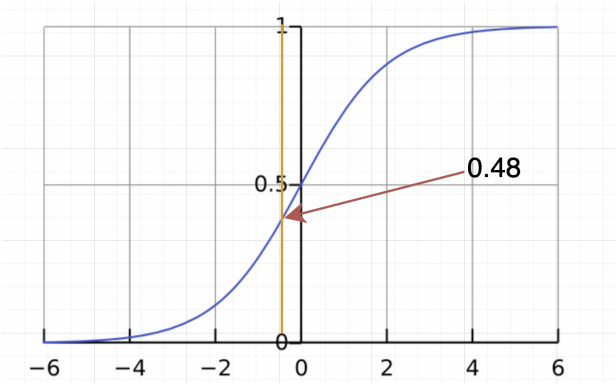

Fig. 3-1 b: Sigmoid Function

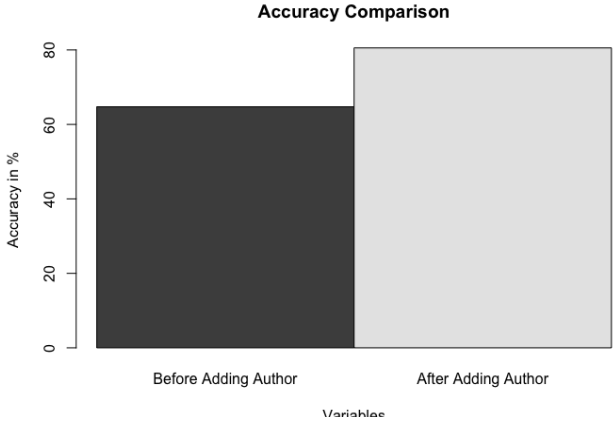

Fig. 3-2: Accuracy Comparison

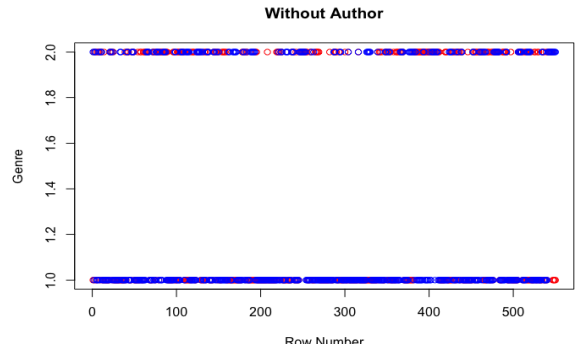

Fig. 3-3: Scatter Plot without Author

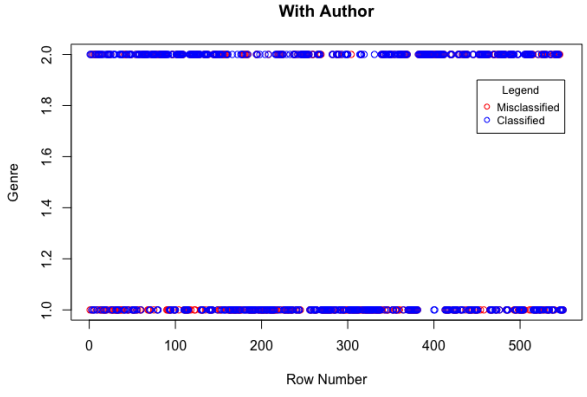

Fig. 3-4: Scatter Plot with Author

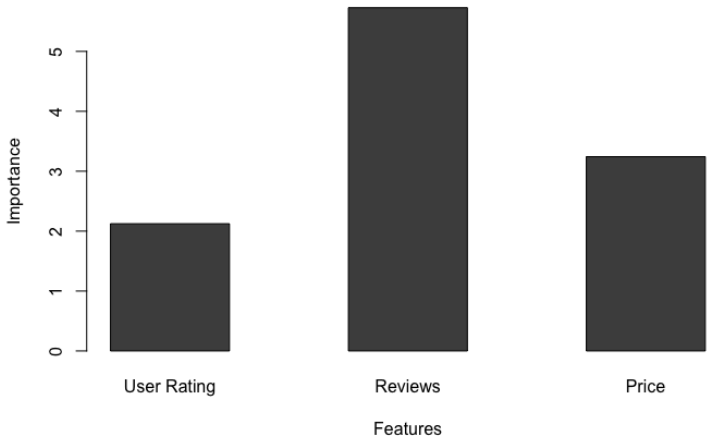

Fig. 3-5: Feature Importance

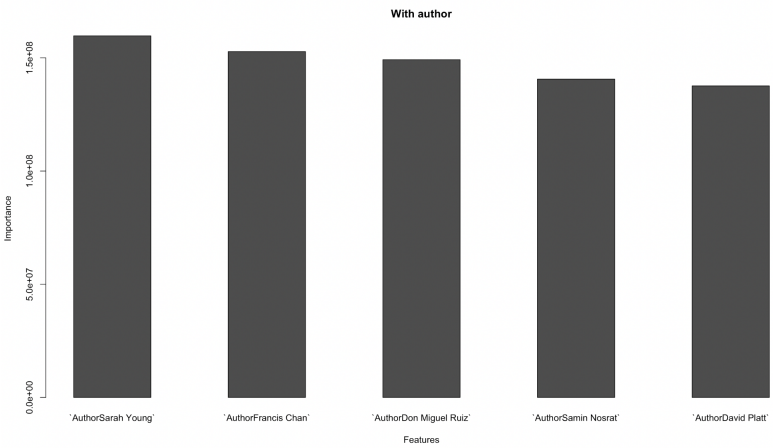

Fig. 3-6: Feature Importance

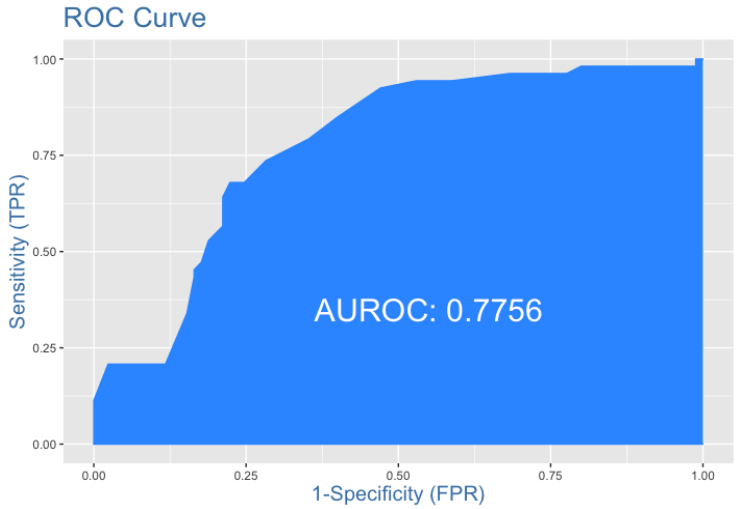

Fig. 3-7: ROC Curve

		Actual Values	
		Non Fiction	Fiction
Predicted Values	Non Fiction	67	19
	Fiction	18	34

Table. 3-1: Confusion Matrix