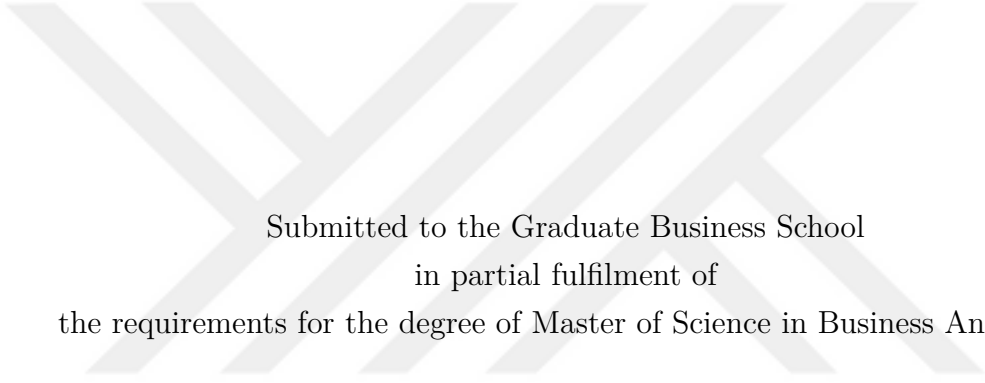


# THE EFFECT OF RELEASE DATES ON THE BOOK SALE RANKS

by  
EDA EYLÜL AKDEMİR



Submitted to the Graduate Business School  
in partial fulfilment of  
the requirements for the degree of Master of Science in Business Analytics

Sabancı University  
December 2020

# THE EFFECT OF RELEASE DATES ON THE BOOK SALES

Approved by:



Date of Approval: December 24, 2020



EDA EYLÜL AKDEMİR 2020 ©

All Rights Reserved

## ABSTRACT

### THE EFFECT OF RELEASE DATES ON THE BOOK SALE RANKS

EDA EYLÜL AKDEMİR

Business Analytics M.Sc. Thesis, DECEMBER 2020

Thesis Supervisor: Prof. Dr. Abdullah Daşcı

Keywords: book sale ranks, time series prediction, supervised learning, lagged variables, linear regression, ridge regression, random forest, light gradient boosting machine, neural networks.

In this study, we examined the effect of a book's publishing date on its sales ranking with a Linear Regression model by using Amazon's daily book ranking and price data for 67 days. We found out that the release date of a book is an important factor in evaluating the book rankings. We also studied the prediction of rankings using the lagged variables of price and ranking. To transform this time series prediction problem into a supervised learning problem, we used the sliding window approach. We used four machine learning and one deep learning approach to predict the rankings. To compare the results, two evaluation criterias;  $R^2$  and root mean squared error were used. When tuning the hyperparameters, we used k-fold Cross Validation. We found out that linear regression outperformed the rest of the models, which are Ridge Regression, Random Forest, Light Gradient Boosting Machine, and Neural Network.

## ÖZET

KITAPLARIN ÇIKIŞ TARİHLERİNİN, SATIŞ SIRALAMASINA OLAN ETKİSİ

EDA EYLÜL AKDEMİR

İŞ ANALİTİĞİ YÜKSEK LİSANS TEZİ, ARALIK 2020

Tez Danışmanı: Prof. Dr. Abdullah Daşcı

Anahtar Kelimeler: satış sıralaması, yayınlanma tarihi, zaman serisi tahmini, güdümlü öğrenme, gecikmeli değişkenler, lineer regresyon, ridge regresyonu, rastgele orman, gradyan arttırma makinesi, yapay sinir ağları.

Bu çalışmada, Amazon web sitesinin 67 günlük kitap satış sıralaması ve fiyat kayıtlarına ait veriyi kullanarak kitapların yayınlanma tarihlerinin, satış sıralamasına olan etkisini Lineer Regresyon yöntemi ile inceledik. Kitapların yayınlanma tarihinin, satış sıralamasını saptamakta önemli bir rol oynadığı sonucuna vardık. Ek olarak, satış sıralamalarını ve fiyat kayıtlarını kullanarak otoregresif bir şekilde satış sıralamalarını tahmin etmeye çalıştık. Bu zaman serisi tahmini problemini bir güdümlü öğrenme problemine dönüştürmek için "kayan pencere" metodunu uyguladık. Satış sıralamasını tahmin etmek için dört makine öğrenimi modeli ve bir derin öğrenme modeli kurduk. Karşılaştırma için iki kriter olan  $R^2$  ve karesel ortalama hata ölçütlerini hesapladık. Modellerdeki parametre seçimlerini yapmak için k-katlamalı Çapraz Geçerlilik yöntemini kullandık. Lineer Regresyonun diğer dört model olan Ridge Regresyonu, Rastgele Orman, Hafif Gradyan Arttırma Makinesi ve Yapay Sinir Ağları'ndan daha iyi performans gösterdiği sonucuna ulaştık.

## ACKNOWLEDGEMENTS

Many people gave their precious support for this work. First of all, I would like to thank Prof. Dr. Abdullah Daşcı, for being my thesis supervisor. Without his precious support, this work would be incomplete. I would also like to thank him for his emotional support as well. As a panic worker, his calmness and trust in me guided me very well during this process.

Secondly, I would like to thank Assoc. Prof. Cenk Kocaş, for giving me the chance to work with this dataset. I am also grateful for his suggestions and ideas.

I also would like to thank Assoc. Prof. Mümtaz Karataş and Assist. Prof. Melek Akın Ateş for their time and invaluable comments that helped me to finalize the thesis.

I am deeply grateful to my parents and my brother, for their love and belief in me. Their unconditional support has brought me to this path.

Next, I would like to express my gratitude to my best friends Burcu Sarı, Gergely Buda, and Elif Saraçoğlu. Their emotional support and suggestions made my way clearer.

Also, I thank Ahmet Mikail Bayındır and Edin Yalçın for their professional support. Without their help, I could not have managed to get the extra data set I needed for my analysis.

Last but not least, I would like to express my sincere gratitude to my best friends; Özge Özkır and Ahmet Alp Softa for their emotional support in my hard times.



*To my family*

## TABLE OF CONTENTS

<b>LIST OF TABLES</b> .....	<b>x</b>
<b>LIST OF FIGURES</b> .....	<b>xi</b>
<b>1. INTRODUCTION</b> .....	<b>1</b>
<b>2. LITERATURE REVIEW</b> .....	<b>3</b>
2.1. Book Sales and Rankings .....	3
2.2. Time Series Prediction as a Supervised Learning Problem .....	4
<b>3. DATA, DATA CLEANING AND PREPROCESSING</b> .....	<b>6</b>
3.1. Uncleaned Data.....	6
3.2. Data Cleaning and Preprocessing - Part One.....	8
3.3. Data Preprocessing - Part Two .....	9
3.4. Train-Test Split .....	10
<b>4. DESCRIPTIVE STATISTICS</b> .....	<b>11</b>
4.1. Listprice .....	11
4.2. Listprice vs. Genre, Age Group and Physical Format .....	12
4.3. Genre, Age Group and Physical Format .....	13
4.4. Publisher.....	15
4.5. Publishing Date.....	15
4.6. ABRank .....	17
4.7. Price .....	20
<b>5. METHODOLOGY</b> .....	<b>22</b>
5.1. Linear Regression .....	22
5.1.1. Backward Elimination Model .....	23
5.2. Ridge Regression.....	24
5.3. Random Forest .....	24
5.3.1. Hash Encoding .....	25
5.4. LightGBM .....	26



5.5. Multilayer Perceptron.....	27
5.6. Evaluation Criteria .....	28
<b>6. RESULTS.....</b>	<b>30</b>
6.1. Release Date Effect on ABRank .....	30
6.2. Linear Regression .....	31
6.3. Ridge Regression.....	33
6.4. Random Forest .....	34
6.5. LightGBM .....	35
6.6. Multilayer Perceptron.....	36
6.7. Overall Results .....	37
<b>7. CONCLUSION .....</b>	<b>38</b>
<b>BIBLIOGRAPHY.....</b>	<b>40</b>
<b>APPENDIX A .....</b>	<b>42</b>

## LIST OF TABLES

Table 3.1. Raw Data Variables and Their Explanations .....	7
Table 3.2. Cleaned Data Variables and Their Explanations.....	9
Table 4.1. Listprice Statistics .....	11
Table 4.2. Average Listprice for Genre, AgeGroup, and Physical Format ..	13
Table 4.3. Weekday Frequency of Publishing Dates .....	16
Table 4.4. Descriptive Statistics for ABRank Variable Sample .....	18
Table 4.5. Descriptive Statistics for Price Variable Sample .....	20
Table 5.1. Linear Regression Full Variables Model .....	23
Table 5.2. Linear Regression Final Model .....	24
Table 5.3. Parameter Trials for Random Forest Regressor .....	25
Table 5.4. Parameter Selection for Random Forest Regressor .....	25
Table 5.5. Parameter Combinations for LightGBM .....	27
Table 5.6. Parameter Selection for LightGBM .....	27
Table 6.1. Results for Full Data Linear Regression .....	31
Table 6.2. Linear Regression Result Model.....	32
Table 6.3. Evaluation Criteria for Linear Regression .....	32
Table 6.4. Ridge Model vs. Linear Regression Model .....	33
Table 6.5. Evaluation Criteria for Ridge Regression.....	33
Table 6.6. Variable Importances for Random Forest .....	34
Table 6.7. Evaluation Criteria for Random Forest .....	34
Table 6.8. Variable Importances for LightGBM .....	35
Table 6.9. Evaluation Criteria for LightGBM .....	36
Table 6.10. Evaluation Criteria for Multilayer Perceptron .....	36
Table 6.11. Overall Results .....	37
Table A.1. Publisher Counts .....	43

## LIST OF FIGURES

Figure 3.1. Sliding window with a window size of 6 .....	10
Figure 4.1. Histogram for listprice .....	12
Figure 4.2. AgeGroup and Genre with Numbers .....	14
Figure 4.3. Physical format and Genre with Numbers .....	14
Figure 4.4. Physical Format and AgeGroup with Numbers .....	15
Figure 4.5. Publishing Frequency of Books.....	16
Figure 4.6. Publishing Percentages According to the Months .....	17
Figure 4.7. A Sample of Book Rankings .....	18
Figure 4.8. Peak Rank Frequency .....	19
Figure 4.9. Released Status on Peak Day .....	19
Figure 4.10. A Sample of Book Prices .....	20
Figure 5.1. Normal Boosting Algorithms from (Mandot, 2018) .....	26
Figure 5.2. LightGBM from (Mandot, 2018) .....	26
Figure 5.3. Two Hidden Layered MLP architecture (Kain, 2019) .....	28
Figure 6.1. Feature Importances for LightGBM.....	36

## 1. INTRODUCTION

The shopping channels evolve as the e-commerce industry develops. Brick-and-mortar stores gave their place to online dealers in many markets and one of them is the book industry. There are many types of online dealers when it comes to books. There are small dealers and publishers, who only sell books like Barnes & Noble, and there are large e-commerce sites that sell a great variety of products; like Amazon. Even way, these websites produce a great amount of data every day.

In our study, we are using an Amazon Book dataset, which consists of 2454 books and 67 days of data points for each book. Each data point has unique pricing and ranking data. The ranking data is called ABRank, which is an abbreviation for Amazon Book Rank. In Amazon, every product which is sold one unit enters the ranking system (McMullen, 2018). This applies to books as well. In our data, we have the save amounts, the genre, physical format, age group, and release status information of the book as well.

This study aims to answer two questions. The first question is about the release dates and whether they affect the rankings, and the second question is about the practicality of predicting the ABRank variable from the rest of the variables. We used Linear Regression for the first question, and five supervised learning approach for the second question. When we examined the literature, we encountered a shortage of papers. Although there are papers that aim to predict book sales, not many of them handled the prediction problem as a supervised learning problem. So we searched for papers which studies time series prediction as a supervised learning problem and we again encountered a shortage of papers.

We applied Linear Regression for measuring the effect of release status on ABRank and used Linear Regression, Ridge Regression, Random Forests, Light Gradient Boosting Machine, and Multilayer Perceptron to predict ABRank from other variables. We found out that the release status/date is a significant factor in predicting the ranking. We also concluded that Linear Regression outperforms other prediction models in terms of R-squared and root mean squared error (RMSE).

The rest of the thesis goes as follows: In the second chapter, we discussed the literature and previous works. The third chapter explains our data, the cleaning, and preprocessing steps. The fourth chapter is about data exploration and descriptive statistics. The fifth chapter explains the methodologies we used to achieve our results. The sixth chapter discusses the results and the seventh chapter concludes our study.



## 2. LITERATURE REVIEW

In this section, we will review the literature. Since there are few studies in this field, we will examine the literature on two main topics. First of all, the papers about book sales and rankings will be covered. Secondly, we will review selected papers that take a time series prediction as a supervised learning problem.

### 2.1 Book Sales and Rankings

Book sales and rankings are popular issues that researchers choose to study. Sornette, Deschâtres, Gilbert & Ageon (2004) found out that exogenous causes similar to advertisements in newspapers could make a book's ranking to jump in a very short time which is followed by a sudden fall, while endogenous causes affect the rankings in a much slower pattern, both in positive and negative ways, using a power-law transformation. A Pareto distribution was used to transform the sales ranks to sales quantities to measure the price competition between two online retailers; Amazon and Barnes and Noble (Chevalier & Goolsbee, 2003). Although Sornette et al. (2004) and Chevalier & Goolsbee (2003) claimed that a simple power-law could be used to transform book rankings into sales, Fenner, Levene & Loizou (2010) showed that a simple power-law could not catch the long tail effect in book sales and could lead to biased results. Alternatively, they proposed a generative model which resulted in the asymptotic power-law distribution in book sales.

For book sales prediction, a great variety of methods were used in the literature. For example, a recent study found out that in the pre-release period, the publisher variable plays an important role for book sales, and authors reached this result by using Learning To Place (L2P) algorithm (Wang, Yucesoy, Varol, Eliassi-Rad & Barabási, 2019). Wang et al. (2019) also came to result that the authors' selling history plays a significant role, while Chang & Lai (2005) found the same outcome

by using Self-Organizing Map of neural network with Case-Based Reasoning (SOM-CBR). SOM with CBR outperformed traditional CBR and K-mean CBR in terms of both accuracy and computation while predicting the optimal volume of book orders (Chang & Lai, 2005). Five different machine learning algorithms include M5P, Random Forest, Linear Regression, k-Nearest-Neighbour, and Support Vector Machine Regression were used by Castillo, Mora, Faris, Merelo, García-Sánchez, Fernández-Ares, De las Cuevas & García-Arenas (2017) and they found out that both decision tree methods give the best results and can be easily used by a publisher when predicting the newly published books' sales.

## **2.2 Time Series Prediction as a Supervised Learning Problem**

Many methods in the literature handle the time series forecasting as a supervised or unsupervised learning problem. For instance, Hota, Handa & Shrivastava (2017) used a sliding window-based Radial Basis Function Network (RBFN) model which is an Artificial Neural Network model to a time series stock data. Another study found out that the sliding window approach allows the time series predictors to be efficient for machine learning algorithms such as Evolutionary Extreme Learning Machines (E-ELMs) while predicting the vehicle speeds (Mozaffari, Mozaffari & Azad, 2015).

In a comparison study, researchers found out that two out of eight machine learning models that are Multi-Layer Perceptron (MLP) and Gaussian Processes were best for time-series predictions and the preprocessing technique of time series were quite significant for the results (Ahmed, Atiya, Gayar & El-Shishiny, 2010). Also, Qian & Gao (2017) concluded that the machine learning models including MLP, Logistic Regression and Support Vector Machines (SVM) outperformed the traditional time series forecasting methods such as ARIMA in precision, by experimenting on Dow 30, S&P 500, and Nasdaq stock datasets. In a survey study, authors claimed that SVM is a popular methodology to predict time series, especially in financial market prediction and electric utility forecasting fields (Sapankevych & Sankar, 2009).

In another Neural Network study, it is found that the heuristics used at finding the optimal size of sliding window and sample ratio, improved the time series prediction results (Frank, Davey & Hunt, 2001). In addition to back propagated neural networks and statistical models such as AR, ARIMA, ARMA and MA, Kayacan, Ulutas & Kaynak (2010) found out that grey system theory based models such as

GM(1,1) which is called as the Grey Verhulst model, can perform better than the methods stated before in a daily currency exchange rates data.

Our work showed that unpopular methods such as Linear and Ridge Regression, Random Forest, Gradient Boosting Machines can be used to predict time series after transforming the data with sliding window approach. Also, we applied MLP to our dataset, which is a popular method in the literature and found out that Linear Regression performed the best.





### **3. DATA, DATA CLEANING AND PREPROCESSING**

In this section, the data will be explained and some descriptive statistics will be examined. First, the raw data will be presented. Then, the cleaning and preprocessing stages will be described. Last, the final data will be explained.

#### **3.1 Uncleaned Data**

The raw data has 7332 unique books, 846,405 rows and 32 columns. Data was collected by Kocas, Pauwels & Bohlmann (2018) between June 1, 2011 to Sept 3, 2011 from Amazon's website, under New releases > coming soon. The columns are explained below on Table 3.1.

Table 3.1 Raw Data Variables and Their Explanations

Variable	Explanation
ID	The ID number given by the data collectors
title	The name of the book.
ISBN10	10 digit unique ISBN number.
ISBN13	13 digit unique ISBN number
ASIN	ASIN number
listprice	The listprice of the book. This variable does not change over time
price	The price of the book according to the timestamp. It changes over time
yousave	Difference between list price and price
yousave %	The amount of saving in terms of percentage
ABRank	Amazon Book Rank. The ranking of the book according to its sales
retailers	Number of retailers that sells a particular book
soldbyamazon	Whether the book is sold by Amazon. Binary variable
physical_format	The actual format of the book. It changes from Audiobook to Paperback books. It has 59 categories
Publisher	The publisher of the book
publishing_date	The release date of the book
Language	The language of the book. Audiobooks have multilingual support
avg_cus_rate	Average customer rating of the book
numberoflike	Number of likes related to the book
Category	Empty column
link	The Amazon link of the book.
date	The retrieval date of the row
time	Retrieval time of the row
total reviews	Number of total reviews in the given date of a particular book
5, 4, 3, 2, 1 Star reviews	The number of star reviews
Amazon extra Rank 1, 2, 3, 4	The first column of the data consists of some genre knowledge, while other columns are mostly empty

A glimpse of the data can be found in Appendix A.

### 3.2 Data Cleaning and Preprocessing - Part One

The data has a great majority of unnecessary and irrelevant data and as a result, most of the data was deleted. As the study mainly examines the book's rankings, first the rows that have missing ABRank values were dropped. Eventually, about 260,000 rows were deleted. For consistency, the books that are not English were deleted, as other languages may have quite different dynamics.

We decided to study the books that are published during the data collection time. To eliminate the other years, the publishingdate column was divided into three columns called year, month, and day. First, the years that are not 2011 were removed from the data. Then, the months that were not equal to 6,7 and 8 were removed. We decided to focus on the dates between June 7, 2011, and August 2, 2011 to capture the effect of pre and post release periods, as the data has a large range of release dates. Finally, the days were filtered and rows that contain irrelevant days were deleted. After this step, approximately 420,000 rows remained. To make sure that each book has the same amount of daily data points, the timestamp date (date column) was divided into three columns called Ryear, Rmonth, and Rday. Then, multi-level sorting was applied to the data. Data is sorted according to their ISBNs, Ryear, Rmonth, and Rday values, respectively. Next, an artificial column called the samedate was created to check the consecutive rows if they have the same day information because, in the data, some of the books have duplicate data points that were gathered two times on a certain day. To avoid the misinformation on price, in duplicate data points, an average of two prices were taken. The rows that have 1 for the samedate variable was dropped from the dataset. Hence, the duplicate variables issue was resolved. Afterward, the books which have less than 67 days of data points, the books that have 0 value for price and list price, and the rows that have a different value for physical format then Hardcover, Audio, and Paperback were removed from the dataset. Using the ISBN13 codes, the genre (fiction and non-fiction) and age group (adult and children) were added via scraping the information from a website called alibris.com. The release status of the books was simply gained by subtracting the publication date from the retrieval date. The date variable was turned into a numerical variable at the range of 1 to 67, as there are 67 data points for each book and finally, a categorical variable called bookno is created to specify distinct books rather than the ISBN13 variable due to its complexity. All the unnecessary columns except date, ISBN13, listprice, price, yousave, ABRank, and physicalformat was dropped. New columns called bookno, Genre, AgeGroup, and Releasedornot were added.

Cleaned data consists of 164,352 rows and 12 columns. Explanation of the variables can be found in Table 3.2.

Table 3.2 Cleaned Data Variables and Their Explanations

Variable	Explanation
date	Date variable. Ranges between 1 and 67
panelid	Short version of ISBN13. Represents distinct books. Used for dummification
bookno	Longer version of panelid. Used for hashing
ISBN13	Unique book identifier
listprice	Initial price of the book
price	The price of the book at the given date
save	The difference between listprice and price
ABRank	Ranking according to the given date
Genre	Genre of the book. 1 for Fiction, 2 for Non-Fiction
AgeGroup	Age group of the book. 1 for Adult, 2 for Children
Phyfmt	Physical format of the book. 1 for Hardcover, 2 for Paperback, 3 for Audiobook.
Releasedornot	Release status of the book. 1 for released, 0 for not released.

### 3.3 Data Preprocessing - Part Two

After the data was cleaned and new columns were added, we transformed the problem from time-series learning to supervised learning by using the sliding window method.

Sliding Window Method is a popular time series segmentation technique that is used in the fields of weather prediction, finance, and medical applications (Yahmed, Bakar, Hamdan, Ahmed & Abdullah, 2015). The window size can be increased or decreased according to the desire or until a certain error threshold is met (Hota et al., 2017). The method is shown in Figure 3.1.

Figure 3.1 Sliding window with a window size of 6

	ABRank	lagrank1	lagrank2	lagrank3	lagrank4	lagrank5	lagrank6
	199688	149062	94970	68020	72165	195985	214547
Window size=6	149062						
	94970						
	68020						
	72165						
	195985						
	214547						

Here, the window size is equal to 6. For both the ABRank variable and price variable, since they are the two variables that depend on time, this method is applied. This means, to predict the ABRank of day N, the previous 6 days' ABRank and price value will be used in the models. For example, to predict day 67's ABRank, the past price and ABRank values of days 66, 65, 64, 63, 62, and 61 are used. After the sliding window, the number of columns rose to 24.

### 3.4 Train-Test Split

After the data was preprocessed, we splitted it into train and test sets. We made the split according to the date variable. The train set only contains days from 7 to 47, while the test set contains days from 48 to 67.

## 4. DESCRIPTIVE STATISTICS

In this section, we will present descriptive statistics on important variables.

### 4.1 Listprice

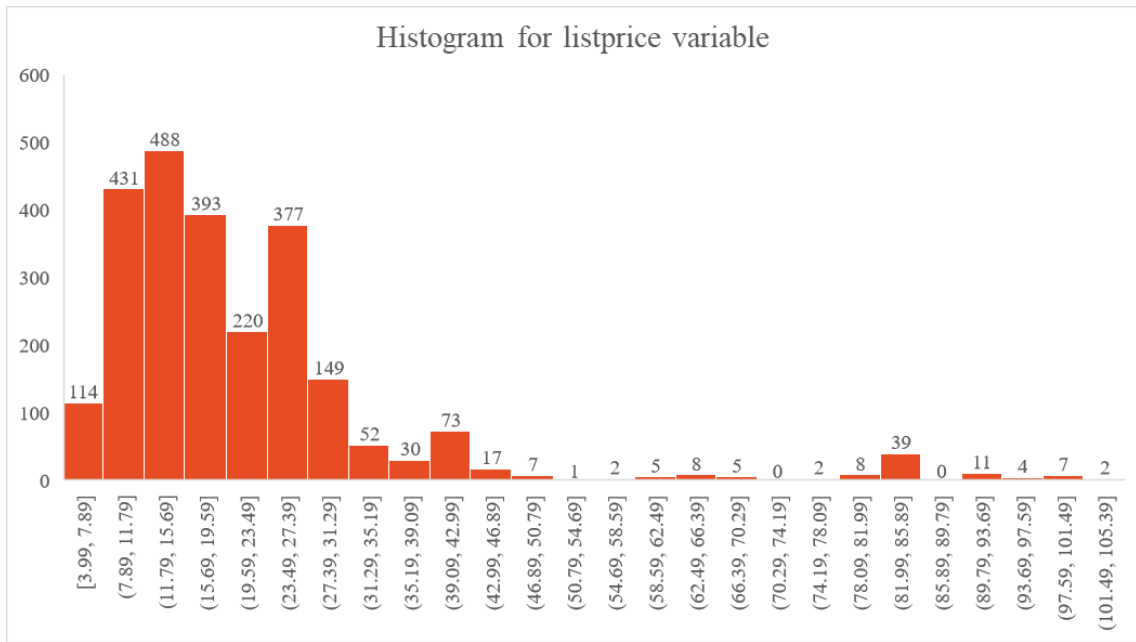
For unique books, descriptive statistics for listprice variable is given below.

Table 4.1 Listprice Statistics

Statistic	Value
Mean	22.08
Standard Error	0.57
Median	16.99
Mode	7.99
Standard Deviation	28.29
Sample Variance	800.50
Kurtosis	486.02
Skewness	18.23
Range	871.01
Minimum	3.99
Maximum	875
Sum	54168
Count	2453

Since the mean is greater than the median, the data is skewed to the right. The positive kurtosis value implies that the listprice value is leptokurtic, meaning that this variable has a profusion of outliers.

Figure 4.1 Histogram for listprice



The histogram was drawn by omitting the outliers since it would be difficult to observe the shape of the listprice variable's distribution.

## 4.2 Listprice vs. Genre, Age Group and Physical Format

There are some interesting results when the listprice is interpreted with other categorical variables. For example, for the genre, the average listprice for fiction books is 17.73\$, and for non-fiction books the average listprice is 30.35\$, meaning that non-fiction books are indeed more expensive than fiction books. For the age group, books for adults are more expensive than books for children. Finally, for the physical format, the most expensive books are hardcover books, followed by audio and paperback books, respectively. The summary table is given below.

Table 4.2 Average Listprice for Genre, AgeGroup, and Physical Format

Labels	Average Listprice
Fiction	17.73
Non-Fiction	30.35
Adult	23.30
Children	15.17
Audio	28.01
Hardcover	31.75
Paperback	14.54

### 4.3 Genre, Age Group and Physical Format

The data is unbalanced in terms of genre, age group, and physical format variables. For genre, 1608 out of 2453 books are fictional and 845 books are non-fictional. The reason for this circumstance might be the challenge of writing non-fictional books. However, this will not be examined in this study since it is out of scope. The number of adult books in this dataset is 2085, while the quantity of childrens books is only 368. The number of audio, hardcover and paperback books are 375, 781 and 1297, respectively.

The ratio of books for genre vs. age group, genre vs. physical format and age group vs. physical format are given on the pie charts below.



Figure 4.2 AgeGroup and Genre with Numbers

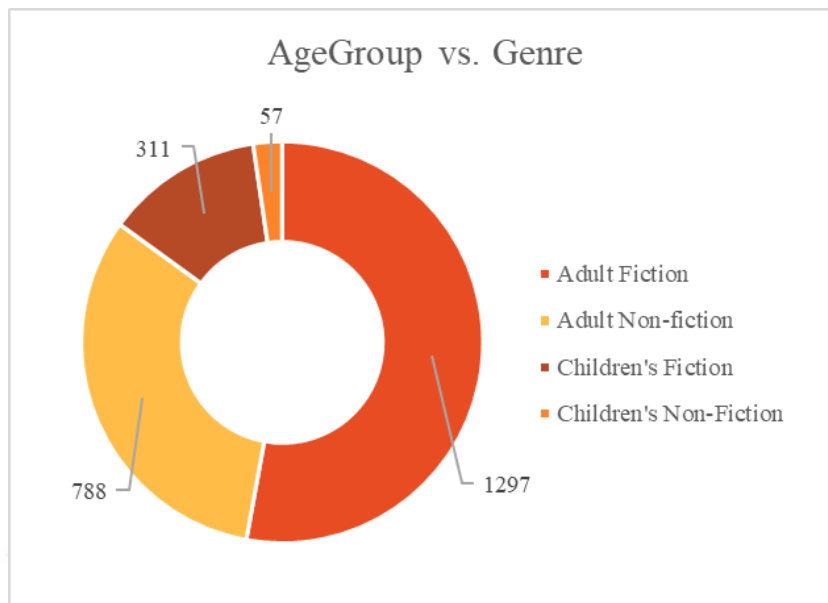


Figure 4.3 Physical format and Genre with Numbers

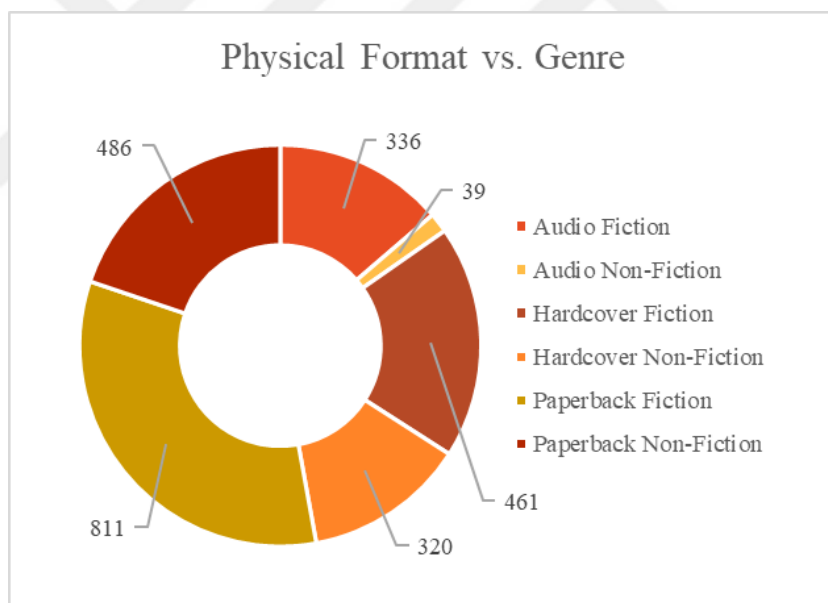
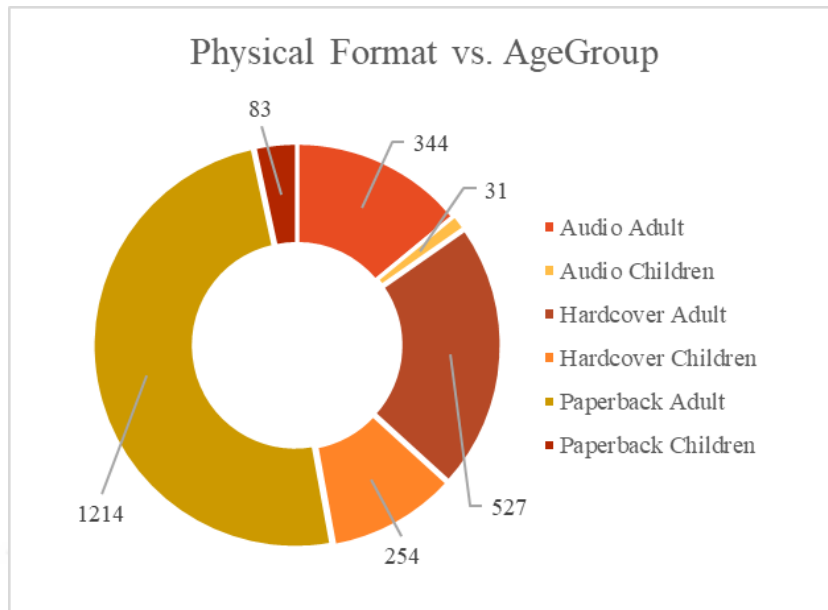


Figure 4.4 Physical Format and AgeGroup with Numbers



#### 4.4 Publisher

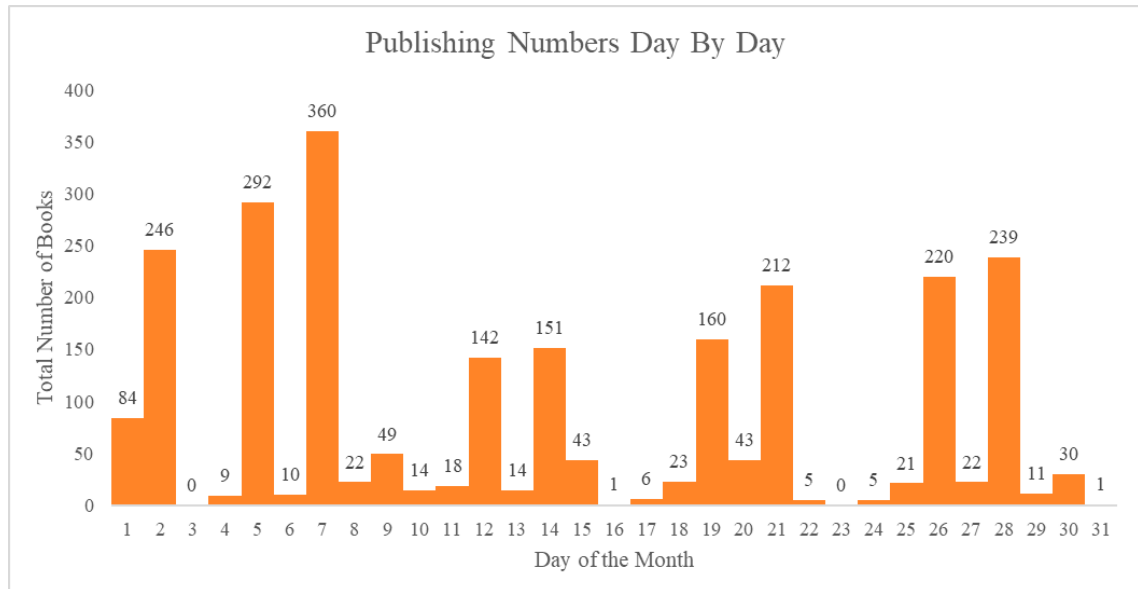
Normally, the publisher variable is not included in the models, however, to understand the books on a deeper level, we include it in this section. There are 303 different publishers for 2453 books. 245 of those publishers only published 10 or fewer books of our dataset. The first publisher is Brilliance Books with 232 books, and the second is Harper Collins, with 148 books. A detailed list of publishers is given in Appendix A.

#### 4.5 Publishing Date

In our dataset, the `releasedornot` variable is used to represent the publishing date. It is in the form of a categorical variable that changes from day to day. For example, if a given book's publishing date is on day 25, the `releasedornot` variable's value until the 25th day is 0, meaning that the book is on the pre-release period and the values after the 25th day are 1, meaning that the book is on post-release period. The 25th

day is the day that comes in 25th place out of 31 days. Since the variable cannot be used in this way, we went back to the main dataset to extract the exact publishing dates. The publishing date's frequency in terms of a month is given in Figure 4.5.

Figure 4.5 Publishing Frequency of Books



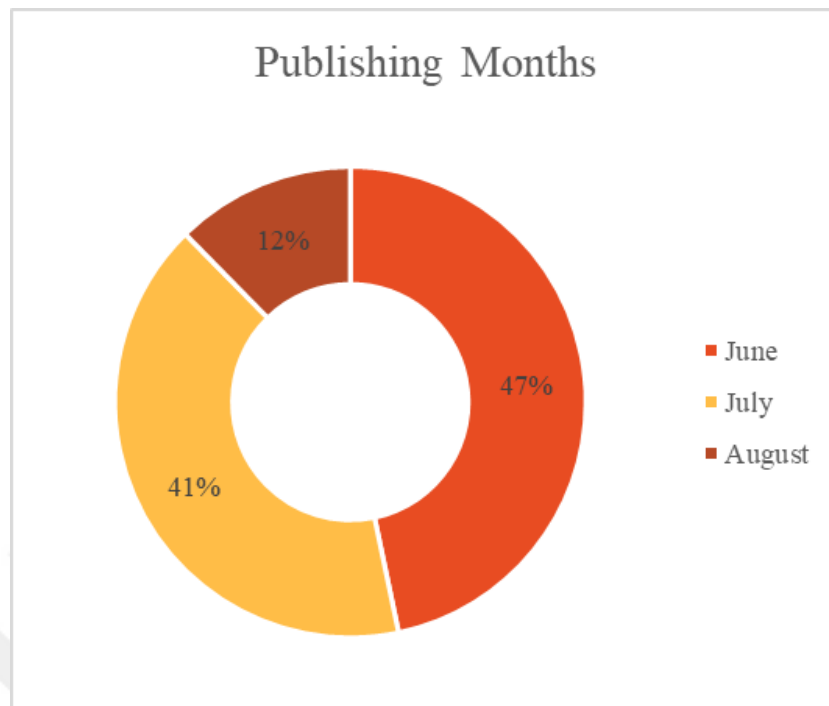
A great majority of books were released at the first weeks of three months. When we look at the weekdays of release dates, we encounter an interesting result, majority of the books were released on Tuesdays. Only five books were released on Sunday.

Table 4.3 Weekday Frequency of Publishing Dates

Day	Frequency
Monday	190
Tuesday	1984
Wednesday	76
Thursday	111
Friday	79
Saturday	8
Sunday	5

In terms of months, the majority of books were published in June and July as seen in Figure 4.6.

Figure 4.6 Publishing Percentages According to the Months

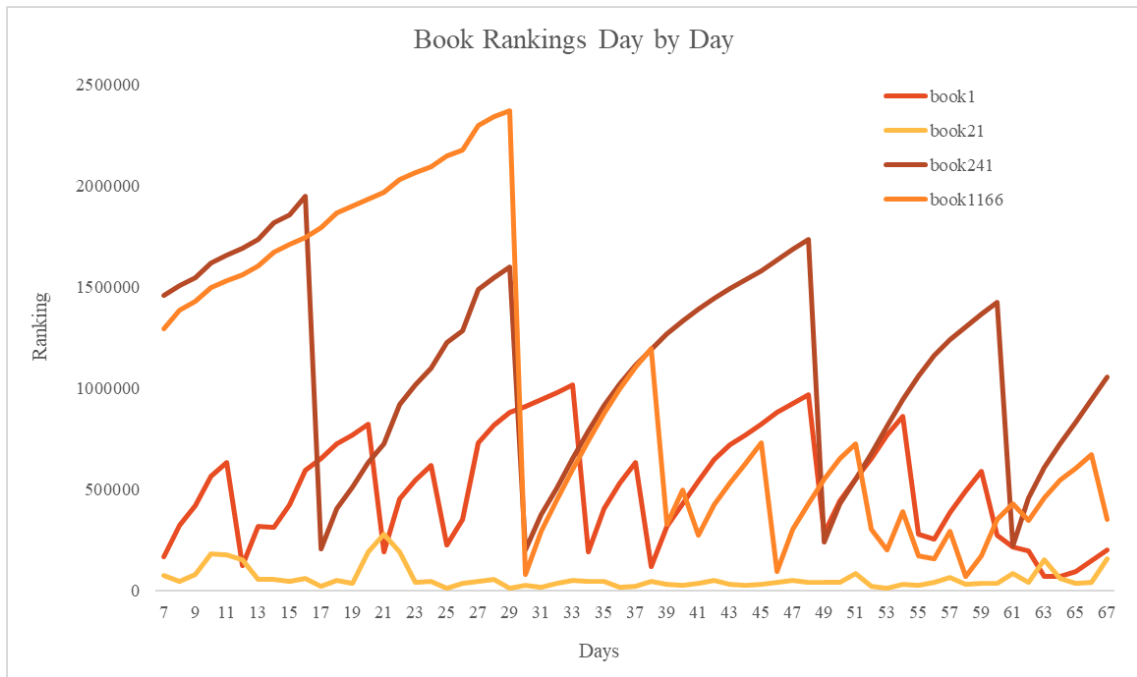


However, it should not be forgotten that our data has only two days for August, so it would be better to compare only June and July.

#### 4.6 ABRank

Amazon puts each book that sells at least one to the ranking system and constantly update the first 10,000 books rankings according to their recent and entire sales (Sornette et al., 2004). Hence, ABRank is a direct indicator variable that shows sales. An example of book rankings across days is given in Figure 4.7. The four books on the graphic are chosen randomly. Book21 has an overall minimum ranking.

Figure 4.7 A Sample of Book Rankings



After the sliding window method application with a window size of 6, we are left with 61 days, ranging between 7 and 67.

We also calculated the descriptive statistics of the sample books, given in Table 4.4. The data is skewed right in terms of the mean and median ratio in book21 and book1166, left skewed in book1 and book241. The CV is telling us the relative size of the standard deviation compared to the mean. For example, for book21, the size of the standard deviation is 88% of the mean, indicating that book21 has the relatively largest standard deviation in terms of ABRank, among the rest of the sample books.

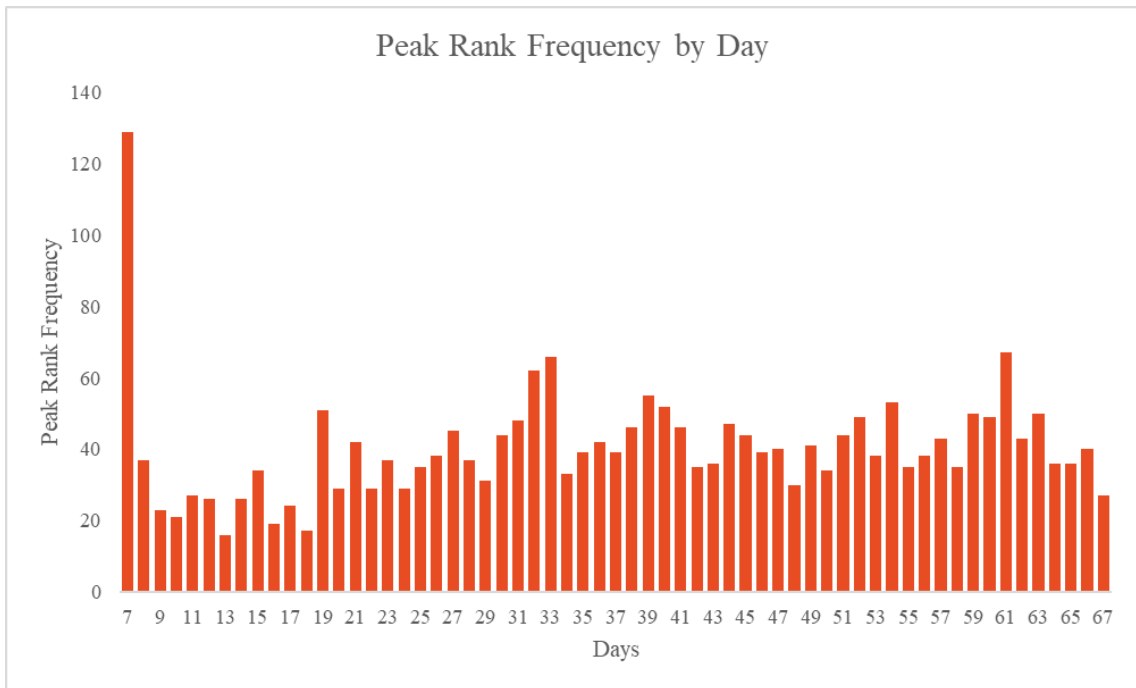
Table 4.4 Descriptive Statistics for ABRank Variable Sample

ABRank	book1	book21	book241	book1166
Mean	512708	60480.02	1105661	991052.5
Median	533151	43065	1163647	670284
Std.Dev.	273575	53630	481541	728662
Mean/Median	0.9617	1.4044	0.9502	1.4786
CV	0.5336	0.8867	0.4355	0.7352

We calculated the peak rankings by taking minimum of each 61 rows in the dataset. Then, we calculated the peak frequency. As seen in Figure 4.8, 7th day has the

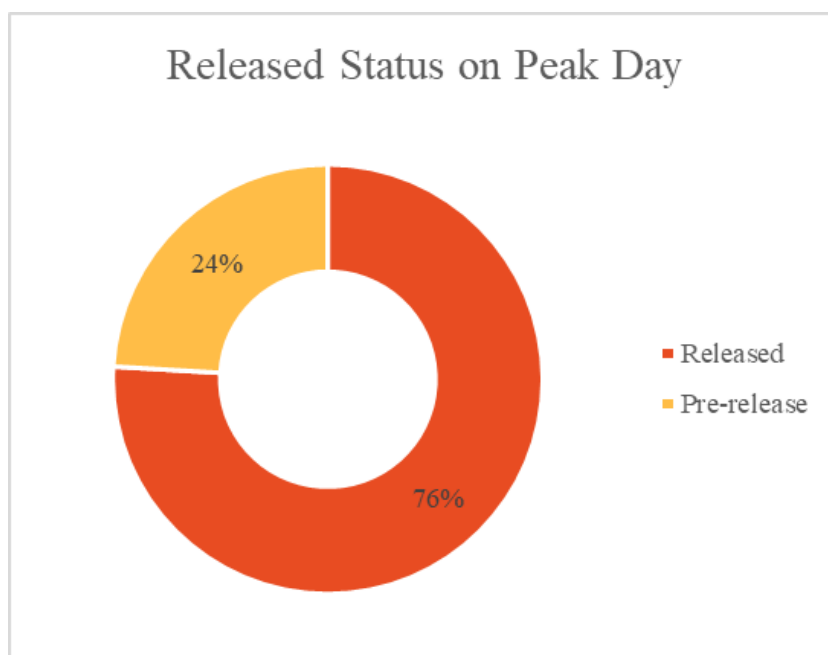
greatest peak frequency. However, this might arise from the absence of the first 6 days. On the other hand, the frequency distribution is balanced.

Figure 4.8 Peak Rank Frequency



We can observe that approximately 76% of books, which makes 1863 out of 2453 made their peak in the post-release period. Only 24% of them made their peak in pre-release period.

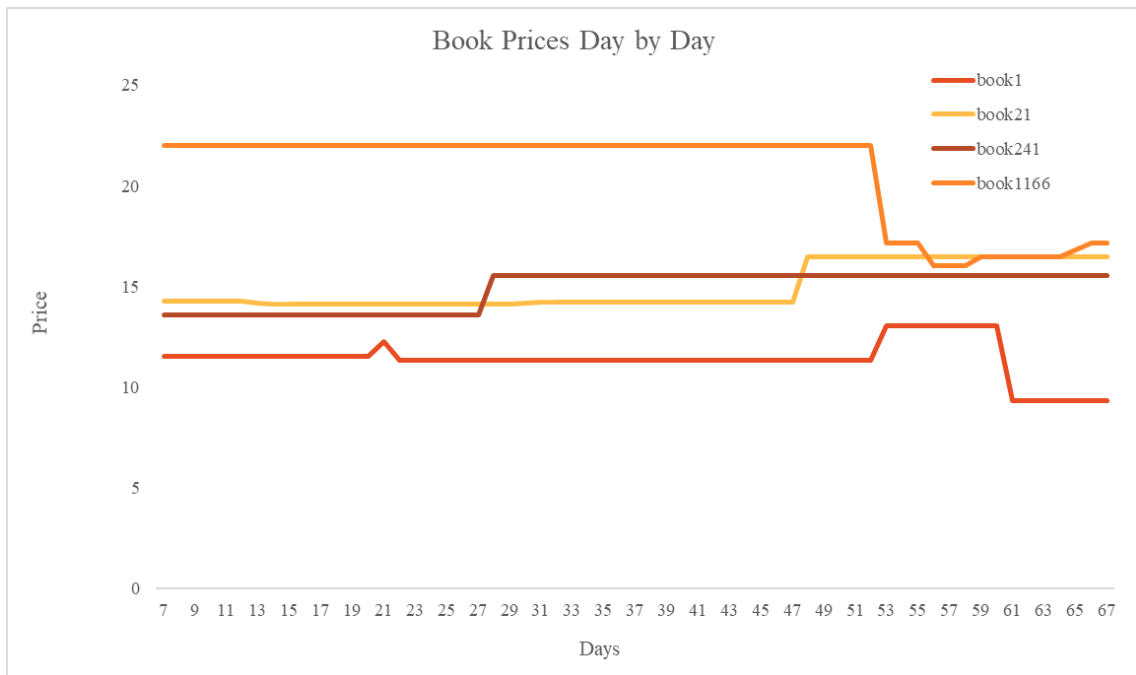
Figure 4.9 Released Status on Peak Day



## 4.7 Price

Like the ABRank variable, price variable changes over time. The price fluctuations of the same books that are used to explain ABRank is given below on Figure 4.10. We can observe that the price variable does not have major changes compared to the ABRank, which can be considered as normal.

Figure 4.10 A Sample of Book Prices



We also calculated the ratio between mean and median, and the coefficient of variation.

Table 4.5 Descriptive Statistics for Price Variable Sample

price variable	book1	book21	book241	book1166
Mean	11.390	14.935	14.875	20.675
Median	11.320	14.210	15.560	21.990
Std.Dev.	0.940	1.096	0.953	2.330
Mean/Median	1.006	1.051	0.956	0.940
CV	0.083	0.073	0.064	0.113

The mean and median ratio shows us the skewness of the data. For example, book1 and book21 are skewed to the right, while book241 and book1166 are skewed to

the left. Relatively, book1166 has the largest standard deviation compared to the mean.





## 5. METHODOLOGY

In this section, the methods that will be used will be explained. The first model that is used is Linear Regression. To select the features, Backward Elimination will be used and explained as a subsection. The second model is Ridge Regression, which is a type of Linear Regression with a quadratic shrinkage. Next, we will explain the two tree based models; Random Forest and Light Gradient Boosting Regressor. The data has many levels of different books, to fit the data properly for the tree based methods, hashing is needed. Hence, Hash Encoding will be explained as well. Finally, the feed forward Multilayer Perceptron will be explained.

### 5.1 Linear Regression

Our dataset has 24 variables, however, 16 of them are usable for linear regression. For example, we did not use the save, Genre, Physical\_format and ISBN13 variables.

One of our usable variables represents different books with 2453 categories, so we are left with 15 statistical units. Given this, one can create the following linear regression form for our dataset:

$$\begin{aligned} ABRank = & \beta_1 listprice + \beta_2 price + \beta_3 Releasedornot + \beta_4 lagprice1 \\ & + \beta_5 lagprice2 + \beta_6 lagprice3 + \beta_7 lagprice4 + \beta_8 lagprice5 \\ (5.1) \quad & + \beta_9 lagprice6 + \beta_{10} lagrank1 + \beta_{11} lagrank2 + \beta_{12} lagrank3 \\ & + \beta_{13} lagrank4 + \beta_{14} lagrank5 + \beta_{15} lagrank6 + \varepsilon \end{aligned}$$

In linear regression, there are many methods to make feature selection that is going to be used in the model, such as step-wise regression, forward selection, and

backward elimination. In this study, the backward elimination model is used.

### 5.1.1 Backward Elimination Model

The first step of backward elimination is to include all features in the model. After the first run, the feature with the highest p-value, which should also be greater than the threshold of 0.05, is eliminated. In each iteration, the same process is repeated until all the remaining features have lower p-values than the threshold. Since this data has many categories of books, the dummified categorical variables except the released status are omitted during this stage.

Table 5.1 Linear Regression Full Variables Model

	coef	std err	t	P> t	[0.025	0.975]
<b>Intercept</b>	123100.0	58900	2.089	0.037	7595.984	239000
<b>C(Releasedornot)</b>	-38750	3524.17	-10.995	0	-45700	-31800
<b>listprice</b>	4474.44	8841.87	0.506	0.613	-12900	21800
<b>price</b>	-6724.85	3611.73	-1.862	0.063	-13800	354.105
<b>lagprice1</b>	9647.236	5262.311	1.833	0.067	-666.831	20000
<b>lagprice2</b>	-5069.38	5339.05	-0.949	0.342	-15500	5395.105
<b>lagprice3</b>	-1025.00	5330.24	-0.192	0.848	-11500	9422.21
<b>lagprice4</b>	1447.24	5361.19	0.27	0.787	-9060.63	12000
<b>lagprice5</b>	-1329.45	5290.63	-0.251	0.802	-11700	9040.13
<b>lagprice6</b>	-367.01	3610.46	-0.102	0.919	-7443.48	6709.46
<b>lagrank1</b>	0.8543	0.003	267.88	0	0.848	0.861
<b>lagrank2</b>	0.0235	0.004	5.617	0	0.015	0.032
<b>lagrank3</b>	0.0064	0.004	1.53	0.126	-0.002	0.015
<b>lagrank4</b>	0.0033	0.004	0.791	0.429	-0.005	0.011
<b>lagrank5</b>	-0.0004	0.004	-0.102	0.918	-0.008	0.008
<b>lagrank6</b>	-0.0054	0.003	-1.749	0.08	-0.011	0.001

As it can be seen from the first fit results, the lagprice6 variable has the highest p-value which is 0.919. Therefore this variable is omitted in the second model.

Table 5.2 Linear Regression Final Model

	coef	std err	t	P> t	[0.025	0.975]
<b>Intercept</b>	123100	58900	2.089	0.037	7595.984	239000
<b>C(Releasedornot)</b>	-38810	3523.712	-11.015	0	-45700	-31900
<b>price</b>	-2783.829	1390.193	-2.002	0.045	-5508.59	-59.068
<b>lagrank1</b>	0.8543	0.003	267.934	0	0.848	0.861
<b>lagrank2</b>	0.0237	0.004	5.656	0	0.015	0.032
<b>lagrank3</b>	0.0084	0.003	2.424	0.015	0.002	0.015
<b>lagrank6</b>	-0.0047	0.002	-2.295	0.022	-0.009	-0.001

After 10 iterations, above is our final model. 10 out of 15 variables are not usable because of their p-values bigger than the threshold of .05, which left us the following five variables. price, lagrank1, lagrank2, lagrank3 and lagrank6.

## 5.2 Ridge Regression

Ridge Regression requires the data to be scaled to produce better results. The data is scaled to have the unit standard deviation. To find the best shrinkage coefficient, we tried 10-fold Cross Validation with values of 0.001, 0.01, 0.1, 1, 10, 50, 100, 150 as alphas. This makes  $8 \times 10 = 80$  fits. The best alpha value chosen is 0.1. Python's Scikit-Learn library and RidgeCV function is used to decide on the best alpha.

## 5.3 Random Forest

Random Forest is a decision tree-based ensemble algorithm that produces the prediction results based on the results of decision trees that are in the forest. To prevent overfitting, the trees in the random forest choose random samples from the original dataset. Random Forest algorithm is applied by using Python's Scikit-Learn library. There is a great variety of parameters to tune, so 3-fold cross-validation is used. The parameters and their values of the trial are given in Table 5.3.

Table 5.3 Parameter Trials for Random Forest Regressor

Parameter Name	Values
max_depth	{60, 80, 100}
min_samples_leaf	{50, 60, 70}
min_samples_split	{100, 150, 200}
n_estimators	{100, 200, 300}

There are  $1*3*3*3*3 = 81$  different parameter combinations and  $k = 3$  folds, which means there are 243 different model fits. Grid Search Cross-Validation chose the following parameters.

Table 5.4 Parameter Selection for Random Forest Regressor

Parameter Name	Value
max_depth	80
min_samples_leaf	50
min_samples_split	200
n_estimators	100

### 5.3.1 Hash Encoding

The dataset has 2453 different books, which means that there are 2453 different categories. In linear models and ridge regression, the high dimensionality due to dummification does not cause a problem, however, in tree-based methods such as Random Forest, high cardinality would likely result in the trees to be too deep and cause memory issues. To resolve that, hash encoding can be used. Hash encoding uses a hash function to transform the multi-leveled categories into a desired number of categories that are called components. However, decreasing the number of categories leads to the issue of collision. Collision rises from two or more categories having the same hash function value. For example, two or more different books might have the same hash function value, which means that they belong to the same category even if they are not. Although it has negative sides, it still works well with most of the algorithms as it reduces the number of dimensions. This is a trade off between the curse of dimensionality and collision.

## 5.4 LightGBM

LightGBM is a fast, low memory consuming gradient boosting algorithm which is based on decision trees (Mandot, 2018). The main difference between LightGBM and other tree-based algorithms is that LightGBM grows decision trees from leaves (leaf-wise) rather than branches (level-wise). Similar to other boosting methods, this algorithm grows decision trees by learning from the previous trees that are built. Therefore, the learning rate is added to the parameters to be tuned for this algorithm. Python's LightGBM library is used.

Figure 5.1 Normal Boosting Algorithms from (Mandot, 2018)

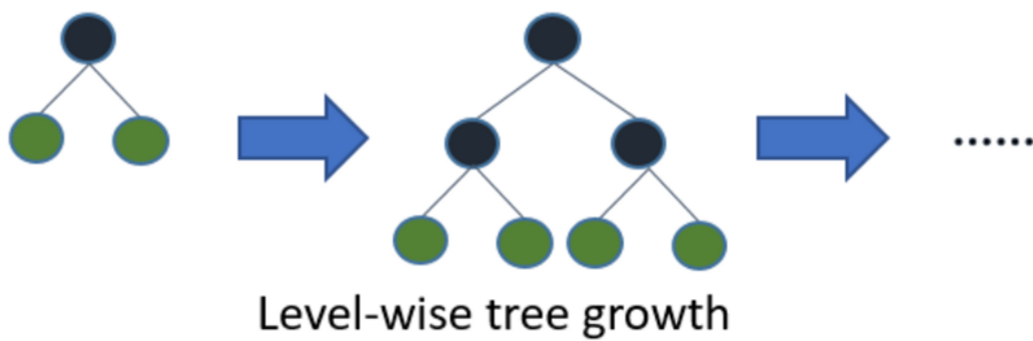
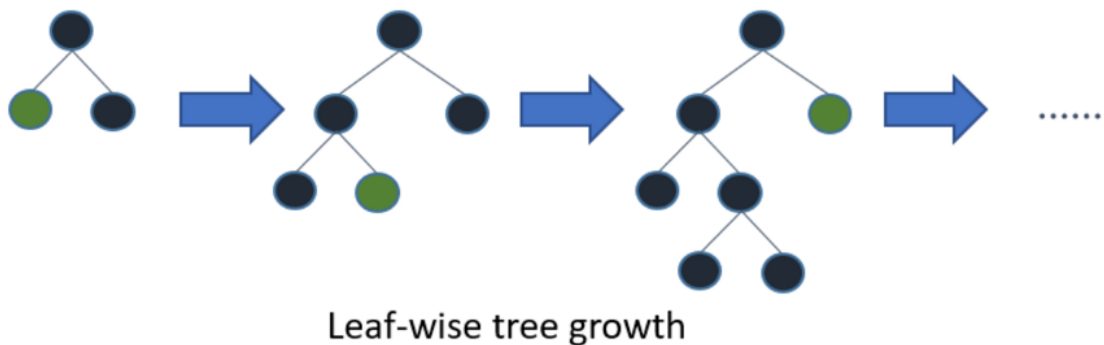


Figure 5.2 LightGBM from (Mandot, 2018)



The parameter tuning is again made with Grid Search Cross Validation. Parameters

and their values of trials are given on Table 5.5

Table 5.5 Parameter Combinations for LightGBM

Parameter Name	Values
max_depth	{30, 40, 50, 60}
learning_rate	{0.001, 0.05, 0.02, 0.01}
num_leaves	{5, 10, 20, 30, 40}
n_estimators	{100, 200, 300, 400, 500}

There are  $4 \times 4 \times 5 \times 5 = 400$  different parameter combinations and  $k = 5$  folds, meaning that there are 2000 model fits. Cross Validation chose the following parameters.

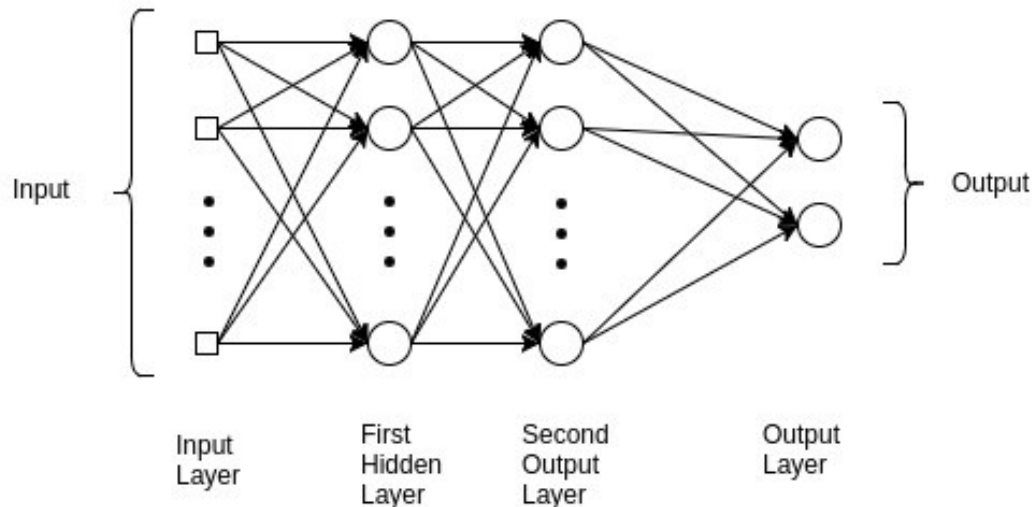
Table 5.6 Parameter Selection for LightGBM

Parameter Name	Value
max_depth	30
learning_rate	0.02
num_leaves	5
n_estimators	400

## 5.5 Multilayer Perceptron

Multilayer Perceptron is a feed forward kind of an Artificial Neural Network, which does not have any inner cyclic connections between its nodes because it feeds the data in one direction. An MLP has at least three layers of nodes. Each node can also be called as neurons. The first layer, which is the input layer, does not have an activation function that fires the neurons. The hidden layers have a nonlinear activation function and they carry the information by cyphering them from layer to layer. The last layer, which is called the output layer, gives the results. Below is an example of a two hidden layered MLP model.

Figure 5.3 Two Hidden Layered MLP architecture (Kain, 2019)



The process of a feed-forward neural network is given as follows: First, a random set of weights is constituted and with the initial values, they are fed to the first hidden layer. Then, an activation function is applied to the input. This process is repeated from a hidden layer to another until we reach the output layer. The output layer gives the predictions. After that, the backpropagation stage begins. With the learning rate, the weights are updated by using stochastic gradient descent optimization. The number of updates will go on until a certain threshold is met or the number of iterations is reached.

We used Python's Tensorflow library and used the GPU to get the results faster. We did not make hyperparameter tuning, because the techniques that are used to tune the hyperparameters are out of this study's scope.

## 5.6 Evaluation Criteria

There are two evaluation criterias used while comparing the models. First one is the  $R^2$  measurement or the Coefficient of Determination, which shows the explained variance of the output variables by accounting the input variables of the model.  $R^2$

measure is calculated as follows:

$$(5.2) \quad R^2 = SSR/SST$$

where;

$$SSR = \sum_i (\hat{y}_i - \bar{y})^2$$

$$SST = \sum_i (y_i - \bar{y})^2$$

The bigger  $R^2$  we have, the better model we have. Second evaluation criteria is the Root Mean Squared Error. The RMSE formula is given below.

$$(5.3) \quad RMSE = \sqrt{\left(\sum_{i=1}^n (\hat{y}_i - y_i)^2 / n\right)}$$

The smaller RMSE means better prediction results.

$R^2$  is a relative indicator while RMSE is an absolute indicator of goodness of fit.



## 6. RESULTS

In this section we will present the results. First we will explain how the Release Date affects the book rankings through linear regression. Then, we will interpret the prediction results for Linear Regression, Ridge Regression, Random Forest, LightGBM and Multilayer Perceptron respectively. Lastly, we will compare the results and decide on the best model.

### 6.1 Release Date Effect on ABRank

We are investigating whether the release status or release date affects the ABRank. According to the linear regression that ran on the full dataset, we can conclude that the release status of a book has a positive effect on ABRank. The result table is given below. The p-value for the variable Releasedornot is extremely low, very close to zero, indicating that this variable's effect on predicting ABRank cannot be zero. This means the rejection of the null hypothesis. From Table 6.1, we can also interpret that by 95% chance, the real beta coefficient of Releasedornot variable is between -38,100 and -29,000. Also, holding all the other predictors fixed, we can see that one unit change in this variable causes the ABRank to decrease 33,530 units, meaning that it has a positive effect on rankings.

Table 6.1 Results for Full Data Linear Regression

variable	coef	std err	t	P> t	[0.025	0.975]
Intercept	-200500	94800	-2.114	0.035	-386000	-14600
Releasedornot	-33530	2309.31	-14.518	0	-38100	-29000
listprice	15870	4928.266	3.22	0.001	6208.786	25500
price	-1930.88	2621.412	-0.737	0.461	-7068.8	3207.036
lagprice1	7126.182	3868.305	1.842	0.065	-455.62	14700
lagprice2	257.8939	3916.202	0.066	0.947	-7417.78	7933.572
lagprice3	-7398.18	3922.185	-1.886	0.059	-15100	289.228
lagprice4	2804.104	3943.581	0.711	0.477	-4925.24	10500
lagprice5	-2176.24	3929.729	-0.554	0.58	-9878.43	5525.951
lagprice6	682.3867	2675.943	0.255	0.799	-4562.41	5927.181
lagrank1	0.8651	0.003	333.622	0	0.86	0.87
lagrank2	0.0219	0.003	6.373	0	0.015	0.029
lagrank3	0.0107	0.003	3.144	0.002	0.004	0.017
lagrank4	0.0058	0.003	1.71	0.087	-0.001	0.012
lagrank5	-7E-05	0.003	-0.021	0.983	-0.007	0.007
lagrank6	0.0014	0.003	0.571	0.568	-0.003	0.006

From the table above, we can transform the Equation 5.4 to the following equation.

(6.1)

$$\begin{aligned}
 ABRank = & -200,500 - 33,500 * releasedornot + 15,870 * listprice \\
 & -1,930.88 * price + 7,126 * lagprice1 + 257.89 * lagprice2 - 7,398.18 * lagprice3 \\
 & + 2,804.1 * lagprice4 - 2,176.24 * lagprice5 + 682.38 * lagprice6 + 0.86 * lagrank1 \\
 & + 0.0219 * lagrank2 + 0.0107 * lagrank3 + 0.0058 * lagrank4 - 7e^{-5} * lagrank5 \\
 & + 0.0014 * lagrank6 + \sum_{i=1}^{2453} \beta_i panel_i
 \end{aligned}$$

## 6.2 Linear Regression

The table for the resulting linear regression is given below. All of the low p-values indicate that these variables' effect on ABRank variable cannot be zero, means the rejection of null hypothesis. Releasedornot: Holding all of the variables fixed, a unit

change in this variable effects the ranking in a positive way, decreasing it 38,810 units. By 95% chance, the real change that this variable will cause is between 45,700 and 31,900. Price: The decreasing effect of price shows that more expensive books have a better ranking. However, the standard error term for this variable is relatively large, making the real decrease that it causes changes from 5,508 to 59 by 95% chance. LagRanks: All of the lagrank variables except lagrank6 has a negative effect on ranking. For example, a unit change in lagrank1 variable causes the ABRank shift by 0.8543 units. This may seem like a small change, but the rank terms are usually stated with thousands, so the change will not be small.

Table 6.2 Linear Regression Result Model

	coef	std err	t	P> t	[0.025	0.975]
<b>Intercept</b>	123100	58900	2.089	0.037	7595.984	239000
<b>C(Releasedornot)</b>	-38810	3523.712	-11.015	0	-45700	-31900
<b>price</b>	-2783.829	1390.193	-2.002	0.045	-5508.59	-59.068
<b>lagrank1</b>	0.8543	0.003	267.934	0	0.848	0.861
<b>lagrank2</b>	0.0237	0.004	5.656	0	0.015	0.032
<b>lagrank3</b>	0.0084	0.003	2.424	0.015	0.002	0.015
<b>lagrank6</b>	-0.0047	0.002	-2.295	0.022	-0.009	-0.001

The equation of the resulting model is given below:

$$\begin{aligned}
 (6.2) \quad ABRank = & -123,100 - 38,810 * releasedornot - 2,783.82 * price \\
 & + 0.854 * lagrank1 + 0.0237 * lagrank2 + 0.0084 * lagrank3 \\
 & - 0.0047 * lagrank6 + \sum_{i=1}^{2453} \beta_i panel_i
 \end{aligned}$$

The evaluation criteria for linear regression is given on Table 6.3

Table 6.3 Evaluation Criteria for Linear Regression

	RMSE-train	r2-train	RMSE-test	r2-test
Linear Regression Model	358817.58	95%	290288.16	95%

The model predicted the test set with an  $R^2$  value of 95% which reveals that 95% of the variance in the data can be explained by this model. The fact that  $R^2$  of training set is close to the  $R^2$  of the test set means that this model does not suffer from overfitting. The lower root mean squared error of the test set (RMSE) also supports that claim.

### 6.3 Ridge Regression

In order to interpret the Ridge Regression, we ran a Linear Regression model with the scaled variables. The resulting coefficients table is given below.

Table 6.4 Ridge Model vs. Linear Regression Model

Variable	Ridge Model	Linear Model
listprice	121447.05	126579.12
price	-178393.75	-177576.20
Releasedornot	-19247.47	-19413.65
lagrank1	1381395.81	1381377.50
lagrank2	38317.67	38463.76
lagrank3	10454.02	10316.32
lagrank4	5389.58	5379.43
lagrank5	-691.30	-687.56
lagrank6	-8958.20	-8966.42
lagprice1	254415.17	258527.62
lagprice2	-132736.50	-137842.21
lagprice3	-28480.43	-22598.30
lagprice4	37962.69	35355.32
lagprice5	-34884.11	-40508.16
lagprice6	-10251.11	-4928.95

As it can be seen above, the listprice, Releasedornot, lagrank2, lagrank6, lagprice1, lagprice2, and lagprice5 variables lose their absolute effect on the ABRank variable while the remaining variables gain more absolute effect. However, the losses and gains are not very significant due to a small shrinkage penalty chosen by Cross-Validation.

The model performance is given below on Table 6.5.

Table 6.5 Evaluation Criteria for Ridge Regression

	RMSE-train	r2-train	RMSE-test	r2-test
Ridge Regression	358808.43	95%	524373.3	83%

Ridge Regression performed slightly better than Linear Regression, however it performed poorly on the test set.

## 6.4 Random Forest

According to Table 6.6, the most significant variable in predicting ABRank is lagrank1, which means the previous day's rank. Second, third and the fourth most important variables are lagrank2, lagrank3, and lagrank6, which are also included in the linear regression model. We can claim that the importance results proves the necessity of these variables' predictive power on ABRank.

Table 6.6 Variable Importances for Random Forest

Variable	Importance
lagrank1	0.992156
lagrank2	0.004579
lagrank3	0.001043
lagrank6	0.000637
lagrank5	0.000412
lagrank4	0.000409
listprice	0.000178
price	0.000125
lagprice6	9.8E-05
lagprice1	7.94E-05
lagprice5	6.67E-05
Releasedornot	6.34E-05
lagprice4	6.13E-05
lagprice3	4.82E-05
lagprice2	4.39E-05

Random Forest model performed better than Ridge Regression and Linear Regression in terms of RMSE in training set, however it performed poorly than Linear Regression in terms of both  $R^2$  and RMSE.

Table 6.7 Evaluation Criteria for Random Forest

	RMSE-train	r2-train	RMSE-test	r2-test
Random Forest	356983.25	95%	497323.43	85%

## 6.5 LightGBM

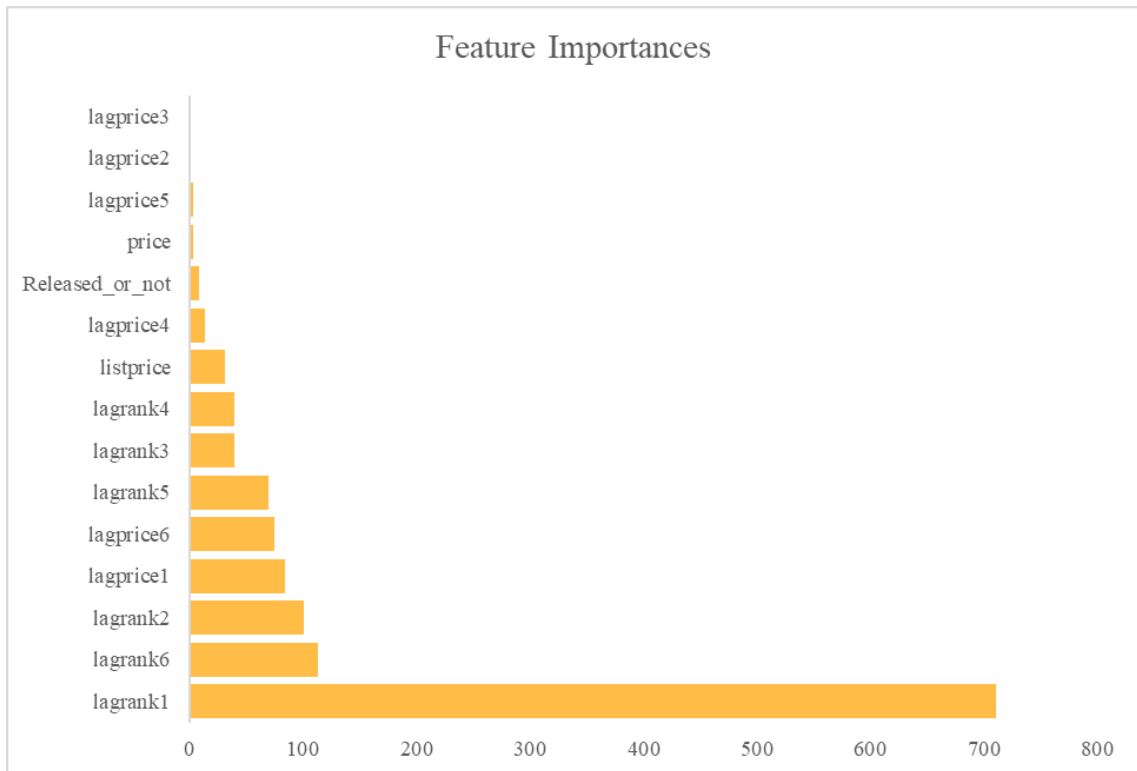
LightGBM, the gradient boosting machine algorithm gave similar results as random forest. According to Table 6.8, the most important feature is lagrank1 again. However, lagrank3 is not as important as in the Random Forest. We saw that lagprice1 variable, which is the price from the previous day gained importance.

Table 6.8 Variable Importances for LightGBM

Variable	Importance
lagrank1	711
lagrank6	113
lagrank2	101
lagprice1	84
lagprice6	75
lagrank5	70
lagrank3	39
lagrank4	39
listprice	31
lagprice4	13
Releasedornot	8
price	3
lagprice5	3
lagprice2	1
lagprice3	1

Below, the bar graph of importances can be seen.

Figure 6.1 Feature Importances for LightGBM



The gradient boosting algorithm performed slightly worse than the Random Forest. The results are given below on Table 6.9

Table 6.9 Evaluation Criteria for LightGBM

	RMSE-train	r2-train	RMSE-test	r2-test
LightGBM	372525.99	95%	567593.43	80%

## 6.6 Multilayer Perceptron

With 4 hidden layers of 100, 75, 50 and 25 nodes respectively, the Multilayer Perceptron gives the following results.

Table 6.10 Evaluation Criteria for Multilayer Perceptron

	RMSE-train	r2-train	RMSE-test	r2-test
MLP	373769.56	95%	584787.38	87%

## 6.7 Overall Results

All of the results stated in the previous sections are summarized in the following table.

Table 6.11 Overall Results

Model	RMSE-train	r2-train	RMSE-test	r2-test
Linear Regression	358817.58	95%	290288.16	95%
Ridge Regression	358808.43	95%	524373.29	83%
Random Forest	356983.25	95%	497323.43	85%
LightGBM	372525.99	95%	567593.43	80%
MLP	373769.56	95%	584787.38	87%

To sum up, Linear Regression gave the best performance, followed by Multilayer Perceptron and Random Forest in test dataset.



## 7. CONCLUSION

We conducted a study about measuring the effect of release dates on book sale ranks and found out that the release status of a book has a significant effect when predicting the ranking. We also tried to predict the rankings using five methods, which are Linear Regression, Ridge Regression, Random Forest, Light Gradient Boosting Machine, and Multilayer Perceptron. Linear Regression outperformed the rest of the models in both terms of RMSE and  $R^2$ . We found out that the price and lagged price variables do not have a significant effect on predicting ABRank while lagged ranks play a critical role.

Although it is expected that a deep learning model and tree-based models will outperform the traditional models, linear regression outperformed all of the other methods. The reason behind this circumstance might be the hyperparameter selection. For example, the neural network could result better if the number of hidden layers and nodes were chosen correctly. Also, if the parameter combinations of tree methods were increased, the results could be better however, the memory consumption and computation time is a great obstacle for this study.

There are some limitations and assumptions to our study. First of all, we eliminated a great amount of the data set due to the lack of data points. Secondly, the data involves only newly released books, so one might say that the data is biased because we could not observe the rest of the books. Also, we believe that the publisher of a book plays a crucial role in predictions however, we were not able to include them in our models because there is a great variety of publishers for a small sample of books. We believe that larger publishers are more likely to sell more books. We were not able to measure the effect of positive and negative reviews, which we think is important in predicting a book's rank due to missing data. We assumed that rankings are direct indicators of sales. Since we do not know the basket information, we could not know whether there is a cross effect between books.

This study might be repeated in a genre base with a larger dataset. Since the dataset is small and genre information is limited to fiction and non-fiction categories, it would

be meaningless to separate the data. A study that is based on genre might help the publishers or e-commerce sites to make better and personalized advertisements that will help to increase the sales.

In addition to the current study, daily numbers of negative and positive comments could be added and the predictions for ABRank could be improved. Furthermore, sentiment analysis of comments could be conducted and a score of effective comments can be measured. With market basket information and comments, one could measure the word of mouth effect and behavior of the buyers.

Also, the causes that make a book sell during the pre-release period could be researched. There might be several reasons behind a sale before release, for example, the choice of the cover material might affect the pre-sales. Hardcover books might have more potential than Paperback books in the pre-release period or vice versa. These kinds of information might give publishers an insight about increasing the sales, or in this case, increasing the rank.

## BIBLIOGRAPHY

- Ahmed, N. K., Atiya, A. F., Gayar, N. E., & El-Shishiny, H. (2010). An empirical comparison of machine learning models for time series forecasting. *Econometric Reviews*, 29(5-6), 594–621.
- Castillo, P. A., Mora, A. M., Faris, H., Merelo, J., García-Sánchez, P., Fernández-Ares, A. J., De las Cuevas, P., & García-Arenas, M. I. (2017). Applying computational intelligence methods for predicting the sales of newly published books in a real editorial business management environment. *Knowledge-Based Systems*, 115, 133–151.
- Chang, P.-C. & Lai, C.-Y. (2005). A hybrid system combining self-organizing maps with case-based reasoning in wholesaler’s new-release book forecasting. *Expert Systems with Applications*, 29(1), 183–192.
- Chevalier, J. & Goolsbee, A. (2003). Measuring prices and price competition online: Amazon. com and barnesandnoble. com. *Quantitative marketing and Economics*, 1(2), 203–222.
- Fenner, T., Levene, M., & Loizou, G. (2010). Predicting the long tail of book sales: Unearthing the power-law exponent. *Physica A: Statistical Mechanics and Its Applications*, 389(12), 2416–2421.
- Frank, R. J., Davey, N., & Hunt, S. P. (2001). Time series prediction and neural networks. *Journal of intelligent and robotic systems*, 31(1-3), 91–103.
- Hota, H., Handa, R., & Shrivastava, A. (2017). Time series data prediction using sliding window based rbf neural network. *International Journal of Computational Intelligence Research*, 13(5), 1145–1156.
- Kain, N. K. (2019). Understanding of multilayer perceptron (mlp).
- Kayacan, E., Ulutas, B., & Kaynak, O. (2010). Grey system theory-based models in time series prediction. *Expert systems with applications*, 37(2), 1784–1789.
- Kocas, C., Pauwels, K., & Bohlmann, J. D. (2018). Pricing best sellers and traffic generators: the role of asymmetric cross-selling. *Journal of Interactive Marketing*, 41, 28–43.
- Mandot, P. (2018). What is lightgbm, how to implement it? how to fine tune the parameters?
- McMullen, C. (2018). How does amazon.com sales rank work?
- Mozaffari, L., Mozaffari, A., & Azad, N. L. (2015). Vehicle speed prediction via a sliding-window time series analysis and an evolutionary least learning machine: A case study on san francisco urban roads. *Engineering science and technology, an international journal*, 18(2), 150–162.
- Qian, X.-Y. & Gao, S. (2017). Financial series prediction: Comparison between precision of time series models and machine learning methods. *arXiv preprint arXiv:1706.00948*.
- Sapankevych, N. I. & Sankar, R. (2009). Time series prediction using support vector machines: a survey. *IEEE Computational Intelligence Magazine*, 4(2), 24–38.
- Sornette, D., Deschâtres, F., Gilbert, T., & Ageon, Y. (2004). Endogenous versus exogenous shocks in complex networks: An empirical test using book sale rankings. *Physical Review Letters*, 93(22), 228701.
- Wang, X., Yucesoy, B., Varol, O., Eliassi-Rad, T., & Barabási, A.-L. (2019). Success

in books: predicting book sales before publication. *EPJ Data Science*, 8(1), 31.

Yahmed, Y. B., Bakar, A. A., Hamdan, A. R., Ahmed, A., & Abdullah, S. M. S. (2015). Adaptive sliding window algorithm for weather data segmentation. *Journal of Theoretical and Applied Information Technology*, 80(2), 322.



## APPENDIX A

### A Glimpse of The Data

ID	16	18
title	The Hare with Amber Eyes: A Hidden Inheritance (9780312569372): Edmund de Waal	Cradle of Gold: The Story of Hiram Bingham a Real-Life Indiana Jones and the Search for Machu Picchu (9780230112049): Christopher Heaney
ISBN10	a-0312569378	a-0230112048
ISBN13	978-0312569372	978-0230112049
ASIN		
listprice	16	17
price	10.3	11.19
You Save	5.7	5.81
You Save %	36	34
ABRank	1175	70554
retailers		50
soldbyamazon	YES	YES
physical_format	Paperback	Paperback
Publisher	Picador; First Edition edition	Palgrave Macmillan; Reprint edition
publishing_date	August 2 2011	July 5 2011
Language	English	English
avg_cus_review	4.5	4.2
numberoflike	4	1
Category		
link	<a href="http://www.amazon.com/Hare-Amber-Eyes-Hidden-Inheritance/dp/0312569378/ref=sr_1_1?s=books&amp;ie=UTF8&amp;qid=1312124592&amp;sr=1-1">http://www.amazon.com/Hare-Amber-Eyes-Hidden-Inheritance/dp/0312569378/ref=sr_1_1?s=books&amp;ie=UTF8&amp;qid=1312124592&amp;sr=1-1</a>	<a href="http://www.amazon.com/Cradle-Gold-Bingham-Real-Life-Indiana/dp/0230112048/ref=sr_1_1?s=books&amp;ie=UTF8&amp;qid=1312124592&amp;sr=1-1">http://www.amazon.com/Cradle-Gold-Bingham-Real-Life-Indiana/dp/0230112048/ref=sr_1_1?s=books&amp;ie=UTF8&amp;qid=1312124592&amp;sr=1-1</a>
date	31/07/2011	31/07/2011
time	00/01/1900	00/01/1900
total reviews	42	44
5 Star reviews	30	21
4 Star reviews	4	15
3 Star reviews	8	6
2 Star reviews	0	2
1 Star reviews	0	0
Amazon extra Rank 1	1 Books > Arts & Photography > Other Media > Ceramics	56 Books > History > Americas > South America
Amazon extra Rank 2	2 Books > Arts & Photography > Schools Periods & Styles > Modern	
Amazon extra Rank 3	3 Books > Arts & Photography > History & Criticism	
Amazon extra Rank 4		

## Publisher Counts of the Data

Table A.1 Publisher Counts

<b>publisher</b>	<b>book_count</b>
Brilliance	232
HarperCollins	148
St. Martin's Griffin	92
Palgrave Macmillan	85
Berkley	56
Arcadia Publishing	55
Signet	55
Random House	41
Berkley Trade	40
Penguin	38
DK	35
Tor Books	33
Minotaur Books	31
NAL Trade	30
Broadway	25
Tantor Media	25
Hyperion	24
Candlewick	23
Kensington	23
Avon	21
Bloomsbury USA	21
Crown	21
Knopf	21
Ballantine Books	20
Thorndike Press	20
Ace	18
Farrar Straus and Giroux	18
Vintage	18
Large Print Press	17
William Morrow	17
Macmillan Audio	16
NYU Press	16
Simon & Schuster	16
Zebra	16
Bantam	15

Dutton	15
Forge Books	15
Henry Holt and Co.	15
Princeton Review	15
Roc	15
Alpha	14
DAW	13
Hachette Audio	13
Pinnacle	13
Barron's Educational Series	12
Del Rey	12
Delacorte	12
Grand Central Publishing	12
Jove	12
Kingfisher	12
Picador	12
Roaring Brook Press	12
Shire	12
Urban Books	12
Viking	12
Anchor	11
Disney Press	11
Feiwel & Friends	11
Dafina	10
Dial	10
Five Star	10
North Atlantic Books	10
Walker Childrens	10
Katherine Tegen Books	9
Mira	9
Shambhala	9
Wizards of the Coast	9
Albert Whitman & Company	8
Doubleday	8
First Second	8
Forever	8
HQN Books	8
Overlook Hardcover	8

Papercutz	8
Puffin	8
teNeues	8
Thomas Dunne Books	8
Tor	8
WaterBrook Press	8
Blackstone Audio Inc.	7
College Board	7
Dell	7
Ecco	7
Melville House	7
Orbit	7
Perigee Trade	7
Pocket	7
Razorbill	7
Rodale Books	7
The University of North Carolina Press	7
Titan Books	7
Wheeler Publishing	7
A&C Black	6
AmazonEncore	6
EgmontUSA	6
Faber & Faber	6
Gale	6
Gotham	6
It Books	6
Little Brown and Company	6
Modern Library	6
Prestel Publishing	6
Putnam	6
Scholastic	6
Three Rivers Press	6
Angry Robot	5
Avery Trade	5
Center Point Pub	5
Cogito Media Group	5
Ember	5
Gallery	5



Greenhaven Press	5
Grosset & Dunlap	5
Listening Library (Audio)	5
Manchester University Press	5
McClelland & Stewart	5
Methuen Drama	5
Pantheon	5
Philomel	5
Plume	5
Portfolio Trade	5
Riverhead Trade	5
Spectra	5
Spiegel & Grau	5
Square Fish	5
Touchstone	5
Tricycle Press	5
Zondervan	5
AmazonCrossing	4
Aphrodisia	4
AudioGO	4
Berg Publishers	4
Bolinda Audio	4
Christian Large Print	4
Genesis Press	4
Graywolf Press	4
Greenwillow Books	4
HCI	4
I. B. Tauris	4
Mark Batty Publisher	4
Multnomah Books	4
powerHouse Books	4
Rutgers University Press	4
Seven Stories Press	4
Soho Crime	4
University of Washington Press	4
Voice	4
Balzer Bray	3
Beacon Press	3

British Film Institute	3
Collins Reference	3
HighBridge Company	3
House of Collectibles	3
Leisure Arts Inc.	3
Marvel Press	3
Monthly Review Press	3
NavPress	3
Oceanview Publishing	3
Osprey Publishing	3
Other Press	3
Quirk Books	3
Readers Digest	3
Speak	3
Tauris Academic Studies	3
The Colonial Radio Theatre on Brilliance Audio	3
Walker & Company	3
Wendy Lamb Books	3
Yearling	3
Amistad	2
Atria	2
Back Bay Books	2
Blue Apple Books	2
Citadel	2
DAAB MEDIA	2
DOM PUBLISHERS	2
Dreamscape Media	2
Europa Editions	2
Gospel Light	2
Hatherleigh Press	2
HP Trade	2
Kennebec Large Print	2
Kuperard	2
National Geographic Children's Books	2
Overlook TP	2
Pluto Press	2
Potter Craft	2
Price Stern Sloan	2

Regal	2
Riverhead Hardcover	2
Schwartz & Wade	2
Skira	2
Skira Rizzoli	2
Sleeping Bear Press	2
Smithsonian Books	2
Soho Constable	2
Soho Press	2
Tarcher	2
Ten Speed Press	2
Threshold Editions	2
Turtleback	2
Tyndale House Publishers	2
Urban Trade Paper	2
Vertical	2
Vision	2
Washington Square Press	2
Westminster John Knox Press	2
Akashic Books	1
Aladdin	1
Amphoto Books	1
Amy Einhorn Books/Putnam	1
Applause Theatre and Cinema Books	1
Archetype	1
Archie Comics	1
Arcturus	1
Arden Shakespeare	1
Atlantic Monthly Press	1
Avery	1
Barrytown/Station Hill Press Inc.	1
Beaufort Books	1
Between the Lines	1
Bluefire	1
BRADY GAMES	1
Brava	1
Broadside Books	1
Campfire	1

Center Street	1
Central Recovery Press	1
Chicken Soup for the Soul on Brilliance Audio	1
Cicerone Press Limited	1
Collins Design	1
Dasan Books	1
David Fickling Books	1
Dover Publications	1
Dundurn	1
Eos	1
EVOLVER EDITIONS	1
Flammarion	1
Fodor's	1
Golden Books	1
Hard Case Crime	1
Harmony	1
Hill and Wang	1
History Publishing Company LLC	1
HNL Publishing	1
Holt Paperbacks	1
Hudson Street Press	1
Igniter	1
Image	1
INDEX BOOKS	1
Level 4 Press Inc.	1
LifeStories	1
Little Bookroom	1
Little Brown Book Group	1
LucasBooks	1
Lucent Books	1
Mad Norwegian Press	1
McGraw-Hill; 1 edition	1
Metropolitan Books	1
Nan A. Talese	1
Nancy Paulsen Books	1
New Chapter Publisher	1
NYR Children's Collection	1
NYRB Classics	1

Overdue Media	1
Pamela Dorman Books	1
Pelican Publishing	1
Philip Wilson Publishers	1
Prentice Hall Press	1
PublicAffairs	1
Reagan Arthur Books	1
Reagan Arthur Books	1
RED DOT EDITION	1
Regnery Publishing	1
RH/Disney	1
Rizzoli	1
Robert Reed Publishers	1
Rough Guides	1
Scribner	1
Select Books	1
Sports Improper Publications	1
Starscape	1
Steerforth	1
Sterling	1
Tauris Parke Paperbacks	1
The Dial Press	1
The Domino Project	1
The Experiment	1
The Monacelli Press	1
Think and Grow Rich on Brilliance Audio	1
Top Shelf Productions	1
Transmedia Publishing	1
Transworld Publishers	1
TROLLEY BOOKS	1
Trumpeter	1
Tyndale House Publishers Inc.	1
Upper Access Inc.	1
UXL	1
Villard	1
W. W. Norton & Company	1
Watson-Guptill	1
Wiley	1

WND Books  
Zed Books

1  
1

