The Price of Books :

A Study of Books Published By Penguin Random House

Jordan Baldwin

September 2020

The Question:

Can we accurately predict the price of a book based on the data available from the Penguin Random House public api?

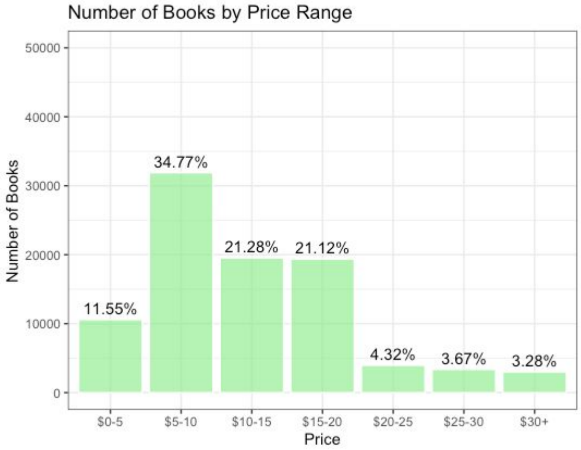
The Answer:

We can predict the price of a book with approximately 60% accuracy, and typically within $5. We are more likely to accurately predict the price for books under $30.

Introducing the Data

Our data consisted of 91K unique books available on sale from Jan-2010 to Feb-2018.

The price of our books ranged from $.99 to $100, after removing outliers, with the greatest concentration of books being between $5 and $10. The average price of a book was $13.28, with a median price of $10.99. Books greater than $30 were specialty editions, leather bound, or otherwise rare books. This information was discovered by independently researching the isbns from a random selection.



The key data points that we used to determine our predicted price were: the type of book (ebook, hardback, paperback, ect.), the publishing division, the category of book (Fiction, Nonfiction, Historical, ect.), the number of pages, the suggested age range, and the year and month that the book went on sale. Detailed information on the categorical data points can be found in Appendix A.

Method and Results of Analysis

Logistic Regression was the method of choice to create our final predictive model.

A first Model 1 found that the data point of suggested age range was not statistically significant and was cut from a second Model 2's coefficients. It was also found that the accuracy was greatly increased for the < $40 range, so a final Model 3 was created using only those. Model 2 and Model3 had very similar results.

	Adjusted R-squared	p-value	RMSE (All)	RMSE (<$40)	Avg. RMSE
Model 2	0.5956	2.2e-16	5.39	4.16	4.77
Model 3	0.6506	2.2e-16	5.56	3.98	4.77

Model 3 was our final choice of model based on the greater degree of accuracy for smaller prices, which was the majority of our data. The final coefficients can be seen in Appendix C.

Final Thoughts

Based on the high amount of spread in our data (CV = 1.5), and a lack of any significant correlation between price and our available data (correlation can be seen in Appendix B), we did not expect to be able to achieve a high degree of accuracy. Still, our analysis could benefit from a more accurate model or the inclusion of additional correlated data points. Text analysis of the books' themes could be a data point to explore in the future.

Appendix

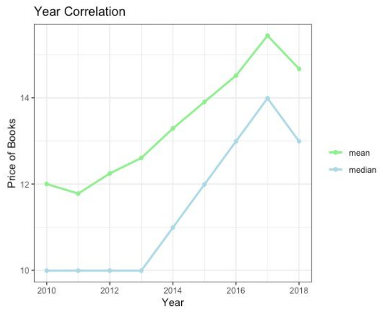
A) Categorical Data Details

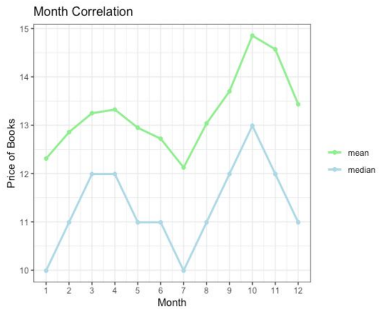
Division	Count	Mean	Median	Min	Max
Other	18024	16.66	13.99	0.99	99.99
Penguin Adult HC/TR	10858	14.48	14.00	1.99	90.00
Berkley / NAL	9328	9.20	7.99	0.99	54.99
RH Childrens Books	7664	8.56	6.99	0.99	50.00
Knopf	7430	12.77	12.99	0.99	100.00
Penguin Young Readers	5403	9.66	8.99	0.99	54.99
Potter/TenSpeed/Harmony/Rodale	3877	15.07	12.99	0.99	100.00
Ballantine Group	2848	10.18	7.99	0.99	74.99
Random House Group	2156	13.45	12.99	0.99	60.00
DK.	2127	13.71	10.99	0.99	60.00
Crown	2043	12.50	12.99	0.99	38.50
Bantam Dell	2028	10.40	8.99	0.99	62.99
Candlewick	1931	13.54	14.99	3.99	100.00
PRH Grupo Editorial	1921	16.27	15.95	5.99	89.99
National Geographic Society	1870	15.18	13.90	2.99	70.00
Christian/Forum	1834	10.53	9.99	0.99	30.00
Dark Horse Comics	1691	14.71	11.99	2.99	100.00
Titan	1535	14.55	9.95	0.99	99.99
Charlesbridge	1403	9.63	7.95	4.95	35.00
Verso Books	1331	21.93	16.95	4.99	100.00
Hay House	1313	11.43	9.99	0.99	69.95
Watkins Media	1090	12.15	9.99	0.99	90.00
Shambhala	1068	20.38	17.95	6.99	100.00
IDW Publishing	964	25.75	19.99	7.99	100.00

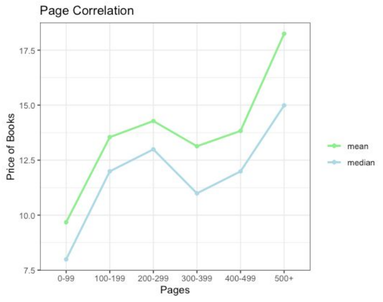
Type	Count	Mean	Median	Min	Max
eBook	45990	9.42	8.99	0.99	74.99
Board	1124	8.62	7.99	3.99	21.99
Hardcover	11375	24.90	19.99	3.99	100.00
Hardcover Library Binding	840	20.05	19.99	12.90	95.00
Paperback	4707	8.39	7.99	3.95	14.99
Trade Paperback	27701	15.73	16.00	3.99	90.00

Category	Count	Mean	Median	Min	Max
FIC	27463	11.14	9.99	0.99	75.00
ART	1404	25.39	19.95	3.99	95.00
BIO	4393	14.17	13.99	0.99	99.99
BUS	2052	16.64	15.00	0.99	99.95
CGN	4718	18.44	14.99	2.99	100.00
CKB	2889	17.58	15.99	0.99	100.00
CRA	1173	17.28	15.99	0.99	45.00
FAM	972	12.59	12.99	0.99	28.00
HEA	2059	14.89	13.99	0.99	90.00
HIS	2652	16.01	14.99	0.99	95.00
HUM	928	13.86	12.95	0.99	75.00
JNF	5784	10.81	8.99	0.99	40.00
JUV	13752	9.69	7.99	0.99	100.00
OCC	1399	13.31	11.99	0.99	78.00
OTH	9319	17.29	14.99	0.99	100.00
POL	1336	16.47	14.95	0.99	100.00
REL	2442	14.39	12.99	0.99	100.00
SEL	1805	13.05	12.99	0.99	32.95
SOC	1214	15.63	14.95	0.99	95.00
SPO	928	15.51	14.99	0.99	75.00
YAF	3055	11.26	9.99	0.99	54.99

B) Correlation of Numerical Coefficients







C) Coefficients for Model3

Coefficients:

(Intercept)	-1.619e+02
Division - Ballantine Group	-2.772e+00
Division - Bantam Dell	-2.877e+00
Division - Berkley / NAL	-2.818e+00
Division - Candlewick	-3.108e+00
Division - Charlesbridge	-3.617e-01
Division - Christian/Forum	-2.888e+00
Division - Crown	-2.761e+00
Division - Dark Horse Comics	2.412e-01
Division - DK.	-2.582e+00
Division - Hay House	-3.104e+00
Division - IDW Publishing	3.737e+00
Division - Knopf	-1.907e+00
Division - National Geographic Society	-4.769e-01
Division - Penguin Adult HC/TR	-1.485e+00
Division - Penguin Young Readers	-1.251e+00
Division - Potter/TenSpeed/Harmony/Rodale	-2.304e+00
Division - PRH Grupo Editorial	-2.249e+00
Division - Random House Group	-2.020e+00
Division - RH Childrens Books	-2.074e+00
Division - Shambhala	3.285e+00
Division - Titan	-1.699e+00
Division - Verso Books	3.816e-01
Division - Watkins Media	-2.762e+00
Type - Board	4.703e+00
Type - Hardcover	1.323e+01
Type - Hardcover Library Binding	1.371e+01
Type - Paperback	-7.586e-01
Type - Trade Paperback	5.695e+00
Pages	1.034e-02
Category - ART	6.699e+00
Category - BIO	1.391e+00
Category - BUS	2.680e+00
Category - CGN	6.892e-01
Category - CKB	4.217e+00
Category - CRA	5.579e+00
Category - FAM	1.205e+00
Category - HEA	2.648e+00
Category - HIS	2.010e+00
Category - HUM	8.787e-01
Category - JNF	-2.510e+00
Category - JUV	-2.624e+00
Category - OCC	1.939e+00
Category - OTH	2.567e+00
Category - POL	1.740e+00
Category - REL	1.981e+00
Category - SEL	1.299e+00
Category - SOC	2.207e+00
Category - SPO	2.942e+00
Category - YAF	-2.273e+00
Year	8.411e-02
Month	3.651e-02