FORDHAM | Gabelli School of Business
THE JESUIT UNIVERSITY OF NEW YORK

Creating A Recommendation System Based On User Reviews In The Amazon Kindle Store

Big Data Analytics
Prof. Dawit Demissie
ISGB-7990
Spring 2021

Group 10:
Prannoiy Chandran
Kristel Kalm
Abdullah Ahmad
Zev Rosenbaum
Yan Li
Jietong Guo

Executive Summary

The main objective of this project was to apply machine learning techniques to build a recommendation system based on user reviews in the Amazon Kindle store. The recommender was built using sentiment analysis and collaborative filtering and implemented with Python, Spark and Google Cloud Platform. These techniques provided descriptive and predictive insights into user behaviour. Data preprocessing was performed to remove missing values and replace alphanumeric ID values with numeric index values. Natural language processing techniques were used to derive polarity and subjectivity scores for the contents of each review to gauge whether each reviewer had enjoyed the ebook in question and to account for the subjectivity of user reviewers. The alternating least squares (ALS) algorithm was trained on the dataset and used to output recommendations for subsets of users and ebooks.

The recommender's performance was evaluated using root mean square error as well as qualitative analysis of the output. The recommender largely met the functional requirements and business goals of the project. Potential improvements were also discussed to make the system more appropriate for enterprise-scale deployment.

Business Goal

Long tail marketing can be defined as trying to target large numbers of niche marketing segments once a company has reached maturity. This can result in added sales for the company and a deeper relationship with its customers. Amazon is the largest bookseller in the world with an estimated 6 million books available for purchase on its website. The yearly revenue for Amazon in 2020 was $386 billion. With that level of supply and demand, long tail marketing is an integral part of Amazon's business model and recommendation systems are a key part of perfecting long tail marketing. By perfecting the recommendations systems, users will not only buy more products from Amazon, but will also likely increase their lifetime expectancy with the company, thus lowering churn rates for Amazon and increasing revenue. The business goals are to improve revenue and reduce costs, while offering personalized recommendations to customers.

Amazon has the advantage of access to one of the largest volumes of customer data in history. While conventional data processing and handling techniques may be limited in drawing actionable insights from such a large dataset, big data techniques can help to discover accurate patterns from historical data. These insights can be leveraged to make the customer experience on the Kindle store more personalized, and the techniques used can be applied to Amazon's other consumer-facing businesses as well. With the number of competitors entering the ebook reader market today, it is imperative for companies that already have large volumes of historical data to fully leverage that advantage. Existing advantages like brand loyalty may be increasingly less reliable for reducing churn and sustaining user growth as the market sees new entrants who can learn from the mistakes of their predecessors.

The functional requirements for the recommender are as follows: (1) identify users who are likely to rate a particular ebook highly, (2) identify ebooks that each user is most likely to rate highly based on the user's historical ratings.

Dataset Description

	asin	helpful	overall	reviewText	reviewTime	reviewerID	reviewerName	summary	unixReviewTime
0	B000F83SZQ	[0, 0]	5	I enjoy vintage b...	05 5, 2014	A1F6404F1VG29J	Avidreader	Nice vintage story	1399248000
1	B000F83SZQ	[2, 2]	4	This book is a re...	01 6, 2014	AN0N05A9LIJEQ	critters	Different...	1388966400
2	B000F83SZQ	[2, 2]	4	This was a fairly...	04 4, 2014	A795DMNCJILA6	dot	Oldie	1396569600
3	B000F83SZQ	[1, 1]	5	I'd never read an...	02 19, 2014	A1FV0SX13TWVXQ	"Elaine H. Turley...	I really liked it.	1392768000
4	B000F83SZQ	[0, 1]	4	If you like perio...	03 19, 2014	A3SPTOKDG7WBLN	Father Dowling Fan	Period Mystery	1395187200
5	B000F83SZQ	[0, 0]	4	A beautiful in-de...	05 26, 2014	A1RK2OCZDSGC6R	ubavka seirovska	Review	1401062400
6	B000F83SZQ	[0, 0]	4	I enjoyed this on...	06 10, 2014	A2HSAKHC3IBRE6	Wolfmist	Nice old fashione...	1402358400
7	B000F83SZQ	[1, 1]	4	Never heard of Am...	03 22, 2014	A3DE6XGZ2EPADS	WPY	Enjoyable reading...	1395446400
8	B000FA64PA	[0, 0]	5	Darth Maul workin...	10 11, 2013	A1UG4Q4D3OAH3A	dsa	Darth Maul	1381449600
9	B000FA64PA	[0, 0]	4	This is a short s...	02 13, 2011	AQZH7YTWQPOBE	Enjolras	Not bad, not exce...	1297555200
10	B000FA64PA	[0, 0]	5	I think I have th...	01 27, 2014	A1ZT7WV0ZUA0OJ	Mike	Audio and book	1390780800
11	B000FA64PA	[0, 0]	4	Title has nothing...	09 17, 2011	A2ZFR72PT054YS	monkeyluis	Darth Maul...the ...	1316217600
12	B000FA64PA	[0, 0]	3	Well written. Int...	12 31, 2013	A2QK1U70OJ74P	Sharon Deem	Not bad; it is we...	1388448000
13	B000FA64PK	[0, 0]	3	Troy Denning's no...	03 15, 2012	A3SZMGJMV0G16C	"Andrew Pruette "...	Han and Leia reun...	1331769600
14	B000FA64PK	[0, 0]	5	I am not for sure...	05 12, 2013	A3H8PE1UFK04JZ	Caleb Watts	Possibly Important	1368316800
15	B000FA64PK	[0, 0]	5	I really enjoyed ...	01 2, 2014	A2EN84QHDRZLP2	Carl craft	Another read	1388620800
16	B000FA64PK	[0, 0]	5	Great read enjoye...	10 29, 2013	A1UG4Q4D3OAH3A	dsa	Recovery	1383004800
17	B000FA64PK	[4, 4]	3	Another well writ...	04 16, 2009	A38Z3Q6DTDIH9J	"Jimmy J. Shaw ""...	Star Wars: The Ne...	1239840000
18	B000FA64PK	[0, 1]	5	This one promises...	01 27, 2014	A1ZT7WV0ZUA0OJ	Mike	my collection	1390780800

Figure 1. Dataset sample

The dataset, downloaded from the Kaggle website, consisted of 982,286 reviews collected across nearly 62,000 products on Amazon's Kindle store. Each row consisted of the ebook's ASIN (Amazon Standard Identification Number; assigned by Amazon for product identification), a helpfulness rating, the 'overall' rating assigned by the reviewer out of 5 stars, a timestamp at the time the review was posted, an alphanumeric ID for the reviewer, the reviewer's account username and the review's contents.

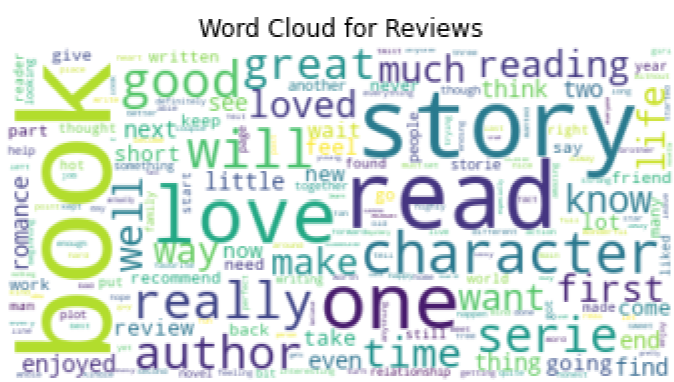

Figure 2. Word Cloud for Review

Figure 2 shows the most common words in the reviews. Words such as story, author, characters, good, love, series, and great are some of the most frequently used words in the reviews, which indicates that most of the reviews are positive reviews about the story, author, and characters of books.

```
+--------------+-----+
|    reviewerID|count|
+--------------+-----+
|A2WZJDFX12QXKD| 1002|
|A13QTZ8CIMHHG4|  929|
|A3PTWPKPXOG8Y5|  787|
|A37LY77Q2YPJVL|  741|
| A320TMDV6KCFU|  731|
|A2JZCZYHNQHSCP|  624|
| A2YJ8VP1SSHJ7|  579|
|A2VXSQHJWZAQGY|  556|
| ANOSVLTGRKABQ|  542|
|A328S9RN3U5M68|  510|
|A3A7FF87LEVCQ1|  488|
| A5JZNVV4TEAWU|  485|
|A1JLU5H1CCENWX|  474|
| AWAP0KEX6POQV|  462|
| AR9RKLQQC2L6K|  453|
|A3KH1OB5BYYQ8H|  448|
|A2APWUAUDWEGEC|  444|
| A3LXRIY0HE71K|  441|
|A14PRVP4JK88E7|  432|
|A3MT311233OHMQ|  409|
+--------------+-----+
only showing top 20 rows
```

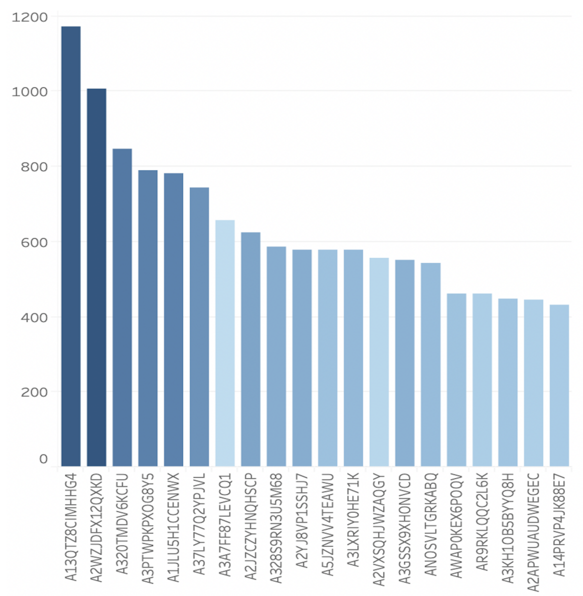

Figure 3. Most active reviewers

The top 5 reviewers in terms of number of reviews posted all had more than 700 reviews, with the top reviewer having posted more than a 1000 reviews. There is a steep dropoff in total review count after the top 5.

```
+----------+-----+
|      asin|count|
+----------+-----+
|B006GWO5WK| 1113|
|B00BTIDW4S|  781|
|B00BT0J8ZS|  516|
|B00JDYC5OI|  502|
|B00H0V069M|  481|
|B00BTIDXVU|  454|
|B00BSX4U04|  452|
|B00BTIDOO6|  428|
|B007R5YDYA|  422|
|B005DOK8NW|  408|
|B005C5YZ86|  394|
|B005ME39HU|  394|
|B00CCRTFSC|  390|
|B00HYQJPC2|  364|
|B00KF0URBM|  354|
|B004E10W0E|  341|
|B008IL48BK|  337|
|B00DJB6KE2|  334|
|B00EYMXM2I|  330|
|B00J1SP9FI|  324|
+----------+-----+
only showing top 20 rows
```

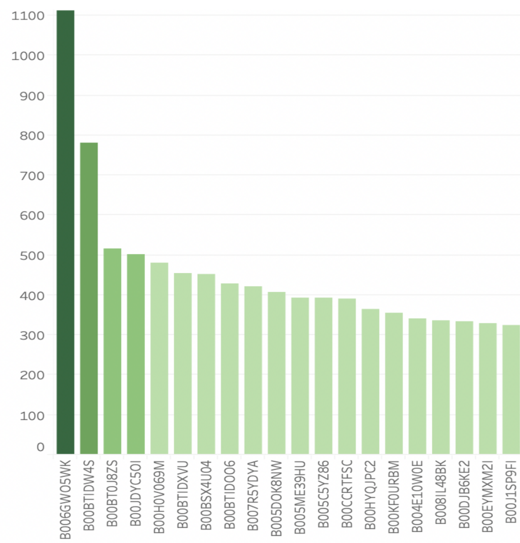

Figure 4. Most reviewed ebooks

Similarly, the ebook with the most reviews had over a 1000 reviews, and there was a sharp dropoff below in terms of number of reviews per ebook.

System Design and Methodology

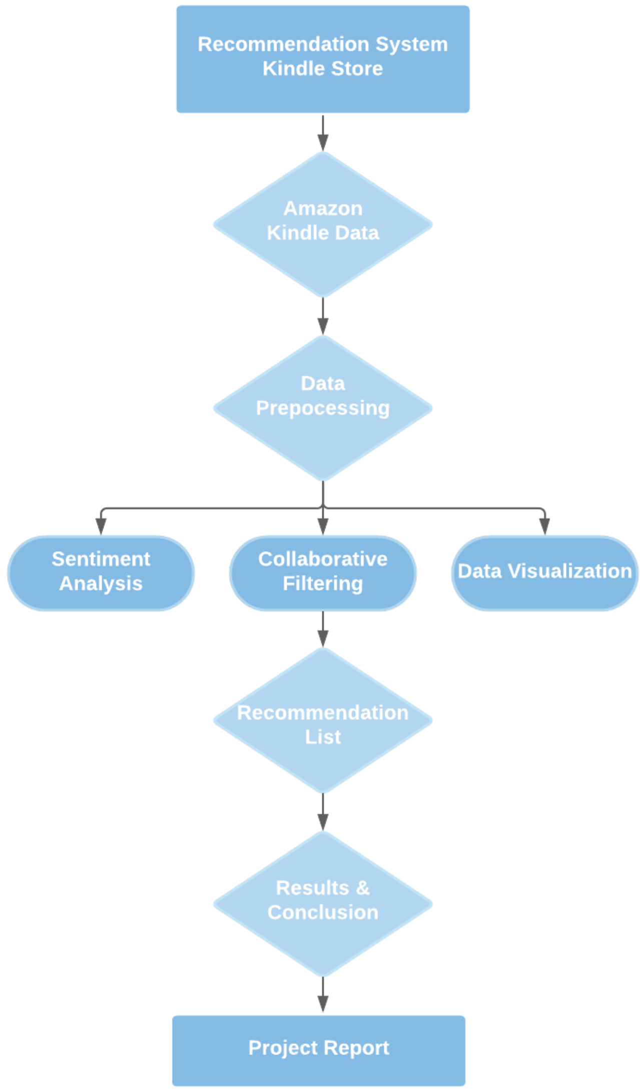

Figure 5. Recommendation project workflow

The dataset was loaded and pre-processed using Spark to convert it into a dataframe and to remove rows with missing values. Sentiment analysis was performed to examine the distribution of user ratings and to extract polarity and subjectivity scores. Collaborative filtering was then used to build the recommender and provide recommendations for subsets of ebooks and users.

Quantitative metrics and qualitative analysis of the output were performed to gain insights into user behaviour and evaluate the performance of the recommender. Potential improvements for future deployments were also discussed, as were the recommender's scope and ideal use cases.

System Implementation

a. Sentiment analysis

Unsupervised natural language processing techniques were used to pre-process and analyze the text in each review's contents. The NLTK and Punkt packages were used for tokenizing sentences and converting them into n-grams and bigrams for more efficient handling. The WordNet lemmatizer was also used to handle context by grouping together the inflected forms of each word. Group used NLTK and TextBlob to generate sentiment scores, and TextBlob was used to derive polarity and subjectivity scores generated by TextBlob for each review.

b. Collaborative filtering

The alternating least squares (ALS) algorithm was selected to build the recommendation engine. This was imported from Spark's machine learning library, MLlib. The system was implemented using Python and PySpark code run on a virtual machine on Google Cloud Platform.

ALS is grouped under matrix factorization, a class of collaborative filtering algorithms. Collaborative filtering computes the similarity between users' historical preferences and uses those as a basis for recommendations. This technique does not require information about the content of the product, relying instead on the assumption that users who had similar tastes in the past (peers) would have similar tastes in the future. Some of the most sophisticated enterprise-grade recommendation engines in use today (by companies including Netflix and Youtube) employ collaborative filtering techniques to study user behaviour.

Matrix factorization offers the advantages of high predictive accuracy, scalability and the ability to detect latent (hidden) features. ALS in particular is well-suited to use cases that call for parallelization, which is useful for large datasets such as this.

asin	overall	reviewerID	asin_index	reviewerID_index
B000F83SZQ	5	A1F6404F1VG29J	34516.0	19938.0
B000F83SZQ	4	AN0N05A9LIJEQ	34516.0	2064.0
B000F83SZQ	4	A795DMNCJILA6	34516.0	8432.0
B000F83SZQ	5	A1FV0SX13TWVXQ	34516.0	9200.0
B000F83SZQ	4	A3SPTOKDG7WBLN	34516.0	39078.0
B000F83SZQ	4	A1RK2OCZDSGC6R	34516.0	35445.0
B000F83SZQ	4	A2HSAKHC3IBRE6	34516.0	1135.0
B000F83SZQ	4	A3DE6XGZ2EPADS	34516.0	15436.0
B000FA64PA	5	A1UG4Q4D3OAH3A	52417.0	35607.0
B000FA64PA	4	AQZH7YTWQPOBE	52417.0	68819.0
B000FA64PA	5	A1ZT7WV0ZUA0OJ	52417.0	43564.0
B000FA64PA	4	A2ZFR72PT054YS	52417.0	37621.0
B000FA64PA	3	A2QK1U70OJ74P	52417.0	57417.0
B000FA64PK	3	A3SZMGJMV0G16C	34517.0	61649.0
B000FA64PK	5	A3H8PE1UFK04JZ	34517.0	47343.0
B000FA64PK	5	A2EN84QHDRZLP2	34517.0	44594.0
B000FA64PK	5	A1UG4Q4D3OAH3A	34517.0	35607.0
B000FA64PK	3	A38Z3Q6DTDIH9J	34517.0	59445.0
B000FA64PK	5	A1ZT7WV0ZUA0OJ	34517.0	43564.0

Figure 6. Transformed dataset with index values for alphanumeric IDs

As Spark's ALS algorithm only accepts integer inputs, the ebook ASINs and reviewer IDs were converted to index values. The transformed dataset was used as the input for the ALS model, with a training-testing split of 80-20. The implicit preferences parameter was set to "false" as explicit feedback (users' ratings) were available as inputs. The rank (number of latent features) was set at 100, and the regularization parameter (to reduce overfitting) was set at 0.15. While rows with missing values were already dropped during pre-processing, the cold start strategy parameter was set to "drop", instructing the model to not include rows with missing values in its computation.

Evaluation

a. Sentiment analysis

```
Review                                   : Polarity   : Subjectivity
I enjoy vintage books and movies so I en :    0.45    :   0.70
This book is a reissue of an old one; th :    0.28    :   0.38
This was a fairly interesting read.  It  :    0.14    :   0.59
I'd never read any of the Amy Brewster m :    0.20    :   0.20
If you like period pieces - clothing, li :    0.05    :   0.45
A beautiful in-depth character descripti :    0.24    :   0.70
I enjoyed this one tho I'm not sure why  :    0.22    :   0.64
Never heard of Amy Brewster. But I don't :    0.05    :   0.24
Darth Maul working under cloak of darkne :    0.53    :   0.45
This is a short story focused on Darth M :    0.19    :   0.41
I think I have this one in both book and :    0.70    :   0.60
Title has nothing to do with the story.  :    0.22    :   0.48
Well written. Interesting to see Sideous :    0.14    :   0.40
Troy Denning's novella Recovery was orig :    0.18    :   0.59
I am not for sure on how much of a diffe :    0.32    :   0.35
I really enjoyed the book. Had the norma :    0.21    :   0.45
Great read enjoyed every minute of it .  :    0.65    :   0.72
Another well written eBook by Troy Denni :    0.21    :   0.42
This one promises to be another good boo :    0.40    :   0.80
I have a version of "Star by Star" that  :   -0.00    :   0.32
Excellent! Very well written story, very :    0.20    :   0.61
```

Figure 7. Sample of polarity and subjectivity scores of reviews

Polarity is an expression of sentiment within a range of [-1, 1]. A polarity of -1 indicates a very negative sentiment, while 0 is neutral and a score of 1 is very positive. Within the sample sentiment scores generated by TextBlob, most reviews were either positive or close to neutral. However, sentiment alone is not enough for reliable insights; users' subjectivity must also be accounted for. The subjectivity score increases proportionally to the level of subjectivity in each review, with a range of [0.0, 1.0]. . Hence, a lower subjectivity score indicates a more objective review. A more subjective sentiment does not necessarily mean it is less likely to be a trusted data source, but we can set a threshold to filter out over-subjective reviews (negative polarity in most of time) and over objective sentiment (robots generated possibly).

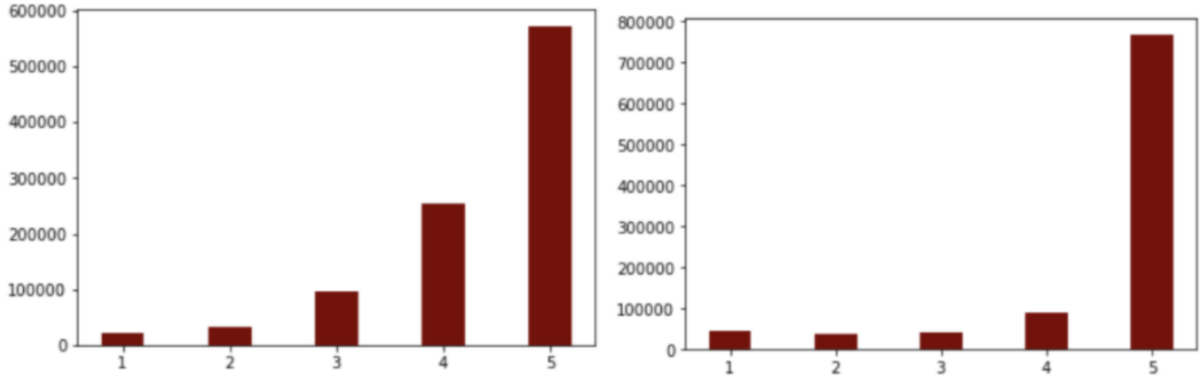

Figure 8. Distribution of stars of reviews

By dividing the sentiment scores generated by NLTK to 5 levels, it can be seen from the distribution bar charts that most sentiment scores were positive. Some negative values were wrongly predicted as positive ones. However, the overall predictive accuracy was sufficient to use the dataset as the input for machine learning algorithms.

b. Collaborative filtering

RMSE = 1.9171296719848283

asin	overall	reviewerID	reviewerID_index	asin_index	prediction
B007UAUPT4	4	A1LM14JMSUTXEZ	4356.0	148.0	2.8183692
B00IIYHLM0	5	AD6ROXTU7305R	5759.0	463.0	3.4258456
B00IIYHLM0	5	ADK827JF6CKJ0	5764.0	463.0	3.9775424
B00IIYHLM0	4	A3JHMDPDDFUSLV	2631.0	463.0	4.2370834
B00IIYHLM0	5	A396UOEM7XPB5	8518.0	463.0	2.5575194
B0095612HK	5	A2U8YWPP1PYHJM	53.0	496.0	4.188701
B0095612HK	5	A371U2QY2N2R66	15995.0	496.0	4.0281844
B00E0QS0D4	5	A3KG5TKIRPXXD7	16968.0	833.0	3.180298
B00E0QS0D4	5	A2HUVS5F8FQKBG	14143.0	833.0	2.4019878
B00E0QS0D4	4	A3JGY8WYUZBU6D	1964.0	833.0	3.120148
B006YTTVHO	4	A14XDOWJ9L4I4B	1302.0	1645.0	2.05153
B00FC09OJQ	5	A1Z9EFZT7D0U2T	12810.0	1959.0	3.9107964
B00KJ5Z1YG	5	AB21PLZLWGDE5	5724.0	2122.0	3.2286897
B00LLI4V82	5	A1F5ZAUCP4KDVG	193.0	2142.0	4.3018727
B00LLI4V82	5	A2OUFEQSKK5PDZ	3359.0	2142.0	4.163117
B00LLI4V82	5	A2Y18PKYE2B11Z	2519.0	2142.0	4.671639
B00E9QPFTM	5	A20VRO0FKD3ST2	7167.0	2659.0	3.7569532
B0065M6OCK	5	A1GQOV7SE71TB0	886.0	3175.0	3.154716
B00D0AGZQK	5	A1I5JKKMD84DRU	195.0	3749.0	3.1614285
B004EPYUOE	4	A2N3X7514J2COM	3346.0	4519.0	1.4145609

only showing top 20 rows

Figure 9. RMSE and sample ALS output

Within the sample ALS output, most predicted ratings appear to be close to the reviewer's actual rating (particularly when the predicted values are rounded up or down) but there are some rows where the predicted and actual values diverge greatly.

Root mean squared error (RMSE) was selected as the primary metric to evaluate the ALS model's performance. While the RMSE of 1.92 was judged to be sufficient for the purposes of this project, a lower value would be ideal for enterprise-scale deployment. Future deployments of the system would benefit from tuning hyperparameters to determine the optimal mix of parameters and possibly reduce the RMSE. In addition, including a bias factor (to distinguish among average, biased and critical users) in the model's parameters may improve its predictive accuracy.

```
+----------------+--------------------+
|reviewerID_index|     recommendations|
+----------------+--------------------+
|            7554|[{31535, 5.061816...|
|           16916|[{4943, 3.8377805...|
|           29811|[{37164, 4.335366...|
|            1051|[{9460, 4.8500547...|
|             596|[{8850, 4.9752464...|
|           28153|[{21544, 4.939356...|
|           39817|[{26625, 3.695491...|
|             305|[{11787, 5.019263...|
|           30867|[{11999, 3.915433...|
|           40999|[{11824, 2.954538...|
|           10561|[{15843, 4.850876...|
|            6433|[{16218, 4.894594...|
|            4142|[{6601, 2.8786082...|
|            8649|[{27225, 3.962237...|
|            3597|[{26760, 4.831100...|
|           10129|[{2148, 4.93216},...|
|           11924|[{2369, 4.772122}...|
|            7782|[{5682, 4.693063}...|
+----------------+--------------------+
```

Figure 10. Top 10 recommended ebooks per user

To test the system's recommendations, a subset of 20 users was selected to generate recommendations for. For each user in the subset, the 10 ebooks they were most likely to rate highly were identified. These recommendations also included their predicted ratings.

	reviewerID_index	recommendations	recommended_ebooks
0	7554	[(31535, 5.061816692352295), (549, 5.009945392...	31535,549,25232,8196,18559,1743,10888,5629,417...
1	16916	[(4943, 3.837780475616455), (7013, 3.618249893...	4943,7013,10976,1208,6062,5019,3512,2210,6253,...
2	29811	[(37164, 4.3353657722473145), (23177, 4.289042...	37164,23177,17410,15442,2685,34027,9597,19622,...
3	1051	[(9460, 4.850054740905762), (1040, 4.827525138...	9460,1040,36069,2752,11142,3217,10836,5629,111...
4	596	[(8850, 4.975246429443359), (5971, 4.963124752...	8850,5971,10632,16098,7009,3077,19980,6301,935...
5	28153	[(21544, 4.939355850219727), (2842, 3.06026959...	21544,2842,17311,18944,11856,4237,2609,4756,35...
6	39817	[(26625, 3.695491075515747), (14071, 3.4853115...	26625,14071,5421,7057,3490,3945,2428,15217,757...
7	305	[(11787, 5.019262790679932), (22347, 4.9408221...	11787,22347,11514,21607,32388,6903,3217,6978,4...
8	30867	[(11999, 3.915433406829834), (5920, 3.65094876...	11999,5920,9738,19328,19111,11388,9978,2884,13...
9	40999	[(11824, 2.954538345336914), (22131, 2.4487910...	11824,22131,6032,4720,20086,12312,13336,19865,...
10	10561	[(15843, 4.850876331329346), (350, 4.807676792...	15843,350,12009,7272,5530,10870,6501,6597,1300...
11	6433	[(16218, 4.894594669342041), (28493, 3.9156761...	16218,28493,27287,39126,10389,36836,5717,20580...
12	4142	[(6601, 2.878608226776123), (23225, 2.84687399...	6601,23225,2099,11048,6711,17796,1394,11360,73...
13	8649	[(27225, 3.96223783493042), (4863, 3.958763122...	27225,4863,4188,10667,4618,10648,10596,3217,20...
14	3597	[(26760, 4.8311004638671875), (25724, 4.022906...	26760,25724,1056,1692,6889,5943,8889,8692,2814...
15	10129	[(2148, 4.932159900665283), (9461, 4.059145927...	2148,9461,21147,3582,8901,2458,11409,15751,125...
16	11924	[(2369, 4.772121906280518), (19094, 4.70031309...	2369,19094,7470,4230,8075,16701,3990,2057,7847...
17	7782	[(5682, 4.693062782287598), (4356, 4.651572227...	5682,4356,2870,10687,2289,9143,4814,10857,5292...

Figure 11. Top 10 recommended ebooks per user (with and without predicted ratings)

The Spark dataframe was converted to a pandas dataframe which was then run through a function to isolate the ASIN index values of each recommended ebook (while omitting the predicted ratings). The recommendation system's output can therefore be handled as a collection of recommendations and the predicted ratings ("recommendations" column) or as a collection of ebook ASIN's alone ("recommended_ebooks" column). The second option would be the likely output used to display a list of recommendations on a user interface.

	asin_index	recommendations	users_recommended_to
0	11757	[(39110, 2.906642198562622), (9093, 2.51824808...	39110,9093,9680,36169,41623,40854,22152,19699,...
1	7171	[(27766, 4.169920444488525), (23524, 4.1699204...	27766,23524,19915,7127,39082,10,20776,35806,40...
2	11766	[(9093, 5.500514984130859), (9687, 5.385972976...	9093,9687,38834,15808,22152,41623,40854,39082,...
3	22984	[(9680, 4.315417289733887), (38615, 4.14181375...	9680,38615,39082,21973,8662,7004,31887,104,233...
4	21606	[(27672, 3.9276347160339355), (12634, 3.372845...	27672,12634,13315,39117,21888,20639,36176,9563...
5	22195	[(32424, 4.936210632324219), (14559, 4.1451697...	32424,14559,11697,13123,30761,12943,7827,5678,...
6	12172	[(7474, 4.859525680541992), (9746, 4.616447448...	7474,9746,18080,9687,31437,15808,7701,36169,11...
7	21911	[(10662, 1.237143635749817), (18926, 1.2156422...	10662,18926,22969,2835,8115,19968,38037,26303,...
8	22797	[(8199, 4.968967914581299), (8100, 4.927923202...	8199,8100,12081,4550,39082,1271,37355,9746,968...
9	11935	[(19137, 4.4604949951171875), (16156, 4.246379...	19137,16156,3414,2857,4273,1420,36169,38615,52...
10	11967	[(9869, 5.423186779022217), (9687, 5.253345489...	9869,9687,15808,1271,38037,1487,38615,11376,16...
11	22331	[(40899, 3.9249448776245117), (36169, 3.569956...	40899,36169,35021,17285,38615,25111,16559,7985...
12	21791	[(9687, 4.337490081787109), (41179, 4.31685876...	9687,41179,38740,38834,1271,4407,10717,15808,1...
13	22274	[(3550, 4.893342971801758), (9746, 4.699467658...	3550,9746,36169,39304,13744,1271,8199,9869,160...
14	7115	[(21415, 3.8939127922058105), (29373, 2.920434...	21415,29373,38615,18365,30688,36169,9093,9687,...
15	21933	[(5391, 2.0466504096984863), (2135, 2.01352071...	5391,2135,296,9746,35845,28219,7070,16115,1137...
16	12275	[(31424, 4.942534923553467), (8674, 4.94214725...	31424,8674,7004,1487,29061,13315,11376,16115,2...
17	11924	[(2, 5.025899887084961), (21973, 5.00674962997...	2,21973,17513,9746,6797,29020,38615,9687,16790...
18	21825	[(19498, 4.9250640869140625), (1235, 4.5386557...	19498,1235,9869,11538,26428,17764,39327,28497,...

Figure 12 . Top 10 recommended users per ebook (with and without predicted ratings)

Similarly, a subset of ebooks was used to identify the top 10 users who were predicted to rate each ebook highly. The output can be handled as a collection of recommended users and their predicted ratings (“recommendations” column) or as a collection of user ID indices alone (“users_recommended_to”) column.

The recommender could potentially have a direct impact on revenue and profit if it helps to influence user behaviour. Users who find the recommendations relevant will be likely to purchase additional ebooks that they might not have otherwise. In addition, satisfied customers may recommend Kindle ebooks and devices to others. Due to the strong Kindle ecosystem, every new customer translates to a significant amount of money spent on a Kindle device as well as several ebooks. These changes, replicated over multiple users who end up purchasing multiple ebooks (while also recommending Kindle devices to an estimated 2-3 friends each), may have a multiplier effect on metrics such as monthly recurring revenue (MRR), customer lifetime value (CTV) and customer acquisition cost (CAC).

Regarding the scope of the system, it was built on a dataset with explicit ratings, and the performance level may not necessarily translate to a dataset with implicit ratings. The system's performance may also be different if the initial dataset is sparse and results in a "cold start" situation. Alternative techniques may be required if the model will not be able to learn from a large volume of user ratings and historical preferences. The current system's ideal use case is when a company has access to a large volume of historical data and the majority of users have rated multiple items, making their preferences very clear. This certainly applies to Amazon, but other companies may not have the same volume of customer data or the capabilities to deploy machine learning and big data techniques on large datasets.

For future implementations, the recommender's output could be improved by converting the index values back to their original alphanumeric values. However, the necessity of this step depends on how the output data will be handled and the requirements of the backend and frontend systems the data will be fed to. In addition, training the model on a different dataset (perhaps one with a larger volume of data or with fewer missing values) may result in a smaller (more desirable) RMSE and more relevant recommendations. In such cases, the costs involved in collecting and handling the required data will have to be weighed against the potential benefits from implementing the system.

Conclusion

Big data techniques can help to uncover significant patterns in large datasets, making them useful for consumer-facing businesses like Amazon that have access to large volumes of customer data to leverage. In this case, the alternating least squares algorithm was used and implemented using PySpark and Google Cloud Platform. ALS was chosen because it is a collaborative filtering method that performs robustly on large datasets with historical data and explicit ratings. Additionally, the matrix factorization aspect allowed the model to find hidden features.

Overall, the recommendation system built was judged to have high scalability and predictive accuracy. It satisfied the functional requirements and business goals outlined in the beginning, and the approach can be translated to other companies seeking to draw insights from historical customer data. The potential costs involved in implementing the proposed big data solution are highly likely to be recouped by the gains in revenue, customer satisfaction and brand loyalty.

References

Data source: https://www.kaggle.com/bharadwaj6/kindle-reviews

https://www.statista.com/statistics/266282/annual-net-revenue-of-amazoncom/#:~:text=The%20time%20series%20shows%20the,billion%20US%20dollars%20in%202019.