"Sales Forecasting of Books & Customer Profiling Based on Sentiment Analysis of Customer Reviews"

Pratap Roy Choudhury, Raj Bhowmik

MS in Data Science, Spring 2022

Indiana University Bloomington

Abstract: *Sentiment analysis is one of the most important tasks of Natural Language Processing. It deals with the text classification to determine the intention of the end user of the text. The paper presents different preprocessing methods like HTML tags and URLs removal, punctuation, whitespace, special character removal to remove noise and use Text Blob and VADER techniques for the sentiment analysis of book reviews to find the top-rated books based on the most positive review scores. The data used for this research was weekly for a time of 12 months, on which time series forecasting was performed to predict future sales of the top-rated books across all states in the United States of America. By leveraging the Decision Tree and Clustering technique we binned people of various age groups together to classify which group of people purchases which specific top-rated books. In this paper, we examine various levels of Sentiment Analysis to find the top selling books, along with a Time Series forecasting technique to predict the unit of sales in upcoming months.*

Keywords: Sentiment Analysis, Machine Learning, TextBlob, VADER, Forecast, SARIMA, Pmdarima, Customer Profiling

1. INTRODUCTION

Customer reviews are an integral part for any product that is being used. They are readily available on the internet, which helps in choosing the best of products. This paper talks about the reviews of books and how it helps the audience to select the best of books based on ratings. These top-rated books are intended to find using positive sentiment scores or the customer reviews [1][2]. Additionally, it also helps the retailers to gather information regarding their top selling books and forecast future sales of their books and thus improving their growth strategy [3]. Based on that, a customer profiling is built, targeting the potential customers thus tailoring their strategy to meet the customers' requirements [4].

2. AGENDA

This research analyses will mainly focus on the following findings to help the book retailer plan more precisely their growth strategy and increase the book sale.

- Analysis of the most positively reviewed books and ranking of the top 4 books.
- Sales forecasting for the top shelf books for the upcoming 2 months across all the 21 states of US.
- Customer profiling based on the most positive reviews.

3. DATA COLLECTION

Based on the research and agenda, there are three different datasets required which are collected from online available repositories. Couple of the datasets are created manually by taking reference from the actual industrial dataset.

3.1. Customer review dataset

The customer reviews are the textual data which are analysed to perform sentiment analysis. Most positive reviews for the books are found by sorting the positive sentiment scores at a book code level and the top rated books are identified. This 'Reviews.xlsx' data has been collected from online source such as

Kaggle, GitHub and other sources of Amazon review data for book product category. There are about 112k reviews available for 20 unique book codes.

Below are the details of each of the fields of the customer review dataset:

Field	Description
Comment	Book reviews posted by the customers
BookCode	Unique book ID

Table 1: ***Data dictionary of the customer reviews collected from Kaggle***

Comment	BookCode
This is a self-published book, and if you want to know why--read a few paragraphs! Those 5 star reviews must have been written by Ms. Haddon's far	52979
I was a dissapointed to see errors on the back cover, but since I paid for the book I read it anyway. I have to say I love it. I couldn't put it down. I reac	22722
A complete waste of time. Typographical errors, poor grammar, and a totally pathetic plot add up to absolutely nothing. I'm embarrassed for this a	52720
I feel I have to write to keep others from wasting their money. This book seems to have been written by a 7th grader with poor grammatical skills fo	32722
Excellent stockings for long shifts on your feet - not too tight, not too loose...garment integrity is longer than package states (with proper care).	84987
It took almost 3 weeks to recieve the two pairs of stockings I ordered and when they arrived they were not in a box nor did they have any tags on th	94987
sizes are much smaller than what is recomended in the chart. I tried to put it and sheer it!. I guess you should not buy this item in the internet..it is	52720
Steven Wardell's book is a pure delight and I've recommended itto friends of all ages. In Rising Sons and Daughters, we learn thatJapanese young pe	22979
You said "...but the charge only lasts a very short time."Did you know that new Ni-MH batteries must be cycled 3 to 5 times before they hold a full c	22979
Many useful concepts of digital compression can be found in this book. It is easy to read and understand, especially for students and engineers in EE	22720
After you watch a few episodes, it will be fairly clear to you who the two central characters are. If you enjoy a fitting but bleak ending, stop at Seaso	22979

Table 2: ***Data sample of the customer reviews collected from Kaggle***

3.2. Sales dataset

The sales dataset is a collection of book sales units, available stocks, price, state, weekly dates, holiday flags. This dataset is used to analyze time series attributes and forecast the book sales for the next 2 months. The 'Sales.xlsx' data is created manually by illustrating few features and factors referenced from sales dataset available on GitHub for various product categories [7]. For train set, it has 50k datapoints across 21 states and 20 unique book codes for 52 weeks. The test set has 3780 datapoints for all the book codes to forecast for next 8 weeks.

Details of each of the fields of the sales dataset:

Field	Description
Unit Sold	Number of books sold
Available Books in the shelf	Book stock available on shelf
Price	Cost of one book unit
State	State name from where book is sold
Week	Week of the book slae
New Year,MLK,President,GoodFriday,Memorial	Holiday variables
BookCode	Unique book ID

Table 3: ***Data dictionary of the sales dataset***

Units Sold	Available Books in the shelf	Price	State	Week	NewYear	MLK	Presidents	GoodFriday	Memorial	Independence	Labor	Thanksgiving	Christmas	BookCode
6187	10146.68	14.4745434	Arizona	2019-01-06 00:00:00	1	0	0	0	0	0	0	0	0	22722
9123	15144.18	14.174504	Arizona	2018-01-07 00:00:00	1	0	0	0	0	0	0	0	0	22979
6236	8356.24	14.1500962	Arizona	2019-01-13 00:00:00	0	0	0	0	0	0	0	0	0	84987
8080	11312	14.0782178	Arizona	2018-01-14 00:00:00	0	0	0	0	0	0	0	0	0	27720
5695	7289.6	14.2382792	Arizona	2019-01-20 00:00:00	0	0	0	0	0	0	0	0	0	37722
7429	9509.12	13.7892045	Arizona	2018-01-21 00:00:00	0	1	0	0	0	0	0	0	0	94987
5633	8449.5	13.7791585	Arizona	2019-01-27 00:00:00	0	1	0	0	0	0	0	0	0	32770
6455	10650.75	14.2938807	Arizona	2018-01-28 00:00:00	0	0	0	0	0	0	0	0	0	22720
6068	9951.52	13.6821028	Arizona	2019-02-03 00:00:00	0	0	0	0	0	0	0	0	0	42727
6644	11294.8	14.2655027	Arizona	2018-02-04 00:00:00	0	0	0	0	0	0	0	0	0	49279
6270	9969.3	14.0987241	Arizona	2019-02-10 00:00:00	0	0	0	0	0	0	0	0	0	10987
7163	8810.49	14.0065615	Arizona	2018-02-11 00:00:00	0	0	0	0	0	0	0	0	0	22729

Table 4: ***Data sample of the sales data created manually***

3.3. Customer dataset

The customer dataset is a collection of customer ID and their age mapped to each book codes which they have either bought or reviewed. This dataset is created by randomly generated several customer ID which are mapped with the book code and ages are created randomly for each individual customer ID.

Below are the details of each of the fields of the customer dataset:

Field	Description
CustomerID	Unique customer ID
BookCode	Unique book ID
Age	Age of the customers

Table 5: ***Data dictionary of the customer dataset***

CustomerID	BookCode	Age
17850	32722	19
17850	32979	19
16098	22926	50
18074	22457	47
17420	22663	46
13767	22727	53
16218	22469	50

Table 6: ***Data sample of the customer data***

4. DATA ANALYSIS

As the entire process requires three different segments of analyses, the plan is to perform sentiment analysis on the customer review dataset to identify the most positively reviewed books and find the top-rated books, apply time series forecasting on the sales data and perform customer profiling using clustering and binning techniques on customer dataset.

4.1. Sentiment Analysis - to find the top-rated books

For sentiment analysis of the customer reviews, the texts are needed to be cleaned to remove the unwanted characters and form a meaningful document. The complete analyses steps are explained below.

4.1.1. Text Preprocessing

A quick manual observation of the review dataset showed that there are 40% of the reviews unrelated to the books. A simple filter is applied to find the relevant book reviews and filter out the rest by tagging reviews based on whether they contain words commonly found in book reviews. Some of the filters viz. book, words, author, publish, read, wrote, write, thesaurus, dictionary, series, plot, tale, story, written, novel, page, classic, bestseller, guide, test, edition, magazine, issue etc. are applied which narrow down the review data to 72158 unique reviews.

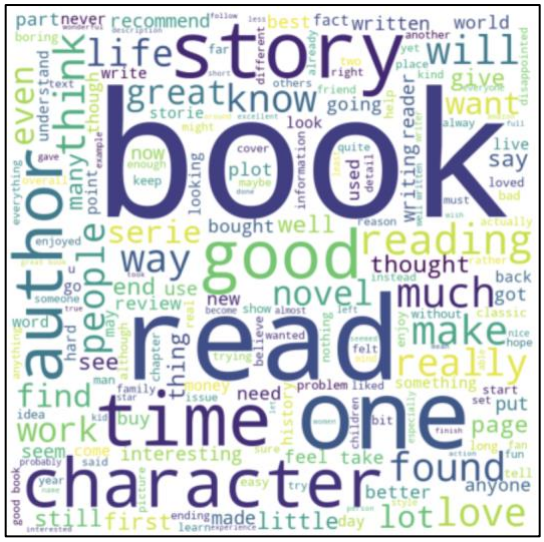

Figure 1: ***Word cloud relevant to the filtered book reviews***

In Natural Language Processing, the textual data is needed to be cleaned and processed before modeling. With the help of some manual observation, it is found that the tweets contain various characters, symbols etc. which should be removed. The prerequisite steps are as follows -

I. The texts are transformed to lower-case and then any username (starting with @), # symbols, numbers, URLs, characters that are non-alphanumeric are removed from the text using regular expression.

II. Using the stopwords package of NLTK, stopwords like and, or etc., are removed. Words like not, did not, doesn't etc. are not removed as these words can impact the sentiment of the text.

III. Tokens in the reviews are lemmatized using the WordNetLemmatizer function.

4.1.2. Sentiment score using TextBlob

The TextBlob sentiment analyzer is an inbuilt package of NLTK that generates polarity and subjectivity scores of a text. This method is applied on the cleaned reviews which gives +1 and -1 scores based on the positive and negative sentiment. Below are the steps that are followed.

- Using the TextBlob package, the sentiment scores for all the books are generated, and converted the scores to 0 and 1, if the reviews were negative and positive respectively.
- Since the concentration is on the most positive reviews for the books, the scores of each review are summed up at a book code level and that gives a direct reflection of the scores of the positive reviews of each unique book.

	BookCode	scores
2	22979	5573
1	22722	5563
15	84987	5515
0	22720	5479
4	32722	1404
3	32720	1400
11	52979	1389
18	114987	1365
7	42722	1362

Figure 2: ***TextBlob positive sentiment score at a book code level***

- The dataset has different count of reviews for each book. To remove the bias in the sentiment scores, these scores are normalized, by dividing the scores of the books by the number of reviews each book had in the data.

	BookCode	scores	Count	normalised_score
2	22979	5573	7788	0.715588
0	22720	5479	7671	0.714248
1	22722	5563	7812	0.712110
15	84987	5515	7806	0.706508
7	42722	1362	2514	0.541766
5	32979	1360	2540	0.535433
6	42720	1343	2524	0.532092
4	32722	1404	2647	0.530412
17	104987	1316	2495	0.527455
18	114987	1365	2589	0.527231

Figure 3: ***Normalized TextBlob positive sentiment score at a book code level***

- Normalized scores are sorted and the top 4 books with the highest positive review scores are selected

4.1.3. Alternative scoring analysis using VADER sentiment scores

To remove any sort of biasness within the process to generate the most positive sentiment scores, VADER is also used as an alternative process to cross verify the result generated using TextBlob. VADER analysis model from python NLTK library is utilized [6]. This implementation technique is chosen because it is especially attuned to the sentiment expressed in social media and sensitive to both polarity (positive/negative) and intensity (strength) of emotion.

Using VADER's positive, negative, Neutral and Compound scores, and TextBlob polarity and subjectivity scores, various statistical aggregators are calculated and formed as a dataframe.

BookCode	ReviewCount	TBPolarity-Mean	TBPolarity-Median	VdComposite-Mean	VdComposite-Median	VdPositive-Mean	VdPositive-Median
22979	7788	0.262690944	0.2489375	0.675650603	0.8591	0.197506549	0.186
22720	7671	0.25962104	0.25	0.67682443	0.8588	0.197533698	0.185
84987	7806	0.25957707	0.246145105	0.662397758	0.8591	0.196394953	0.185
22722	7812	0.258864454	0.246063763	0.667374834	0.85545	0.196406426	0.183
32979	2540	0.094894668	0.1	0.244783976	0.4454	0.127396457	0.119
42720	2524	0.093856367	0.099166667	0.263988312	0.4678	0.126944532	0.119
94987	2591	0.091883814	0.1	0.257836164	0.4574	0.12708298	0.119

Table 7: ***Statistical aggregators of polarity and subjectivity scores***

4.2. Time Series Forecasting - demand forecast of top 4 books for 2 months

The sales dataset that is available is a 52-week book sales and availability of the books data which is analyzed for time series attributes. Firstly, the data is filtered for the top 4 positively rated books. Below are the time series analyses and forecasting steps explained in detail.

- All the holiday variables are combined into one feature called 'isHoliday' by summing up the holiday flags at every book code, state and week level.
- Since there are many combinations available for the 21 states for 4 book codes, an initial analyses is done for one of the state and one of the book code to check the time series trends. It is done for 'Arizona' state and '22722' book code. Periodic Dickey-Fuller test is performed (to check for stationarity) after normal and seasonal differencing the unit of book sold over weeks.

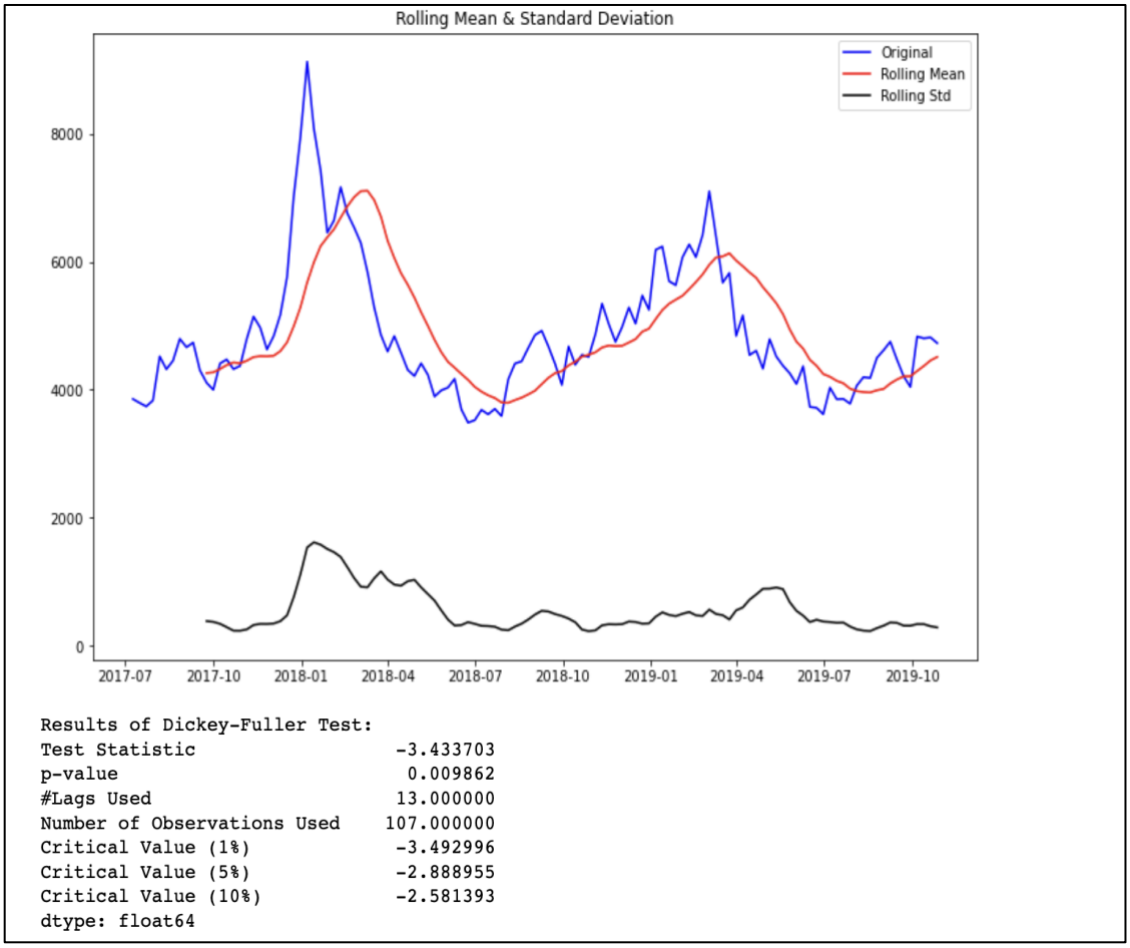

Figure 4: ***'Test Statistic' and the 'Critical Value' to check for the stationarity of the time series using Dickey-Fuller test***

- As it can be seen that the hypothesis is significant as the test statistic is lower than the critical values at 5% or 105 level.
- Next is tried to identify the time series components such as trend, seasonality, and residuals by observing the seasonal decompose plots.

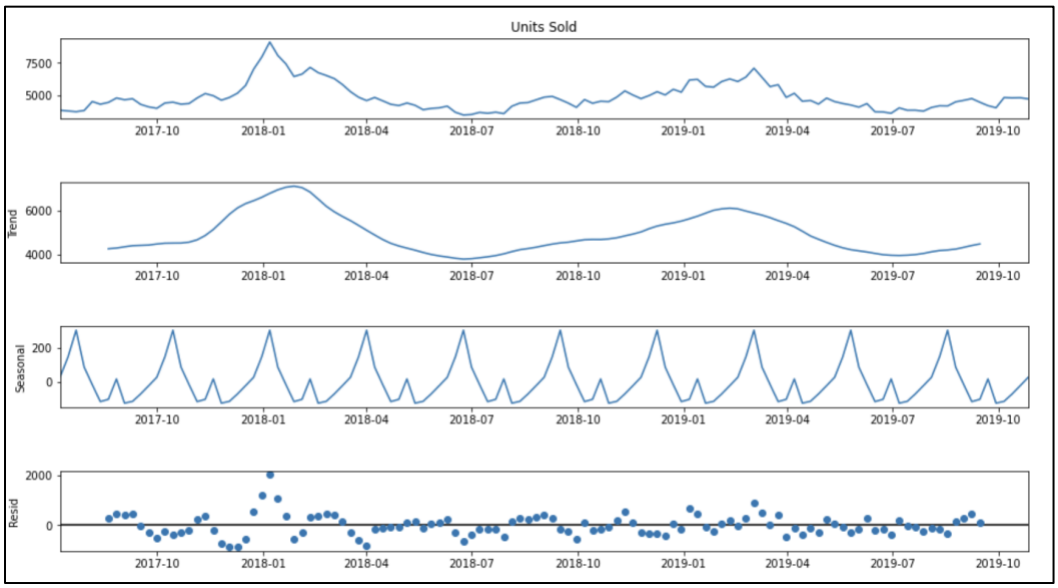

Figure 5: ***Seasonality, Trend and residual plot***

- ACF and PACF plots created to derive the orders (AR, MA coefficients) and the seasonality order if present.

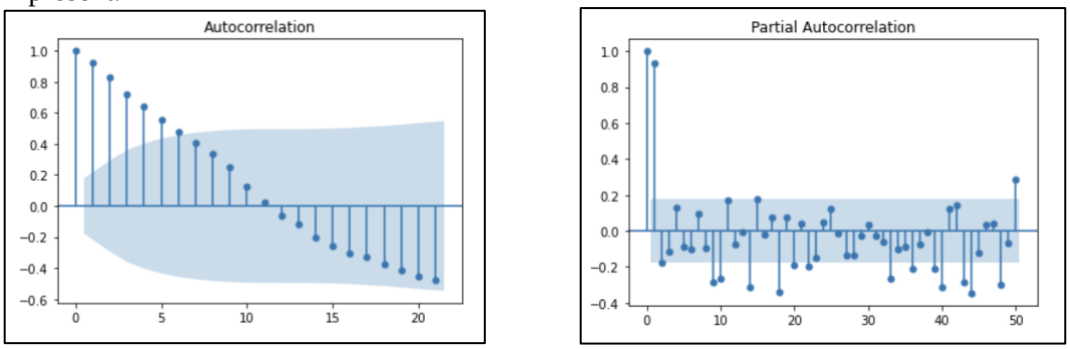

Figure 6: ***ACF and PACF plots***

- Pmdarima (pyramid-arima) is a statistical library in Python for time series analysis, which is equivalent of R's auto.arima functionality - automatically detects the order (AR,MA coefficients) and the seasonality order if present.
- Now to generate the forecasting model, iterating for individual book codes and states and automating the process using 'pmdarima' with all possible combinations of exogenous variables ('Available Books in the shelf', 'Price' and 'isHoliday') and without any exogenous variable. An exogenous variable is a variable that is not affected by other variables in the system.
- Running for the train test split to check the performance of our model by validating with AIC and BIC scores.
- In this model, the combination of variables are chosen based on the AIC and BIC scores. A good model has lower AIC and BIC score compared to other models.
- The lowest AIC/BIC score was observed for the one with exogenous variable combination 'Available Books in the shelf' and 'Price' for most of the Book Code and State combination.

```
Running for State:Florida and Bookcode:84987
Without exogenous
AIC:2103.4640111616254,BIC:2111.125641836427
With exogenous columns ('Available Books in the shelf',)
AIC:2105.447033500176,BIC:2115.6625410665783
With exogenous columns ('Price',)
AIC:2091.7728758029475,BIC:2101.98838336935
With exogenous columns ('isHoliday',)
AIC:2110.063693271468,BIC:2120.27920083787
With exogenous columns ('Available Books in the shelf', 'Price')
AIC:2090.210613638494,BIC:2100.4261212048964
With exogenous columns ('Available Books in the shelf', 'isHoliday')
AIC:2111.141190751637,BIC:2121.3566983180394
With exogenous columns ('Price', 'isHoliday')
AIC:2092.600806414637,BIC:2107.9240677642406
With exogenous columns ('Available Books in the shelf', 'Price', 'isHoliday')
AIC:2093.068951562067,BIC:2105.8383360200696
```

Figure 7: ***Model AIC and BIC scores for the exogenous variable combinations***

- The finalized model is trained on the train dataset using ‘pmdarima’ including 'Available Books in the shelf', 'Price’ as exogenous variables.
- Once the model is trained, it’s applied on the test dataset which is basically the top 4 book codes across 21 states for the next 8 weeks ie. 2 months. The unit book sold demand is forecasted for the books for the next 2 months.

4.3. Customer Profiling

A brand that understands its audience and is profiling its customers can tailor its offerings to suit those customers. This includes the ability to personalise communications and marketing across all channels to make them more relevant to the people receiving them. Knowing your customers allows a brand to recognize customers and make them feel valued.

Here the target is to identify the customer groups who are the most potential buyers and reviewers across any states and of which age groups. Also, how the holiday season effects the top books to sell in any state and bought by any customer group. Below are the initial steps performed on the Customer dataset.

- The initial EDA shows that the mean and median of the age grouping by the top book codes are almost similar and within the range of 46-47 and 47-48 respectively.
- A histogram plot of the age groups shows that the most of the customers are between 30 to 60 years of age who are the primary customer.

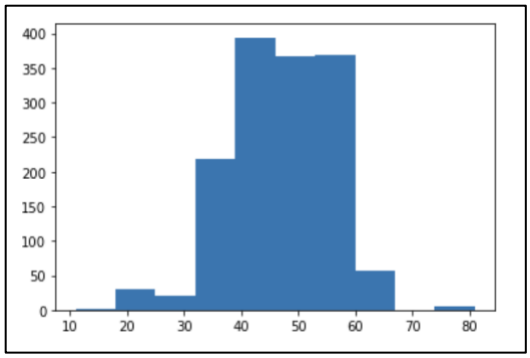

Figure 8: ***Histogram plot of the customer age groups distribution***

- Built a balanced Decision Tree Classifier with age variable as X and the top 4 BookCodes as Y. Expected output was the binned ages in the tree with BookCodes in the leaf to provide the age

range for each of the Book Code. But that did not make sense because of no clear distinction between the age range for each Book Code and the approach is complicated for a single variable.

4.3.1. State wise profiling

Using the Sales data, the top 4 book codes are filtered, and total number of books sold are calculated at a book code and state level. Percentage distribution of 'Units Sold' by State for each book are calculated to determine which were the top 2 states from where customers would buy each book.

State	22720	22722	22979	84987	22720_dist%	22722_dist%	22979_dist%	84987_dist%
Arizona	39530353.0	589011.0	16128367.0	4921576.0	3.890475	3.059885	4.131745	4.557353
California	137358824.0	2345784.0	73349815.0	13364194.0	13.518501	12.186238	18.790666	12.375173
Colorado	33895335.0	431207.0	13176354.0	3968926.0	3.335891	2.240100	3.375502	3.675205
Florida	147989775.0	3212136.0	49399044.0	13925137.0	14.564772	16.686896	12.654987	12.894603
Illinois	79292490.0	1430364.0	28813257.0	8810239.0	7.803762	7.430674	7.381345	8.158235
Indiana	21857432.0	356291.0	5759931.0	2825560.0	2.151152	1.850915	1.475572	2.616454
Kansas	18128590.0	293016.0	5361513.0	2323227.0	1.784169	1.522204	1.373506	2.151296
Maryland	61617864.0	1230131.0	24397233.0	7119549.0	6.064271	6.390473	6.250054	6.592665
Minnesota	24543015.0	241213.0	9434767.0	3042123.0	2.415460	1.253090	2.416987	2.816990
Missouri	24526163.0	435872.0	6841594.0	3190521.0	2.413802	2.264335	1.752671	2.954406
Nevada	15015297.0	284243.0	6906519.0	1753553.0	1.477767	1.476629	1.769304	1.623781
New Mexico	39822262.0	958182.0	13368443.0	3878667.0	3.919204	4.977711	3.424711	3.591625
New York	126166572.0	2879161.0	56034071.0	12198648.0	12.416988	14.957106	14.354740	11.295884
Ohio	19747715.0	290350.0	5271889.0	2718993.0	1.943519	1.508355	1.350546	2.517773
Oklahoma	23638654.0	450783.0	6193967.0	2472489.0	2.326455	2.341797	1.586763	2.289512
Oregon	19973057.0	196049.0	7697551.0	2077801.0	1.965697	1.018465	1.971949	1.924033
Tennessee	24159506.0	556126.0	5750650.0	3250432.0	2.377716	2.889048	1.473195	3.009883
Texas	118285360.0	2504990.0	42346013.0	11291527.0	11.641340	13.013306	10.848150	10.455894
Utah	15532788.0	210424.0	5677414.0	1961335.0	1.528697	1.093143	1.454433	1.816186
Washington	4833889.0	52321.0	1592912.0	530881.0	0.475739	0.271805	0.408070	0.491593
Wisconsin	20165350.0	301798.0	6851082.0	2366599.0	1.984622	1.567826	1.755102	2.191458

Figure 9: ***Units Sold distribution of the top 4 books across all states***

4.3.2. Holiday wise profiling

GiftingHoliday' flag is created to signify holidays when a book is sold, especially during 'ThanksGiving', 'NewYear', 'Christmas'. Similar to the state wise profiling, the total number of book sold is calculated at GiftingHoliday flag and Book code level. A pivot table of average number of book units sold to shows whether any of the books were popular gift choices.

GiftingHolidays	22720	22722	22979	84987
False	400727.683230	7515.386335	153917.916356	42711.851346
True	383515.365079	8728.523810	147941.412698	38435.365079

Figure 10: ***Units Sold distribution on holiday vs not a holiday***

4.3.3. Customer Age profiling

From the customer data, the customers are binned into seven logical division : Children (1-12 yrs), Teenagers (13-17 yrs), Adults (18-29, 30-39, 40-49 yrs) and Seniors (50-59, 60+ yrs). The age bin and grouping procedures are explained below.

- Customer ID distribution grouping by Age bin and Book Code.
- Filter out groups of Book code and Age bin having most of the customers than the average of customers in each groups.

- Finding the percentage distribution of customers across each Book code and Age bin by total customers in each Age groups.

	BookCode	AgeBin	CustomerID	CustomerID_totalsByAgeBin	CustomerID_totalsByBookCode	CustomerIDDist%
3	22720	30-39	119	278	640	42.805755
4	22720	40-49	239	546	640	43.772894
5	22720	50-59	228	523	640	43.594646
10	22722	30-39	70	278	393	25.179856
11	22722	40-49	151	546	393	27.655678
12	22722	50-59	142	523	393	27.151052
17	22979	30-39	51	278	225	18.345324

Figure 11: ***Customer ID distribution by total customers in each Age bin and by total customers for each book code***

- Customer Age groups are structured after sorting the percentage distribution table by Book code and percentage customer distribution and then picking up age groups the top for each Book code aggregating by Book code

	BookCode	AgeBin	CustomerID	CustomerID_totalsByAgeBin	CustomerID_totalsByBookCode	CustomerIDDist%
26	84987	50-59	77	523	203	14.722753
17	22979	30-39	51	278	225	18.345324
11	22722	40-49	151	546	393	27.655678
4	22720	40-49	239	546	640	43.772894

Figure 12: ***Customer profiles after sorted by book code and percentage customer distribution***

5. RESULTS

5.1. Top 4 Books

As the top positive sentiment scores of the customer reviews indicated the highest rated books, the top 4 books are picked up and these books are the following ones.

Ranking	Book Code
1	22979
2	22720
3	84987
4	22722

Table 8: ***Top 4 positively rated books across the United States***

5.2. Sales Forecast of top 4 books for next 2 months

The sales forecast for the top 4 books across all the 21 states for the next 8-9 weeks are predicted. This forms a forecast of 757 data points. This forecast will help the retailer to plan for the demand and stock of these books for the upcoming 2 months and that can help in planning for the required growth and strategy making for sales.

State	Week	BookCode	Units Sold
Arizona	2019-11-03 00:00:00	22722	5162
Arizona	2019-11-10 00:00:00	22722	5078
Arizona	2019-11-17 00:00:00	22722	5076
California	2019-11-17 00:00:00	22722	18120
California	2019-11-24 00:00:00	22722	18142
California	2019-12-01 00:00:00	22722	15303
Ohio	2019-11-17 00:00:00	84987	40110
Ohio	2019-11-24 00:00:00	84987	39676
Ohio	2019-12-01 00:00:00	84987	38115

Table 9: ***Sample of the sales forecast of few books in different States of America***

5.3. Customer Profiling

After completing the state and holiday wise profiling for different age groups of customers, it has been found that there are certain combinations of age group customer who buy any one of the top 4 books during a holiday season.

Top 4 BookCode	Consumer Profile
22720	People in the age range 40-49 from 'Florida', 'California' are more likely to buy book 22720. This book has more or less a similiar average distribution of units sold so we cant say its popular during gift season.
22722	People in the age range 40-49 from 'Florida', 'New York' are more likely to buy book 22722 as a gift as the average units sold is higher during gift season.
22979	People in the age range 30-39 from 'California', 'New York' are more likely to buy book 22979 This book have more or less a similiar average distribution of units sold so we cant say its popular during gift season.
84987	People in the age range 50-59 from 'Florida', 'California' are more likely to buy book 84987 and customers are least likely to buy book 84987 as a gift as the average units sold is lower during gift season.

Table 10: ***Customer profiling for each of the top 4 book code***

6. DISCUSSION

The study intends to find out the top selling and reviewed books for a book retail company. Also, to help the company by predicting the book sale demand forecast for the next 2 months and targeting the most potential buyers by profiling the customer based on their age groups, state and holidays. It is also discussed about the sentiment analysis of the customer reviews of the book using Machine Learning

technique of sentiment analysis. TextBlob and VADER sentiment analysis are performed to find the sentiment scores. Time series analyses has been done and book sale forecast is performed.

There are some obvious limitations to these analyses as most of the sales and customer data are created manually by illustrating some of the industrial data taken as reference. The Sold units of books and price and the available stocks are illustrated which can be different for any of the book retail company. Also, the book reviews are taken from various online repositories which can be biased based on the customers from different regions and the reviews for actual books. The customer ages are also randomly generated and mapped with the book codes. But these entire analyses can give a fare demonstration of the entire analytical steps to perform the sentiment analysis and sales, or demand forecast of any product and to perform customer profiling for any industrial use case.

REFERENCES

[1] Mudambi, S. M., & Schuff, D. (2010). Research Note: What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.com. *MIS Quarterly*, *34*(1), 185–200. https://doi.org/10.2307/20721420

[2] Addanki, Mounika, Saraswathi, Dr,Scholar, Research. 2019/07/12. CLASSIFICATION OF BOOK REVIEWS BASED ON SENTIMENT ANALYSIS: A SURVEY. doi 10.13140/RG.2.2.11576.29447

[3] Wang, X., Yucesoy, B., Varol, O. et al. Success in books: predicting book sales before publication. EPJ Data Sci. 8, 31 (2019). https://doi.org/10.1140/epjds/s13688-019-0208-6

[4] The art of customer profiling Why understanding audience is important and how to do it. Experian. https://www.experian.co.uk/assets/marketing-services/white-papers/wp-the-art-of-customer-profiling.pdf

[5] S. G. Kapoor, P. Madhok, S. M. Wu. Modeling and Forecasting Sales Data by Time Series Analysis. First Published February 1, 1981, Research Article https://doi.org/10.1177/002224378101800110

[6] Bajaj, Aryan. "Can Python understand human feelings through words? – A brief intro to NLP and VADER Sentiment Analysis". Data Science Blogathon, Analytics Vidya, 17 June 2021. https://www.analyticsvidhya.com/blog/2021/06/vader-for-sentiment-analysis/

[7] Book Publishing Dataset. https://gist.github.com/apietrick24/bfffc6c0d47abf00029790381e89626d