Elizabeth Earl

DSC550-T301

Course Project

Final writeup

Introduction:

Initially my project started as potential Amazon ads targeted based on a shopper's buying trends, quickly I learned this project would be difficult because of consumer data protection. My project became a way for Amazon to use predictive models to target consumers with top rated items. The purpose of this project was to solve the problem consumers had on not knowing what to buy on Amazon or ads being irrelevant.

When people spend hours scrolling on a site it is great as it brings in traffic, but that traffic only brings in so much revenue for a company, which is what a company like Amazon needs to guarantee sales. By targeting an audience with ads of top-rated products scrolling audiences become consumers spending real money and thus creating a bigger revenue for Amazon. When you (Amazon) begin pushing top rated and reliable items (based on previous reviews) you gain consumers trust which leads to greater sales in the future.

Gaining trust of buying is the best way to gain revenue and although it may be an unethical way to push company created products onto consumers it still brings in the sales/money which is the ultimate goal. By spending money on a predictive model

product Amazon will essentially be investing in their own future. All revenue gain must start with a slight revenue loss but, in this case this modeling project will only create gain from the start.

As mentioned above, consumer's shopping trend data was hard to come by so instead top-rated book and electronics data sold on Amazon in previous years was used. Both datasets were found on Kaggle, links have been provided below.

- Top Books data: Amazon Top 50 Bestselling Books 2009 - 2019 | Kaggle
- Top electronics data: Amazon Top 100 Best Sellers in Electronics 2021 | Kaggle

Summary of Milestones:

As mentioned, initially (during milestone 1) I was only working on data based on the top 50 bestselling books on Amazon. In the beginning the route this project was taking was a track based on COVID. Originally, I was assuming COVID caused consumers to resort to indoor activities such as reading thus leading to an influx of book sales and using such data to create future ads based on the popular books over the years.

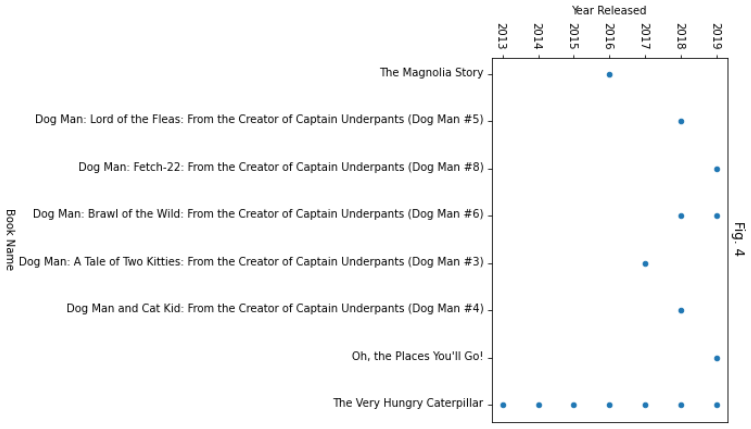


Fig. 4

Because the figure above, I noticed it would be difficult to continue this project with only books data as often books were rereleased in different years.

For milestone 2 I introduced the top electronics dataset to add move values to my overall project. I relied heavily on data cleaning for this milestone to where I made the datasets similar as shown below:

```
In [187]: books.head()
```

Out[187]:

	Name	Author	Rating	Reviews	Price	Year	Genre
534	Where the Crawdads Sing	Delia Owens	4.8	87841	15	2019	Fiction
382	The Girl on the Train	Paula Hawkins	4.1	79446	18	2015	Fiction
383	The Girl on the Train	Paula Hawkins	4.1	79446	7	2016	Fiction
32	Becoming	Michelle Obama	4.8	61133	11	2018	Non Fiction
33	Becoming	Michelle Obama	4.8	61133	11	2019	Non Fiction

```
In [188]: electronics.head()
```

Out[188]:

	Year	Rank	Name	Rating	Reviews	Price
13476	2021	77	Echo Dot (3rd Gen) - Smart speaker with Alexa ...	4.7	1154421	$34.99
13403	2021	4	Echo Dot (3rd Gen) - Smart speaker with Alexa ...	4.7	1154421	$24.99
13305	2021	6	Echo Dot (3rd Gen) - Smart speaker with Alexa ...	4.7	1151460	$24.99
13205	2021	6	Echo Dot (3rd Gen) - Smart speaker with Alexa ...	4.7	1150403	$24.99
13274	2021	75	Echo Dot (3rd Gen) - Smart speaker with Alexa ...	4.7	1150403	$24.99

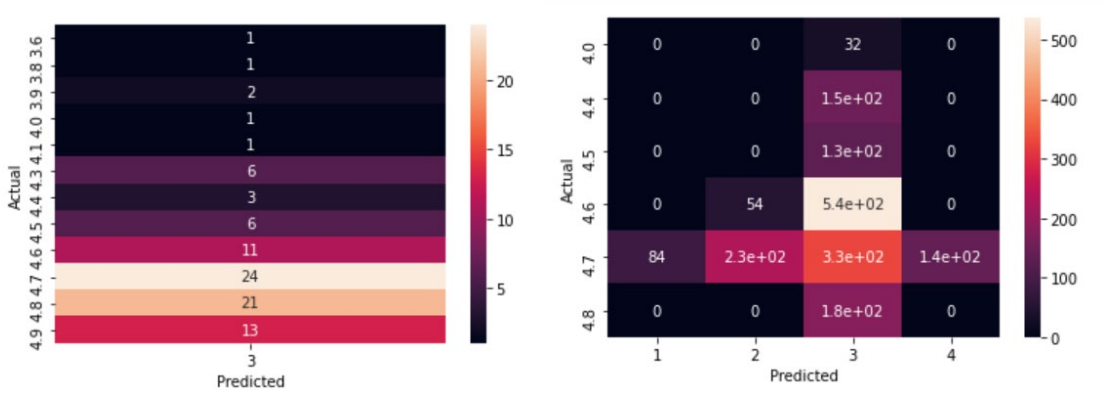
Neither of the datasets I was working with contained null/NAN values, so I simply focused on books/electronics that met certain criteria. The criteria I wanted each item to meet was over 2.0 rating and contain reviews over the average number of reviews per dataset. As noticed in milestone 1 (with the figure above) some items were duplicated, I removed duplicates but only if the years were the same. This idea that something was

rereleased in following years seemed to be important as the electronics had similar trends.

I decided not to merge my books and electronics datasets as some beneficial data was exclusive to each data (such as genre for books and pricing for electronics). Therefore, milestone 3 still required me to work on each dataset separately. When comparing books and electronics dataset I noticed that the electronics had less variety in rating scale versus the books dataset as can be seen below (where left is electronics dataset and right is books dataset).

Rating	Count	%
4.7	1545	41.18
4.6	1191	31.74
4.8	382	10.18
4.4	296	7.89
4.5	259	6.90
4.0	79	2.11

Rating	Count	%
4.8	51	28.33
4.7	44	24.44
4.9	25	13.89
4.6	19	10.56
4.5	14	7.78
4.3	8	4.44
4.4	6	3.33
4.0	4	2.22
4.1	3	1.67
3.9	2	1.11
3.8	2	1.11
4.2	1	0.56
3.6	1	0.56



This information was crucial in determining what rating I wanted to focus on. The top ratings from both databases were a rating from 4.6-4.9. Based on such information and after calculating through a Logistic Regression model I concluded the perfect item to suggest to people is one with a 4.7(if an electronic) and 4.6(if a book).

Conclusion:

After calculations and model fitting, I determined the best items that should be advertised to consumers to bring up Amazon's revenue. Below are the ideal items, based on both tables that should be advertised to consumers:

```
books = books.sort_values(by='Reviews', ascending=False)
books.head()
```

	Name	Author	Rating	Reviews	Price	Year	Genre
21	All the Light We Cannot See	Anthony Doerr	4.6	36348	14	2015	Fiction
20	All the Light We Cannot See	Anthony Doerr	4.6	36348	14	2014	Fiction
75	Divergent	Veronica Roth	4.6	27098	15	2013	Fiction
76	Divergent	Veronica Roth	4.6	27098	15	2014	Fiction
465	The Subtle Art of Not Giving a F*ck: A Counter...	Mark Manson	4.6	26490	15	2018	Non Fiction

```
electronics = electronics.sort_values(by='Reviews', ascending=False)
electronics.head()
```

	Year	Rank	Name	Rating	Reviews	Price
13476	2021	77	Echo Dot (3rd Gen) - Smart speaker with Alexa ...	4.7	1154421	$34.99
13403	2021	4	Echo Dot (3rd Gen) - Smart speaker with Alexa ...	4.7	1154421	$24.99
13305	2021	6	Echo Dot (3rd Gen) - Smart speaker with Alexa ...	4.7	1151460	$24.99
13274	2021	75	Echo Dot (3rd Gen) - Smart speaker with Alexa ...	4.7	1150403	$24.99
13205	2021	6	Echo Dot (3rd Gen) - Smart speaker with Alexa ...	4.7	1150403	$24.99

This information will help give the best suggestion to consumers which will result in sales. If Amazon uses this method to push top items for other categories (besides books and electronics) they can monopolize on all products being sold. This predictive method will allow consumers with a better shopping experience as they are buying reliable items. The Logistic Regression model helped predict how accurate it would be that a given item will be a top rated and thus frequently purchased item. As mentioned at the start of this report, this can also be used unethically by Amazon to use such information to then push their own branded “knock off” version of top items being sold to increase revenue.