Baby_products_ML

Muneeb Hassan

December 2024

Contents

1 EDA

The project involves working on amazon_baby_products data set from kaggle. This is a sentiment analysis project which predicts the ratings of the products based on the review left by the user.

1.1 Understanding the data

The following features I observed in our dataset:

- The dataset conssit of 3 rows known as **name** (Name of the product), **review** (The review left by the user), and **rating** (The rating of the product which ranges from 0-5).
- This is a big dataset with 183,530 columns.
- Seems like the **name** column might not be necassary for our analysis so we might drop it.

1.2 Cleaning the data

First we will need to deal with null values. The following table shows the total number of null values in each row.

name	318
review	829
rating	0

There are no null values in the **ratings**. There are some null values in **name** and **review**. However, total numer of these values are almost non-existent as compared to the total number of rows in the dataset. Thereferore, it's better to remove these columns with null values from our dataset.

Using the **pandas'** bult-in **dropna()** function, we can drop all the rows which have null values.

We probably do not need the **name** as it won't help to predict our model. The following is a snapshot of our dataset after removing null values and dropping the **name** column:

	review	rating
0	These flannel wipes are OK, but in my opinion ...	3
1	it came early and was not disappointed. i love...	5
2	Very soft and comfortable and warmer than it l...	5
3	This is a product well worth the purchase. I ...	5
4	All of my kids have cried non-stop when I trie...	5
...	...	...
183526	Such a great idea! very handy to have and look...	5
183527	This product rocks! It is a great blend of fu...	5
183528	This item looks great and cool for my kids.....	5
183529	I am extremely happy with this product. I have...	5
183530	I love this product very mush . I have bought ...	5

Figure 1: Snapshot of our dataset

Since the rating can only range from 0-5, we donot need to manage outliers.

We'll plot the total number of each ratingsusing bar graph to get the mode. The following bar graph depicts the value count for each rating.

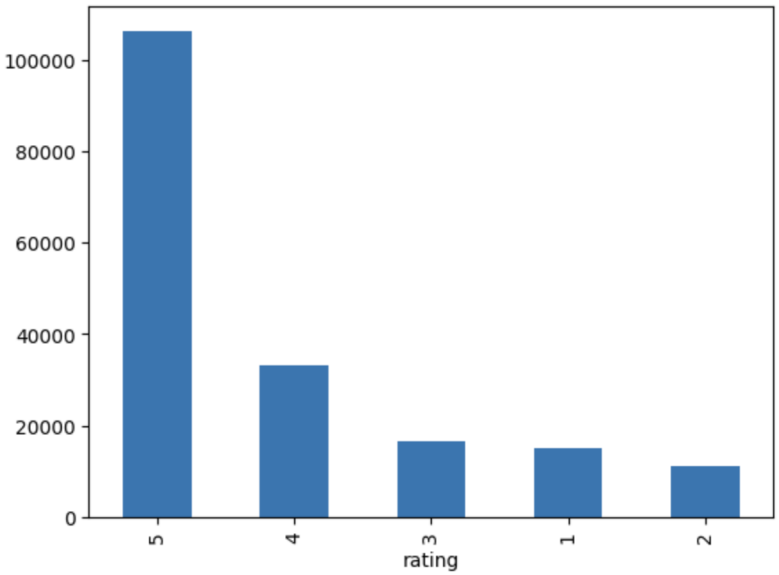

Figure 2: Bargraph for our value counts.

As you can observe, the dataset is little biased towards higher ratings.