Student Name: Brett Petch

Student Number: 251038051

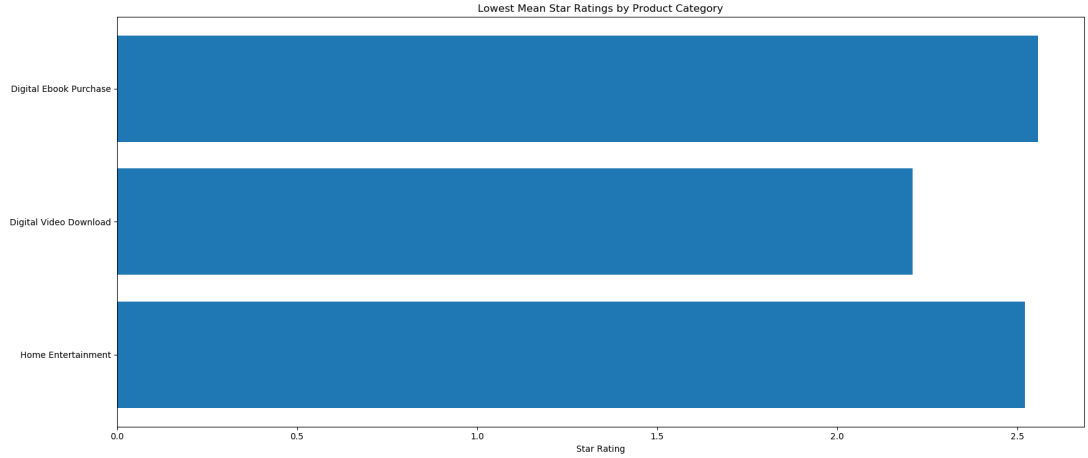
The dataset I chose was the Amazon Review Dataset, US multilingual. I chose the dataset due to the multi-star ratings, large amount of data and text to be analyzed. The data is quite interesting, especially as the machine learning approaches I took were very automated due to the high number of entries. The first visualization was a correlation of helpful reviews, but the lowest rated average by department on Amazon. From here, I continued with the dataset and created a bag of words approach for data analysis.

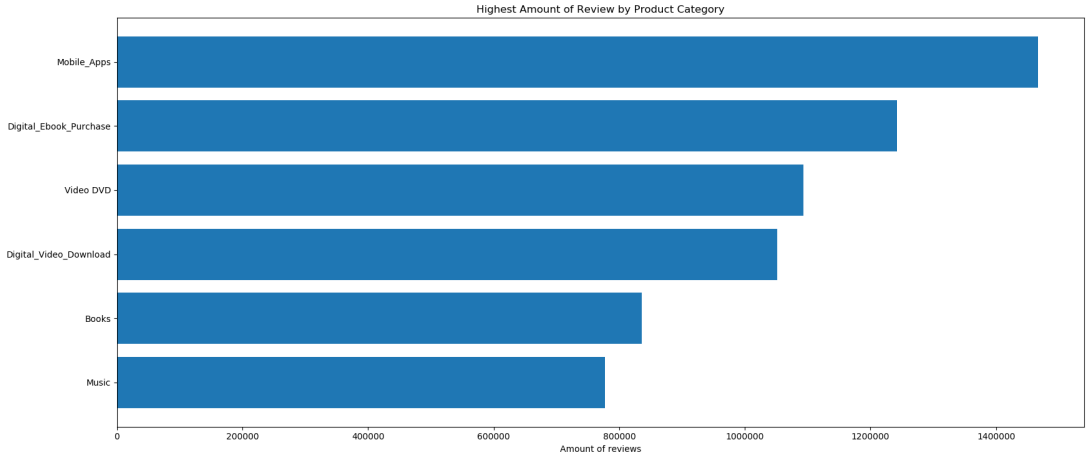
The machine learning tasks verify the validity of reviews, create a bag of words representation, then model around that to create a vector to compare phrases to. From the machine learning methods used, I discovered that because the dataset was so large, the category of the reviews matter. I was unable to solve this within the time constrains of the project.

The different algorithms used resulted in different outcomes, including a lower AUC for the 2nd used.

The first visualization (as provided below) show the correlation of product category and lowest average star ratings. The 2nd visualization is a representation of the highest amount of reviews by top 5 product categories.

The dataset will automatically download itself into the directory where python is running upon running.


Lowest Mean Star Ratings by Product Category
Digital Ebook Purchase
Digital Video Download
Home Entertainment
0.0
0.5
1.0
1.5
2.0
2.5
Star Rating


Highest Amount of Review by Product Category
Mobile_Apps
Digital_Ebook_Purchase
Video DVD
Digital_Video_Download
Books
Music
0
200000
400000
600000
800000
1000000
1200000
1400000
Amount of reviews