Daniel Lung
DS210
12/15/23

Final Project Write-Up

I originally proposed to work on a dataset of the top 20 movies released annually from 1989 to 2014, but I changed it to a dataset containing information about books, ordered by sales rank. The steps I will take are no different from the steps I initially proposed for my original dataset. My code is split into 2 modules and 1 main file. All it takes to run is the dataset in the same folder as the directory and “cargo run”. It outputs several things to check if the file is properly read and handled, such as the first entry of the dataset as well as outputting my desired features and target. Then it outputs other data that can be interesting to look over. In the end, it outputs the accuracy.

The first choice I had to make while working with the dataset was deciding how to handle the data. I initially considered making a Book struct with all of the relevant features, but opted to use HashMaps. My first instinct for this dataset was to focus on regression as a way to predict how well books with chosen features would do for the target. I spotted some relevant features that I felt would be good, such as “Book_average_rating” and “gross sales”. However, I felt that it would be interesting to analyze 2 features further and introduce the "Publisher_rating" feature by calculating the average book ratings per unique publisher. This seemed like a feature that would be very useful, and would also provide relevant information to someone interested in seeing which publishers produced highly rated books.

I encountered many difficulties trying to implement regression in Rust. The process in Python is very straightforward thanks to tools like scikit-learn which streamlined the whole process into just a few lines of code. I was also struggling with selecting the right regression methods, leading to many failed attempts while trying to code. One of the greatest problems I had was figuring out why my matrix dimensions were mismatched and I could not calculate predictions until I resolved that issue. Another issue I faced, despite being minor, was that some of the headers had whitespace that I did not notice, and this caused me to spend a lot of time figuring out what was wrong.

Once I finally got the code to successfully output something, I was faced with a huge issue. The MSE and MAE values I calculated were very high. I initially suspected it was an issue with my data cleaning, and I was partially correct. At first, I was removing every entry that had a missing value. I realized that was unnecessary as I only needed to remove entries that had missing values in the specific features I was looking for. However, this only made a small difference in my results. Then, I experimented with scaling the features to see if that would affect the results. I tried min-max scaling, and although it helped, it was also only by a little bit. Then I tried using z-score normalization, which contributed to better MSE and MAE scores. However, I still felt the MSE and MAE scores were unsatisfactory. I played around with feature selection for a while longer, but in the end, I was left with 5 features, “Publishing Year”, “Book_average_rating”, “gross sales”, “sale price”, and my feature, “Publisher_rating”.

In summary, my code essentially:

1. Reads the data into HashMaps,
2. Cleans data by
    a. trimming any whitespaces in the feature headers
    b. filtering out entries without required features
3. Computes and appends “Publisher_rating” to the dataset
4. Select the most relevant features and target variable
5. Normalizes the data
6. Applies linear regression
7. Creates predictions
8. Calculates for the MSE and MAE

The final result was MSE: 323894943 and MAE: 17389.

This was extremely unsatisfactory, and despite various attempts to refine the model, I concluded that regression is not the optimal tool for this dataset. However, this attempt has provided valuable insights into making use of the Rust language and I have recognized how important it is to understand the dataset I picked and how important that is for choosing the appropriate methods to analyze it.