Cheng-Shao(Eric), Lin

Prof. Hongchang Gao

Principles of Data Science

5 April 2024

Final Project Progress Report I

Project title: Dog Breed Rating Plot

**Abstract**

My project process is moving smoothly so far, thanks to a calculated approach that ensures a successful completion. While I began work on this by doing research, I quickly started applying new methods and techniques that I was researching, resulting in a multi-step process for the beginning of my project. The overview of my project is creating a system in which the user, a college student, can select a dog breed specific to their needs based on their preferences. The system would then show the user what kind of dog they should get, and give them a match rating from 1-5, with 5 being the best possible match for the user. Additionally, the system will redirect the user to a URL with information about the dog breed that is the best match for them.

**Process**

The first step of my process was loading the dataset containing the dog breed trait information from the CSV file. Then I checked to confirm if there were any missing values, and I found 7 missing. The values were ‘weight’ and ‘height’. Therefore, I removed the non-numeric characters from them and converted the cleaned values to numeric format. I used errors=’coerce’ to ensure that whether any values could not be converted to numeric, they would be replaced with NaN which acts as a missing value to avoid error. By calculating the mean of the already given values, I was able to replace the missing values with it. The second step in my process was checking the categorical features of the given data. The features that I checked are breed name, a website url that shows the user what dog is recommended to them, the breed group, and life span of the dog. Therefore, I picked breed group and life span in order to convert them into the correct format. In addition, I set up an example of user preferences that would make these dogs suitable for college students. The user preferences I selected were dogs that can adapt well to apartment living, it being able to tolerate being alone, whether the dog is friendly, and general health, in addition with a rating, from a scale of 1 to 5, with 5 being the maximum score. I chose these specific preferences based on typical college student traits and living situations. Additionally, I defined a function that could calculate the similarity of the user preferences and the ranking for each user from a scale of 1-5. In order to achieve that, I initialized the similarity value to zero, and used a for-loop with if-statement, so if the attribute is in the user preferences, then it would calculate the similarity.

Additionally, I applied the similarity function into each row in the DataFrame and created a new similarity column. Using the similarity column as a foundation, I adjusted the order in ascending order, resulting in a recommendation of the top 10 breeds for the selected user. As the result, the output will give the user a breed name, the URL of their preferred dog and the similarity score based on their preferences.

**Conclusion**

Loading the data, finding the missing values and converting the categorical features was the easiest part of this assignment as it is something we have worked on during lab exercises. However, finding the similarities, obtaining the user preferences and ranking the preferences based on the similarity score with the breeds was a more time consuming task, that could be refined in the future of this assignment as well.

Moving forwards, I am going to obtain the linear regression by using the user similarity score. Next I will be working on classification methods in order to classify user preferences and refine the recommendations given to them. Additionally, I will be calculating the accuracy, recall, and F-1 score in order to compare the performance of each different breed model. I will also be working on visualizing the plot for recommended breeds based on the score they receive. By using all these methods, I will be refining the recommendation system and making it more efficient.