Aditya Srikanth
aks9136

Data Analysis Project Two

Introduction

In this report, we will discuss the general process and specific decisions made in order to answer the assigned queries. The dataset used was a movie reviews dataset that contained both movie ratings and personality-adjacent questionnaire response for thousands of users. Missing data was imputed using the computed mean of the missing data point's respective column. The various data science methods outlined in the report were implemented using Python and specifically utilized the Scikit-Learn library. The dataset was stored and processed primarily using the Pandas library.

User Correlation

In order to answer various questions regarding user correlation, the first step taken was to transpose the dataset given to yield a structure that represented the movie ratings and questionnaire responses of each user as a column. As such, the dataset was made up of thousands of columns, each representing a single, unique user. The built-in Pandas method corr() computed the correlation between two columns (two users). Using this, we were able to understand how strongly any two users were correlated. For each user, we could then find their most correlated "buddy" user; to organize the information, we generated a dictionary that contained user and "buddy" user pairs as key-value pairs in the dictionary. The built-in function also reported the correlation strength, which allowed us to find the most correlated pair by searching through all correlations in our dataset.

Generally speaking, the correlations between users were widely varied, but some users had very strong correlations. It is, however, worth noting that when two users had many missing values, their correlation was reported as very strong. This is because their missing values were imputed identically and as a result, two users with many imputed values would be similar in nature. Given the specifications of this assignment, this was largely unavoidable.

Aditya Srikanth
aks9136

Modeling Ratings and Personality

Given a wealth of data clearly demarcated as movie ratings and personality, attempting to model a relationship between these two parts of the data was a worthwhile endeavour. We began with a simple linear regression, using the movie ratings as predictors for various personality traits. The linear regression utilized a train-test split where 20% of the data was withheld from training in order to serve as test data. The linear regression model performed fine (clearly showed trends, but did not solidly establish a correlation beyond doubt), but optimizations to the model were clearly available.

The first modification to the model that we employed was to modify the model into a Ridge Regression, utilizing a regularization parameter in order to make the model more generalizable. This was seen in basic testing; while training saw more error, the test data outperformed the simple linear regression model as it was designed with generalization in mind.

Next, we tried a Lasso Regression model. This model showed an improvement as well, specifically peaking at a specific alpha parameter. It is possible that any of the three aforementioned models could benefit from any number of additional optimizations. However, as a general inquiry and an introductory prototype, we stop at this level of model implementation.

Conclusion

Overall, it appears that we can make some tie between movie ratings and personality traits. However, in order to confidently make this type of connection, we believe that more data (specifically less reliance on imputing) and additional model optimizations would be necessary. Otherwise, we observe uncertainty and some weakness with correlation-based claims.