Further Findings and Curiosities from Published Movie Data (Wallisch & Whritner, 2017)

as reported by: Guilherme Albertini

Summary

I continue from where work was left off on Project 1: here, we are interested in correlational study among users and developing linear regression models with the movie rating and personality data. Once again, to cut down on false positives, the per-test significance level α is set to 0.005 (as per Benjamin et al., 2018). Some key findings include:

- User 831 and 896 are the most correlated with coefficient 0.9995
- Lasso regression model with hyperparameter of 0.1 is the best-performing model
- Ordinary least squares and ridge regression models did not generalize well
- Lasso models reduced features enough to yield lower mean absolute errors, suggesting much fewer than the 400 features are explaining the variance in the data

Discussion

Given the feedback of 1097 participants ("users"), I find the maximum absolute value of the correlation coefficient. For every user in the data, I found its most correlated user and find the most correlated pair. Note that throughout this project, instead of row-wise elimination of missing data, column averages filled in the missing entries. Users gave a rating (0 to 5) to each of the 400 films and a rating (1 to 5) for a set of personality questions detailed in the notebook. I assume the data are derived from a random, or at least representative, sample; I assume all these 477 variables are continuous, jointly normally distributed, random variables and any pair follows a bivariate normal distribution in the population from which they were sampled. *Using the pandas correlation method for Pearson's coefficient, the most correlated pair of users were* ***User 831 and 896*** *with a coefficient of* ***0.9995***. This would suggest there is an almost perfect linear relationship between the two! The first 10 correlated user pairs are shown in Figure 1.

User 0	User 1	User 2	User 3	User 4	User 5	User 6	User 7	User 8	User 9
User 583	User 831	User 896	User 364	User 896	User 99	User 239	User 896	User 896	User 1004

Figure 1: *Correlation pairings of the first 10 users. The maximum absolute Pearson coefficient is used.*

It is very bizarre that User 896 is the most correlated with respect to these specific users, but this analysis was done with replacement; that even if the coefficient of correlation was higher for pairing User 2-User 896 than User 7-User 896, both would still show the most correlated user of all users. Further analysis is warranted.

I then attempt to find a linear relationship between a user's movie ratings data and the personal questions that were asked. An 80:20 split of the data from training to test is considered in this analysis.

An ordinary least squares regression (OLS) is fit first: the columns of sensation-seeking behaviors (Columns 401-421) query from a set of questions that lie on a risky/high energy-to-risk-averse spectrum and were aggregated to form our dependent vector. The 400 columns of the movie reviews became the independent variables with their respective weights becoming predictors matrix. Next, both ridge regression and lasso regression models were tried with specific hyperparameter values. The mean absolute error for all models (both training and test data) is shown in **Figure 2**. Consult the notebook for further information.

The lasso model with a hyperparameter ("alpha") of 0.1 gave the lowest overall error of all models considered. This finding suggests that many of the features (or many of the movie reviews) were not explaining the variance in the model; dimensionality reduction should be studied.

With this in mind, the lasso models were compared against the baseline OLS model to view how the predictor weights decreased (or became 0) with the different weights of hyperparameters, with special attention placed on the model mentioned prior. As expected, an aggressive value of 1 for alpha would cancel out many features leading to a sparse predictor weight matrix but a very low value would keep features (essentially certain movie review columns) that were not impactful when it came to predictive insight: they did not explain the variability in the data as much as other features did. The first few of these feature values are shown in **Figure 3.** See notebook for further information.

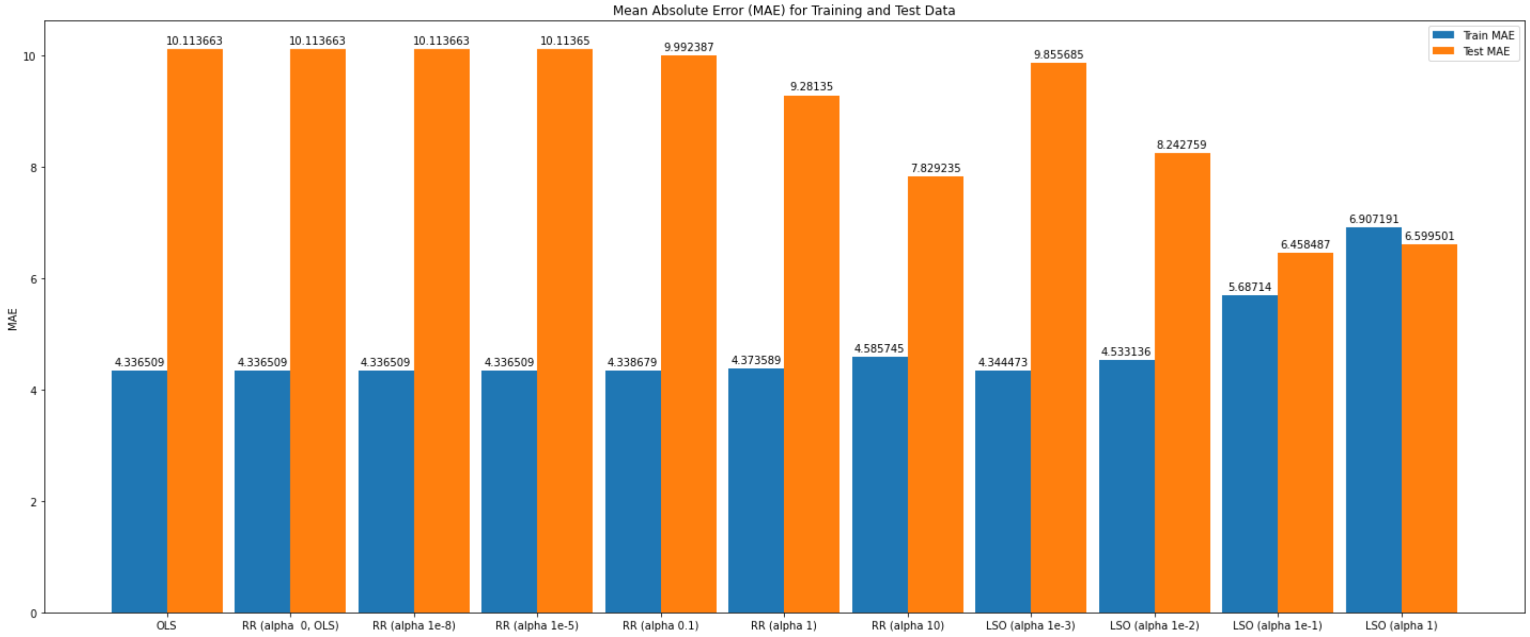

Figure 2: *Mean absolute error (MAE) for all models considered for both training and test data. Note that the Lasso model with alpha at 0.1 gave the best-fitting model.*

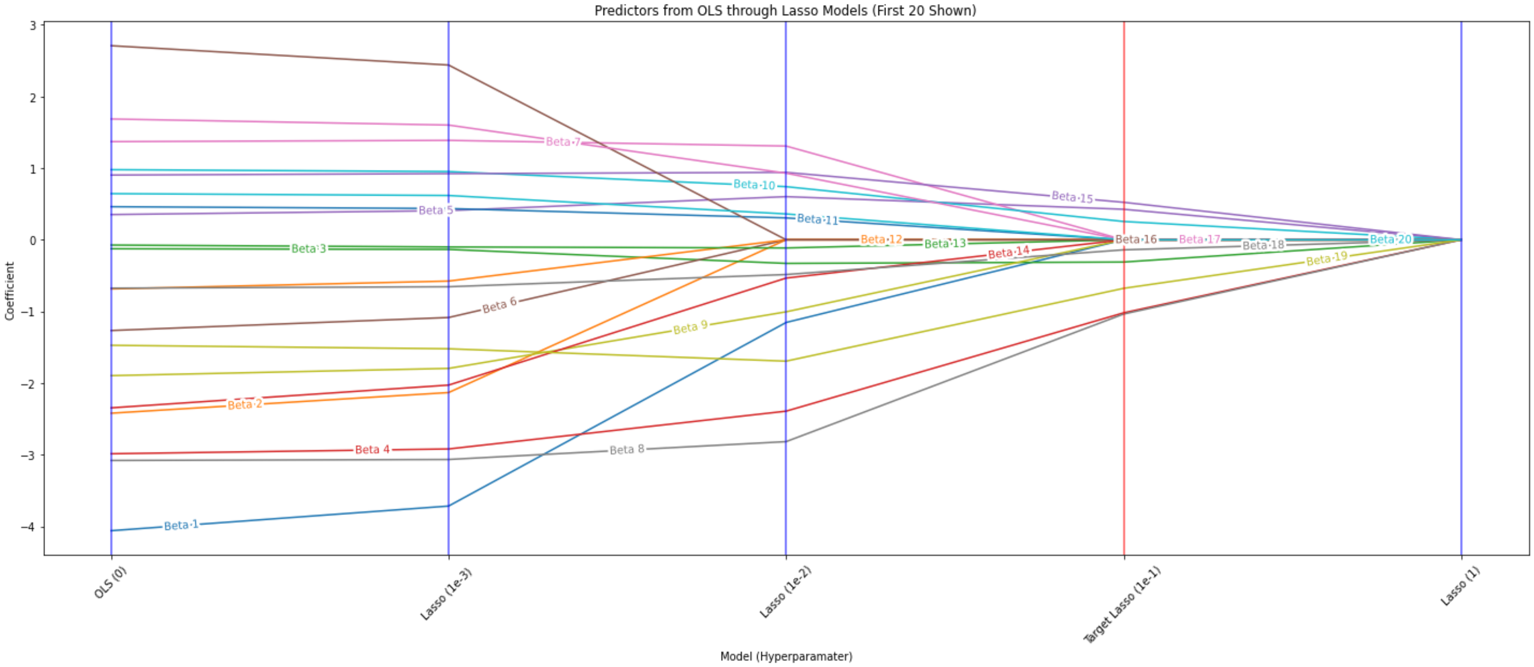

Figure 3: *OLS and Lasso models showing the predictor values over several different penalty terms. Features would tend to 0 (or drop out altogether) when increasing lasso hyperparameter.*