Final Project Report

Raleigh Goodwin

Introduction

For this project (Github link), I am working with a data set that was collected in my lab last year. This study examined the potential relationship between participants' evaluation of politicized scientific information and trait-level science curiosity. Months prior to data collection for this particular study, the participants completed a baseline assessment that included two cognitive ability measures — Raven's Progressive Matrices (`rav_scored`; ranged 1-10) and verbal logic (`vl_scored`; ranged 1-4) — and cognitive processing style measures — Need for Cognitive (`nfc_scored`; ranged 1-6) and Actively Open-minded Thinking (`aot_scored`; ranged 1-7). They also provided demographic information (i.e., age, gender, race, education) and reported their political ideology (`conservatism_f`; ranged 1-7, where higher values indicated greater levels of conservatism).

Later, the same participants (N=538) were invited back to participate in this study. This time, they completed a task in which they saw a series of graphs depicting longitudinal data about a variety of different environmental metrics. They were then asked to make projections about relative future values (e.g., whether future levels will be higher or lower than current levels) of those metrics, given the information given to them by the graphs. These graphs differed in two crucial ways. The first variable was the political or apolitical nature of the metrics each graph displayed. The political graphs displayed metrics related to climate change (e.g., atmospheric CO_2 levels); the apolitical graphs showed environmental metrics unrelated to any political platform (e.g., population levels of harmless, lake-dwelling bacteria).

Second, the graphs differed in whether the slope between the final two data points was congruent or incongruent with the global trend of the data. In instances where the local and global trends are incongruent, people can be susceptible to endpoint bias, a phenomenon in which people overweight recent, local trends in longitudinal data as opposed to overarching, robust global trends. In this study (or in the aspect of the study I'm focusing on for this project), we tested whether conservative participants responded with more endpoint bias on politicized items when local trends were incongruent with global trends but congruent with conservative ideology. Participants provided estimates for future values of each metric 10 years from now and 20 years from now. Performance on these items is represented in the data by the variables ending in `_ch`, which indicate the extent to which a participant responded in line with the local trend over the global trend (1 = heavily weighted global trend, 7 = heavily weighted local trend). After this task, participants completed a science curiosity inventory (the score of which is represented by `sc_scored` in the data, where higher values indicate greater science curiosity). This score was calculated using a formula developed by the creator of the scale.

Lastly, they answered three open response questions: 1) one in which they reported what they were thinking and feeling during the graph task, 2) one in which they reported what they think the graph portion of the experiment may have been studying, and 3) one in which they reported what they thought the science curiosity inventory may have been studying (it was embedded in a larger "general interest" inventory). I generated sentence embeddings for each question and they are included here, labeled `o1`, `s1`, and `s2`, respectively. From my initial visual inspection, the responses to these items seemed interesting because there seemed to be a distinct split between participants on whether they picked up on the political nature of the study. Some provided responses that were very much at face value (e.g., "the graph portion was studying how we understand graphs"; "the interest inventory measured what kind of topics I'm interested in"). Others correctly guessed that we were interested in their evaluation of climate change information

based on political ideology, though neither climate change nor politics was ever explicitly mentioned in the study (political ideology was only reported in the baseline).

In the case of this project, I chose an outcome variable that was not included in our preregistered analyses: climate change risk perception (`cc_risk_sc`, where higher values indicate greater climate change risk perception). While not featured in our formal hypotheses, this variable is (obviously, by definition) a more proximal measure of climate change beliefs than political ideology. Studying climate change risk perception has obvious broader implications from a social and environmental context; however, I also chose to focus on this variable out of curiosity. My preregistered analyses were focused on predicting endpoint bias, using information like political ideology, and the cognitive variables included in this data frame were also not part of my preregistered analyses. I thought this project could be a beneficial opportunity to see the data through a different lens, especially since I would like to examine this data in a more exploratory fashion for hypothesis generation.

Below are tables of basic descriptives for some of the major distinct variables. This data does not contain any missing values that would necessitate imputation; the variables I'm examining had validation requirements on the Qualtrics survey, so participants were not able to move on in the study without answering each question.

|	mean	sd	median	min	max	skew	kurtosis	se
climate_risk	8.478	2.863	9	1	11	-1.149	0.332	0.123
conservatism	3.705	1.807	4	1	7	0.149	-1.208	0.078
science_curiosity	24.621	5.774	25	9	40	-0.161	-0.299	0.249
need_for_cognition	21.364	6.681	23	6	30	-0.634	-0.623	0.288
actively_openminded_thinking	33.056	5.272	34	9	42	-0.745	0.996	0.227
verbal_logic	2.918	0.918	3	0	4	-0.458	-0.603	0.040
ravens_matrices	5.625	1.563	6	1	10	-0.089	0.427	0.067
arctic_ice_bias	2.582	1.311	2	1	7	1.124	0.909	0.057
global_temperature_bias	2.355	1.160	2	1	7	1.302	2.031	0.050
co2_levels_bias	1.916	1.017	2	1	7	1.621	3.660	0.044
icesheets_mass_bias	1.924	1.196	2	1	7	1.829	3.728	0.052

Methods

I tested three modeling approaches: linear regression (unpenalized), linear regression with ridge penalty, and linear regression with elastic net penalty

For every approach, I employed 10-fold cross-validation to examine model performance. The latter two approaches all required hyperparameter tuning. For all hyperparameters, I triangulated the optimal value through trial and error of testing different ranges and levels of granularity. I chose ridge regression over lasso because I believe (?) that the ridge penalty can work well when you suspect that a large portion of features in the data may be unrelated to the outcome variable, and I suspected that the sentence embeddings would add a significant amount of noise to the data.

For ridge regression hyperparameter tuning, `lambda` was tuned, and `alpha` was fixed to 0:
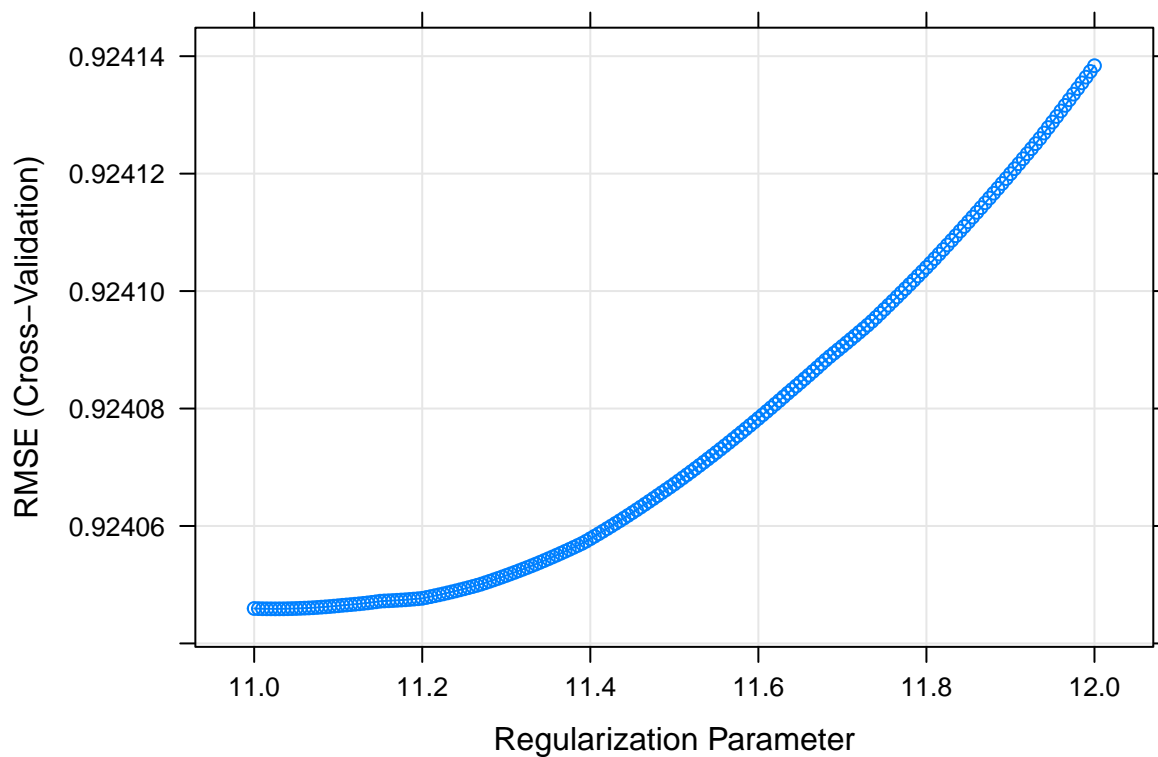
For elastic net, both `alpha` and `lamba` were tuned:
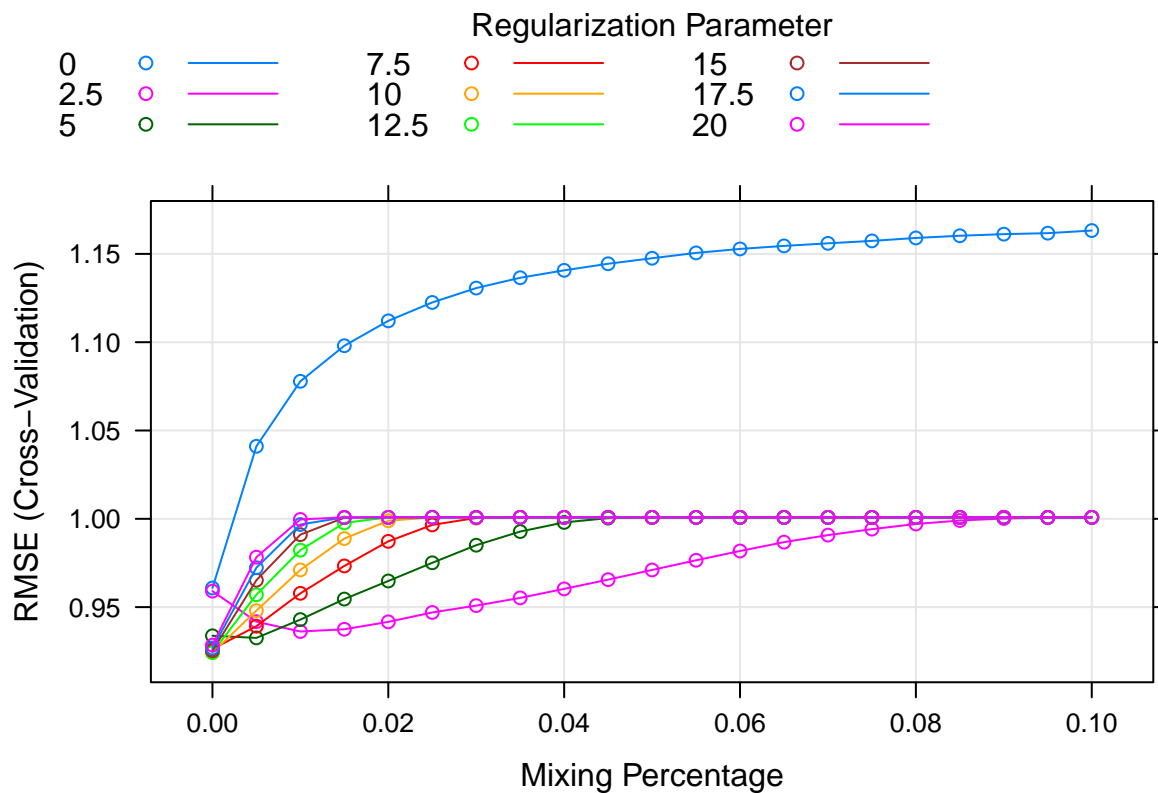
For each model, I planned to judge performance based on comparing $RMSE$, MAE, and R^2.

I also attempted to task a decision tree algorithm, though I have not yet been successful in producing functioning predictive models. I have included the code in the Github repo, but I will not discuss the findings

of that model in this report. Nonetheless, for the decision tree algorithm, in addition to the `complexity parameter`, I attempted to write some for loops to tune `minsplit` and `maxdepth` as well. While I **believe** (?) these loops ended up working correctly, I haven't been able to get R to actually successfully create models that generate predictions without throwing errors. This is something I'd like to continue working on, though, as well as applying Random Forests as well.

Results

In short, the results of all models I initially tested were underwhelming.

V1	Model	Rsq	MAE	RMSE
1	Linear Regression	0.00168	19.78599	26.30839
2	Linear Regression with Ridge Penalty	0.00001	8.89166	9.22951
3	Linear Regression with Elastic Net Penalty	0.00004	8.89141	9.23002

Of these models, the model using elastic net penalty technically performed best, and both regularized regression models significantly outperformed the traditional linear regression approach in terms of lowering RMSE, though they also had even lower R^2 values than the already-small non-reguarlized model. Thus, while it appears that regularization did help reduce overfitting, the amount (or lack thereof) of variance these models explain make them essentially useless. The most important variables in the elastic net model (and, indeed, in all of the models) were all sentence embedding variables; the highest-ranking non-NLP variable (one of the gender dummy variables) was ranked 556 in terms of importance. In fact, in the top 1,000 "important" variables, there is only one other non-NLP predictor (one of the race dummy variables). Given the poor performance of these models, I decided to test these same three approaches with a data set that did not contain any of the sentence embedding features.

V1	Model	Rsq	MAE	RMSE
1	Linear Regression	0.01503	8.87538	9.22931
2	Linear Regression with Ridge Penalty	0.04873	8.88975	9.23565
3	Linear Regression with Elastic Net Penalty	0.08037	NA	9.23429

Interestingly, the RMSE values for all of these models were comparable to the original regularized regression models I created. However, while these models' R^2 values were still relatively small, they were significantly larger than those of the previous models. Once again, the best performing model was the elastic net model. The most important variables (according to the `vip()` function) were largely gender and race, though performance on Raven's Progressive Matrices was ranked 8.

```
vip(elastic2, num_features = 10, geom = "point") + theme_bw()
```
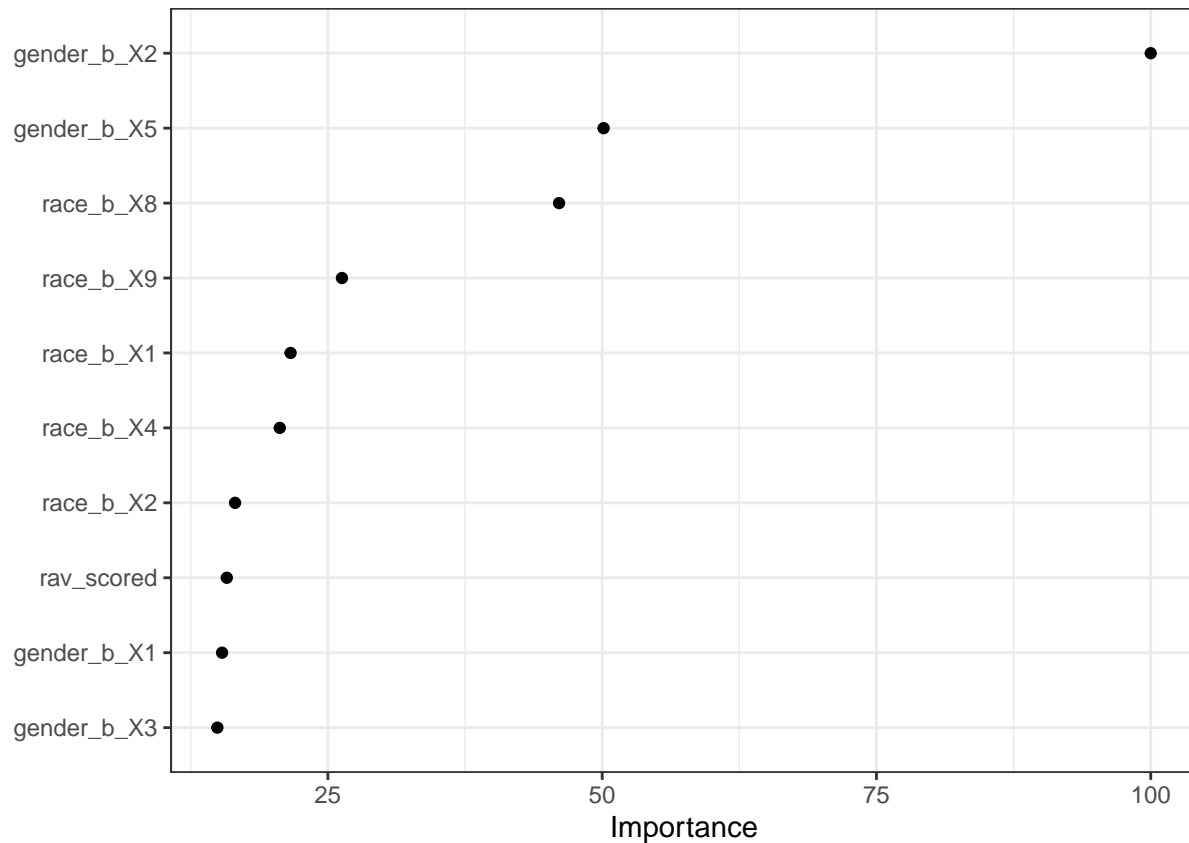
Discussion

The linear regression without penalty performed the worst by far, at least in terms of the original models. Without any penalty, it's unsurpising that a model containing thousands of sentence embedding features loosely related to the outcome variable would be significantly overfit.

All of that being said, I do not believe any of the models I present in this project are close to optimal, and I'm curious to keep working on this further. This is not necessarily surprising as the model I tested through these approaches was not based on specific theoretical rationale. On the other hand, based on my previous/preregistered analyses of this data, I would not have expected the demographic variables to be ranked most important. I think it's possible that, if I were to work on this longer, I could create a model with more predictive power. Though not the main goal of this assignment, I think that attempting to work through those `for` loops for the decision tree models (even if I still haven't gotten them to fully cooperate) was actually quite a helpful exercise in functional programming in addition to the manual tuning of hyperparameters.