Final_Report

Valid Models: Edwin Chau, Joanna Jin, Roger Yuan

12/8/2019

The goal of this project was to predict whether a home team won a basketball game given the values for each predictor. We generated a design matrix using the training data to create dummy variables for our team and league variables. We also reformatted the date variable to reflect the number of days since the first date, though this turned out to not impact the model performances.

Lasso

	No	Yes
No	0.1938025	0.1060924
Yes	0.2128151	0.4872899

Lasso Training MSE: 0.6810924

Our first attempt was to try a LASSO logistic regression model, which would perform both variable selection and prediction simultaneously. This model was a good start, but was the worst of our three top attempts. It had a public and private score of 0.6614.

The LASSO model produced 36 predictors, setting the insignificant ones to zero. We fitted this model using 10 fold cross validation, which gave us an optimal λ that was then used to predict wins for the test set. Normally we would need to standardize our predictors, as ones with large magnitudes will have smaller coefficients. LASSO would then filter predictors with extremely small coefficients, as they are seen to be insignificant. However, the glmnet function standardizes internally, so we don't need to format the data ahead of time.

Our hope for the LASSO model was that it would make the daunting task of modeling 217 predictors a bit easier to manage by filtering out the less "important" ones. However, while LASSO makes a model more interpretable by filtering, a Ridge logistic model would ultimately have the edge when it comes to predicting accuracy. Thus, we moved on to fitting a Ridge model in the hopes it would improve our score.

Ridge With Interactions

	No	Yes
No	0.2122899	0.1101891
Yes	0.1943277	0.4831933

Ridge Training MSE: 0.6954832

Our Ridge model had a public score of 0.66747 and private score of 0.67597. This was a slight improvement on LASSO, which makes sense due to the fact that it does not filter out predictors. By merely shrinking their coefficients and therefore their influence on predictions, the Ridge model could still keep that information around to give it a slight edge over LASSO in terms of accuracy.

Fitting the Ridge model had an identical process to the LASSO. We used cross validation to identify the optimal λ before making our predictions. In an effort to improve the Ridge model further, we added interaction

terms to the existing training data. We did this by computing the correlations between predictors and selecting the ones with a correlation greater than 0.7. This resulted in a model with a public score of 0.66868 and a private score of 0.67839.

Adding interaction terms did not improve our original Ridge model by much in terms of predictive accuracy. This makes sense in hindsight because Ridge has a regularization term that shrinks coefficients towards 0. Linear regression struggles with high correlation between predictors because standard error estimates for these predictors would increase and make predictions more variable. However, Ridge reduces this problem with a regularization term, thus including interactions did not ultimately change much, though it did give the model more features to predict with and increased our accuracy ever so slightly.

Elastic Net

	No	Yes
No	0.2126050	0.1122899
Yes	0.1940126	0.4810924

Elastic Net MSE: 0.6936975

To achieve a balance between the variable selection feature of LASSO regression and the higher prediction accuracy of ridge regression, we looked into Elastic Net regularization as another option. Elastic Net combines the penalties of Ridge and LASSO to get the best of both worlds. By setting the alpha term in glmnet to 0.3, we opted for an Elastic Net model that performs more similar as a Ridge model, encouraging grouping for correlated variables, and also reduces random noise brought along by insignificant predictors. This model yielded a public score of 0.66626 and a private score of 0.68082.

CV Error for all Models