math_158_semesterproject_part4

Kevin Loun & Tesfa Asmara

5/9/2022

Contents

1 Introduction

The dataset for this project contains 10,000 League of Legends ranked matches from the North American region with 775 variables offered through the Riot Games API, provided on Kaggle (Games 2021a)(James 2020). Each match is pulled from players who rank Gold in the League system, a ranking system that matches players of a similar skill level to play with and against each other. Amongst North American players, the Gold skill level was the second most common tier, achieved by 27.7 percent of players, or approximately 49.86 million players when considered against Riot Games' player base of 180 million ("Distribution of League of Legends (Lol) Summoners in North America as of October 2021, by Tier" 2021)(Games 2021b). This dataset will be referred to as `lol10`.

For this project, the following variables are of interest: time spent crowd controlling others, map side, longest time spent living, kills, gold earned, and total damage dealt. A figure including all the relevant variables and their description is attached at the end.

2 Shrinkage & Smoothing Models

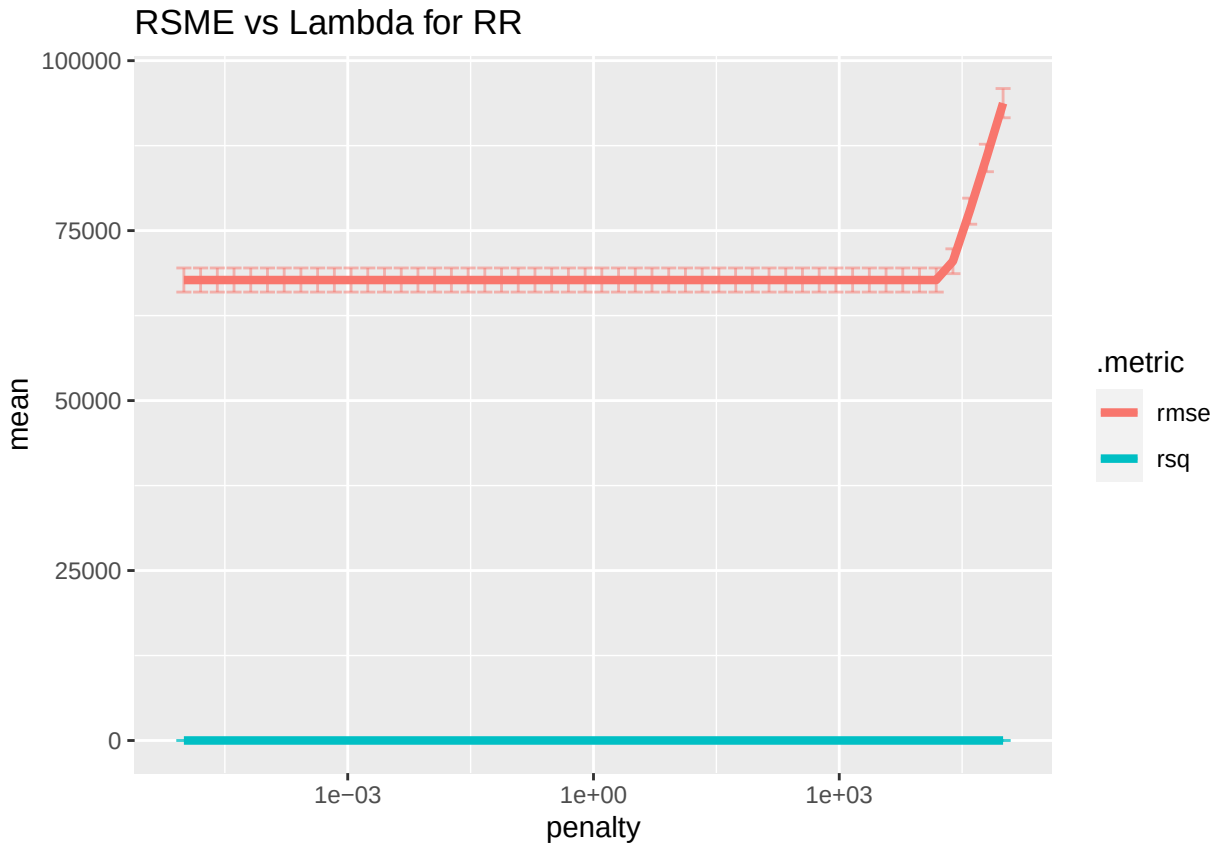
2.1 Normalizing Data

Since we are running a Ridge Regression and LASSO model on our data we need to ensure that our data is normalized to ensure that all variables contribute equally to the penalized coefficients in our models.

2.2 Ridge Regression

Ridge Regression optimization provides a trade-off between two different criteria: variance and bias.It seeks to find coefficients that minimize the SSE of the data set. Ridge Regression attempts to shink the coeffcients in our model close to zero but does not actually remove any coeffcients. This is done using a tuning parameter $\lambda > 0$. To develop a Ridge Regression model for our data we created a recipe that normalized our data and then used cross validation to find the penalty, λ value that would best minimize the SSE of our data. After using cross validatation we found that the value of λ that best minimized the SSE and our coefficients was 0.000001. The plot bellow showcases how our MSE and R^2 value change as a function of λ and our final ridge regression model can also be found bellow.

```
## # A tibble: 5 x 3
##   term                         estimate penalty
##   <chr>                           <dbl>   <dbl>
## 1 (Intercept)                   551340. 0.00001
## 2 b_gold_earned                 180281. 0.00001
## 3 b_kills                       -17192. 0.00001
## 4 b_time_c_cing_others           16638. 0.00001
## 5 b_longest_time_spent_living    15542. 0.00001
```

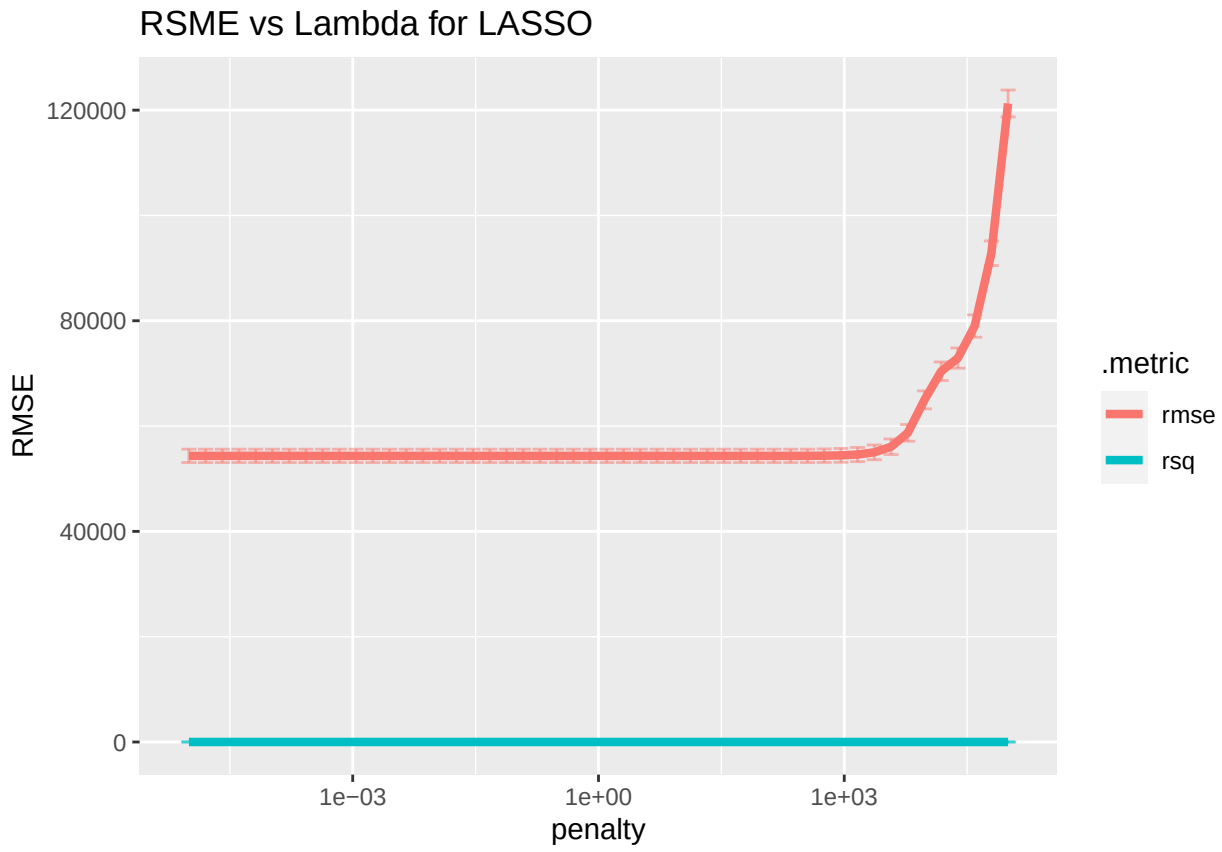


2.3 LASSO Regression

Similar to Ridge Regression, LASSO optimization provides a trade-off between two different criteria: variance and bias.It seeks to find coefficients that minimize the SSE of the data set. LASSO attempts to shink the coeffcients in our model to zero but does not keep all predictors, only important ones. This is also done using a tuning parameter $\lambda > 0$. To develop a LASSO model for our data we created a recipe that normalized

our data and then used cross validation to find the penalty, λ value that would best minimize the SSE of our data. After using cross validatation we found that the value of λ that best minimized the SSE and our coefficients was also 0.000001. The plot bellow showcases how our MSE and R^2 value change as a function of λ and our final ridge regression model can also be found below.

```
## # A tibble: 5 x 3
##   term                        estimate penalty
##   <chr>                          <dbl>   <dbl>
## 1 (Intercept)                  551340. 0.00001
## 2 b_gold_earned                269336. 0.00001
## 3 b_kills                      -79530. 0.00001
## 4 b_time_c_cing_others              0  0.00001
## 5 b_longest_time_spent_living   -659. 0.00001
```

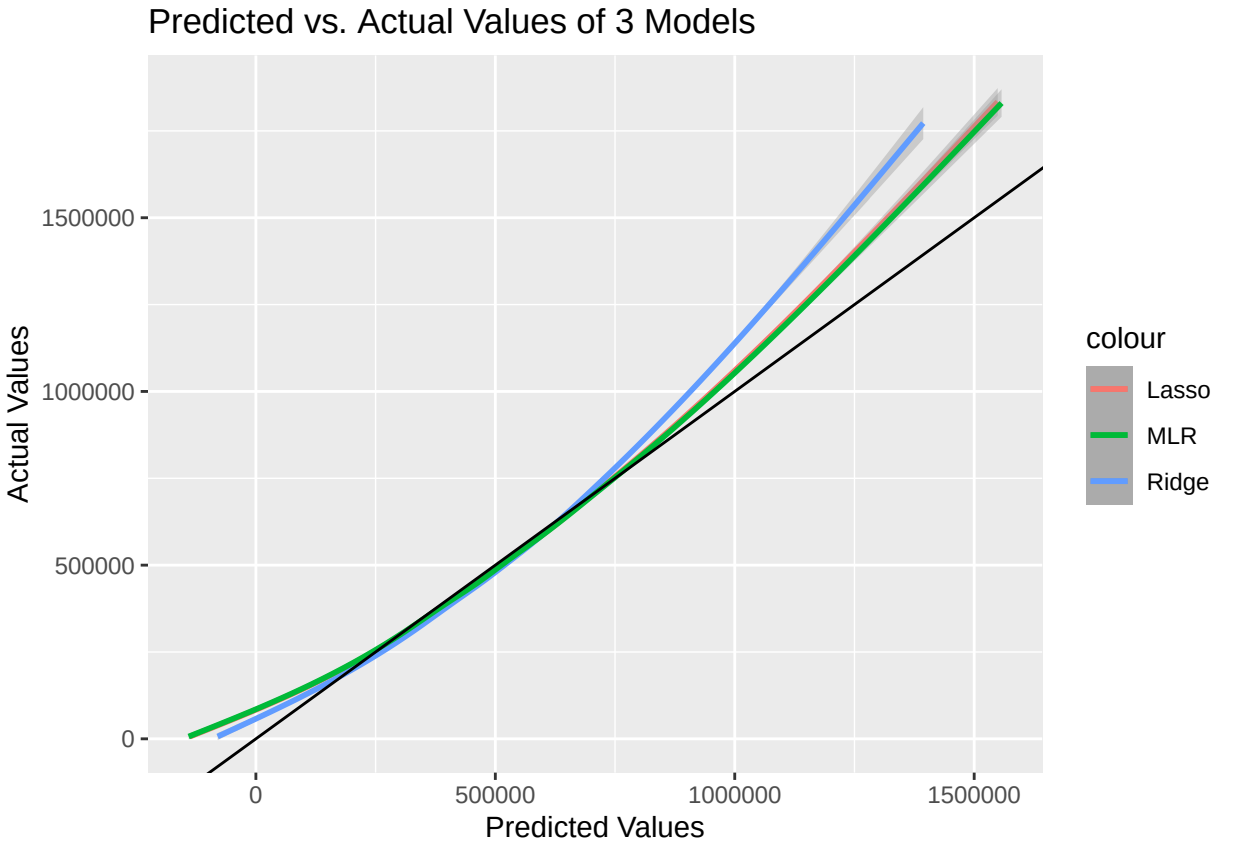


2.4 Comparing Models

When comparing MLR, RR, and LASSO we can look at the coefficients of the individual models. In the LASSO Model the coeffcient of `b_time_c_cing_others` has been penalized to 0 while in MLR it is a negative value with and in RR it is a large coefficient. In the MLR and LASSO model `b_kills` has a relatively large coefficient while it becomes more penalized in the RR model. Finally, the coefficient estimate for `b_longest_time_spent_living` in MLR and LASSO are relatively small compared to RR which is odd because of how RR and LASSO penalized the coefficient differently despite using the same λ value. The remaining variables appear to remain similar throughout the three models.

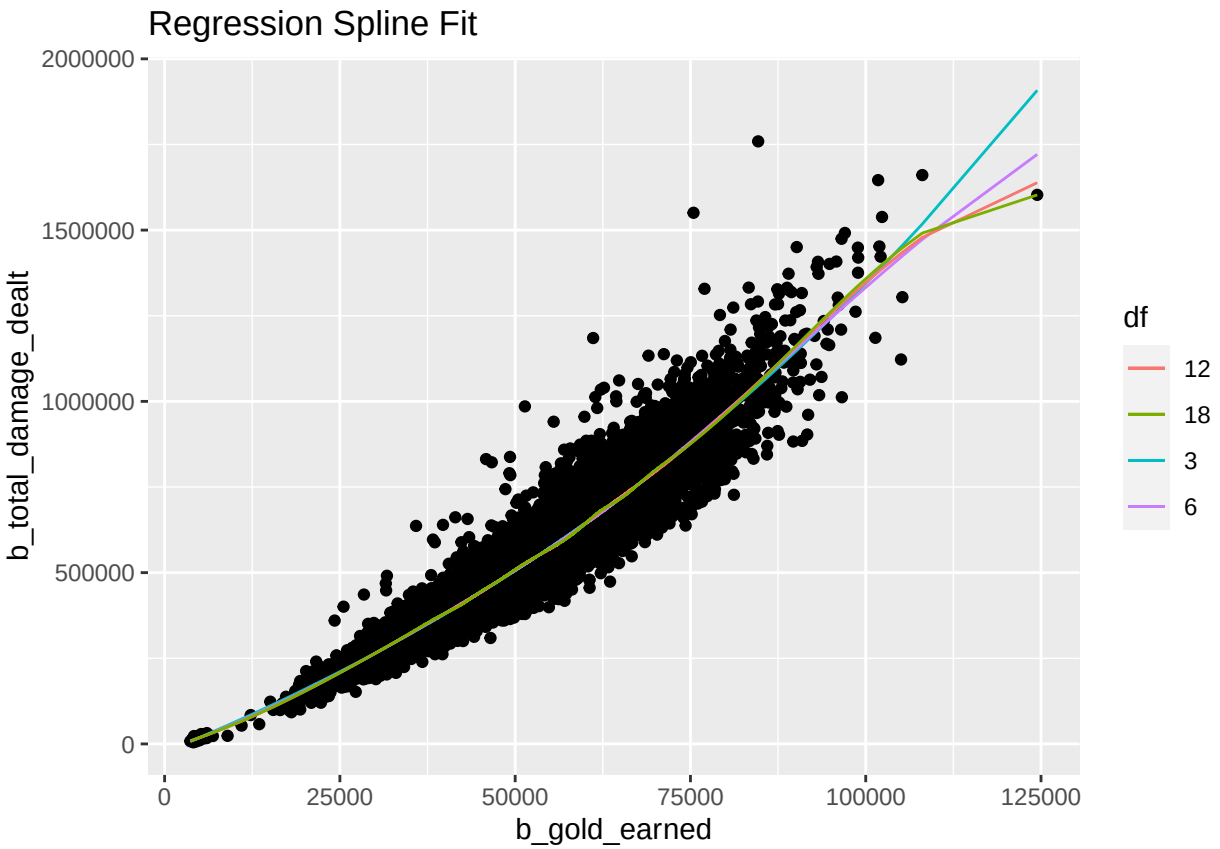
2.5 Plotting Predicted vs Actual for 3 Models

When comparing the predictions of the MLR, Ridge Regression, and LASSO models visually we can see that the MLR and LASSO model produce predictions that are closest to the actual value of the test set while the Ridge Regression model is the least accurate. However, it should be noted that this could be due to overfitting of the models, so while the models may appear to have more predictive accuracy they may not have the same accuracy on new data. Note that smooth lines were used in place of individual points for visual clarity.



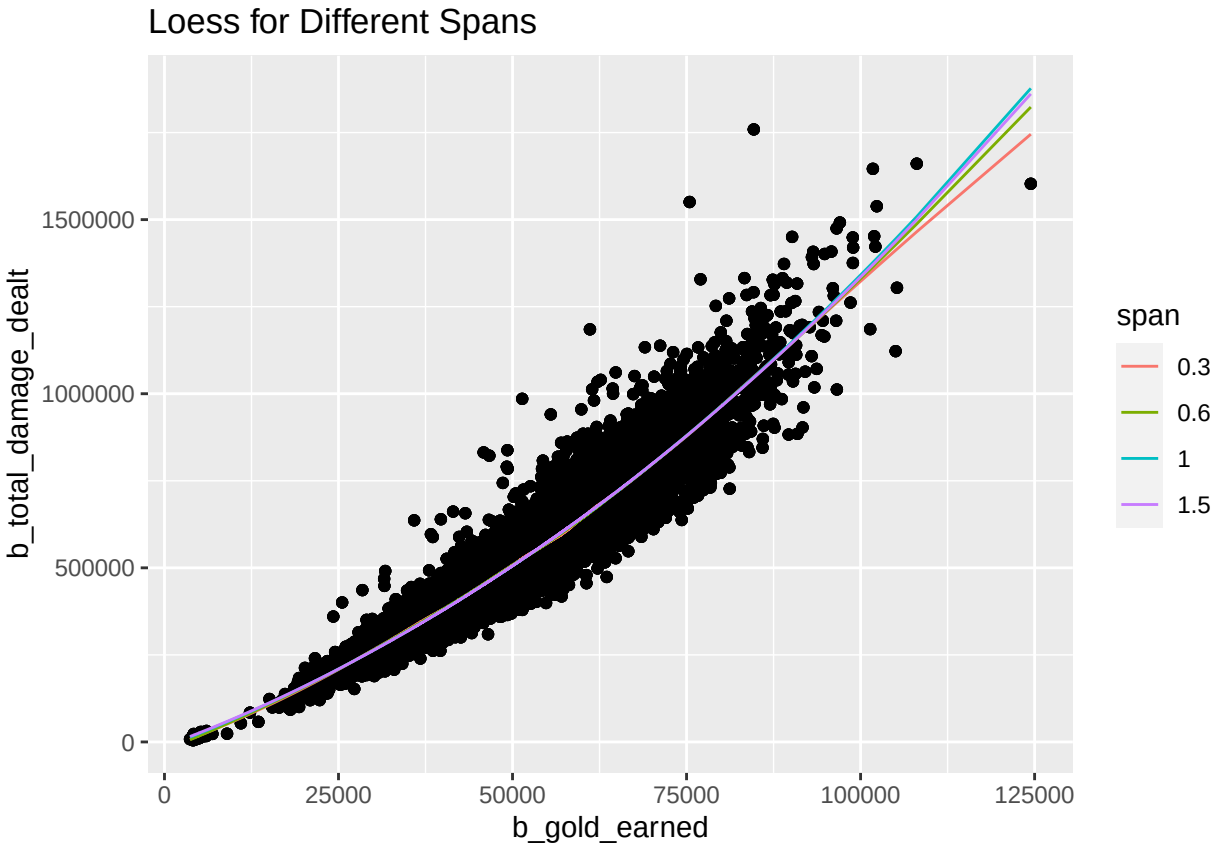
2.6 Regression Spline

In our Regression Spline Model we chose to run a model on `b_gold_earned` and `b_total_damage_dealt`. Since Regression Splines use Degrees of Freedom as the primary paramter to determine the number of knots used in the model we decided to use four different values of degrees of freedom:3,6,12,and 18. Below is a plot of the 4 Regression Spline models with data from our dataset. Visually, it appears that DF=6 produces the best model because it is not as affected by the outlier as df=12 and df=18 and it is not indiffernet to outliers as df=3 is.



2.7 Loess

In our Loess model we chose to run a model on `b_gold_earned` and `b_total_damage_dealt`. Since Loess models use span asa parameter to determine weights of points, we choose to use four span values to create four different models. Our span values are .3 , .6, 1.0 , and 1.5. Below is a plot of our four models and visually it appears that the best model has a span of .6 because the other models appear to affected by outlier points or indifferent to them.



2.8 Best Smooth Model

2.9 Conclusion

After building or MLR model, we began to look at our data through Shrinkage and Smoothing methods, as such we built and fitted our data to Ridge Regression, LASSO, Regression Spline, and Loess Models. We found that interestingly in our LASSO and Ridge Regression models our MSE did not change much with a change in our penalty value until large values of λ which interested us because it made us wonder if this was simply due to our data or if it was because we had mutated our variables. When comparing MLR, LASSO, and Ridge Regression we found that there were significant differences in the coefficients estimated by the three models, but when compared visually MLR and LASSO had very similar predictive accuracy while Ridge Regression appeared the least accurate. When looking at our Smoothing models individually, we found that our Ridge Rrgression model using degrees of freedom 6 and Loess model uwing a span of .6 to be the best models for our data. Based on our results we would ask the question: would the results of our Shrinkage models be different if we had not mutated our data but instead used our initial data, and would it have been different if we had used more data from the initial data set?

2.9.1 TO DO

~~• Introduction (briefly refresh the reader's mind as to the variables of interest). Remember that you should include a reference for the original data source, and the reader should know to what population you are inferring your results.~~

~~• Run both ridge regression and LASSO on the full variable set (use cross validation to find lambda). Compare and contrast the models (i.e., coefficients) with the final MLR model from the previous project assignment.~~

~~• Make a single plot with the observed response variable on the x-axis and the predicted response variable on the y-axis. Overlay (using color with a legend) 3 different predictions: MLR, RR, LASSO. Comment on the figure.~~

~~• Choose a single variable and run both smoothing spline and kernel smoother models. Change the parameters so that you have at least four different models for each method.~~

~~• Plot the (8+) smoothed curves on either one plot or two plots (depending on which looks better for your data. Comment on the figure(s).~~

• Without cross validating, which of the 8 smoothed models would you choose to use for future predictions? Your argument might include smoothness, interpretation of coefficients, ability to include variability of the predictions, etc.

• A Conclusion (Summarize your results. Comment on anything of interest that occurred. Were the data approximately what you expected or did some of the results surprise you? What other questions would you like to ask about the data?)

"Distribution of League of Legends (Lol) Summoners in North America as of October 2021, by Tier." 2021. *Statista*. https://www.statista.com/statistics/807298/league-of-legends-player-tier/.

Games, Riot. 2021a. "Riot Games Api." *Riot Developer Portal*. https://developer.riotgames.com/apis.

———. 2021b. *Twitter*. Twitter. https://twitter.com/riotgames/status/1455172784938651649?s=20&t=AQmQGrTa1ijf6u3cEDPZcg.

James. 2020. "League of Legends Ranked Match Data from Na." *Kaggle*. https://www.kaggle.com/jamesbting/league-of-legends-ranked-match-data-from-na.