A Comparison of Statistical Learning Methods for Regression

OLS, LASSO, bagging: A Comparison

Andrew Bates

December 20, 2018

Executive Summary

This is the executive summary.

1 Introduction

In 2001 Leo Breiman described two approaches to analyzing data with statistical models (Breiman 2001). In *data modeling* we specify a stochastic model for the data, one that has known theoretical properties, and the main purpose is usually to make inferences on the population of interest. In the other approach, what Breiman calls *algorithmic modeling*, the focus is less on the the structure of the data itself and more about the output of the model. We are not concerned about finding a model that satisfies the theoretical assumptions needed to make inferences. We are interested in whether the model can make accurate predictions on newly collected data. Data modeling is the method typically taught in statistics and is probably what most who have studied statistics think of when they think about modeling. However, there has been increased interest in algorithmic modeling recently[1] with some, including Breiman, diminishing traditional statistical modeling.

In this paper we compare Breiman's two modeling paradigms by analyzing Major League Baseball data with the goal of developing a model to predict a players salary. We examine one data model (linear regression), one algorithmic model (random forest), and one model at the intersection of the two approaches (LASSO). For each model, we discuss some advantages and disadvantages in terms of both predictive capability and interpretability. The primary aim is to construct a predictive model but, although prediction and interpretability are often seen at odds with one another (Breiman 2001, 206), in some situations one may be interested in finding a balance between the two.

2 Methods

In this analysis we use the `Hitters` data from the R package `ISLR` (James et al. 2017), a companion package to *An Introduction to Statistical Learning with Applications in R* (James et al. 2013) containing the data sets used in the book. The Hitters data contains information on 322 players from the 1986 and 1987 Major League Baseball (MLB) seasons. There are 19 covariates included in this data set that can mostly be broken down into two categories: performance metrics for the 1986 season (number of at bats, number of home runs, etc.), and performance metrics based on a given players career (career runs, career hits, etc.). There are 16 continuous variables and three

[1] See https://trends.google.com/trends/explore?date=all&geo=US&q=machine%20learning for example which shows web interest in machine learning since 2004 (accessed 12/9/2018).

categorical variables. For the 1987 season we have the player's salary on opening day along with their league (American or National) at the beginning of the season.

The salary variable has 59 missing observations, 18% of the data. This was too many observations to ignore so we imputed the values using k-nearest neighboors before proceeding with the analysis. After examining histograms of the continuous variables, it was evident that transformations were in order. Salary, along with several covariates, were heavily right-skewed. We chose log transformations for these variables because it is a common technique and allows us to readily interpret linear regression coefficients. In all, 11 of the 20 variables were log transformed. All numeric variables were then subsequently centered about the mean and scaled by the standard deviation.

Prior to model fitting the data was split into a training and testing set with 20% reserved for testing. All three models were trained using 5-fold cross-validation with the final model chosen to be the one with the lowest root mean squared error (RMSE). In each cross-validation run for linear regression a stepwise procedure was used with Akaike Information Criteria (AIC) as the model selection criterion. Diagnostics were run on the model chosen via cross-validation and some covariates were subsequently ommited based on variance inflation factors and correlations between the covariates. Grid search was used for hyperparameter tuning of the lasso and random forest with 10 values in each grid. For the lasso the hyperparameter is the lasso penalty and for random forest the hyperparameter is the number of randomly selected covariates considered at each split. A comparison of predictive ability for the three models was made through RMSE on the testing set. We investigate interpretability via coefficient interpretation for linear regression and lasso and variable importance for lasso and random forest.

The analysis was conducted using the statistical software R (R Core Team 2018). This document was written using the R packages `R Markdown` (Allaire et al. 2018), `knitr` (Xie 2018b), and `bookdown` (Xie 2018a). All materials used to conduct the analysis and compose the report can be found at `https://github.com/asbates/stat696/tree/master/reports/baseball`.

3 Analysis

3.1 Exploratory Analysis

first talk about missing value imputation. explain what method and why. also explain why even impute. if the data set was bigger, might consider just removing them but in this case i think we need to impute them.

3.1.1 Feature Engineering

look at plots of all the variables. for skewed ones (salary for sure), do transformations. log would be simplest but also consider box cox and look into other transformation methods. should we look at correlations? we don't really need to be concerned with collinearity here do we? again, don't need to go into a ton of detail here but highlight a few variables, especially salary and any extreme cases. put a table in the appendix listing out all the final features and mention it here.

3.2 Modeling Fitting

explain model fitting method. for each method, k-fold cross validation was performed with the final model being the one that minimized the mean cross validation error. We used rmsle as a performance metric because it penalizes large and small errors more evenly compared to rmse. don't forget to discuss hyperparameter tuning. for liner regression, the only real hyperparameter is the number of variables. within each cross validation run, maybe we can do some sort of stepwise regression variable selection. how would this work in code? can this be done with recipes? can this be done with mlr? will i have to implement a custom method via MASS? for the others, does recipes handle random search or only grid search? can it even do hyperparameter tuning? (i assume yes). maybe list out the the range of hyperparameters considered for random forest. with lasso, would it be ok to extend this to elastic net or does it have to be strictly lasso?

3.3 Prediction Results

4 Conclusion

limitations: the most obvious is the timeline. the model constructed here should clearly (?) not be used to estimate player salaries for the next season (2019). besides the obvious issue of inflation, player salaries have experienced an inflation rate beyond that of the overall inflation rate (reference). Additionaly, elite level athletes have generally seen a tremendous growth in performance (skills, ability) over the last 32 years (reference). The game itself has also experienced an evolution (changes) that may have an impact on model performance. From rule changes, increases in overall althletic performance, and changes in player population, among others, the game of baseball has surely changed since 1986. That is not to say that the model developed here is useless however. We just want to caution the reader that one should not expect exceptional performance if this model is applied to a current MLB season (although that might be an interesting experiment). The usefulness in this analysis comes from the overall model construction procedure. This gives future analysts a starting point to work with.

Another limiting factor of this analysis is in the available variables in the data set. Baseball statistics has improved vastly over the last 32 years (reference) and there are substantially more metrics available currently (reference). This will no doubt improve the likelihood of providing the model with the 'right' features; features that could have a major impact on model performance. Often times predictive modeling is more about coming up with (obtaining) appropriate features than a particular model (reference). another thing that might help is player position and stats associated with various positions (i'm thinking pitchers here)

If future analysts deisire to construct a model for estimating contemporary MLB player salaries, we recommend the following approach.

A Supplementary Material

B R Code

References

Allaire, JJ, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, and Winston Chang. 2018. *Rmarkdown: Dynamic Documents for R.* https://CRAN.R-project.org/package=rmarkdown.

Breiman, Leo. 2001. "Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author)." *Statist. Sci.* 16 (3). The Institute of Mathematical Statistics: 199–231. doi:10.1214/ss/1009213726.

James, Gareth, Daniela Witten, Trevor Hastie, and Rob Tibshirani. 2013. *An Introduction to Statistical Learning Wiht Applications in R.* Springer.

———. 2017. *ISLR: Data for an Introduction to Statistical Learning with Applications in R.* https://CRAN.R-project.org/package=ISLR.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Xie, Yihui. 2018a. *Bookdown: Authoring Books and Technical Documents with R Markdown.* https://CRAN.R-project.org/package=bookdown.

———. 2018b. *Knitr: A General-Purpose Package for Dynamic Report Generation in R.* https://CRAN.R-project.org/package=knitr.