The Lahman Database

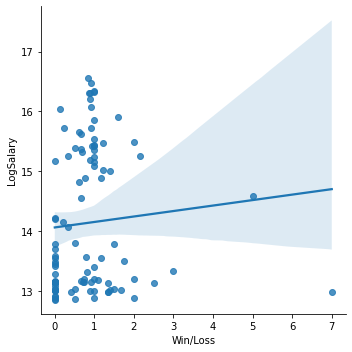
IST5220 Group 7

Joshua Jager, Maxwell Butler, Austin Funcheon, Conor Joyce, Casey Kemerer, David Poteet

The Lahman Database for MLB database is a large source of background data for information on many different aspects for the different statistics associated with baseball. These performance measures drive many aspects of baseball- from the front office management of the team, to contracts, to fantasy baseball for the everyday fan. Analysis of this sort of data has million, if not billion dollar impact to the industry, so it makes it a very varied and interesting set to interact with.

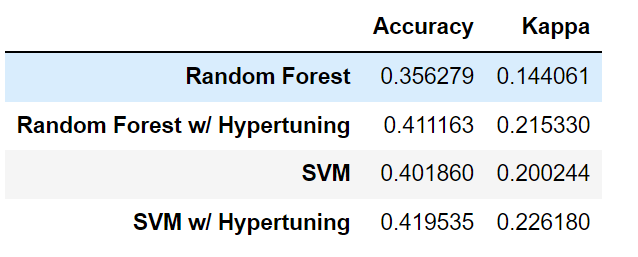
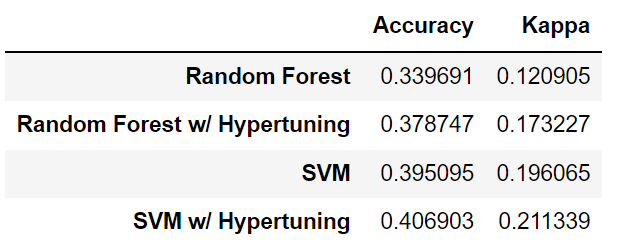
The crucial first step was trying to understand the datasets; while an over abundant amount of data might lead to a better conclusion on the back end, it definitely made the beginning of the effort a bit overwhelming. As a group, I think we understood how complex data could be, but before digging into the dataset, it was perhaps noteworthy how raw some of the dataset was. A fairly significant amount of time for several team members went into understanding the datasets, which were (likely) relevant, and sorting through the data to import and clean the data to provide a meaningful dataset. From there, logically sorting thorough the data with several methods such as PCA to reduce the redundant / colinear measures. This lead to some reprocessing and transformation of several data sets- for example, creating a W/L ratio, since they generally total to the same number of games, so having both measures doesn’t add much relevance to the data.

Early visualization of the salary information made it very clear there was a heavy time dependence present in the data as well; salaries have continued to drift up over the years, thus requiring other transformations to make sure data was normalized to account for inflationary measures over time. Additionally, there was a large degree of skew, so salary information was transformed log(y) to better linearize the dataset.

For the data analysis portion of the program, we evaluated several methods to assess the data, for both the pitching data and the hitting data: Random Forest and SVM, each with hypertuning. Visualization of the data revealed a couple of interesting details for pitching. For W/L ratio, there was a visual delineation of two clusters within the data, that is otherwise unexplained. This makes us wonder if there is an additional hidden variable that turns out to be quite meaningful in explaining this binomial split into two seemingly distinct, yet unexplained populations.

Odd Bimodal distribution with W/L

Reviewing each of the regressions, it appears there is a pretty good convergence among the methods- both RF and SVM for the pitching data set came out to be about a 40% accuracy / .2 kappa score. This implies our model is not an abundantly good predictor of salary given the data available in the data set- only mildly better than a coinflip. This does not bode well for this set of statistics being good enough to have a meaningful predictor for pitching salary; there appears to be uncaptured variables that significantly contribute to the outcome of pitcher salary. Likewise, we saw similar performance of the hitting data, with a ~40% accuracy score and ~.2 Kappa score. This likewise implies there is a very weak correlation with our model to hitter salary being predicted by the dataset we have available.



Hitting Model Results Pitching Model Results.

Overall, while the output model was not exactly a groundbreaking model to make us millionaires, it was a very complex dataset that allowed us to apply just about every aspect of this course: bringing the data into python, visualizing the data in a meaningful way, cleansing the data, reducing the number of inputs to reduce noise, testing out several machine learning algorithms methodologies, and being able to evaluate the success of our model with real world data. It challenged the team in many ways and served as a great learning device that shall serve us well as we move through the real world.