**Josh, Laura, & Connor**

**Deliverable #2**

**6/22/2021**

Our next goal is going to be using RandomForest where we can test the outcome of a game between 2 teams. We will run 3 models using TEAM ELO-related variables and another 3 using PITCHER-related variables. These models will take a different combination of each related variable to determine the best combination. We will then compare these outcomes to the real outcomes of these past games and determine the most accurate model. Tuning and further adjustments will follow in order to get the best model possible. We are aiming for 80% accuracy after tuning our models. In these 6 RandomForest models we started using a train set of 10%. All the initial parameters of these models were the same to prevent bias. “Home Win” will be our y-variable that we are attempting to predict.

Starting with the PITCHER-related data, three models were made containing 1, 2, and 4 variables, respectively. Model “x” represents the first model run using the lone variable, “Starting PItcher Win %”. The result was, as expected, extremely low with an r-squared of 0.0198. Model “x2” contained 2 variables, “'Starting Pitcher Win %” and “Starting Pitcher Rating”, which ran at an equally unimpressive r-squared of 0.0247. We started to believe that pitcher-related variables, by themselves, are not statistically significant for our predictions. Model “x3” combined 4 different variables, a model we hoped would give us decent results. However, the variables, “Starting Pitcher Win %”, “Starting Pitcher Rating”, 'Home %”, “Starting Pitcher ADJ”, only yielded a r-squared of 0.0252. This cemented our previous suspicions concerning pitcher variables.

ELO-related data initially did not yield a better result than previous variables. Almost following the same path of 3 models containing 1, 2, 3 variables, respectively, the last model will only have 3 variables. “Home Post Elo” is the lone variable used in the first model, Model “z”. To our disbelief, the r-square yielded 0.02736, not any better than the pitcher variable models. The next model, “Model z2”, contained the variables, “Home Post Elo”, ”Home Pre Elo” and yielded a better r-squared of 0.1456. This proved to us that ELO-related variables are already more significant by themselves. The 3-variable model including, “Home Post Elo”, “Home Pre Elo”, “Home %” yielded our best r-squared of 0.2422. Now, we found our best model, but we decided to try one last model where we combined all the variables from each 3-variable model. We decided to focus on tuning the 3-variable ELO model first, since it is the most promising.

Our first decision for tuning the model was to run a RandomizedSearchCV to figure out the most favorable parameters of the random forest. The results, however, ended up yielding a worse r-squared than before. We then decided to tune the parameters ourselves in order to yield the best results. We started with tuning the max\_depth of the model. We found that the larger the depth, the better the yield. However, this is clearly overfitting and can not be an accurate representation of our true best model. So, we started with a single decision tree (n\_estimators) and worked up to what we felt was not over-fit. We ended on n\_estimators of 11 with a max\_depth of 6 resulting in an r-squared of 0.3113.

We followed this method again with the Elo plus Pitcher Win % and rating model just to provide a mode of comparison and check if the two subjects can be more significant in tandem. The setup was roughly the same, and we committed to a max\_depth of 6 once again. The optimal number of trees for this model was 29. While we thought the Pitcher information was a completely bogus predictor, this combo model surprised us with a very comparable R-squared, ranging from .3074 to 0.3149. (We discovered that each running of a Random Forest--even holding all parameters and inputs constant--produces different MSE and R^2 values).

With our two best models having very similar performance, a winner must still be chosen. Given that the Elo + Pitcher model has only a slightly larger success rate and is not consistently more accurate, we have to go with the Elo 3-variable Random Forest model as our best bet. This model is more economical in the sense that it produces a nearly identical result with fewer resources. This lends to a more time- and space-efficient model, which is desirable for massive datasets.

With these being the best models, it is clear that our models are not prolific for predicting the outcome of MLB games. Another issue could be that this data set does not offer enough variables that are statistically significant enough to make precise predictions. With the current form of the dataset, there is not much more tuning of the hyperparameters that could be done. Other predictors could be brought in from a larger source, but this game-by-game prediction yields a decent amount of significance to start with.