ISDS 7070 – JAMES DAVIS

Deliverable #1

We are using machine learning to find the best model to predict outcome of MLB games. We started with a rough dataset where no data was missing and wrangled our data using several grouping and aggregate functions. Data was relatively clean when we got it. We have several variables to compare in different models and use to find the most accurate combination. These include the average ELO number per team per season (), team win probability () and starting pitcher win probability ().

Our next goal is going to be using RandomForest where we can test the outcome of a game between 2 teams, we will run one model using TEAM ELO and another using WIN %. We will then compare these outcomes to the real outcomes of these past games and determine the most accurate model. We are aiming for 80% accuracy. In these 2 RandomForest models we will be using a test set of a random 10%.

We have made assumptions that a team will be consistent throughout a season, that playoff data will produce outliers and that the data gathered by FiveThirtyThree is accurate. The most obvious issues that we expect to run into in this problem will be inconsistent performances from teams (playing well half of a season and poorly the other half) and inconsistent performances by pitchers. The pitcher’s performance will be our largest problem because they are an individual, not a group like the team, and this will make them much less consistent.