## Michael Wegan G45192209

## Zheying Chen G43061235

## Weiwei Li G27494462

## Xueli Zhou G29955935

After we built the ensemble model to predict Most Valuable Player (MVP) status in Major League Baseball in Milestone 5, we realized that our model results were unstable. This was because the results were very sensitive to the data partitioning process we implemented to break the data into testing and training sets. In order to understand the variance caused by partitioning the data differently and develop stable predictions for each player, we implemented a Monte Carlo Cross Validation. This partitioned the data 100 times and re-ran the entire modeling process. We averaged the predictions across each of those 100 iterations to generate more robust final predictions with strong results. Our modeling process predicts the MVP correctly in 53% of opportunities. The MVP is within the top six predictions 86% of the time.

The Monte Carlo Cross Validation process incorporated variance caused by different data partitions but created more stable predictions for each player. We have included at the end of our analysis a plot that will show how the estimates for 2 players change with the incorporation of additional runs of the cross validation. The early runs lead to very sudden changes in the mean prediction but the rate of change slows as more iterations are incorporated, thus stabilizing the results. The area under the ROC curve varied from .59 to .98 in the individual runs, but when the results were averaged across all 100 iterations, it climbed to .989. We would have preferred to run the cross validation 1000 times but we lacked the computational power to perform that many iterations. Results would have further stabilized and may have been better with that number of iterations.

Overall, the exercise of attempting to predict the MVP was valuable. We learned much through the implementation of models that we have only learned about in class. When faced with unexpected issues such as the variance in results caused by different test and training data partitions, we had to figure out why we were getting different results than what we expected and decide how to move forward. The technical understanding of the modeling process is important, but the actual implementation can only be learned by working through a project like this. The fact that we were able to get solid results was a bonus. Additionally, we learned the importance and time-intensive nature of data cleaning. Much of our effort and time was invested in the cleaning process, but this paid-off when we were developing models. The data were well organized and easy to work with because we had cleaned the data well.

Moving forward with this project, we think that adding additional features may add predictive power. By running this model several times during the baseball season, we believe we can generate realistic predictions for the MVP that may financially aid sports media or oddsmakers. We hope to refine this model so that we may monetize it.