### Sean Hoffman - Final Project – Proposal

# Machine Learning vs Las Vegas

## Analyzing College Basketball Results Past to Model the Future

**Outline:** For my final project, I plan on training a model to use historical college basketball game statistics and results compared to the betting odds, lines, and point spreads. I hypothesize that these results can be used to generate projections for upcoming games that would allow us to make decisions in placing lower-risk wagers and boost our returns. I plan on running tests of the model against game results from the 2019-2020 basketball season to show if the model can perform successfully. I’m hoping to be able to output information for the upcoming day’s schedule of games that provides a confidence score for that day’s betting odds, lines, and point spreads to guide a wagering strategy that could result in consistent gains over time.

**Technologies:** This project will use Python Pandas, Python Numpy, Python Matplotlib, Scikit-Learn, and Keras at a minimum to be able to execute the analysis.

**Data Sets:** The primary data sets I’ll be using will be captured by leveraging the sportsreference API for Python by Robert Clark (<https://github.com/roclark/sportsreference>). This API pulls statistics from [www.sports-reference.com](http://www.sports-reference.com) which will allow my code to capture the necessary game and player data for analysis. Additionally, I’ll be using data sets available at <https://www.sportsbookreviewsonline.com/scoresoddsarchives/ncaabasketball/ncaabasketballoddsarchives.htm> to provide the sports book betting odds, lines, and point spread data for past seasons which allows us to see how the model is stacking up.

**Potential Issues:** The number one foreseeable issue with this project is the possibility of overfitting due to the overwhelming number of tracked statistics in a college basketball game. I will need to avoid overfitting so that the model can generalize to future games. One method I will attempt to incorporate is by calculating the importance of the feature set using LOFO Importance (<https://github.com/aerdem4/lofo-importance>) to iteratively remove each feature from the set and evaluate that feature’s importance to the performance of the model against model performance with all features included. I’ll also attempt to measure some of these linear relationships using Pearson’s and Spearman’s correlation coefficients to support removal of some features.

The other major concern is with data completeness. There is a possibility of box scores for games missing data in some situations, and for there to be no sports book information for certain games. There will need to be some cleaning performed in order to remove some of these outliers that could skew the model.