MIST 5635 Group Project Deliverable 2 (500-1000 words)

By: Ryan Cullen, Alec Zerona, Warren Paintsil, and Will Hensley

1) Identify the type of problem we are solving.

The problem is the United States men's soccer team performs poorly when competing at a high level. Here are a few statistics on the U.S. men's soccer team, providing context to why we care about exploring the most statistically significant indicators of a team's success in the tournament.

a. The team has made 12 World Cup appearances:

i. The team has a record of 5-6-2 (Wins-Losses-Draws).

ii. The best result in general for the team was reaching the 1930 semi-finals.

iii. The best relevant result for the team was reaching the 2022 quarterfinal.

b. 2024 U.S. Men's Soccer Team Statistics shows the team currently has a losing record and negative goal differential (updated November 15, 2024)

We hope to potentially improve the future of the U.S. men's soccer team with data analytics by statistically proving what the team should focus on. Example future analysis includes exploring offensive strategies such as possession and total passes, defensive strategies such as forced turnovers and pressure applied.

We will perform exploratory analysis with a variety of statistical methods outlined next to determine the statistically significant indicators of the teams' success.

2) List potential methods for solving the problem.

a. Logistic regression

b. Bagging/random forest

c. Boosting/xgboosting

3) Describe the steps we plan to take to ensure the conclusions are as accurate and robust as possible. Note that a viable analysis plan should detail at least three viable approaches to solving the problem, as well as specifics about training/testing procedures.

Procedures for All 3 Approaches/Models Described After This

For all three viable approaches below to solve the problem defined above, we have specific procedures explained here that will apply to the training and testing datasets. We will utilize the caret package in R, specifically the create data partition method to split the dataset into training (75%) and testing (25%) data as well as perform K-fold cross-validation and parameter tuning for reducing risk of overfitting the dataset. We plan on calculating accuracy such as the misclassification rate for logistic regression.

a. Logistic regression

Looking at the data, we can create a new feature that represents an indicator for teams that won in the FIFA World Cup (team1_win = 1 means “win”, team1_win = 0 means “loss”). This allows us to build a logistic regression model with the available features trying to predict what features are significant in winning and what features can help us predict teams that win. We can narrow down our features to make the model perform at a 95% confidence level and draw predictions from there. However, there are a couple regularization methods that we can implement for our

feature selection to help prevent overfitting. Using LASSO, we can create a model that penalizes unimportant features by devaluing them to zero. Given that the dataset is initially wide, this might be the best way to go about conducting our analysis. Using RIDGE, we can preserve features that might still impact our model prediction, but devalue their weights. By comparing these models, we can test how each method performs on the training and test error, but ultimately improve our model by narrowing down to the most important features.

b. Random forest

Random forests will be another viable approach to determine the statistically significant indicators of the teams' success. Random forests incorporate extra randomness during the process of growing the trees. Random forest should improve the decision tree by decorrelating the numerous trees to reduce the variance when we average the trees. We will use bootstrap sampling (with replacement) to draw new datasets. The number of features/columns/predictors considered at each split will probably be close to the square root of said features/columns/predictors. The downside is trading interpretability for accuracy, but with the variable importance, the downside is not bad.

c. Boosting

We can utilize boosting by creating multiple copies of the original training data set. We will generate each tree from previous trees and identify the mistakes in the previous models. Weak learners are brought together and create one strong learner. The weak learners are likely to be over-fitting and their accuracy is not always there. Since we do not have a large amount of observations, we will try to utilize a lambda to incorporate a slow learning model to gain

thorough insights of the information. Extreme gradient is the type of boosting that will be used. It will give us the biggest and best return that we are looking for.