Joshua Harkness

Assignment 04

STA6704

Prof. Smith

<https://github.com/darkhark/2019JAX_PlayByPlay/tree/master/Assignment03>

1. Done through code.
2. I chose to use AIC for my scoring measurement. In all reality though, my scores all ended up in the same order each time anyway, with the scoring model being the best in each scenario due to it having the closest value to zero. The reason I chose AIC over BIC is BIC will give a higher penalization for more complex models. With this data, many easily understand correlations exist between the data and I want to find the relationships that may be more complex. One example of this is how yards\_gained and yards\_after catch are made into a relationship. It seems obvious that bigger plays in the NFL likely also have more yards after the initial catch unless the quarterback is able to consistently make deep touchdown passes where the receiver is already in the endzone. We’re talking about the Jaguars here though, so that is unlikely the case. One of the less obvious and likely weaker relationships I noticed was between pass\_middle and air\_yards. There are few possibilities for this that I will cover during my project.

A close up of a logo

Description automatically generated

interval

hc -6940.355

h2pc -7186.785

iamb.fdr -7280.212

aracne -8524.373

1. Training model on training data – These are on predicting if a first down was made.

Reference

Prediction 0 1

0 187 76

1 52 70

Accuracy : 0.6675

95% CI : (0.618, 0.7144)

No Information Rate : 0.6208

P-Value [Acc > NIR] : 0.03219

Training Model on Test Data

Reference

Prediction 0 1

0 66 26

1 16 21

Accuracy : 0.6744

95% CI : (0.5864, 0.7543)

No Information Rate : 0.6357

P-Value [Acc > NIR] : 0.2060

I do not believe this model would be useable by an NFL coach. Most of the variables used to predict first down are not new information to the coach (more yards gained therefore more likely it was a first down). More pruning will need to be done in an attempt to find information that is not obvious. The model may be able to work for training people new to the sport what to look for in order to understand how the sport works, but honestly it’s easier just to watch the game and learn that way. The relationships that I was hoping to be meaningful did not prove to be so. I was hoping pass location, time of the game, and depth would have a stronger relationship to the result of the play and if it ended up breaking big. With that said, there are a few improvements I will make and include in my project.

I did not include the quality variable this time just out of learning how this modeling works. I think what I may do in the future is use the Quality variable I have to remove other variables that don’t really help with predicting a play’s outcome. When I ran the dimension reduction without the quality variable though, it did not bode well for determining the quality. Many factors have weak relationships though so I believe removing more of those will help. I also believe it may be beneficial to not remove some dummy columns like pass\_location\_left because when I did remove that, it creates too strong of a relationship between the two remaining pass locations that they may not look elsewhere for a more hidden relationship.