**Zach Feldman - OSU SSI 2020**

**Abstract - NFL Expected Completion Research**

In recent years, completion percentage over expected has proved to be an effective statistic to evaluate quarterback performance. There has also been an explosion in public and reproducible research surrounding the NFL thanks to packages like nflscrapR and events like the Big Data Bowl. This research examines different modeling techniques, using publicly available data and reproducible methods, for predicting completion probability in the NFL. Using regular season play-by-play data from 2009 to 2019, I attempt to find the best completion probability model, weighing both model performance and quarterback evaluation performance. I utilize generalized additive (GAM) and extreme gradient boosted (XGBoost) models trained on regular season data from 2009 to 2018. The data was split into train and test sets of 70 percent and 30 percent of the data, respectively. The XGBoost model also utilized 5-fold cross validation for hyper-parameter tuning. 2019 regular season data was used as another method of out of sample testing. While the XGBoost model performed slightly better in all phases, the GAM provided easier interpretation of variables. The distance of the pass, air yards, was far and away the most important variable, which was as expected. Other significant variables include: yardline the play started from, if the quarterback was hit on the play, if the pass was over the middle, and if the pass was tipped. We also see the season is an important variable to include in the models as the game of football has changed over the last decade. With publicly available data and reproducible methods, we can determine factors that influence completion percentage to further understand quarterback play in the NFL.