Capstone Project Abstract Proposal

Author: Christopher Rowe

The National Football League started collecting data on the field in 2015 by placing computer chips in the equipment of all the players. This results in accumulated data about each player on the field. Recently, the NFL uploaded data sets onto Kaggle for analysis, including a collection of all running plays from Week 1 of 2017 to Week 12 of 2019, resulting in over 33000 plays, and 45 attributes. This project will be looking at the yards gained from a running play, subtract the distance needed to collect a first down, and determine whether a first down was successful or not. A regression analysis will be performed to see which factors are most important to determine if a running play will get the first down. As well, predictive models will be created to see if we can enter the details of a play and see if we can predict if a first down will be gained or not if we ran. A logistical regression model will be created with all the attributes of the data, and K-Nearest Neighbor model will be created alongside Random Forest and Decision Tree, and compare the accuracies of each to determine which is better to use, using Python and R. What is hoped to gain is whether we have sufficient information in the collected data to determine if teams can make a first down with the running plays they have in their current playbooks, and which factors can improve to for more successful running plays.

Dataset link: <https://www.kaggle.com/c/nfl-big-data-bowl-2020/data>