Noah Caldwell-Gatsos
Vamsi Banda
Michael Cantu
Eric Yang

ML Final Project
Dr. Downey
11.06.2018

Northwestern University's football team currently has little work done in applying statistic methods to achieve predictive analytics of future games' outcome – for our final project in EECS348, we would like to assist their efforts by providing a method of achieving helpful predictions of future game results that they might be able to apply in their planning, coaching, and training. Through this assistance, we anticipate that our group will be able to provide results that Northwestern's own analytics teams can use to distribute resources more effectively and alter their strategies in tandem with our results.

The data used in this project will be provided by Northwestern's football analytics group, on recommendation from our professor – which we will receive after we have passed the necessary background checks needed to view this information. However, we have already received a data dictionary to assist in analyzing potential attributes we will be working with to help their decisions.

After analyzing the given dictionary and relevant features, we've compiled a list of terms that we believe are key in assisting our further research:

1. "Week" attribute which contains the week of the matches taking place
    a. These also possess a classifier that determines the nature of the game itself, i.e. Preseason Game, Wild Card (WC), Divisional Playoffs, Conference Championships, or Super Bowl (for the NFL), and for the NCAA, Conference Championships, Bowl Games, Playoffs, Championships, and All-Star Games.
2. Other attributes we wish to examine further are
    a. Gameseasons – the NFL Season Year
    b. Offsuccess (Offensive Play Successes) – shows all offensive plays that were a success (or had a neutral or failing effect) in the database
    c. List of Touchdowns
    d. Time to Throw – timing from when the ball is snapped to the conclusion of the Quarterback's participation in the play.
3. An additional list of maneuvers is also under consideration, with terms such as
    a. Pump Fake – a move by the quarterback to deceive the defense
    b. Stunts – a defensive stunt on a pass play or punt rush, which is a planned maneuver by a pair of players of the defensive team where they exchange roles to better slip past the offensive team's blockers at the beginning of the play.
    c. Shotgun – a type of formation the offensive team can take, used mainly for passing plays. In a shotgun formation, the quarterback receives the snap from the center at the line of scrimmage.

There are many additional attributes within the data dictionary that we would like to examine, but these were the main options from a preliminary examination of the potential data.

Because of the structure that the data will be provided in, we can postulate which methods we will apply to analyze said data. Pre-processing obviously includes data cleaning –

namely through data integration by removing anomalies (such as duplicates) and normalizing the data. We would eventually like to do explorative data analysis to find the significance of each of the features we listed above. We are planning on using a classification model such as decision trees, k-nearest neighbors, SVM, and logistic regression to determine winning game outcomes and help offer predictions on future games. We would like to run the data on all types of machine learning classification models by using GridSearchCV's API. Our main task will involve finding the best hyper-parameters and quality metrics to apply – such as precision, recall, and accuracy, to the test set.