Predicting Quarterback Success in the National Football League

Project Proposal

Group 97

Joshua Mayanja

Master of Science in Analytics, Georgia Institute of Technology

ISYE 64740: Computational Analysis

Professor Xie

Summer 2024

Problem Statement

The National Football League is a multi-billion dollar enterprise, with millions of fans all over the world. Every Sunday from September to January, these fans watch their teams, some cheering on a Super Bowl contender, and others suffering through a losing season. For many in the latter, looking forward to the annual NFL draft in April is much more exciting than the current games, as their team will likely have a high pick, and the chance to select a franchise changing player. Oftentimes, the players selected at the top of the draft are Quarterbacks, as they are the leader of a team's offense.

Finding a quarterback that can become a franchise changing player has actually proven to be a very challenging task. Since 1970, over 500 quarterbacks have been selected, and the probability that one becomes a First Team All-Pro selection is just 4.6%. This probability rises to 9.4% for quarterbacks in the first round, and 11.9% for quarterbacks drafted in the Top 5 (Fox Sports). Given the fact that quarterback is such an important position, it is vital for a team's long term success to draft a high end quarterback when given the opportunity. The current evaluation methods in place for draft eligible quarterbacks are not very strong, and have missed on many quarterbacks every year. These systems are mostly visual, with teams sending scouts to watch each player. Most prospect evaluations are based mainly off of this visual evaluation, and then potentially an additional interview if the team is interested.

This project aims to enhance the current quarterback evaluation methods through the use of machine learning. Machine learning has the ability to reveal hidden trends in data, and a strong model would be able to reveal which quarterbacks are NFL ready, and which simply exposed weak points in the college game that disappear in the NFL. This model would provide that much needed extra insight, enabling teams to select a franchise quarterback, and save them from the minimum three years of misery that comes with selecting a bad one.

Data

The data for this project will be scraped from the *Sports Reference* websites. 20 years of NFL draft history (from 2001-2021) will be scraped from *Sports Reference*'s *Pro Football Reference* website. From this list, each quarterback's collegiate statistics will be scraped from their page on the *SRCFB* website. Data on the school that each quarterback attended (record, rank, and conference) will also be scraped from *SRCFB.*

Methodology

This project will utilize multiple machine learning methodologies. In the EDA phase, all of the quarterbacks that have been drafted in this 20 year window will be clustered to classify them into tiers. These tiers will range from elite to practice squad, and is based on their NFL success. Once classified, these quarterbacks will be split into a training and a test set. From there, an XGBoost model, a Random Forest model, and a Neural Network will be fit to create a model to classify each quarterback into the correct tier. The feature vectors generated for each quarterback here will also be used to generate quarterback comparisons for prospects in the upcoming drafts (2023, 2024, and 2025) through the use of cosine similarity.

Methods for dealing with missing or incomplete data points will be evaluated as EDA is performed. At the moment, there appears to be very few incomplete data points, so it would make most sense to remove these data points. However, if it is discovered that there is a much larger number of incomplete data points, then imputation will be explored.

Evaluation

Evaluation of this model will take place at two different points in time. In the short term, evaluation of the model will be based on its performance against the test set of data generated. A high accuracy score and a high R-squared value will be the indicators of success.

In the long term, accuracy and R-squared will once again be measured, once each quarterback drafted in the 2023, 2024, and 2025 classes has been in the NFL for 5 years. These sets of quarterbacks will serve as secondary test sets. Once again, a high accuracy score and a high R-squared will be the indicator for success. This second test set is arguably more important to see better performance in, as this is the live data that teams will be making decisions off of moving forward. If the model performs strongly on both, then it will be able to reliably provide teams with extra insights needed when they are looking to draft a quarterback.

References

"College Football Stats, History, Scores, Standings, Schedule & Records: College Football at Sports." SRCFB, www.sports-reference.com/cfb. Accessed 29 June 2024.

"Pro Football Stats, History, Scores, Standings, Playoffs, Schedule & Records." Pro Football Reference, www.pro-football-reference.com. Accessed 29 June 2024.

"Where Are Great NFL Qbs Picked? Analyzing 54 Years of History by Draft Position." FOX Sports Research, FOX Sports, 3 Mar. 2024, www.foxsports.com/stories/nfl/where-great-nfl-qbs-are-picked-analyzing-54-years-of-history-by-draft-position. Accessed 29 June 2024.