NFL Quarterback Predictions: An Analytical Approach Using Environmental Data

By Xaric Rios, Aaron Palomin, and Santiago Mendiola
Advised by Dr. Erik Enriquez

In a time where data analytics drives decision making across all domains, sports analytics has seen rapid growth particularly in football. Our goal for this project was to explore how machine learning would be used to predict NFL quarterback behaviors. We focused first specifically on a quarterback's rushing attempts as a proof of concept, then fully transitioned to passing and rushing touchdowns made by quarterbacks. Originally, we wanted to create a system capable of predicting full game outcomes, but the scope proved to be too much given our constraint of time and resources. Instead, we refined our focus to concentrate solely on quarterback rushing and passing metrics.

Motivated by the NFL's growing body of publicly available data and the popularity of player statistics among both fans and analysts, we saw an opportunity to come up with something important to use in sports data science. We were inspired not just because of the potential technical challenge, but also by the possibility of uncovering patterns in each quarterback's decision making that are not immediately obvious. Platforms like ESPN and StatMuse provide a lot of statistics, but few solutions exist that examine players' tendencies in the context of environmental causes like wind, temperature, and the field conditions.

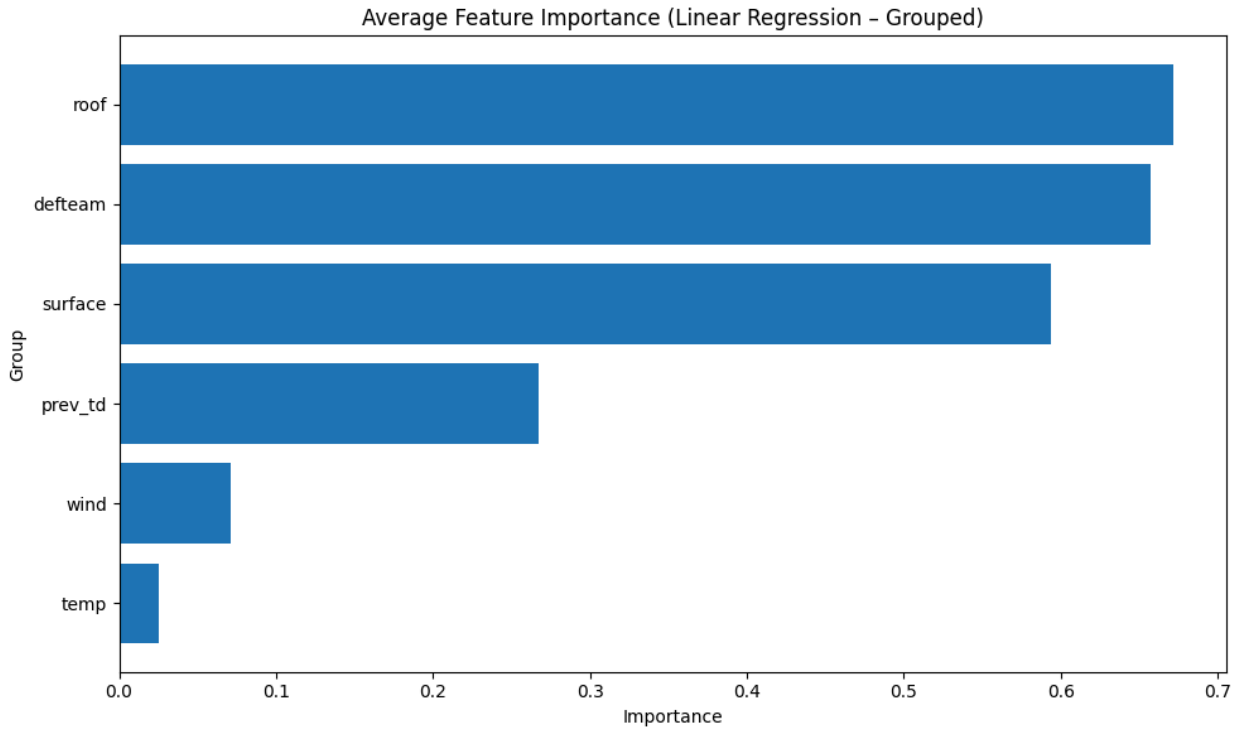
To address this, we wanted to propose a system built around the nfl verse GitHub repository, which contains detailed play by play data from different NFL games. Our pipeline was constructed using different Python libraries such as Pandas, NumPy, Sklearn, and Matplotlib within Google Colab. The architecture we wanted was relatively simple: start by loading and

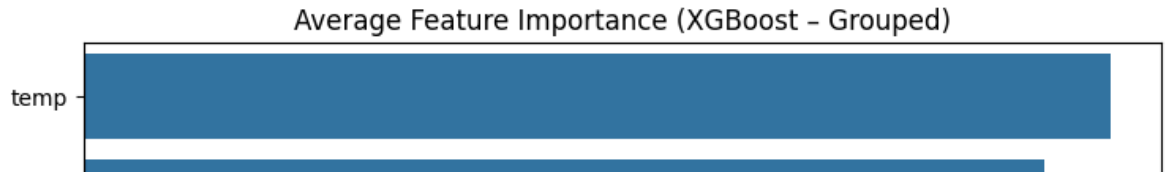
cleaning data, exploring what data we had , engineer different features (e.g roof and surface), train machine learning models, and then evaluate their performance. We trained and compared different models such as linear regression, xgboost random forest, and dummy models, each chosen for their balance of performance and how easy they are to interpret.

For those who may be unaware, Linear Regression is one of the most straightforward machine learning models. It works by fitting a straight line through the data that best explains the relationship between the input features and the target variable. Each feature is assigned a weight (or coefficient) that reflects its influence on the prediction. This makes linear regression highly interpretable. Changes in predictions can be directly traced to specific input variables. In contrast, xgboost Random Forest is a more advanced, tree based model that builds an ensemble of decision trees to capture complex, non linear relationships in the data. It iteratively improves its predictions by focusing on the errors of previous trees, which makes it powerful and often more accurate, especially with larger or noisy datasets. However, this added complexity can make it more difficult to interpret compared to linear models.

The most important architectural decisions involved are feature selection and model choice. In comparison, while linear regression provided a straightforward approach with an interpretable baseline, random forests offered complexity and accuracy at the expense of transparency. Throughout our development, we continuously improved our preprocessing to be able handle any missing values, normalize scales, and be able to encode categorical data, which were necessary for the improvement of model performance.

We evaluated our system both statistically and qualitatively. We confirmed our model outputs against known statistics reported by ESPN.com, and the accuracy of predictions was measured through cross validation of events and games that took place beyond the scope of our training data. Despite meticulous data preparation and model tuning, we found that depending on which model is observed different insights can be taken away. While comparing feature importance across both models, Linear Regression and xgboost highlight different strengths. Linear Regression finds more importance in categorical features like roof, defteam, and surface, suggesting these variables have strong, consistent individual effects on the target. Now in contrast with that xgboost puts emphasis on continuous variables such as temp, wind, and prev_td. In short what we found is that if the goal is explainability and transparency, Linear Regression is the more interpretable choice. But for predictive accuracy and modeling complexity, xgboost offers better flexibility at the cost of some interpretability. It would ultimately depend on what aspect you are looking for, i.e predictive power or explanatory power.





Project management played a critical role in the success of our project. We used Jira for keeping track of task assignments and GitHub to continuously control our updates, allowing us to organize each team member's responsibilities and to synchronize our code. The weekly meetings helped the team maintain a shared understanding of our progress, this enabled us to recognize when obstacles came up. Documentation and communication tools like Google Docs and Discord ensured that all team members could collaborate and communicate effectively, especially to resolve scheduling conflicts and assignment inquiries.

In summary, our project successfully delivered a working system that was capable of analyzing NFL quarterback data. From cleaning data and preparing for an analysis, to generating predictive models with which we were able to measure data accuracy. While the hypothesis regarding the influence of environmental conditions was not strongly supported by the data, the project offered valuable insights into the complexities of sports data analysis. This investigation helped us understand real world machine learning workflows, from data engineering to algorithm evaluations.

Ultimately, our most significant takeaway was the importance of flexibility in both design and expectations. We adapted to our setbacks, redefined our goals, and still produced a system that met functional requirements and formed a strong foundation for future research. In a field as unpredictable as football, adaptability proved just as important as any technical tool.