Predicting NFL Game Outcomes Using Time-Series Analysis

Ankith Rajendran akrajend@usc.edu
Mourad Zeynalov zeynalov@usc.edu
Manuj Kumar manujkum@usc.edu

Abstract

In recent years sports betting has become one of the fastest growing industries in the United States. The National Football League (NFL), with over $13 billion in revenue in 2022, is one of the primary ventures in sports betting.Predicting the outcomes of NFL games would be of significance to consumers, NFL franchises, and the gaming industry alike. Our research aims to contribute to the field of football analytics by leveraging machine learning techniques. We have chosen time-series as a focal point due to the nature of sports data being seasonal. Accurate predictions for a team's performance during a given season can have a significant impact on betting odds. Our approach uses statistics from previous teams and determines the best indicators of success, using them to build a successful predictive model for determining the outcome of an NFL game. Therefore, this project will be useful to teams and fans in improving the game dimensions and as a case that is beneficial to all wagering market agents.

1 Introduction

Forecasting, for example, the outcome of an upcoming NFL game is a complex problem given the myriad factors that can impact who wins or loses. These factors include team dynamics, individual player performance, location, weather, and other conditions tied to the game. Moreover, while offensive and defensive strategies, skill levels, injuries and other factors change as well over time, each game typically consists of several of these elements in concern with each other. The problem takes on a form similar to predicting the stock market, since there are deterministic elements embedded in the game that can change meaningfully over a day, and that a good predictive model should be able to capture. However, the NFL data associated with games is considered 'temporal' – which means that some patterns change over time. Team dynamics and styles of play evolve with each game played over the course of several weeks or months, so we must design a predictive model that can be exposed to these patterns and learn in a dynamic way to generate the best possible forecasts. Addressing this problem could have widespread applications: a stable model could help NFL teams fine-tune strategy, offer actionable insights to fans and assessments to analysts, and be leveraged by the betting industry, providing a data-driven alternative to its traditional methods of forecasting game outcomes.

Importance

Considering current trends, predicting games outcome has numerous benefits;for instance it has a major importance to respective teams (to build their strategies to perform better from previous stats). It has a promising impact in fantasy sports and the broader sports betting market,even though betting is a very traditional field.Nevertheless, advancement in data, technology and development of these kind of models that is capable of analyzing and predicting these outcomes requires overcoming key challenges in data variability and sequence modeling, making it both a challenging and valuable area of study.

Challenges

By their nature, NFL games are complex, there could have been up to hundreds of factors contributing to the outcome of a game, including, but not limited to, up-trends/downtrends of a team, players with different coaching ideas, and uncontrollable factors like weather. It is hard to model these features in a way that considers sequential game nature. This problem also includes working with a really large dataset with a high number of features. If we do not generalize well, there is a risk of overfitting the model. Other than the above-mentioned challenges, a few other reasons such as recent performance trends and integration of non-quantifiable factors such as team bonding/morale or coaching strategies into a model which works primarily on statistical data resulted in a complex model.

Beneficiaries

A strong model for predicting the outcomes of NFL games would be very beneficial in a wide scope. Predictive insights could help NFL teams/coaches make strategic decisions in order to optimize game-day performances. It will add to the depth of understanding of the fans and sports analysts about the possible outcomes of the games, hence their interest in the sport. Such an outcome would, therefore, be of great benefit to the sports betting industry, along with fantasy sports platforms, in terms of enhanced data-driven predictions that might make odds setting better-informed and boost market trust. Overall, such a project could give a really valuable contribution to sports analytics development by providing a basic replicable method for game prediction.

2 Related Works

Several prior studies have explored the application of machine learning and statistical techniques to predict NFL game outcomes. *Stephen et al.* uses a logistic regression approach using team performance metrics as features. XGBoost is generally considered more powerful than logistic regression. The strength of their approach is that its interpretable and identifies key predictive features. It creates separate models for each team to capture team-specific patterns. However, logistic regression may not capture complex nonlinear relationships as well as XGBoost could. The model struggles with inconsistent teams and recent changes in team performance, while a non-linear approach may address these weaknesses by capturing more complex patterns. The evaluation metrics they used were NFL game data from 2001-2018, with 2001-2016 used for training and 2017-2018 for testing. Their resulting accuracy ranged from 30-70% depending on the team, with an average around 60%, showing some predictive power, but significant room for improvement, especially for inconsistent teams.

Stekler et al. uses the New York Times power scores as predictors of NFL game outcomes. Power scores are a simpler, more interpretable rating system for teams – and accounts for home field advantage. However, power scores may not capture temporal trends as well as XGBoost could. The model's performance declined over the study period, suggesting it may not adapt well to changes in the league over time. The model used NFL game outcomes from 1994-2000 seasons as the evaluation set, measured by percentage of games correctly predicted. The power score model correctly predicted about 66% of games overall, but performance declined from 69% in 1994 to 63% in 2000 – showing difficulty in adapting to changes in the league over time.

Koopman et al. develops a dynamic multivariate model using score-driven time series. The model captures temporal dynamics in predicting match outcomes – similar to our time-series approach. It modeled the number of goals scored by each team as a Poisson distribution. It then used the Ranked probability score (RPS) for accurate predictions from match results for the top divisions of England, Germany, France, and Spain over multiple seasons. One of the issues is that The Poisson assumption may not always hold for high-scoring games, whereas XGBoost may be more flexible in modeling score distributions. The results demonstrate the success of time-series forecasts, while still lacking compared to machine learning based approaches.

3 Methods

Considering the nature of work required to predict games, extensive care has to been given to quality of data and the type of statistical information they represent. In order to create a robust model that is able to generalize well, the first step of the pipeline was to gather data from reliable sources that is compact but also is descriptive in the amount of information it conveys. Therefore, the type of data we considered was team data, primarily split into offense and defense data as units, rather than individual players. Data of individual players is helpful, but it adds additional complexity, especially in football games where there are different lineups based on whether a team is on the offense or is trying to defend the opposing team. As a result of this consideration, the FiveThirtyEight dataset on team matchups was one of the datasets chosen for training our models. This dataset revolves around assigning teams a *Elo* rating, which is an indication of the power rating of a particular team. It is a common approach in game prediction tasks to better capture the performance of a team, taking into accounts variables like current form, head-to-head win loss ratio and so on.

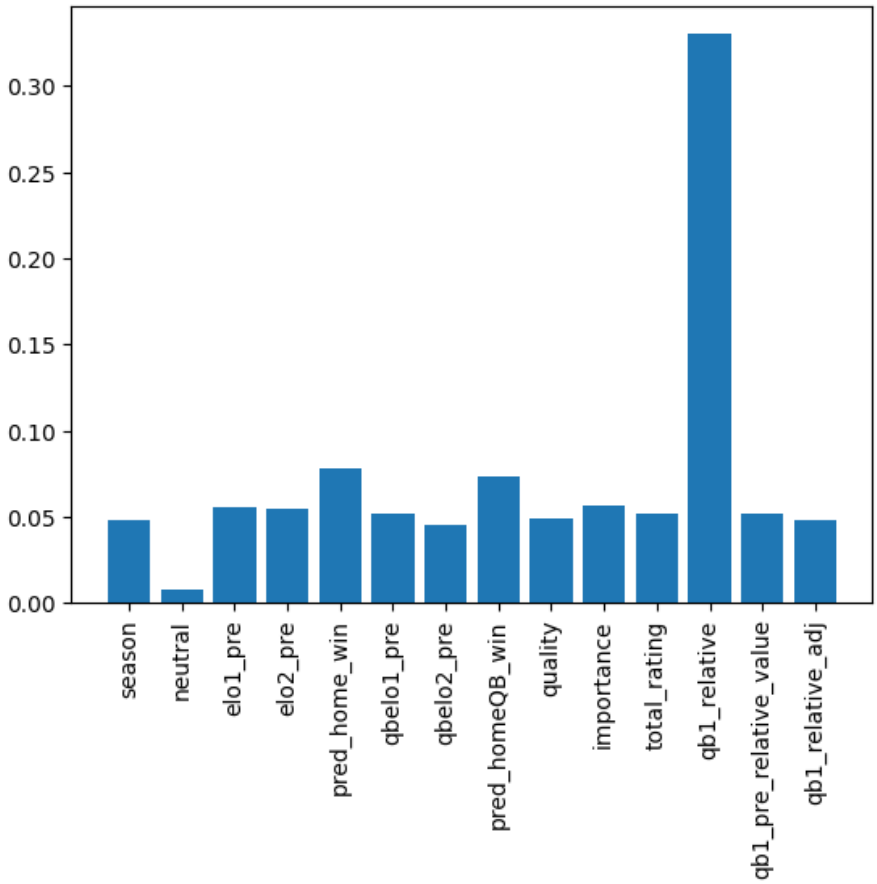

Figure 1: XGBoost feature importance

An additional advantage of using *Elo* ratings are that these ratings are dynamic, which means they change after each and every match based on the outcome. This helps measure the strength of a team relative to its competition in a context-sensitive manner. *Elo* ratings also an indicator of certain matchups where the overwhelming favorite of the match might not win because the opponent relies on certain strategies that nullify the attacks of the team and so on, capturing changes in outcomes due to strategical tweaks and so on. The modification done to the *Elo* rating is how the weighted average is calculated, instead of giving each feature equal importance, we use the XGBoost classifier to denote the effect of each feature onto the outcome variable, and the resulting graph can be seen in *Figure 1*. As expected, the feature with the most importance was the *qb1_relative* feature, a difference measure between the form of the two quarterbacks, which is considered as the most important role in

football. As expected, the feature with the most importance was the *qb1_relative* feature, a difference measure between the form of the two quarterbacks, which is considered as the most important role in football. Using this information, we calculated the new *Elo* rating, and used them to train our models.

For the baseline model, Logistic Regression and a MultiLayer Perceptron (MLP) were chosen as models to train on. These models were chosen as these are simple but efficient, and give us a base indication of the accuracy on which we can test and compare our more sophisticated model/techniques against. These models were implemented successfully, with their results illustrated in the `Experiments` section. All of the models were first trained on individual features, to replicate the state-of-the-art implementation, and then moved on to the base Elo dataset, and then finally the updated dataset with the weighted average of the *Elo* rating concatened with annotated weather data during those matches.

To add onto this, based on popular methods mentioned in the *Literature Review*, XGBoost models were also a suitable model to use, and we implemented the same. Cross validation was performed on the validation set and the best hyperparameters were then chosen for training. As discussed, to take advantage of the sequential nature of data, RNNs and LSTM architectures were also explored.

Once all of these methods were tested on the base dataset, our next step was to append annotated weather data onto the dataset, and monitor the effect of these features on the outcome of the game. The spatial and temporal features chosen were *Temperature, Weather Conditions, Humidity,Precipitation and Wind Speed*. Our hypothesis was that these features, while not as important as game statistics, are still features that contribute to the performance of a team and can ultimately decide games. This information was scraped from spreadspoke, a popular third-party website for sports stats. This dataset, with the above additional features were trained using the same models and parameters, and their performance can be seen in the table below. Further discussion of these results will be given in the Experiments section.

4 Experiments and Results

As mentioned above in the *Methods* section, to get a grasp of the data and to understand the distribution better, we initially opted for data visualization techniques in the form of distribution graphs and trends through time to highlight the importance of particular features. As shown in *Fig 2* and *Fig 2*, these illustrate the win probability distribution of home and away teams over the course of time and the changes in *Elo* ratings of a particular team, in our case the `LA Rams` from 1950-2022. As shown in Fig. 3, the peaks of this graphs corresponds to the years the LA Rams ended up winning the championship, and this proves that the *Elo* rating is a suitable indicator of a team's form.

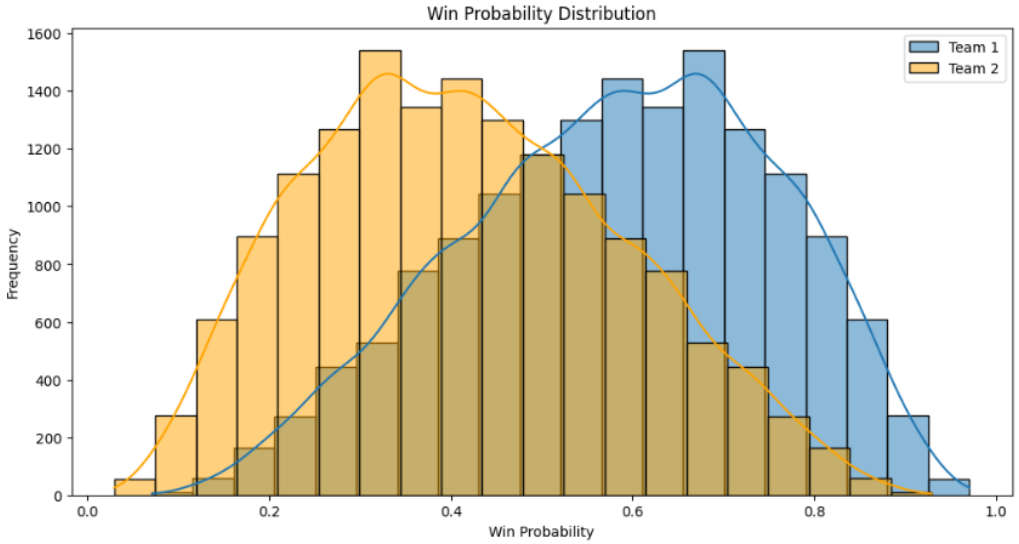

Figure 2: Win Probability Distribution of home and away teams

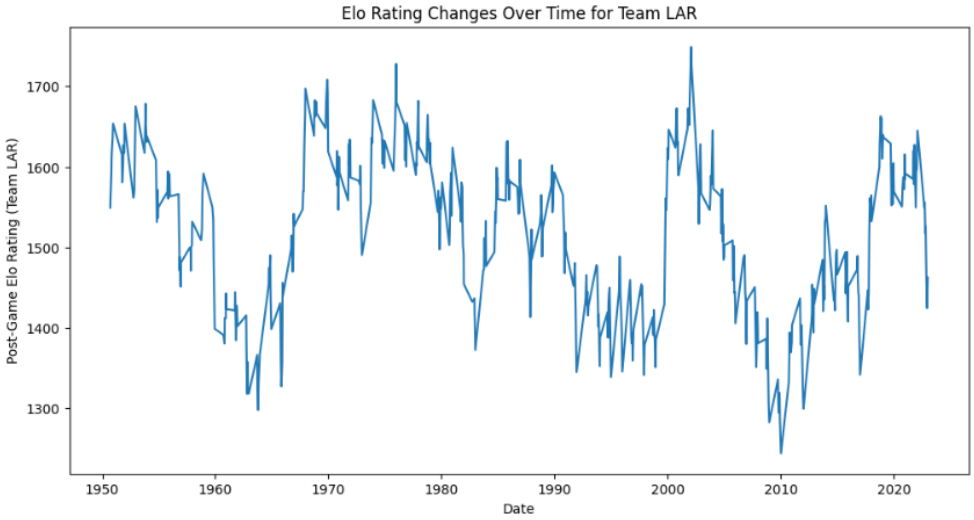

Figure 3: Elo change over time for the LA Rams

Once various viz. techniques were implemented and correlation between features were determined, model training and evaluation was the next step. To begin with, the first model that was considered was a Logistic Regression model. The train set for the same was the data from the year 1950-2018, and the test set was the schedule for the next four years till 2023. This would ensure that a team's past form would be given priority, and also prevent any data leakage. The accuracy metrics chosen for this task was the F1-score, with the results shown below in `Table 1`. To supplement this, a 5-layer MLP with a *tanh* activation was implemented and run for 50000 epochs on the train data.These models gave us an indication that the data pre-processing techniques adopted are effective and the project is headed in the right direction. These results also ultimately give us a baseline on which we can compare the next set of models against, which were the sequential based models - the RNNs and the LSTMs.

As mentioned in the `Methods` section, the sequential nature of data, from one season to the next, can be exploited by use of recurrent neural networks, with the principles of weight sharing and sequential processing providing a benefit to performance. These models were first tested on the base dataset, without the weighted average of the `Elo` rating, and then was trained again on the updated dataset to compare the performance between the two methods, and also observe if our modification led to an increase/decrease in performance. The custom RNN and LSTM architectures consisted of 34 layers in total, with 30 RNN blocks followed by a dropout layer and two dense layers, with residual connections in an attempt to tackle the vanishing gradient problem. We choose the total number of layers to be 34 since the total number of NFL games in a season is 272, which would be divisible by 34 and would lead to more efficient computation. As expected, all the models showed a ~2-3 % increase in accuracy, which is a significant improvement over previous state of the art models.

Table 1: Model Performance

	Logistic Regression	XGBoost	MLP	RNN	LSTM
Individual Features	61.33%	62.94%	51.40%	63.03%	63.27%
Base Elo Dataset	66.47%	69.03%	52.41%	67.24%	70.09%
Updated Elo Dataset + Annotated Weather Data	67.39%	72.67%	54.32%	72.13%	74.63%

From `table 1`, we can see that the model achieves the least accuracy when data is represented as just raw statistics, and not segregated into offensive - defensive - QB related actions. As a result, the model is unable to actually learn something of value and results in a very high error of ~0.8. An improvement of this is titled *Base Elo Dataset*, which segregates game statistics into offensive and defensive actions, and also incorporates the *Elo* ratings of the two teams, with seperate ratings computed for the quaterback of the two teams as well. This combination of individual features

resulted in a significant increase in model performance, as shown in table 1. The final experiment explored was basing the *Elo* rating on feature importance, and changing the weights of features based on the correlation with respect to game outcome. In addition, spatial and temporal data, mentioned in the Methods section were also added and then the different models were tested again. The amplification of the most significant features proved to be successful, seeing a ~ 4 % increase in the best performing model, the LSTM.

5 Discussion

The results presented in Table 1 highlights the comparative performance of various machine learning and deep learning models across three datasets: Individual Features, Base Elo Dataset, and Updated Elo Dataset with Annotated Weather Data. We can infer the following from our experiments:

The improvement in performance from the Individual Features dataset to the Base Elo Dataset emphasizes the importance of using the Elo rating system as a foundational metric for predictive modeling. All models demonstrated an increase in accuracy when trained on the Base Elo Dataset, with LSTM achieving the highest improvement (6.82 %) compared to Logistic Regression, which showed a relatively modest gain (5.14 %).

It was also observed that among the models, LSTM consistently outperformed others in all three datasets. Its ability to capture sequential dependencies and model temporal patterns is a likely cause of its high accuracy. The RNN also performed well, but the gap between RNN and LSTM becomes more pronounced when using the Updated Elo Dataset with Annotated Weather Data, where LSTM achieves a significant accuracy of 74.63 %. On the other hand, MLP showed limited improvement across datasets, suggesting that it struggles to leverage the sequential and contextual nature of the data. Logistic Regression and XGBoost, while relatively robust, were outperformed by the deep learning models, especially on the latter datasets.

The Updated Elo Dataset, augmented with annotated weather data, substantially improved model performance for all the models. For instance, the inclusion of weather data boosted XGBoost's accuracy by 3.64 % and LSTM's by 4.54 %. This highlights the significance of integrating contextual features such as weather conditions, which can influence game outcomes. The added features seem to particularly benefit models capable of capturing complex interactions, such as XGBoost and LSTM.

However,while LSTMs delivered the highest accuracy, the performance comes with higher computational cost and training complexity compared to simpler models like Logistic Regression or XGBoost. For use-cases where a portion of accuracy can be sacrificed for quick real time predictions or limited computation at hand, feature based models such as XGBoost might be better suited for the task.

6 Future Scope

This study demonstrates the value of integrating domain-specific features, such as Elo ratings and contextual weather data, for improving predictive accuracy in sports based domains. This project also highlighted several avenues for further explorations, which will be discussed below:

The annotated weather data proved beneficial, but other contextual features like player injuries, team strategies, or crowd attendance could further enhance model performance. Another set of features would the offensive and defensive ranking amongst the entire league as a unit, and also the number of rookie players in the squad whose performance might be hindered due to pressure of performance. Additionally, another factor to consider would be how the absence of one player, either due to injuries or being transferred to another team, would affect the *Elo* rating of that team. It is dependent on numerous factors - position, individual stats and team comp being some of those factors.

The models, particularly deep learning-based models like LSTMs and RNNs, will benefit from additional hyperparameter optimization. Optimization techniques such as Bayesian optimization or grid search can be employed to fine-tune parameters like learning rates, number of layers, and hidden layers necessary for optimal performance.Beyond LSTMs and RNNs, more sophisticated architectures such as Transformer models or hybrid approaches combining CNNs and RNNs could potentially capture even more sophisticated patterns in the data, and thereby increasing accuracy. However, with added accuracy also comes the problem of the 'explainability' factor. Complex models like the ones mentioned above do not explicitly outline how predictions are made, and thus it is

difficult to actually know which features to maximize or limit based on performance. An example of this would be when we incorporated weather data to the dataset, we can see that there is an increase in performance, but we do not know exactly how it is used, what teams benefit the most from adverse weather conditions, and to what degree does weather actually affect matches. These factors, if known, can be utilized to better train models and achieve even higher accuracies.

Future attempts could also focus on deploying these models in real-time environments to predict game outcomes dynamically as soon as new data becomes available.

References

[1] Koopman, S.J. and Lit, R., 2019. Forecasting football match results in national league competitions using score-driven time series models. International Journal of Forecasting, 35(2), pp.797-809.

[2] Rakytyanska, H. and Demchuk, M., 2020. Football Predictions based on Time Series with Granular Event Segmentation. In Lecture Notes in Computational Intelligence and Decision Making: Proceedings of the XV International Scientific Conference "Intellectual Systems of Decision Making and Problems of Computational Intelligence"(ISDMCI'2019), Ukraine, May 21–25, 2019 15 (pp. 478-497). Springer International Publishing.

[3] Andrienko, G., Andrienko, N., Anzer, G., Bauer, P., Budziak, G., Fuchs, G., Hecker, D., Weber, H. and Wrobel, S., 2019. Constructing spaces and times for tactical analysis in football. IEEE Transactions on Visualization and Computer Graphics, 27(4), pp.2280-2297.

[4] Min, B., Kim, J., Choe, C., Eom, H. and McKay, R.B., 2008. A compound framework for sports results prediction: A football case study. Knowledge-Based Systems, 21(7), pp.551-562.

[5] Wheatcroft, E., 2021. Forecasting football matches by predicting match statistics. Journal of Sports Analytics, 7(2), pp.77-97.

[6] Boulier, B. L., & Stekler, H. O. (2003). Predicting the outcomes of National Football League games. International Journal of Forecasting, 19(2), 257–270. doi:10.1016/s0169-2070(01)00144-3

[7] Bouzianis, Stephen. "Predicting the Outcome of NFL Games Using Logistic Regression." Honors Theses and Capstones, University of New Hampshire, 2019,

[8] O'Leary, M., 2014. FiveThirtyEight: data vs. the gut. Information Today, 31(6), pp.20-22.

[9] Collins, M., Moir, G., Muresan, J. and Savage, D., 2017. NFL Data Analytics For A New Era.

[10] Vu, S., 2015. The Era of Analytics in the NFL: Application of Modern Portfolio Theory.

[11] Klein, J., Frowein, A. and Irwin, C., 2018. Predicting Game Day Outcomes in National Football League Games. SMU Data Science Review, 1(2), p.6.

[12] Gallagher, M., 2019. A Better Predictor of NFL Success: Collegiate Performance or the NFL Draft Combine? (Master's thesis, East Tennessee State University).

[13] Bosch, P., 2018. Predicting the winner of NFL-games using Machine and Deep Learning. Vrije universiteit, Amsterdam.

[14] Zhong, Y., Zhang, S., Yi, Q. and Ruano, M.Á.G., 2024. The influence of meteorological factors on the technical performance of football teams during matches. Biology of Sport, 41(4), pp.165-172.

[15] Joly, B. and Dik, M., 2021. Cold weather teams in the National Football League and home-field advantage. Proceedings of International Mathematical Sciences, 3(1), pp.10-24.

[16] Wolfson, J., Addona, V. and Schmicker, R.H., 2011. The quarterback prediction problem: Forecasting the performance of college quarterbacks selected in the NFL draft. Journal of Quantitative Analysis in Sports, 7(3).

[17] Reyers, M. and Swartz, T.B., 2023. Quarterback evaluation in the national football league using tracking data. AStA Advances in Statistical Analysis, 107(1), pp.327-342.