**Project Final Report- Predicting and Visualizing NFL Player Salaries- Group 165**

Walker Harrison Butin, Ivan Eduardo Del Valle, Roman Erick Emde, Faiza Mehboob Jamani

*CSE 6242 Project Progress Update- Group 165*

**Introduction**

NFL salary data is highly volatile and influenced by both external and cyclical factors. The NFL salary cap (the maximum amount a team can spend on player salaries) was introduced in 1994, which turned what was formerly a simple supply-demand negotiation between players and management into a more complicated optimization problem.

Apart from individual player performance, it is commonly hypothesized in sports media that factors such as overall team performance, position, the market for a particular position in a given offseason, and the quality of incoming prospects can all affect the size of the contract that ends up being offered to an NFL player when they are looking to renew.

Additionally, because NFL player salaries are often signed for variable lengths of time (typically from 1-5 years, with some instances of up to 10 years), the contract that a player signs may not necessarily be indicative of their true value to a team, but rather an effort to secure an adequate player in a position of need for multiple years. Therefore, players can often be overpaid when considering what a true fair market value for their services would be. A novel approach using time series techniques, predictive modeling methodologies, and intuitive data visualization techniques could improve the ability of NFL teams to better value and prioritize players that would improve their team.

**Problem Definition**

For the purposes of this project, we have modeled NFL player salary using several performance metrics with the primary objective of identifying undervalued and overvalued players, offering insights that can help NFL teams make more informed decisions and maximize their roster value. We have focused primarily on free agents, as they present a more competitive and dynamic market.

Formal Problem Definition

Given a set of NFL players P = {p₁, p₂, ..., pₙ} where each player p belongs to position set POS = {QB, RB, WR}, and each player has associated performance statistics S = {s₁, s₂, ..., sₘ} over time period T = {t₁, t₂, ..., tₖ}, our objective is to develop a function f: S → V that maps a player's performance statistics to their market value V, measured by the average annual value (AAV) of their contract.

Formally, for each player p with position pos ∈ POS and statistics vector *S\_p = [s₁, s₂, ..., sₘ*], we aim to find: *V\_p = f\_pos(S\_p)*, where f\_pos is a position-specific prediction function that minimizes the error function: *E = ∑(V\_p - V\_p^actual)²* across all players, where V\_p^actual is the player's actual contract value. Additionally, we define a player's valuation status as:

Fairly valued: V\_p ≈ V\_p^actual (within a predetermined threshold)

Undervalued: V\_p > V\_p^actual

Overvalued: V\_p < V\_p^actual

The problem's complexity is increased by the temporal volatility of NFL salary data, requiring normalization methods to account for market trends and inflation over time.

This framework allows NFL teams to identify market inefficiencies and optimize resource allocation within salary cap constraints by targeting undervalued players whose predicted market value exceeds their current contract value.

**Literature Survey**

*Does one Simply Need to Score to Score?:* (Schmidt, Berri, & Brook, 2007) An analysis of NBA salary valuation, where the primary argument made is that teams overvalue points scored, as opposed to more sophisticated metrics. While a different sport, the principles and methodologies are relevant.

*Exploring Patterns Behind Sports:*(Liu, Ma, & Zhou, 2025) Research paper with extensive statistical analysis on general sports patterns, with applications of time-series analysis that could be of relevance.

*Machine Learning Model to Evaluate the Salary of Football Players:*(Li, Kampakis, & Treleaven, 2022) Research paper on salary valuation of European Football (Soccer) players, using techniques including regression and random forest models which could be implemented in our project.

*A review of football player metrics and valuation methods: a typological framework of football player valuations:*(Hill, Skinner, & Grosman, 2025)This paper creates a decision framework to help researchers pick the valuation typology best suited to the type of data they have available. This will likely be quite useful when designing our algorithm to value players.

*Using Tracking and Charting Data to Better Evaluate NFL Players: A Review:*(Eager, et al., 2023)This paper uses NFL charting data combined with newly available tracking data such as RFID chips to identify better performing players. This is closely aligned with our project, using some of the same data sources we intend to use.

*Do Players Perform for Pay? An Empirical Examination via NFL Players’ Compensation Contracts***:** (Kim, Sarin, & Sarin, 2018)This paper examines whether incentive-based compensation leads to better results than flat-fee compensation. They use two metrics that might be useful in our project, as they measure the likelihood that a particular play leads to a score or a win.

*XGBoost: A Scalable Tree Boosting System:*(Chen, Guestrin, 2016) This is the original published paper for one of the ML algorithms we plan to test.

*A Non-Linear Approach to Predict the Salary of NBA Athletes using Machine Learning Technique:* (Jain, Jain, Pancinovia, George, 2022) This paper ran a similar analysis to what we plan to do, except for the NBA. They got the best prediction quality from RandomForest, and we can directly apply some of their learnings to our own project.

*NFL Career Success as Predicted by NFL Scouting Combine:*Szekely, Brian, et al. (2023). This paper investigates the use of data from the NFL Scouting Combine to predict future success. These are valuable considerations when seeking to analyze a player’s physical abilities in relation to performance and salary.

*Is the Blind Side Tackle Worth It?: An Analysis of the Salary Allocation of the NFL Offensive Line:* Froelich, R. (2016). This paper uses logistic regression to evaluate the importance of the positions in the offensive line in relation to salary allocation. We may choose to apply these methods to our project, which overlaps with the scope of this study.

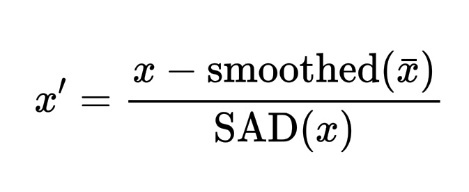
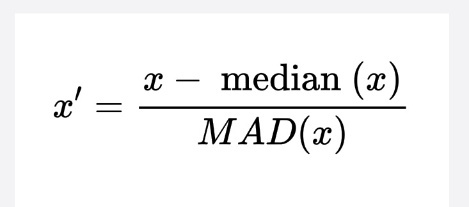
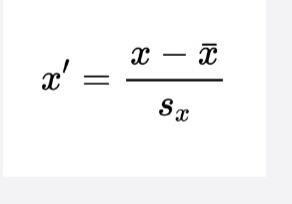
*Quantifying NFL Players’ Value with the Help of Vegas Point Spreads Values:* Hoffer, A., & Pincin, J. A. (2023). This study aims to set player value based on Las Vegas point spreads. The methods used highlight the metrics that significantly correlate to player value, which prompts further analysis for our valuation methods.

*Forecasting Trends and Seasonal by Exponentially Weighted Averages:* Holt, C. C. (1957). Exponential smoothing methodology first introduced the extension of an upward trend component.

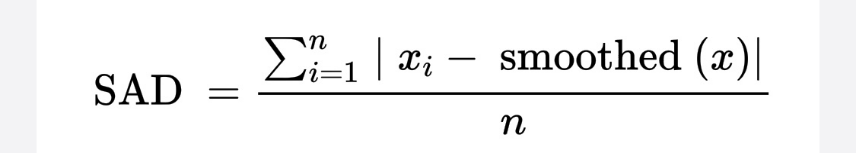
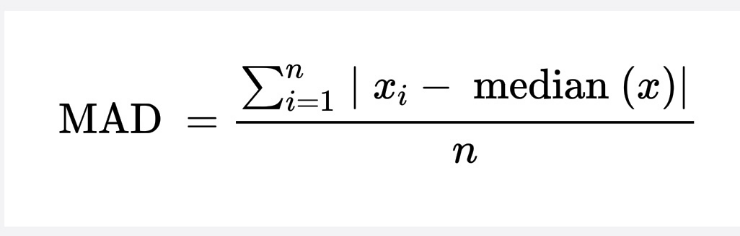
**Proposed Methods**

As NFL salary data is highly dynamic and influenced by a variety of factors, our group tested three different normalization methods to reduce errors while simplifying data analysis and interpretation. These included mean centering, median centering (Kappal, 2019), and exponential smoothed average centering.

*Mean Centering*  *Median Centering Exponentially Smoothed Centering*



Median absolute deviation (MAD) is used to calculate median centered values, which takes the average absolute deviation from the sample median. Smoothed absolute deviation (SAD) takes a similar approach, taking the average absolute deviation from the smoothed sample mean. The formulas for both are shared below for convenience:



Mean and median centering are standardized methods used extensively in research. However, with the heavy amount of volatility in the mean salary signed by players by year, and the inability of the median to capture the top end of the market, we propose a novel method for evaluating market trends of salary data in the NFL. Using exponential smoothing on mean salary values by year, we can still capture market trends while reducing the variability.

After normalizing the data and partitioning the data into testing and training sets, where we test on players who signed in 2024, we then built GBM and regression models for each position, each with their own strengths and weaknesses. First, we built linear regression models for each of the three positions, using LASSO for variable selection. Regression is popular for its simplicity and interpretability but generally has difficulty when dealing with highly correlated data.

We used gradient boosting for the next set of models, which is a tree-based ensemble method that combines multiple models to generate more accurate results. While random forest builds each decision tree/model independently, gradient boosting builds models sequentially with each new tree improving upon the previous one. The method is robust and tends to achieve higher accuracy than linear regression and is more capable of capturing non-linear complexities within the data. We fine-tuned the parameters for the ensemble methods using a grid search for each position-specific GBM to achieve the best results on our training and testing data.

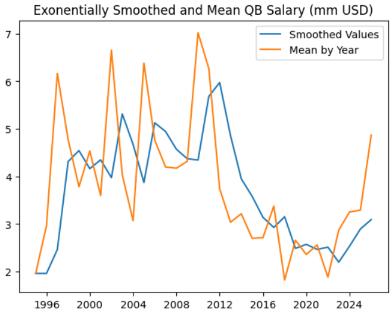
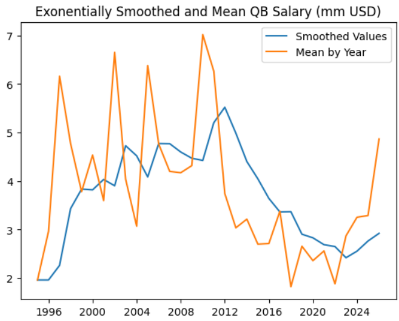
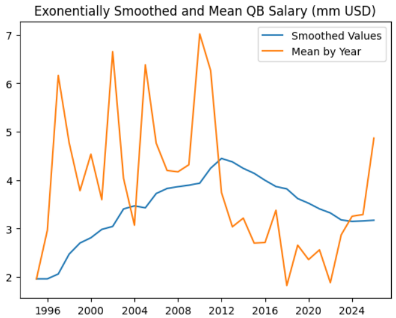
After running the models, we converted the predicted exponentially smoothed salaries back into present day USD, so that we could compare our salary valuations with the most recently signed contract for a player, allowing users of the dashboard to identify potentially undervalued or overvalued players. We did this by applying the inverse of the scaling formulas outlined above to the predicted values.

To visualize these modeled player valuations and allow users to interact with and gain insights from them, we created both a heatmap and an interactive dashboard to visualize our results for our users. The heatmap compares a player’s current salary against our model’s predicted salary, using a metric we built called Valuation Status to portray if a player is undervalued, overvalued, or fairly compensated. This metric is presented by the color of each circle (player), and the size of each circle depicts the degree to which each player is under or overvalued. Additional calculated fields were added to denote the difference between Actual Salary and Predicted Salary in the heatmap tooltips. We also included interactive filters such as Position (QB, WR, RB), Salary Range (Low, Medium, or High), and Valuation Status.

**Evaluation**

For exponential smoothing, we needed to pick a smoothing level and decide whether we wanted to include trend and seasonal parameters. There was no reason to expect seasonal trends, and while we expected an overall upward trend in the progression of mean salaries, this was not clear when looking at the data. Therefore, we just needed to evaluate the smoothing level, alpha, for how much we wanted to discount older observations. Several different values of alpha were evaluated visually:

*Alpha = 0.1 Alpha = 0.3 Alpha = 0.5*



Based on these plots, an alpha of 0.3 was selected. Values were then normalized using the exponentially smoothed normalization method outlined previously. These results were fed into down-stream regression models, and compared to mean-normalized and median-normalized salary values.

Observations

When visualizing the data, it became apparent that the salary range for running backs was much smaller than that of quarterbacks and wide receivers. The highest-salaried and most overvalued data points are associated with quarterbacks. Furthermore, when visualizing the top 5, 10, and 15 highest undervalued players, only one of them is a quarterback as the ninth most undervalued. This supports the concept that quarterbacks are the most “valuable” position in terms of salary.

For model evaluation, we used standard metrics such as R-squared, which returns the amount of variation in the dataset that is explained by the model, and Mean-Squared-Error (MSE), which measures the average of the squared differences between the predicted and actual values (in this case, smoothed salaries). The first three regression models were able to achieve R-squared scores between .5 and .65 on the features returned from LASSO. Feature selection on the quarterback-specific regression model, for example, returned four features: passing completions, passing attempts, passing yards, and rushing yards. The test-data R-squared squares were .644 for the quarterback model, .534 for the running back model, and .548 for the receiver model. The GBM models performed much better. The test data R-squared values ranged from .75 - .9 when modeled on the exponentially smoothed salaries with tuned hyperparameters. The grid search results for the quarterback specific model, for example, returned a learning rate of .01 and a max depth of 2. This increase in performance here is likely due to the GBM’s ability to capture nonlinear trends and interactions with the data. After deciding on using GBM for our final predictions, we then tested the models on each alternate time series method. The results for each position-specific GBM model on each type of normalization method are shown below:

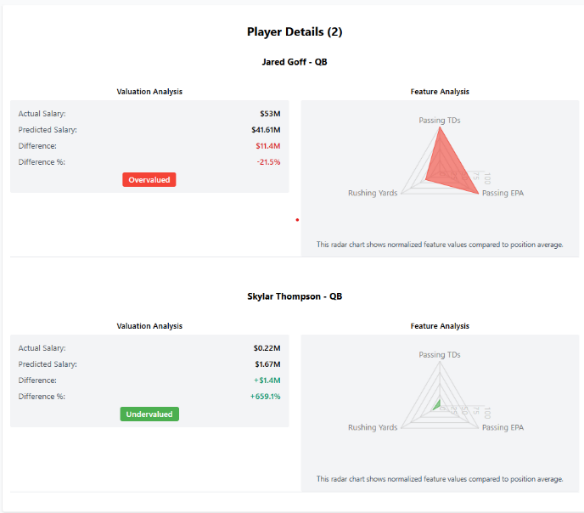
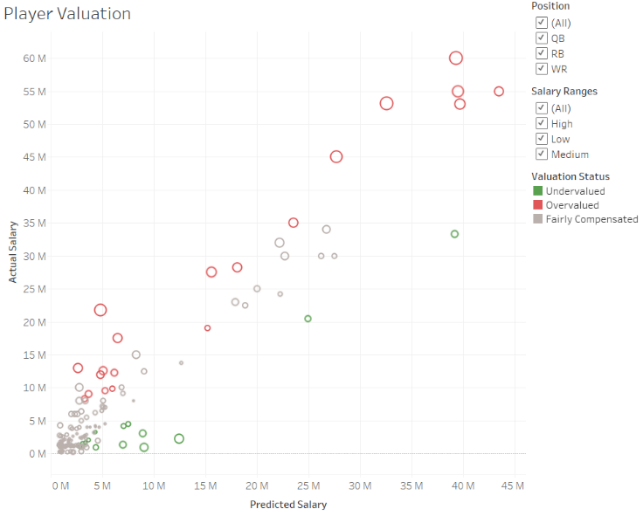
|  |  |  |  |
| --- | --- | --- | --- |
|  | QB | RB | WR |
| Mean-  centered | R2: 0.83  MSE: 1.27 | R2: 0.66  MSE: 1.61 | R2: 0.84  MSE: 1.72 |
| Median-centered | R2: 0.84  MSE: 1912.2 | R2: 0.76  MSE: 16.02 | R2: 0.81  MSE: 83.4 |
| Smooth-centered | R2: 0.842  MSE: 1.42 | R2: 0.75  MSE: 5.8 | R2: 0.9  MSE: 4.8 |

While each method performed well, based on these results, we decided on using smoothing for our final salary predictions. We expect this is due to the ability of exponentially smoothed mean salary values to capture the market for a position in recent years, instead of just a single mean value.

We then stored the three most important features returned from the GBM model, to be used in our final dashboard. For the quarterback-specific model, these were passing touchdowns thrown, passing EPA (Expected Points Added), and rushing yards gained. For the running back model: rushing yards gained, target share, and receiving yards after a catch. For the wide receiver model: receiving yards, target share, and receiving first downs. All of these made sense intuitively, as these features represent common and important ways to evaluate the performance of a player at each position.

After converting the smoothed salaries back to present-day USD, the final dataset with the most important historical player statistics, most recent player salary amount, and our predicted player salary valuation, was ready to be analyzed visually with our dashboard and heatmap

. *Heatmap* *Dashboard Drilldown*





**Conclusions and Discussion**

Limitations

Our models do not account for player injuries throughout the season; this can skew the performance metrics used to build the models. As those metrics would not be achieved to the same degree as other players due to lost playing time, our model outputs and visualizations could prematurely label a player as “overvalued.” Therefore, players with significant injuries should be noted or removed from model evaluation. Another limitation of our approach includes decreased accuracy for elite outlier players (particularly quarterbacks) whose market value often exceeds what performance metrics alone would predict, likely due to intangible qualities and market scarcity.

Additionally, using current player salaries as a baseline is not necessarily a fair representation of the market they will demand in the following off-season. We did exclude rookie contracts (i.e., first contracts when players enter the league), as these are standardized by the collective bargaining agreement between players and the NFL. However, players’ values can shoot up significantly if they have had recent success, for which our analysis did not account. In the case that a real NFL franchise was using this data, however, they would have access to a player’s actual contract demands, which would resolve this issue.

Results

Our analysis demonstrates that exponentially smoothed salary normalization combined with position-specific gradient boosting models provides a robust framework for predicting NFL player market value. Using our most performant centering methodology and regression model, we can capture a fair valuation of a player's worth based on past performance. NFL teams can then use this to better negotiate with, or target, certain players. The models achieved high predictive accuracy (R² values between 0.75-0.9) and identified key performance metrics that drive player valuation for each position.

Significance

Our interactive visualization tools allow teams to easily identify undervalued free agents, optimizing their roster construction within salary cap constraints. This approach provides NFL front offices with a data-driven advantage in player acquisition strategy, potentially leading to more efficient resource allocation and improved team performance. Future work could expand this methodology to defensive positions and incorporate additional metrics such as advanced player tracking data.

*Note: All team members have contributed a similar amount of effort*

**References**

**Kappal, S.** (2019). *Data normalization using median & median absolute deviation (MMAD) based Z-score for robust predictions vs. min–max normalization*. *London Journal of Research in Science: Natural and Formal*, 19(3), 10–17. <https://d1wqtxts1xzle7.cloudfront.net/61796863/Data_Normalization_Using_MAD_Z-Score20200115-15558-1ttyo6e-libre.pdf>

**Di Cellio Dias, P. C., Forti, M., & Witarsa, M.** (2018). *Credit score: A comparison of gradient boosting with logistic regression in practical cases* (Paper No. 1857-2018). Proceedings of the SAS Global Forum 2018. SAS Institute Inc.​[SAS Customer Support Site+1SAS Customer Support Site+1](https://support.sas.com/resources/papers/proceedings18/allResults.html?utm_source=chatgpt.com)

Retrieved from <https://support.sas.com/resources/papers/proceedings18/1857-2018.pdf>

**Seabold, S., & Perktold, J. (**2010). *statsmodels: Econometric and statistical modeling with Python*. Proceedings of the 9th Python in Science Conference. <https://www.statsmodels.org/dev/generated/statsmodels.tsa.holtwinters.ExponentialSmoothing.html>

**scikit-learn developers.** (n.d.). *sklearn.ensemble.GradientBoostingClassifier*. scikit-learn. <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html>