NFL Player Salary and Performance Analysis (2017)

Abraham Reay II
Computer Science Wentworth Institute Of Technology

Abstract—This paper analyzes the relationship between NFL player performance statistics and their corresponding salaries using a simulated 2017 dataset. By applying a multiple linear regression model, the study aims to determine whether individual performance metrics can effectively predict salary cap hits. Despite logical assumptions about high-impact positions correlating with compensation, the results demonstrate poor predictive power, suggesting the need for more complex modeling or richer features. This analysis serves as a foundation for evaluating roster efficiency and understanding valuation dynamics in professional sports.

Keywords—NFL, regression, salary prediction, performance metrics, data science

I. INTRODUCTION

NFL player salaries vary dramatically, influenced by a mix of performance, experience, and market dynamics. This project investigates whether on-field performance metrics such as passing yards, rushing yards, tackles, and sacks can reliably predict player salary, specifically cap hit. A regression-based approach was chosen to evaluate the strength and direction of these relationships. Previous research and common team practices suggest high-impact roles like quarterbacks or pass rushers may drive strong salary associations. The goal of this project is to build a quantitative model to test these assumptions and highlight which metrics contribute most to salary outcomes. In the broader context of sports analytics, understanding salary determinants supports better contract negotiations, talent evaluation, and salary cap management. Analysts and team managers are increasingly leveraging data science methods to uncover hidden value in player contracts. By establishing a statistical foundation, this project connects on-field contributions with financial impact a growing area of interest in professional sports economics.

II. DATASETS

The primary salary data was sourced from a Kaggle dataset titled 'NFL 2017 Player Salaries'. Due to the lack of full performance statistics, player metrics were simulated based on typical values from Pro Football Reference 2017 records. The final dataset contained 1,999 rows and included the following features: Player Name, Position, Team, Cap Hit, Passing Yards, Rushing Yards, Receiving Yards, Tackles, Sacks, and Interceptions. All performance statistics were generated using position-based logic. No missing values were present, and the data was cleaned and engineered to ensure consistency. The dataset did not include variables such as age, years of experience, draft position, or injury history all of which could offer valuable insights. Including such variables in future versions could allow for deeper regression diagnostics and improved model performance. The performance metrics were chosen to reflect typical indicators of contribution by position group (e.g., sacks for defensive linemen, passing yards for quarterbacks), ensuring positional relevance in the simulated stats.

TABLE I.

Feature	Description	Unit	Data Type
Player Name	Player's full name	—	Text
Pos	Player's position (e.g., QB, RB, LB)	—	Categorical
Tm	Team abbreviation	—	Categorical
Cap Hit	Salary cap hit for 2017 season	USD	Numeric
Passing Yards	Total passing yards (simulated)	Yards	Numeric
Rushing Yards	Total rushing yards (simulated)	Yards	Numeric
Receiving Yards	Total receiving yards (simulated)	Yards	Numeric
Tackles	Total tackles (simulated)	Count	Numeric
Sacks	Total sacks (simulated)	Count	Numeric
Interceptions	Total interceptions (simulated)	Count	Numeric

III. METHODOLOGY

A Multiple Linear Regression model was trained using scikit-learn's LinearRegression class. The dataset was split 80/20 for training and testing. Evaluation metrics

included R^2, MAE, and RMSE. The model assumes linearity, normal distribution of errors, and feature independence. While interpretable, linear regression is limited in capturing non-linear interactions and real-world salary complexity. In addition to the baseline linear regression, exploratory tests using regularization techniques (like Ridge and Lasso regression) were considered. These methods help mitigate multicollinearity and overfitting by penalizing large coefficients. However, given the weak fit of the baseline model and limited data richness, further model experimentation was deferred to future work. Feature scaling was deemed unnecessary as all metrics shared similar numerical magnitudes.

IV. RESULTS

The model yielded an R^2 of -0.045, MAE of $1.25M, and RMSE of $1.9M. These results suggest poor predictive accuracy. Among the features, 'Sacks' and 'Passing Yards' correlated positively with salary. 'Tackles' and 'Interceptions' showed weaker or negative contributions. The model failed to capture critical nonlinear drivers like experience, injury, or market dynamics. The negative R^2 score also implies that the model performed worse than simply predicting the mean salary for all players. While 'Sacks' had the highest positive influence among predictors, some features (e.g., 'Interceptions') may act as noise due to inconsistent positional relevance. These findings support the notion that salary is not linearly tied to performance metrics alone especially when stats are simulated and lack contextual factors like game impact or playoff performance.

V. DISCUSSION

The poor model performance reflects the limitations of using only raw performance data. NFL salaries depend on multiple layers of evaluation beyond stats. Future work should explore complex models like Random Forests or XGBoost and integrate additional features such as experience, draft round, or contract year to better capture real-world conditions. Additionally, positional biases in how contracts are structured could skew salary patterns. For example, kickers and special teams players may have relatively low statistical output yet receive stable contracts. Integrating contract type (rookie, veteran, franchise tag) and performance-based incentives could help isolate pure performance influence. Data augmentation using real-world NFL game logs and PFF-style advanced metrics would elevate both the accuracy and relevance of future models.

VI. CONCLUSION

This study attempted to estimate NFL player salaries using simulated 2017 performance data. The regression model demonstrated poor accuracy, reinforcing the idea that salary is influenced by factors beyond performance alone. However, this project lays groundwork for more advanced analysis with richer features and more complex modeling techniques. From a team management perspective, being able to approximate player value using machine learning could eventually assist in contract decisions or salary cap planning. Although this project did not yield an effective salary prediction tool, it highlighted several important data science principles including the challenge of modeling economic outcomes using incomplete or noisy inputs.

REFERENCES

[1] Kaggle, 'NFL Player Salaries and Statistics', https://www.kaggle.com/datasets/adhurimquku/nfl-player-salaries-and-statistics

[2] Spotrac, 'NFL Salary Rankings (2017)', https://www.spotrac.com/nfl/rankings/_/year/2017

[3] Pro Football Reference, '2017 NFL Player Stats', https://www.pro-football-reference.com/years/2017/

[4] NFL.com, 'Player Stats', https://www.nfl.com/stats/player-stats/

[5] Scikit-learn, https://scikit-learn.org/stable/