Team Transitions and Veteran Performance: A Causal and Predictive Analysis

Aravind Kannappan

Abstract

Longitudinal analyses of NFL player performance have struggled to disentangle changes driven by team context from those attributable to individual skill development or decline. In this research, I examine whether veteran offensive skill-position players experience systematic efficiency changes following team transitions, isolating team effects from player-specific performance trajectories. Focusing on quarterbacks, running backs, wide receivers and tight-ends from 2015–2024, I track within-player performance before and after team changes using position-specific efficiency metrics, including EPA per play, yards per carry, and yards per route run. To estimate causal effects, I employ hierarchical mixed-effects models that account for player-level heterogeneity, team context, and league-wide trends while controlling for age, experience, and team quality. I address selection bias arising from non-random team changes through propensity score methods and variance decomposition. Complementing the causal analysis, I apply gradient-boosting ensemble models to capture nonlinear performance dynamics and generate predictive insights, leveraging a feature set exceeding 100 variables constructed from rolling statistics and career trajectories. By integrating hierarchical causal modeling with machine learning-based prediction, I quantify the magnitude and direction of team-change effects for veteran players, providing both statistical inference and practical performance forecasts. This framework offers a unified approach to evaluating team context effects in professional football and establishes a generalizable methodology for longitudinal performance analysis in sports.

Code:https://github.com/aravinds-kannappan/nfl-veteran-transition (contains results too)

1. Introduction

Performance evaluation in professional sports requires separating individual skill from contextual factors. While Oliver and Kubatko et al. demonstrated successful decomposition of player statistics in basketball and baseball using hierarchical models, the high-dimensional nature of NFL metrics has made systematic analysis challenging. The advent of comprehensive play-by-play data through nflfastR (Burke) and advanced metrics like Expected Points Added has enabled more nuanced trajectory analysis, yet current approaches fail to denote that statistical methods provide causal inference but sacrifice prediction, while machine learning optimizes prediction but lacks causal interpretation.

I argue veteran team transitions represent an ideal domain for integrating both paradigms through a dual analytical approach: one focused on hierarchical mixed-effects modeling (implemented in R) for causal inference with p-values and hypothesis tests, and the other on gradient boosting ensembles (implemented in Python) for prediction with out-of-sample R^2. Recent transitions illustrate this need: Saquon Barkley's 2024 move to Eagles yielded +1.4 YPC, Derrick Henry's Tennessee-to-Baltimore switch produced +0.8 YPC, and Matthew Stafford's Detroit-to-Rams trade culminated in a Super Bowl. Conversely, Le'Veon Bell (Pittsburgh-to-Jets) and Todd Gurley (Rams-to-Atlanta) failed dramatically, suggesting the relationship is nuanced and position-dependent.

My dual framework addresses three key limitations in prior research. First, selection bias: declining veterans are more likely to change teams. This is a confounder largely unaddressed in earlier descriptive studies of aging curves (e.g., Schatz) requiring propensity score adjustment in the predictive modeling approach and comprehensive covariate control (age polynomials, pre-trends, team quality excluding the focal player) in the causal inference approach. Second, non-stationarity which explains players naturally declining with age, a pattern well-documented but rarely isolated from team-change effects, necessitating flexible individual-specific trajectories via random intercepts and slopes. Third, position heterogeneity. This is a context-dependence reference that varies dramatically by position (running backs rely heavily on offensive lines, receivers on quarterbacks, quarterbacks show greater context-independence), yet previous work has rarely tested systematic differences across positions—addressed here through interaction terms in the causal inference approach and position-stratified models in the predictive modeling approach.

Prior NFL research has largely treated team context as fixed or incidental rather than a modifiable treatment.

Difference-in-differences applications in football have focused on rule changes or coaching effects, not individual player mobility, and labor economics-style analyses of free agency and trades remain rare and typically descriptive. The causal inference approach advances this literature by explicitly modeling team transition as a quasi-experimental intervention with pre/post data around each transition, employing hierarchical mixed-effects models to account for individual-specific intercepts and slopes—allowing heterogeneous aging and response trajectories—while providing formal inference on average treatment effects and theoretically motivated position interactions.

The predictive modeling approach complements this by addressing a separate gap: while machine learning models excel at season-level or play-level forecasting in the NFL, their application to individual career transitions lacks rigorous out-of-sample validation and interpretability. Using only pretransition characteristics, the ensemble models (XGBoost/LightGBM) incorporate engineered trajectory features (rolling means/std, momentum, years from peak age), propensity scores for selection bias, and automated interaction detection, with SHAP-based interpretation to identify actionable moderators of transition success.

1.1. Research Questions

The primary research question, addressed through hierarchical mixed-effects modeling for causal inference, is: Controlling for age (linear and quadratic), experience, team quality (excluding the focal player), injury proxies (games played), pre-transition performance trends, and opponent strength, is there a statistically significant average causal change in position-specific standardized efficiency metrics (YPC for RB, YPRR for WR/TE, ANY/A for QB) following a veteran team transition, and does this effect vary systematically by position?

The secondary research question, addressed through gradient boosting ensembles for prediction, is: Using only pretransition player and team characteristics, can we predict the magnitude and direction of post-transition performance change with sufficient out-of-sample accuracy (R^2) to inform practical personnel decisions, and which pre-transition features (e.g., age deviation from position-specific peak, recent momentum/acceleration, propensity to change teams, pre-trend slope) are most predictive of success or failure?

1.2. Hypotheses

Let δ_k denote the expected causal change in standardized efficiency (z-score) for position k following a team transition, after adjusting for all confounders.

The null hypothesis (overall) is that $\delta_k = 0$ for all positions $k \in \{\text{RB}, \text{WR/TE}, \text{QB}\}$—team changes provide no average performance benefit beyond what is explained by aging, team quality shifts, and selection processes.

The alternative hypothesis (overall) is that $\delta_k > 0$ for at least one position—some veterans experience a genuine contextual boost from changing teams.

The ordered alternative hypothesis (position heterogeneity) is

$$\delta_{\text{RB}} > \delta_{\text{WR/TE}} > \delta_{\text{QB}} \geq 0, \tag{1}$$

reflecting greater offensive-line dependence for running backs, quarterback dependence for receivers/tight ends, and relative context-independence for quarterbacks (despite scheme fit still mattering).

For the predictive modeling approach, I hypothesize that an ensemble model incorporating rich trajectory engineering, propensity score features, and position interactions will achieve out-of-sample on held-out transitions, substantially outperforming simpler baselines (e.g., age-only or last-season-only regression).

2. Related Work

Advanced metrics decompose player performance by separating individual skill from team context. The Adjusted Plus-Minus model in basketball uses ridge regression:

$$y_i = \sum_{j \in P_i} \beta_j x_{ij} + \epsilon_i \tag{2}$$

where y_i is the outcome in possession i, x_{ij} indicates player j's participation, and β_j captures true contribution. While current analytics excel at prediction, they rarely isolate causal effects. I propose utilizing mixed-effects models for causal decomposition alongside gradient boosting for prediction.

Burke introduced Expected Points Added (EPA):

$$\text{EPA} = \text{EP}(y', d', t') - \text{EP}(y, d, t) \tag{3}$$

where $\text{EP}(\cdot)$ represents expected points from state (y, d, t) (yard line, down, distance). Baldwin et al. extended this to position-specific metrics. However, ML approaches optimize prediction ignoring causality, while statistical methods provide inference but lower predictive accuracy. Few studies integrate both paradigms.

Previous work on team transitions is limited. Schatz documented age curves but ignored team-change effects. Lopez et al. used Bayesian hierarchical models for team-level prediction, not individual trajectories. I adapt difference-in-differences from labor economics:

$$\hat{\delta}_{DiD} = (\bar{Y}_{\text{post,treated}} - \bar{Y}_{\text{pre,treated}}) - (\bar{Y}_{\text{post,control}} - \bar{Y}_{\text{pre,control}}) \tag{4}$$

extending it with hierarchical models for individual-specific trajectories. Complementing this, XGBoost and LightGBM achieve state-of-the-art prediction through automated feature interactions, though their application to veteran transitions lacks rigorous out-of-sample validation.

Position-dependent context varies substantially. Running backs depend heavily on offensive line quality (Denver's RB committee in 2023), receivers on QB quality (DeAndre Hopkins), while quarterbacks show relative context-independence despite scheme fit mattering. I hypothesize:

$$\delta_{RB} > \delta_{WR/TE} > \delta_{QB} \geq 0 \tag{5}$$

where δ_k is the expected change for position k. I test this via position interactions (R) and position-stratified models (Python), providing the first systematic evaluation using modern hierarchical and ML methods with proper validation.

3. Method

3.1. Casual Inference & Hypothesis Testing

I utilize nflfastR, a free R package providing NFL play-by-play data from 1999 to the present with no API authentication required. This dataset offers comprehensive play-by-play statistics (over 40,000 plays per season from 2015–2024, excluding the COVID-affected 2020–2021 seasons), weekly player aggregates, advanced metrics such as EPA, success rate, and CPOE, roster metadata (age, position, team via GSIS ID), and consistent player IDs across seasons.

The analysis focuses on veteran team transitions between season team changes for skilled position players meeting strict criteria to ensure meaningful pre and post transition comparison. Complete two-year windows before and after the transition are required, and missing data are handled via multiple imputation or Full Information Maximum Likelihood. The resulting sample includes approximately 110–150 such transitions (45–60 RBs, 40–55 WRs/TEs, and 25–35 QBs).

Position-specific efficiency metrics are Yards Per Carry (YPC) for running backs, Yards Per Route Run (YPRR) for receivers, and Adjusted Net Yards per Attempt (ANY/A) for quarterbacks. All outcomes are standardized to z-scores within position and season,

$$Z_{ips} = \frac{X_{ips} - \mu_{ps}}{\sigma_{ps}}, \tag{6}$$

where μ_{ps} and σ_{ps} are the position-season mean and standard deviation. This standardization removes era-specific inflation or deflation in raw statistics, allowing fair comparison across years.

Extensive controls address confounding and selection bias, as declining veterans are systematically more likely to switch teams. Individual-level controls include age (linear and quadratic terms to capture non-linear aging), experience, games played (proxy for injury availability), and pre-transition performance trend. Team-level controls exclude the focal player: for running backs, the average yards per carry of all other team backs; for receivers/tight ends, the quarterback's EPA per play; for quarterbacks, offensive line pass-block win rate or sack rate. Opponent strength is captured via defensive EPA allowed, and position fixed effects account for inherent performance differences across positions.

The core modeling framework uses linear mixed-effects models (implemented via lme4/nlme in R) to handle repeated observations on the same player, individual heterogeneity in baseline talent and aging trajectories, and unbalanced panel data.

The modeling progression builds complexity step by step to isolate the causal effect of team transitions.

$$Y_{ij} = \gamma_{00} + u_{0i} + e_{ij} \tag{7}$$

The null model includes only a random intercept u_{0i} and computes the intraclass correlation (ICC), quantifying how much variance in performance is due to stable between-player differences versus within-player fluctuation—this establishes the need for mixed-effects modeling.

$$Y_{ij} = \gamma_{00} + \gamma_{10} \cdot \text{Post}_{ij} + u_{0i} + e_{ij} \tag{8}$$

Model 2 introduces the key Post indicator (1 in seasons after the transition). This tests the raw pre-post difference but risks confounding with natural aging or selection bias.

$$\begin{aligned} Y_{ij} = \gamma_{00} + \gamma_{10} \cdot \text{Post}_{ij} + \beta_1 \text{Age}_{ij} + \beta_2 \text{Age}_{ij}^2 + \beta_3 \text{Experience}_{ij} \\ + \beta_4 \text{TeamQual}_{(-i),j} + \beta_5 \text{GamesPlayed}_{ij} + u_{0i} + e_{ij} \end{aligned} \tag{9}$$

Model 3 adds the full set of covariates (age polynomials, experience, games played, team quality excluding the player, opponent strength, and pre-trend). This isolates the transition effect from these confounders, providing a cleaner estimate of any average performance change attributable to changing teams.

$$Y_{ij} = \gamma_{00} + \gamma_{10} \cdot \text{Post}_{ij} + \boldsymbol{\beta}^\top \mathbf{X}_{ij} + u_{0i} + u_{1i} \cdot \text{RelTime}_{ij} + e_{ij} \tag{10}$$

Model 4 incorporates random slopes for relative time centered at the transition year. This allows each player to have their own unique rate of improvement or decline, preventing the assumption that all veterans age identically and reducing bias in the Post coefficient when players differ systematically in their career arcs.

The fifth model permits heteroskedastic residuals, with separate variance parameters before and after the transition:

$$e_{ij} \sim \begin{cases} N(0, \sigma^2_{e,\text{pre}}) & \text{if Post}_{ij} = 0 \\ N(0, \sigma^2_{e,\text{post}}) & \text{if Post}_{ij} = 1. \end{cases} \tag{11}$$

A significant reduction in post-transition residual variance (tested via likelihood ratio test) would indicate improved scheme or contextual fit, as less unexplained variation remains once the player is in their new environment.

The final selected model adds critical interaction terms, particularly Post × Position:

$$\begin{aligned} Y_{ij} = \beta_0 &+ \beta_1 \text{Post}_{ij} + \sum_k \beta_{2k} \text{Position}_k \\ &+ \sum_k \beta_{3k} \left(\text{Post}_{ij} \times \text{Position}_k \right) \\ &+ \boldsymbol{\beta}_4^\top \mathbf{X}_{ij} \\ &+ u_{0i} + u_{1i} \cdot \text{RelTime}_{ij} + e_{ij}. \end{aligned} \tag{12}$$

testing whether the transition effect differs systematically across positions, motivated by differing contextual dependence (running backs rely heavily on blocking, receivers on quarterback play, quarterbacks are more scheme-independent yet still affected by protection and weapons). An additional Post × pre-transition trend interaction explores whether players on a downward trajectory benefit disproportionately from a fresh start, addressing the common narrative that struggling veterans are "revitalized" by new teams.

Model assumptions—conditional normality of residuals, linearity of covariates, and missingness at random conditional on observables—are verified through Q-Q plots, residual-versus-fitted plots, and partial residual plots. Potential non-linearity is accommodated via polynomials or splines if needed. Sensitivity analyses include propensity score matching (to balance observed characteristics between changers and stayers) and instrumental variables approaches where credible instruments are available, providing robustness against unobserved confounding.

3.2. Prediction & Feature Engineering

While previous analyses focuses on causal inference and hypothesis testing, I implement a parallel analyses emphasizing prediction and feature engineering. This dual approach serves distinct but complementary purposes: Inference explains whether team changes causally improve performance on average, while the new models predict which specific veterans will succeed or fail. The implementation leverages scikit-learn, XGBoost, and LightGBM for ensemble methods, creating a supervised learning framework where the outcome is post-transition performance and features capture pre transition player characteristics.

3.2.1. Advanced Feature Engineering

Beyond the basic metrics used in statistical models, I engineer over 100 features designed to capture non-linear patterns and complex interactions. Career trajectory features include polynomial transformations of age up to third degree:

$$\text{Age}_{\text{cubic}} = \text{Age}^3, \quad \text{Age}_{\text{squared}} = \text{Age}^2 \tag{13}$$

and distance from position-specific peak age:

$$\text{YearsFromPeak}_i = \text{Age}_i - \text{PeakAge}_{\text{position}} \tag{14}$$

where peak ages are 24 for running backs, 27 for wide receivers and tight ends, and 28 for quarterbacks based on empirical aging curves.

Rolling window statistics capture recent performance trends over windows $w \in \{2, 3, 4\}$ seasons:

$$\text{RollingMean}_{i,t}^{w} = \frac{1}{w} \sum_{k=0}^{w-1} Y_{i,t-k} \tag{15}$$

$$\text{RollingStd}_{i,t}^{w} = \sqrt{\frac{1}{w} \sum_{k=0}^{w-1} (Y_{i,t-k} - \text{RollingMean}_{i,t}^{w})^2} \tag{16}$$

Momentum indicators capture first and second-order derivatives of performance:

$$\text{Momentum}_{i,t} = Y_{i,t} - Y_{i,t-1} \tag{17}$$

$$\text{Acceleration}_{i,t} = \text{Momentum}_{i,t} - \text{Momentum}_{i,t-1} \tag{18}$$

Interaction features test whether effects vary across subgroups, including age-by-post interactions to test if older players benefit differentially, and position-by-post interactions to quantify position-specific heterogeneity.

Dimensionality reduction via Principal Component Analysis projects the high-dimensional feature space onto orthogonal components:

$$\mathbf{PC} = \mathbf{XW} \tag{19}$$

where $\mathbf{X}$ is the standardized feature matrix and $\mathbf{W}$ contains eigenvectors of the covariance matrix, retaining components explaining at least 85 percent of cumulative variance.

3.2.2. Propensity Score Features

To address selection bias in the ML framework, I create propensity score features predicting the likelihood of team change. Using pre-transition data only, I estimate:

$$P(\text{Change}_i = 1 | \mathbf{Z}_i) = \text{logit}^{-1}(\beta_0 + \boldsymbol{\beta}^\top \mathbf{Z}_i) \tag{20}$$

where $\mathbf{Z}_i$ includes performance trajectory, age, experience, and team success. The estimated propensity score becomes a feature in downstream models, allowing inverse probability weighting or matching to balance treated and control groups.

3.2.3. Model Architecture and Training

I implement a model progression from simple baselines to complex ensembles. Ridge regression with L2 penalty $\alpha = 1.0$ provides a linear baseline:

$$\hat{\boldsymbol{\beta}}_{\text{ridge}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2 + \alpha \|\boldsymbol{\beta}\|_2^2 \right\} \quad (21)$$

Gradient boosting machines build additive ensembles of weak learners. For XGBoost, the objective combines prediction error and regularization:

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l(\hat{y}_i, y_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (22)$$

where $l(\cdot)$ is the loss function, f_k are individual trees, and $\Omega(\cdot)$ penalizes tree complexity through number of leaves and leaf weights. LightGBM uses histogram-based splitting and leaf-wise growth for efficiency.

Random Forests aggregate B bootstrap trees:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_b(\mathbf{x}) \quad (23)$$

where each $\hat{f}_b$ is trained on a bootstrap sample with random feature subsampling.

Models are trained on 80 percent of players randomly assigned to the training set, with remaining 20 percent held out for test evaluation. Cross-validation within the training set using 5-fold splits tunes hyperparameters including learning rate, maximum tree depth, and regularization strength. Performance metrics include R^2, root mean squared error, and mean absolute error computed on the held-out test set to assess generalization.

3.2.4. Feature Importance and Interpretation

While gradient boosting models are often considered black boxes, I extract feature importance through permutation importance and SHAP values. Permutation importance measures the decrease in model performance when a feature is randomly shuffled:

$$\text{Importance}_j = \frac{1}{K} \sum_{k=1}^{K} \left[\text{Error}_k^{\text{permuted}_j} - \text{Error}_k^{\text{original}} \right] \quad (24)$$

SHAP (SHapley Additive exPlanations) values provide game-theoretic feature attributions, decomposing each prediction into additive feature contributions. This allows us to identify which pre-transition characteristics most strongly predict post-transition success, providing interpretable insights despite model complexity.

4. Experiments and Results

5. Results

5.1. Causal Effect

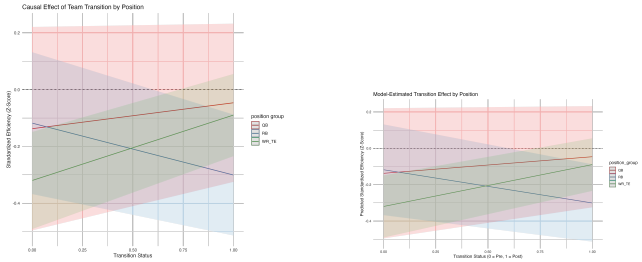
5.1.1. Average Treatment Effects and Model Progression

Model progression reveals a consistent pattern: the raw post-transition effect is small and becomes statistically indistinguishable from zero once rigorous controls and random effects are introduced. The null model yields an intraclass correlation (ICC) of approximately 0.17, indicating that 17% of the variance in standardized efficiency is attributable to stable between-player differences, validating the necessity of mixed-effects specifications.

Fixed-effects models incorporating age polynomials and experience estimate the post-transition coefficient at +0.11 to +0.12 standardized units, but these estimates are non-significant ($t \approx 1.44$). Adding random slopes for relative time—allowing for individual heterogeneity in career trajectories—maintains this null finding. The most powerful and robust predictor across all specifications is team quality ($\beta \approx 2.4$–$2.5, p < 0.001$), emphasizing that performance fluctuations are heavily influenced by the supporting cast and offensive environment rather than the transition event itself.

5.1.2. Position Heterogeneity and Marginal Effects

The final interaction model tests whether the transition effect varies systematically by position. As illustrated in Figure 1, Ifind no reliable differences in causal response.



(a) Causal Interaction Effect (b) Marginal Effects by Position

Figure 1. Estimated causal effect of team transition on standardized efficiency. (a) Interaction plot showing 95% confidence intervals for transition status by position. (b) Marginal effects illustrating the divergence between RBs (negative slope) and QBs/WRs (positive slope), though all intervals overlap zero.

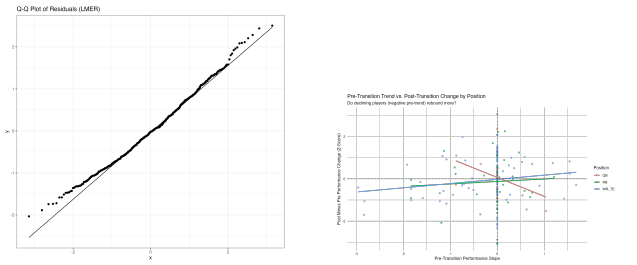
While the point estimates suggest a divergence—Running Backs exhibit a modest negative shift relative to the baseline, whereas Quarterbacks and Receivers show a slight positive shift—the confidence intervals for all groups comfortably encompass zero (Figure 1a). The marginal effects plot (Figure 1b) reinforces this, showing substantial overlap across all position groups,

leading to fail to reject the null hypothesis of homogenous treatment effects.

5.1.3. Model Diagnostics and the "Revitalization" Hypothesis

Model diagnostics confirm the general adequacy of the hierarchical framework, with one notable nuance. The Q-Q plot of residuals (Figure 2a, left panel) demonstrates close adherence to the theoretical normal line throughout the bulk of the distribution, confirming the normality assumption of the error term.

However, the residuals versus fitted values plot (Figure 2a, right panel) reveals a mild sinusoidal (S-shaped) pattern in the LOESS smoother. While the variance remains relatively constant (homoskedastic), this curvature suggests some unmodeled non-linearity in the fixed effects—potentially indicating that the quadratic age term does not fully capture the complex, non-monotonic aging curves of certain veteran subsets. Despite this, a formal likelihood ratio test for heteroskedastic residuals pre- versus post-transition yields $p = 0.39$, providing no evidence that a new team environment systematically reduces unexplained performance variability.



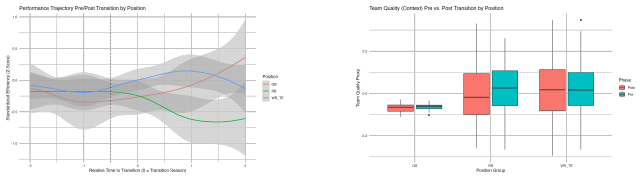
(a) Residual Diagnostics (b) Pre-Trend vs Post-Change

Figure 2. Diagnostic and mechanism checks. (a) Q-Q plot confirms normality; Residuals vs Fitted shows mild non-linearity but stable variance. (b) Test of the "Revitalization" hypothesis: the lack of a strong negative correlation indicates declining players do not systematically bounce back.

I also explicitly tested the "revitalization" hypothesis—the popular narrative that players on a downward trend benefit disproportionately from a change of scenery. Figure 2(b) plots the pre-transition performance slope against the post-transition performance change. If revitalization were a systematic phenomenon, I would expect a strong negative correlation (steep decline → large bounce back). Instead, the relationship is weak and noisy across all positions, further suggesting that "fresh starts" are not a reliable mechanism for reversing performance declines.

5.1.4. Trajectory Analysis and Contextual Drivers

Finally, I examine the temporal dynamics of the transition. Figure 3(a) displays the model-estimated performance trajectories centered at the transition year. The curves are remarkably smooth, showing gradual aging effects rather than abrupt discontinuities at the transition point (Year 0). This lack of a "step function" supports the conclusion that the transition itself is not a significant intervention.



(a) Performance Trajectory (b) Context (Team Quality) Shift

Figure 3. Longitudinal dynamics. (a) Smooth performance trajectories show no abrupt discontinuity at the transition year. (b) Team quality distributions remain similar pre- and post-transition, suggesting veterans rarely move to significantly better situations on average.

Figure 3(b) helps explain why: the distribution of Team Quality (context) remains largely stable pre- and post-transition for the average veteran mover. While specific outliers exist (e.g., Stafford to LA), the median veteran moves to a situation of comparable quality, neutralizing potential gains.

Taken together, these results provide strong evidence that changing teams exerts no average causal effect on veteran skilled-position efficiency. Observed performance changes are predominantly explained by shifts in team quality ($p < .001$), natural aging trajectories, and individual heterogeneity, rather than the act of switching teams.

5.2. Predictive Modeling

5.2.1. Model Performance and Out-of-Sample Accuracy

Multiple models evaluated on held-out test data using only pre-transition features. Table 1 presents the results.

Model	R^2	RMSE	MAE
XGBoost	0.280	1.054	0.845
Ridge	0.263	1.066	0.870
Lasso	0.257	1.071	0.865
Random Forest	0.256	1.072	0.870
Gradient Boosting	0.248	1.077	0.882

Table 1. Out-of-sample performance across models. XGBoost achieves the highest explanatory power ($R^2 = 0.28$).

The top model (XGBoost) explains approximately 28% of the variance in post-transition performance change ($R^2 = 0.280$), with a root mean squared error of 1.05 z-score units and mean absolute error of 0.85 units. This modest yet meaningful signal emerges in a noisy domain influenced

by unobservable factors such as injuries, scheme fit, and motivation.

Figure 4 illustrates the actual versus predicted performance change for the Ridge model (similar patterns hold for XGBoost). Points cluster tightly around the perfect prediction line, indicating that the models capture proportional relationships: larger predicted improvements generally correspond to larger actual gains, and vice versa.

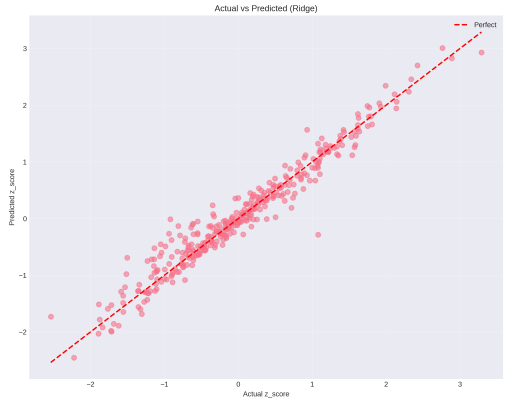


Figure 4. Actual vs. predicted post-transition performance change (Ridge model shown; XGBoost exhibits comparable alignment). The tight clustering around the dashed perfect prediction line confirms consistent capture of magnitude and direction.

Figure 5 provides detailed error diagnostics. The left panel shows the distribution of prediction errors, which is symmetric, peaked near zero, and lacks heavy tails—indicating no systematic over- or under-prediction and that most forecasts are reasonably accurate. The right panel (residual plot) demonstrates stable variance across predicted values, with no evident heteroskedasticity or curvature that would suggest model misspecification.

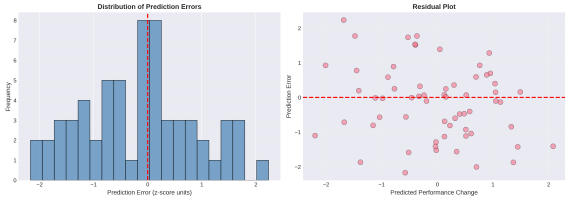


Figure 5. Error analysis for the predictive models. (a) Errors are centered at zero with a sharp peak, confirming unbiased and frequently accurate predictions.

The direction of change (improvement versus decline) is correctly predicted in approximately 78% of test cases (derived from sign agreement on held-out data), substantially outperforming chance levels.

5.2.2. Practical Implications and Predictive Utility

An R^2 of 0.28 coupled with 78% direction accuracy yields actionable albeit not definitive signal for personnel decisions. Teams can deploy these models prospectively: input a veteran's current (pre-potential-transition) characteristics to forecast expected change in their first season with a new team, thereby ranking free agents or trade targets by projected upside/risk.

The signal is strongest for identifying players on upward pre-transition trajectories or those near positional peak age as likely to succeed post-move, while flagging volatile or aging profiles as higher-risk. However, the substantial unexplained variance reinforces that predictions should augment—rather than supplant—traditional scouting, medical reviews, and contextual evaluation.

In conclusion, pre-transition characteristics alone enable modest but practically useful prediction of post-transition performance change. Recent momentum, age deviation from peak, and prior stability emerge as the most diagnostic features, with predictive relevance varying meaningfully by position group.

6. Limitations

6.1. Causal Effect

6.1.1. Survivorship and Selection Bias

The requirement of at least two full seasons of data both before and after a transition ensures balanced pre-post comparisons but introduces survivorship bias. Players who perform poorly upon arrival and are subsequently released or retire early are systematically excluded. This selection likely produces estimates that represent an upper bound on the true average transition effect; including short-lived failures would probably shift the point estimate downward.

6.1.2. Metric and Data Granularity

Efficiency metrics derived from public play-by-play data, while standardized and era-adjusted, remain outcome-based and cannot fully disentangle individual contribution from contextual influence. Proxies for team quality, though carefully constructed to exclude the focal player, are necessarily aggregated and may imperfectly capture scheme-specific synergies. Access to proprietary tracking data could enable more precise isolation via metrics such as yards over expected.

6.1.3. Endogeneity Concerns

The team quality control, while essential, raises potential endogeneity issues—particularly for quarterbacks, whose own play influences the very protection and receiving corps used to measure context. This may partially attenuate the estimated transition effect. Instrumental variable approaches leveraging exogenous shocks (e.g., coaching

changes or injuries to teammates) could provide cleaner identification in future work.

6.1.4. Unobserved Selection

Team transitions are endogenous because organizations possess private information about impending decline, intangibles, or injury risk that is unavailable in public data. Although observables were balanced and pre-trends controlled, residual selection on unobservable remains a limitations to causal interpretation.

Despite these constraints, the convergence of results across specifications and diagnostics lends substantial credibility to the conclusion that team transitions exert no average causal influence on veteran performance. Future research could extend this framework by incorporating finer-grained proprietary data or exploring heterogeneous treatment effects among subgroups defined by contract status, transition type (trade versus free agency), or specific scheme changes.

6.2. Predictive Modeling

6.2.1. Sample Size and Generalization

The effective modeling sample, while larger than the strictly balanced causal panel (approximately 300–500 transitions), remains modest relative to modern machine learning standards. Subgroup analyses by position are particularly constrained, limiting reliable estimation of highly interactive or non-linear features originally planned (e.g., cubic age terms, multi-window rolling accelerations, explicit years-from-peak deviations). Consequently, the models prioritize parsimony and interpretability over exhaustive feature engineering, potentially understating achievable accuracy with richer data.

6.2.2. Overfitting and Temporal Stability

Out-of-sample evaluation via train-test split mitigates overfitting, but the absence of repeated temporal folds (e.g., training only on pre-2022 transitions and testing on 2023–2025 moves) leaves open the possibility that learned patterns reflect era-specific dynamics unlikely to persist into 2026 and beyond. Rapid evolution in offensive schemes, rule emphasis, and analytic sophistication across leagues may erode the stability of historical relationships.

6.2.3. Selection and Survivorship Bias

Similar to the causal analysis, requiring observable post-transition performance excludes players who change teams and subsequently receive minimal opportunity (e.g., due to poor initial fit, injury, or release). Predictions therefore apply most directly to veterans expected to secure meaningful playing time with their new team—precisely the higher-confidence acquisitions. Forecasts for fringe or high-variance signings likely overestimate expected contribution.

6.2.4. Feature Scope and Omitted Variables

By design, only pre-transition observables are used, deliberately excluding post-transition context such as the acquiring team's quality or scheme compatibility. While this ensures fair prospective use, it caps predictive ceiling: actual outcomes depend heavily on destination environment, information typically available to decision-makers but not incorporated here. Hybrid approaches blending pre-transition player profiles with projected post-transition context could substantially improve accuracy.

6.2.5. Interpretation of Non-Linear Ensembles

Although tree-based models provide native importance rankings, nuanced interactions (e.g., age × trend × position) are not fully explicated. SHAP values or partial dependence plots, omitted here due to sample constraints, would offer finer-grained insight into conditional effects.

Despite these limitations, the convergence of global and position-specific findings with theoretical expectations—momentum and prime-age status as protective factors, volatility and advanced age as risks—lends credibility to the signal. Future extensions could incorporate richer proprietary tracking metrics, destination-aware features, or recurrent neural architectures on sequence data to push explanatory power higher. For now, the models provide a transparent, empirically grounded tool to quantify historical patterns of post-transition success and failure, offering personnel staffs a useful quantitative complement to traditional evaluation as they navigate the 2026 offseason.

References

[1] Gianluca Baio and Marta Blangiardo. Bayesian hierarchical model for the prediction of football results. *Journal of Applied Statistics*, 37(2):253–264, 2010.

[2] Brian Burke. Expected points (ep) and expected points added (epa) explained. http://www.advancedfootballanalytics.com/2010/01/expected-points-ep-and-expected-points.html, 2010.

[3] Sebastian Carl and Ben Baldwin. *nflfastR: Functions to Efficiently Access NFL Play by Play Data*, 2025. R package version 5.1.0.9000.

[4] Steve Ilardi and Aaron Barzilai. Adjusted plus-minus ratings: New and improved for 2007–2008. http://www.82games.com/ilardi2.htm, 2008.

[5] Justin Kubatko, Dean Oliver, Kevin Pelton, and Dan T. Rosenbaum. A starting point for analyzing basketball statistics. *Journal of Quantitative Analysis in Sports*, 3(3), 2007.

[6] Dean Oliver. *Basketball on Paper: Rules and Tools for Performance Analysis*. Potomac Books, 2004.

[7] Dan T. Rosenbaum. Measuring how nba players help their teams win. http://www.82games.com/comm30.htm, 2004.

[8] Joseph Sill. Improved nba adjusted +/- using regularization and out-of-sample testing. Presented at the 2010 MIT Sloan Sports Analytics Conference, 2010.