Research Article

Predicting NFL Head Coach Tenure Using Ordinal Classification

https://doi.org/10.1515/jqas-YYYY-XXXX
Received Month DD, YYYY; revised Month DD, YYYY; accepted Month DD, YYYY

Abstract: What if NFL teams could predict which head coach candidates would be successful? This project predicts the tenure classification of NFL head coach hires using statistics available at the time the hire was made. Using 150 engineered features spanning coach experience, historical team performance during prior coordinator and head coaching roles, and hiring team context, we implement ordinal classification via the Frank-Hall binary decomposition method with XGBoost base classifiers. This approach respects the natural ordering of tenure classes (short, medium, long) and penalizes distant misclassifications more heavily than adjacent errors. The ordinal model, optimized using Quadratic Weighted Kappa (QWK), achieves a QWK of 0.754, mean absolute error of 0.307, and 98.4% adjacent accuracy on held-out test data, which outperforms a standard multiclass approach on all ordinal metrics and achieves a 5.3× improvement over human baseline decisions. SHAP-based feature importance analysis reveals that defensive metrics are approximately 1.7 times more predictive than offensive metrics, with this pattern consistent across eras and coach backgrounds. These findings provide actionable insights for NFL teams evaluating head coaching candidates.

Keywords: head coach, tenure prediction, sports analytics, NFL

1 Introduction

What if the Denver Broncos could have avoided hiring Nathaniel Hackett? Or if the Raiders had not hired Josh McDaniels? Certainly, these teams would be in a different position today if they had hired different candidates. But more broadly, what if NFL teams could predict which head coach candidates would be successful? That is the aim of this project.

The stakes of head coaching decisions are substantial. NFL teams invest significant resources in coaching searches, and a failed hire carries consequences beyond the win-loss record: organizational instability, wasted draft capital on players who don't fit the new system, and the opportunity cost of not hiring a more successful candidate. Since 2000, approximately half of all NFL head coaching hires have lasted two years or fewer, suggesting that teams systematically struggle to identify candidates who will succeed in the role. Yet despite this high failure rate, the hiring process remains largely subjective, driven by interviews, references, and gut instinct rather than data-driven analysis.

Tenure serves as a more robust measure of coaching success than win-loss percentage because it implicitly accounts for organizational context and expectations. A coach who maintains a .450 winning percentage with a rebuilding franchise may be considered successful and retain their position, while a coach with the same record on a team expected to contend may be fired after two seasons. Tenure captures this reality: it reflects not just on-field performance but also the organization's assessment of whether the coach is meeting expectations given the circumstances. By predicting tenure rather than wins, the model learns patterns that generalize across different team situations without requiring explicit modeling of each organization's unique context.

This project addresses this gap by developing a machine learning model to predict head coach tenure classification using only information available at the time of hiring. By analyzing 150 engineered features spanning coach experience, historical team performance during prior coordinator and head coaching roles, and hiring team context, the model provides an objective, data-driven assessment of coaching candidates. The ordinal classification approach respects the natural ordering of tenure outcomes, distinguishing between short (1–2 years), medium (3–4 years), and long (5+ years) tenures, while penalizing distant misclassifications more heavily than adjacent errors.

2 Literature Review

At the time of writing, there have been no journal publications that attempt to predict the success of NFL coaching hires through statistical learning techniques. Currently, the NFL is only just beginning to implement artificial intelligence (AI) in play calling prediction (DataRobot, 2020).

There are few papers that examine the impact of individual features on NFL head coaching success. Roach (2016) used a linear regression with seven features to attempt to predict the number of wins of head coaches in their first three years to understand if prior NFL head coaching experience impacts success in position. This paper found that previous head coaching experience had a negative impact on the success of new head coaches. Despite this finding, the model supported an adjusted R^2 of only 0.336. This low value, the lack of regularization, and the small number of features decreases confidence in the study's findings.

Mielke (2007) reviews research in sports economics and suggests that hiring decisions made solely on playing success are unlikely to be optimal given financial (resource) inequality among sports franchises.

3 Methods

Using statistics available at the time of hiring, this project attempts to predict the tenure classification of NFL head coach hires using XGBoost models (Chen and Guestrin, 2016). Raw data was collected by scraping pro-football-reference.com. All data processing, model implementation, and analysis were performed using Python with scikit-learn (Pedregosa et al., 2011).

3.1 Predicting Coach Tenure Classification

The tenure of a coach hire is defined as the number of years the hired coach remains in the same position before being fired, leaving for another role, or retiring. Equation (1) shows the mapping between the coach tenure t (in years) and the three coach tenure classification labels $C(t)$.

$$C(t) = \begin{cases} 0 & \text{if } t \leq 2 \\ 1 & \text{if } 2 < t \leq 4 \\ 2 & \text{if } t > 4 \end{cases} \tag{1}$$

This project groups coach tenures into classes for classification rather than predicting the number of years with a regression as there is little apparent difference between coaches with similar tenures. For example, a coach that remains in position for 15 years is not 50% more successful than a coach who remains in position for 10 years. These different coach classifications are intended to indicate different levels of coaching success based on the number of years they maintain their position.

Importantly, these tenure classes exhibit a natural ordering (Class 0 < Class 1 < Class 2), making ordinal classification more appropriate than standard multiclass methods. Standard multiclass approaches treat all misclassifications equally, but in this domain, predicting Class 0 for a true Class 2 coach is a more severe error than predicting Class 1.

3.1.1 Frank-Hall Ordinal Classification

This project implements ordinal classification using the Frank-Hall binary decomposition method (Frank and Hall, 2001). For K ordinal classes, this approach trains $K-1$ binary classifiers, each predicting the probability that an instance exceeds a given threshold. For our 3-class problem:
- **Classifier 1**: $P(Y > 0)$ — distinguishes Class 0 from Classes 1 and 2
- **Classifier 2**: $P(Y > 1)$ — distinguishes Classes 0 and 1 from Class 2

Class probabilities are then derived from these cumulative probabilities:

$$P(Y = 0) = 1 - P(Y > 0) \tag{2}$$
$$P(Y = 1) = P(Y > 0) - P(Y > 1) \tag{3}$$
$$P(Y = 2) = P(Y > 1) \tag{4}$$

The Frank-Hall method offers several advantages: it works with any base classifier (preserving XGBoost's strengths), produces interpretable probability distributions, and naturally penalizes distant misclassifications.

3.1.2 Evaluation Metrics

This project uses Quadratic Weighted Kappa (QWK) as the primary evaluation metric because it is specifically designed for ordinal classification problems. QWK measures agreement between predicted and true classes while weighting disagreements by their squared distance; a prediction of Class 0 for a true Class 2 instance is penalized four times more heavily than a prediction of Class 1. This property aligns with the practical reality that predicting a short tenure for a coach who achieves long tenure (or vice versa) represents a more consequential error than being off by one class. QWK also served as the optimization target during hyperparameter tuning, ensuring that the model explicitly learns to minimize ordinal prediction errors. Secondary metrics include mean absolute error (average class distance between predictions and truth), adjacent accuracy (proportion of predictions within one class), and macro F1 score (for comparison with standard classification approaches).

3.1.3 Cross-Validation Strategy

To prevent data leakage, this project implements coach-level stratified cross-validation. Since individual coaches may appear multiple times in the dataset (e.g., Bill Belichick was hired as head coach in both 1991 and 2000), all instances for a given coach are kept together in either the training or test set. This ensures the model cannot learn from a coach's prior hiring outcomes when predicting their tenure in a different role.

3.2 Data Description

This project utilizes 150 engineered features for each head coaching hire, organized into five categories as shown in Table 1. Appendix A provides the complete list of features.

Core Experience Features (8 features) capture fundamental coaching background: age at time of hire, number of previous head coaching stints, and years of experience at each level (college position coach, college coordinator, college head coach, NFL position coach, NFL coordinator, and NFL head coach).

Coordinator and Head Coach Statistics (132 features) capture team performance during the coach's prior roles. For each role (OC, DC, HC), 33 statistics are recorded spanning offensive production (points, yards, turnovers), passing efficiency (completions, yards, touchdowns, interceptions, net yards per attempt), rushing performance (attempts, yards, touchdowns), drive metrics (scoring percentage, turnover

Tab. 1: Feature categories and counts

Category	Features	Count	Description
Core Experience	1–8	8	Age, prior HC hires, years at each coaching level
OC Statistics	9–41	33	Team offensive performance during OC tenure
DC Statistics	42–74	33	Opponent offensive performance during DC tenure
HC Statistics	75–107	33	Team offensive performance during HC tenure
HC Opponent Stats	108–140	33	Opponent offensive performance during HC tenure
Hiring Team Context	141–150	10	Hiring team's recent performance metrics

percentage, plays per drive, yards per drive), and situational effectiveness (third-down conversion rate, fourth-down conversion rate, red zone percentage). For defensive coordinators, these statistics reflect opponent performance (i.e., points allowed, yards allowed). For head coaches, both team and opponent statistics are captured, resulting in 66 features.

Hiring Team Context (10 features) capture the state of the team at the time of hiring: winning percentage, points scored, points allowed, yards of offense, yards allowed, yards per play, yards per play allowed, turnovers forced, turnovers committed, and playoff appearances; averaged over the two seasons prior to the hire.

All performance statistics (features 9–150) are normalized using z-scores relative to league averages for each season, enabling meaningful comparisons across eras. For example, a coach whose offense ranked one standard deviation above league average in 1985 is comparable to a coach whose offense ranked one standard deviation above average in 2020.

This project utilizes SVD-based matrix factorization to impute missing values, which occur when coaches lack experience at certain levels (e.g., a first-time head coach has no prior HC statistics). Figure 1 shows the correlation matrix among all 150 features post-imputation. Notable correlation exists within feature categories, particularly among offensive statistics during OC tenure and among statistics during HC tenure, reflecting the interdependence of team performance metrics. Appendix B shows the distribution of coach tenure classifications across all hiring instances.

4 Results

The dataset contains 635 coaching hire instances with known tenure outcomes, featuring 150 engineered features per instance. The class distribution is imbalanced: Class 0 (1–2 years) comprises 49.0% of instances, Class 1 (3–4 years) comprises 26.8%, and Class 2 (5+ years) comprises 24.3%.

Data was split into training (508 instances) and test (127 instances) sets using coach-level stratified sampling to prevent data leakage. Hyperparameters were tuned via 5-fold coach-level cross-validation on the training set (final values shown in Appendix C). All reported metrics are evaluated on the held-out test set.

4.1 Predicting Coach Tenure Classification

4.1.1 Ordinal Classification Model Performance

Table 2 shows the performance metrics for the ordinal XGBoost classifier on the held-out test set (127 instances). The model achieves strong performance across all ordinal metrics.

The ordinal model achieves a quadratic weighted kappa of 0.754, indicating substantial agreement between predictions and true labels while accounting for ordinal distance. The model achieves 98.4% adjacent accuracy, meaning nearly all predictions are within one class of the true label, which is a critical property for ordinal classification.

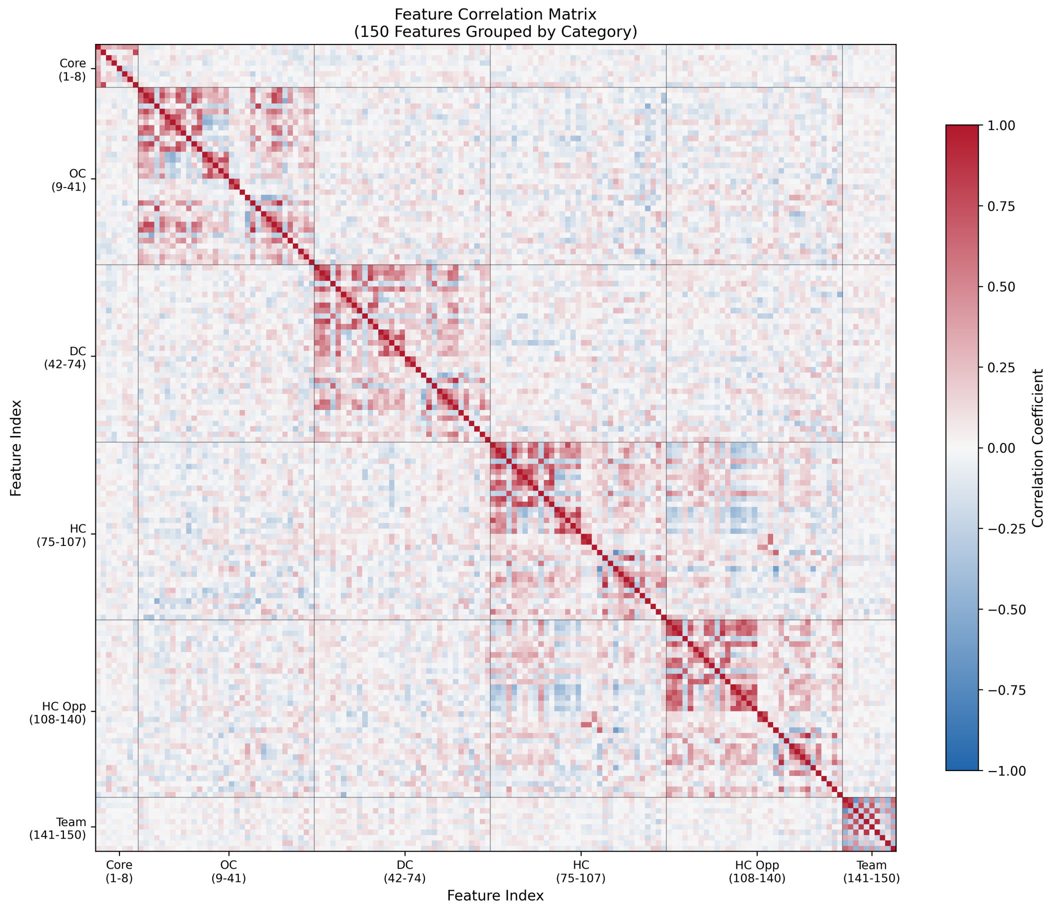

Fig. 1: Feature correlation matrix showing relationships among 150 features grouped by category: Core Experience (1–8), OC Stats (9–41), DC Stats (42–74), HC Stats (75–107), HC Opponent Stats (108–140), and Hiring Team Context (141–150).

Tab. 2: Coach tenure classification prediction results (Ordinal Model)

Metric	Test Set Performance
Mean Absolute Error (MAE)	0.307
Quadratic Weighted Kappa (QWK)	0.754
Adjacent Accuracy (±1 class)	98.4%
Exact Accuracy	72.4%
Macro F1 Score	0.695
AUROC (macro OVR)	0.881
Human Baseline F1[1]	0.130
Model Improvement[2]	5.3×

[1]Assuming all GMs believe their selected HC is Class 2
[2]Model F1 vs. human baseline

4.1.2 Comparison with Standard Multiclass Classification

To validate the ordinal approach, we compare against a standard multiclass XGBoost classifier trained with the same hyperparameters. Table 3 shows that the ordinal model outperforms multiclass on most metrics, particularly those that account for class ordering.

Tab. 3: Ordinal vs. Multiclass model comparison on held-out test set

Metric	Ordinal	Multiclass	Better
MAE	**0.307**	0.402	Ordinal
QWK	**0.754**	0.672	Ordinal
Adjacent Accuracy	**98.4%**	96.9%	Ordinal
Exact Accuracy	**72.4%**	63.0%	Ordinal
Macro F1	**0.695**	0.589	Ordinal
AUROC	**0.881**	0.836	Ordinal
Class 1 F1	**0.581**	0.358	Ordinal (+62.3%)

The ordinal model shows consistent improvement across all metrics, with the most notable improvement in Class 1 (middle class) F1 score. The middle class is typically most difficult to predict because it can be confused with both Class 0 and Class 2; the ordinal model's 62.3% improvement (0.581 vs. 0.358) demonstrates that the Frank-Hall decomposition, combined with QWK-based hyperparameter optimization, substantially helps distinguish the intermediate tenure class.

Figure 2 shows the sorted validation set with corresponding marks for the ground truth values and the predicted values.

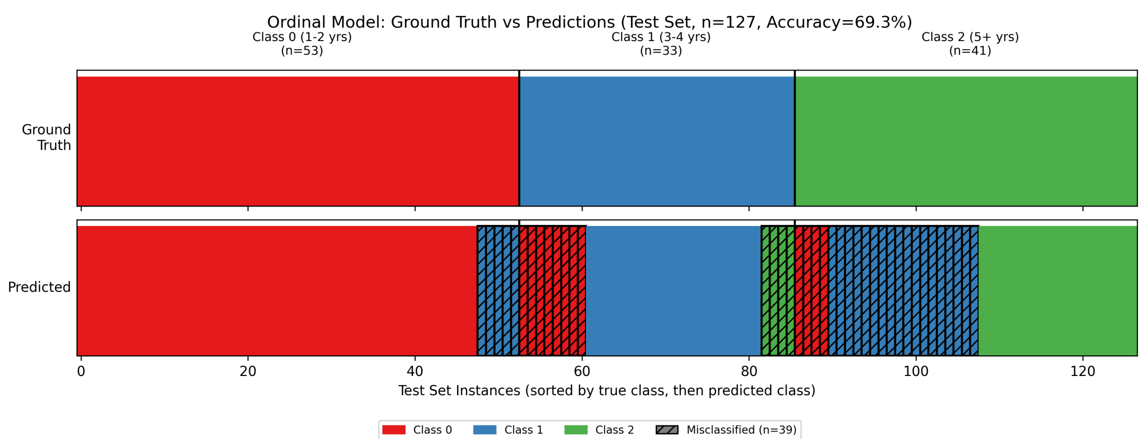

Fig. 2: Ordinal model predictions versus ground truth on the held-out test set. Top row shows true class; bottom row shows predicted class. Hatched bars indicate misclassifications. Instances are sorted by true class, then by predicted class within each true class.

Figure 3 shows the total and average feature importance aggregated by category.

The HC Stats category (66 features) contributes the highest total importance (0.459), reflecting the large number of features capturing prior head coaching performance. This aligns with intuition: coaches who have previously succeeded as head coaches carry demonstrated evidence of their capabilities. DC Stats (0.256) and OC Stats (0.166) contribute less total importance, while Core Experience (0.064) and Hiring Team (0.056) contribute the least due to their smaller feature counts.

When examining average importance per feature, a Kruskal-Wallis test reveals no statistically significant differences across categories ($H = 2.42$, $p = 0.66$). The mean importance values are similar: Core Experience (0.0080), DC Stats (0.0078), HC Stats (0.0069), Hiring Team (0.0056), and OC Stats (0.0050). This finding suggests that the model draws broadly on information from all feature categories rather than relying heavily on any single type of predictor. The predictive signal for tenure classification is distributed across coaching experience, prior performance at multiple levels, and team context.

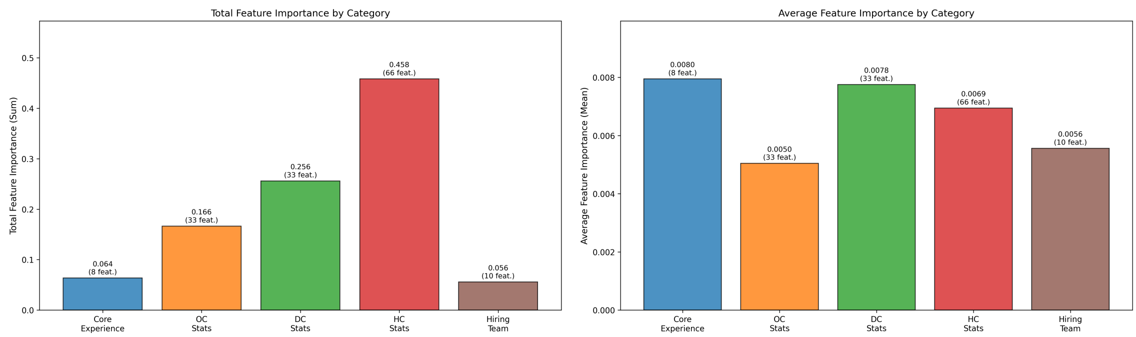

Fig. 3: Feature importance aggregated by category. Left: Total importance (sum across all features in category). Right: Average importance (mean per feature). HC Stats contribute highest total importance due to the large number of features, while average importance per feature is similar across categories.

4.1.3 Predicting the Tenure of Recent Head Coach Hires

Table 4 shows the ordinal model's predictions for coach tenure for the 21 head coaches hired in the last four years. This table also shows the probabilities associated with each class prediction; these probabilities sum to 1, and the class with the greatest probability is the final predicted class. As a reminder, Class 0 represents coaches who remain a head coach for 1–2 years, Class 1 represents coaches who remain a head coach for 3–4 years, and Class 2 represents coaches who remain a head coach for 5+ years.

Seven coaches in the prediction set (marked with *) have prior head coaching stints that appear in the training data. To prevent data leakage, these coaches are predicted using a model retrained with their prior instances excluded. This ensures predictions are based solely on their characteristics at time of hire, not on the model having seen their past outcomes.

Tab. 4: Ordinal classifier coach tenure predictions for 21 recent head coach hires

Coach Name	Year	Pred.	P(C0)	P(C1)	P(C2)
Aaron Glenn	2025	0	56.4%	42.1%	1.5%
Ben Johnson	2025	0	68.8%	30.8%	0.4%
Brian Callahan	2024	0	68.5%	28.9%	2.6%
Brian Daboll	2022	0	77.1%	22.6%	0.3%
Brian Schottenheimer	2025	1	37.4%	62.0%	0.6%
Dan Quinn*	2024	1	6.7%	91.4%	1.9%
Dave Canales	2024	0	55.9%	43.1%	1.0%
DeMeco Ryans	2023	1	4.6%	93.6%	1.7%
Jim Harbaugh*	2024	0	69.6%	0.0%	30.4%
Jonathan Gannon	2023	1	48.0%	49.4%	2.6%
Kellen Moore	2025	0	73.1%	26.5%	0.4%
Kevin O'Connell	2022	1	24.0%	68.5%	7.5%
Liam Coen	2025	0	51.4%	44.8%	3.8%
Mike Macdonald	2024	1	21.9%	69.0%	9.1%
Mike McDaniel	2022	0	75.0%	24.1%	0.9%
Mike Vrabel*	2025	2	0.6%	40.7%	58.7%
Pete Carroll*	2025	1	8.8%	64.3%	26.9%
Raheem Morris*	2024	0	95.7%	2.3%	1.9%
Sean Payton*	2023	1	10.2%	89.2%	0.6%
Shane Steichen	2023	0	62.8%	36.9%	0.4%
Todd Bowles*	2022	0	86.8%	12.3%	0.8%

*Predicted with model retrained to exclude coach's prior HC data

After applying the data leakage fix, the ordinal model predicts only 1 of the 21 recent hires to achieve Class 2 (5+ years): Mike Vrabel (58.7% confidence). Notably, Pete Carroll, who would have been predicted Class 2 without the leakage fix due to the model having learned from his successful 14-year Seattle tenure, is instead predicted Class 1 (64.3%) when his prior outcomes are excluded from training. This demonstrates the importance of preventing data leakage for fair predictions. The model shows highest confidence (95.7%) that Raheem Morris will have a short tenure, while DeMeco Ryans and Dan Quinn receive strong Class 1 predictions (93.6% and 91.4%, respectively), suggesting moderate expected tenure with high confidence.

5 Feature Importance Analysis

While the ordinal classification model demonstrates strong predictive performance, understanding *which* features drive predictions provides actionable insights for team decision-makers. We employ SHAP (SHapley Additive exPlanations) values (Lundberg and Lee, 2017) to quantify each feature's contribution to model predictions. For the Frank-Hall ordinal classifier, we compute SHAP values for each binary classifier and aggregate them using mean absolute values across classifiers.

5.1 Offensive vs. Defensive Metrics

We categorize the 150 features into six groups based on their source and type: Core Experience (8 features), OC Statistics (33 offensive features from coordinator tenure), DC Statistics (33 defensive features from coordinator tenure), HC Team Statistics (33 features reflecting team offensive performance during head coaching tenure), HC Opponent Statistics (33 features reflecting opponent performance, i.e., defensive effectiveness, during head coaching tenure), and Hiring Team Context (10 features). Table 5 presents the mean absolute SHAP values by category.

Tab. 5: Feature Importance by Category

Category	# Features	Total |SHAP|	Avg |SHAP|
HC Opponent Stats (Defense)	33	0.131	0.0040
DC Stats (Defense)	33	0.088	0.0027
HC Team Stats (Offense)	33	0.080	0.0024
Hiring Team Context	10	0.022	0.0022
OC Stats (Offense)	33	0.049	0.0015
Core Experience	8	0.012	0.0015

Aggregating offensive metrics (OC Stats + HC Team Stats) and defensive metrics (DC Stats + HC Opponent Stats), we find that defensive features exhibit substantially higher predictive importance. Defensive metrics average 0.0033 SHAP per feature compared to 0.0020 for offensive metrics—a ratio of 1.69.

To assess statistical significance, we compute a per-coach defensive-to-offensive SHAP ratio and test whether this ratio systematically exceeds 1.0 (Table 6). Of the 635 coaching hires, 503 (79.2%) have defensive metrics contributing more to their tenure prediction than offensive metrics. The median per-coach ratio is 1.70, with a tight bootstrap 95% confidence interval of [1.64, 1.83]. A Wilcoxon signed-rank test strongly rejects the null hypothesis that the median ratio equals 1.0 ($p < 0.0001$), and a sign test confirms that the proportion of coaches with defensive-dominant predictions significantly exceeds 50% ($p < 10^{-50}$). These results provide strong evidence that defensive metrics are more predictive of coaching tenure than offensive metrics.

Tab. 6: Statistical Tests: Offensive vs. Defensive Feature Importance

Test	Result	Interpretation
Coaches with Def > Off	503/635 (79.2%)	Strong majority
Median Def/Off Ratio	1.70	—
Bootstrap 95% CI (Median)	[1.64, 1.83]	Excludes 1.0
Wilcoxon Signed-Rank	$p < 0.0001$	Highly significant
Sign Test	$p < 10^{-50}$	Highly significant

5.2 Consistency Across Eras

One concern is whether the defensive emphasis reflects historical biases in coaching evaluation that may have diminished in the modern passing-oriented NFL. We segment the data into three eras: Pre-1970 ($n = 262$), 1970–1999 ($n = 210$), and 2000+ ($n = 163$). Table 7 presents the defensive-to-offensive importance ratio by era.

Tab. 7: Defensive Bias by Era

Era	n	Def/Off Ratio	Per-Sample Median
Pre-1970	262	1.69×	1.68
1970–1999	210	1.76×	1.78
2000+	163	1.61×	1.68

A Kruskal-Wallis test comparing per-sample defensive/offensive ratios across eras yields $H = 1.83$, $p = 0.401$, indicating no significant difference. The defensive bias in tenure prediction is consistent across all eras, suggesting it reflects persistent organizational evaluation patterns rather than era-specific coaching philosophies.

5.3 Coach Background Analysis

We further investigate whether coaches are evaluated based on their area of expertise by classifying coaches according to their pre-head-coaching career: offensive background (primarily offensive coordinator or position coach experience, $n = 201$), defensive background ($n = 159$), or other/mixed ($n = 275$). The "other" category consists predominantly of pre-1970 coaches (219 of 243) for whom detailed career histories are unavailable; for hires since 2000, background classification is complete. If organizations evaluate coaches on their expertise, we would expect offensive metrics to matter more for offensive-background coaches and defensive metrics to matter more for defensive-background coaches.

Table 8 presents the results. Both offensive-background and defensive-background coaches show significantly higher importance for defensive metrics (Wilcoxon signed-rank $p < 0.0001$ for both groups). Crucially, the magnitude of defensive bias does not differ between groups (Mann-Whitney $p = 0.660$; bootstrap 95% CI for difference: $[-0.22, 0.31]$).

5.4 Top Predictive Features

Table 9 presents the ten most predictive individual features. Notably, situational defensive metrics dominate: red zone attempts faced (as DC), fourth-down conversion rate allowed, and third-down metrics. These suggest that organizations, implicitly or explicitly, weight high-leverage defensive situations when evaluating head coaches.

Tab. 8: Defensive Bias by Coach Background

Coach Background	n	Median Def/Off Ratio	Wilcoxon p
Offensive Background	201	1.81×	< 0.0001
Defensive Background	159	1.77×	< 0.0001
Mann-Whitney (between groups)			$p = 0.660$

Tab. 9: Top 10 Most Predictive Features

Rank	Feature	Mean |SHAP|
1	Red Zone Attempts (DC)	0.0342
2	4th Down % (HC Opponent)	0.0222
3	3rd Down Attempts (HC Opponent)	0.0176
4	Scoring % (HC Opponent)	0.0174
5	3rd Down % (HC Team)	0.0165
6	Rushing Attempts (HC Opponent)	0.0119
7	Passing Attempts (DC)	0.0095
8	3rd Down % (HC Opponent)	0.0071
9	Yards/Drive (HC Team)	0.0058
10	Total Yards (HC Team)	0.0051

5.5 Discussion

The feature importance analysis reveals several insights relevant to coaching evaluation.

Across all analyses, defensive performance metrics are approximately 1.7 times more predictive than offensive metrics, with 79.2% of coaches showing higher defensive than offensive SHAP contributions ($p < 10^{-50}$). This aligns with the conventional wisdom that “defense wins championships,” but extends it to job security: defense also wins job tenure.

Core experience features (age, years as coordinator, number of prior head coaching stints) rank lowest in predictive importance. This suggests that quality of prior performance matters more than quantity of experience, a finding with implications for hiring practices that weight experience heavily.

Statistical tests confirm that defensive emphasis is consistent across eras (Kruskal-Wallis $p = 0.401$) and coach backgrounds (Mann-Whitney $p = 0.660$). This persistence suggests the pattern reflects deep organizational preferences rather than artifacts of historical data or coaching pipelines.

The top predictive features emphasize high-leverage situations (red zone, third/fourth down), suggesting that organizations value coaches who perform in critical game situations.

Figure 4 visualizes the category-level importance, and Figure 5 shows the consistency of defensive bias across eras.

6 Conclusion

The ordinal classification approach using the Frank-Hall binary decomposition method, optimized with Quadratic Weighted Kappa (QWK), demonstrates strong predictive performance for NFL head coach tenure. The model achieves a QWK of 0.754, AUROC of 0.881, and 98.4% adjacent accuracy on held-out test data, indicating that predictions are both accurate and, when incorrect, typically only off by one class. Compared to a standard multiclass approach, the ordinal model shows improvements across all metrics, with a notable 62.3% improvement in the challenging middle-class (3–4 year tenure) F1 score (0.581 vs. 0.358).

SHAP-based feature importance analysis reveals that defensive metrics are approximately 1.7 times more predictive than offensive metrics, with 79.2% of coaches showing higher defensive than offensive

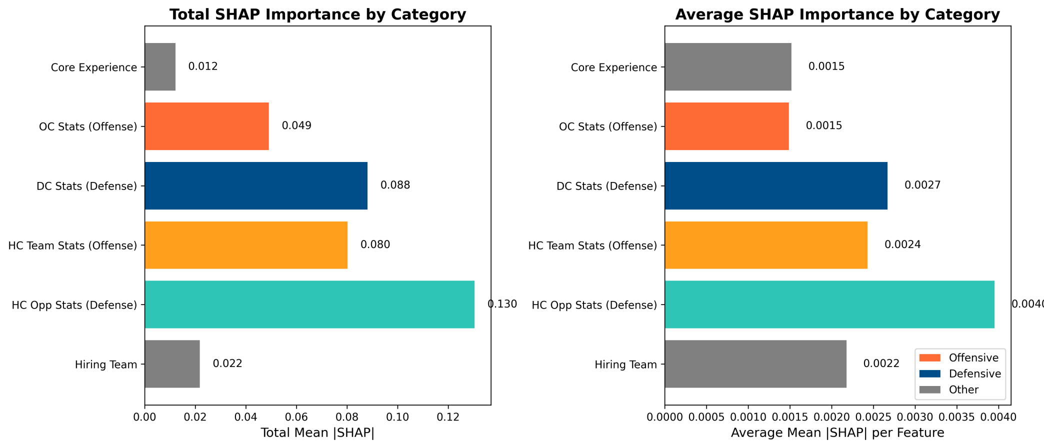

Fig. 4: SHAP feature importance by category, split into offensive and defensive metrics. Defensive categories (blue) show higher average importance than offensive categories (orange).

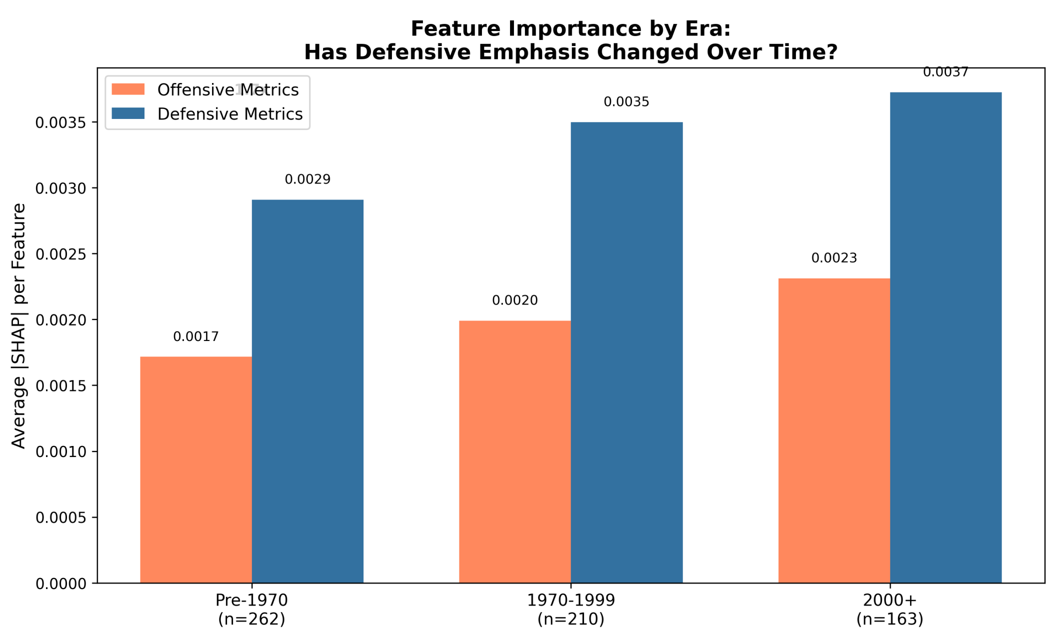

Fig. 5: Defensive-to-offensive importance ratio by era. The defensive bias is consistent across all eras (Kruskal-Wallis $p = 0.401$).

feature contributions ($p < 10^{-50}$). This defensive emphasis is consistent across all eras (Kruskal-Wallis $p = 0.401$) and coach backgrounds (Mann-Whitney $p = 0.660$), suggesting it reflects persistent organizational evaluation patterns. Notably, situational defensive metrics—red zone defense, third-down efficiency, fourth-down conversion rates—rank among the most predictive features, indicating that organizations implicitly weight high-leverage defensive situations when evaluating head coaches.

The practical implications of this research extend beyond academic interest. With approximately half of NFL head coaching hires lasting two years or fewer, teams clearly struggle to identify successful candidates through traditional evaluation methods. The model's 5.3× improvement over the human baseline (Macro F1 of 0.695 vs. 0.130) demonstrates that data-driven approaches can substantially outperform intuition-based hiring decisions. While no model can guarantee successful hires, integrating predictive analytics into the evaluation process could help teams avoid high-risk candidates and identify undervalued coaching talent. For a league where a single additional win can mean the difference between playoff contention and a losing season, even modest improvements in hiring decisions translate to meaningful competitive advantages. The model's predictions for recent hires, such as high confidence that Mike Vrabel will achieve long tenure and that Raheem Morris faces a short tenure, illustrate how these insights can inform real-world evaluations and set appropriate expectations for coaching transitions.

References

Chen, T. and Guestrin, C. (2016). XGBoost: a scalable tree boosting system. In: *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794.

DataRobot (2020). Using machine learning to peek inside the minds of NFL coaches. Available at: https://www.datarobot.com/blog/using-machine-learning-to-peek-inside-the-minds-of-nfl-coaches/

Frank, E. and Hall, M. (2001). A simple approach to ordinal classification. In: *Machine Learning: ECML 2001*, Lecture Notes in Computer Science, vol. 2167. Springer, Berlin, Heidelberg, pp. 145–156.

Lundberg, S. M. and Lee, S. (2017). A unified approach to interpreting model predictions. In: *Advances in Neural Information Processing Systems*, vol. 30, pp. 4765–4774.

Mielke, D. (2007). Coaching experience, playing experience, and coaching tenure: a commentary. *International Journal of Sports Science & Coaching*, **2**(2), pp. 117–118.

Pedregosa, F. et al. (2011). Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, **12**, pp. 2825–2830.

Roach, M. (2016). Does prior NFL head coaching experience improve team performance? *Journal of Sport Management*, **30**(3), pp. 298–311.

A Feature Descriptions

Tab. 10: Feature descriptions (Features 1–41)

No.	Feature Description
1	Age at hiring
2	Number of times previously hired as head coach
3	Number of years' experience as college position coach
4	Number of years' experience as college coordinator
5	Number of years' experience as college head coach
6	Number of years' experience as NFL position coach
7	Number of years' experience as NFL coordinator
8	Number of years' experience as NFL head coach
9	During years as NFL OC, team's average points scored
10	During years as NFL OC, team's average yards
11	During years as NFL OC, team's average yards/play
12	During years as NFL OC, team's average turnovers
13	During years as NFL OC, team's average 1st downs
14	During years as NFL OC, team's average passing completions
15	During years as NFL OC, team's average passing attempts
16	During years as NFL OC, team's average passing yards
17	During years as NFL OC, team's average passing touchdowns
18	During years as NFL OC, team's average passing interceptions
19	During years as NFL OC, team's average NY/A
20	During years as NFL OC, team's average passing first downs
21	During years as NFL OC, team's average rushing attempts
22	During years as NFL OC, team's average rushing yards
23	During years as NFL OC, team's average rushing touchdowns
24	During years as NFL OC, team's average rush yards per play
25	During years as NFL OC, team's average rushing 1st downs
26	During years as NFL OC, team's average number of penalties
27	During years as NFL OC, team's average penalty yards
28	During years as NFL OC, team's average penalty 1st downs
29	During years as NFL OC, team's average number of drives
30	During years as NFL OC, team's average scoring percentage
31	During years as NFL OC, team's average turnover percentage
32	During years as NFL OC, team's average drive duration
33	During years as NFL OC, team's average plays per drive
34	During years as NFL OC, team's average yards per drive
35	During years as NFL OC, team's average points per drive
36	During years as NFL OC, team's average number of 3rd down attempts
37	During years as NFL OC, team's average third down conversion percentage
38	During years as NFL OC, team's average number of 4th down attempts
39	During years as NFL OC, team's average 4th down conversion percentage
40	During years as NFL OC, team's average red zone attempts
41	During years as NFL OC, team's average red zone percentage

Tab. 11: Feature descriptions (Features 42–74)

No.	Feature Description
42	During years as NFL DC, opponent team's average points scored
43	During years as NFL DC, opponent team's average yards
44	During years as NFL DC, opponent team's average yards/play
45	During years as NFL DC, opponent team's average turnovers
46	During years as NFL DC, opponent team's average 1st downs
47	During years as NFL DC, opponent team's average passing completions
48	During years as NFL DC, opponent team's average passing attempts
49	During years as NFL DC, opponent team's average passing yards
50	During years as NFL DC, opponent team's average passing touchdowns
51	During years as NFL DC, opponent team's average passing interceptions
52	During years as NFL DC, opponent team's average NY/A
53	During years as NFL DC, opponent team's average passing first downs
54	During years as NFL DC, opponent team's average rushing attempts
55	During years as NFL DC, opponent team's average rushing yards
56	During years as NFL DC, opponent team's average rushing touchdowns
57	During years as NFL DC, opponent team's average rush yards per play
58	During years as NFL DC, opponent team's average rushing 1st downs
59	During years as NFL DC, opponent team's average number of penalties
60	During years as NFL DC, opponent team's average penalty yards
61	During years as NFL DC, opponent team's average penalty 1st downs
62	During years as NFL DC, opponent team's average number of drives
63	During years as NFL DC, opponent team's average scoring percentage
64	During years as NFL DC, opponent team's average turnover percentage
65	During years as NFL DC, opponent team's average drive duration
66	During years as NFL DC, opponent team's average plays per drive
67	During years as NFL DC, opponent team's average yards per drive
68	During years as NFL DC, opponent team's average points per drive
69	During years as NFL DC, opponent team's average number of 3rd down attempts
70	During years as NFL DC, opponent team's average third down conversion pct.
71	During years as NFL DC, opponent team's average number of 4th down attempts
72	During years as NFL DC, opponent team's average 4th down conversion pct.
73	During years as NFL DC, opponent team's average red zone attempts
74	During years as NFL DC, opponent team's average red zone percentage

Tab. 12: Feature descriptions (Features 75–107)

No.	Feature Description
75	During years as NFL HC, team's average points scored
76	During years as NFL HC, team's average yards
77	During years as NFL HC, team's average yards/play
78	During years as NFL HC, team's average turnovers
79	During years as NFL HC, team's average 1st downs
80	During years as NFL HC, team's average passing completions
81	During years as NFL HC, team's average passing attempts
82	During years as NFL HC, team's average passing yards
83	During years as NFL HC, team's average passing touchdowns
84	During years as NFL HC, team's average passing interceptions
85	During years as NFL HC, team's average NY/A
86	During years as NFL HC, team's average passing first downs
87	During years as NFL HC, team's average rushing attempts
88	During years as NFL HC, team's average rushing yards
89	During years as NFL HC, team's average rushing touchdowns
90	During years as NFL HC, team's average rush yards per play
91	During years as NFL HC, team's average rushing 1st downs
92	During years as NFL HC, team's average number of penalties
93	During years as NFL HC, team's average penalty yards
94	During years as NFL HC, team's average penalty 1st downs
95	During years as NFL HC, team's average number of drives
96	During years as NFL HC, team's average scoring percentage
97	During years as NFL HC, team's average turnover percentage
98	During years as NFL HC, team's average drive duration
99	During years as NFL HC, team's average plays per drive
100	During years as NFL HC, team's average yards per drive
101	During years as NFL HC, team's average points per drive
102	During years as NFL HC, team's average number of 3rd down attempts
103	During years as NFL HC, team's average third down conversion percentage
104	During years as NFL HC, team's average number of 4th down attempts
105	During years as NFL HC, team's average 4th down conversion percentage
106	During years as NFL HC, team's average red zone attempts
107	During years as NFL HC, team's average red zone percentage

Tab. 13: Feature descriptions (Features 108–150)

No.	Feature Description
108	During years as NFL HC, opponent team's average points scored
109	During years as NFL HC, opponent team's average yards
110	During years as NFL HC, opponent team's average yards/play
111	During years as NFL HC, opponent team's average turnovers
112	During years as NFL HC, opponent team's average 1st downs
113	During years as NFL HC, opponent team's average passing completions
114	During years as NFL HC, opponent team's average passing attempts
115	During years as NFL HC, opponent team's average passing yards
116	During years as NFL HC, opponent team's average passing touchdowns
117	During years as NFL HC, opponent team's average passing interceptions
118	During years as NFL HC, opponent team's average NY/A
119	During years as NFL HC, opponent team's average passing first downs
120	During years as NFL HC, opponent team's average rushing attempts
121	During years as NFL HC, opponent team's average rushing yards
122	During years as NFL HC, opponent team's average rushing touchdowns
123	During years as NFL HC, opponent team's average rush yards per play
124	During years as NFL HC, opponent team's average rushing 1st downs
125	During years as NFL HC, opponent team's average number of penalties
126	During years as NFL HC, opponent team's average penalty yards
127	During years as NFL HC, opponent team's average penalty 1st downs
128	During years as NFL HC, opponent team's average number of drives
129	During years as NFL HC, opponent team's average scoring percentage
130	During years as NFL HC, opponent team's average turnover percentage
131	During years as NFL HC, opponent team's average drive duration
132	During years as NFL HC, opponent team's average plays per drive
133	During years as NFL HC, opponent team's average yards per drive
134	During years as NFL HC, opponent team's average points per drive
135	During years as NFL HC, opponent team's average number of 3rd down attempts
136	During years as NFL HC, opponent team's average third down conversion pct.
137	During years as NFL HC, opponent team's average number of 4th down attempts
138	During years as NFL HC, opponent team's average 4th down conversion pct.
139	During years as NFL HC, opponent team's average red zone attempts
140	During years as NFL HC, opponent team's average red zone percentage
141	Hiring team's average winning percent in previous two years
142	Hiring team's average points scored in previous two years
143	Hiring team's average points allowed in previous two years
144	Hiring team's average yards of offense in previous two years
145	Hiring team's average yards of offense allowed in previous two years
146	Hiring team's average yards / play in previous two years
147	Hiring team's average yards / play allowed in previous two years
148	Hiring team's average turnovers forced in previous two years
149	Hiring team's average turnovers in previous two years
150	Hiring team's number of playoff appearances in previous two years

B Data Distributions

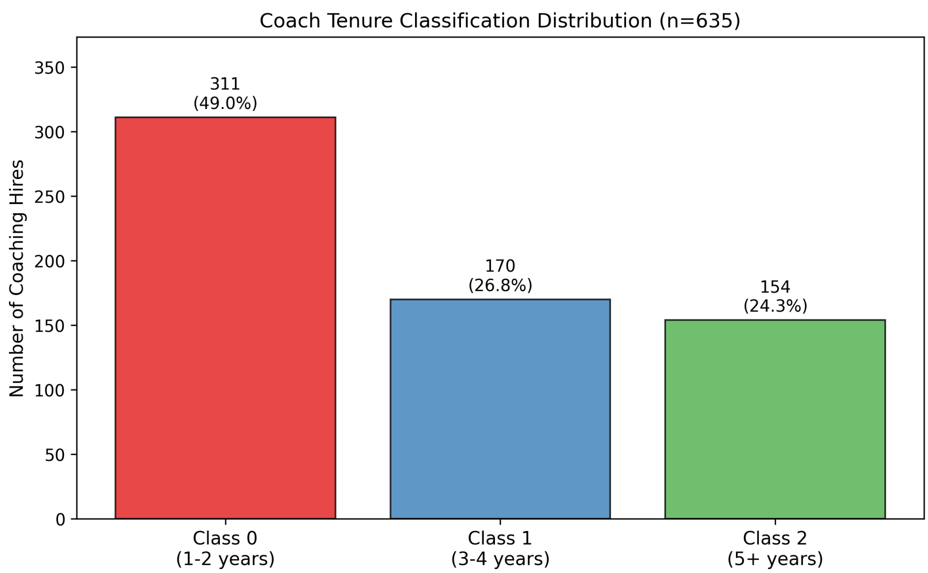

Fig. 6: Coach tenure classification frequency distribution across all 635 coaching hire instances with known tenure outcomes.

C Model Hyperparameters

Tab. 14: Final hyperparameters for the ordinal XGBoost classifier model (QWK-optimized)

Hyperparameter	**Value**
Classification Method	Frank-Hall Ordinal
Optimization Metric	Quadratic Weighted Kappa
Base Classifier Objective	binary:logistic
Number of Binary Classifiers	2
Random State	42
Number of Estimators	200
Learning Rate	0.25
Max Estimator Depth	2
Gamma	0
Lambda (L2 Regularization)	0.1
Alpha (L1 Regularization)	0.01
Subsample	0.80
Colsample by Tree	0.90
Minimum Child Weight	3