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: This study develops an ordinal classification model to predict the tenure of NFL head coaching hires using features available at the time of hiring. From an initial set of 150 engineered features spanning coach experience, historical team performance during prior coordinator and head coaching roles, and hiring team context, a feature selection analysis evaluating subsets of 5 to 150 features across 50 cross-validation splits identifies 40 as the best-performing feature count. The model employs the Frank-Hall binary decomposition method with XGBoost base classifiers, respecting the ordering of tenure classes (short: 1–2 years, medium: 3–4 years, long: 5+ years). Optimized using Quadratic Weighted Kappa (QWK), the 40-feature model achieves a mean QWK of 0.744 (95% CI: [0.731, 0.757]) averaged across 50 train/test partitions, with a mean absolute error of 0.307 (95% CI: [0.294, 0.320]) and 98.1% adjacent accuracy—outperforming the full 150-feature model on all metrics. The ordinal model outperforms a multiclass approach across all metrics ($p < 0.001$ by paired t-test), with the largest advantage on class 1 F1 (mean $\Delta = +0.106$, 95% CI: [+0.088, +0.124]). SHAP-based feature importance analysis reveals that defensive metrics are more predictive than offensive metrics (mean defensive/offensive SHAP ratio: 1.14, 95% CI: [1.10, 1.17]).

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

1 Introduction

Head coaching decisions represent one of the most consequential personnel choices in professional sports. Coaching changes are frequent, yet evidence suggests that head coach turnover yields little if any improvement in team performance (Bryson et al., 2024), and the high rate of turnover in the NFL suggests that teams struggle to identify candidates who will succeed in the role. In our dataset of 635 head coaching hires (1920–2021), 49% lasted two years or fewer. Despite this high failure rate, the hiring process relies primarily on interviews, scheme presentations, and qualitative assessments rather than quantitative analysis. The NFL's adoption of the Rooney Rule in 2003, which mandates interviews with minority candidates for head coaching vacancies, was motivated in part by evidence that hiring decisions were driven by professional networks rather than systematic evaluation of coaching credentials (Collins, 2007). Quantitative methods offer a complementary approach to evaluating coaching candidates by identifying statistical patterns associated with tenure outcomes.

To that end, this study uses coach tenure, defined as the number of years a hired coach remains in the position before being dismissed, departing, or retiring, as the primary outcome variable. 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 retained, while a coach with the same record on a team expected to contend may be dismissed after two seasons. Tenure reflects not only 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 captures patterns that generalize across different team situations without requiring explicit modeling of each organization's context. This approach is related to survival analysis methods used to study coaching tenure in professional soccer (Gilfix et al., 2020), though we adopt an ordinal classification framework rather than a time-to-event model.

2 Literature Review

Quantitative analysis of coaching in professional sports has received growing attention in the sports analytics literature. Gilfix et al. (2020) studied coaching tenure in Major League Soccer using survival analysis, finding that performance improvement relative to predecessors, league context, and age at hire all affect tenure duration. Cannon et al. (2025) proposed a method for isolating NBA head coaches' in-game scheme effects on win probability using Bayesian additive regression trees. These studies demonstrate increasing interest in quantifying coaching value, though neither addresses prediction of tenure at the time of hire.

Within NFL analytics, several studies have employed machine learning methods for prediction tasks. Lock and Nettleton (2014) used random forests to estimate win probability before each play, and David et al. (2011) applied neural network committees to NFL game outcome prediction using team-level statistics similar to those employed in the present study. Wolfson et al. (2011) examined the analogous problem of forecasting NFL quarterback career success from pre-career observable features, finding that much of the variation in future performance is driven by unobservable factors. Yurko et al. (2019) developed a reproducible Wins Above Replacement (WAR) methodology for evaluating NFL offensive players, establishing a framework for measuring individual contributions within a team sport.

The literature on NFL coaching specifically is sparse. Roach (2016) used linear regression with seven predictors to examine whether prior head coaching experience predicts wins in a coach's first three years, finding a negative association, though the model's adjusted R^2 of 0.336 and limited feature set constrain the generalizability of this finding. Mielke (2007) reviews research in sports economics and argues that hiring decisions based solely on prior success are unlikely to be optimal given resource inequality among franchises. Urschel and Zhuang (2011) analyzed NFL coaching decision-making through the lens of prospect theory, demonstrating that coaches exhibit risk and loss aversion in strategic choices.

To our knowledge, no prior study has applied machine learning to predict the tenure or success of NFL head coaching hires at the time of hiring. The present study addresses this gap by combining a large feature set with ordinal classification methods designed for the ordered nature of tenure outcomes.

3 Methods

Using statistics available at the time of hiring, we predict the tenure classification of NFL head coach hires using XGBoost models (Chen and Guestrin, 2016). Raw coaching histories and team performance statistics are scraped from Pro-Football-Reference.com (Pro-Football-Reference, 2025). All data processing, model implementation, and analysis are performed using Python with scikit-learn (Pedregosa et al., 2011).

3.1 Data and Features

The feature engineering pipeline produces 150 candidate features for each head coaching hire, organized into five categories (Table 1). Appendix A provides complete descriptions of all 150 features.

Core experience features capture coaching background (age, prior head coaching stints, years of experience at each level). Coordinator and head coach statistics record a common set of 33 team performance metrics—covering production, passing, rushing, penalties, drive efficiency, and situational play—for each of four role–side combinations (OC, DC, HC Team, HC Opponent). DC and HC Opponent statistics reflect the performance of the opposing offense, so they capture defensive effectiveness indirectly. All statistics are computed as unweighted means across seasons in the respective role. Hiring team context features capture the state of the hiring team (win percentage, points, yards, turnovers, playoff appearances) averaged over the two seasons prior to the hire.

Tab. 1: Feature categories and counts for the 150 engineered candidate features. Each coaching hire is characterized by core experience metrics, performance statistics from prior coordinator and head coaching roles, and the hiring team's recent performance.

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 Team Stats	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

3.1.1 Feature Normalization

The coordinator and head coach performance statistics (features 9–140) are normalized relative to league averages to enable meaningful comparisons across eras. For each statistic s in season t, the normalized value is computed as:

$$z_{i,s,t} = \frac{x_{i,s,t} - \mu_{s,t}}{\sigma_{s,t}} \tag{1}$$

where $x_{i,s,t}$ is the raw value of statistic s for coach i's team in season t, and $\mu_{s,t}$ and $\sigma_{s,t}$ are the league-wide mean and standard deviation of statistic s across all teams in season t. For coaches whose role spans multiple seasons, the normalized values are averaged across seasons. This normalization ensures that a coach whose offense ranked one standard deviation above league average in 1985 is treated equivalently to a coach whose offense ranked one standard deviation above average in 2020. The hiring team context features (141–150) are not normalized, as they capture the absolute state of the hiring team (e.g., win percentage, playoff appearances) rather than a coach's relative performance in a prior role.

3.1.2 Missing Data Imputation

Missing values arise when coaches lack experience at certain levels. A first-time head coach has no prior HC statistics (66 missing features), and a coach who served as a coordinator on only one side of the ball will have an additional 33 missing features for the other coordinator role. The most common profile, a first-time head coach with experience at one coordinator position, thus has 99 of 150 features missing. Across all 635 coaching hire instances, the median number of missing features is 99, reflecting that most coaches have not held all three role types (OC, DC, and HC) prior to their head coaching hire.

We use truncated SVD (singular value decomposition) to impute missing values via low-rank matrix factorization. The 635 × 150 feature matrix $\mathbf{X}$ is decomposed as $\mathbf{X} \approx \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$, where the rank is selected to minimize held-out reconstruction error using 5-fold cross-validation on observed entries. The imputed matrix is then used for all subsequent feature selection and model training. Appendix B shows the distribution of coach tenure classifications across all hiring instances.

3.2 Ordinal Classification Model

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 (2) shows the mapping between the coach tenure t (in years) and the three coach tenure classification labels $C(t)$.

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

We group coach tenures into classes for classification rather than predicting continuous tenure duration, as the relationship between tenure length and coaching success is nonlinear. For instance, a 15-year tenure does not represent a proportionally greater success than a 10-year tenure. The ordinal classes instead capture qualitatively different outcomes: early dismissal, moderate tenure, and long-term retention.

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.

We implement 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, the first classifier estimates $P(Y > 0)$, distinguishing Class 0 from Classes 1 and 2, and the second estimates $P(Y > 1)$, distinguishing Classes 0 and 1 from Class 2.

Class probabilities are then derived from these cumulative probabilities:

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

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

$$P(Y = 2) = P(Y > 1) \tag{5}$$

The Frank-Hall method offers several advantages: it works with any base classifier, produces interpretable probability distributions, and naturally penalizes distant misclassifications.

Each binary classifier in the decomposition is an XGBoost gradient-boosted tree ensemble (Chen and Guestrin, 2016). XGBoost minimizes a regularized objective function over T trees:

$$\mathcal{L} = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i) + \sum_{t=1}^{T} \left(\gamma |\mathcal{T}_t| + \tfrac{1}{2} \lambda \|w_t\|^2 \right) \tag{6}$$

where ℓ is the logistic loss for binary classification, $|\mathcal{T}_t|$ is the number of leaves in tree t, w_t is the vector of leaf weights, and γ and λ are regularization parameters controlling tree complexity and leaf weight magnitude, respectively. Trees are added sequentially, with each new tree fitted to the negative gradient of the loss with respect to the current ensemble predictions.

3.3 Evaluation Methodology

3.3.1 Metrics

We use Quadratic Weighted Kappa (QWK) as the primary evaluation metric because it is specifically designed for ordinal classification problems. For K classes, QWK is defined as:

$$\kappa_w = 1 - \frac{\sum_{i,j} W_{ij}\, O_{ij}}{\sum_{i,j} W_{ij}\, E_{ij}}, \quad \text{where} \quad W_{ij} = \frac{(i - j)^2}{(K - 1)^2} \tag{7}$$

Here O is the observed confusion matrix, E is the expected confusion matrix under independence (computed from the outer product of row and column marginals of O, normalized to sum to the total number of instances), and W is the quadratic weight matrix. A prediction of Class 0 for a true Class 2 instance receives weight $W_{0,2} = 1$, while an adjacent error receives weight $W_{0,1} = 1/4$, so distant misclassifications are penalized four times more heavily. QWK ranges from -1 (systematic disagreement) through 0 (chance agreement) to 1 (perfect agreement), and also serves as the optimization target during hyperparameter tuning. 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).

3.3.2 Cross-Validation Strategy

To prevent data leakage, we implement 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. The dataset is split into training (508 instances, 80%) and test (127 instances, 20%) sets using a coach-level stratified procedure: each coach is assigned to the class corresponding to their most frequent tenure outcome, coaches are grouped by class, and coaches are greedily allocated to the test set within each class until the target proportion is reached. This ensures approximate class balance in both splits while preventing coach overlap. To avoid dependence on a single partition, the primary performance metrics reported in Section 4 are averaged across 50 independent train/test splits generated with different random seeds.

Within each training set, hyperparameters are tuned via 5-fold cross-validation using the same coach-level grouping procedure, ensuring approximate class stratification in every fold. We perform randomized search with 1,000 iterations over the hyperparameter space defined in Appendix D, scoring each candidate configuration by Quadratic Weighted Kappa (QWK) on the held-out folds. The $K - 1$ binary classifiers in the Frank-Hall decomposition share the same hyperparameters but are trained independently on their respective binary targets.

3.3.3 Model Comparison

To compare the ordinal (Frank-Hall) classifier against a standard multiclass XGBoost baseline, both models are trained from scratch on each of the 50 train/test partitions using the same hyperparameters and feature set. For each partition, we record the difference (ordinal minus multiclass) on every evaluation metric, yielding 50 paired observations per metric. We test whether the mean difference is significantly different from zero using a one-sided paired t-test, with the alternative hypothesis that the ordinal model outperforms the multiclass model.

3.4 Feature Selection

To reduce noise and improve model performance, we employ a feature selection procedure based on SHAP (SHapley Additive exPlanations) importance values (Lundberg and Lee, 2017). An initial ordinal model is trained on all 150 features, and mean absolute SHAP values are computed across all training instances. Features are ranked by their SHAP importance, and reduced models are trained using only the top K features for $K \in \{5, 10, 20, \ldots, 150\}$. Each feature count is evaluated across the same 50 independent train/test splits used for the primary analysis. This procedure identifies $K = 40$ as the best-performing feature count (Section 4.1).

4 Results

4.1 Feature Selection

Table 2 presents model performance across feature counts. Features are ranked by mean absolute SHAP value and reduced ordinal models are trained using only the top K features. Each feature count is evaluated across 50 independent train/test splits.

Tab. 2: Model performance with SHAP-ranked feature subsets across 50 independent train/test splits. 95% CI = confidence interval for the mean (t-distribution). Bold indicates peak QWK.

# Features	QWK [95% CI]	MAE [95% CI]	Adj. Acc. [95% CI]	Macro F1 [95% CI]
5	0.473 [.454, .493]	0.524 [.509, .538]	90.5% [89.8, 91.2]	0.544 [.534, .554]
10	0.587 [.572, .602]	0.434 [.422, .447]	93.5% [93.0, 93.9]	0.605 [.595, .616]
20	0.707 [.695, .719]	0.336 [.324, .347]	96.8% [96.4, 97.2]	0.673 [.663, .684]
30	0.729 [.715, .744]	0.314 [.300, .328]	97.3% [96.9, 97.7]	0.691 [.677, .704]
40	**0.744** [.731, .757]	**0.307** [.294, .320]	**98.1%** [97.8, 98.5]	**0.691** [.679, .703]
50	0.726 [.714, .739]	0.317 [.304, .329]	97.4% [97.1, 97.8]	0.688 [.675, .700]
60	0.719 [.704, .734]	0.325 [.312, .339]	97.5% [97.1, 97.9]	0.678 [.666, .690]
70	0.722 [.709, .735]	0.325 [.313, .337]	97.8% [97.4, 98.1]	0.675 [.664, .686]
80	0.726 [.713, .738]	0.325 [.313, .337]	98.0% [97.7, 98.3]	0.673 [.662, .684]
90	0.713 [.700, .726]	0.330 [.317, .342]	97.5% [97.2, 97.8]	0.674 [.663, .685]
100	0.707 [.693, .720]	0.339 [.325, .352]	97.6% [97.2, 97.9]	0.664 [.651, .677]
110	0.699 [.686, .712]	0.348 [.336, .360]	97.6% [97.2, 98.0]	0.653 [.643, .664]
120	0.696 [.683, .710]	0.346 [.334, .358]	97.4% [97.0, 97.7]	0.658 [.647, .670]
130	0.693 [.679, .708]	0.351 [.338, .364]	97.3% [97.0, 97.7]	0.652 [.641, .663]
140	0.688 [.674, .701]	0.352 [.340, .364]	97.2% [96.8, 97.5]	0.653 [.642, .664]
150	0.692 [.678, .707]	0.350 [.337, .363]	97.3% [96.9, 97.6]	0.656 [.643, .668]

The final model uses the 40 features listed in Table 3. All five original categories are represented, with HC Opponent Stats contributing the most features (14), followed by HC Team Stats (10), DC Stats (6), Hiring Team (5), OC Stats (3), and Core Experience (2). Head coaching statistics account for 24 of the 40 selected features.

4.2 Ordinal Classification Model Performance

Table 4 reports performance metrics for the ordinal XGBoost classifier using the 40 SHAP-ranked features, averaged across 50 independent train/test partitions.

The ordinal model achieves a mean quadratic weighted kappa of 0.744 (95% CI: [0.731, 0.757]) across 50 partitions. Adjacent accuracy averages 98.1%, meaning nearly all predictions fall within one class of the true label.

Figure 2 presents the confusion matrix averaged across all 50 train/test partitions. Class 0 (short tenure) is classified most accurately (84.5%), while class 1 (medium tenure) is hardest to distinguish (59.8%), with errors spreading in both directions. Distant misclassifications between class 0 and class 2 are rare (1.9% and 4.9%), consistent with the high adjacent accuracy.

4.3 Comparison with Standard Multiclass Classification

To validate the ordinal approach, we compare against a standard multiclass XGBoost classifier trained with the same hyperparameters. For each of the 50 random partitions, both an ordinal and a multiclass

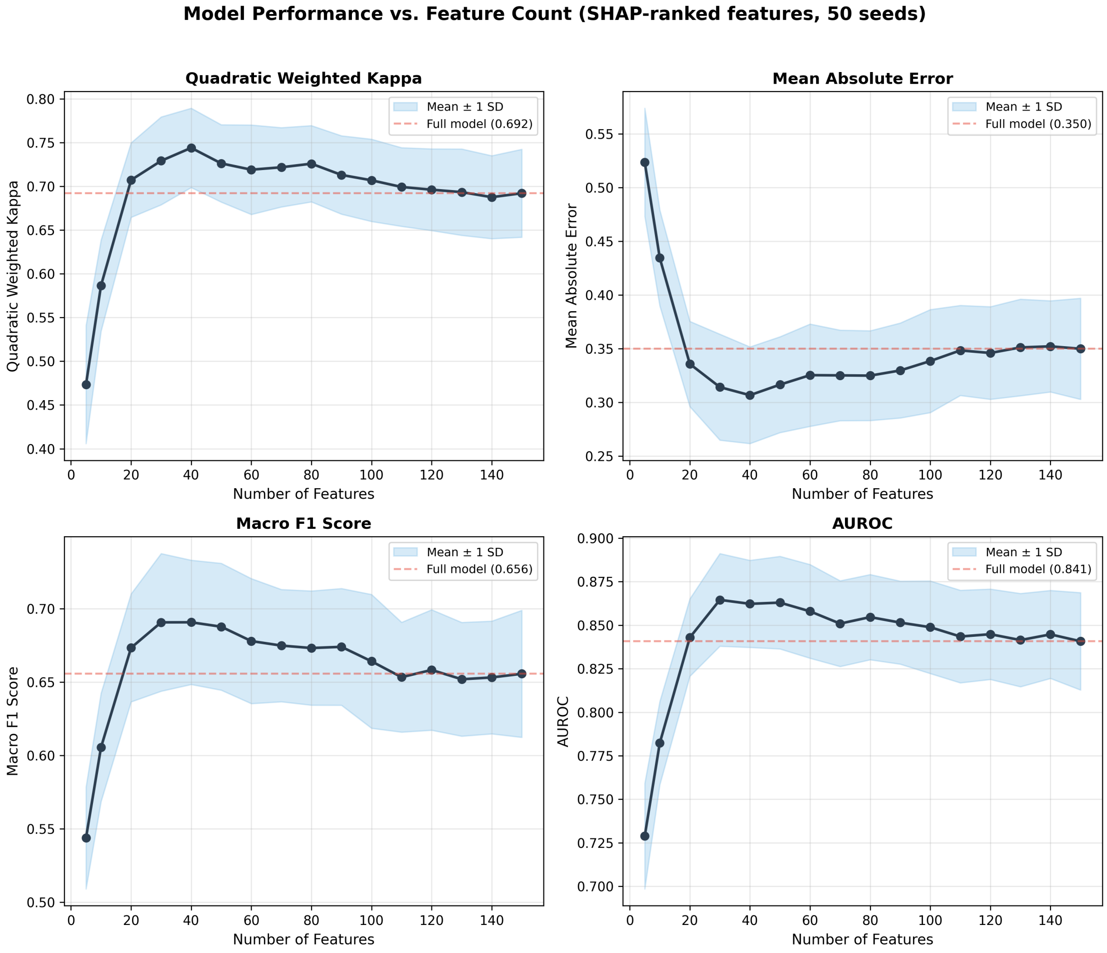

Fig. 1: Model performance as a function of feature count (SHAP-ranked). Shaded bands show ±1 standard deviation across 50 independent train/test splits. Dashed red line indicates full-model (150 feature) performance. Performance peaks at 40 features and generally declines beyond 50 features.

model are trained from scratch on the training set and evaluated on the test set. Table 5 summarizes the comparison across all partitions. Statistical significance is assessed via one-sided paired t-tests on the 50 seed-level differences (ordinal minus multiclass).

The ordinal model outperforms the multiclass approach on all seven metrics, with every difference statistically significant at $p < 0.001$ by paired t-test and all confidence intervals excluding zero. The largest advantage appears on class 1 (medium tenure) F1, with a mean improvement of +0.106 (95% CI: [+0.088, +0.124]).

4.4 Feature Importance

We categorize the 40 selected features into six groups based on their source and type: Core Experience (2), OC Statistics (3), DC Statistics (6), HC Team Statistics (10), HC Opponent Statistics (14), and Hiring Team Context (5). Table 6 presents the mean absolute SHAP values by category.

Aggregating offensive metrics (OC Stats + HC Team Stats) and defensive metrics (DC Stats + HC Opponent Stats), defensive features exhibit higher average per-feature importance (ratio of mean |SHAP| per defensive feature to mean |SHAP| per offensive feature = 1.14, 95% CI: [1.10, 1.17], across 50 partitions). The SHAP-based feature selection itself reflects this asymmetry: 20 of the 40 selected features (50%) come

Tab. 3: The 40 SHAP-selected features used in the final model, ranked by normalized importance (mean |SHAP| across 50 train/test partitions, normalized to sum to 1) with 95% confidence intervals. Category abbreviations: HC Opp = HC Opponent Stats, HC = HC Team Stats, DC = DC Stats, Hire = Hiring Team, OC = OC Stats, Core = Core Experience.

	Feature	Imp.	95% CI		Feature	Imp.	95% CI
1	3D% (HC)	.081	[.077, .085]	21	4D Att (HC Opp)	.016	[.015, .018]
2	3D Att (HC Opp)	.073	[.069, .077]	22	Int (HC)	.016	[.014, .018]
3	RZ Att (DC)	.067	[.064, .070]	23	RZ Att (HC Opp)	.015	[.014, .017]
4	4D% (HC Opp)	.062	[.058, .065]	24	Yds Off (Hire)	.013	[.012, .015]
5	Sc% (HC Opp)	.054	[.051, .057]	25	Pen 1D (DC)	.013	[.012, .014]
6	Rush Att (HC Opp)	.052	[.049, .055]	26	Rush 1D (HC Opp)	.013	[.011, .014]
7	3D% (HC Opp)	.052	[.048, .055]	27	TO Forced (Hire)	.012	[.011, .014]
8	Yds/Dr (HC)	.047	[.044, .050]	28	Pass TD (HC)	.012	[.011, .013]
9	Yds/Dr (HC Opp)	.042	[.039, .045]	29	TO Comm (Hire)	.011	[.010, .012]
10	Y/P (OC)	.030	[.028, .032]	30	Yrs NFL Coor (Core)	.011	[.010, .012]
11	Cmp (HC)	.030	[.027, .032]	31	Pass 1D (DC)	.011	[.010, .012]
12	Pass Att (DC)	.029	[.027, .032]	32	Pts Allow (Hire)	.011	[.009, .012]
13	TO (DC)	.025	[.023, .027]	33	Yrs NFL Pos (Core)	.010	[.009, .011]
14	TO% (HC)	.025	[.023, .027]	34	Cmp (HC Opp)	.010	[.009, .011]
15	TO (HC)	.021	[.019, .023]	35	RZ% (HC Opp)	.009	[.008, .011]
16	Plays/Dr (OC)	.019	[.017, .022]	36	Y/P (DC)	.009	[.008, .010]
17	#Dr (HC Opp)	.019	[.017, .021]	37	Rush Y/A (HC Opp)	.007	[.007, .008]
18	RZ Att (OC)	.018	[.016, .019]	38	Pts (HC Opp)	.007	[.006, .008]
19	Pts (HC)	.017	[.016, .019]	39	Y/P (Hire)	.007	[.006, .008]
20	Yds (HC)	.017	[.015, .019]	40	Sc% (HC)	.006	[.006, .007]

Tab. 4: Coach tenure classification prediction results (Ordinal Model, 40 features). Metrics are averaged across 50 independent train/test partitions; 95% CI = confidence interval for the mean (t-distribution).

Metric	Mean	95% CI
Mean Absolute Error (MAE)	0.307	[0.294, 0.320]
Quadratic Weighted Kappa (QWK)	0.744	[0.731, 0.757]
Adjacent Accuracy (±1 class)	98.1%	[97.8%, 98.5%]
Exact Accuracy	71.2%	[70.0%, 72.4%]
Macro F1 Score	0.691	[0.679, 0.703]
AUROC (macro OVR)	0.862	[0.855, 0.869]
Optimistic Baseline F1*	0.130	—

*All-Class-2 baseline: reflects the assumption that every hire will succeed

Tab. 5: Ordinal vs. Multiclass model comparison across 50 independent train/test partitions (40-feature model). Mean Δ = Ordinal − Multiclass. 95% CI = confidence interval for the mean difference (based on the t-distribution). p = one-sided paired t-test.

Metric	Mean Δ	95% CI	p
MAE	−0.044	[−0.054, −0.034]	<0.001
QWK	+0.033	[+0.023, +0.042]	<0.001
Adj. Accuracy	+0.9pp	[+0.5, +1.2]pp	<0.001
Exact Accuracy	+3.5pp	[+2.5, +4.5]pp	<0.001
Macro F1	+0.050	[+0.039, +0.060]	<0.001
AUROC	+0.022	[+0.017, +0.027]	<0.001
Class 1 F1	+0.106	[+0.088, +0.124]	<0.001

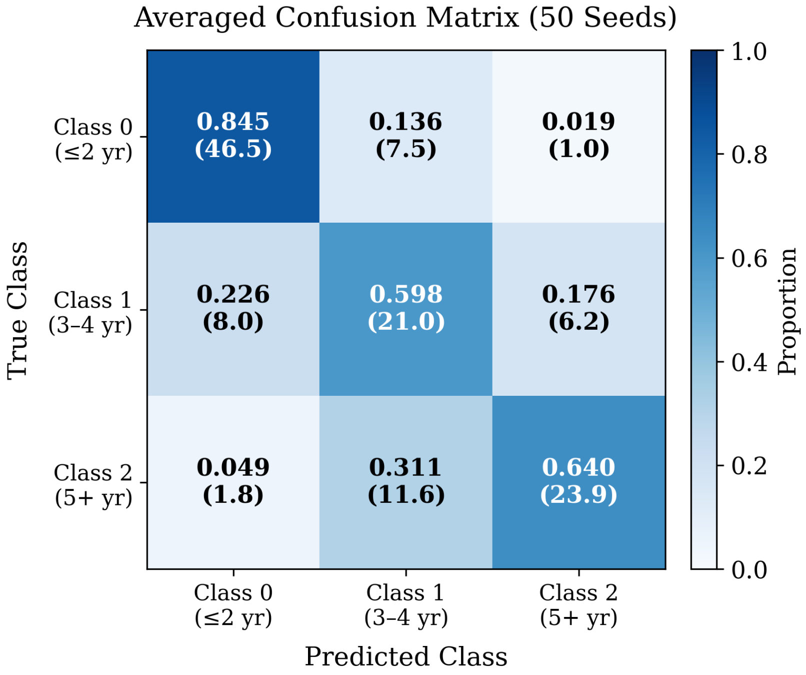

Fig. 2: Confusion matrix averaged across 50 independent train/test partitions. Each cell shows the proportion of instances with that true/predicted class combination, with the average raw count per partition in parentheses. Rows are normalized to sum to 1.0.

Tab. 6: Feature importance by category measured using mean absolute SHAP values across 50 train/test partitions. Total |SHAP| is the sum across all features in each category; Avg |SHAP| is the mean per feature. 95% CI = confidence interval for the mean (t-distribution).

Category	# Feat.	Total |SHAP|	95% CI	Avg |SHAP|	95% CI
HC Stats (Team + Opp.)	24	0.407	[.400, .413]	0.0169	[.0167, .0172]
DC Stats (Defense)	6	0.089	[.087, .092]	0.0149	[.0144, .0153]
OC Stats (Offense)	3	0.039	[.036, .041]	0.0128	[.0121, .0136]
Hiring Team Context	5	0.032	[.030, .033]	0.0063	[.0060, .0066]
Core Experience	2	0.012	[.012, .013]	0.0062	[.0058, .0066]

from defensive categories versus 13 (33%) from offensive categories, despite equal representation in the full feature set. This asymmetry suggests that a coach's ability to limit opposing offenses carries more weight in tenure decisions than offensive production.

Of the eight core experience features, only years as NFL position coach and years as NFL coordinator appear in the top 40, ranking 30th and 33rd respectively. Age, number of prior head coaching stints, years as NFL head coach, and all college experience features fall below the SHAP selection threshold. This is consistent with the finding of Roach (2016) that prior head coaching experience does not positively predict future success.

Appendix C presents partial dependence plots for the six most predictive features, illustrating how each feature's value affects the predicted class probabilities.

5 Conclusion

This study applies ordinal classification via the Frank-Hall binary decomposition method to predict NFL head coach tenure. Using 40 SHAP-ranked features selected from an initial set of 150 candidates, the model achieves a mean QWK of 0.744 (95% CI: [0.731, 0.757]), mean absolute error of 0.307 (95% CI: [0.294, 0.320]), and 98.1% adjacent accuracy (95% CI: [97.8%, 98.5%]) averaged across 50 independent train/test partitions. Feature selection analysis demonstrates that the 40-feature model outperforms the full 150-feature model by +0.052 QWK (95% CI: [+0.038, +0.066]). The ordinal model consistently outperforms a standard multiclass approach across all metrics ($p < 0.001$ by paired t-test), with the largest improvement on class 1 F1 (mean $\Delta = +0.106$, 95% CI: [+0.088, +0.124]).

SHAP-based feature importance analysis shows that defensive features carry higher average per-feature importance than offensive features (ratio = 1.14, 95% CI: [1.10, 1.17]). The top-40 feature set contains 20 defensive features versus 13 offensive features, and 6 of the top 10 features measure opponent performance during prior head coaching tenure.

Several limitations should be noted. The model relies on publicly available aggregate statistics and cannot capture qualitative factors such as leadership ability, organizational fit, or interpersonal dynamics that may influence tenure decisions. The tenure outcome itself reflects organizational decisions that may be driven by factors unrelated to coaching quality, such as ownership changes or front office turnover. Additionally, the relatively small sample size of 635 coaching hires, combined with class imbalance (49% Class 0), limits the statistical power for detecting differences in the minority classes.

Coaching history data was scraped from Pro Football Reference. Processed data and analysis code are available at https://github.com/jwilliamson7/CoachingProject or upon request to the corresponding author.

Two directions may yield the largest improvements. First, the current features average statistics across seasons in a given role, discarding temporal information: a coordinator whose units improved each year is indistinguishable from one whose units declined. Trajectory features capturing year-over-year trends could add discriminative power, particularly for the medium-tenure class where misclassifications are concentrated. Second, the model's representation of organizational context is limited to five hiring-team features; ownership stability, general manager tenure, roster quality, and salary cap position all plausibly affect tenure but are not currently modeled.

References

Collins, B. W. (2007). Tackling unconscious bias in hiring practices: the plight of the Rooney Rule. *NYU Law Review*, **82**, pp. 870–912.

Bryson, A., Buraimo, B., Farnell, A. and Simmons, R. (2024). Special ones? The effect of head coaches on football team performance. *Scottish Journal of Political Economy*, **71**(3), pp. 295–322.

Cannon, A. J., Fisher, J. D., Fellingham, G. W. and Page, G. L. (2025). Analyzing the effects of NBA head coaches. *Journal of Quantitative Analysis in Sports*. DOI: 10.1515/jqas-2025-0025.

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.

David, J. A., Pasteur, R. D., Ahmad, M. S. and Janning, M. C. (2011). NFL prediction using committees of artificial neural networks. *Journal of Quantitative Analysis in Sports*, **7**(2), pp. 1–15.

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.

Gilfix, Z., Meyerson, J. and Addona, V. (2020). Longevity differences in the tenures of American and foreign Major League Soccer managers. *Journal of Quantitative Analysis in Sports*, **16**(1), pp. 17–26.

Lock, D. and Nettleton, D. (2014). Using random forests to estimate win probability before each play of an NFL game. *Journal of Quantitative Analysis in Sports*, **10**(2), pp. 197–205.

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. *International Journal of Sports Science & Coaching*, **2**(2), pp. 105–108.

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

Pro-Football-Reference.com (2025). Pro-Football-Reference. Sports Reference LLC. Available at https://www.pro-football-reference.com/.

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

Urschel, J. D. and Zhuang, J. (2011). Are NFL coaches risk and loss averse? Evidence from their use of kickoff strategies. *Journal of Quantitative Analysis in Sports*, **7**(3).

Wolfson, J., Addona, V. and Schmicker, R. H. (2011). The quarterback prediction problem: forecasting the performance of college quarterbacks selected in the NFL draft. *Journal of Quantitative Analysis in Sports*, **7**(3), pp. 1–22.

Yurko, R., Ventura, S. and Horowitz, M. (2019). nflWAR: a reproducible method for offensive player evaluation in football. *Journal of Quantitative Analysis in Sports*, **15**(3), pp. 163–183.

A Feature Descriptions

Tab. 7: 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. 8: 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. 9: 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. 10: 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

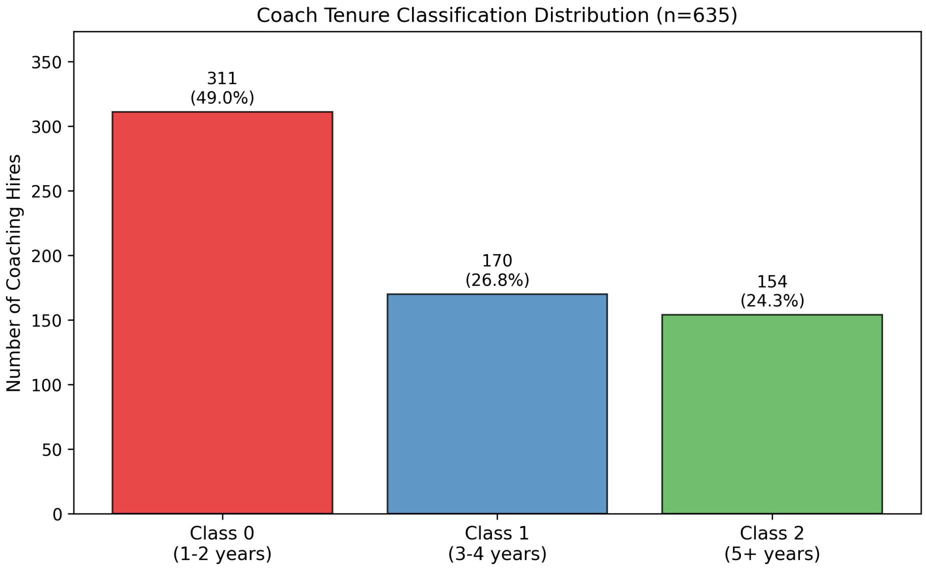

Fig. 3: Coach tenure classification frequency distribution across all 635 coaching hire instances with known tenure outcomes. Class 0: short tenure (1–2 years, 49.0%); Class 1: medium tenure (3–4 years, 26.8%); Class 2: long tenure (5+ years, 24.3%).

C Partial Dependence Plots

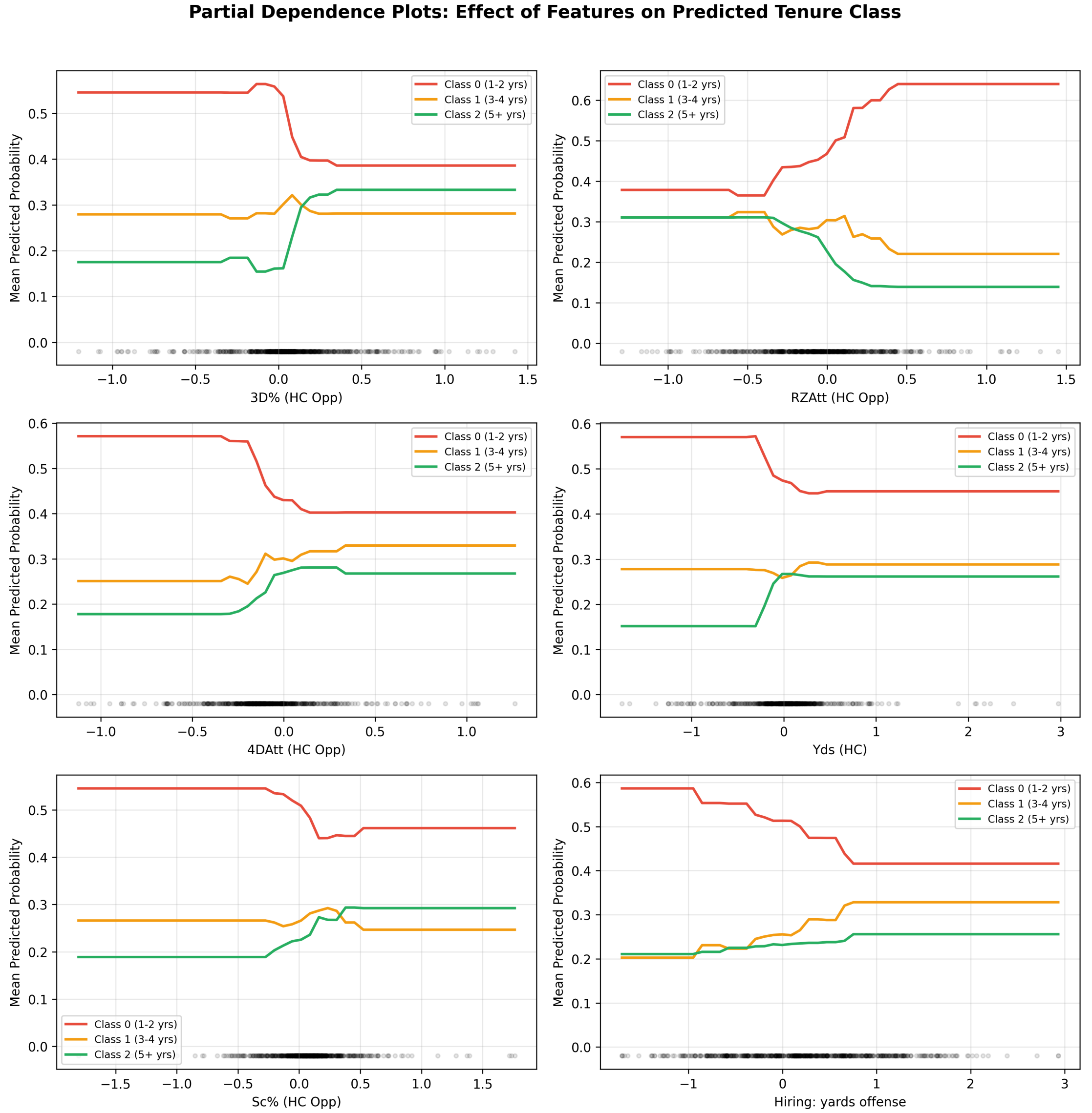

Fig. 4: Partial dependence plots for the six most predictive features. Each panel shows how the predicted probability of each tenure class changes as the feature value varies, with all other features held at their observed values. Feature values are normalized relative to league averages.

D Model Hyperparameters

Tab. 11: Hyperparameter search space for randomized search (1,000 iterations, 5-fold coach-level CV, scored by QWK). Values are sampled uniformly from each discrete set.

Hyperparameter	Candidate Values
Number of Estimators	{25, 50, 100, 200}
Learning Rate	{0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40}
Max Depth	{2, 3, 4}
Gamma	{0, 0.01, 0.05, 0.1}
Lambda (L2 Regularization)	{0, 0.01, 0.1, 0.5}
Alpha (L1 Regularization)	{0, 0.01, 0.1}
Subsample	{0.80, 0.85, 0.90, 0.95, 1.00}
Colsample by Tree	{0.80, 0.85, 0.90, 0.95, 1.00}
Minimum Child Weight	{1, 2, 3, 4, 5}

Tab. 12: 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
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