BAX452: Machine Learning Final Project Report

Kyle Ayisi, Rohan Saxena, Jeremy Ampofo

Executive Summary

This project develops a machine learning model to predict NFL “breakout” players—young athletes poised to dramatically improve their performance. In a high-stakes environment like fantasy football and sports betting, correctly forecasting breakout seasons offers significant competitive advantage. Using over 21 seasons of data (2003–2023) and focusing on key positions (running backs, wide receivers, tight ends), we utilized more than 70 features combining traditional statistics and advanced efficiency metrics. Our best-performing ensemble model achieved approximately 91% accuracy and identified nearly 47% of actual breakouts in hold-out tests. These predictive insights can guide team roster decisions, inform betting strategies, and improve fantasy draft outcomes.

Background

With the growing sports analytics field, understanding and even predicting athlete performance is at the forefront of the minds of not only analysts, but the everyday fan. Identifying a breakout player—a young athlete who leaps into an elite performance tier—is a persistent challenge in NFL analytics. Traditionally, scouting and subjective analysis have been used; however, the increasing availability of advanced performance metrics now enables data-driven predictions.

What is a Breakout Player?

In our work, a breakout is defined as:

- A player with minimal experience (years 1-3)
- A player who is “elite” relative to their peers at their position (e.g., top 15 for RBs/WRs or top 10 for TEs)
- A player who exhibits at least a 25% improvement in key production metrics.

The focus on offensive skill positions is due to their historical propensity for dramatic year-over-year improvement, offering a clear target for modeling and tangible benefits for stakeholders such as teams, fantasy managers, and betting markets. By defining a breakout star this way, we ensure that the threshold is high enough to recognize the truly great new players in the league.

Analysis

Data and Feature Engineering

Our dataset spans 21 NFL seasons and includes both traditional statistics (e.g., total yards, touchdowns, snap counts) and advanced metrics (e.g., yards per route run, efficiency ratios, and usage trends). We preprocess data through:

- **Cleaning and Scaling:** Addressing missing values and standardizing metrics.
- **Feature Engineering:** Creating opportunity metrics (targets per game, snap percentage), efficiency indicators (yards per route, per-touch production), and interaction features that capture position-specific trends.

Exploratory analysis showed that per-opportunity performance (e.g., a wide receiver’s yards per route run) was more predictive of future success than raw volume statistics.

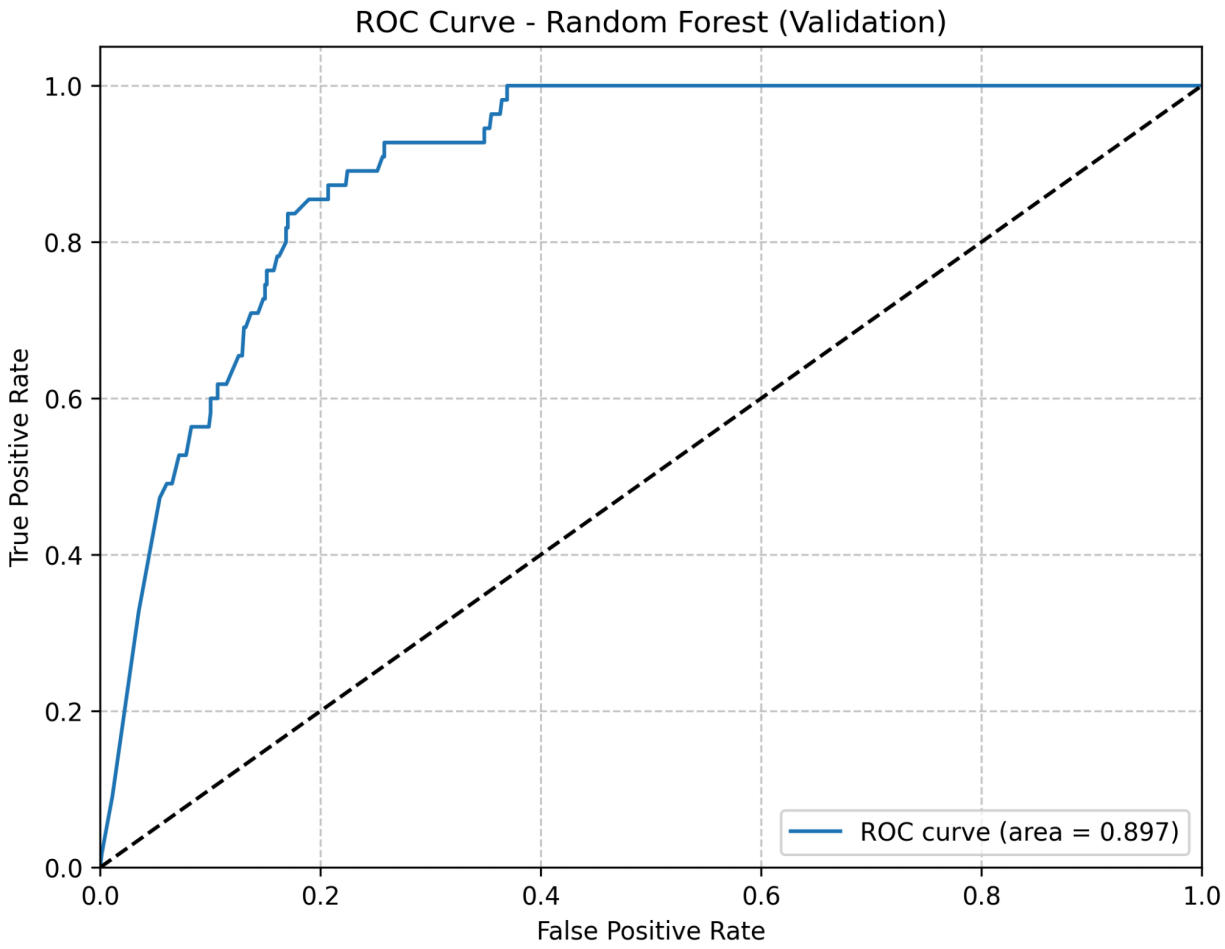
Model Selection and Evaluation

We experimented with several models:

- **Logistic Regression:** With L1 regularization and class weighting to handle imbalance, offering clear interpretability of feature contributions. This was our base model used to give us a general idea of whether our train of thought provided a valid approach to classifying players.
- **Random Forest Ensemble:** Configured with 200 trees (max depth 10) to capture nonlinear interactions and mitigate overfitting. These hyperparameters were reached through grid-search optimization.
- **Other Approaches:** Preliminary tests with gradient boosting and a simple neural network were performed, but the tree-based methods provided a balance of accuracy and explainability.

A time-aware validation scheme was employed (training on earlier seasons and testing on more recent seasons) to simulate real-world forecasting. Evaluation metrics included accuracy, recall (critical for catching as many true breakouts as possible), precision, and ROC-AUC. Our random forest model achieved an AUC of approximately 0.90 and a recall rate near 47%, which speaks to its ability to capture emerging stars, though reflecting its need for improvement.

Figure 1. ROC Curve for Random Forest Model



Recommendations

Based on our findings, we propose the following:

- **For NFL Teams and Scouts:** Use the model to identify undervalued players showing efficiency and rising usage, enabling proactive roster moves before market value increases. With this analysis and forthcoming refinement, NFL executives can peek into the future and set their teams up for success.

- **For Fantasy Sports Platforms:** This analysis highlights potential breakout candidates, allowing managers to gain a competitive edge by drafting low-cost, high-upside players, leading to increased league success.
- **For Betting Markets:** Integrate model predictions to adjust player prop lines. Accurate forecasting of breakouts can inform bets on player performance, benefiting both sportsbooks and bettors. For this in particular, high accuracy is important so as to best inform lines. This use case is less concrete, as betting is a highly-regulated industry.

It is recommended that stakeholders treat the model's outputs as one part of a broader decision-making process, integrating qualitative scouting and contextual factors.

Limitations and Future Directions

While promising, the project has limitations:

- **Breakout Definition and Scope:** Our criteria (top tier, 25% improvement) are somewhat subjective and focus only on RBs, WRs, and TEs. Extending the model to include quarterbacks or defensive players may require redefining features and metrics.
- **Data Constraints:** Despite spanning 21 seasons, the dataset remains modest by machine learning standards, and advanced metrics are only available for recent years. Future work should integrate more comprehensive data and address missing values.
- **Trade-offs in Model Performance:** Our focus on high accuracy comes at the cost of either a high precision or high recall. We choose to balance precision and recall to increase overall model accuracy. Users must be aware of potential false positives and adjust decision thresholds based on specific use cases. The nature of our business question does lend itself to false positives because there are so many things that can cause an upward trending player to have an off year and not breakout.

- **Unmodeled Factors:** External influences such as injuries, coaching changes, and off-field issues remain unaccounted for, limiting the model's predictive power in real-world, dynamic environments.

Future enhancements could involve expanding the positional scope, incorporating time-series models (i.e. LSTMs) to capture player trajectory more effectively, and integrating additional contextual data like injury history and coaching strategies. By addressing these limitations, the model's accuracy and practical utility can be further improved.

Conclusions

Our study demonstrates that a machine learning model can effectively predict NFL breakout players by combining decades of historical data with advanced performance metrics. The model confirms that efficiency and opportunity metrics—rather than sheer volume—are key predictors of future performance. Although our best model (a random forest ensemble) is not infallible, it provides meaningful guidance by identifying a high percentage of true breakout cases. This predictive capability can transform how teams scout talent, how fantasy managers draft players, and how sportsbooks set odds. Ultimately, the work paves the way for further refinement and wider adoption of data-driven strategies in NFL player evaluation.