EVALUATING MACHINE LEARNING MODELS AND STACKING TECHNIQUES FOR PREDICTING NFL GAME OUTCOMES AND BETTING TRENDS

Sanchay Bhutani
sanchay.bhutani@student-cs.fr

Sagar Vishishta
sagar.vishishta@student-cs.fr

Ayushmaan Khalkho
ayushmaan.khalkho@student-cs.fr

ABSTRACT

This project aims to explore and compare multiple machine learning models for predicting National Football League (NFL) game outcomes and betting results. Using historical game, weather, and betting data from 1966–2024, we evaluate baseline models such as Logistic Regression, Random Forests, Decision Trees, and Gradient Boosting methods. We further investigate whether stacking these models yields statistically significant improvements in predictive performance. The results of this study can offer practical insights for sports analytics, betting markets, and data-driven decision-making in sports forecasting. We further investigate whether stacking these models yields statistically significant improvements compared to individual models, highlighting the impact of ensemble learning on structured sports data.

1 MOTIVATION AND PROBLEM DEFINITION

Sports analytics has become a major domain of machine learning applications, especially in predicting outcomes, optimizing strategies, and informing betting markets. The National Football League (NFL), with its rich historical data, offers a unique opportunity to examine how data-driven models can predict game results and betting outcomes.

Formal Problem Definition: Let $X \in \mathbb{R}^{n \times m}$ denote the characteristic matrix of NFL games, including team stats, weather, and betting odds, and $y \in \{0, 1\}$ denote whether the favorite team covered the spread. We aim to learn a function $f : X \to y$ that maximizes predictive performance (e.g., ROC AUC).

Motivation: Accurate prediction of NFL results not only improves analytical understanding of team performance, but also supports betting markets, fan engagement, and sports journalism. Beyond practical utility, this problem is also an ideal testbed for comparing supervised learning algorithms on structured, real-world, and temporally correlated data.

Applications:

- Sports analytics for forecasting win probabilities and betting line movements.
- Development of ensemble-based predictive models for financial and risk-related domains.
- Benchmarking of stacking ensembles across heterogeneous ML models.

Related Work: Prior studies have examined the prediction of sports outcomes using logistic regression and ensemble learning methods. Recent research has expanded into hybrid and deep learning techniques to integrate team statistics, odds, and weather information ((McHale et al., 2011), (Horvat et al., 2019), (Tsiligiridis et al., 2020)). However, few have systematically evaluated the comparative effect of stacking classical models on sports data sets. Logistic regression is interpretable but may underfit; Random Forest and XGBoost capture nonlinear interactions but risk overfitting; KNN models local relationships but is sensitive to scaling. Stacking ensembles are designed to combine

the complementary strengths of these models ((McHale et al., 2011; Horvat et al., 2019; Tsiligiridis et al., 2020; **?**)).

2 METHODOLOGY

Data: The dataset contains NFL game results since 1966, including game scores, weather conditions, and odds of a bet from multiple public sources such as ESPN, Pro Football Reference, and NFL.com. Betting lines are sourced from `spreadspoke.com` and other historical repositories.

Preprocessing: We will clean and encode relevant features, normalize numerical attributes, and handle categorical variables such as team names and locations. The outcome variable will be binary — whether the favorite team covered the spread. Missing values are imputed, categorical variables (team names, locations) are one-hot encoded, and numerical features are normalized. Temporal order is preserved to prevent data leakage.

Baseline Models:

- Logistic Regression (LRG)
- K-Nearest Neighbors (KNB)
- Gaussian Naive Bayes (GNB)
- Extreme Gradient Boosting (XGB)
- Random Forest (RFC)
- Decision Tree (DTC)

Model Justification: Logistic Regression (linear), KNN (local patterns), GNB (probabilistic baseline), RFC/DTC (nonlinear interactions), XGB (efficient gradient boosting). Stacking ensembles combine these models to leverage complementary decision boundaries.

The models will be trained and evaluated using 5-fold cross-validation with ROC AUC as the primary metric.

Stacking Ensemble: A metamodel (e.g., logistic re.g.ression or gradient enhancement) will be trained on the out-of-fold predictions of base learners to evaluate performance improvement. The comparison will highlight whether model stacking leads to statistically significant gains in predictive accuracy.

Tools: We will employ Python with libraries including `scikit-learn`, `xgboost`, and `pandas` for data preprocessing, training, and evaluation.

3 EVALUATION

Performance Metrics:

- ROC AUC for model discrimination.
- Mean and standard deviation across 5-fold CV splits.
- Statistical tests (e.g., paired t-test) to evaluate whether stacking provides a significant improvement over individual models.

Experimental Design:

1. Train and validate six baseline classifiers.
2. Implement a stacking ensemble using their predictions as meta-features.
3. Compare ROC AUC and accuracy across models.
4. Analyze the importance of the features and the calibration of the model.

Expected Outcome: We expect ensemble methods, particularly stacking, to outperform individual classifiers by effectively combining complementary decision boundaries. However, we also

anticipate that model complexity and correlation among base learners will influence the extent of improvement. The correlation among the base learners and the overfit may reduce the stacking benefit.

4 REFERENCES

REFERENCES

I. G. McHale and D. E. Morton. Predicting sports outcomes using statistical models. *Journal of the Royal Statistical Society*, 2011.

T. Horvat and A. Job. Machine learning methods for sports result prediction: A review. *Information*, 2019.

T. Tsiligiridis and D. Ntakolia. Data Mining and Machine Learning in Sports: A systematic review. *Applied Sciences*, 2020.

T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD Conference*, 2016.

I. Witten, E. Frank, M. Hall, and C. Pal. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2017.