sports_betting_ml

Sports Betting using ML

CS4641 Project Group 74 Final

Akina, Maxwell, Elliot, Tyler, James

Recent studies show the potential of machine learning (ML) in sports betting, particularly in football. Stübinger et al. (2018) developed a ML framework to predict match outcomes. Their study on over 8,000 games found that ML models outperform both traditional betting strategies, yielding statistically and economically significant returns. Hubáček et al. (2019) introduced a convolutional neural network-based model for sports betting that incorporates player-level statistics. By reducing correlation with bookmaker predictions, their approach systematically generated profits with NBA data. Stübinger et al. (2020) combined player characteristics with ML in over 47,000 matches, delivering a consistent 1.58% profit per match. Ruzička and Chovanec (2019) found that using long-term data improved prediction accuracy, with random forest and logistic regression being particularly effective.

Our dataset comes from Pro Football Focus' Premium position stats for NFL players. It includes advanced metrics such as adjusted completion percentage and allowed pressure for quarterbacks, receiving vs scheme for wide receivers, and PFF's own grades per player. This dataset will help us make predictions for player prop bets based on historical and projected stats. Link.

Problem: Predicting sports outcomes is tough, and many bettors rely on gut feelings or basic stats, often leading to inconsistent results. While plenty of data exists, it's too complex for manual analysis and not fully utilized.

Motivation: Machine learning offers a way to improve sports betting predictions. Leveraging ML to analyze trends can lead to smarter, data-driven bets, moving away from intuition. It's also an exciting use of modern technology in sports analytics.

Methodology

This project aimed to predict NFL player performance metrics for the following week, focusing on receiving yards, receptions, and touchdowns. The workflow was structured into the following steps:

1. Data Collection and Transformation

   The data was sourced from graphical outputs and manually converted into CSV format. While automating the process for all data types was explored, it proved time-intensive, so the focus remained on receiving data.

   Weekly receiving data was collected for individual players, including metrics such as targets, receptions, receiving yards, and touchdowns.

2. Data Preprocessing

   Weekly data was appended to create a time-series structure, organized by player and pass type (e.g., short, deep).

Aggregation methods were applied to summarize stats for each player across the weeks. For example, key metrics were averaged or totaled to provide meaningful inputs for the models.

Missing values were addressed using mean imputation, ensuring consistency and reducing bias in the dataset.

The dataset was updated weekly as new data became available, reflecting ongoing player performance trends.

3. Feature Engineering

   Relevant features such as receiving depth, pass type, and yards after catch (YAC) were included to capture the nuances of player performance.

   By aligning weekly stats with specific players and play types, the dataset provided a granular view of receiving performance.

4. Model Development

   Three predictive models were developed to estimate next week's performance:

   Ridge Regression: A linear model with L2 regularization to handle multicollinearity and avoid overfitting.

   Random Forest: A non-linear ensemble learning model that excels in capturing complex interactions between features.

   Neural Network: A deep learning approach designed to identify intricate patterns in the data.

5. Neural Network Hyperparameter Tuning

   The Neural Network underwent hyperparameter tuning to optimize its performance. Adjustments were made to:

   The number of layers and neurons.

   Learning rate and optimization algorithm.

   Regularization techniques (e.g., dropout or L2 penalty).

   Batch size and epochs for training.

   This iterative tuning process aimed to balance model complexity with generalization, minimizing overfitting while capturing complex trends.

6. Model Evaluation

   All models were evaluated using Root Mean Squared Error (RMSE) to assess their predictive accuracy. RMSE was chosen for its interpretability and sensitivity to large errors, which is critical in sports prediction.

   Comparative analysis was conducted to understand the strengths and weaknesses of each model:

   Ridge Regression offered stability and consistency.

   Random Forest captured game-to-game variability but required sufficient data to stabilize predictions.

   Neural Networks demonstrated superior performance on non-linear patterns but were sensitive to small datasets and required extensive tuning.

7. Iterative Updates and Improvements

As new weekly data became available, the dataset and models were updated to reflect the latest player performance trends.

Observations from model performance guided further preprocessing adjustments and model refinements, such as re-engineering features or re-evaluating hyperparameters.

By following these steps, the project successfully built a pipeline for predicting NFL player performance metrics, balancing data limitations with model complexity to achieve meaningful insights.

Quantitative Metrics

We developed a machine learning pipeline that predicts NFL player receiving yards on a per-game basis, focusing on Atlanta Falcons players using PFF data. Our evaluation, based on Root Mean Squared Error (RMSE), highlighted the following:

- Ridge Regression yielded the best average RMSE across all time-slice folds with an average RMSE of 102.
- Random Forest Regressor produced a slightly lower RMSE in the final fold, with a score of 44 compared to 66 for Ridge Regression, though it averaged higher across earlier folds at 124 RMSE.
- Neural Network achieved an average RMSE of 80.89 for yards, 7.61 for receptions, and 1.18 for touchdowns, making it the strongest performer overall for receptions and touchdowns.

RMSE is a strong metric for evaluating NFL player performance predictions because it directly measures the average prediction error in the same units as the target variables (e.g., yards, receptions, touchdowns), making it easy to interpret. It penalizes large errors more heavily due to squaring, which is important for capturing the impact of outlier performances that are common in football. RMSE works well with continuous variables like yards and receptions and aligns with the optimization objectives of many machine learning models. Also, its single scalar value allows for straightforward comparison across different models, helping identify the best-performing approach.

Analysis of Models

1. Ridge Regression: Using Ridge Regression, we observed a stable performance across time-slice folds, with an average RMSE of 102. Ridge Regression's regularization appears effective in handling the limited data available per game, reducing the influence of outlier weeks and providing balanced predictions across the season. This stability in predictive performance, though yielding higher RMSE in the last fold, demonstrates the model's consistency in generalizing trends from past games.

2. Random Forest Regressor: Random Forest performed more variably across the time-slice folds, averaging a higher RMSE of 124 over earlier folds but outperforming Ridge Regression in the final fold with an RMSE of 44. This suggests that Random Forest may be better at capturing game-to-game variability, though it requires sufficient data to balance the effects of high-variance game weeks. The model's stronger performance on the last fold indicates its ability to capture complex patterns as more data accumulates, though it remains sensitive to smaller datasets.

3. Neural Network: The Neural Network performed competitively, particularly excelling in predicting receptions and touchdowns, with RMSEs of 7.61 and 1.18, respectively. For yards, it achieved an average RMSE of 80.89, outperforming Ridge Regression and Random Forest overall. The Neural Network's ability to model intricate patterns in the data makes it particularly effective for these metrics, though it exhibited variability across folds, such as a high RMSE of 156 in earlier folds. This suggests a sensitivity to limited data in certain weeks, possibly due to overfitting or instability during training.

Comparison of Models

Ridge Regression demonstrates consistent performance with stable RMSE values across most time-slice folds, suggesting it effectively balances predictive accuracy and regularization in the presence of limited data. However, its higher RMSE in the final fold highlights a potential weakness in adapting to sudden shifts or trends in recent weeks. On the other hand, Random Forest Regressor exhibits more variability across folds, initially producing higher RMSEs but significantly outperforming Ridge Regression in the final fold. This suggests that Random Forest excels in capturing complex game-to-game variability but requires a larger dataset to stabilize its predictions.

The Neural Network stands out with superior performance on receptions and touchdowns, leveraging its ability to model non-linear patterns effectively. Its competitive RMSE for yards also displays its potential to outperform traditional models like Ridge Regression and Random Forest in capturing intricate relationships within the data. However, its variability in earlier folds, with notably higher RMSE, indicates a sensitivity to small sample sizes and potential overfitting. While the Neural Network shows the most promise overall, its performance depends heavily on careful tuning and sufficient data, whereas Ridge Regression offers a more stable but less dynamic alternative, and Random Forest serves as a middle ground with potential for late-season adaptability.

Insights and Interpretation

Switching to game-by-game data provided a more dynamic and responsive model, able to adjust predictions based on recent performance trends rather than aggregating over a season. This approach allows for a more granular analysis of player performance, although the limited amount of data (only thirteen weeks) affects the models' ability to generalize.

- Ridge Regression remains the most stable and consistent across folds, providing a reliable baseline for predictions when data is limited.
- Random Forest captures week-to-week variability well, but its higher variability across folds highlights its reliance on larger datasets.
- Neural Networks demonstrate strong potential, particularly for predicting receptions and touchdowns, but require further refinement to stabilize performance in early folds.

Next Steps

1. **Expand Data Collection**: As the NFL season progresses we will have access to more data, the models, especially Random Forest and Neural Networks, will have improved performance and reduced variance.
2. **Additional Features**: Including more contextual features, such as game conditions, injuries, and team dynamics may enhance predictive accuracy
3. **Ensemble Models**: Combine the strengths of Ridge Regression, Random Forest, and Neural Networks into an ensemble to maximize predictive performance and balance their trade-offs

Project Evolution and Justification for Changes

Initially, our project aimed to predict season totals across multiple player positions, including passing, rushing, and receiving grades, to provide comprehensive insights for sports betting. However, the broad scope and complexity of this approach led us to narrow our focus to receiving metrics. This shift allowed us to refine our analysis and focus on game-by-game receiving data specifically for Atlanta Falcons players, enabling more granular and dynamic predictions.

As the project evolved, we introduced several key enhancements to improve model performance and contextual accuracy: 1. Incorporation of Opponent Defensive Rank: We added weekly opponent defensive rankings as a feature

in our dataset. This contextual information provides critical insight into how the strength of opposing defenses influences player performance, allowing the models to make more nuanced predictions. 2. Expanded Model Selection: In addition to Ridge Regression and Random Forest Regressor, we implemented a Neural Network to explore its potential for capturing complex, non-linear relationships in the data. This addition provided a significant boost in accuracy, particularly for metrics like receptions and touchdowns. 3. Time-Slice Cross-Validation: To better simulate real-world, week-to-week predictions, we adopted a time-slice cross-validation technique. This approach tests the models on unseen future data, mimicking the dynamic nature of predicting player performance in a live season context.

These changes have enhanced the adaptability and interpretability of our machine learning pipeline. Ridge Regression remains a reliable baseline for stable predictions, while Random Forest excels at modeling variability with larger datasets. The Neural Network has demonstrated its strength in capturing intricate patterns and provided the most accurate results overall for receptions and touchdowns. Combined with the addition of contextual features like defensive rank, our pipeline now offers a more comprehensive and responsive framework for predicting game-by-game player performance.

References

- Matt Gifford, Tuncay Bayrak, *A predictive analytics model for forecasting outcomes in the National Football League games using decision tree and logistic regression*, Decision Analytics Journal, Volume 8, 2023, 100296, ISSN 2772-6622.
- Stübinger, J., & Knoll, J. (2018). *Beat the Bookmaker: Winning Football Bets with Machine Learning*. In M. Bramer & M. Petridis (Eds.), Artificial Intelligence in Science and Industry, 219-233. Springer.
- Hubáček, O., Šourek, G., & Železný, F. (2019). *Exploiting the Sports-Betting Market Using Machine Learning*. International Journal of Forecasting, 35(3), 783–796.
- Stübinger, J., Mangold, B., & Knoll, J. (2020). *Machine Learning in Football Betting: Prediction of Match Results Based on Player Characteristics*. Applied Sciences, 10(1), 46.
- Ruzička, M., & Chovanec, M. (2019). *The Application of Machine Learning Principles in Sports Betting Systems*. Acta Electrotechnica et Informatica, 19(3), 16-20.
- PFF Weekly NFL Receiving Data

Member Contribution

| Member | Contribution | |———–|————————————–| | Elliot | Model Selection, Data Preprocessing, Model Implementations, References| | Tyler | Data Preprocessing | | Akina | Exploratory Data Analysis & Visualizations, Data Sourcing | | Maxwell | Results Evaluation and Analysis, Data Sourcing | | James | Data Preprocessesing, Model Testing and Implementation |

Reflection on Project

This project provided valuable insights into modeling NFL player performance, showing the challenges and rewards of working with week-by-week data. Collaborating as a group allowed us to tackle data collection, preprocessing, and analysis, also allowing for diverse perspectives to refine our models. While the variability in player performance and limited data posed challenges, our exploration of Ridge Regression, Random Forest, and Neural Networks displayed the trade-offs between stability, complexity, and adaptability. Overall, this experience deepened our understanding of predictive modeling in sports and demonstrated the importance of aligning models with the nuances of the data and problem domain.

———–OLD————-

CS4641 Project, Proposal. (scroll for Midterm Checkpoint)

Akina, Maxwell, Elliot, Tyler, James

Recent studies show the potential of machine learning to improve sports betting strategies, particularly in football. Stübinger et al. (2018) developed a machine learning framework using large-scale data to predict football match outcomes. Their study on over 8,000 matches found that machine learning models outperform both linear regression and naive betting strategies, yielding statistically and economically significant returns. Hubáček et al. (2019) introduced a convolutional neural network-based model for sports betting that incorporates player-level statistics. By reducing correlation with bookmaker predictions, their approach systematically generated profits with NBA data. Stübinger et al. (2020) combined player characteristics with machine learning, applied to over 47,000 matches, delivering a consistent 1.58% profit per match. Ruzička and Chovanec (2019) found that using long-term data improved prediction accuracy, with random forest and logistic regression being particularly effective.

Our dataset comes from Pro Football Focus' Premium position stats for NFL players. It includes advanced metrics for each position and PFF's in-house grades for every player. This dataset will help us make predictions for player prop bets based on historical and projected stats. Link (requires subscription).

Problem: Predicting sports outcomes is tough, and many bettors rely on gut feelings or basic stats, often leading to inconsistent results. While there's plenty of data available, it's too complex for manual analysis, meaning most bettors don't fully utilize it.

Motivation: Machine learning offers a way to improve sports betting predictions. Leveraging ML to analyze trends can lead to smarter, data-driven bets, moving away from intuition. It's also an exciting use of modern technology in sports analytics.

Methods

Data Preprocessing:

- Handle missing data
- Normalize odds
- Convert categorical data (team names, player positions) into numerical values
- Handle time-series data

Machine Learning Models:

- Logistic Regression to predict game outcomes
- Random Forest to capture relationships between player statistics and game outcomes
- Neural Network (need specifics from EDA for model type)

(Potential) Results and Discussion

We expect logistic regression to serve as a reliable baseline for predicting NFL game outcomes. Random forests, given their ability to handle complex player statistics, should improve prediction accuracy. Neural networks, using player metrics, could further enhance performance.
Our models aim to leverage player-level data to improve sports betting predictions. We expect random forests and neural networks to outperform simpler models like logistic regression by capturing deeper patterns in player performance data. If successful, these models could lead to more informed, data-driven betting strategies.

References

- Stübinger, J., & Knoll, J. (2018). *Beat the Bookmaker: Winning Football Bets with Machine Learning*. In M. Bramer & M. Petridis (Eds.), Artificial Intelligence in Science and Industry, 219-233. Springer.
- Hubáček, O., Šourek, G., & Železný, F. (2019). *Exploiting the Sports-Betting Market Using Machine Learning*. International Journal of Forecasting, 35(3), 783–796.
- Stübinger, J., Mangold, B., & Knoll, J. (2020). *Machine Learning in Football Betting: Prediction of Match Results Based on Player Characteristics*. Applied Sciences, 10(1), 46.
- Ruzička, M., & Chovanec, M. (2019). *The Application of Machine Learning Principles in Sports Betting Systems*. Acta Electrotechnica et Informatica, 19(3), 16-20.

Member Contribution

| Member | Contribution | |———–|————————————————–| | Elliot | Literature Review, Problem/Motivation, Results & Discussion, References, GitHub Page | | Tyler | Methods | | Akina | Video Presentation, Results & Discussion | | Maxwell | Dataset, Dataset Description | | James | Video Presentation, Literature Review, Gantt Chart |

Midterm Checkpoint

Results and Discussion

Quantitative Metrics

We developed a machine learning pipeline that predicts NFL player receiving yards on a per-game basis, focusing on Atlanta Falcons players using PFF data. Our evaluation, based on Root Mean Squared Error (RMSE), highlighted the following:

- Ridge Regression yielded the best average RMSE across all time-slice folds with an average RMSE of 102.
- Random Forest Regressor produced a slightly lower RMSE in the final fold, with a score of 44 compared to 66 for Ridge Regression, though it averaged higher across earlier folds at 124 RMSE.

Analysis of Models

1. Ridge Regression: Using Ridge Regression, we observed a stable performance across time-slice folds, with an average RMSE of 102. Ridge Regression's regularization appears effective in handling the limited data available per game, reducing the influence of outlier weeks and providing balanced predictions across the season. This stability in predictive performance, though yielding higher RMSE in the last fold, demonstrates the model's consistency in generalizing trends from past games.

2. Random Forest Regressor: Random Forest performed more variably across the time-slice folds, averaging a higher RMSE of 124 over earlier folds but outperforming Ridge Regression in the final fold with an RMSE of 44. This suggests that Random Forest may be better at capturing game-to-game variability, though it requires sufficient data to balance the effects of high-variance game weeks. The model's stronger performance on the last fold indicates its ability to capture complex patterns as more data accumulates, though it remains sensitive to smaller datasets.

Insights and Interpretation

Switching to game-by-game data provided a more dynamic and responsive model, able to adjust predictions based on recent performance trends rather than aggregating over a season. This approach allows for a more granular analysis of player performance, although the limited amount of data (only nine weeks) affects the models' ability to generalize. The Ridge Regression model's consistency suggests it may be a more reliable option when limited data is available, while Random Forest shows potential for capturing week-to-week variability given a larger dataset.

Next Steps

1. **Expand Data Collection**: We will get more data throughout the season as weeks progress, which we can then use to further refine our model and make more accurate week-to-week projections.
2. **Additional Features**: Incorporate game context features such as opponent strength which may improve predictions by providing context for each game's performance.

Project Evolution and Justification for Changes

Initially, the project aimed to use season totals across multiple positions, including passing, rushing, and receiving grades, to provide comprehensive player predictions for sports betting comparisons. However, the complexity of such a broad approach led us to narrow our scope to receiving metrics. Ultimately, we chose to focus on game-by-game receiving depth data specifically for Atlanta Falcons players to achieve a more granular and responsive analysis.

This pivot allowed us to implement a time-slice cross-validation technique that better simulates real-world, week-to-week predictions. Our Ridge Regression and Random Forest models now provide game-level insights, with a focus on optimizing predictions for individual games rather than season totals. This change has improved model interpretability, providing us with an adaptable framework that can be further refined as more data becomes available.

References

- Matt Gifford, Tuncay Bayrak, *A predictive analytics model for forecasting outcomes in the National Football League games using decision tree and logistic regression*, Decision Analytics Journal, Volume 8, 2023, 100296, ISSN 2772-6622.
- Stübinger, J., & Knoll, J. (2018). *Beat the Bookmaker: Winning Football Bets with Machine Learning*. In M. Bramer & M. Petridis (Eds.), Artificial Intelligence in Science and Industry, 219-233. Springer.
- Hubáček, O., Šourek, G., & Železný, F. (2019). *Exploiting the Sports-Betting Market Using Machine Learning*. International Journal of Forecasting, 35(3), 783–796.
- Stübinger, J., Mangold, B., & Knoll, J. (2020). *Machine Learning in Football Betting: Prediction of Match Results Based on Player Characteristics*. Applied Sciences, 10(1), 46.
- Ruzička, M., & Chovanec, M. (2019). *The Application of Machine Learning Principles in Sports Betting Systems*. Acta Electrotechnica et Informatica, 19(3), 16-20.
- PFF Weekly NFL Receiving Data

Member Contribution

| Member | Contribution | |———–|————————————————–| | Elliot | Model Selection, Data Preprocessing, Model Implementations, References| | Tyler | Data Preprocessing | | Akina | Exploratory Data Analysis & Visualizations, Data Sourcing | | Maxwell | Results Evaluation and Analysis, Data Sourcing | | James | Data Preprocessesing, Model Testing and Implementation |