Jonathan Koski

Final Report:
NFL Matchup Prediction Analysis

1. Problem Statement

Predicting the outcome of an NFL game is an inherently complex task due to the dynamic and high-variance nature of football. Seasons have small pools of data being limited to sixteen regular season games. Coaching strategies, injuries, weather and psychological factors all interact to impact the outcome of a game. Due to the high noise of football games, a data driven approach is valuable.

Building a model for NFL outcomes offers several practical benefits for teams, analysts, sportsbooks, bettors, and fans alike. Overall identifying which factors influence the outcome of a game most can turn raw data into actionable insights for these parties.

2. The Data

Game level data sourced from pro-football-reference.com was scraped for the 2017-2024 seasons and merged into a single dataframe. Our initial data frame consisted of 2200+ rows each representing an individual game, with 14 columns covering the data and time of matchups, the teams facing off, points, yards, and turnovers by the winning and losing teams, and an away game indicator for the winning team.

3. Data Wrangling

The Data Wrangling phase began with basic data cleaning, including standardizing historical team names by mapping them to their current counterparts to ensure consistency across all seasons. Next, the dataset was restructured to eliminate directional indicators such as “winner/loser” labels and the “away game” flag. Instead, statistics were aligned by home and away teams, and a new boolean variable, home_win, was introduced to represent the game outcome from the home team's perspective.

Following this restructuring, we conducted initial feature engineering. Specifically, we calculated rolling averages over the previous five games for key performance metrics such as yards gained, points scored, and turnovers. These rolling features were computed separately for both the home and away teams, providing a snapshot of each team’s recent form heading into a game. In addition, we derived differential features capturing the difference in recent performance between the competing teams.

To ensure data quality and model reliability, we dropped outliers, and excluded tie games from the dataset, as they do not provide a clear outcome for binary classification. Finally, the cleaned and engineered dataset was saved and subsetted for modeling purposes. Our new Data Frame consisted of 2034 rows, 36 columns.

4. Data Exploration

Initial data exploration was kept relatively minimal. Focus was primarily on verifying data quality and understanding general distributions. Visualizations used to confirm rolling averages were behaving as expected. While this step did not involve deep exploratory analysis it was a useful sanity check to have confidence in the data moving forward.

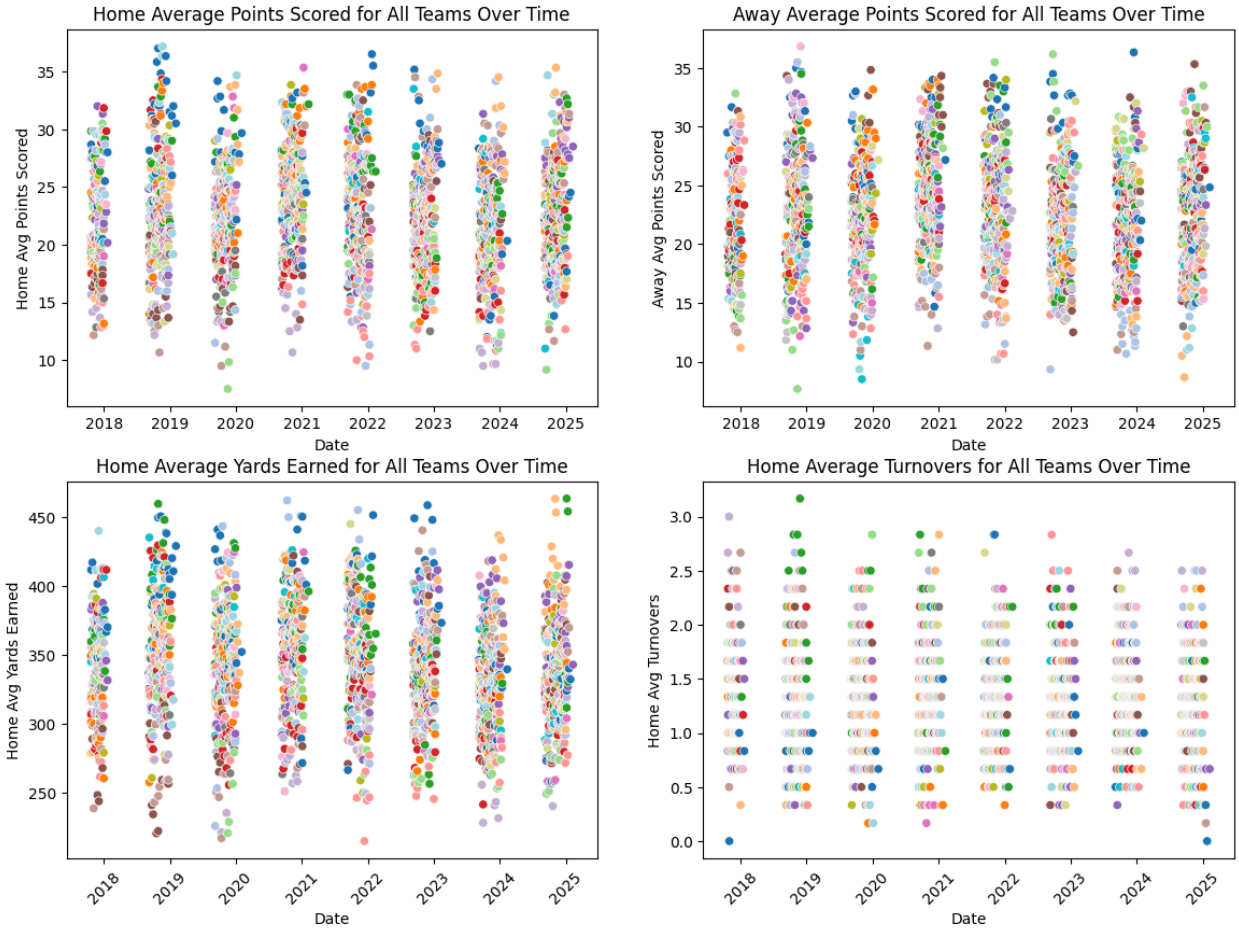

Figure 1. (a) Home Average Points Scored for All Teams over Time, (b) Away Average Points Scored for All Teams Over Time, © Home Average Yards Earned for All Teams Over Time, (d) Home Average Turnovers for All Teams Over Time

5. Model and Findings

Data was split into a 80%-20% train-test split using Scikit Learn. Features were scaled using StandardScaler to account for varying ranges in features.

Target distribution:

- Home Win (class 1): 55%
- Home Loss (class 0): 45%

Models tested (pre-tuning):

Logistic Regression

- Accuracy: 61.2%

Random Forest

- Accuracy 60.9%

Both models outperformed the 55% baseline. Logistic regression showed strong recall for wins but struggled with losses. Random forest produced more balanced results across both classes.

Hyperparameter Tuning:

Tuning performed using grid search with 5-fold cross-validation.

Logistic Regression:

- Best parameters:
  - C= 0.01
  - solver='liblinear'
- Best Cross-Validated Accuracy: 61.8%

Random Forest:

- Best parameters:
  - n_estimators = 100
  - max_depth = 5
  - max_features = 'sqrt'
  - min_samples_leaf = 2
  - min_samples_split = 2
- Best Cross-Validated Accuracy: 62.3%

Feature Importance:

Top 4 features ~40% of decision weight

- Avg_Points_Diff
- Avg_Yards_Diff
- Away_Avg_YardsEarned
- WinRate_Diff

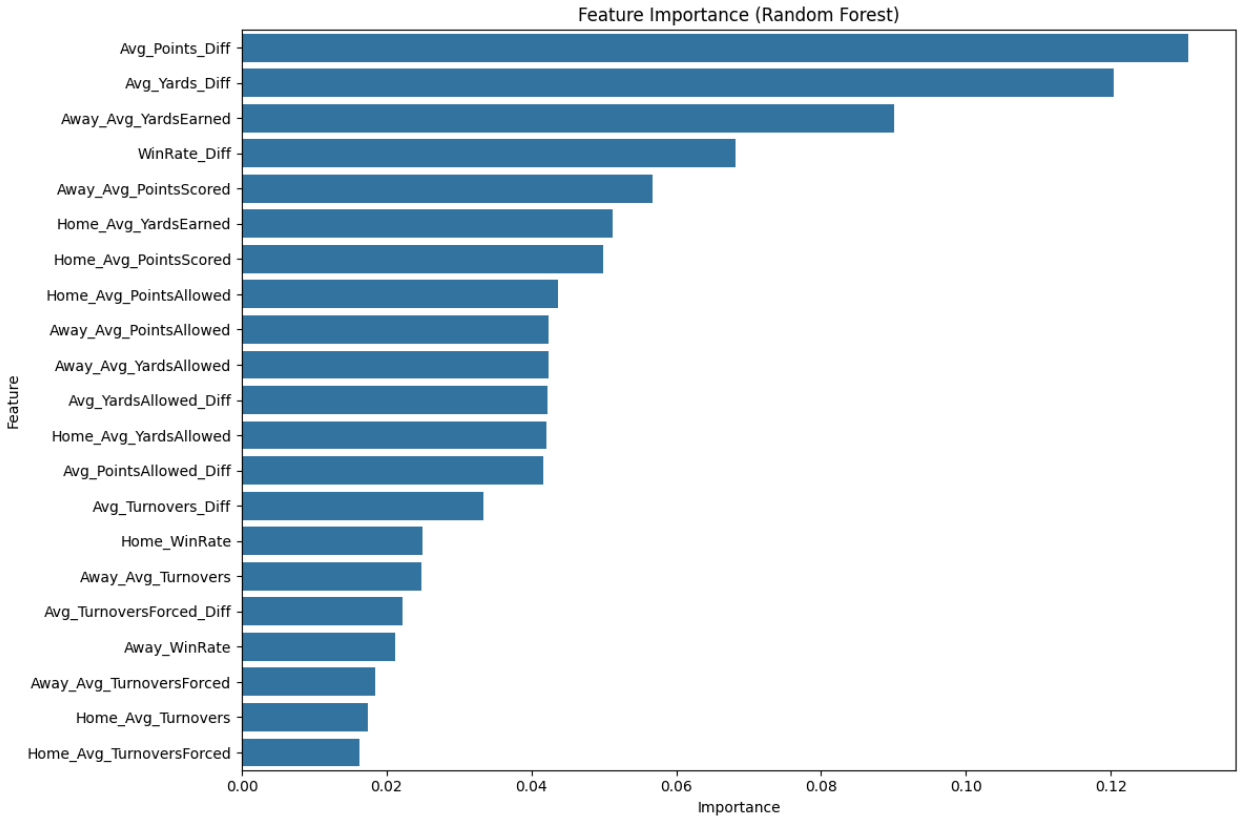

Fig 2: Feature Importance for Random Forest Classifier Model

Model Evaluation:

Given its slight improvement in accuracy and lower tendency to overfit, I selected the Random Forest model to generate predictions on the unseen data subset. The model was retrained on this updated data and achieved a 67% accuracy in predicting home team victories—successfully capturing several notable outcomes, including major upsets such as the Detroit Lions (9-point favorites) losing to the Washington Commanders in the playoffs. While the model's AUC score of 0.63 may appear modest, it reflects the inherent complexity and unpredictability of NFL outcomes. Overall, the model demonstrated moderate predictive power,

outperforming both a baseline strategy of always picking the home team and random guessing.

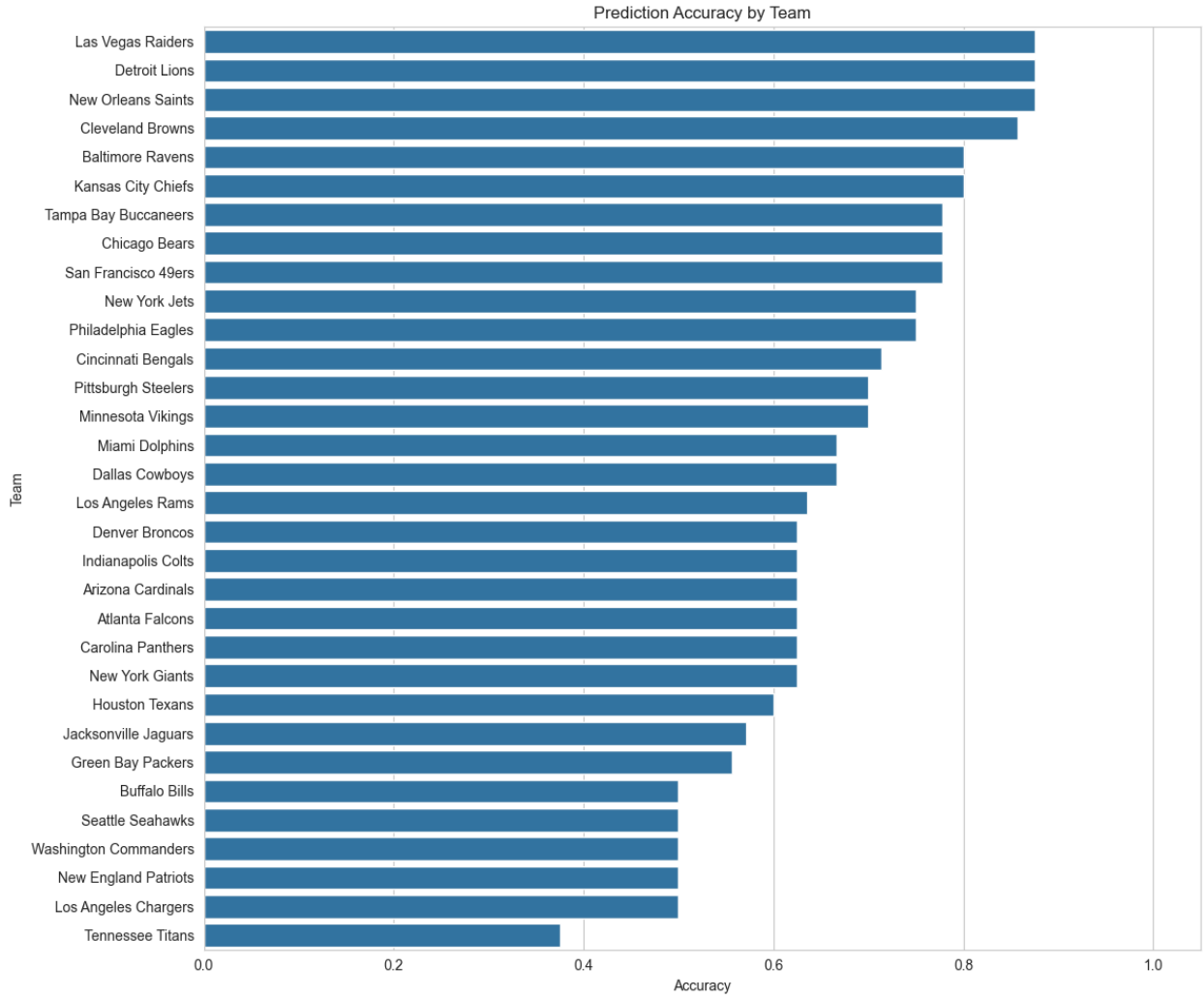

Figure 3: Prediction Accuracy by Team

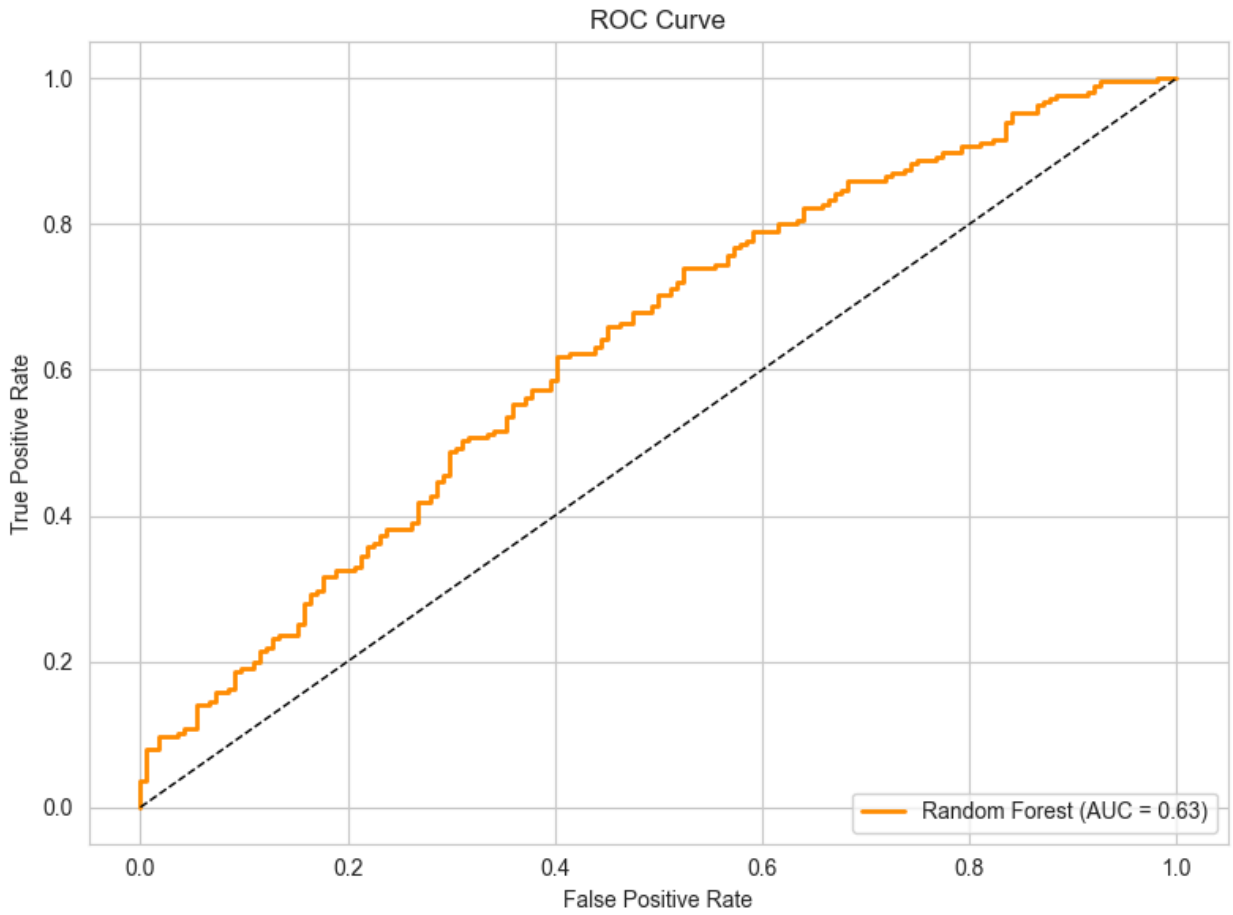

Figure 4: ROC curve

6. Next Steps

Given the models moderate predictive ability with the simple data pool used, I believe data driven matchup prediction should be pursued further. An ensemble of models may work better, along with more tuning. I think the first step should be to get more granular with the data. Metrics like player performance, injuries and suspensions, weather, and matchup history could all improve performance.