Travel Stats

Effects of Travel on Team Performance

Jacky Jiang

Problem Statement

The goal of this project is to determine if the travel schedule of MLB teams affects their performance. Key factors include travel distance, consecutive road games, and rest days, e.t.c.

Data Source

Data Set

The provided data source is a comprehensive dataset of Major League Baseball (MLB) games, comprising 56,775 entries.

Key Features

Game date
Teams involved
Game venue details
Game type
Scores
Performance statistics

Feature Engineering

Location Information

Add geographical coordinates for each MLB game venue to the dataset.

Query **Google Maps API** with the venue and city names, retrieving their geographical coordinates.

Travel Information

The travel distances and time since the last game for each MLB team, both home and away.

Uses the **Haversine** function to compute the travel distance between a team's last game venue and the current game venue.

For each game, record the travel distance and time since the last game for both the home and away teams.

Interactive Features

Create interaction terms and a new variable travel_dis_diff

Use the **PolynomialFeatures** object from scikit-learn, which generates interaction terms to capture potential combined effects of different travel distances.

Feature Engineering

Streaks Information

The current winning or losing streaks for both home and away teams.

A positive streak value indicates a winning streak, and a negative value indicates a losing streak.

Aways Information

The number of consecutive days an away team spends on the road.

For each game, the script calculates the number of consecutive days the away team has been on the road. This count is reset at the start of a new road trip or a new year.

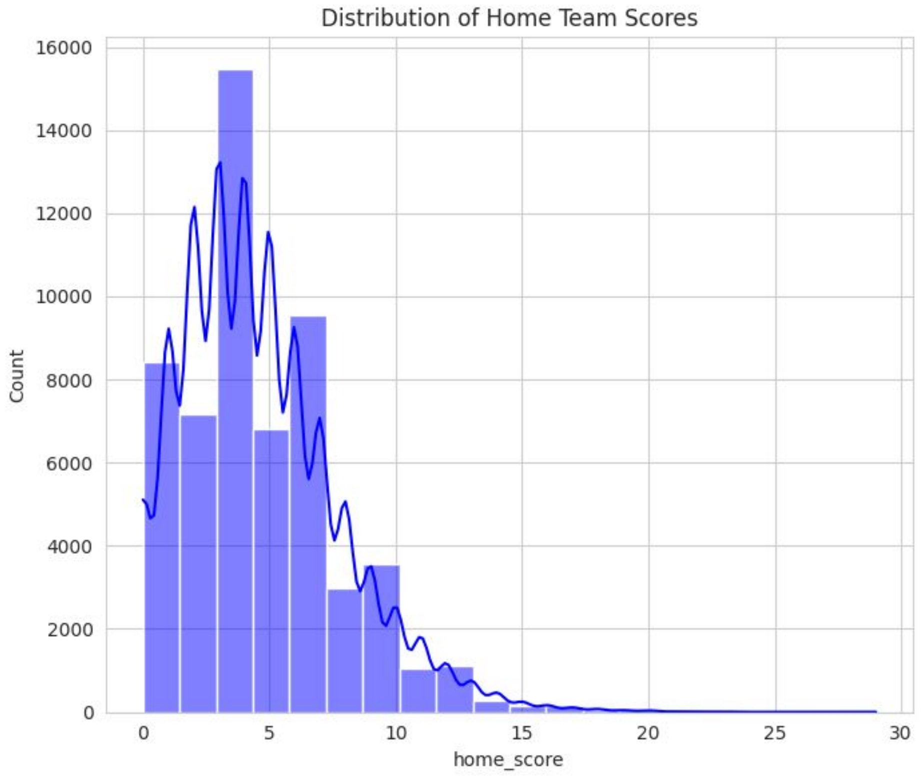
Data Cleaning and Wrangling

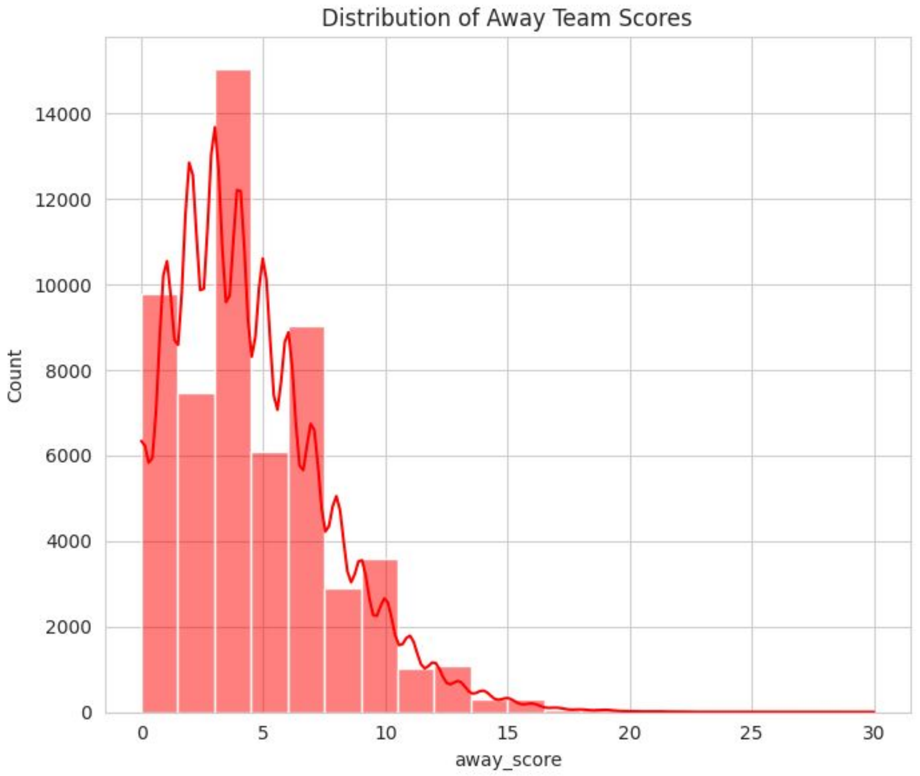
Applied **OneHotEncoder** to transform **categorical variables** like home_team and away_team are converted into a format suitable for modeling.

Used **StandardScaler** to ensure numerical features contribute equally.

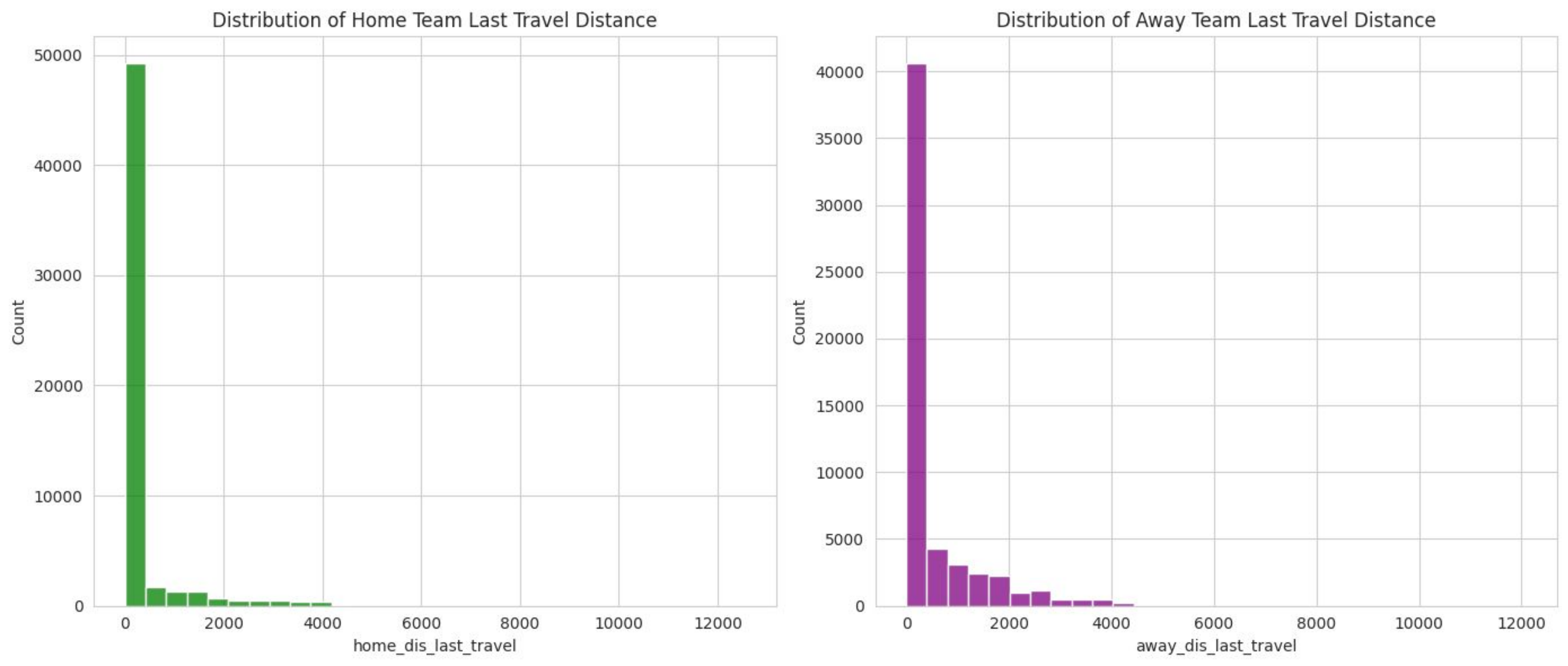
ColumnTransformer facilitated these steps in one go, enhancing our model's ability to learn from both categorical and numerical data effectively.

EDA of Score Distributions

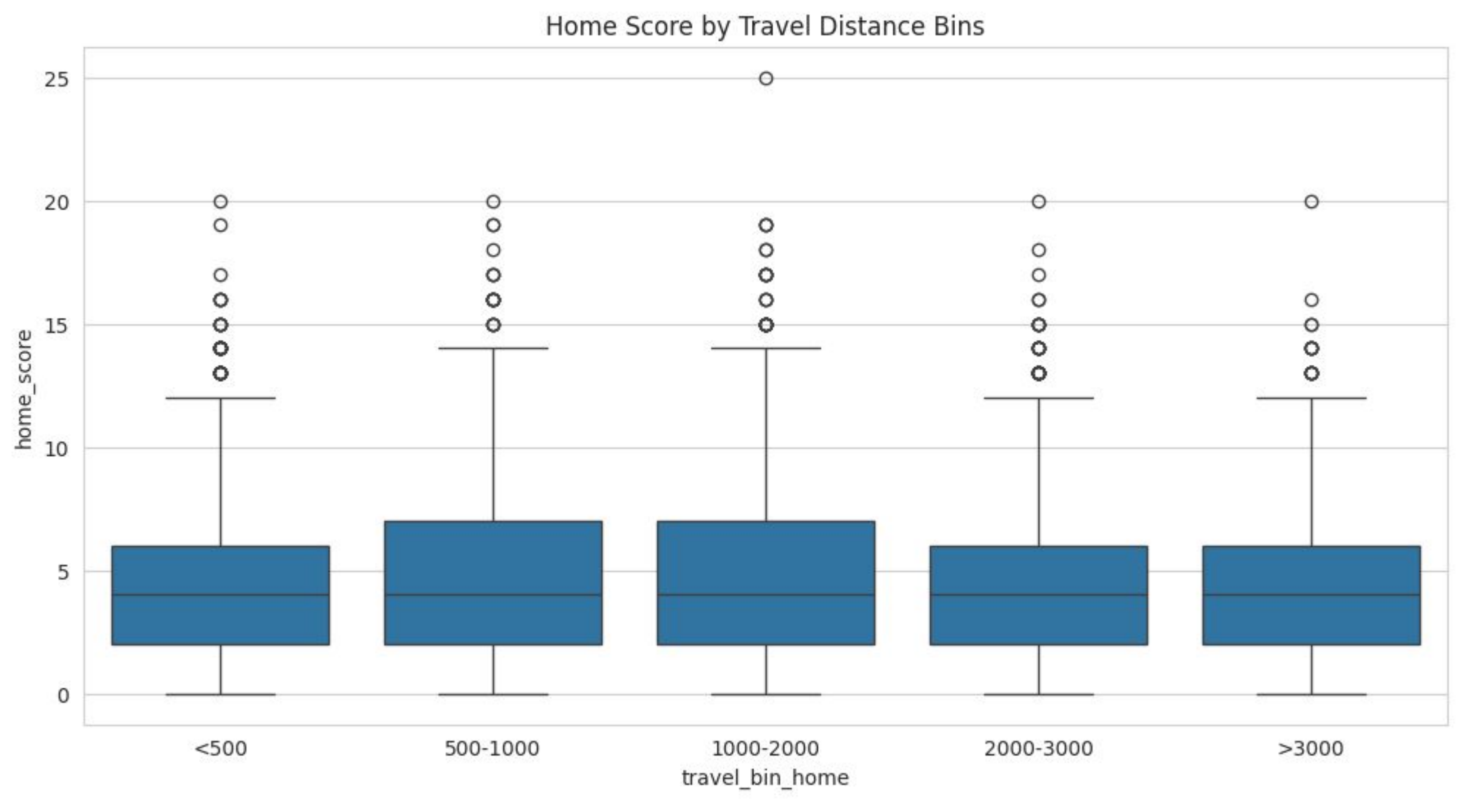




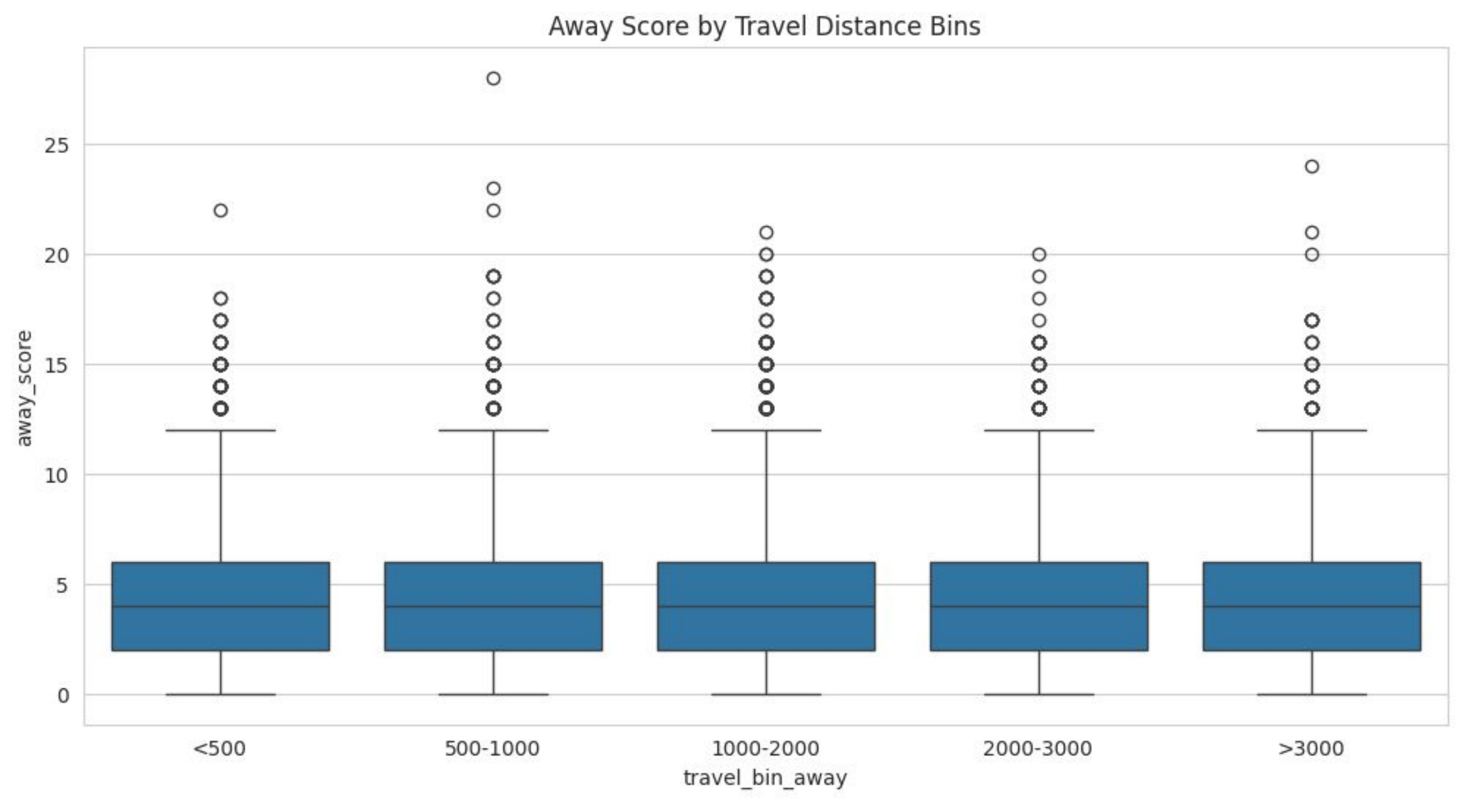
EDA of Home vs. Away Wins



EDA of Travel Distances



EDA of Travel Distances



Choosing the Right Model: XGBoost

We selected the eXtreme Gradient Boosting (XGBoost) model for its high performance in tabular data prediction tasks.

XGBoost is an ensemble learning method renowned for its speed and accuracy, making it ideal for our goal to assess the impact of travel schedules on MLB team performance.

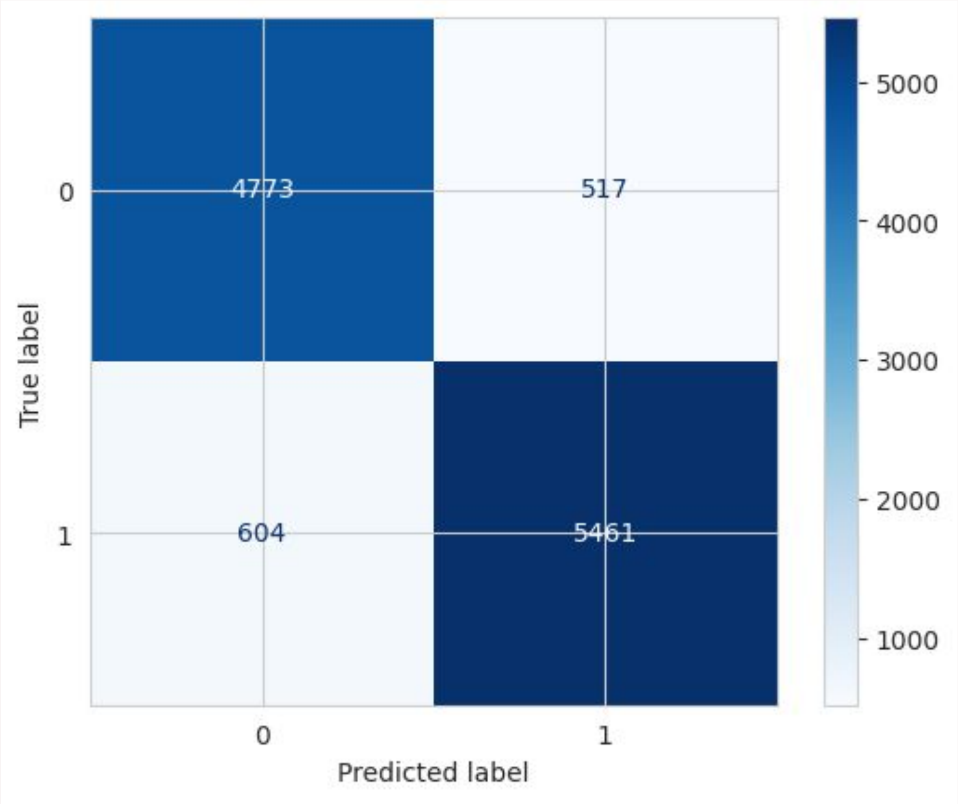
Hyperparameter Tuning with XGBoost

We fine-tuned our XGBoost model using a **RandomizedSearchCV** approach, optimizing over a wide range of hyperparameters to find the best combination.

The parameters included the **number of trees**, **learning rate**, **tree depth**, and **subsampling rates** for rows and columns, tailored to boost our model's predictive accuracy.

The best-performing model achieved a cross-validation accuracy of **89.9%**, with a standard deviation of **0.5%**, indicating robustness in its predictions.

Confusion Matrix



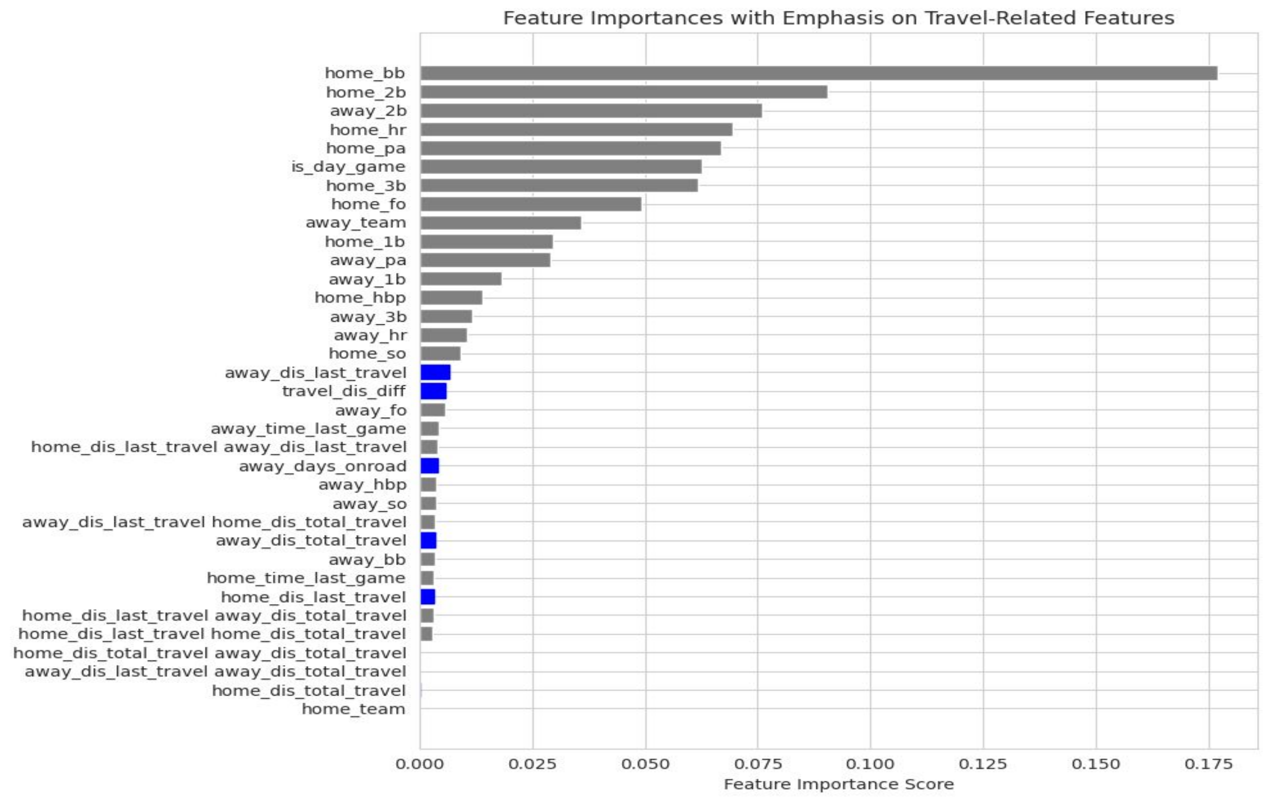
Feature Importance

We analyzed feature importances to understand what drives predictions in our model.

Notably, **in-game statistics** such as bases on balls (home_bb) and doubles (home_2b and away_2b) emerged as top predictors.

Travel-related features, while not at the very top, still showed a significant impact on the model's outcomes, supporting the hypothesis that travel schedules influence team performance.

Feature Importance



Comparing Travel vs. Non-Travel Models

We constructed two models: **one with travel-related features** and **one without**, to directly assess the travel impact.

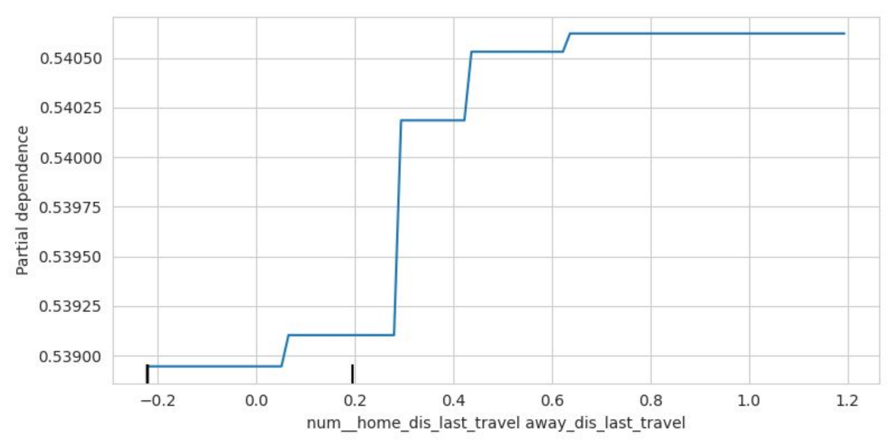
The comparison revealed only a **performance drop** of **15%** when **excluding** travel metrics, a paired t-test yielded a **p-value** of **0.006**.

This statistical test suggests that while travel features are **not the most dominant**, their influence is **significant enough** to affect model performance and, by extension, team performance predictions.

Visualize the Effect of Travel

PDP analysis indicates that certain travel distances correlate with changes in win probability.

Sharp changes in plots like pdp_num__home_dis_last_travel away_dis_last_travel.png demonstrate that specific travel scenarios can have a notable impact on win probability.



The Verdict on Travel Schedules

Our analysis indicates that travel schedules are a **relevant** but **not overriding** factor of MLB team performance.

The statistically significant result from the paired t-test reinforces that **travel has a measurable albeit modest impact**.

Thanks!