



USING MACHINE LEARNING TO PREDICT NBA PLAYER IMPROVEMENT

Sau

FROM SPRINT 2



Improving model performance and complexity

More feature engineerings

Data scrap

Reduce Multicollinearity

Better handling with N/A values

FEATURE ENGINEERING


Target feature	PER Improvement
List of features I created to increase model complexity	Previous PER Average, Previous PER Improvement, Previous PER Improvement Average, Previous OBPM Average, Previous OBPM Improvement Average, Previous TS% Average, Previous PTS Average, Previous USG% Average, Previous USG% Improvement, Previous 3PAr Average

❖ 10 out of 51 features for my models are engineered



DATA WEB SCRAPING

[Sports Reference](#) | [Baseball](#) | [Football \(college\)](#) | [Basketball \(college\)](#) | [Hockey](#) | [Football](#) | [blog](#) | [Stathead](#) | [Questions or Comments?](#)

SPORTS
REFERENCE

Rate Limited Request (429 error)

We apologize, but you have triggered rate limiting by our cloud service provider.

This could be for one of many reasons.

- You accessed more than thirty pages in less than a minute.
- You have written a bot that accessed too many files too quickly.
- Multiple people are accessing the site at the same time via your IP address.
- You are an employee of ESPN.

Generally, we block traffic that we think is from a bot for an hour and traffic that we think is a real user for five minutes.

[See our Bot Traffic page](#) or our [SR and Data Use Page](#).

Report an issue with our site or perhaps [our twitter account](#).

Bot Policy

Sports Reference is primarily dependent on ad revenue, so we must ensure that actual people using web browsers have the best possible experience when using our site. Unfortunately, non-human traffic, ie bots, crawlers, scrapers, can overwhelm our servers with the number of requests they send us in a short amount of time. Therefore we are implementing rate limiting on the site. We will attempt to keep this page up to date with our current settings.

Currently we will block users sending requests to:

- our sites more often than twenty requests in a minute.
- This is regardless of bot type and construction and pages accessed.
- If you violate this rule your session will be in jail for an hour.

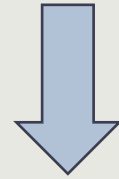
Why Not Provide an API?

- Original Kaggle dataset from 1995 to 2023
- Scraped 13 years more of NBA data from 1982 to 1994
- Increased from 7881 rows in sprint 2 to 11375 rows for final modelling

MODEL IMPROVEMENT

Sprint 2

Model	Train Score	Test R ²	RMSE	MAE
Linear Regression	0.25	0.24	2.48	1.93
Random Forest Regression	0.23	0.22	2.52	1.94



Now

Model	Train Score	Test R ²	RMSE	MAE
Linear Regression	0.327	0.364	2.41	1.82
Random Forest Regression	0.338	0.377	2.39	1.76

New models:

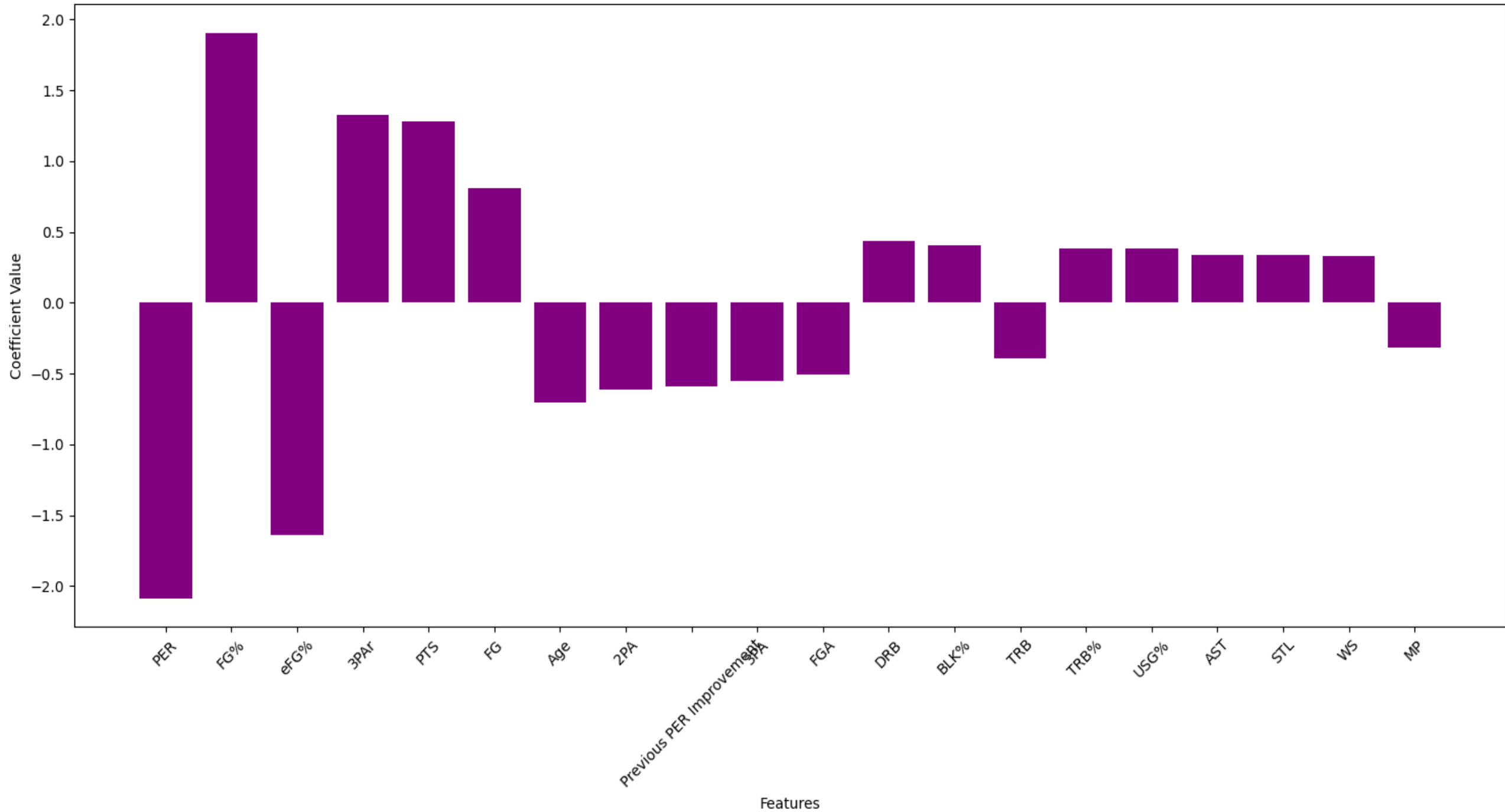
- ❖ XGBoost Regression
- ❖ Support Vector Regression

MODEL
COMPARISONS

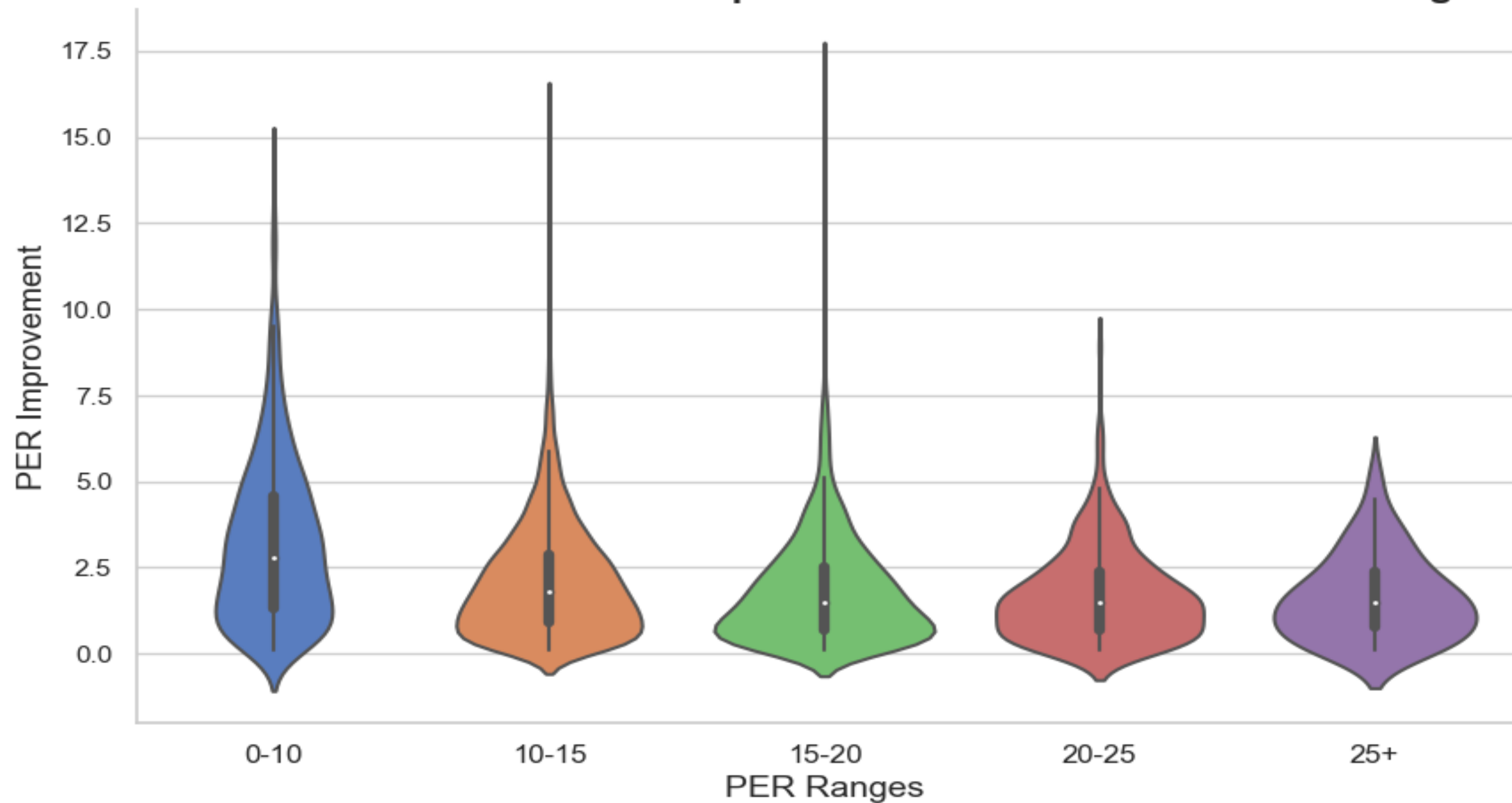


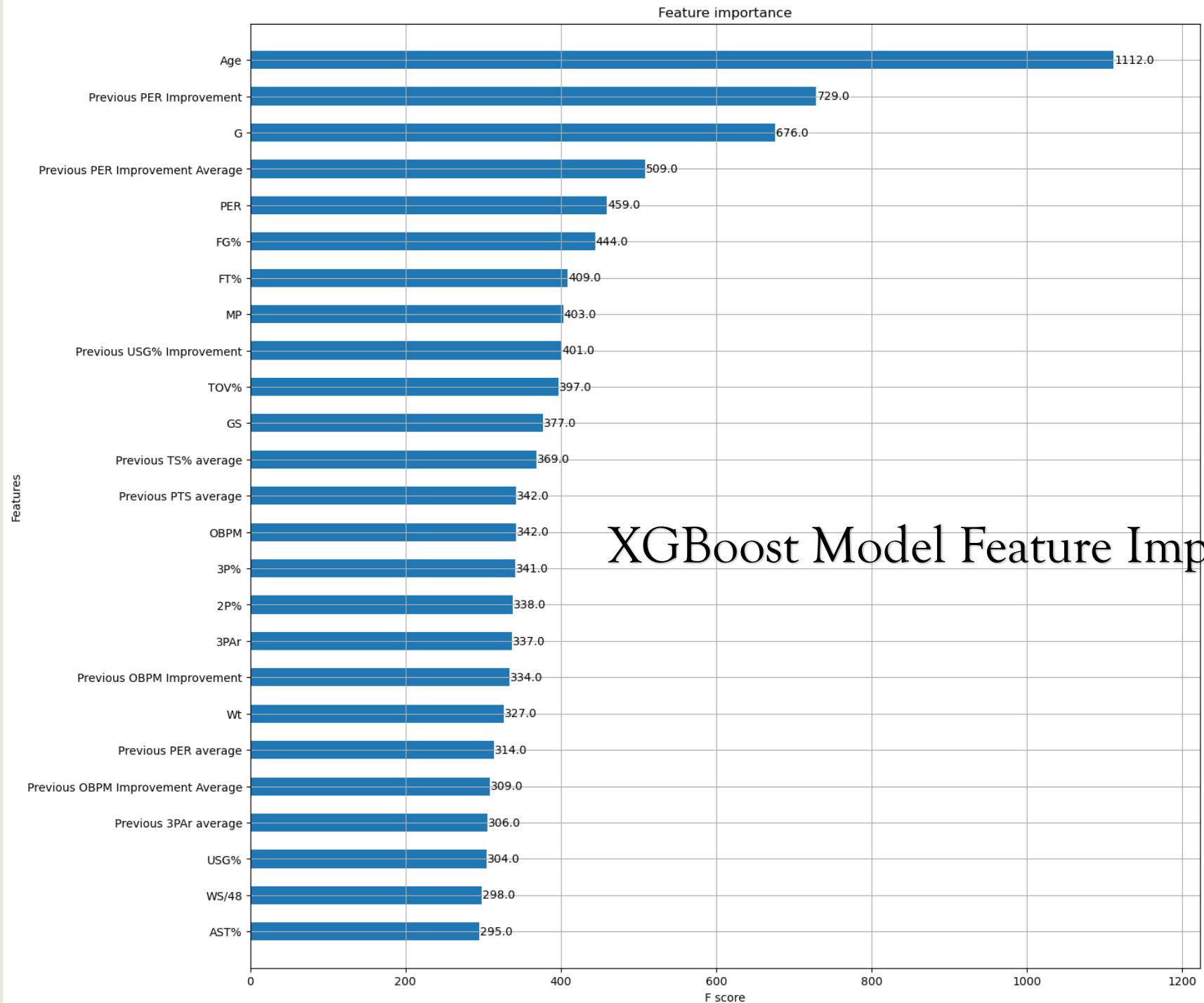
Model	Train Score	Test R ²	RMSE	MAE
XGBoost Regression	0.343	0.389	2.37	1.76
Random Forest Regression	0.338	0.377	2.39	1.76
SVM Regression	0.326	0.360	2.42	1.75
Linear Regression	0.327	0.364	2.41	1.82

Top 20 Features by coef from Linear Regression



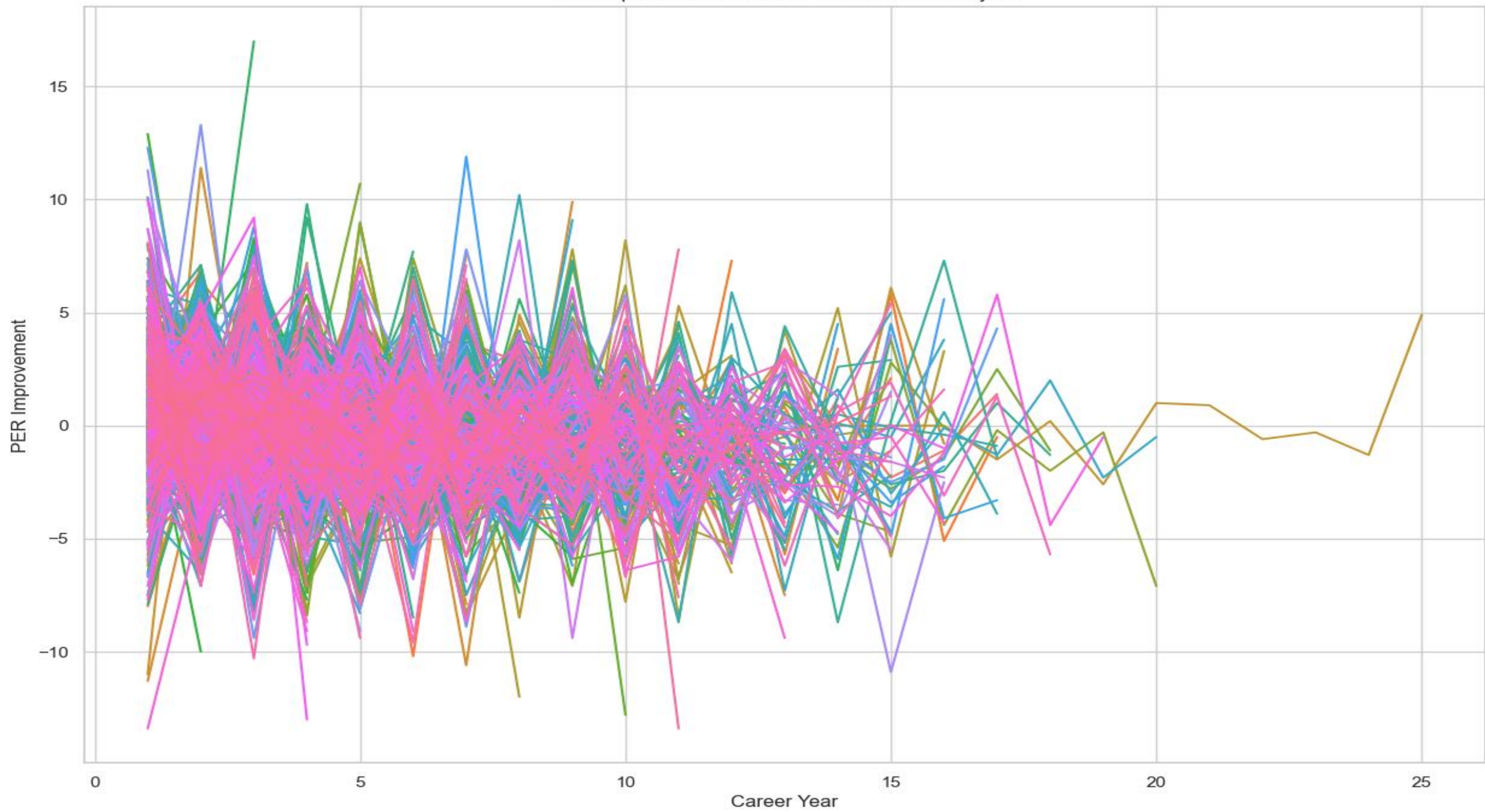
Distribution of Positive PER Improvement Across Different PER Ranges



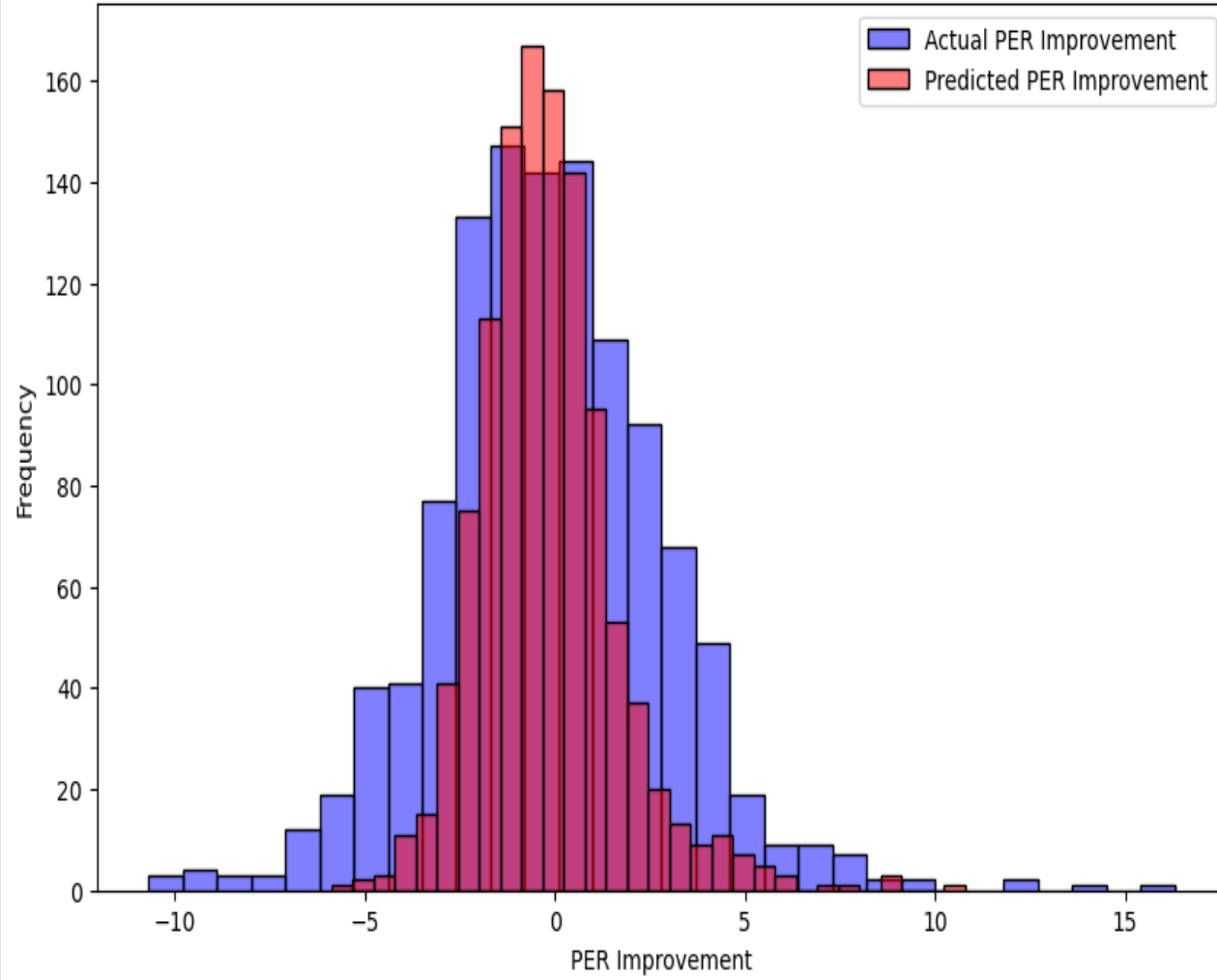


XGBoost Model Feature Importance

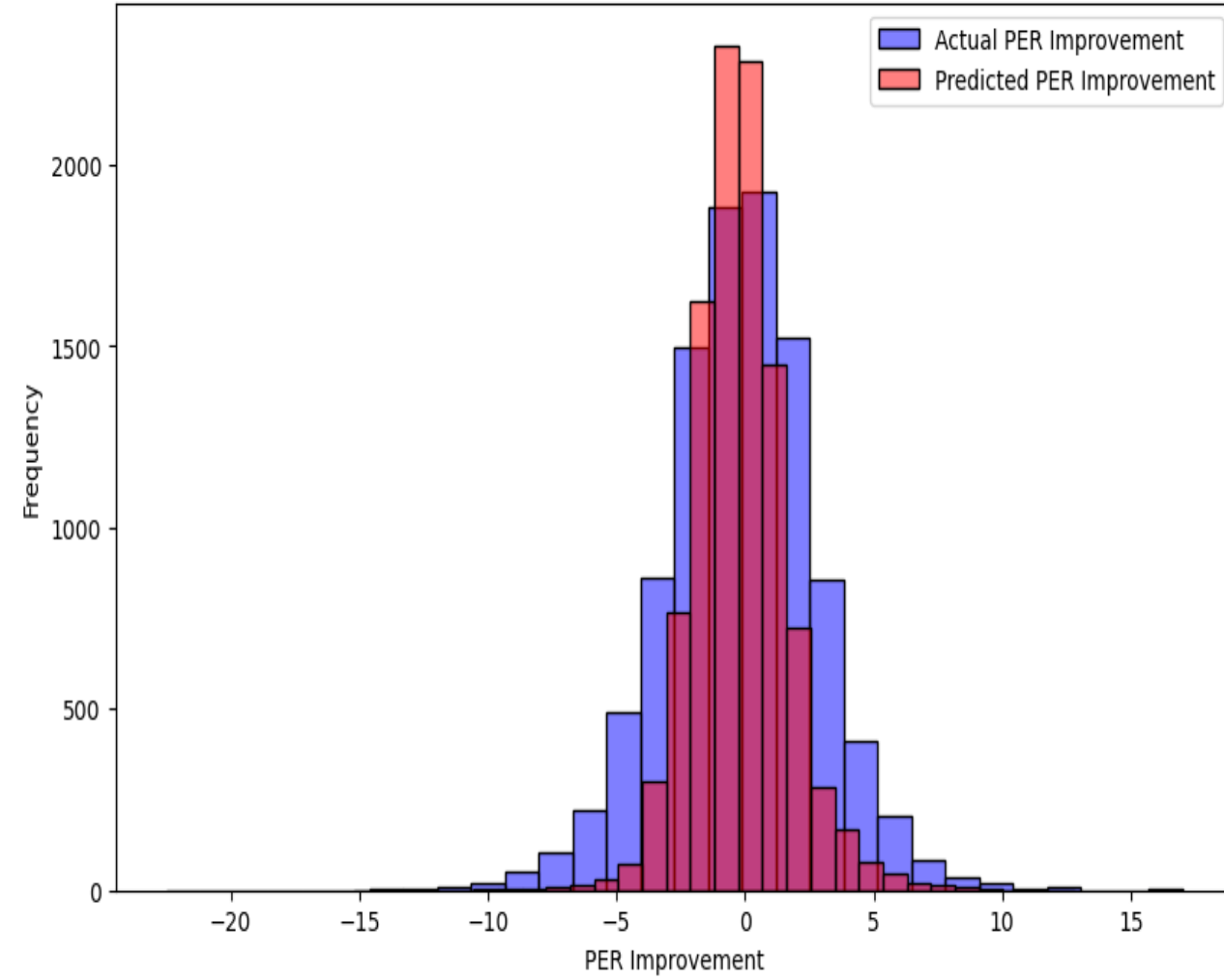
PER Improvement Trends for 1000 Random Players



Histogram of Actual vs. Predicted PER Improvement (XGBoost Test)



Histogram of Actual vs. Predicted PER Improvement (XGBoost Train)



FINAL MODEL EVALUATION

Model	Train Score	Test R ²	RMSE	MAE	Over/Under Estimation	Accuracy of predicting positive PER Improvement
XGBoost Regression	34.3%	38.9%	2.37	1.76	47.80% 52.20%	73.27%

Hyperparameters
from GridSearch:

Learning rate = 0.02
Max_depth = 6
Min_Child_weight = 0.001
N_estimators = 290
Subsample = 0.7



MY LEARNINGS AND NEXT STEPS