Predicting MLB Player Performance with TabPFN and Traditional Models

Chinghao Chang (Ryan)
University of Michigan
Ann Arbor, MI, USA
ryanchc@umich.edu

***Abstract*—This paper investigates the effectiveness of the TabPFN model, a transformer-based predictor for tabular data, in forecasting Major League Baseball (MLB) players' next-season batting average (BA) and home run count (HR). We compare TabPFN with traditional models such as Ridge Regression, Random Forest, and XGBoost using historical data from the Lahman database. Results show that TabPFN significantly improves prediction performance on BA but performs comparably on HR prediction.**

I. Introduction

Baseball statistics have long been a fertile ground for predictive modeling. With the rising popularity of sabermetrics and player performance analysis, accurately forecasting a batter's future performance is of both analytical and commercial value. Motivated by this, our project aims to build predictive models for two key performance metrics: next-season batting average (BA_next) and home run count (HR_next).

Recent advances in transformer-based architectures have shown promise in tabular data tasks. Among these, TabPFN (Tabular Prior-Data Fitted Network) offers a lightweight, pretrained model specifically designed for small tabular datasets. This project evaluates its performance against widely used methods like Ridge Regression, Random Forest, and XGBoost.

II. Method

A. Data and Feature Engineering

We use player-season statistics from the Lahman Baseball Database, combining `Batting.csv` and `People.csv` via `playerID`. Key features include hits, home runs, walks, and age (calculated as `yearID - birthYear`). To avoid division-by-zero errors, we apply conditional logic when computing rate-based metrics.

Engineered features include:

- **BB/SO**: plate discipline ratio.
- **HR rate**: home runs per at-bat.

We randomly sampled 1,000 records to reduce computation time. Categorical variables like `teamID` were one-hot encoded, and all numeric features were standardized using `StandardScaler`.

B. Environment Setup

All experiments were performed in Python 3.10 using a Conda environment with PyTorch 2.0.1 (CPU). The TabPFN model and necessary packages were installed as follows:

```
conda create -n tabpfn-env python=3.10 -y
conda activate tabpfn-env
pip install torch==2.0.1 torchvision==0.15.2 torcha
pip install tabpfn pandas matplotlib
```

Experiments were run locally in JupyterLab on an Intel CPU machine.

C. Modeling Approach

We trained four models:

- **Ridge Regression**: Linear model with L2 regularization.
- **Random Forest**: Ensemble of decision trees.
- **XGBoost**: Boosted gradient trees with regularization.
- **TabPFN**: Pretrained transformer for small tabular data.

All models were evaluated using RMSE and R^2 on a holdout test set. TabPFN was run on CPU with performance constraints bypassed using environment variable overrides.

III. Results

A. Quantitative Metrics and Graphs

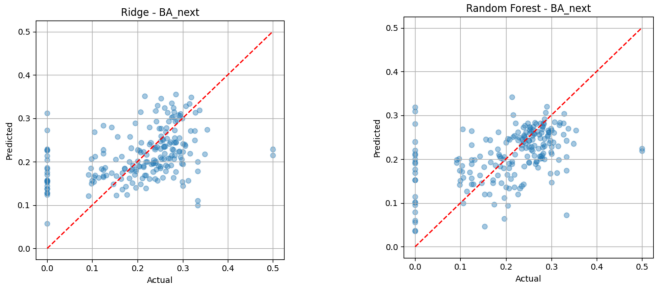

Fig. 1. Ridge Regression performance on BA_next (left) and HR_next (right)

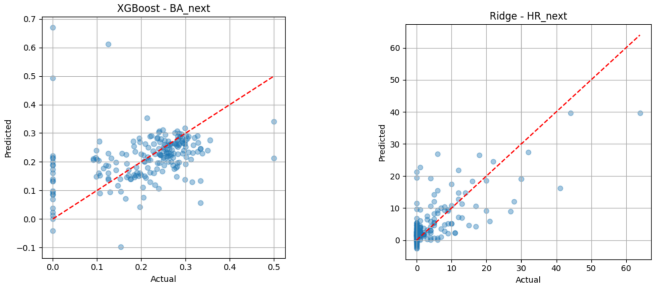

Fig. 2. Random Forest performance on BA_next (left) and HR_next (right)

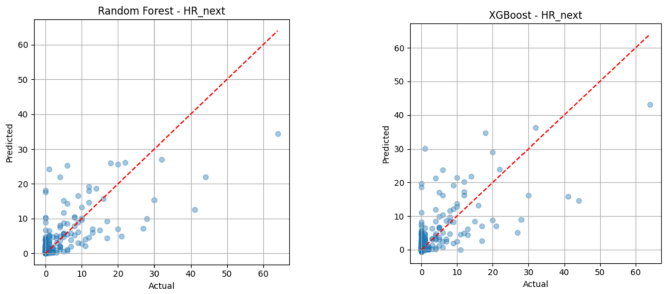

Fig. 3. XGBoost performance on BA_next (left) and HR_next (right)

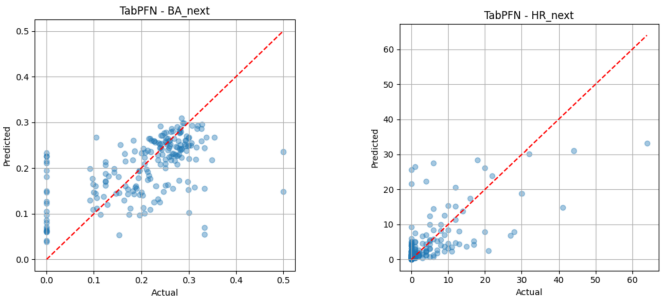

Fig. 4. TabPFN performance comparison on BA_next (left) and HR_next (right)

TABLE I
MODEL PERFORMANCE ON TEST SET

Model	**BA_next**		**HR_next**	
	RMSE	R^2	RMSE	R^2
Ridge	0.0925	0.1373	5.7861	0.5193
Random Forest	0.0897	0.1887	6.1472	0.4574
XGBoost	0.1072	0.1581	6.5264	0.3884
TabPFN	**0.0828**	**0.3084**	6.2511	0.4389

B. Interpretation

TabPFN outperformed all traditional models in predicting BA_next, achieving both the lowest RMSE and the highest R^2. This suggests that transformer-based representations capture subtle statistical patterns in BA that traditional models might overlook.

However, on HR_next prediction, traditional models like Ridge and Random Forest remain competitive. The relatively weaker TabPFN performance here could be attributed to the skewed distribution and larger variance in HR data.

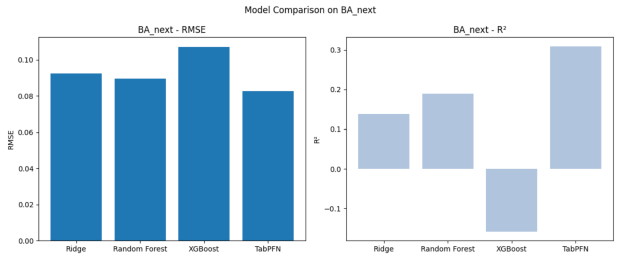

Fig. 5. Model comparison on BA_next: RMSE and R^2

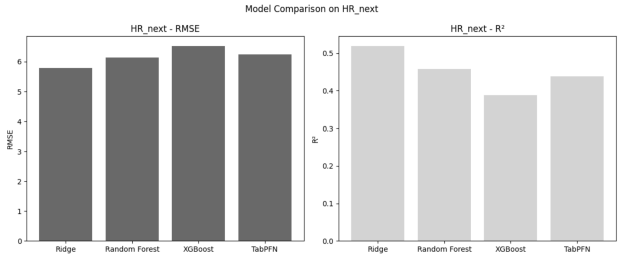

Fig. 6. Model comparison on HR_next: RMSE and R^2

IV. CONCLUSION

This study shows that TabPFN is a promising model for predicting baseball statistics, particularly batting average. While traditional models still perform well for high-variance targets like HR count, TabPFN's performance on BA suggests it can be an effective tool for small tabular datasets in sports analytics.

Future work can explore ensemble combinations of TabPFN and tree-based models or include additional features such as physical metrics, injury history, or team dynamics to improve prediction robustness.

As a long-time baseball fan and former varsity player, I recognize that statistics alone may not capture the full context of a player's performance. For example, factors like player morale, injury history, team dynamics, ballpark dimensions, or even in-game roles (e.g., cleanup hitter vs. leadoff) can significantly affect future outcomes but are not reflected in historical stats.

Moreover, from a fan's perspective, a player's perceived "clutch" ability or resilience under pressure is rarely quantifiable but can influence performance unpredictably. While TabPFN and other models offer strong statistical baselines, integrating domain insights from baseball could enhance interpretability and robustness.

REFERENCES

[1] N. Hollmann, S. Müller, K. Eggensperger, and F. Hutter, "TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second," *arXiv preprint arXiv:2207.01848*, 2022. [Online]. Available: https://arxiv.org/abs/2207.01848

[2] C. Dalzell *et al.*, "Lahman Baseball Database," R package version 12.0.0, Jan. 2025. [Online]. Available: https://cran.r-project.org/web/packages/Lahman/Lahman.pdf

[3] S. Wang, "Learning Contextual Event Embeddings to Predict Player Performance in the MLB," in *Proc. MIT Sloan Sports Analytics Conf.*, 2023. [Online]. Available: https://www.sloansportsconference.com/research/learning-contextual-event-embeddings-to-predict-player-performance

[4] H.-C. Sun, T.-Y. Lin, and Y.-L. Tsai, "Performance Prediction in Major League Baseball by Long Short-Term Memory Networks," *arXiv preprint arXiv:2206.09654*, 2022. [Online]. Available: https://arxiv.org/abs/2206.09654

[5] D. Bailey, "A Regression-Based Approach for Prediction of Major League Baseball Player Performance," Dublin Business School, 2020. [Online]. Available: https://esource.dbs.ie/bitstreams/bcd0ae21-cb4d-4cb1-87b6-061de6792f5d/download

CODE REPOSITORY

The complete code is available at: Final Project- Code.txt