STAT 507 Project Proposal

Project Title: Predicting MLB Player Performance Using TabPFN Transformers

---

- **Overview**

Background & Motivation: MLB is a data-rich sport. With decades of structured statistics, we can build predictive models to estimate how a player might perform next season. This project explores how modern small-data transformers, especially the TabPFN-v2-reg model from Hugging Face, can improve player performance prediction.

Why This Project: As someone passionate about baseball and data analytics, I want to explore how modern machine learning models can improve traditional performance prediction. Explores how few-shot tabular transformers perform compared to traditional models

- **Data & Model**

Dataset: Lahman Baseball Database (processed subset of ~1000–3000 players)

Model: Prior-Labs/TabPFN-v2-reg from Hugging Face

Baseline: Ridge Regression, Random Forest, XGBoost

Expected Insights:

1. Which features (e.g. age, prior season stats) best predict future outcomes
2. Whether TabPFN outperforms simple ML models under limited data

- **Prior Work**

Literature Review:

1. Traditional baseball forecasting uses linear regression, decision trees, and ensemble models.
2. Few studies apply tabular transformers in sports prediction

- **Preliminary Results**

Data Understanding:

1. **Source**: Lahman Database via Kaggle
2. **Shape**: ~20,000 rows, 80+ features ——> ~1000-3000 rows for training
3. **Target**: Feature columns include: ages, batting stats, position, team changes

Tools from Class:

1. pandas for data loading and preprocessing
2. scikit-learn for implementing baseline models
3. matplotlib / seaborn for data visualization
4. Hugging Face Transformers (e.g., Prior-Labs/TabPFN-v2-reg) for advanced modeling

- **Project Deliverables**

1. A fine-tuned Hugging Face transformer model that predicts MLB player performance for the 2023 season
2. Comparative evaluation against classical models (e.g., Linear Regression, Ridge Regression, Random Forest)
3. Visualizations including feature importance, error metrics, and distribution plots
4. Final GitHub repository including Jupyter Notebooks and a project summary report

- **Subgoals:**

1. Clean and preprocess the Lahman Baseball dataset
2. Implement and evaluate classical baseline models
3. Fine-tune and test the Hugging Face tabular transformer model
4. Interpret the results and create visualizations
5. Summarize findings and prepare final deliverables

- **Timeline**

Week 1–2: Literature review, data loading, preprocessing, and exploratory analysis

Week 3: Train baseline models; fine-tune and evaluate the Hugging Face model， Visualize results, interpret models, finalize report and GitHub report

1. Project Workflow Diagram

Lahman Dataset

↓

Data Preprocessing

- Handle missing values
- Select relevant features
- Encode categorical variables

↓

Exploratory Data Analysis (EDA)

- Summary statistics
- Correlation analysis
- Visualization of key variables

↓

Model Training

├── Baseline Models

│ • Ridge Regression

│ • Random Forest

│ • Gradient Boosting

│ ↓

│ Performance Evaluation (RMSE, MAE, R^2)

└── Hugging Face Model

• Model: `Prior-Labs/TabPFN-v2-reg`

• Use subset of training data (≤10,000)

↓

Performance Evaluation (same metrics)

↓

Comparison & Interpretation

- Compare all models' performance
- Discuss pros/cons and insights

↓

Final Deliverables

- Summary report
- GitHub repository

2. Model Architecture (TabPFN-v2)