Nicholas Condos
November 12, 2025
DATA 6560 – Sports Analytics

Checkpoint 3: Data Quality & Source Validation

Project Title & Data Overview

Title: Modeling ATP Match Outcomes Using Serve, Return, and Performance Efficiency Metrics

This project uses ATP Tour match data from 2008 to 2024 to study how serve, return, and efficiency metrics predict match outcomes across hard, clay, and grass courts. Building on Checkpoints 1 and 2, the goal is to create logistic regression and random forest models that show which factors most affect a player's chances of winning.

The dataset—Jeff Sackmann's *tennis_atp* repository—is a trusted, publicly available source used in both academic and professional sports analytics. It provides enough detail and consistency for reliable data analysis and modeling.

Data Source Summary

Primary Source: Jeff Sackmann's *tennis_atp* GitHub Repository

- **Author:** Jeff Sackmann, a leading sports data researcher.
- **Access:** Public and free at github.com/JeffSackmann/tennis_atp.
- **Files Used:** atp_matches_YYYY.csv (2008–2024), plus atp_players.csv, atp_rankings.csv, and matches_data_dictionary.txt.
- **Contents:** Match-level data including player IDs, rankings, serve and return statistics, and results.
- **Source Type:** Official ATP match results compiled from tournament and score data.

This source is consistent, transparent, and commonly cited in sports analytics research, making it ideal for this project.

Data Structure & Content

Each annual file contains one row per match, including player information, rankings, and match statistics.

Key Variables:

- **Tournament Info:** ID, surface, round, date
- **Player Info:** winner_id, loser_id, rank, handedness, age, height
- **Serve Stats:** aces, double_faults, first_serve_%, service_points_won
- **Return Stats:** return_points_won, break_points_saved, break_points_converted
- **Efficiency Ratios:** winners/unforced errors, total_points_won

For 2008–2024, there are roughly 40,000 valid matches. Variables are clearly defined, and both categorical and numeric data are suitable for modeling.

Data Completeness & Consistency

The dataset is complete and standardized after 2008, with minimal missing values.

- **Missing Values:** Fewer than 5% of entries missing serve or return stats, mainly from 2008–2010.
- **Formatting:** Consistent column names, units, and data types across all years.
- **Duplicates:** Minor duplicates (e.g., walkovers) removed using match IDs.
- **Surface Distribution:** Hard ≈ 55%, Clay ≈ 30%, Grass ≈ 15%. Stratified sampling ensures balanced modeling.
- **Time Coverage:** Every season from 2008–2024 included with no gaps.

The dataset is stable and ready for exploratory analysis and model development.

Quality Issues & Potential Biases

1. **Older Records:** Some early matches lack advanced stats, but this affects less than 5% of the sample.
2. **Surface Bias:** Hard courts are most common; this will be adjusted using surface-based sampling.

3. **Player Names:** Spelling and accent inconsistencies standardized using UTF-8 normalization.

4. **Match Status:** Retirements (RET), Walkovers (WO), and Abandoned matches (ABD) excluded.

5. **Sample Bias:** Dataset focuses on main-draw ATP events, not Challenger or Qualifier levels.

These issues are documented and managed to maintain reliability and transparency.

Initial Cleaning & Preparation Plan

Data preparation will be done in **Python (Pandas, NumPy)** following a structured workflow:

1. Import and combine all annual CSVs (2008–2024).

2. Filter out RET, WO, and ABD matches.

3. Remove duplicate records using match IDs.

4. Impute missing numeric values (mean or median).

5. Drop records with more than 20% missing data.

6. Create new features like Serve-Return Balance and Winners/Error Ratio.

7. Encode categorical variables (surface, handedness) and apply z-score scaling to numeric fields.

8. Verify summary statistics and surface balance post-cleaning.

This plan ensures a clean and well-structured dataset for modeling.

Next Steps

- Finalize and export atp_matches_2008_2024_clean.csv.

- Perform EDA and visualizations by surface type.

- Explore correlations between serve, return, and efficiency metrics.

- Prepare modeling dataset and begin logistic regression and random forest analysis for Checkpoint 4.

Summary Statement

The ATP dataset is reliable, detailed, and suitable for predicting match outcomes across surfaces. Minor issues, such as older missing stats and surface imbalance, have been identified and addressed. With consistent formatting and a clear cleaning plan, the dataset is ready for exploratory and predictive analysis.