

## CHECKPOINT 6 — Analysis Memo #1

Nicholas Condos

Dec 5, 2025

[https://github.com/nicholascondos/Tennis\\_Analytics](https://github.com/nicholascondos/Tennis_Analytics)

### 1. Question and Outcome

This project examines:

**Which performance and efficiency factors most strongly differentiate ATP match winners from losers between 2008–2024?**

Building on Checkpoints 1–5, this checkpoint reframes the research question into a predictive modeling task: given a player's serve, return, and efficiency metrics, can we predict whether that player **wins or loses** the match?

The binary outcome variable is `outcome = 1` for winners and `0` for losers, evaluated at the **player-match** level.

### 2. Data Used

The cleaned dataset produced in Checkpoint 4 (`atp_matches_2008_2024_clean.csv`) is used.

It includes one row per completed ATP singles match (2008–2024), with doubles, walkovers, qualifying matches, and round-robin events removed.

For modeling, I converted the match-level dataset into a **player-level dataset** with two rows per match:

- One row for the **winner** (`outcome = 1`)
- One row for the **loser** (`outcome = 0`)

Each row contains:

- `player_srv_ret_balance` – combined serve + return efficiency (CP4)
- `player_first_serve_pct` – first-serve percentage

## CHECKPOINT 6 — Analysis Memo #1

- `player_ret_1st_pct` – first-serve return percentage (CP5 Fig. C)
- `player_1st_points_won` – first-serve points won (CP5 Fig. D)
- `player_rank`, plus surface and match year

After removing rows with missing values, the modeling dataset has roughly **86,612 rows**, balanced evenly between winners and losers.

### 3. Method Choice

The outcome is binary, and the research question seeks interpretable relationships between performance metrics and probabilities of winning.

Therefore, **logistic regression** is the most appropriate modeling method. It provides interpretable coefficients and aligns directly with the KPIs and relationships explored in Checkpoints 1–5.

#### Baseline Model (Model 1)

A logistic regression using one predictor:

- `player_srv_ret_balance`

This continues the CP4/CP5 emphasis on `efficiency_diff` as the strongest separator of winners and losers.

#### Improved Model (Model 2)

Adds two CP5-inspired return and serve indicators:

- `player_srv_ret_balance`
- `player_ret_1st_pct`
- `player_1st_points_won`

This tests whether key return skills and first-serve dominance improve prediction beyond efficiency alone.

## CHECKPOINT 6 — Analysis Memo #1

### 4. Analysis Specification

**Row definition:** one player per row per match

**Outcome:** `outcome`  $\in \{0, 1\}$  (loser vs winner)

#### Predictors

- **Model 1:**
  - `player_srv_ret_balance`
- **Model 2:**
  - `player_srv_ret_balance`
  - `player_ret_1st_pct`
  - `player_1st_points_won`

**Time Frame:** 2008–2024 ATP matches

**Train/Test Split:** 70% / 30%, stratified

**Baseline Accuracy:** With balanced rows, baseline = **50%**

### 5. Results

#### Model 1 — Efficiency Only

- **Accuracy:** 0.787
- **Baseline:** 0.50
- **Coefficient:** positive, large
- **Interpretation:** Combined serve-return efficiency alone yields a strong predictive lift, confirming CP5's finding that `efficiency_diff` is the single strongest descriptive separator between winners and losers.

#### Model 2 — Efficiency + First-Serve Return + First-Serve Points Won

## CHECKPOINT 6 — Analysis Memo #1

- **Accuracy: 0.878**
- **Precision/Recall/F1: all  $\approx 0.88$**
- **Coefficients:**
  - `player_srv_ret_balance`: **29.24**
  - `player_ret_1st_pct`: **10.66**
  - `player_1st_points_won`: **0.0047**

### Interpretation:

Model 2 substantially improves upon Model 1, increasing test accuracy from **78.7% to 87.8%**. Both added predictors are statistically meaningful:

- Higher **first-serve return percentage** greatly increases win probability
- More **first-serve points won** also contributes positively

Together, these metrics capture both defensive (returning first serves) and offensive (winning first-serve points) aspects of match performance - specifically aligning with the patterns discovered in CP5's visual analysis.

## 6. Checks and Reasonableness

### Scaling:

All predictors lie on well-behaved scales (between 0–1 or small integers), so logistic regression behaves accordingly.

### Leakage:

No variable directly encodes match outcome. Predictors are expressed generically ("player\_"), not as winner/loser labels.

### Train/Test Validity:

Stratified splitting ensures a balanced/fair evaluation.

### Coefficient Interpretability:

All coefficient signs match tennis intuition:

## CHECKPOINT 6 — Analysis Memo #1

- Efficiency  $\uparrow \rightarrow$  win probability  $\uparrow$
- First-serve return skill  $\uparrow \rightarrow$  win probability  $\uparrow$
- First-serve dominance  $\uparrow \rightarrow$  win probability  $\uparrow$

## 7. Interpretation

Model 1 and Model 2 together show that ATP match outcomes can be predicted extremely well using performance-based metrics. Efficiency remains the strongest single predictor of winning probability, capturing holistic match control.

However, the improved model shows that **first-serve interactions** - both returning and winning points behind one's own first serve - add substantial explanatory power.

This directly supports conclusions from CP5: modern ATP matches are often decided by who controls the **first-serve exchange** on both sides of the ball.

## 8. Limitations

- Logistic regression may miss nonlinear patterns or interactions, such as efficiency  $\times$  surface or rank  $\times$  serve dynamics.
- Break-point statistics and second-serve performance metrics are not yet included.
- Match-level contextual factors (e.g., opponent strength, fatigue, rally length) are not modeled.

## 9. Next Steps (Checkpoint 7)

- Add **break-point conversion and save rates** into the model to test high-leverage point performance.
- Explore **surface-based interactions**, since CP5 revealed large surface-driven differences in efficiency.
- Compare logistic regression with a **tree-based model** (e.g., random forest) to see whether nonlinear relationships offer additional predictive gain while remaining interpretable.

## CHECKPOINT 6 — Analysis Memo #1