CHECKPOINT 7 — Analysis Memo #2

Nicholas Condos

Dec 8, 2025

https://github.com/nicholascondos/Tennis_Analytics

1. Question and Outcome

This project continues to examine which performance and efficiency metrics best differentiate ATP match winners from losers between 2008 and 2024. In Checkpoint 7, the goal is to build upon the CP6 predictive models by introducing one or two theory-driven improvements to the logistic regression framework. The outcome variable remains a binary indicator of match result at the player level, where 1 represents the winner and 0 represents the loser.

2. Data Used

The analysis uses the CP4 cleaned dataset, atp_matches_2008_2024_clean.csv, which contains all completed ATP singles matches from 2008 to 2024. All preprocessing steps from earlier checkpoints remain unchanged, including removal of doubles, qualifying rounds, walkovers, and incomplete matches.

Following the CP6 structure, the dataset is expanded into a player-level format with two rows per match: one for the winner and one for the loser. Each row includes serve statistics, return statistics, efficiency metrics, break-point data, ranking, surface, and year. No new filters or exclusion rules were applied in CP7 to maintain continuity with prior checkpoints.

3. Summary of CP6 (Baseline Models)

Checkpoint 6 produced two logistic regression models:

Model 1: Efficiency Only

- Predictor: player_srv_ret_balance
- Accuracy: 0.787
- Interpretation: Combined serve-return efficiency alone is a strong predictor of match outcomes.

Model 2: Efficiency + Return + First-Serve Points Won

CHECKPOINT 7 — Analysis Memo #2

- Predictors: player_srv_ret_balance, player_ret_1st_pct, player_1st_points_won
- Accuracy: 0.878
- Interpretation: First-serve return percentage and first-serve points won significantly increase predictive performance.

These two models form the baseline for the CP7 improvements.

4. CP7 Model Upgrades

Checkpoint 7 requires one or two targeted improvements. Two upgrades were added in this checkpoint, both directly supported by CP5 exploration and tennis logic.

Upgrade 1: Break-Point Save Percentage
CP5 analysis showed differences in break points saved between winners and losers, but raw counts do not account for the number of opportunities faced. To improve interpretability, CP7 replaces break-point counts with a percentage measure:

BP Save % = bpSaved / bpFaced

This metric better captures performance under high-pressure moments.

Upgrade 2: Efficiency × Surface Interaction Terms
CP5 surface analysis demonstrated that efficiency differences vary by surface. The gap between winners and losers was largest on grass, smallest on clay, and moderate on hard courts. To model this relationship, CP7 introduces interaction terms:

eff_x_grass
eff_x_clay
eff_x_hard

These terms allow the effect of efficiency on winning probability to change appropriately across surfaces.

5. Analysis Specification (Model 3)

Outcome
Binary win/loss indicator (1 = winner, 0 = loser).

Predictors Included

CHECKPOINT 7 — Analysis Memo #2

Baseline CP6 predictors:

- player_srv_ret_balance
- player_ret_1st_pct
- player_1st_points_won

CP7 additions:

- player_bp_save_pct
- eff_x_grass
- eff_x_clay
- eff_x_hard

Row Definition
One row per player per match.

Sample
ATP singles matches from 2008 through 2024.

Train/Test Split
70 percent training, 30 percent testing, stratified by class.

6. Results

Model 3 (Upgraded CP7 Model)
Accuracy: 0.882

Confusion Matrix
[11473 1467]
[1431 10247]

Classification performance:

- Precision: 0.88 for both classes
- Recall: 0.88 for both classes

- F1-score: 0.88 for both classes

Coefficient Summary:

- player_srv_ret_balance: 23.131379
- player_ret_1st_pct: 21.160654
- player_1st_points_won: 0.001683
- player_bp_save_pct: 2.132935
- eff_x_grass: 5.282919
- eff_x_clay: 5.533910
- eff_x_hard: 5.655250

Interpretation:
Efficiency remains the single strongest predictor of match outcomes, followed closely by first-serve return percentage. Break-point save percentage provides additional explanatory value by capturing high-pressure performance.

All surface interaction terms are positive, indicating that efficiency is beneficial across surfaces, with the coefficients suggesting that its largest marginal effect occurs on hard courts.

7. Stability Check

A stability check was conducted using hard-court matches only.

Hard-court-only accuracy: 0.88

The similarity between the hard-court accuracy and the full-sample accuracy (0.882) demonstrates that the CP7 model generalizes well and that the new features do not destabilize performance across subsets.

8. Interpretation

The CP7 upgraded model demonstrates that:

- Serve-return efficiency remains the most important individual indicator of match success.
- First-serve return percentage is highly influential in determining outcomes.
- First-serve points won contributes positively to winning probability.

CHECKPOINT 7 — Analysis Memo #2

- Break-point save percentage improves predictive performance by capturing success in decisive, high-pressure scenarios.

- The influence of efficiency is surface-dependent, with the largest effect on hard courts and smaller effects on clay.

These results are consistent with earlier descriptive findings and extend CP5 insights into an enhanced predictive model.

9. Limits

The logistic regression framework assumes linearity in the log-odds and does not represent complex nonlinear relationships. Break-point save % may be volatile when the number of break-point opportunities is low.

Additionally, the model does not include opponent-specific effects, strategic patterns, or ball-tracking data, which limits insights into tactical behavior.

10. Next Steps

The next checkpoint will finalize the project by selecting the best-performing model, constructing a polished executive summary, and producing final figures, tables, and documentation. The GitHub repository will be updated to include all completed checkpoints, cleaned datasets, and modeling code.