CHECKPOINT 8 - Executive Summary

Nicholas Condos

Dec 12, 2025

https://github.com/nicholascondos/Tennis_Analytics

Executive Summary

Serve–return efficiency and first-serve return skill are the strongest predictors of ATP match success, driving an 88 percent accurate model built from 17 years of match data. This project analyzes over 43,000 ATP singles matches from 2008 to 2024 to identify which performance metrics most consistently separate winners from losers and to translate these insights into actionable recommendations for coaching and strategy development.

The analysis leverages Jeff Sackmann's ATP dataset and proceeds through systematic cleaning, feature engineering, exploratory data analysis, and predictive modeling. Across all checkpoints, the goal is to build a model that is not only accurate but also interpretable and grounded in tennis performance theory. The final model highlights a small set of key metrics that have outsized influence on match outcomes and provides a clear direction for improving competitive performance.

Data and Approach

The project uses match-level ATP data that is cleaned to remove doubles, qualifying rounds, walkovers, retirements, and incomplete entries. The final dataset includes completed main-draw singles matches with standardized surface labels and a uniquely defined match key.

To enable player-level modeling, each match is transformed into two rows: one for the winner and one for the loser. This creates a balanced dataset of approximately 86,000 player-match observations. Critical features used in the modeling process include:

- Combined serve–return efficiency

- First-serve return percentage

- First-serve points won

- Break-point save percentage

- Surface indicators and efficiency interactions

CHECKPOINT 8 - Executive Summary

Initial exploratory analysis (Checkpoint 5) revealed large and consistent differences in efficiency-related metrics between winners and losers, motivating the inclusion of these variables in predictive models. Subsequent modeling checkpoints (CP6 and CP7) refine the feature set and improve predictive performance.

Modeling Results

The final model (Model 3) is a logistic regression including efficiency, first-serve return percentage, first-serve points won, break-point save percentage, and efficiency-by-surface interaction terms. This upgraded model achieves:

- **Overall accuracy:** 0.882

- **Hard-court-only accuracy:** 0.88

- **Balanced precision, recall, and F1-scores:** 0.88

The model demonstrates strong generalization across surfaces and maintains accuracy even when restricted to the largest subset of matches (hard courts). These results confirm the robustness of the selected predictors and the value of incorporating surface-specific effects.

Exhibit 1. Model 3 Coefficients

(Logistic Regression, 2008–2024 ATP Matches)

Predictor	Coefficient
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

CHECKPOINT 8 - Executive Summary

These coefficients indicate the relative influence each variable has on the likelihood of winning an ATP match after controlling for the other factors in the model. The substantially larger coefficients on serve - return efficiency and first-serve return percentage show that these two metrics carry the greatest explanatory power and align with the patterns observed throughout the exploratory analysis.

The positive but smaller coefficients on break-point save percentage and the surface-specific efficiency interactions suggest that these elements meaningfully refine the model by capturing situational and contextual advantages, though their effects are secondary compared to the primary drivers.

Overall, the coefficient structure reinforces the central finding that point-level efficiency—both on serve and return—remains the most reliable determinant of match outcomes across surfaces and seasons.

Takeaway: Efficiency and first-serve return skill are the dominant determinants of winning probability, while break-point save percentage and surface-adjusted efficiency effects provide meaningful secondary lift.

Key Insights

1. **Serve–return efficiency is the strongest single indicator of match success.**
   Players who consistently win both first- and second-serve points outperform opponents across all surfaces.

2. **First-serve return percentage plays a critical role in determining match outcomes.**
   Winners gain a large advantage by neutralizing opponents' first serves more effectively.

3. **Success in high-pressure moments matters.**
   Break-point save percentage significantly improves model performance and reflects a player's ability to manage decisive points.

4. **Surface conditions modify the impact of efficiency.**
   Efficiency contributes most strongly on hard courts, moderately on grass, and least on clay. Training and strategy should adjust accordingly.

5. **The model is both accurate and interpretable.**
   A logistic regression framework provides transparency into the relative importance of each metric, making the findings accessible for coaches and analysts.

Risks and Limitations

CHECKPOINT 8 - Executive Summary

The model assumes linear effects in the log-odds and may not capture deeper nonlinear relationships or player-specific tendencies. Break-point rates may fluctuate in matches with very few opportunities. The dataset does not include rally-level or ball-tracking information, which limits tactical interpretation. External factors such as fatigue, injuries, opponent identity, or psychological variables are not represented.

What The Next Steps Would Be

With additional time or data, the next steps would include:

- Incorporating richer point-level or rally-level tracking data.

- Modeling opponent-adjusted performance (e.g., return ability relative to average opponent).

- Testing nonlinear approaches such as random forests or gradient boosting for accuracy comparison.

- Exploring time-varying trends (e.g., how serve dominance has evolved).

These extensions would deepen tactical insights and further refine performance predictors.