International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Serve Statistics and Match Outcomes in Tennis: Separating Descriptive Correlates from Genuine Predictors Using Simulation

Samyak Shah

12th Grade A2
Jayshree Periwal International School, Jaipur- Rajasthan

Abstract:
Statistics including first-serve percentage and aces and service points are frequently given as drivers of tennis performance, and machine-learning research frequently classifies match winners based on them. Since these statistics are cumulative up to the time of the match they are supposed to account, they can reflect the outcome but not forecast the outcome. The paper measures such difference through a point-by-point Monte Carlo simulation of 200 seasons (64 players, 50 rounds, best-of-three matches), where the attributes of serve, return and second-serves of each player are the drivers of every point. Five sets of KPI were assessed in logistic regression, linear discriminant analysis, decision trees and support machine with 5-fold cross-validation at n = 90, 180, 360 and 1,000 matches. At n = 180, logistic regression was 90.6% accurate with all post-match KPIs and 86.8% with post-match serve statistics alone, but only 59.2% with pre-match rolling serve statistics (majority-class baseline 53.0%). A pre-match win-rate test score was 66.9 and its inclusion in a test score did not assist (-1.5 points). KPI set accounted 91.4% of accuracy variance; algorithm accounted 1.1%. The choice of features prior to cross-validation resulted in 60.8 percent accuracy using pure noise. Serve statistics are good descriptive correlates of winning but poor prospective predictors; we give a reporting checklist and open code.

Keywords: Tennis analytics; serve statistics; data leakage; match outcome prediction; cross-validation; Monte Carlo simulation

1. Introduction

Tennis has long been analyzed notationally and statistically in search of match properties which distinguish winners and losers (O'Donoghue and Ingram, 2001). Some of the most extensively published are serve-related statistics: the percentage of first serves in, the percentage of points won on first, and second serve, aces, double faults and break points saved are released on virtually every professional match. The hierarchical nature of tennis scoring whereby the result of an outcome of a match is a deterministic function of the result of the outcome of points on serve and return allows one to deduce the probability of winning a game, a set or a match based on the probability of winning points on serve and return (Klaassen and Magnus, 2001; Newton and Keller, 2005). This has inspired models that predict the winner of the match using player-level serve and receive statistics (Barnett and Clarke, 2005; Klaassen and Magnus, 2003) and comparisons of rating-based predictors, including Elo (Elo, 1978; Kovalchik, 2016) and ranking differences (Del Corral and Prieto-Rodrigo, 2010).

In practice two very different questions are often confused. The former is descriptive: do winners display better serve statistics than losers? They nearly never fail to do so, as serve points won are included into the points that determine the match. The second is predictive: can the statistics of serves of a player help to predict the winner before a match? The second is the question that coaches, analysts and forecasters are interested in, but when it comes to the studies of the first type it is often reported that the classification accuracy using post-match statistics answers the question of the second type. This error is referred to as target leakage in machine learning (Kaufman et al., 2012). In a recent basketball study, as an example, the winners of contemporaneous KPIs were classified in 180 games and reported cross-validated accuracies of 78.9 to 85.1% (Liang et al., 2025); the same is true of any sport where the outcome is constructed based on the statistics being used.

The problem is aggravated by small samples and preprocessing. In tennis performance research, samples of one or two hundred matches are common, and the sampling error of cross-validated accuracy is wide (Varoquaux, 2018; Vabalas et al., 2019), and an unbiased measure of its variance is not available (Bengio & Grandvalet, 2004). Estimates are optimistically biased when the full dataset is used to screen the variables prior to cross-validation (Ambroise and McLachlan, 2002). It is impossible to separate these effects in observational data because the actual data-generating process of real matches is unknown. A known process simulation is possible.

Our question is: to what extent is the observed predictive value of tennis serve statistics descriptive circularity, and what is the size of the remaining true pre-match signal relative to a simple results-based benchmark, and how well can small-sample studies separate algorithms? The contributions are four times. In the first stage, we divide descriptive and predictive accuracy on the same simulated matches. Second, we examine whether serve statistics provide predictive information not already contained in a player win record, a standard seldom used in KPI research. Third, we characterize the sampling distribution of cross-validated accuracy at literature-typical sample sizes. Fourth, we have an open code and reporting checklist. There were five hypotheses before the analysis. **H1:** accuracy will rank after match all KPIs > after match serve statistics > before match KPI sets. **H2:** the cross-validated accuracy will be so unstable at n 180 that the rankings of algorithms will be unstable. **H3**: the type of KPI set will explain more variance in accuracy than the type of algorithm. **H4**: pre-match serve statistics will contribute little predictive value over a pre-match win-rate baseline. **H5:** the choice of features prior to cross-validation will inflate accuracy with uninformative features, but the choice of normalisation prior to cross-validation will not.

2. Method

2.1 Design and rationale

We have performed a Monte Carlo study of a fully crossed 5 (KPI set) × 4 (algorithm) × 4 (sample size) design replicated 200 times using independent seasons. This is done with a simulation because we need a known way of generating an outcome, to differentiate the leakage from the signal. There was no involvement with human data or human participants. The data-generating process (DGP) model is an illustrative representation of a professional tour, not a calibrated one, and orders of conditions—not levels of accuracy per se—are the primary focus of inference.

2.2 Data-generating process

Each reproduction modeled a season of 64 players and 50 rounds. Each player had one best-of-three-sets match against a randomly selected opponent in each round (32 matches per round, 1,600 per season). Four

latent attributes were randomly selected by each player, which were N(0, 1): serve power, serve consistency, second-serve quality and return skill. Simulation of matches was point by point. The server served the player with a probability on the first serve on each service point; an ace was served with a probability which depended on the player serving, otherwise the server won the rally with a probability which declined with the skill of the returning player. A first serve that was missed resulted in a second serve, which was a double fault with a player-specific probability, otherwise a rally was decided with second-serverally rally probabilities (Table 1). Games, sets (tiebreak at 6 6 in each set) and matches were scored conventionally. Tiebreak points counted towards total points won but not towards serve statistics. The assignment of player A was random, thus, no structural home or first-listed advantage.

Table 1- ***Parameters of the data-generating process (per service point)***

Component	Probability specification	Latent attributes
First serve in	0.62 + 0.0375 × consistency	Consistency
Ace (given first serve in)	0.08 + 0.0375 × power	Power
First-serve rally won	0.68 + 0.008 × power − 0.0375 × opp. return	Power, return
Double fault (given first serve out)	0.05 − 0.006 × consistency	Consistency
Second-serve rally won	0.55 + 0.030 × second-serve quality − 0.0375 × opp. return	Second serve, return
Latent attributes (×4)	N(0, 1) per player per season; constant within a season	—

Note. Probabilities were clipped to plausible ranges. Parameters were chosen so that serve points won average about 64% and the best-possible pre-match accuracy (an oracle using the true latent attributes) is about 70%; they were not fitted to tour data.

2.3 KPI sets

The following were calculated for each player per match: service points won, first serve in %, first serve points won, second serve points won, ace rate, double fault rate and break points saved (shortened towards 60% with 3 pseudo-observations to avoid undefined values). Five sets of KPIs were compared, all differences between player A and player B. All KPIs: the seven serve stats, as well as breaks of serve and share of total points won, will be posted after match. Post-match serve stats: the 7 serve stats of the match. (c) Pre-match, serve statistics: expanding window referring to the 7 statistics of the player's previous games. (d) Pre-match, win rate: each player's smoothed win over previous games; (wins + 1)/(matches + 2), a results-only measure, similar to rating systems. Pre-match, serve stats + win rate. The sets were available from round 11, and there were 1280 qualifying rounds per season. Genuine prospective prediction is allowed only for sets (c)–(e).

2.4 Algorithms and validation

To determine whether player A won, we used logistic regression (LR), linear discriminant analysis (LDA), a decision tree (DT; maximum depth 5, minimum leaf size 5) and a radial-kernel support vector machine

(SVM; C = 1) implemented in scikit-learn (Pedregosa et al., 2011; see Hastie et al., 2009). The data was scaled by min–max in training folds. Accuracy was the average of 5-fold stratified and shuffled cross validation. For each replication, the number of matches sampled without replacement (90, 180, 360 or 1,000) was recorded. For the majority class, the baseline rate was recorded.

2.5 Experiments

- **Experiment 1** (leakage and algorithm comparison; H1, H3). Cross-validated accuracy of all KPI set × algorithm × n cell.
- **Experiment 2 (sampling variability; H2).** Variance of error among replicates and the fraction of replicates where each algorithm was optimal (n = 180).
- **Experiment 3 (univariate winner-loser tests).** One-sample t-tests of the winner-loser difference in each KPI, post-match versus pre-match, with reports of detection rates (proportion of replications with $p < .05$) and mean Cohen d at n = 180.
- **Experiment 4 (incremental value; H4).** similar comparisons of pre-match sets (d) and (e) with (c) and (d).
- **Experiment 5 (preprocessing leakage; H5).** At n = 180, logistic regression on (i) 50 pure-noise KPIs, (ii) the seven pre-match serve statistics and 43 noise KPIs, choosing the five best by ANOVA F-test on the full data (leaky) or within training folds (correct); and (iii) minmax normalisation on the full data versus within folds.

2.6 Statistical analysis

Paired comparisons were made since conditions were compared against each other on the same simulated datasets. Each contrast is reported with the mean of the paired difference in accuracy and a 95% t-based confidence interval, the standardised paired effect (dz; Lakens, 2013) and a Wilcoxon signed-rank p-value adjusted within a set of contrasts using the Holm (1979) method. The effects were interpreted with respect to the standards of Cohen (1988). There was a descriptive sum-of-squares decomposition ($\eta 2$) to provide the portion of the total variance in accuracy that can be attributed to KPI set, algorithm, sample size and sample interactions; conditions within a replication are correlated so $\eta 2$ is descriptive and not inferential. With 200 replications the Monte Carlo standard error of any cell mean is below 0.5 percentage points. Python analyses with NumPy (Harris et al., 2020), SciPy (Virtanen et al., 2020) and scikit-learn were used.

3. Results

3.1 Descriptive winner–loser differences

Paired comparisons were made since conditions were compared against each other on the same simulated datasets. Each contrast is reported with the mean of the paired difference in accuracy and a 95% t-based confidence interval, the standardised paired effect (dz; Lakens, 2013) and a Wilcoxon signed-rank p-value adjusted within a set of contrasts using the Holm (1979) method. The effects were interpreted with respect to the standards of Cohen (1988). There was a descriptive sum-of-squares decomposition ($\eta 2$) to provide the portion of the total variance in accuracy that can be attributed to KPI set, algorithm, sample size and sample interactions; conditions within a replication are correlated so $\eta 2$ is descriptive and not inferential. With 200 replications the Monte Carlo standard error of any cell mean is below 0.5 percentage points. Python analyses with NumPy (Harris et al., 2020), SciPy (Virtanen et al., 2020) and scikit-learn were used.

Table 2- ***Univariate winner–loser tests at n = 180 (200 replications)***

KPI	Post-match: % rep. p < .05	Post-match: mean |d|	Pre-match: % rep. p < .05	Pre-match: mean |d|
Service points won	100%	1.00	83%	0.23
First-serve points won	100%	0.73	54%	0.17
Second-serve points won	100%	0.60	60%	0.16
First serve in %	49%	0.15	30%	0.10
Ace rate	70%	0.19	54%	0.16
Double-fault rate	52%	0.14	28%	0.09
Break points saved	100%	0.50	37%	0.11
Breaks of serve	100%	1.34	n/a	n/a
Total points won	100%	1.24	n/a	n/a
Win rate	n/a	n/a	100%	0.44

Note. One-sample *t*-tests of the winner-minus-loser differential against zero. Breaks of serve and total points won exist only post-match; win rate only pre-match.

3.2 Predictive accuracy by KPI set and algorithm

The results of Experiment 1 are summarized in Table 3 and Figure 1. Again as in H1, accuracy occurred after each match and serve for all algorithms and sample sizes, and was ordered after the pre-match sets. All post-match KPIs correctly classified 90.6% of matches at n = 180; logistic regression correctly classified 86.8% of the matches with all post-match KPIs, 66.9% of matches with the pre-match win rate only, 65.4% with win rate plus serve statistics and 59.2% with the pre-match serve statistics only, against a majority-class baseline of 53.0%. Compared to post-match KPIs (which reached 93.1% asymptotic ceilings when pooling ten simulated seasons or 12,800 matches), pre-match serve statistics had a ceiling of 58.4%, win rate 67.3%, both 67.5%, and an oracle with the true latent attributes had a ceiling of 69.6%. The omission of return skill from serve statistics will therefore only pick up a portion of the true pre-match signal in this DGP, and this is largely due to it being the other dependent variable driving outcomes.

Table 3- ***Cross-validated accuracy (%) by KPI set and algorithm at n = 180: mean (SD) over 200 replications***

KPI set	LR	LDA	DT	SVM
Post-match, all KPIs	90.6 (2.1)	93.9 (1.5)	91.7 (2.4)	91.0 (2.2)
Post-match, serve statistics	86.8 (2.5)	86.5 (2.6)	82.9 (3.8)	86.2 (2.8)
Pre-match, serve statistics	59.2 (4.2)	58.4 (4.4)	53.7 (4.4)	57.7 (4.8)
Pre-match, win rate	66.9 (3.6)	67.0 (3.6)	61.8 (4.6)	66.5 (3.9)
Pre-match, serve + win rate	65.4 (3.9)	65.3 (4.2)	59.2 (4.7)	64.2 (4.5)

Note. LR = logistic regression; LDA = linear discriminant analysis; DT = decision tree; SVM = support vector machine. Majority-class baseline = 53.0%.

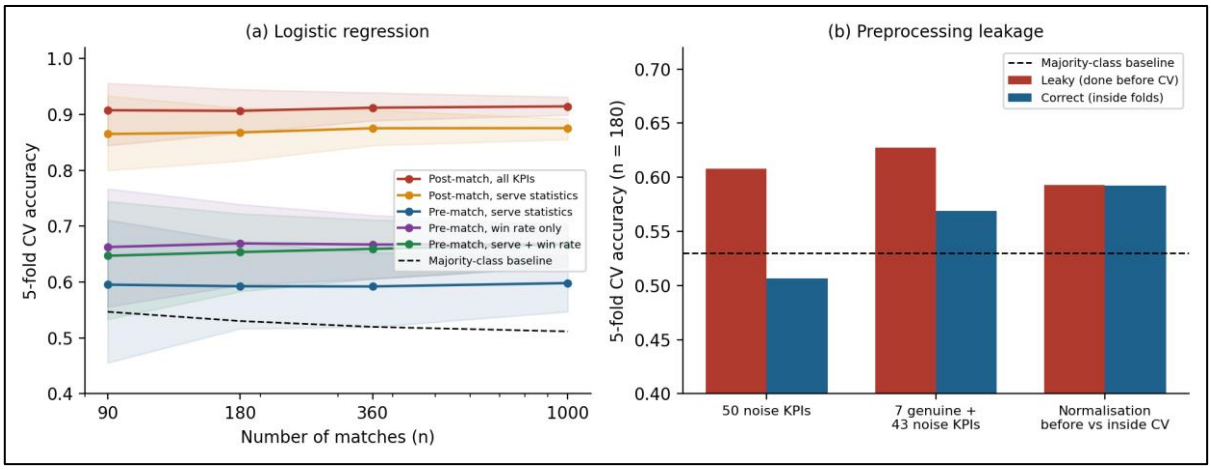


Figure 1. (a) Mean cross-validated accuracy of logistic regression by sample size and KPI set; bands are 2.5th–97.5th percentiles across replications. (b) Mean accuracy at n = 180 when feature selection or normalisation is done before versus inside cross-validation.

Table 4 provides the paired contrasts versus pre-match serve statistics, the only serve-based set that can be used to predict. Post-match serve statistics surpassed it by approximately 28 percentage points when using logistic regression, and all post-match contrasts across the 32 combinations of algorithm and sample size had Holm-adjusted p =.001 (smallest mean difference 27.0 points). The difference did not decline with sample size (27 points at n = 90; 28 at n = 1,000), which would be found in case the difference is due to leakage and not to error in estimating the difference. H1 was supported.

Table 4- *Paired differences in accuracy (percentage points) relative to pre-match serve statistics at n = 180*

Model	Post-match all KPIs − pre-match serve [95% CI]	d z	Post-match serve − pre-match serve [95% CI]	d z
LR	31.4 [30.8, 32.0]	6.94	27.5 [26.9, 28.2]	6.01
LDA	35.5 [34.9, 36.1]	7.80	28.0 [27.4, 28.7]	5.74
DT	38.0 [37.4, 38.7]	7.78	29.2 [28.4, 29.9]	5.24
SVM	33.3 [32.6, 34.0]	6.61	28.4 [27.7, 29.2]	5.42

Note. All Holm-adjusted Wilcoxon $p < .001$.

3.3 Does the algorithm matter?

Variance decomposition showed that 91.4% of the variance in accuracy was due to KPI set, 1.1% was due to algorithm, 0.2% was due to the interaction of KPI set and algorithm, and 6.7% was due to the residual. Therefore it was concluded that KPI set explained about 83 times more variance than algorithm, supporting H3. In pre-match serve statistical results, logistic regression did not significantly differ from logistic regression or LDA at n = 1,000 or n = 360 (mean differences of -0.2 and -0.0 points, respectively; Holm p = 0.26 and 0.26, respectively), while LDA significantly outperformed the decision tree and the SVM by

5.6 and 1.5 points, respectively, at n = 180 (dz = -1.26 and -0.45, respectively; both p < .001). Simple models worked best, like many small-sample problems with a near-linear signal.

3.4 Precision of accuracy estimates and ranking stability

An experiment (Experiment 2) had a large sampling error. The SD of cross validated accuracy for logistic regression was 6.0 points for n = 90, and 4.2 points for n = 180, while corresponding to the 95% interval across replications was 51.7% to 67.2% for n = 180, compared with just 2.6 points for n = 1,000. Logistic regression beat the majority-class baseline by 6.2 points on average (95% CI [5.6, 6.9]), but the decision tree was not distinguishable from it at n = 180 (0.7 points, 95% CI [0.0, 1.3], Holm p = 0.13). The best algorithm per replication on pre-match serve statistics: Logistic regression (44%), LDA (29%), SVM (23%) and DT (5%). On the other hand, it is not implausible that the recommendation was found in only one study in which 180 matches were investigated, supporting H2.

3.5 Incremental value of serve statistics

Experiment 4 (Table 5) indicated that the results-only benchmark was superior to pre-match serve statistics for prediction: the pre-match win rate was superior to the pre-match serve statistics by 7.7 points (logistic regression, n = 180; dz = 1.75). When serve stats were thrown in, it didn't improve. It reduced accuracy by 1.5 points at n = 180 and 0.8 at n = 360 (both Holm p < .001), and made no difference at n = 1,000 (-0.0 points, Holm p = 0.45). The additional predictors add noise but not new information that is not already provided by the win record. H4 was supported, but subject to the DGP (Section 4.4).

Table 5- ***Incremental value of pre-match serve statistics (percentage points of accuracy) at n = 180***

Model	Win rate − serve statistics [95% CI]	d z	Serve + win rate − win rate [95% CI]	d z
LR	7.7 [7.1, 8.3]	1.75	-1.5 [-1.9, -1.2]	-0.60
LDA	8.6 [7.9, 9.2]	1.84	-1.7 [-2.1, -1.3]	-0.56
DT	8.1 [7.3, 8.9]	1.47	-2.6 [-3.3, -1.9]	-0.53
SVM	8.8 [8.1, 9.5]	1.68	-2.3 [-2.9, -1.8]	-0.64

Note. Positive values in column 2 mean the win rate outperformed serve statistics; negative values in column 4 mean that adding serve statistics lowered accuracy.

3.6 Preprocessing leakage

On the original 50 uninformative KPIs, selecting the five best on the complete data prior to cross-validation yielded 60.8% accuracy, 7.8 points better than the 53.0% accuracy of a model with no signal, but the correct in-fold selection yielded 50.6% accuracy, which is at chance (Table 6). Leaky selection overstated accuracy by 5.8 points (95% CI [5.2, 6.4]) with its seven genuine and 43 noise KPIs. Normalising before cross-validation changed accuracy by 0.02 points (95% CI [-0.06, 0.11], p = 0.78). H5 was confirmed: the hazard is target-dependent steps, not scaling.

Table 6- *Effect of performing preprocessing before versus inside cross-validation (n = 180)*

Setting	Leaky accuracy (%)	Correct accuracy (%)	Inflation, points [95% CI]	d z
50 pure-noise KPIs	60.8	50.6	10.14 [9.58, 10.69]	2.55
7 genuine + 43 noise KPIs	62.7	56.9	5.82 [5.24, 6.40]	1.39
Normalisation before vs inside CV	59.3	59.2	0.02 [-0.06, 0.11]	0.04

Note. Logistic regression with five selected features. Majority-class baseline = 53.0%.

4. Discussion

4.1 Principal findings

Serve statistics as descriptors and predictors are very different. In the match, they divide the winners and losers almost perfectly and aid classification effectiveness of approximately 87%. The same statistics, calculated by using previous games, favour the accuracy of around 59 at n = 180 (ceiling 58), marginally above the 53.0% base, and marginally below a result-only benchmark (67). The difference of about 28 points between the two is leakage: service points that are won will be included in the points of determination of the match. The compatibility of high accuracy reports on match-day serve statistics with a weak predictive future result ability is thus compatible. These models stand up as accounts of the manner matches were won.

The claims that are frequently attached to sample size and algorithm choice do not match the size or choice. Sampling error (SD of the order of 4-6 points) is worse than the difference between algorithms at 90180 matches and the algorithm with the highest performance varies with dataset. The small number of trustworthy differences favour very simple models with low variance. Finally, procedural leakage can produce above-chance accuracy as a result of noise.

4.2 Implications for research and practice

We recommend the following be reported in studies using KPIs to predict tennis outcomes:

- Whether a KPI is known before the match; contemporaneous KPIs should be called descriptive correlates not predictors.
- In addition to accuracy, a results-based benchmark (win rate, ranking or an Elo-type rating) and a majority-class baseline should be used to test any claims that add value given statistics (Kovalchik, 2016).
- Avoid using arithmetically relevant metrics as KPIs in prediction models, e.g. service points won, total points won and breaks of serve.
- Confidence intervals of accuracy, not just a single point estimate, and an indication of sample size and whether that is appropriate.
- Feature selection, tuning, or any other target dependent step that is completed in training folds.
- As players appear in multiple matches, chronological validation (train on prior games, test on later games) is recommended, with the matches grouped by player: this can overestimate performance if the folds are random.
- Playing surface/round quality, opponent quality, and open data/code.

The take home points for coaches and analysts are that serve statistics should be used for diagnostic purposes, such as determining why a game was lost, while ratings should be used for predictive purposes,

such as predicting overall ability. Statements made by bettors and broadcasters about a player, such as "wins when first-serve percentage is above a certain value", are descriptive and should not be interpreted as predictions.

4.3 Relation to existing work

It is consistent with the hierarchical model for tennis scoring (Klaassen & Magnus, 2001; Newton & Keller, 2005) and also the player-statistics model (Barnett & Clarke, 2005; Klaassen & Magnus, 2003): there is genuine signal in the serve statistics that explains the pre-match serve beats the baseline. They also subscribe to the conclusion that the simple rating approaches are difficult to beat (Kovalchik, 2016; Del Corral & Prieto-Rodríguez, 2010). The leakage and small-sample issues are those occurring in other domains (Ambroise & McLachlan, 2002; Kaufman et al., 2012; Varoquaux, 2018; Vabalas et al., 2019) that, to our knowledge, have yet to be quantified in tennis serve-statistics studies.

4.4 Limitations

The DGP is intentionally straightforward. It disregards surface, fatigue, injury, serve-first advantage, within-match momentum (points are not quite independent in actual matches; Klaassen and Magnus, 2001), tiebreak serving order and best-of-five matches, and its latent attributes remain constant within a season. The return skill is a significant contributor to results and therefore serve-only pre-match models are constrained by design; in real data the serve share of skill can vary and the absolute levels of accuracy and ceilings are not to be used as approximations to any tour. The result that serve statistics do not contribute to the win record depends on the win record being a good summary of ability, which is the case in this DGP with equal match volume per player. The number of classifiers was fixed with fixed hyperparameters, cross-validation was done by random, and no real-data validation was conducted. The conclusions relate to research design and any given published study can only be evaluated using its own data.

4.5 Future research

The framework should be used on open match level data (such as ATP and WTA match records and point-level serve statistics) to be validated chronologically and by player-grouping, allowing for the comparison of simulated and empirical inflation. Extensions comprise of surface-specific and opponent-adjusted serve statistics, Elo and Bayesian rating benchmarks, proper scoring rule probability forecasts of the actual tour distribution, and calibration of the DGP to the actual tour distribution.

5. Conclusion

In the point by point simulation of tennis, the accuracy based on post-match serve statistics was approximately 28 percentage points higher than the accuracy based on the same statistics measured before the match and the KPI set achieved an accuracy of approximately 91% compared to the algorithm's accuracy of 1%. At 180 matches, serve statistics were poor predictors, did not provide any improvement over the win-rate benchmark and could not be used to reliably rank algorithms. The above chance accuracy from noise was due to selection of features prior to cross-validation. Serve statistics are good descriptive correlates of winning, but poor prospect predictors, and study should be designed and reported accordingly.

REFERENCES:

1. Ambroise, C., & McLachlan, G. J. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences, 99*(10), 6562–6566. https://doi.org/10.1073/pnas.102102699
2. Barnett, T., & Clarke, S. R. (2005). Combining player statistics to predict outcomes of tennis matches. *IMA Journal of Management Mathematics, 16*(2), 113–120. https://doi.org/10.1093/imaman/dpi001
3. Bengio, Y., & Grandvalet, Y. (2004). No unbiased estimator of the variance of *K*-fold cross-validation. *Journal of Machine Learning Research, 5*, 1089–1105.
4. Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
5. Del Corral, J., & Prieto-Rodríguez, J. (2010). Are differences in ranks good predictors for Grand Slam tennis matches? *International Journal of Forecasting, 26*(3), 551–563. https://doi.org/10.1016/j.ijforecast.2009.11.006
6. Elo, A. E. (1978). *The rating of chessplayers, past and present*. Arco.
7. Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature, 585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2
8. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.
9. Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65–70.
10. Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data, 6*(4), Article 15. https://doi.org/10.1145/2382577.2382579
11. Klaassen, F. J. G. M., & Magnus, J. R. (2001). Are points in tennis independent and identically distributed? Evidence from a dynamic binary panel data model. *Journal of the American Statistical Association, 96*(454), 500–509. https://doi.org/10.1198/016214501753168217
12. Klaassen, F. J. G. M., & Magnus, J. R. (2003). Forecasting the winner of a tennis match. *European Journal of Operational Research, 148*(2), 257–267. https://doi.org/10.1016/S0377-2217(02)00682-3
13. Kovalchik, S. A. (2016). Searching for the GOAT of tennis win prediction. *Journal of Quantitative Analysis in Sports, 12*(3), 127–138. https://doi.org/10.1515/jqas-2015-0059
14. Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for *t*-tests and ANOVAs. *Frontiers in Psychology, 4*, Article 863. https://doi.org/10.3389/fpsyg.2013.00863
15. Liang, Z., Gao, Y., Wang, J., & Liu, Z. (2025). Data-driven insights into basketball performance: Unveiling the impact of advanced analytics on player and team efficiency. *Molecular & Cellular Biomechanics, 22*(1), Article 1083. https://doi.org/10.62617/mcb1083

16. Newton, P. K., & Keller, J. B. (2005). Probability of winning at tennis (I). Theory and data. *Studies in Applied Mathematics, 114*(3), 241–269. https://doi.org/10.1111/j.0022-2526.2005.01550.x
17. O'Donoghue, P., & Ingram, B. (2001). A notational analysis of elite tennis strategy. *Journal of Sports Sciences, 19*(2), 107–115. https://doi.org/10.1080/026404101300036299
18. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.
19. Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PLOS ONE, 14*(11), Article e0224365. https://doi.org/10.1371/journal.pone.0224365
20. Varoquaux, G. (2018). Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage, 180*, 68–77. https://doi.org/10.1016/j.neuroimage.2017.06.061
21. Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods, 17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2