Tennis Odds

Andy, Annie, Karol, Maanav, Martin, Shrey, Victor





Dictionary

Definitions from Oxford Languages · Learn more

ten·nis

/ˈtenəs/

noun

a game in which two or four players strike a ball with rackets over a net stretched across a court. The usual form (originally called **lawn tennis**) is played with a felt-covered hollow rubber ball on a grass, clay, or artificial surface.



Overview

- Goal: Build a model to predict tennis match probabilities
- Benchmark: Match or approach Kalshi pricing accuracy

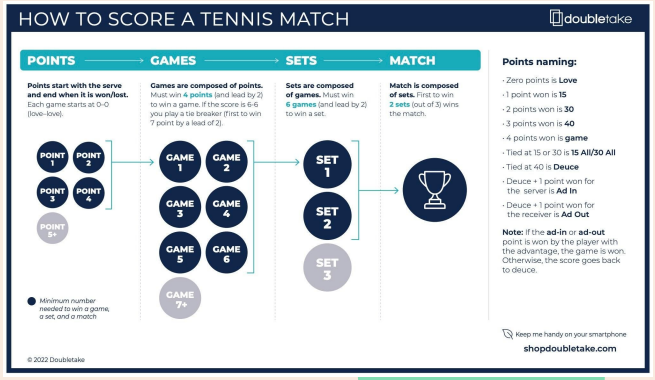
Why Tennis?

- Tennis has clear states:
  - Point → Game → Set → Match
- Individual sport:
  - No teammates → less random
- Potential market inefficiency?
  - Less attention than major sports
  - Maybe inefficient!!



What Tennis?

- Points → smallest unit
  - Goes up in increments of 15:
    - 0 (love) → 15 → 30 → 40 → 60
  - First to 60 (win by 2)
  - Deuce - both players reach 40 → Advantage
- Games → First to 60 points wins a game
- Sets → First to 6 games wins a set (win by 2)
- Match → best of 3 sets
- Association of Tennis Professionals (ATP) - men's
- Women's Tennis Association (WTA) - women's



HOW TO SCORE A TENNIS MATCH

doubletake

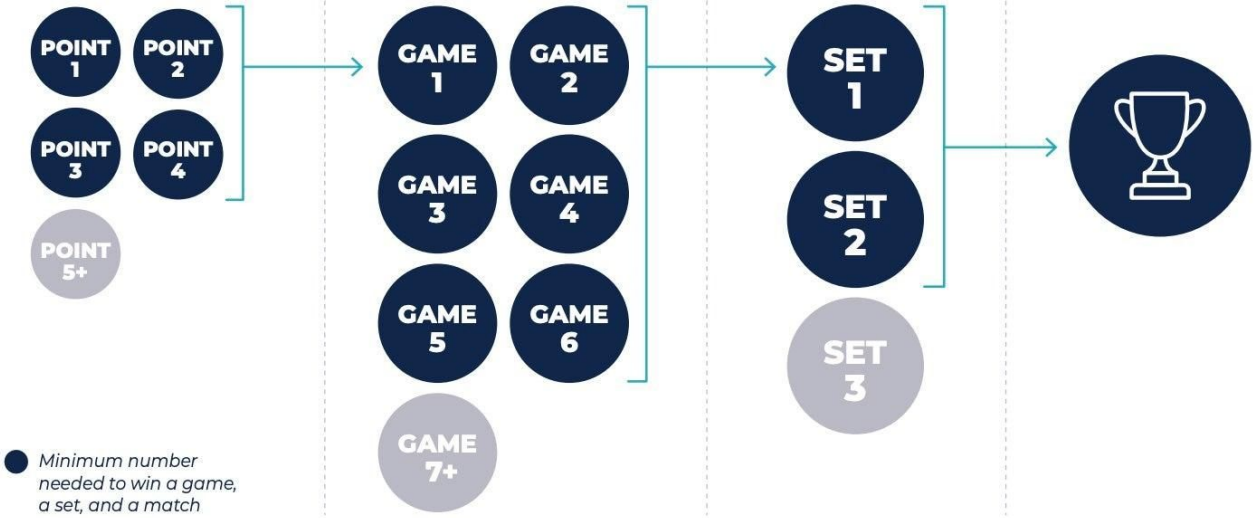
POINTS → GAMES → SETS → MATCH

Points start with the serve and end when it is won/lost. Each game starts at 0–0 (love–love).

Games are composed of points. Must win **4 points** (and lead by 2) to win a game. If the score is 6-6 you play a tie breaker (first to win 7 point by a lead of 2).

Sets are composed of games. Must win **6 games** (and lead by 2) to win a set.

Match is composed of sets. First to win **2 sets** (out of 3) wins the match.



● *Minimum number needed to win a game, a set, and a match*

Points naming:

- Zero points is **Love**
- 1 point won is **15**
- 2 points won is **30**
- 3 points won is **40**
- 4 points won is **game**
- Tied at 15 or 30 is **15 All/30 All**
- Tied at 40 is **Deuce**
- Deuce + 1 point won for the server is **Ad In**
- Deuce + 1 point won for the receiver is **Ad Out**

Note: If the **ad-in** or **ad-out** point is won by the player with the advantage, the game is won. Otherwise, the score goes back to deuce.

Keep me handy on your smartphone

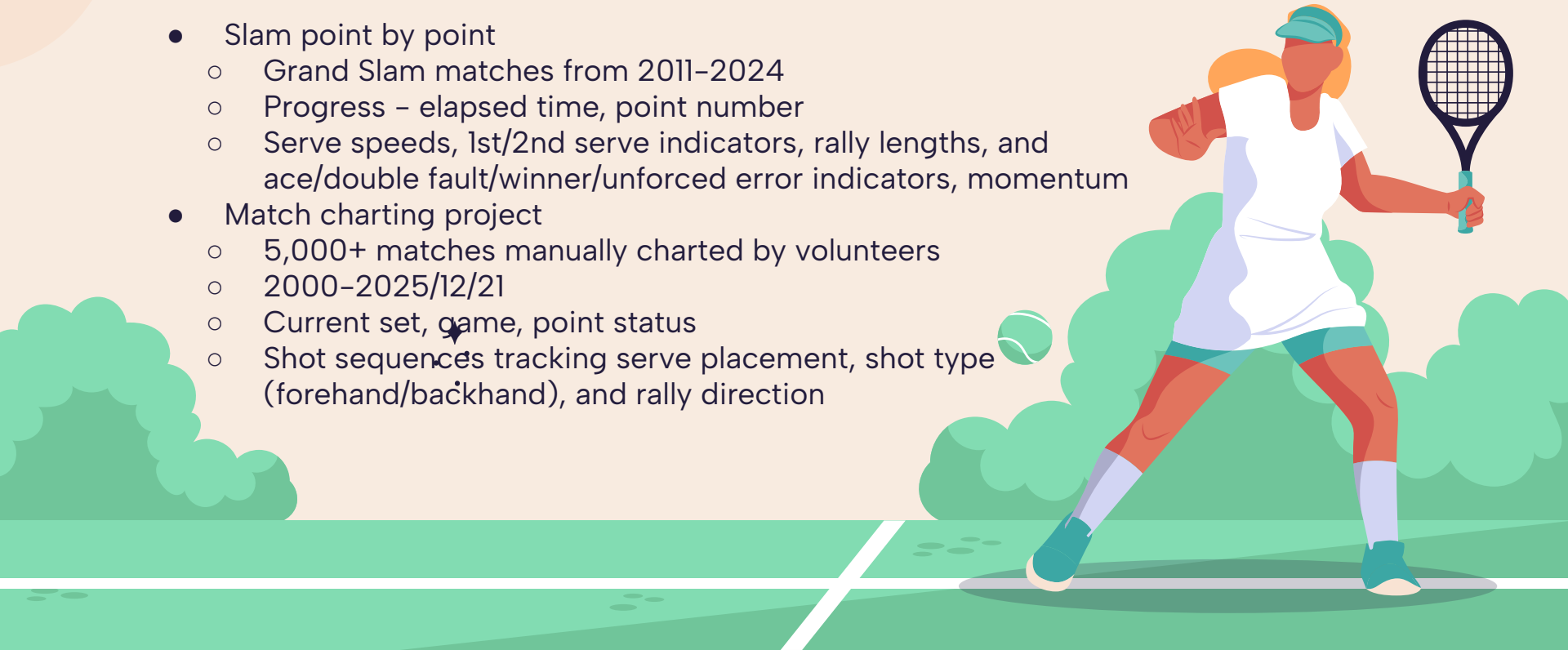
shopdoubletake.com

© 2022 Doubletake

Data Granularity

Point

- Slam point by point
  - Grand Slam matches from 2011–2024
  - Progress – elapsed time, point number
  - Serve speeds, 1st/2nd serve indicators, rally lengths, and ace/double fault/winner/unforced error indicators, momentum
- Match charting project
  - 5,000+ matches manually charted by volunteers
  - 2000–2025/12/21
  - Current set, game, point status
  - Shot sequences tracking serve placement, shot type (forehand/backhand), and rally direction



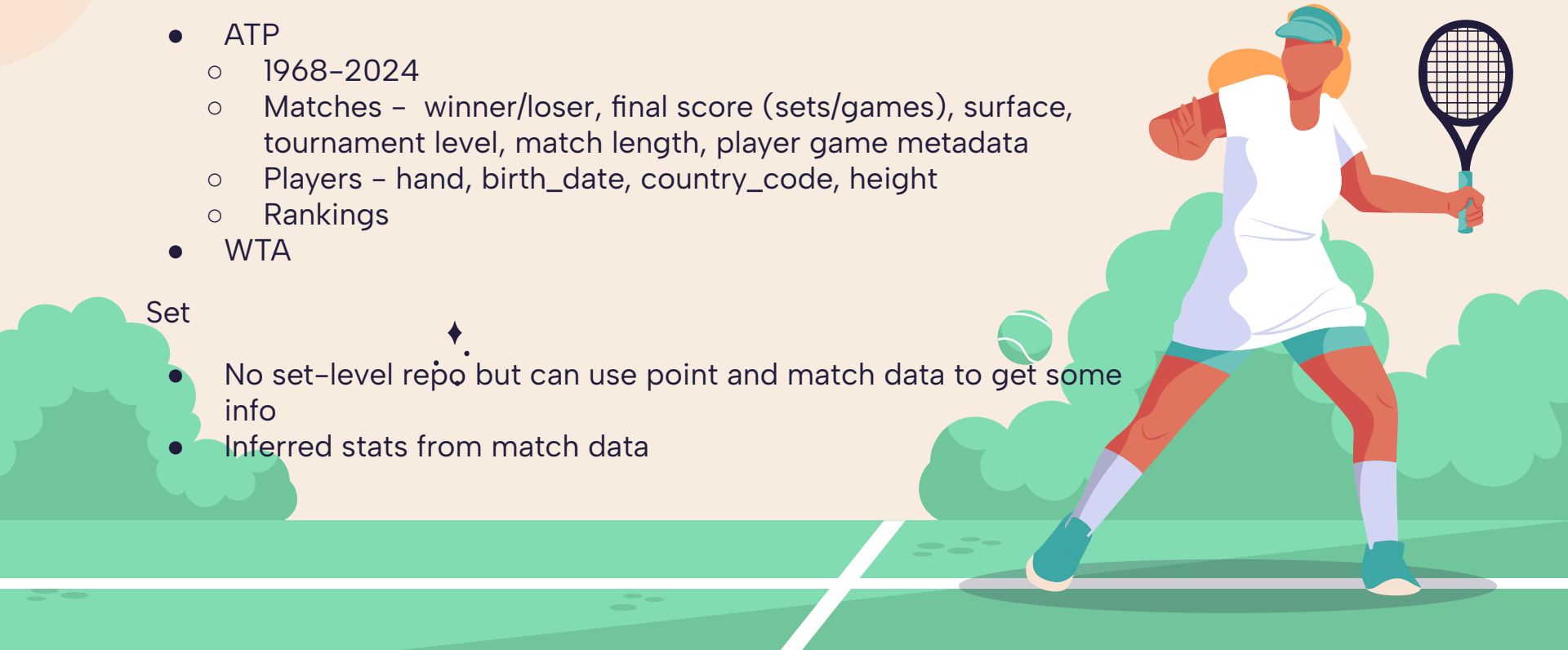
Data Granularity

Match

- ATP
  - 1968–2024
  - Matches – winner/loser, final score (sets/games), surface, tournament level, match length, player game metadata
  - Players – hand, birth_date, country_code, height
  - Rankings
- WTA

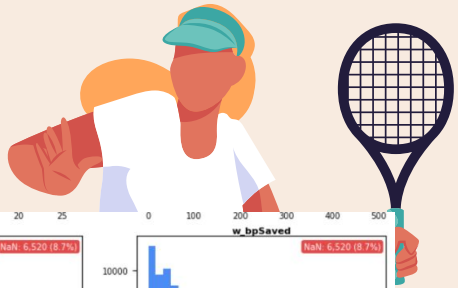
Set

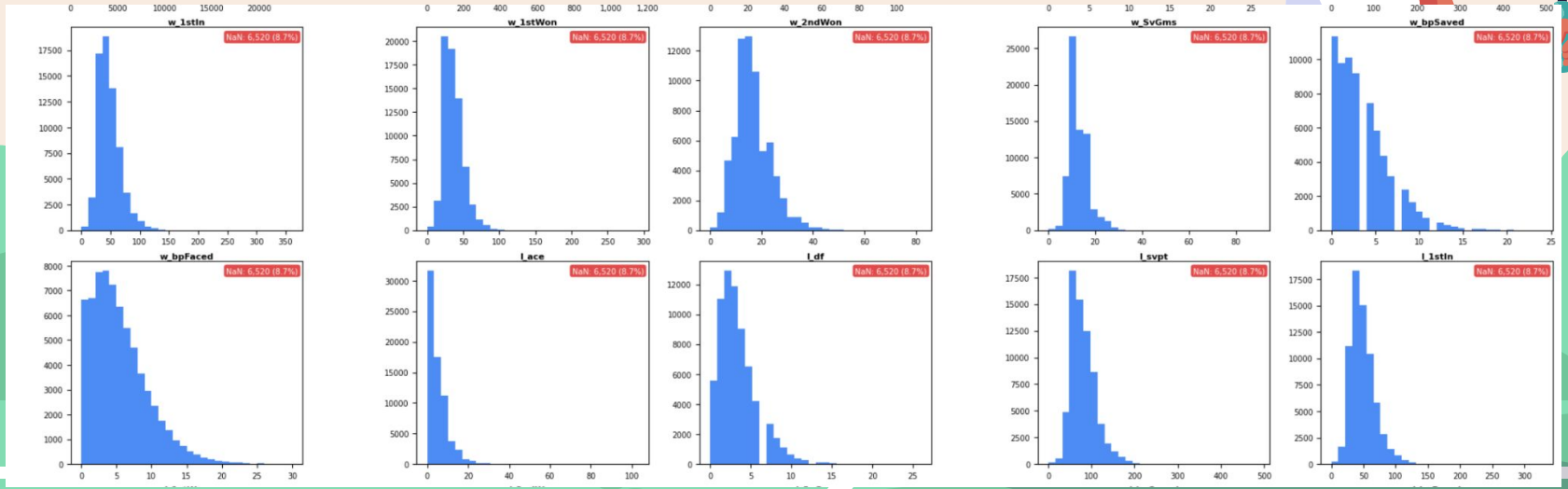
- No set-level repo but can use point and match data to get some info
- Inferred stats from match data



Data Reliability

- Present for most columns
- Less mainstream matches (doubles, qualifying, futures) have sparser data
- Game length missing data
- Older entries (pre 2000) missing more





Bias

- Data sources
- Time
  - Data available for different ranges of years
  - Older data sparser
- Match charting manual submission
  - Popularity bias?
  - Subjectivity – tennis knowledge and personal judgement
  - Fatigue
- Surface homogenization – In the 1990s, grass was faster and clay was slow. Modern surfaces have been "homogenized" (brought closer together)
- ATP
  - Davis Cup matches are included in the tour-level files, but there are no stats for Davis Cup matches until the last few seasons



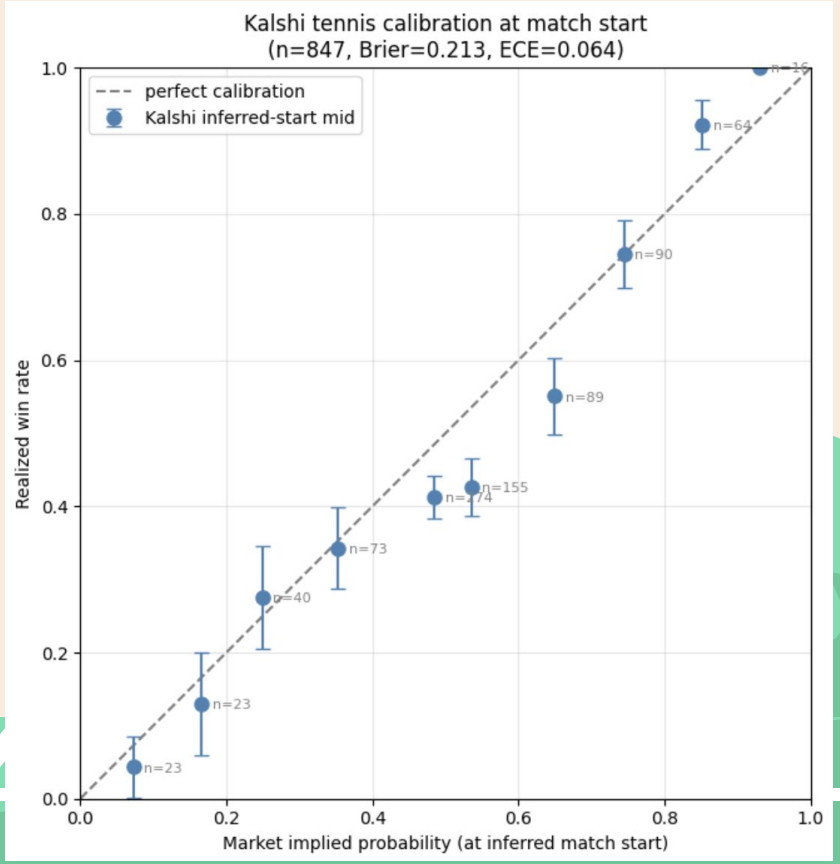
Kalshi market analysis

What is calibration?

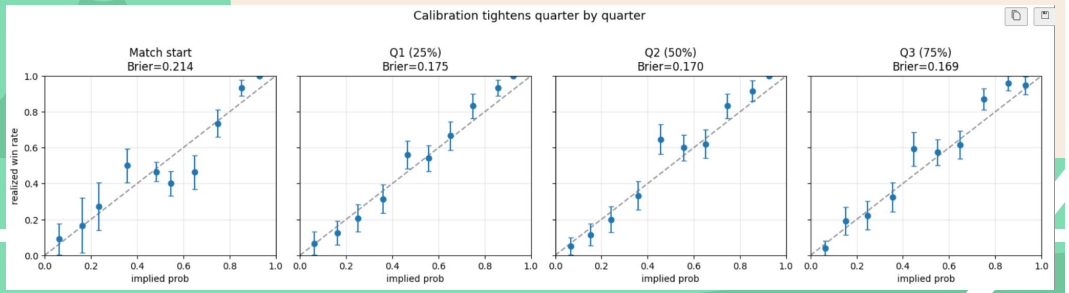
Calibration measures whether a market's stated probabilities match reality. When it says 70%, does the event actually happen 70% of the time?

Is the Kalshi tennis market beatable?

The market is reasonably well-calibrated at match start. Most of its sharpness comes from absorbing live match info.



Dots near the line = market is accurate. Dots below = market is overconfident. Dots above = market is underconfident.



Kalshi market analysis

Where could the edge be?

Strong favorites are underpriced by ~5pp at every phase — the market won't push past ~88% even when it should.

Challengers are noisier than main-tour (Brier 0.223 vs 0.186) — best place to look for pre-match edge.

```
End-game favorite bias (Q3 = 75% phase):
Do people buy favorites too aggressively near the end?

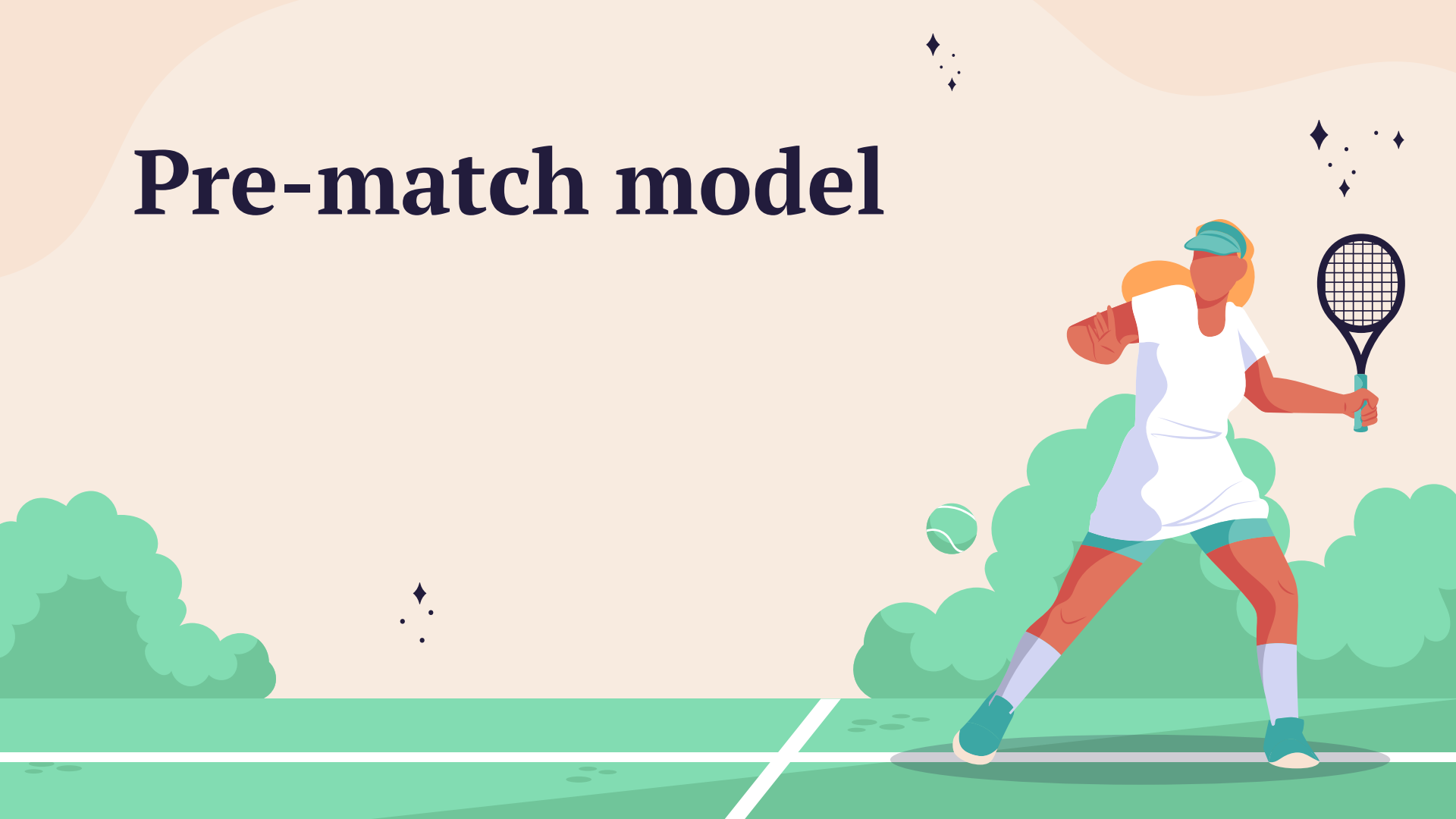
  Match start:
bin_left bin_right  n  mean_pred  mean_actual
     0.8       0.9 36      0.852        0.917
     0.9       1.0 17      0.936        0.941
    Weighted avg: predicted 0.879, realized 0.925 (UNDERCONFIDENT by 4.6 pp)

  Q1 (25%):
bin_left bin_right  n  mean_pred  mean_actual
     0.8       0.9 55      0.852        0.909
     0.9       1.0 25      0.933        0.960
    Weighted avg: predicted 0.877, realized 0.925 (UNDERCONFIDENT by 4.8 pp)

  Q2 (50%):
bin_left bin_right  n  mean_pred  mean_actual
     0.8       0.9 50      0.850        0.900
     0.9       1.0 38      0.931        0.974
    Weighted avg: predicted 0.885, realized 0.932 (UNDERCONFIDENT by 4.7 pp)

  Q3 (75%):
bin_left bin_right  n  mean_pred  mean_actual
     0.8       0.9 53      0.855        0.887
     0.9       1.0 42      0.936        0.952
    Weighted avg: predicted 0.891, realized 0.916 (UNDERCONFIDENT by 2.5 pp)
```

Pre-match model



Data

- Used ATP and WTA CSVs
- Maintained pre-match characteristics only (elo, surface elo, right/left handed, height, age, etc.)
- Old matches missing a lot of data
- 2 ways to calculate elo
  - External elo csv
  - Dynamically updating player elo based off of wins/losses
- Trained model on pre 2018 matches
- Tested model on post 2021 matches



Features

- 5 categories of features
  - Strength (elo, surface elo)
  - Ranking (rank, seed)
  - Win rate (career win rate, surface win rate, 10 match win rate)
  - History (H2H win %)
  - Context (surface, tournament, handedness, height, etc.)



Model setup

$$R'_A = R_A + K(S_A - E_A)$$

New Elo score of player A | Current Elo score of A | Adjustement Rating (K-factor) | Performed score of A | Expected score of A

$$E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}}$$

Expected score of player A | Difference between Elo score of B and A

- Player A vs. Player B, target = 1 if player A wins
- Numeric features stored as differences
  - A elo - B elo
- Convert categorical variables to numerical (surface -> surface_hard, surface_clay, surface_grass)
- 2 models for each elo calculation
  - External elo: Elo csv missing a lot of entries, excluded matches with unknown players
  - Dynamic elo: Chronologically iterated through matches. Current elo used as prematch feature, updates elo accordingly for winner/loser



Top Predictors

	feature	coef	abs_coef
1	num__surface_elo_diff	1.144667	1.144667
0	num__elo_diff	-0.741528	0.741528
60	cat__b_entry_PR	-0.737220	0.737220
53	cat__a_entry_PR	0.737220	0.737220
3	num__rank_points_diff	0.365092	0.365092
51	cat__a_entry_LL	-0.349184	0.349184
58	cat__b_entry_LL	0.349183	0.349183

External Elo

	feature	coef	abs_coef
0	num__elo_diff	0.817586	0.817586
86	cat__a_hand_U	-0.307135	0.307135
90	cat__b_hand_U	0.307119	0.307119
1	num__surface_elo_diff	0.275438	0.275438
89	cat__b_hand_R	-0.262219	0.262219
85	cat__a_hand_R	0.262091	0.262091

Dynamic Elo

Evaluation Metrics

- Log loss: how good predicted probabilities are, heavily penalizes predictions that are confident and wrong
- Brier: Average squared difference between predicted probability and actual result
- Lower = better



	model	log_loss	brier	accuracy@0.5
0	logistic_regression	0.609447	0.211411	0.661632

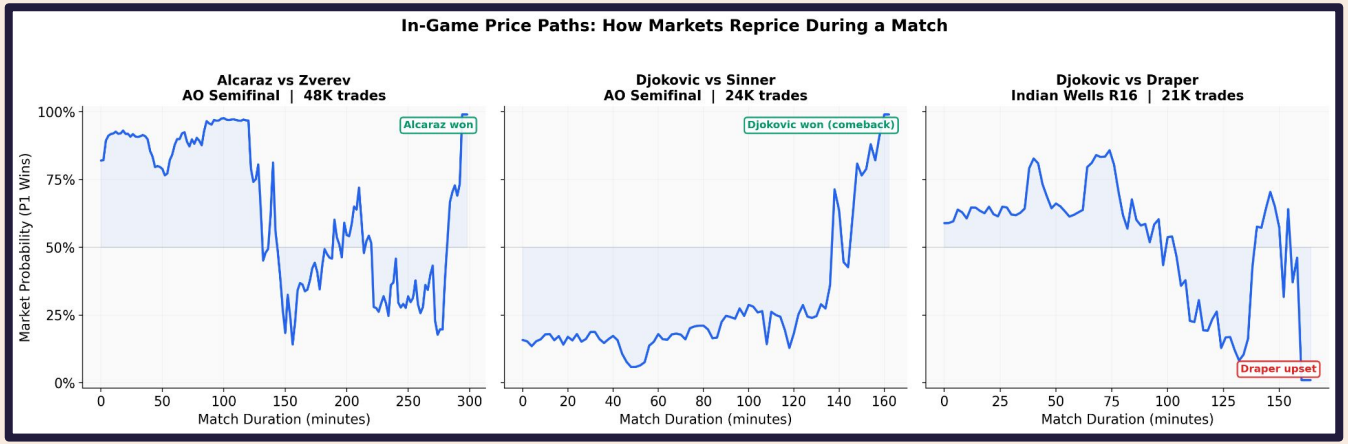
Dynamic

	model	log_loss	brier	accuracy@0.5
0	logistic_regression	0.629512	0.219426	0.648322

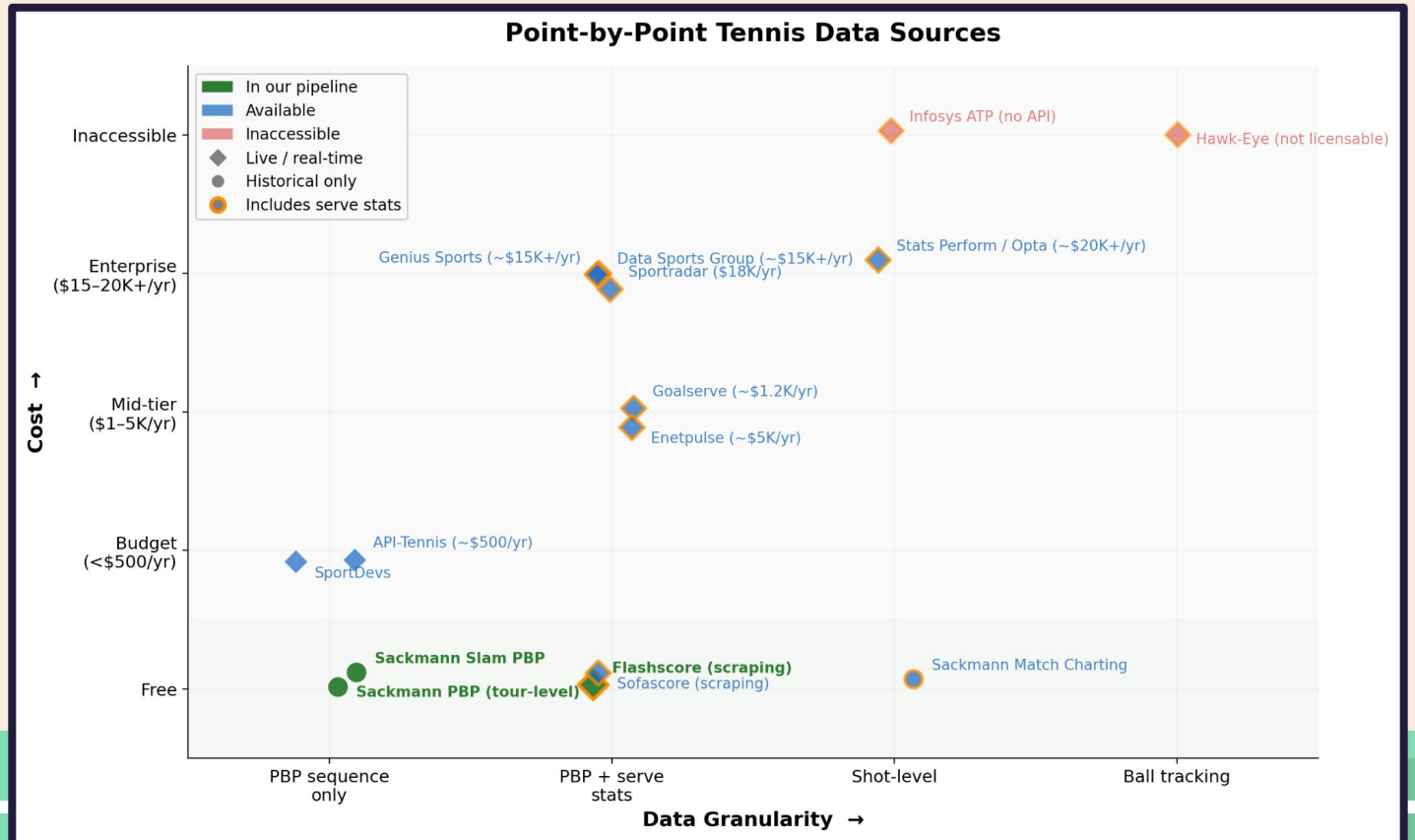
External

In-Game Model

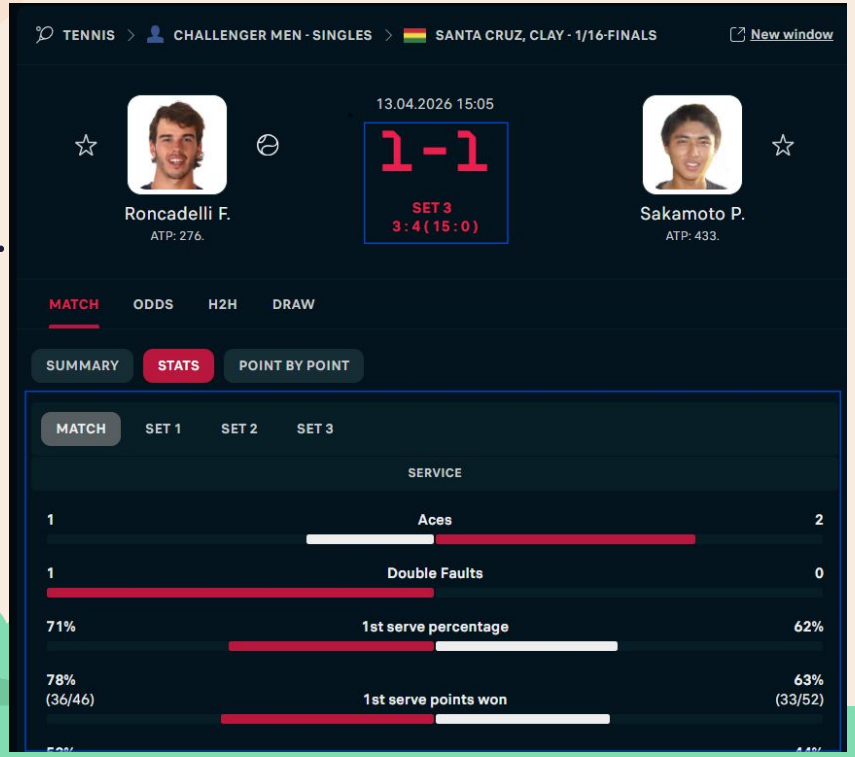
84% of volume is in-match | ~110 model updates



Data Sources



Live Polling





In-game Features

Feature Inventory: 53 Features Across 6 Categories

All features observable from live Flashscore polling or pre-match data

Score State — 7

`set_num`	Current set
`sets_p1 / sets_p2`	Sets won
`games_p1 / games_p2`	Games won
`is_p1_serving`	Server indicator
`total_points`	Points played

Serve Stats (live, cumulative) — 15

`p1/p2_serve_pct`	Overall serve %
`serve_pct_diff`	Serve % gap
`p1/p2_first_pct`	1st serve in %
`p1/p2_first_won_pct`	1st serve won %
`p1/p2_second_won_pct`	2nd serve won %
`p1/p2_deuce_serve_pct`	Deuce-side serve win %
`p1/p2_ad_serve_pct`	Ad-side serve win %
`p1/p2_serve_side_diff`	Deuce vs ad gap

Break Points & Pressure — 8

`p1/p2_bp_save_pct`	BP save %
`p1/p2_bp_conv_pct`	BP convert %
`p1/p2_ace_rate`	Ace rate
`p1/p2_df_rate`	Double fault rate

Momentum — 2

`p1_pts_won_frac`	Overall pt win %
`p1_last_10_won`	Last 10 pts won

Pre-Match (Elo, rankings, H2H) — 16

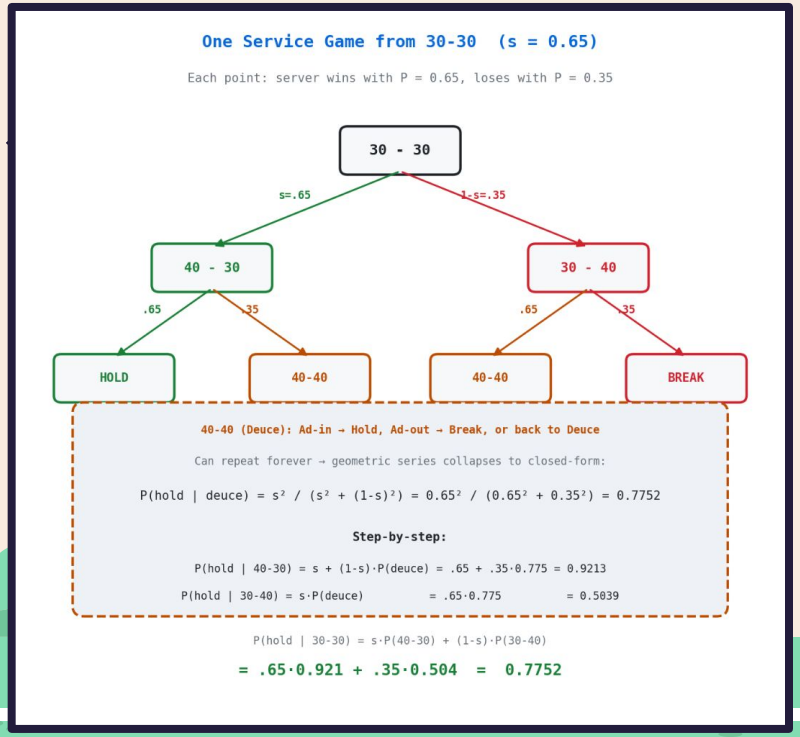
`elo_diff`	Elo rating gap
`surface_elo_diff`	Surface-specific Elo
`log_rank_ratio`	Ranking ratio
`pregame_*_diff`	8 tour-level stat diffs
`height/age_diff`	Physical diffs
`rank_points_ratio`	Points ratio

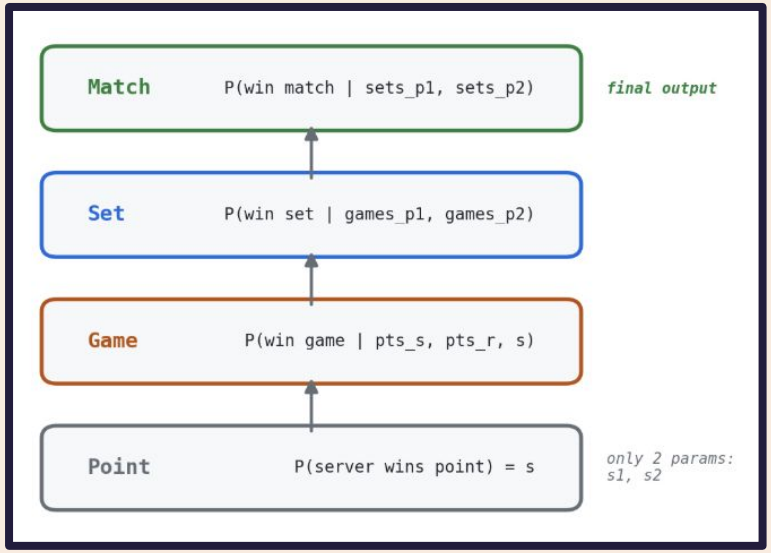
Context — 6

`is_atp / best_of`	Tour & format
`is_clay / is_grass`	Surface indicators
`is_p1/p2_lefty`	Handedness

Noob-Level Model: IID

- **IID (Independently & Identically Distributed) Model:** assumes that each observation is generated independently and comes from the same underlying probability distribution.





Pros: simple & interpretable
Cons: feature complexity

IID

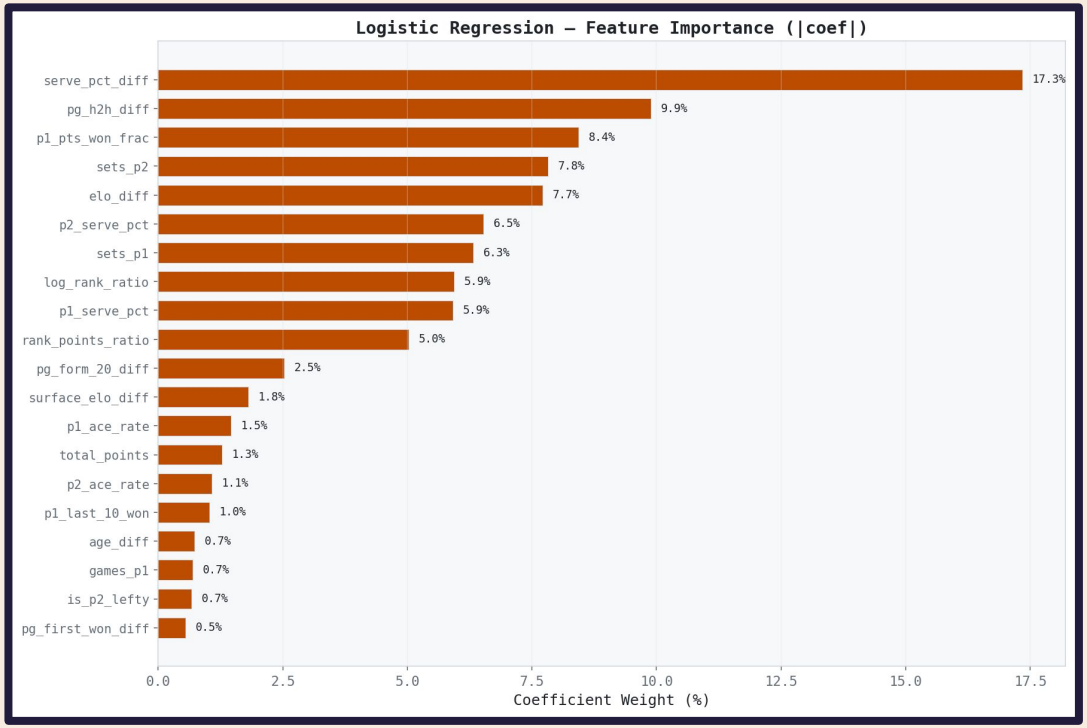
Are Points in Tennis Independent and Identically Distributed? Evidence From a Dynamic Binary Panel Data Model*

Franc J.G.M. Klaassen and Jan R. Magnus

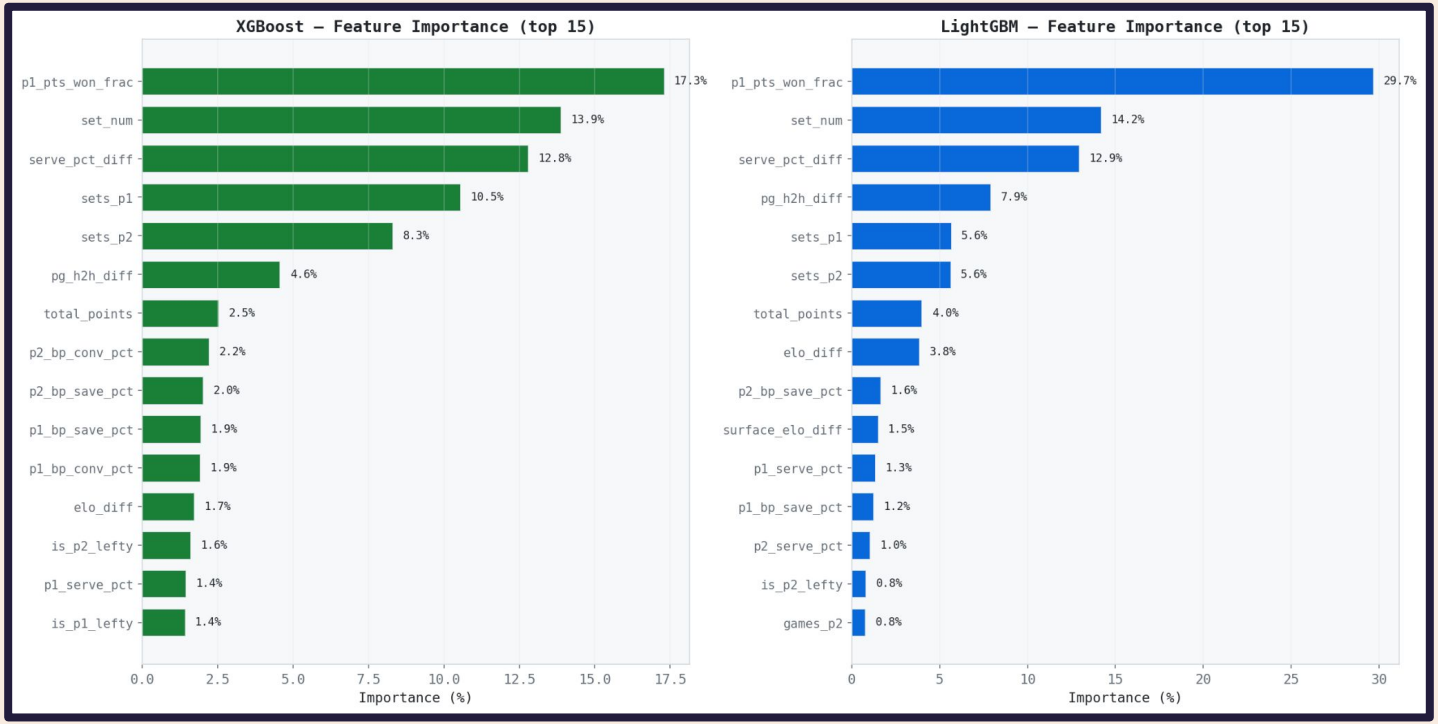
Forthcoming in the
Journal of the American Statistical Association, 96,
June 2001

Even though we show that points are not iid, we also show that the divergence from iid is small. Hence, in many practical applications concerning tennis—such as predicting the winner of the match while the match is in progress—the iid hypothesis will still provide an good approximation, *if we correct for the quality of the players*. This makes our study a useful starting

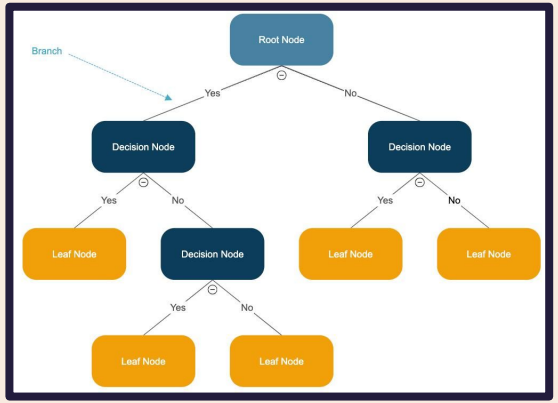
Beginner-Level Model: Logistic Reg

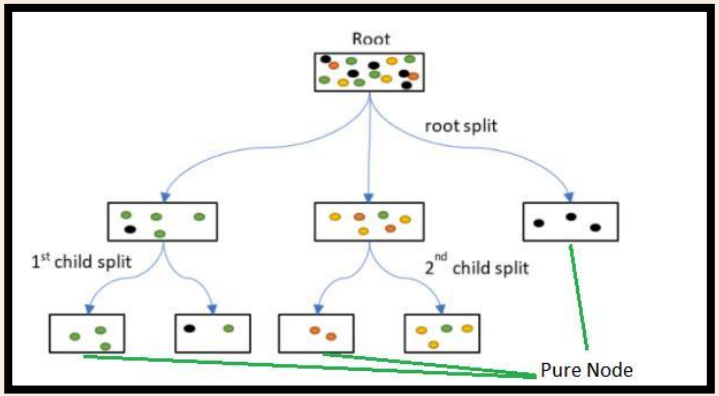


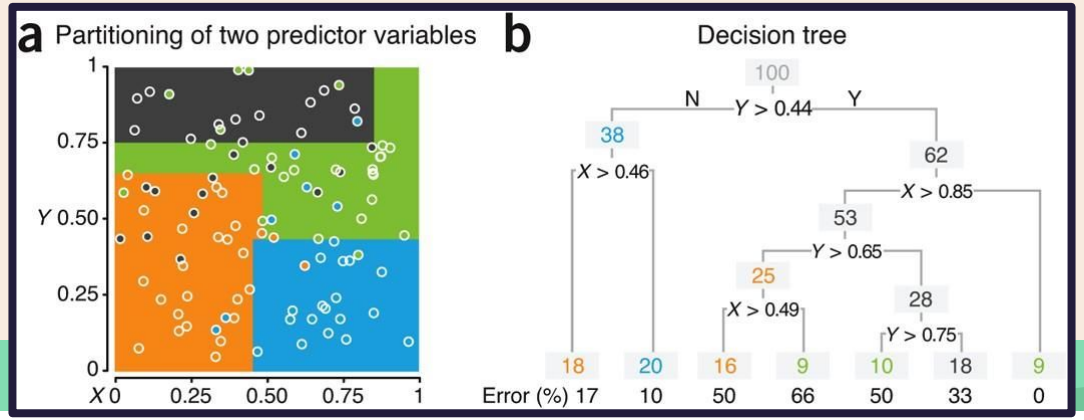
Advanced-Level Model: Boosted Trees



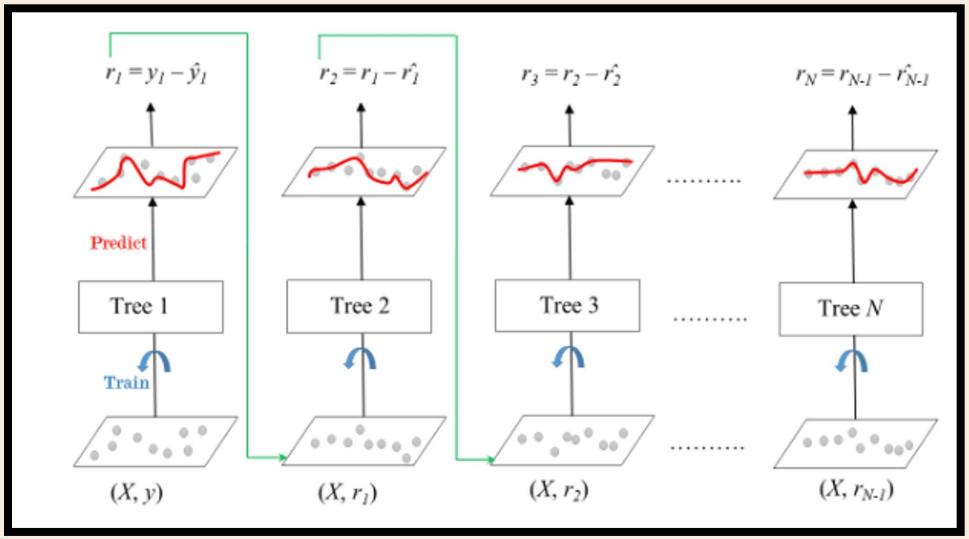
Boosted Trees

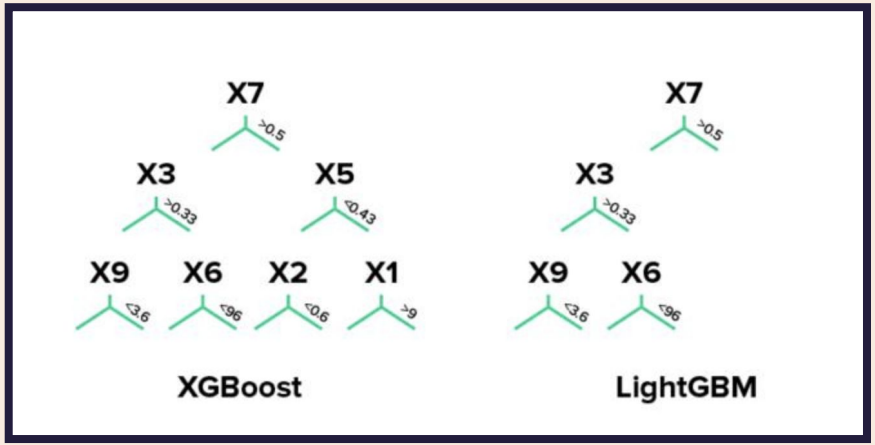




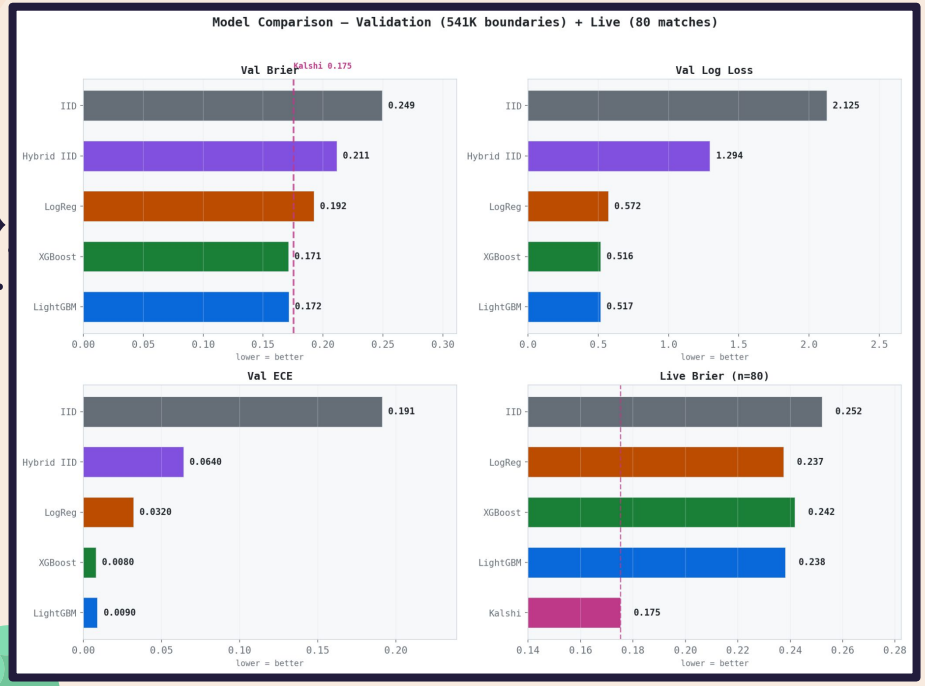


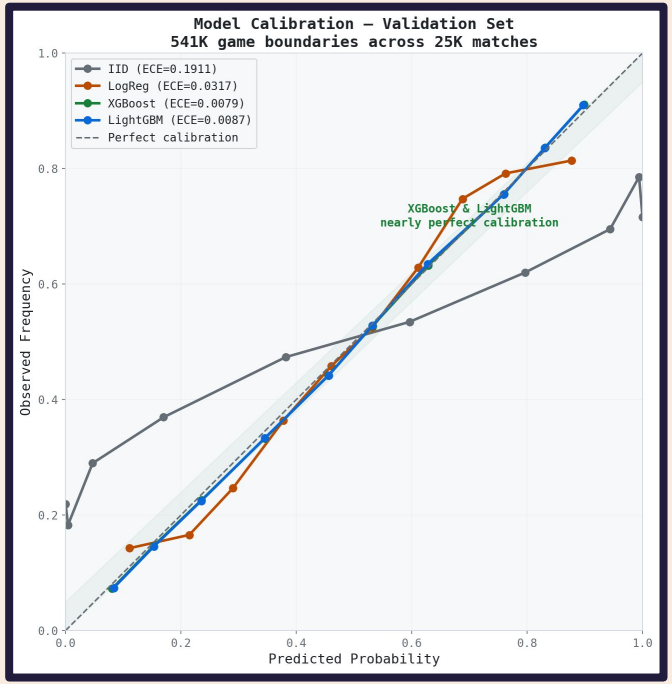
Boosted Trees



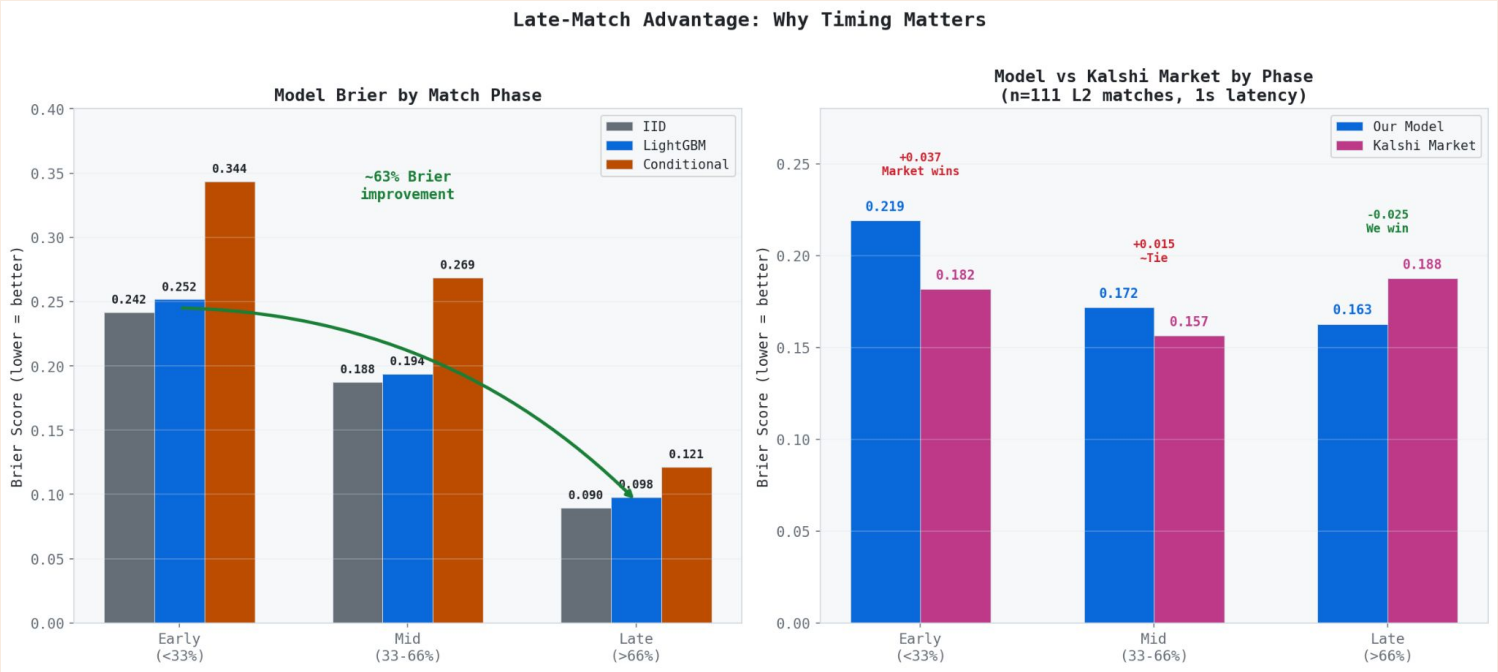


Model Comparison





Model Performance



Conclusion

- Tennis Markets are well calibrated.
- Pre-match odds are difficult to beat
- Matching in-game odds is viable with better data
- Data is the bottleneck, hard to find edge





Thank You