**MONOGRAPHIE**

## REGRESSION

**NOTES DE**: *Machine Learning for the Prediction of Professional Tennis Matches*, M.Sipko (64p)

Mémoire, exactement même thème que moi. Used models : Logistic regression+ Artificial Neural Network

**Part 1: Background:** General info on the game of tennis, tennis betting, statistical models (markov chains for representation of matches), machine learning models

**Part 2: Feature extraction**

\*\*Few different techniques to make certain datapoints more relevant for a given prediction (matches played more recently, on same surface, against same opponent)

- **Historical average**: To have one single feature for something like “ATP rank” per match, the author uses difference in the feature for both players. For historical average stats, difference of average over all matches/ matches against common opponents.

-**Time Discounting:** To account for the fact that recent matches are more relevant to players form than old ones, one can use a formula like Weight= min( f,f^t) with f between 0 and 1 and t= time since match was played  
Time discounting is a hyperparameter that needs to be adjusted

**Surface weighing**: Same as with time, matches played on same surface will be more relevant in predicting a new match: either we only take into account matches played on that surface (but significant data loss for grass for example): author calculates correlation in performance between surfaces for all players to know how to weigh a certain example match towards the one we are predicting

**Uncertainty**: Defined as inverted sum of all the weights. Tends towards 0 the more example matches there are -> lower the more common opponents there are.   
Uncertainty for surface and time weighing: take the sums of all weights of all matches for both players S1 and S2 and compute 1/S1\*S2 to have an indication of uncertainty (see if not enough matches)

**\*\*New feature construction:** using existing features to combine into something simpler/ more useful (combining two stats like serve win%= 1st serve+ 2nd serve win%, serve advantage= serve player 1- serve player )

-modelling fatigue using #games in the previous days, modelling injury using dummy whether player retired from previous match

**Part 3: Data Cleaning**

* Deleting invalid datapoints/ nulls, deleting extreme values due to errors/ due to sample for a player being too small -> certain stats not significant
* Looks at distribution tables for features; use SCALING (centrer et REDUIRE variables)

**Part 4: Logistic Regression**

P 26: separation of dataset in training set, validation set and test set, chronologically in this order: matches are distributed evenly across the year (always the same tourneys at the same times). Crossvalidation not necessary because lot of computation but dataset big enough to avoid overfitting

-evaluation metric for which subset of features selected by strategies below is the best: ROI of betting strat using Kelly criterion, logistic loss  
- choice of feature selection strategy, forward selection/ backward elimination/ recursive selection  
-optimize hyperparameters: grid search is exhaustive search of optimal position of every hyperparameter for global max in accuracy, but demands super computer power: we do greedy heuristic search, fix all other hyperparameters while moving one and finding best value. Not ideal because hyperparameters not independent so might result in local maximum but approach was satisfactory.

**Part 5: Higher Order Models**

-discussion of Bias (underfitting) and variance (overfitting): variance is the sensibility of the model to small variation in the data

-logistic regression with interaction features (produit des variables comme en econo), doenst work cause too many variables (features) -> artificial neural network

-however: training ANN can take up to 10 minutes depending on hyperparameters; discussion of model optimization.

**(Part 6: Implementation overview)** -> explain dataflow+ what libraries/ programming techniques were uses

**Part 7: Evaluation**

-Using betting markets to evaluate short-term+ long-term profitability of the models, compared to a benchmark

**Models used:** Neural network and logistic regression

**Data used:** Jeff Sackmann

**Features:** servewinpct, completeness (serve\*return), serve advantage (difference of differences in serve1 vs return2 pct), fatigue (based on games played in the last 3 days \* discount factor), retirement in last match dummy (not relevant for us), head to head difference

**Results:**

**Discussion:**

**Deployment:** 3 betting strategies, only for betting on the winner: 1. Betting on the predicted winner every match, 2. Betting if model probabilities better than bookmaker probabilities, 3. Kelly criterion: betting a fraction of a standard bet according to the estimated probabilities

**NOTES DE: Machine Learning for Professional Tennis Match Prediction and Betting Andre Cornman, Grant Spellman, Daniel Wright (2017)**

**Models used :** Logistic regression, SVM, random forest, neural network

**Data used:** Sackmann and tennis-data.co.uk (for betting odds)

**Features:** All features were differences between player 1 and player 2: features were: rank, rankpoints, age, height, avg ace/ break point/ double fault (over 5,10, or 20 matches), head to head wins overall + H2H wins specific to surface

**Results:** Highest accuracy: +- 69,6% with random forest hyperparameter adjusting

**Discussion:** Performance of different models was discussed using different measures:  
For every model: accuracy on training set+ 5 fold cross validation accuracy  
-Random forest: feature importance, calibration, error analysis with confusion matrix for higher ranked vs lower ranked aka favorite vs non favorite -> find that model is much more accurate when predicting favorite to win

**Deployment:** Betting market: strategy of betting 1 dollar on either player or not betting at all, 3,3% return per match

**NOTES DE: Neural Networks and Betting Strategies for Tennis ,**Vincenzo Candila 1,\*,† and Lucio Palazzo (2020)

**Models used:** ANN compared to 5 other studies that are point-based/ regression-based/ paired comparison

**Data used:** tennis-data.co.uk and Sackmann, on 6 different time window: from 2005-2013 up to 2005-2018, comparison of results

**Features:** 31 features, again taken as differences between player 1 and player 2, including all sackmann match stat averages + fatigue, completeness (returnpct\*svpct), H2H + some dummies like both lefties/ both righties/ has been in top 10 before -> largest set of features I came across

**Results:** Importance of each variable determined as summed product of connection weights (procedure from Olden et al): head to head, won point return frequency, completeness of player and advantage on service most important

**Discussion:** past models are difficult to compare because the use data from different time frames + different evaluation methods  
-evaluation of this model not through accuracy but only through Diebold Mariano test comparing performance of different models: we don’t get its own accuracy but only know its more accurate than all the other models compared to

**Deployment**

4 different betting strategies, sometimes betting on favorite and sometimes on the underdog, different amounts and with different requirements for placing a bet( going from betting on all matches to only betting if favorite has more than 95% chance to win). The ANN outperforms the only model that has competing accuracy (the lisi and zanella model) but not sure if this betting strategy was very sound, can try to take the best of it and use Kelly criterion.

**NOTES DE: Tennis betting: can statistics beat bookmakers? Francesco Lisi*∗*and Germano Zanella (2017)**

Should we take the absolute difference in rankings instead of Player1-Player2?

This study really dives into each feature and how it should be constructed, how to make the most of it: take the difference, look at distribution of the feature in the sample

**Models used:** Logistic regression

**Data used:** 1081 matches from 2012 ATP

**Features:** ATP points (absolute difference), ATP ranks (to make it less correlated to ATP pt difference, authors make intervals based on a clustering of the sample: these intervals must represent zones with some homogeneity in technical differences among players), ages (different combinations of ages, age difference, age and square age), surface( a dummy for preferred surface between “fast” and “slow” was created based on historical win percentage but was not significant; ) home factor (dummy for tourney in home country or not -> was found to be significant), bookmaker odds (to account for external factors like physical+ psychological conditions, the particular time frame, authors thought this info might be included in bookmaker odds because they must pay attention to this stuff: they add as a feature the odds for the favorite but only if these odds are greater than 2, so the feature only comes in to play when bookmakers give the favorite less than 46% chance to win which means there must be something special: 12% of data for which feature non null)

**Results:** Model with every one of the features added one by one, AIC criterion, pseudo Rsquared and Brier Score. 75,8% accuracy on sample dataset vs 77.5% from bookmakers. On test set of 501 2013 Grand Slam matches: 77,2% vs 78% from bookmakers

**Discussion:  
-**Comparing model with and without the difference of rank interval ( integer between 0 and 4) by using Akaike criterion shows feature to be relevant  
-Age did not seem relevant using multiple different combinations according to the Akaike criterion, except when only adding the age of the favorite, which had a negative coefficient in the regression  
-Comparing model probabilities to bookmaker probabilities derived from odds shows that the distributions are similar and yield similar results on the winner prediction on the sample

**Deployment:** betting market; authors emphasize that only accurate winner prediction is not sufficient to make a profit on betting market, you need to explore strategies. (they cite Mchale and Morton)  
Strategy: bet only if there is a “clear” favorite (more than 55%) and model and bookmakers agree on him; odds of favorite between 1.30 and 1.70 and odds of underdog between 2.40 and 4.30( in order to have high enough return for favorite/ limit risk for underdog) ; ratio model odds/ bookmaker odds must be in between two treshholds : authors want some difference as indicator that a bet might be profitable but they suppose thatwhen the difference is too large bookmakers might take into account a factor that the model overlooks

* The thresholds for odds and odds ratio are calibrated to maximise in-sample return (this was done empirically by trying out different thresholds). Solutions where less than 10% of matches are bet on were neglected.

Strategy was compared to always betting on the favorite and betting randomly and obv performed much better. About 16% ROI on the Grand Slam

**NOTES DE**: *Searching for the GOAT of tennis win prediction*, Stephanie Ann Kovalchik (12p)

-Paper qui teste 11 modèles de prédiction pour leur qualité de prédiction sur 2400 matches de la saison 2014, and using bookmaker predictions as a benchmark

« A literature review was conducted to identify published

models for forecasting wins in tennis. The first stage of

the search strategy queried Google Scholar for articles

containing “tennis” and at least one word containing

“predict” or “forecast.” Hits were reviewed for relevance.

Articles were considered relevant if they mentioned a

strategy for match prediction in the abstract. In the second

stage of the review process, the citations of the relevant

articles were also reviewed. » -utiliser cette tech pour trouver des articles

-3 type of models : regression (mostly probit family), point-based (probability of winning a point extrapolated to whole match using iid assumption), paired comparison; analysis of predictors used by every model. Regression tend to use player rankings, point-based extrapolate probability of winning 1 point to the whole match using Markov chain

-“Four properties of model performance were evaluated:prediction accuracy, calibration, log-loss, and discrimination.” -> utilizer pour évaluer mon modèle

-“This suggests that accounting for recency of play and the quality of opponents are critical elements in predicting the outcome of matches at the elite level with greater accuracy.” -> only Elo model got close to bookmaker model results by having these characteristics; validates recency weighing and common opponents

## POINT-BASED

**NOTES DE: KLAASSEN, F. J. G. M. & MAGNUS, J. R. (2003) Forecasting the winner of a tennis match.**

**Models used:** point-based to predict prob before the start and during the match + logit model based on rank

**Data used:** point probabilities, whether BO3 or BO5, what the current score is and who is serving / logit mode: Wimbledon match dta 1992-1995,495 matches

**Results:** No evaluation of the model

**Discussion:** authors show that the difference in point-winning probs is more important than the sum of probs

**Deployment:** None

**NOTES DE**: *A common-opponent stochastic model for predicting the outcome of professional tennis matches,* Knottebelt & Madurska

This paper uses a point-based, common opponent model to predict WTA and ATP matches; models are validated using betting odds and ROI over a certain number of matches.

-The differential in points won on serve is a significant predictor of match winner ( Fig.1)

“One might note that there can be quite some variation between the probabilities of winning resulting from different

common opponents. This suggests that predictions made with only a small number of common opponents should be

treated with caution.”

**Overall goal/specificity**: Markov model but the data used to compute point-winning probability is based only on matches played against common opponents so that there is no bias due to different levels of opposition when computing the probabilities for both players

**Data used**: Inspired by Barnett and Clarke and O’malley.: Historical data to compute serve probabilities. An insight arising from O’malley is that the match-winning probability mainly depends on the **difference** in serve-winning probabilities between the two players ( for serve-winning probabilities ranging from 0.2 to 0.8 as can be observed in professional tennis) (this is almost a direct quote).  
The data is limited to recent matches (previous 12 months) so that it would reflect the recent form of the players. Moreover, only matches were both players have sufficient matches against common opponents are taking into account.

**Dependent variable/ features**: The dependent variable is the probability of a player to win a point on his serve. The probability is computed using a number of past matches from the players and correcting it with

**Validation**: Application to the betting market, betting if bookmaker odds better than calculated odds -> mentioning of the Kelly criterion. When using only percentages from same surface matches for computation, positive ROI; when using all surfaces, negative ROI

**Discussion**:

-Other possible data sources to : Hawkeye or video + audio processing of tennis matches (G. Hunter, A. Shihab, K. Zienowicz, Modelling tennis rallies using information from both audio and video signals, in: Proceedings of the IMA

International Conference on Mathematics in Sport, 2007, pp. 103–108.)

**NOTES DE:** Probability of winning at tennis I. Theory and Data, Newton n Keller (2005)

**Overall goal/ specificity:** Formula for game, set and matchwinning are computed from serve point win probability+ proof that these probabilities are independent from who serves first+validation on US + Wimbledon data and discussion of the assumption that points are i.i.d

**Data :** Us Open and Wimbledon 2002

**Model type/ Dependent variable/features :** Dependent variable is serve percentage and mathematical model is made to this to match win percentage; big part of study focuses on this mathematical relation as well as for sets/ tiebreakers and how the formula changes given number of sets, less time spent on trying to determine serve percentages as precisely as possible (small data sample). Also computation of each player’s probability of winning a 128man tournament and apply to data for US open and wimbledon

**Validation:**

**Discussion:** Table 3, p 14/29: shows that win probabilities mostly depend on the difference between serve percentages, validates using difference in sp instead of two different features -> we can check this by comparing results

Non- i.i.d. effect : Klaassen and Magnus find first game is the most difficult one to break, Jackson and Mosurski show “hot-hand” (one point calls the next) and “back to the wall” (better when coming from behind) effects which support points not being i.i.d. , would be more precise if used varying game/ point winning probability, for this we need more granular data on specific points / games -> point-by-point db Jeff Sackmann  
Also from C. Morris “The most important points in tennis” -> some points are more important than others and that also makes win percentages of those points move

**NOTES DE:** *Probability Formulas and Statistical Analysis in Tennis,*  O’Malley

**Overall goals/ specificity:** Creating and evaluating the formula’s linking game, set, match probabilities to serve win probabilities (newton n keller 2005 is a reference), also creating various formula’s for probabilities of winning tie-breaker set or coming back from certain deficit / 2nd part is validating by comparing these formula’s to real matches

**Data :** Validation data 2007 Wimbledon

**Model type/ Dependent variable/features :**

**Validation:**

**-** Looking at Wimbledon 2007 data, observed and predicted proportion of games won were strongly correlated.

- However authors find statistically significant difference in proportion of break points saved by server vs overall points won ( 55% break points vs 67% other) which casts doubt on the i.i.d assumption-> (other study to look at: Newton Aslam 2006)

**Discussion:** Formula’s could be interesting for commentators, organizators (reduce variance in length of a match by changing scoring rules) and trainers (evaluate from formula’s if better to focus on improving receiving game or serving game based on respective point-winning probability improvement)

**NOTES DE:** Monte Carlo Tennis : A Stochastic Markov Chain Model, Newton & Aslam 2009

**Overall goals/ specificity:** Create a point based Markov model but point winning probabilities are modelled by taking the data and making a probability density function for “winning on serve” and “winning on return” and considering these percentages as normal variables; mean servepct +- 62%, mean returnpct +- 38%.

**Data :** Men’s ATP circuit matches 2007, 330 players over 59 tournaments

**Model type/ Dependent variable/features :** Dependent variable is % of chance to win the match. The models includes not only the serve win pct but also return win pct and standard deviations (to account for consistency/ lack thereof in serving or receiving, thus being able to make more accurate predictions)

**Validation:** The probability for each player to win on serve is shown to be close to a gaussian; theses values are picked based on the pdf function (adjusted for the return ability of opponent) 30000 times and the associated win prob functions are computed: win probabilities are also modeled into a pdf function that is shown to be gaussian (as predicted in advance by relationship of servewinpct and matchwinpct).

**Discussion:** Authors notice that lack of consistency (high standard deviation on serve) is disadvantageous when being the better server but actually advantageous when lesser server (decrease/ increase chance of winning the match)

-Same as Barnett & Clarke: determine svwpt and rwpt based on the average percentages of the population and the opponent’s ability (in the opposite): pct is adjusted up or down depending on if opponentis better or worse than population average

**NOTES DE**: *a set-by-set analysis method for predicting the outcome of*

*professional singles tennis matches,* Knottebelt & Madurska

This paper takes the model from the paper just above (point-based, Markov Chain) and takes a set-by-set approach, meaning that probability of a given player of winning a point varies from set to set

*An analysis of the game of tennis*, Carter & Crews (1974)

The probabilities of each player winning a game, set and match are calculated through the use of formulas taking the probability of each player winning a point on their serve (pA and pB) as variables.

The point probabilities are also used to determine the duration of a match in this article.

The authors try to simplify the formulas by approximating the probability of winning a point for each player using the average of the probability of winning on their own serve and winning on the opponents serve. (they stress that this is an approximation)

How are the point probabilities estimated? – They are not estimated empirically, the authors create tables for probability of winning games/ sets/ matches and expected duration based on different values of pA and pB. Only one empirical example is used and compared to the predictions given by the formulas.

Furthermore, the authors explore how changes in the rules (number of games required to win a set/ margin required to win a set) affect the game/set/match winning probabilities and expected duration

SUMMARY: This article did give a way of extrapolating the probability of winning a point to that of winning the match, but its application was still rather limited given that the real “difficult” prediction, namely the probability of each player to win a point, was not yet attempted. This is not a surprise given the year of publication of the article (no computers yet). It also relies on the assumption that the probability of each player to win points (on their serve or in general) is constant.

*Optimal strategy in tennis: A simple probabilistic model*, S.L. George (1973)

This article focuses on the fact that players usually have a first “strong” serve and a second “weak” serve that is used if they fault on their first serve. The first serve results in a higher probability of winning the point given that it is good, whereas the second serve has a probability chance of being good. The authors then examine what the optimal serving strategy is depending on what these probabilities are, and conclude that the most common strategy (strong first serve, then weak serve if fault) is not necessarily the most optimal. There is no predictive model created.

1. Modern non ML models

T. Barnett and S. R. Clarke. *Combining player statistics to predict outcomes of tennis matches*. IMA Journal of Management Mathematics, 16:113–120, 2005.

Instead of creating a model for the winner of the match based on the ex post probabilities of each player winning a point, this study uses player statistics to predict the winner of the game. They want to use statistics that are available before the game, contrary to the many studies that exist already that calculate the probability of each player of winning at service to predict the match winner, but therefore use empirical data from the match itself that is only available afterwards. The authors say it is useful to know the estimate the win probabilities (and the duration of the match) before it happens for broadcasters and tournament organizers, who need to schedule accordingly.

They create a specific formula for win probability against a certain player using one player’s serving ability and the other one’s returning ability ( determined as the difference of their average with the population average serve/ return pct)

**DATA**

The authors focus on one match between Roddick and El Aynaoui. The features they use are all serve and return win percentages, based on 70 matches. The statistics are weighed so that recent matches, in particular matches played in the same tournament in which said match was played, are given more importance.

A modified formula is used to create specific serving and returning percentages for the returning and serving percentages of the opponent

* See Klassen n Magnus 2001, exploring i.i.d. assumptions for tennis points

A Markov chain from an earlier study by the same authors is used, with a constant point win probability p: moreover, the “advantage server/ receiver” and “deuce” states are considered to be the same as 40-30 or 30-30.

The study seems a bit weak since the quality of the model in predicting winner and duration is only tested on a single match. Moreover, it seemed that a lot of importance was given to the prediction of length of match rather than only winner.

Are Points in Tennis Independent and Identically Distributed? Evidence From a Dynamic Binary Panel Data Model,2001 KLAASSEN & MAGNUS

This study starts off by rejecting the i.i.d assumption of tennis points, stating that winning a point positively impacts the probability of winning the next point, and that certain “difficult situation” points reduce the servers win probability, in particular for weaker players.

The study is very statistical and does not look to predict the winner of a match, but rather to test the assumption of identical distribution between tennis matches

J. A. O’Malley. *Probability Formulas and Statistical Analysis in Tennis*. Journal of Quantitative

Analysis in Sports, 4(2), 2008.

This paper focuses on modeling tennis matches using novel statistical formulas. The winner of a match is predicted by computing the probability of winning a point and extrapolating it to the entire match using Markov chains. Furthermore, the probability of various scenario’s occurring, like coming back after losing a break point or winning a tiebreaker, are calculated.

The main objective of this study is to create mathematical formulas giving the probabilities of certain scenario’s. The paper also shortly discusses the i.i.d. assumption, and whether probabilities of winning a point are heterogenous, because they vary in certain high-pressure situations. The point-by-point data used to compare the theoretical results to reality came from one Wimbledon tournament.

*A common-opponent stochastic model for predicting the outcome of professional tennis matches* William J. Knottenbelt, Demetris Spanias ∗, Agnieszka M. Madurska

This study cites other studies we have taken notes on, that try to predict the winner of tennis matches, either by using rank or by computing the probability of winning a point per player and extrapolating it to the whole match.

The study emphasizes the fact that the point-win probability method is inherently biased because it uses average point winning probability in its formula, although players have not necessarily faced opponents of the same caliber. The article proposes to base calculations of averages only on common opponents between two players. This technique is applicable because the ATP tour only has a limited number of players playing every year that play multiple matches against each other, so there is a big enough sample to reduce to common opponents.

The O’Malley 2008 model is reused; one of the key findings referred to is the fact that the match winning probability is almost linearly dependent on the difference in winning at serve probabilities for each player. The point-winning probabilities are modified using the common-opponent approach, and a second model is created only using past matches played on the same surface. The model is then tested on the betting market over a 12 month period and yields a positive ROI (using a simplistic constant 1 dollar betting strategy).

*A calibration method with dynamic updates for within-match forecasting of wins in tennis* Stephanie Kovalchik a,b,\*, Machar Reid

The paper assesses the performance of different previous studies, point-based, rank-based an odds-based. It wants to improve the Klaassen and Magnus 2003 model by updating the probabilities during the match.

The ELO rating developed by FiveThirtyEight has been shown to be one of the most accurate (Kovalchik 2016), therefore we will test it on our training data to see what result we get using only this feature.

In essence, this study reuses the Klaassen and Magnus model, but adds a dynamic update component that allows for predictions to evolve at any stage of the match. We will most likely not be using dynamic updating.

*A Bradley-Terry type model for forecasting tennis match results*, Ian McHalea, Alex Mortonb,∗

*Predicting the Outcome of a Tennis Tournament: Based on Both*

*Data and Judgments, Wei Gu, Thomas Saaty (2019)*

This study collects 44 metrics: basic match information, descriptive parameters for each player and performance metrics for each player, plus 11 additional performance measures that were combinations of previous performance measures. To test out if the 20 performance measures are significant for the prediction of the outcome of matches, the author used a Wilcoxon rank-sum test and found that all 20 variables were significant. Then they used logistic regression with these 20 performance variables: they acknowledged that there was a multicollinearity issue but in the ANP method (Analytic Network Process) (-> seems to be a method created by the authors themselves?) this is ok because regression serves to detect correlation between win/loss and every independent variable, “The coefficients of the regression provide references for experts to conduct the pairwise comparison”.  
The study uses “expert judgment”:

Testing Rosen’s Sequential Elimination Tournament Model: Incentives and Player Performance in Professional Tennis, K. Gilsdorf, V. Sukhatme, 2007