Game, Set, Learn: A Machine Learning Approach to Tennis Match Prediction

Nikhil Shenoy, Abhinav Malik, Claire Wang
[1]University of Pennsylvania, [2]School of Engineering and Sciences

Overview

Tennis is one of the most popular sports for the variety of style, strategies against different players and the globe-trotting part of the ATP World Tour. Match prediction attracts not only huge profit but also heated academic interest. In this project, we seek to use machine learning methods to build a model of a match-up between two players based on statistics collected in previous matches, and predict the winner of the match.

Design Goals

- ❑ Do machine learning models perform better than rules of thumb?
- ❑ If so, which is the best model?
- ❑ What are the most important factors that predict match outcome?
- ❑ Create simple baselines as the rules of thumb for match prediction
- ❑ Extract variety of features representative of player and match statistics
- ❑ Construct machine learning models based on extracted features
- ❑ Conduct ablation study to discover most relevant features to match result

Figure 1. Nadal-Federer Australian Open Final

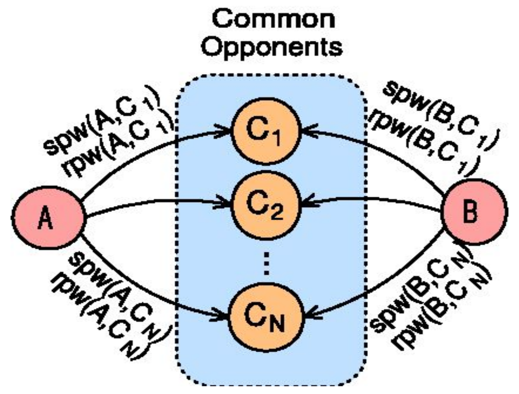


Figure 2. Common Opponent Model [3]

Data Preprocessing

- ❑ Data set: ATP World Tour data compiled by Jeff Sackman, limited to 2000-2017,
  - ❑ training: 2000-2008
  - ❑ validation: 2009-2013
  - ❑ testing: 2014-2017
- ❑ Features are converted using the Common Opponent Model (Figure 2).
  - ❑ Find set of common opponents for the two players
  - ❑ For each player in the matchup, calculate mean of the statistic (ex.: aces) against common opponents.
  - ❑ Difference of means of the statistic generates the feature value.
- ❑ Derived features are functions of raw features. For example, "completeness" rewards players who are consistently good in service points won *and* break points saved.
- ❑ Transform into binary classification problem: Did Player 1 win the match?
- ❑ Symmetry: the result of the match should be the same regardless of how the assignment to Player 1 and Player 2 is made.

Experiments

- ❑ Random subset of data swaps statistics for Player 1 and Player 2
- ❑ Train three models: SVM, Neural Net, Logistic Regression
- ❑ Compare model performance against other models
- ❑ Compare model performance against baseline
- ❑ Determine top features that influence outcome

Machine Learning Models

SVM

- ❑ Standard linear model. Sipko [2] expects it to do well
- ❑ Can account for high dimensional data with kernels
- ❑ Kernels: linear, polynomial (deg. 3), rbf. Linear was the best, with C=0.5

Multi-Layer Perceptron

- ❑ Feed-forward network with three layers; its non-linear activation layer can distinguish data that is not linearly separable
- ❑ Makes a good classifier when response variable is categorical

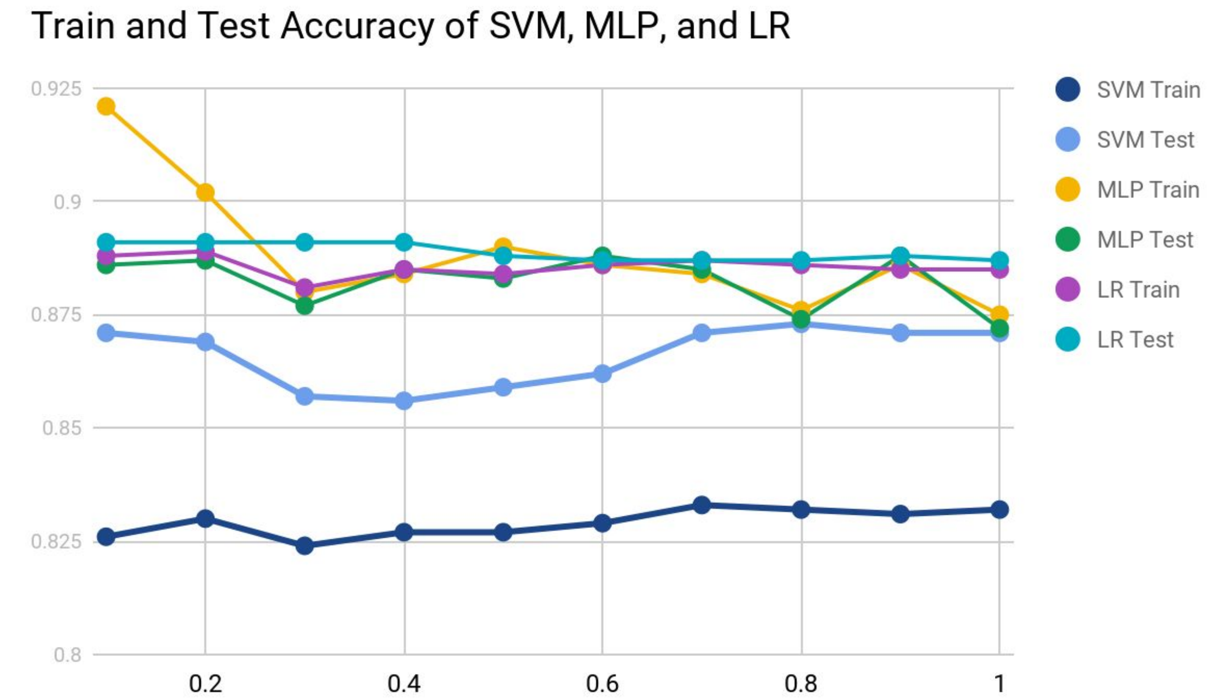
Logistic Regression

- ❑ Linear algorithm. Resistant to over fitting
- ❑ Preferred classifier for binary data and multidimensional feature space

Table 1. Performance of baselines vs. models

Accuracy (%)	Train	Validation	Test
SVM	86.409	88.048	89.487
MLP	86.252	87.608	89.132
Logistic Regression	86.202	87.544	89.001
Rank Baseline	64.079	66.696	66.833
H2H Baseline	64.784	62.070	57.787
Seed Baseline	58.863	58.949	59.169
Age Baseline	47.485	48.812	51.428

Graph 1. Comparison of Classifier Performance



Ablation Study

- ❑ We want to assess which are the most important factors in predicting match outcome
- ❑ Procedure: Test with all features, then test with one feature removed. Compute difference in accuracy
- ❑ Expectation: Player statistics (1stWon, aces, bpSaved) and Completeness should dominate. More even distribution among features
- ❑ Result: Head-to-Head is dominant, then set score and surface. Other features either don't contribute much

Graph 2. Feature comparison with ablation study

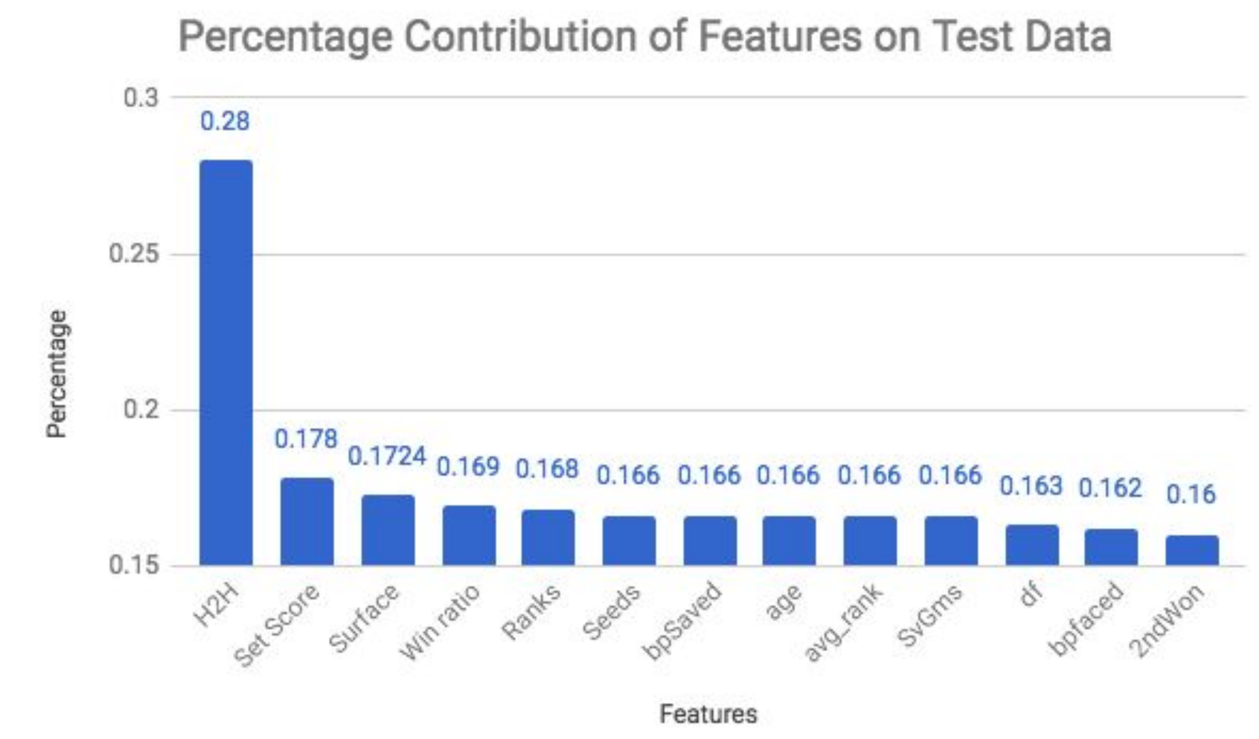


Table 2. Top 5 Contributing Features

Feature	Validation	Test
H2H	0.248	0.280
Set Score	0.134	0.178
Surface	0.1299	0.1724
Win ratio	0.125	0.169
Ranks	0.127	0.168

Conclusion and Future Work

We sought to answer three questions set out in the beginning of the project. The three machine learning models generally performed better than simple baselines by over 20% of accuracy. Even though there is no significant difference in models' performance, SVM is the best performing one among all three. The ablations study also shows that head-to-head, set score, and surface are among the most relevant features to match result.

One remaining issue that can be extended to future work is to evaluate how well the models perform in a full tournament: the importance of certain tournaments over others, the probability of match fixing, and how likely one player loses to another as a tactic to game the system.

Acknowledgements

- ❑ We would like to thank Jeff Sackman, who meticulously gathered and organized the data we used for this project.
- ❑ We would also like to thank Professor Roth and the course staff for their feedback and guidance in the project.

References

1. 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. J. Am. Stat. Assoc., 96, 500–509
2. MIchael Sipko, Machine Learning for the Prediction of Professional Tennis Matches, 2015
3. W.J. Knottenbelt, D. Spanias, and A.M. Madurska. A common-opponent stochastic model for predicting the outcome of professional tennis matches. Computers and Mathematics with Applications, 64:3820-3827, 2012
4. S. Ma, C. Liu, and Y. Tan. Winning matches in Grand Slam men's singles: an analysis of player performance-related variables from 1991 to 2018. Journal of sports sciences, 31(11):1147-55, 2013
5. A. Somboonphokkaphan, S. Phimotares, and C. Lursinsap. Tennis Winner Prediction based on Time-Series History with Neural Modeling. IMECS 2009: International Multi-Conference of Engineers and Computer Scientists, Vols I and II: 127-132, 200
6. A. Madurska. A Set-by-Set Analysis Method for Predicting the Outcome of Professional Singles Tennis Matches. Imperial College London. June 2012