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

Nikhil Shenoy Claire Wang Abhinav Malik

***Abstract*—Tennis is an international sport that people enjoy watching for the variety of style, strategies against different players when playing on different surfaces, and the globe-trotting part of the ATP World Tour. With an increasingly talented group of players at the top of the game, a strong incentive exists, in the betting and purely academic arenas, to predict the outcome of matches. In this project, we seek to use machine learning methods to build a model of a match-up between two players based on statistics about their playing style, and predict the winner of the match. By doing this repeatedly on a tournament draw, we aim to construct a model of which statistics are most important in determining the winner of a match. We limit the scope of our model to the ATP World Tour, which is top level of the mens game.**

I. Introduction

Models for match prediction are useful for a variety of parties. Coaches can use the information to determine where their player is weak and improve those aspects of their game. In the betting world, predictions can be used to place bets on favorite players for big cash payoffs. Even the players themselves can use such a model to assess a matchup with an upcoming opponent, and what their chance is to get to a certain round in the draw. All of these stakeholders can also use the model to analyze what went wrong in matches that were considered upsets, such as Roger Federer, a seven-time Wimbledon champion, losing to Sergiy Stakhovsky in the second round of Wimbledon 2013. Our goal in this project is to create a model that can account for all of these situations and provide useful information to all interested parties.

II. Previous Work

A plethora of work has been done on tournament predictions, and tennis predictions in specific. Due to the nature of the hierarchical scoring system in tennis (4 points in a game, 6 games in a set, win or sets, with variations), structured models are easier to create as opposed to other sports. Additionally, papers have been published about specific properties of the game that we can leverage when building our model. In the first paper we looked at, written by Klaassen and Magnus [2], the authors analyze whether individual points in tennis are independently and identically distributed. They concluded that points are neither independent of each other nor identically distributed;however, they also concluded that deviations are sufficiently small that the IID hypothesis will still provide a good approximation. This fact is extremely useful as it allows us to build models without needing to continuously look at the previous state of a match. Sipko [1] has also done extensive analysis of similar data, and provides a reference for feature extraction and results from models he has tried. One method Sipko uses for feature extraction when comparing two players is to consider a set of common opponents between two players. For a specific raw attribute, such as the percentage of points won when returning serve, and a matchup between Player 1 and Player 2, the value of the attribute is measured between Player 1 and every player in the set of common opponents. The process is repeated for Player 2. Then, the vectors for each player are averaged to generate two scores. These scores can then be used to compare the performance of Player 1 versus Player 2 for this specific attribute. By using this method as one of our feature extractors, we can build much more nuanced models.

III. Methodology

Our experiments seek to answer three questions:

1) Are machine learning models better at predicting match results than a casual watcher using basic rules of thumb?
2) Which machine learning model is the best for prediction?
3) Which are the most important features to consider when predicting a match?

In the first question, we assess whether machine learning models will actually do better than a casual fan using his own instincts. It is very common for a fan to say “this player is ranked higher than his opponent, so I think he will win”, and use no other data for his prediction. We turn this statements into baselines, or rules of thumb, to simulate the performance of a “fan classifier”. We compare the results against machine learning models to determine whether advanced analysis of tennis data is really useful.

We expect that machine learning methods will do better than a simple baseline, so the next question we consider is which model to choose. Many classifiers have been created over the years, and we compare the support vector machine, neural networks, and logistic regression to try and find the best model that suits this particular problem. We will compare the accuracy of each model on a test data set to determine the best model.

Finally, once we have the best model, we attempt to discover the most important features of a match. Such information is useful to the stakeholders because it simplifies the problem of predictions into a handful of statistics that they can use to improve their game or maximize their winnings. To do this,

we perform an ablation study to analyze the effect of removing certain features from the classification.

We decided to use tennis tournament data from a repository created by Jeff Sackmann [8], which contains all data in ATP matches from 1968 to 2017. In each match, a number of statistics such as the number of aces, double faults, and break points saved, are detailed for each player. These statistics provide ample information about the match that our classifiers can train on.

IV. Data Preprocessing

We used a variety of techniques to prepare our data for input into the algorithm. First we checked for the unique values for each feature in the raw data, and removed examples where the value would be unusable in our feature extraction phase. An example of this would be finding "nan" values in the numerical features; performing any operations on data with this value would result in a value that our classifiers would not be able to process. We considered techniques on how best to replace the value, but we eventually decided that the schemes we had come up with could still alter the outcome of the classification. Thus, we decided to drop the inconsistent examples entirely.

We also restricted the data that we considered to the years 2000-2017. Sackmann's repository contained a wealth of tennis data going back to the inception of the ATP World Tour in 1968, but we decided not to use all of the data because of how the game has evolved since then. There are a number of issues that would have cropped up if we had used the full data set:

- **Too much data:** While we would have liked to use the full set, we did not have the computational power to do our feature extraction in a reasonable amount of time. Our feature extraction methods (as will be discussed in the next section) tend to be $O(n^2)$ operations, and they would have become prohibitively expensive when working with $O(100000)$ examples.
- **Difference in technology:** The racquets, shoes, court surface, and ball, among other items, have changed drastically over the years, allowing some players to excel.For example, racquets have allowed players in the last 20 years to hit harder and with more spin, which was not possible before. To avoid this, we limit the years considered so that the issue of varied sports technology is relatively constant.
- **Difference in playing style:** From the year 1968 to around the middle of the 1990s, the playing style for most tennis players was "serve-and-volley". After that period of time, players began to finish more points at the baseline, resulting in the playing style we have today. This difference prioritizes certain statistics, like serves and volleys, more than others such as statistics. We eliminate this particular difference by choosing an era with consistent playing style.

Another decision was made to remove all the Davis Cup matches from the data set. The Davis Cup format requires players to be on a team and represent their country, while countries face each other in a tournament for the final prize. This format has different rules than normal singles plays, as each team has a captain and can decide who plays at what spot in their lineup against another country. Since these matches can be arranged differently than a standard tournament draw, we consider it as noise and remove all non-ATP matches.

Finally, we turned the original data set into a binary classification problem. The raw data set had the two players' statistics and the name of the winner, which isn't an easy format to work with. Instead, we labeled the players "Player 1" and "Player 2", and then added a label column such that the value would be 1 if "Player 1" won the match and zero otherwise. By doing this, the question our classifiers seek to answer is "what are the weights of the features that will help Player 1 win the match?", which is an easier problem to solve.

V. Feature Extraction

We used four different methods to extract features for our problem. After all the features are extracted, we create a feature vector with a length equal to the number of features, and have the aforementioned Player 1 win/loss value as our label. We provide a summary of our feature extraction methods here.

A. Symmetry of Features

In the original data, we are provided with two values per statistic in every example. For example, the "aces" statistic is recorded for both Player 1 and Player 2. This representation provides us with specific information for each player, but it also has a drawback; if we exchanged the labels for Player 1 and Player 2, then our classifier could assign different weights to the individual feature and predict a different outcome. To avoid this, we create a single feature with the raw statistics as shown in Equation 1.

$$\text{Feature}_i = \text{Raw}_{1,i} - \text{Raw}_{2,i} \tag{1}$$

where i is the example index, RAW is the raw feature from the original data set, and FEATURE is our newly generated feature. Previous work by Clark and Dyte [3] have shown that this representation is effective, so we transform all of our raw features using this method.

B. Common Opponent Model

We also use a method called the Common Opponent Model, which was originally proposed by Knottenbelt [4]. In this model, we consider the set of common opponents that both Player 1 and Player 2 have played against. For each of Player 1 and 2, we find the mean of the statistic against each player in the set of common opponents. Finally, we create a difference variable using the two means.

The idea behind this model is that comparing statistics between two players becomes more meaningful when judged against a standard baseline. We get a better idea of the true value of the statistic for each Player since we take into account his performance against a variety of opponents. All of our numerical features were calculated in this way.

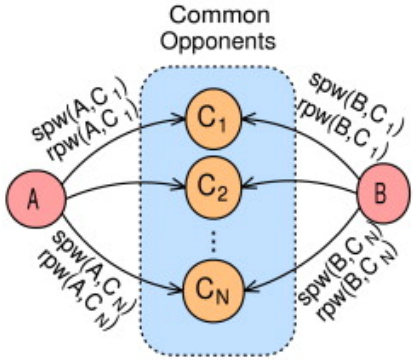

Fig. 1. The Common Opponent Model

C. Derived Features

We created derived features as a function of other features that we have. One noteable example is the Completeness feature. In debates about who the best player of all time is, commentators require that a player have a good serve, good groundstrokes, good returns, and good movement, among other factors. Someone who has all of these factors is considered a "complete" player. We try to simulate completeness in Equation 2 by taking the product of the first serves won, the second serves won, and the fraction of break points won. Unfortunately, our data set did not have information about groundstrokes and court movement, so we limited the definition to just three component features.

$$\text{Completeness}_i = \text{1stWon}_i * \text{2ndWon}_i * \frac{\text{bpSaved}_i}{\text{bpFaced}_i} \quad (2)$$

By multiplying the component features, we generate a derived feature that requires a player to be competent in all areas for the feature to be significant. But if any is lacking, then the feature value will decrease and potentially not have as much impact on the final prediction. We feel that such features provide additional information about the problem based on domain-specific knowledge and can bolster the accuracy of our models.

D. Categorical Features

While the idea of categorical variables is not new, we use it to include information that was not used in previous work. These works always included statistics about a player's performance, but never included information about the surface of the court. Tennis is played on four types of surfaces; hard, grass, clay, and carpet. The tactics used to win are different on each surface, and some players' styles are more suited to one surface than another. To account for the court surface in our feature vector, we use a 1-Hot encoding to indicate which type of surface a particular match is played on.

VI. Experiments

Our experiments seek to answer the three questions we set in the beginning of the project:

1) Are machine learning models better at predicting match results than a casual watcher using basic rules of thumb?
2) Which machine learning model is the best for prediction?
3) Which are the most important features to consider when predicting a match?

We briefly describe the models that we use, and then detail the results of our experiments.

A. Models

1) Support Vector Machines: We chose to use a Support Vector Machine (SVM) for classification for two main reasons; first, no previous published work has utilized the SVM for this problem, and we saw an opportunity to examine how effective this classifier is. Also, the SVM does not encounter local minima during the calculation, which provides an advantage over classifiers like neural networks. We used the Soft SVM formulation in Equation 3 to build the classifier, where w is the weight vector, y_i is the label of example i, x_i is the i-th example, and C is a hyperparameter that scales the penalty of performing misclassifications.

$$\min_w \frac{1}{2} w^T w + C \sum_i \max(0, 1 - y_i w^T x_i) \quad (3)$$

To find the optimal hyperparameters for the SVM, we used a grid search over three kernels (linear, polynomial, and the radial basis function) and the scaling factor C. We trained on our training set and tested on the validation set, giving the results in Table I.

Kernel	C	Accuracy
Linear	0.25	.867
Linear	0.5	.871
Linear	1.0	.870
Linear	10	.869
Polynomial	0.25	.859
Polynomial	0.5	.860
Polynomial	1.0	.861
Polynomial	10	.864
RBF	0.25	.864
RBF	0.5	.867
RBF	1.0	.864
RBF	10	.861

TABLE I
Results from Tuning SVM Hyperparameters

From this grid search experiment, we found that the Linear Kernel with $C = 0.5$ was the optimal pair of hyperparameters for our model.

B. Ablation Study

To find out what the most important features in determining match outcome are, we perform an ablation study. In such a study, we first get the accuracy of the classifier with all features included. Then, we remove features one at a time, to examine the contribution each feature has to the final accuracy of the model. We performed this technique on all of our models (ACTUALLY NEED TO DO THIS FOR MLP AND LR), with

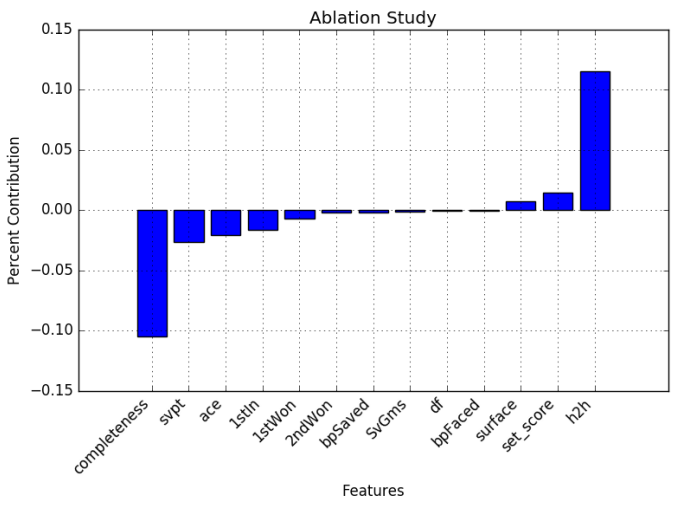

Fig. 2. Ablation Study with SVM

results from the ablation study on our SVM shown in Figure VI-B.

The results from this study are interesting because of the signifiance placed on certain features. We observed that the Head-to-Head feature had the most significant contribution to the test accuracy, boosting it by about 12%. The implication of this is that the history of the pair's past matches outweigh any of the statistics gathered during the current match. The next two most significant features were the set score and the court surface. Intuitively, the court surface must have significance due to the way the ball bounces on different material. But the more interesting trend is that the remainder of the features that we used either did not contribute to the test accuracy or actually hindered the result. From this, a conclusion is that our model provides no information about which match statistic a player should prioritize in order to win, implying that so long as the player can win, he can play whichever way he likes. The most notable of the remaining features is completeness, which is one of the derived features that we created to simulate a player's ability in multiple aspects of the game. The implication from the study is that a player training to be good at each type of shot is actually detrimental to his success, allowing one to conclude that that a player should only focus on improving certain shots, perhaps those he has a natural talent for, to increase his chances of winning. We had expected the ablation study to show that each of the match statistics contributed some positive percentage to the test accuracy an dsuggest what areas a player should focus on improving. The final results, however, was that the most important feature in determining match outcome is the past Head-to-Head history and the court surface for the match.

VII. Conclusion

The conclusion goes here.

Acknowledgment

The authors would like to thank...

References

[1] M. Sipko, "Machine learning for the prediction of professional tennis matches," Master's thesis, Imperial College London, June 2015.

[2] J. G. M. Klaassen and J. R. Magnus, "Are points in tennis independent and identically distributed? evidence from a dynamic binary panel data model," *Journal of the American Statistical Association*, vol. 96, pp. 500–509, June 2001.

[3] S. Clarke and D. Dyte, "Using official ratings to simulate major tennis tournaments," *International Transactions in operational Research*, vol. 7, pp. 585–594, 2000.

[4] 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*, vol. 64, pp. 3820–3827, 2012.

[5] A. M. Madurska, "A set-by-set analyssis method for predicting the outcome of professional singles tennis matches," Master's thesis, Imperial College London, June 2012.

[6] A. Somboonphokkaphan, S. Phimoltares, and C. Lursinsap, "Tennis winner prediction based on time-series history with neural modeling," *International MultiConference of Engineers and Computer Scientists*, vol. 1, March 2009.

[7] 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 2008," *Journal of Sports Sciences*, vol. 31, pp. 1147–1155, 2013.

[8] J. Sackmann, "Tennis atp." https://github.com/JeffSackmann/tennis_atp, 2018.