Predicting Tennis Match Outcomes using Machine Learning

Prateek Talukdar
Department of Computer Engineering
Rochester Institute of Technology
pxt8351@rit.edu

***Abstract*—As is the case with any sport, a large amount of data is generated from every tennis match played on tour. At first glance, it may be assumed that this data is useful only for documenting results. However, given the wide variety of factors that determine the outcome of any match, valuable insights into how players match up can be learned by delving deeper into the data. This information can then be exploited to determine the likely outcome between any two players in future matches. The objective of this project is to assimilate data of professional tennis matches played over the past few years, and train a model using machine learning techniques to predict the winner of a match between two players.**

***Index Terms*— Fantasy Sports, Machine Learning, Neural Networks, Predictive model, Support Vector Machines (SVM), Tennis**

I. Introduction

Tennis is one of the most popular sports globally, with tournaments drawing hundreds of thousands of spectators at the venue. Additionally, millions more follow their favorite players through digital broadcasts. The BBC reported a total viewership of almost a billion people [1] for the professional men's tour in 2015, and these numbers are steadily rising every year. Given the widespread international interest in the sport, it is unsurprising that the fantasy sports market continues to gain prominence. To be successful, participants need to accurately determine the winner of any given match. It may also involve additional parameters such as the predicting score and estimating player statistics.

Given the huge number of variables that come into play while making a forecast, it may be beneficial to use sophisticated learning techniques to make better predictions rather than relying on gut instinct. The objective of this project is to use historical match data along with head-to-head statistics between players to predict the likely outcome professional men's tennis matches on the ATP World Tour.

The rest of this paper is organized as described below. Following the introduction, Section 2 expands upon the significance of the different parameters used to make predictions. Section 3 introduces the learning methods being utilized for this paper. Section 4 gives an overview of the dataset, and also presents the experimental results. Section 5 correlates the obtained results with real-world predictions and lays out the scope for future work.

II. Background

A tennis court is a rectangle 78 feet long and 27 feet wide, which is separated into two equal halves by the net. However, it may be played on a variety of surfaces including clay, grass and hard. Further the courts may be indoors, or exposed to the elements outdoors. Each of the above drastically alters the speed and bounce of the ball, lending itself to different playing styles. Generally, players tend to favor the surface they trained on during their formative years, and consequently achieve their best results on the surface. A player adept at the serve-and-volley game style might find themselves struggling on the long drawn out rallies on a clay court. Hence, the surface of the court is always an important consideration while making a prediction.

Intuitively, the ranking of the players is also a crucial factor. Higher ranked players tend to be more skilled, and win a higher percentage of their matches. They also tend to have better head-to-head records against most other players, thereby achieving the better ranking. Although it would be easy to conclude that the higher rank player will always prevail, practically this is not the case as there are upsets in nearly every tournament.

Any point in a match begins with a player serving the ball across the net. This is the only shot wherein a player has complete control, and is not reacting to the ball struck across the net by his opponent. Players with higher serving speeds or accuracy can protect their own serve better, making it significantly harder to beat them. Hence, serving statistics go a long way in determining the outcome of a matchup between two players.

There are several other factors which can be considered, including but not limited to age, dominant hand, fitness levels and length of matches. Knowing which features to select from the dataset can make or break the predictive model. Thus supervised learning techniques are preferred.

A. Related Work

There has already been some research into predictive models

for tennis matches. Barnett and Clarke [7] use historical data to determine the probability of a player winning a single point, which is then extended to the match as a whole. Clarke has also come up with a model that uses the difference in the rankings of two players to make predictions as to the outcome. These methods use rely on a single feature to make predictions and leave room for improvement. Wagner and Narayanan [6] use SVM to help predict the results in a popular fantasy tennis competition which attracts thousands of participants each year. Sipko and Knottenbelt [4] provide a comprehensive look at the process of extracting meaningful features from the raw dataset. The authors also introduce the common opponent model in the paper. Barnett and Pollard [8] carry out studies to highlight the significance of the court surface on a player's performance. Cornman et al. [5] use tennis betting odds to develop a single shot decision model to maximize the projected return on bets.

B. Project Setup

A singles match is contested between two players, Player 1 and Player 2. Player numbers are assigned sequentially as per the official tournament draw. The target outcome between the players can be defined as {1, 0} where 1 is the outcome of Player 1 winning the match and 0 is the outcome of his opponent (Player 2) winning the match.

As only supervised learning methods are used, the input vector X comprises of the details of the player and the match while the output y is the outcome of the match. This convention has been used throughout the project.

C. Feature Engineering

A major issue to overcome before the model could be trained was the fact that all match results in the dataset were originally in the form '*Player1* beat *Player2* by *score*'. If the model were to be trained in this way, there would be an inherent bias towards Player 1, and the resulting predictions would be completely skewed. To avoid this, half of all matches (selected at random) had the names and corresponding attributes of both players interchanged. Hence, the player previously labeled as Player 1 would now be known as Player 2. Additionally, a column specifying if Player 1 was victorious was added. This was not present in the source data as it defaulted to Player 1 always winning.

Since the player attributes were interchanged to avoid bias, the score representation became meaningless (tennis scores always list the winning player first). Hence, a custom feature called match coefficient was introduced, which takes into account the overall difference in score across the total number of sets played.

To reduce dimensionality, certain attributes like ranking or points was represented as the difference between the corresponding features of both players. This also helps prevent overfitting due to the prevalent symmetry in the original dataset.

The decision threshold used for predictions was 0.5. Hence a predicted label with a value greater than 0.5 implied the model favored Player 1 to be the winner, while a lower value meant Player 2 was picked as the winner.

III. PROPOSED METHODS

Different machine learning techniques were used to predict the outcomes. A brief summary of each model is included below.

A. Linear Regression

Linear regression is a method of mapping the output of one dependent variable based on one or more defining (independent) variables. It can be used to fit a predictive model based on an observed set of values. After training, when additional unlabeled points (dependent variable unknown) are considered, the fitted model is used to make a prediction of the response. Since our desired prediction is if a given player wins or loses a match, Linear Regression is the basis of a good classification technique.

B. Support Vector Machine (SVM)

SVM is a supervised learning algorithm that distinctly classifies points on either side of a separating hyperplane. A hyperplane is nothing but a decision boundary used to separate input data points. After learning the training data, an SVM builds a model that assigns unforeseen test samples into one category or the other. A linear SVM model was used in this project as the dimensionality of data was not very high.

C. Random Forest

A random forest model is a supervised learning technique. It may be thought of as an extension of the random tree model in which multiple decision trees are trained. In random forests, trees are 'grown' from random samples of the training set by continually splitting the data. This means that a subset of features is chosen at random and the split reflects the best partition based on those features. The number of trees is a hyperparameter specific to the model. For any data point, the output of this algorithm is an average of the outputs of each tree containing that data point.

D. Naïve Bayes

Naïve Bayes classifier is a probabilistic model based on Bayes' theorem. The defining characteristic of this algorithm is that all features used for classification are assumed to be independent from one another. It is further assumed that each feature contributes equally to the final outcome, which may not always hold true. However, due to the simplicity of the model, predictions made are generally close to those obtained via other learning methods.

IV. RESULTS

The data used for this project was obtained primarily from [3]. The data here was separated based on year, with each year containing the results of approximately 2600 matches. All matches from 2012-2016 were used to train the model, 2017 was used for testing and finally the results were validated on all matches contested in 2018. Hence, the data was approximately partitioned in a 70-15-15 ratio for training, testing and

validation respectively. Each match recorded in the dataset includes details of the two players including ranking at the time of the match, the tournament and round at which the matchup occurred, the court surface, and the final score. However, some records were incomplete or missing data, and these entries were manually removed before training. Approximately 95% of the aggregated data was usable.

The results for the different learning techniques described in Section 3 are summarized in the table below.

TABLE I
PREDICTION ACCURACIES FOR DIFFERENT MODELS

Model	Prediction Accuracy
Linear Regression	68.29%
SVM (Linear)	68.06%
Naïve Bayes	66.92%
Random Forest	64.94%

Just observing the accuracies gives us an overview of which models are best suited to this problem, but they offer few insights with respect to making sense of *what* the models are actually predicting. Additional analysis of the predictions was done to come up with metrics to better understand the results.

The predictions from the four methods were combined into a single result set for the following.

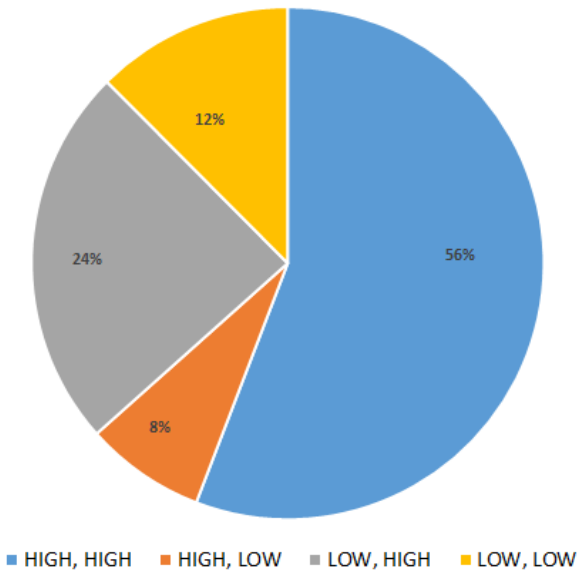

Fig. 1. Prediction accuracy based on player ranking. The first parameter represents the predicted player while the second represents the actual winner of the match. HIGH, HIGH indicates the higher rank player was predicted, and actually went on to win the match.

The ability of the models to predict upsets is captured in Fig. 1. Higher ranked players were favored 64% of the time, while an upset was predicted on the remaining 36%. Comparing these to the actual results, we can see the models predicted the winner correctly in approximately 68% of all matches (H, H and L, L). A lower ranked opponent prevailed on only 8% of the matches when the higher ranked player was predicted, while higher ranked players prevailed on 24% of all matches when an upset was predicted. Thus, the blue and yellow regions correspond to accurate predictions while the orange and grey regions represent the error.

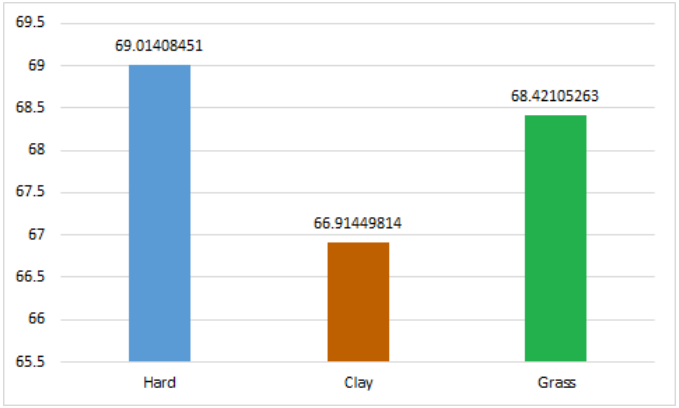

Fig. 2. Prediction accuracy by playing surface

The next metric to evaluate is prediction based on playing surface. As seen in Fig. 2, the predicted accuracy for the three surfaces are similar, with hard courts proving to be the best. This can be explained by the fact that a greater number of matches are contested annually on hard courts (60%) compared to clay (30%) and grass (10%). Hard courts are also considered to be the most 'neutral' surface due to their consistent pace and bounce. Given the specialized nature of clay and grass courts, these results are in line with what we expect to see.

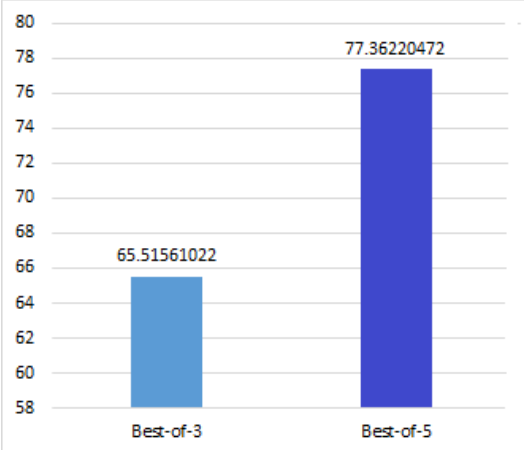

Fig. 3. Prediction accuracy by format

Fig. 3 illustrates the accuracy of the models across different formats. Apart from the four grand slams where matches are contested in the 3-out-of-5 set format, all other tournaments use the standard 2-out-of-3 sets. Hence, this is a good indicator of the predictions for the most prestigious tournaments. The

results in the best-of-5 format is significantly higher simply because it is that much more difficult to pull off upsets in the longer matches. A player might contest a portion of a match at a very high level (better than their statistical average), and in the shorter format this may provide enough momentum to close out the match. However, the longer format favors more skillful players who can sustain a high level of performance, and thus fewer upsets are seen. Here again the results are in line with the expected outcomes.

V. Conclusion

The obtained results present a fair representation of the 2018 ATP season. It will be an interesting experiment to repeat the process with the data from the 2019 tennis season and see how the models fare with an additional year's worth of training data. Custom datasets can be created by manually scraping and merging data from several online sources. These expanded datasets would include more fine grain statistics and increase the dimensionality of the model. However, it would allow for better hyperparamter tuning, and result in higher accuracies. Additionally, models of higher complexity like non-linear SVM or neural networks can be developed.

References

[1] BBC US & Canada. (2016, Mar. 21) *Novak Djokovic questions equal prize money in tennis* [Online]. Available: https://www.bbc.com/news/world-us-canada-35859791

[2] V. A. Freeman. (2017, Jul. 25). Kaggle ATP Tennis Dataset [Online]. Available: https://www.kaggle.com/m3financial/atp-tennis-data-from-201201-to-201707/version/9

[3] Tennis-Data.co.uk (2018, Nov.). ATP Men's Tour data [Online]. Available: http://www.tennis-data.co.uk/alldata.php

[4] M. Sipko and Dr. W. Knottenbelt, "Machine Learning for the Prediction of Professional Tennis Matches," M.Eng. Final Project, Dept. Computing, Imperial College, London, 2015

[5] A. Cornman et al., "Machine Learning for Professional Tennis Match Prediction and Betting," CS229 Term Project, Dept. Comp. Sc., Stanford Univ., CA, 2018

[6] A. Wagner and D. Narayanan, "Using Machine Learning to predict tennis match outcomes," 6.867 Final Project, Dept. Elec. Eng. And Comp. Sc., MIT, Cambridge, MA

[7] T. Barnett and S. R. Clarke, "Combining player statistics to predict outcomes of tennis matches," IMA Journal of Manage. Math., 16:113120, 2005

[8] T. Barnett and G. Pollard, "How the tennis court surface affects player performance and injuries," Medicine Science Tennis, 12(1):34–37, 2007

[9] ATP World Tour Performance Zone [Online]. Available: https://www.atpworldtour.com/en/performance-zone