Tennis Prediction Model

PREPARED FOR

Seng 474 Data Mining Project

PREPARED BY

John Schriemer
Shayla Grymaloski
Amaan Makhani
Bryan Travers
Ricky Huang

1. Problem Overview

The global sports market reached a value of nearly $488.5 Billion in 2018 [1] and has been growing steadily. Some of this growth is attributed to the increase in sports analytics utilizing statistics and, more recently, machine learning. Using machine learning to analyze and predict sports results has proved both helpful and lucrative for data scientists, coaches, and bookkeepers alike. In particular, making predictions using machine learning have been utilized by those in the growing sports betting market, which has an estimated market value of USD 155.49 billion by 2024 [2]. The demand for accurate predictive models in these markets has prompted a new problem space for machine learning, with many challenges yet to be solved.

Tennis poses some unique challenges concerning machine learning. With three different playing surfaces, varying climates and conditions, and diverse playing styles and strategies, the potential for noisy data is nearly guaranteed. Also, it has been difficult to predict a match outcome when the momentum of any tennis match can swing in the matter of a few successfully placed strokes. David Foster Wallace has well-described tennis' unpredictability as "Just one single shot in one exchange in one point of a high-level match is a nightmare of mechanical variables". The large number of variables that define a match has made the problem of using machine learning to predict match outcomes intriguing and worthwhile for our team.

Our team plans to use a combination of a few well-established ATP tour result data sets from the last 20 years from Kaggle and Opendatasoft. The final report will document the training of different machine learning models and methods using platforms TensorFlow and Scikit Learn. The risk of obtaining and utilizing the data set is minimal as the data set is already retrieved from these hosts. In terms of data quality, if there are missing rows, we will deal with these rows by entering an appropriate default value. If time allows, we could add in some model optimization, like trimming the 20-year data set by various amounts to see if we can obtain comparable results with less data. The two datasets used for training come from two different perspectives. One dataset is obtained from the ATP tour while the other includes bookkeeping match odds. We plan on experimenting with the ATP dataset, comparing the last 20 years of data with the full 63 years of match data. Due to the minimal risk of obtaining the data, there are no current backup sources of data other than our two datasets. We are planning to then predict the outcome of the 2017 tennis season with our program and compare the results with the actual outcome of the 2017 season.

2. Goals

The complexity of tennis as a sport, and its propensity for anomalies, makes match results difficult to predict. Our project will obtain two measurements of success to combat this complexity. These two measurements are correct match predictions and correct tournament predictions. The rationale for breaking the overall measure of success into smaller steps is to increase the power of our model. These results evaluated in regards to a weighted distribution measure of success would only classify if the prediction of the grand slam is right or wrong. Although this data is ideal for evaluating the entire project, it doesn't account for the proximity to the correct prediction. The proximity to the

result is simply the number of rounds separating a player's elimination compared to the actual round the player was eliminated. Using this measure, we will then be able to decrease the proximity to the prediction, which will, in turn, impact the overall prediction accuracy in an ideal scenario. We hope to achieve a match prediction rate of 70% and a tournament prediction rate of 25%; these measures of success are based on similar studies conducted for tennis match prediction [3][4]. If time permits will compare our historic vs more recent data sets to see which model gives the most accurate predictions.

3. Project Plan

Date	Goal
June 22-28	Clean the data (Fill in missing data, convert data types to be consistent, and add/delete attributes as needed)
June 29- July 5	Try out different models (LinearRegression, DecisionTreeRegressor, RandomForestRegressor, Neural Network) using various platforms such as TensorFlow, Scikit Learn, PyTorch, Keras, etc. and training them using cleaned data to find and shortlist a handful of promising models. Examine the factors of each model above including efficiency rate, training time, data size, and other factors. Create and finish the **Progress Report** outline the project status including trial data such as models used and preliminary model examination.
July 6-12	Fine-tune the parameters of the selected model and examine the effect of different parameters on the efficiency of the system. Examine classification and deal with the overfitting of data.
July 13-19	Evaluate a preliminary model on the test set and examine the viability of the results. This will include retrieving the predictors and the labels from the test set, running them through the pipeline to transform the data and evaluate the final model on the test set.
July 20-26	Examine the reproducibility of results and gathering the results in the form of tables to be used in the final report. Structure of the final report and prepare the code and data sets for submission. Outline presentation focuses as well as the medium to which the presentation will be presented. Analyze the results in the final report in order to sum the outcome of the project.
July 27- August 2	Complete the **Final Presentation** & **Final Report** of findings and the solution (highlighting what we've learned, what worked and what did not, what assumptions were made, and what the system's limitations are) all while creating visualizations of our data analysis.

In order to compare the various models used in the preliminary report results, we will modify the learning rate, the size of the dataset, and other parameters that affect the categorization of a data set. This will be performed as pruning methods in decision trees. To guide us in our project, we have found a prior article that used neural networks and SVM's to predict match outcomes [3]. Another report is also used to guide the feature vector used and the data needed [4]. These articles will help guide us and allow us to compare our results through the use of a measure of success. In addition to the articles, we will use the guidance of our professor, Nishant Mehta. We will also be using the guidance of our teaching assistants Mahdi Hajiabadi, Bingshan Hu and Hamid Shayestehmanesh.

4. Task Breakdown

All the members of this project team will play a vital role in the success of the team. The task breakdown will ensure that all team members are involved in a machine learning aspect of this project. We aim to select two models, namely model 1, and model 2 to produce the final report. These models will then be used to split our group up into two groups, a group of two and a group of three. The smaller group will be responsible for modelling the data in Scikit Learn. The larger group of three will work on a new method of modelling we have not discussed in class. Currently, the plan is for the large group of three to explore and use the model Tensorflow.

We intend to use multiple datasets, all of which require cleaning. As such, each team member will be responsible for cleaning their assigned data. In terms of the presentation and report, since the formats are unknown, it is assumed all members will be assigned a section of the report and a similar section in the presentation. These will not necessarily be used, but are intended as backup data. The task breakdown of individual team members is shown below:

Amaan: Cleaning dataset 1, program model 1, change the parameters of model 1, keep track of data produced by model 1, contribute to the report, and contribute to the presentation.

Bryan: Cleaning dataset 2, program model 2, change the parameters of model 2, keep track of data produced by model 2, contribute to the report, and contribute to the presentation.

John: Cleaning dataset 1, program model 1, change the parameters of model 1, keep track of data produced by model 1, contribute to the report, and contribute to the presentation.

Shayla: Cleaning dataset 2, program model 2, change the parameters of model 2, keep track of data produced by model 2, contribute to the report, and contribute to the presentation.

Ricky: Cleaning dataset 2, program model 2, change the parameters of model 2, keep track of data produced by model 2, contribute to the report, and contribute to the presentation.

5. References

1) ltd, R., 2020. *Sports Global Market Opportunities And Strategies To 2022*. [online] Researchandmarkets.com. Available at: <https://www.researchandmarkets.com/reports/4770417/sports-global-market-opportunities-and-strategies?utm_source=BW&utm_medium=PressRelease&utm_code=ctvc8g&utm_campaign=1244426+-+Sports+-+%24614+Billion+Global+Market+Opportunities+%26+Strategies+to+2022&utm_exec=joca220prd> [Accessed 20 June 2020].
2) R. ltd, "Sports Betting Market by Platform, by Type, and by Sports Type: Global Industry Perspective, Comprehensive Analysis, and Forecast, 2017-2024", *Researchandmarkets.com*, 2020. [Online]. Available: https://www.researchandmarkets.com/reports/4853933/sports-betting-market-by-platform-by-type-and. [Accessed: 20- Jun- 2020].
3) A. Cornman, G. Spellman and D. Wright, "Machine Learning for Professional Tennis Match Prediction and Betting", *Cs229.stanford.edu*, 2020. [Online]. Available: http://cs229.stanford.edu/proj2017/final-reports/5242116.pdf. [Accessed: 20- Jun- 2020].
4) A. Wagner and D. Narayanan, "Using Machine Learning to predict tennis match outcomes", *Deepakn94.github.io*, 2020. [Online]. Available: http://deepakn94.github.io/assets/papers/6.867.pdf. [Accessed: 20- Jun- 2020].