

CS 5805 Project Proposal
Nick Ponticelli
Dataset: Jeff Sackmann Tennis ATP/WTA Match Data

Introduction

The dataset is ATP matches from 2007 to 2024. This dataset is publicly available and found on github. Each year has roughly 3,000 observations meaning in total this project will have 50,000 observations, with the potential for more as the data goes back to 1975. In addition, there are 48 features of the dataset, with 32 being numerical and 16 being categorical. Some examples of categorical features include type of tennis court surface, player nationality, and tournament name. There are numerous numerical features such as player physical and performance measurables. This data is relevant today and has clear applications in sports betting, coaching, and sport performance analysis. In addition, point by point data from various grand slam tennis tournaments will be used for feature engineering to add further statistical analysis of players. Each grand slam tournament has on average 50,000 singles points played which can be leveraged to further create a player profile based on tendencies and traits. It will be important to develop a strategy to clean the ATP match dataset as there are cases such as player retirements and injuries leading to null values and outliers for competitive match score.

Regression Analysis

The objective is to predict the score margin of a ATP match, which is defined as total games won by the winner minus the total games won by the loser. This is a continuous target variable that quantifies how competitive a match was. This has applications in match quality analysis as well as sportsbook margin modeling. Independent variables will include court surface, heights of players, player rankings, and ages. Feature engineering will occur to create rolling averages for players based on a previous number of most recent matches. This will add past performance to the model to make it more robust. Encoding will be required, one such example of this is court surface, which can be hardcourt, grass, clay, or carpet. In addition, handedness will be encoded as well.

Clustering and Classification

Unsupervised clustering will be done to group players into similar play styles. The variables used for this clustering will include winners, unforced errors, aces, point length, and match time. In projecting the clusters, it is expected that clusters will mirror human-recognized tennis play styles such as aggressive baseliners (high winners and unforced errors), serve-bots (high aces), and counter attackers (low unforced errors and longer points and match time). This data will be collected from the grand slam tournament datasets as this data contains more detailed point by point data for analyzing this. This data is significant for sports performance analysis, as different play styles necessitate different strategies.

Classification will be a logistic regression model which will predict the winner of the match as a binary one or zero. If the model projects the higher seeded player to win, the prediction will be 1, otherwise it will be 0. It will be important to review the dataset to ensure it is balanced as the data skews towards players of higher seeds winning more often. Feature engineering will be done to include past performance results as a rolling average, as well as historical win loss percentage versus their opponent in the match.

Association Rule Mining

Association rule mining will be used to find combinations of match conditions that are associated with winning outcomes. These rules produce insights that are useful for analysts, coaches, or broadcasters. Examples of these rules could include a high likelihood of a player winning on a grass surface if they are over 1.85m and left handed, or if a player on clay is a counter attacker and their opponent is an offensive baseliner. To make this more feasible, continuous variables such as aces, match length, etc will be transformed into discrete bins. This will allow us to identify which conditions are meaningful predictors of winning outcomes for all variable types.

Summary

ATP tennis data provides a great opportunity for applying machine learning over many problem types. This project will utilize regression for match competitiveness, clustering to identify different player styles, classification to predict winners. In addition, association rule mining will provide insights into combinations that are linked to certain match outcomes. Through all of these algorithms, a sophisticated level of tennis performance analysis will be reached. Specifically, all of these results can be used by coaches for player development and performance evaluation, sports broadcasting for creating data-driven narratives for audiences, and lastly for sports betting and analytics.