**Prediction of Tennis Match Outcomes Project Summary– 05/26/2022**

Situational Overview

Our data consultancy has been approached by LV Sports Consultants (LVSC), a group providing recommendations on wagering lines and betting limits to a number of Las Vegas sports books. LVSC has requested for us to build a quantitative model of the outcomes of tennis matches within the next 3 months, and to provide a semi-technical summary report on aspects of the model that we construct that were useful in optimizing the accuracy of our predictions. They gave us leeway in terms of exactly what level and variety of the game we focused on, and further indicated that they hoped to use our initial modeling exploration as the core of a more comprehensive effort that we could potentially become involved with long-term depending on our level of success and ideas for refinements. As the final stage of the 3-month project, we built a machine learning model to predict match outcomes, and presently share our main findings and proposed next steps with the LVSC team.

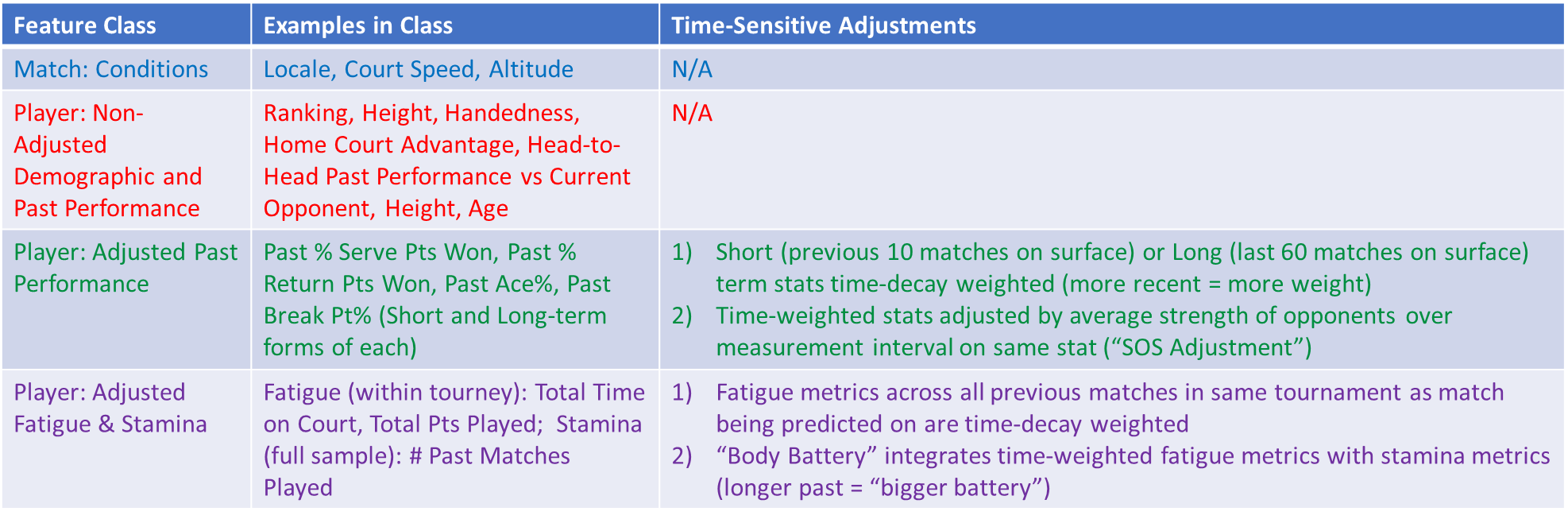
Methodology

We decided to focus on the top level of men’s professional singles tennis, the [Association of Tennis Professionals](https://www.atptour.com/) main tour, and obtained a large amount of raw match outcomes data going back over a decade from the [Github site](https://github.com/JeffSackmann/tennis_atp) of a well-regarded tennis statistician named Jeff Sackmann. Using a portion of this data archive (~35,000 matches from 2008-2019) and our knowledge of the game, we built a large set (~150) of individual player performance, player demographic, and match conditions-derived features to predict the outcomes of individual matches

The specific target feature (TF) that we aimed to predict was the percentage of points won by a given player [0-100%] in a given match. To predict this TF, we generated ~150 features comprising 4 general classes (**Table 1**). These four classes were 1) Match: Conditions; 2) Player: Non-Adjusted Demographic and Past Performance; 3) Player: Adjusted Past Performance; 4) Player: Adjusted Fatigue and Stamina. **Table 1** provides example features from each class, and describes the adjustments made to features in classes 3 and 4.

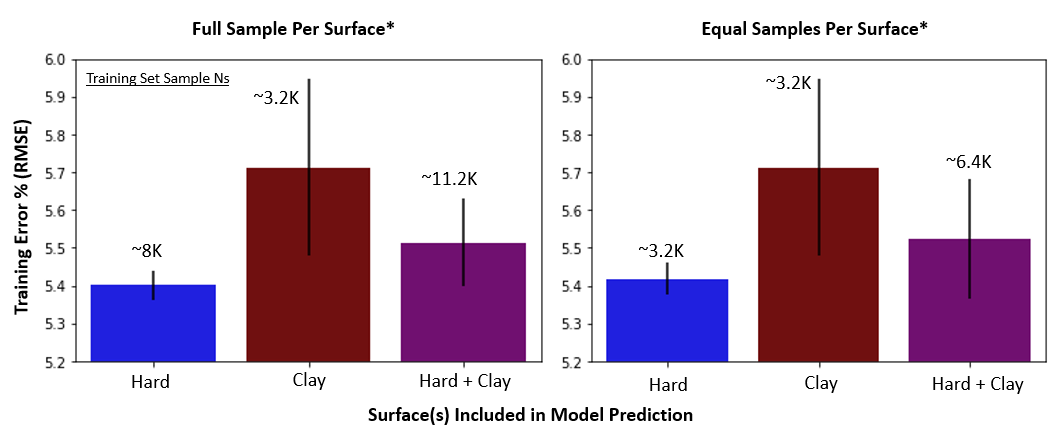
All features in the 3 Player classes were generated in both Raw and Differential forms. To take player height as an example, the Raw version would be the player’s actual height (e.g., 180 cm) and the Differential version would be 180 cm minus the height of his opponent in the match being predicted on (e.g., 180 -170 = 10). The opponent’s Differential height would appear in his record for the same match as -10 (170-180).

One additional critical methodological point is that prevention of data leakage was a priority across the feature generation and modeling stages. All predictive features were comprised *only of data that would have been available prior to the match being predicted*. For example, “Long-Term Past Serve Points Won%” (see **Table 1**) is an adjusted mean of a player’s data from the 60 matches he played prior to but not including the match being predicted. Thus, past performance features used to predict TF for each player in each match were generated with retrospective, rolling-windows of several different sizes.

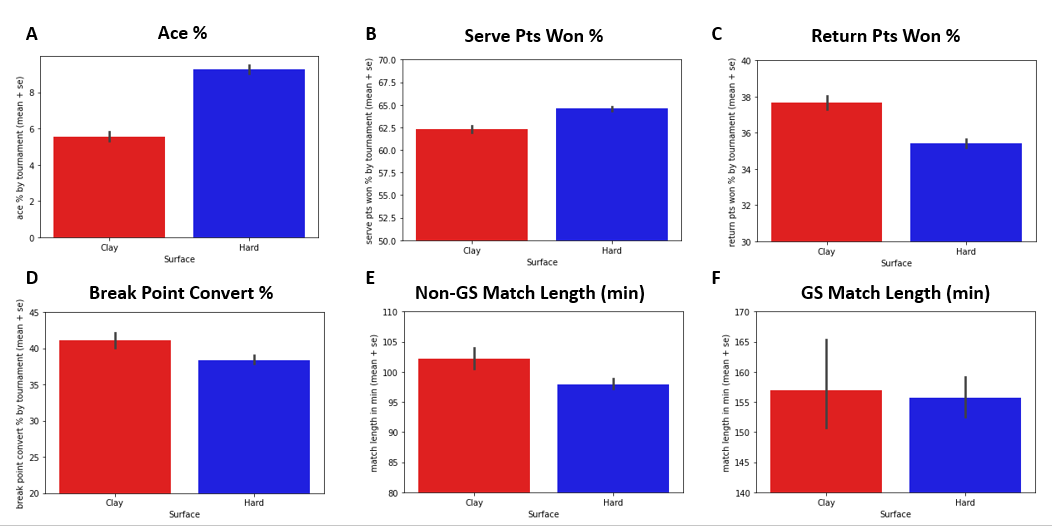
Table 1. Summary of Predictive Feature Generation.

Key Findings

1. Hard court tennis match outcomes were modeled more successfully than those on clay courts. **Figure 1** shows that the best model prediction quality was for hard court tennis alone, even when sample size was made equivalent for both surfaces.

Figure 1. Best model prediction quality by play surface inclusion. The left panel shows best prediction quality for hard (blue bar) and clay court (maroon bar) modeled separately, as well as hard and clay court tennis modeled together (purple bar). In this panel, all records (on per player per match) that passed initial filters were included. This resulted in nearly a 3:1 ratio of hard court to clay court matches in the sample.

To evaluate whether sample size disparity likely accounts for the prediction quality advantage for hard court vs clay court matches, a second analysis was conducted where the hard court records set was randomly reduced to be on identical size to the clay court record set (right panel). As can be seen in the right panel, the hard court prediction quality was only very slightly decreased when sample sizes were matched. The combined model is also reduced in prediction quality relative to the hard court model, despite having an increased sample size relative to the hard court model. This provides additional evidence that sample size alone is not the likely explanation for better prediction for hard court matches.

1. There are clear differences in the statistical description (**Figure 2**) and correlational underpinnings (**Figure 3**) of hard court and clay court tennis that may underlie the difference in prediction quality on the two surfaces.

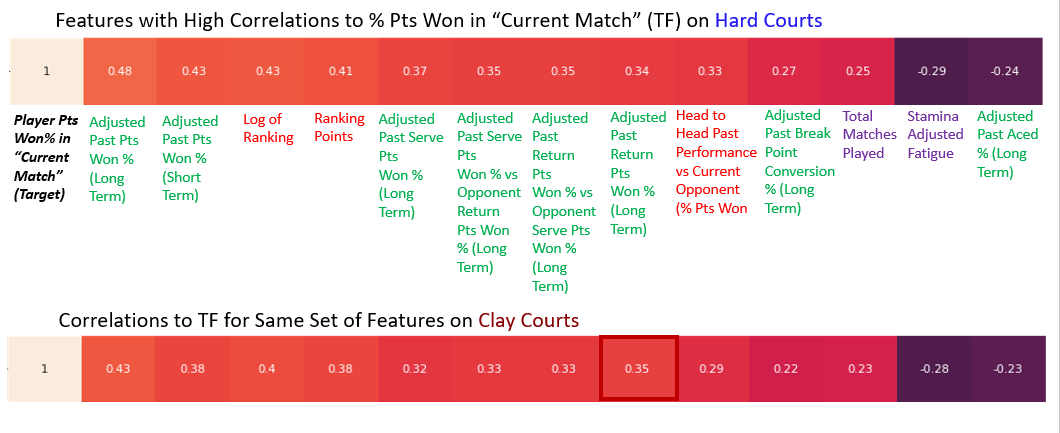
Figure 2. Statistical properties of hard and clay court tennis. On average, matches on hard courts are substantially more dominated by the serve (Panels A-D), and take longer than matches on clay (Panels E-F). Longer, more physically demanding matches and less dominance by the serve may lead to less predictability of outcomes on clay, and hence lower prediction quality. Furthermore (not shown), indoor matches are much more common on hard courts than on clay. Reduced impact of weather conditions, as well as less air resistance (leading to more aces) may lead to more predictable outcomes overall, which would also lend itself to higher prediction quality on hard courts.

Figure 3. Comparison of the strength and relative importance of selected predictive features and target feature (TF; percentage of points won by a given player in a given match) between hard court and clay court tennis. These features are among the most correlated to the target feature on both hard and clay courts. All of the highest correlation to TF predictive features are Differential features (Player Raw Feature Value – Opponent Raw Feature Value). Overall correlation strengths are lower on clay (only past return performance bucks this trend), which supports the idea that clay court tennis is harder to predict. The sequence of correlation strengths for these high correlation to TF features is quite different as well for the two surfaces, which may present a challenge when attempting to model both surfaces together.

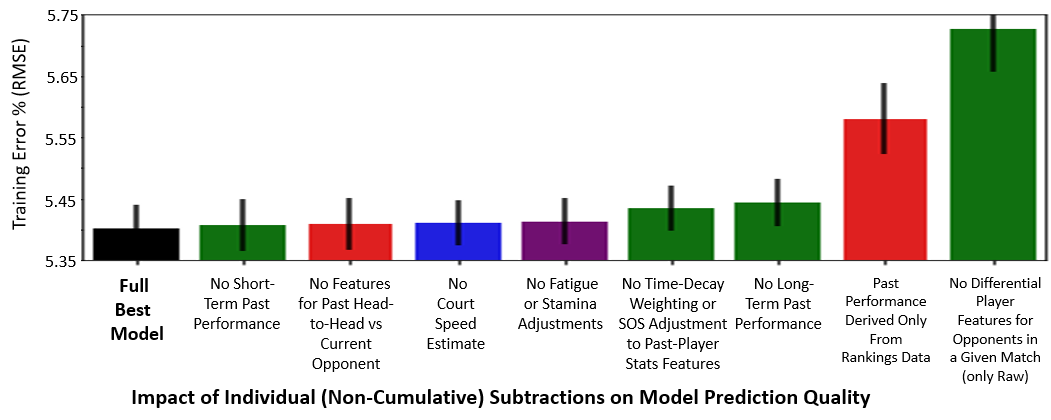
1. After playing surface, what matters most to prediction quality is direct comparison of opponents in the match being predicted. Especially important is to include features derived directly from past match statistics (**Figure 4**).

Figure 4. Dependence of prediction quality on key features classes and data adjustment types. The most important elements for best model prediction quality (two rightmost bars, left to right) were using 1) Previous match statistics, not just rankings and derivatives thereof, to create player performance features.; 2) Differential feature values between opponents in a given match that TF is being predicted (e.g., if Player A has a 10% Past Ace % [Raw Ace% A] and Player B has a 6% Past Ace% [Raw Ace% B], the Player A Differential Past Ace% = 10-6 = +4%). This depiction was constructed from the best hard court model, but the same trends apply to the clay court model, albeit with higher prediction error across model inclusions.

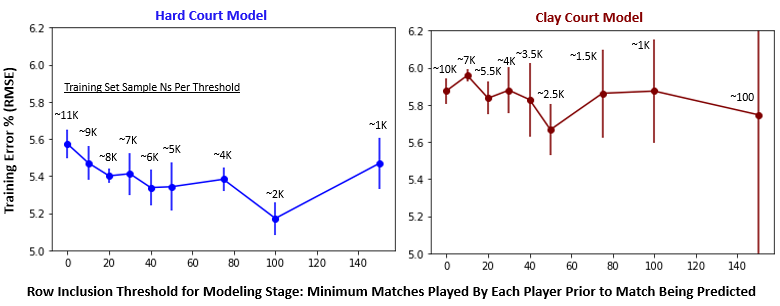
4) More data does not necessarily lead to improved match outcome prediction quality (**Figure 5** and **Figure 6**).

Figure 5. Model prediction quality as a function of per player previous matches played threshold. The best prediction quality occurs at a fairly high threshold of past matches for both players in the match being predicted on (100 for hard court; left panel). This occurs despite the fact that the overall amount of data included in model training at this threshold is ~1/5 of the total amount of data in the model sample. This threshold yields high prediction quality because it filters out matches between one or more players with low amounts of prior match data from which to construct predictive features for the match to be predicted on.

However, there is a clear tradeoff with setting the minimum threshold of past matches too high. A relatively high percentage of matches are played between players without a lot of experience at the top level of the ATP tour, and a model trained only on players with a large amount of previous data may be overfitted to matches played between more experienced players and will not perform well on the many matches between less experienced players it will “see” later. To counter the risk of overfitting the model to matches between only highly experienced opponents, a threshold of 20 previous matches for both players in a given match to be predicted on was applied for finding the best model.

One critical technical note: matches between players not reaching minimum threshold of prior matches in the overall sample to include in modeling were included in the earlier stages of accrual of past match stats and predictive feature creation for later matches in the sample. Thus, for example, a match between players both in their 21st match in the sample would be predicted using stats from their previous 20 matches in the sample. This despite the fact that, at a minimum threshold of 20 previous matches, those previous 20 matches would not themselves be included in the modeling stage.

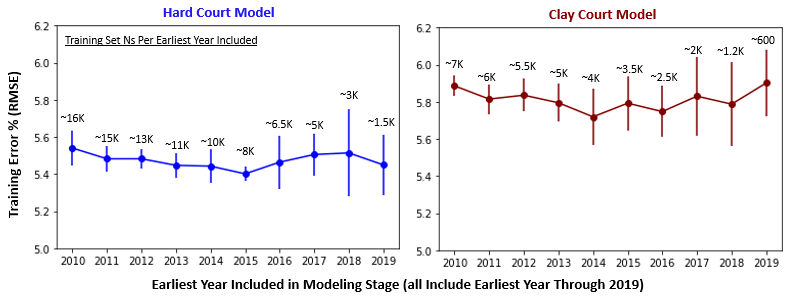


Figure 6. Model prediction quality as a function of data inclusion time range. Best prediction for hard court matches (left panel) occurred with 5 consecutive years of data included in the model. This was true despite the fact that this range (2015-2019) included half as much data as the maximum range (2010-2019). However, data availability did begin to matter with too small a sample (below ~2K samples), so there was a data “relevance” vs quantity tradeoff.

As an important technical note, a minimum of 2 years of data was used for accrual of past match stats and predictive feature creation for later matches in the sample. For example, when matches from 2010-2019 are included in model generation, match stats from the years 2008 and 2009 were used to generate predictive features for 2010-2019 even though the records from 2008 and 2009 were filtered out prior to the model generation stage.

**Recommendations to Client (Based on Key Findings)**

1. Maintain the separation of hard court and clay court models moving forward, and continue to strike a balance for data inclusion between maximal sample size and maximal relevance

Rationale: We were able to model hard court tennis more accurately, and with less variability, even when the sample size of matches was reduced to that of clay court matches. Whether or not hard court tennis will be more profitable to invest in will depend on the accuracy of the respective surface models *relative to the market*. Ascertaining this will require comparison of present model predictions to those derived from historical pre-match wagering line data

Along with this continued separation of the surfaces, we recommend based on our analyses to continue to limit the time range of modeling data inclusion to ~5 years and to exclude matches between players with very limited previous experience before a match being predicted (<20 matches on the same surface). This will ensure modeled data remain relevant to the current game, and that the model will not be overly influenced by either extremely experienced players or by players with too little previous data to accurately estimate ability.

1. Obtain historical pre-match wagering data and evaluate effects of integrating into current model

Rationale: These data can be added to the existing model as one or more new predictive features. If these features improve prediction quality (they almost certainly will, as the market is very wise), the logical next step would be to look closely at the features of matches for which prediction quality improved the most (see **Figure 7** for possible examples). The match features for which the market improves the model most will become primary focus for new data acquisition and integration for inclusion in the present model.

1. Obtain and integrate “micro-conditions” data into current model

Rationale: Clearly, where the current model is most lacking data-wise is in match-specific conditions that would be known only in the immediate lead-up to a given match being predicted. This “micro-conditions” data would include weather variables (e.g., humidity, temperature, wind speed and direction), court speed and ball bounce height estimation derived from “on the ground” sources (currently the proxy from this comes from data from the previous year at the same tournament), and real-time injury information on individual player as a tournament evolves. This approach synergizes with the second recommendation (see **Figure 7** for visual description) in that the market wisdom insights will refine which of these micro-conditions to prioritize for data acquisition and model integration.

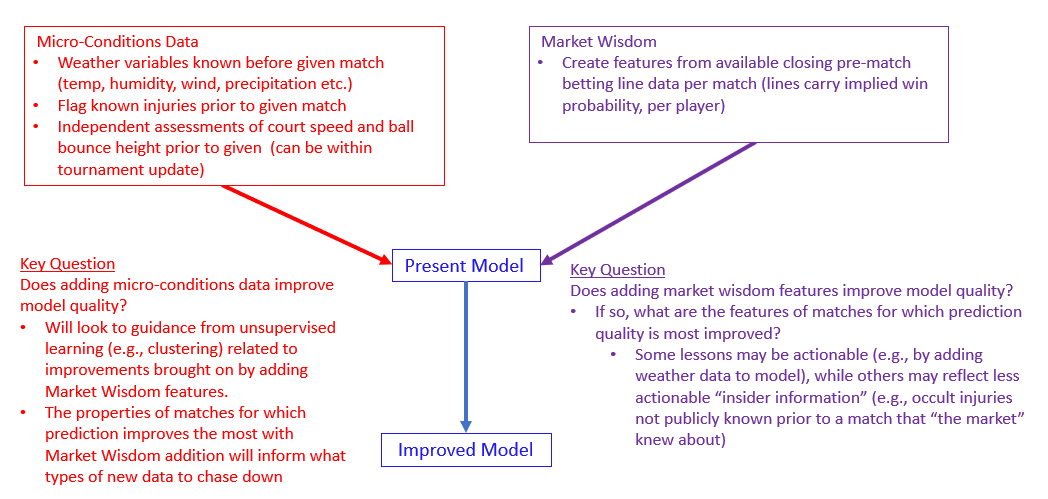


Figure 7. Future Directions Summary: Integration of “Micro-conditions” and “Market Wisdom” data to improve model.