Tennis Analytics: Using Machine Learning to Improve Performance of Emerging Players

Almas Rizvi, Arjun Chakraborty, Rajat Mittal, Shubham Arora, Matthew A. Lanham

Purdue University Krannert School of Management

rizvi7@purdue.edu; chakra38@purdue.edu; mittal51@purdue.edu; arora90@purdue.edu; lanhamm@purdue.edu

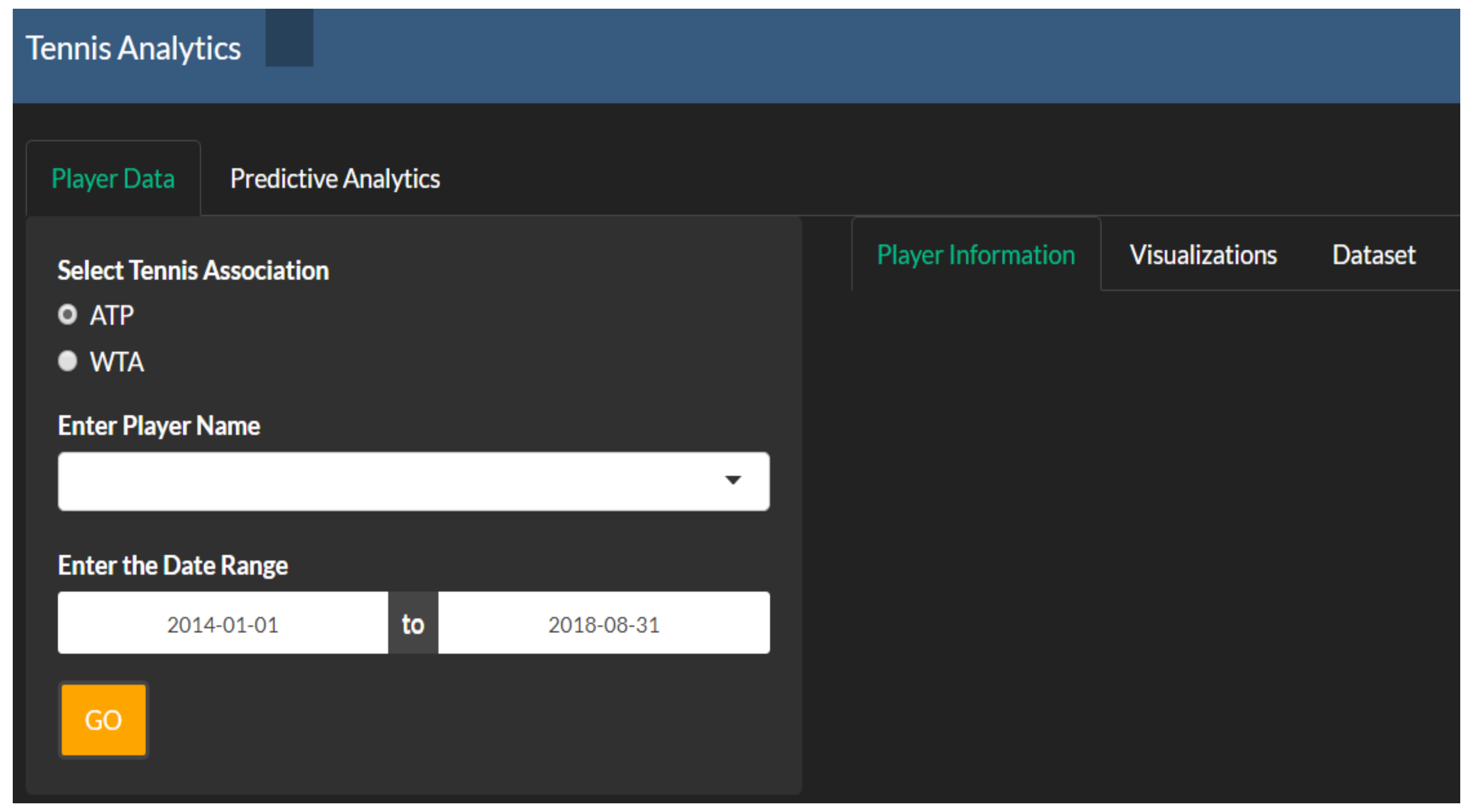
Abstract

Analytics has penetrated many sports including NBA, NHL, and NFL, but most of these sports focus on team level dynamics. In contrary, the unique analytics solution provided in the study is based upon individual player statistics in Tennis. Our approach utilizes the performance history of emerging players in helping them to improve their winning chances against the better ranked players in future matches. Using features like aces, double faults, first serve points, second serve successful and other match statistics, games won by player per match are predicted and it is used to calculate the winning score of player.

Introduction

The popularity of data-driven decision making in sports has revolutionized the way sports is being viewed today. Analytics in individual sports is somewhat lacking compared to that of large revenue generating team sports. Hence it was decided to explore the extent to which analytics could be applied in tennis. Match analytics could provide players and coaches with objective data to help them improve and strategize game tactics based on their opponents.

Having the ability to benchmark a player against other players, and identify areas of a player's game that need to be improved upon, would provide useful decision-support.



Select Tennis Association
○ ATP
● WTA

Enter Player Name

Enter the Date Range
2014-01-01 to 2018-08-31

Shiny App Demonstration

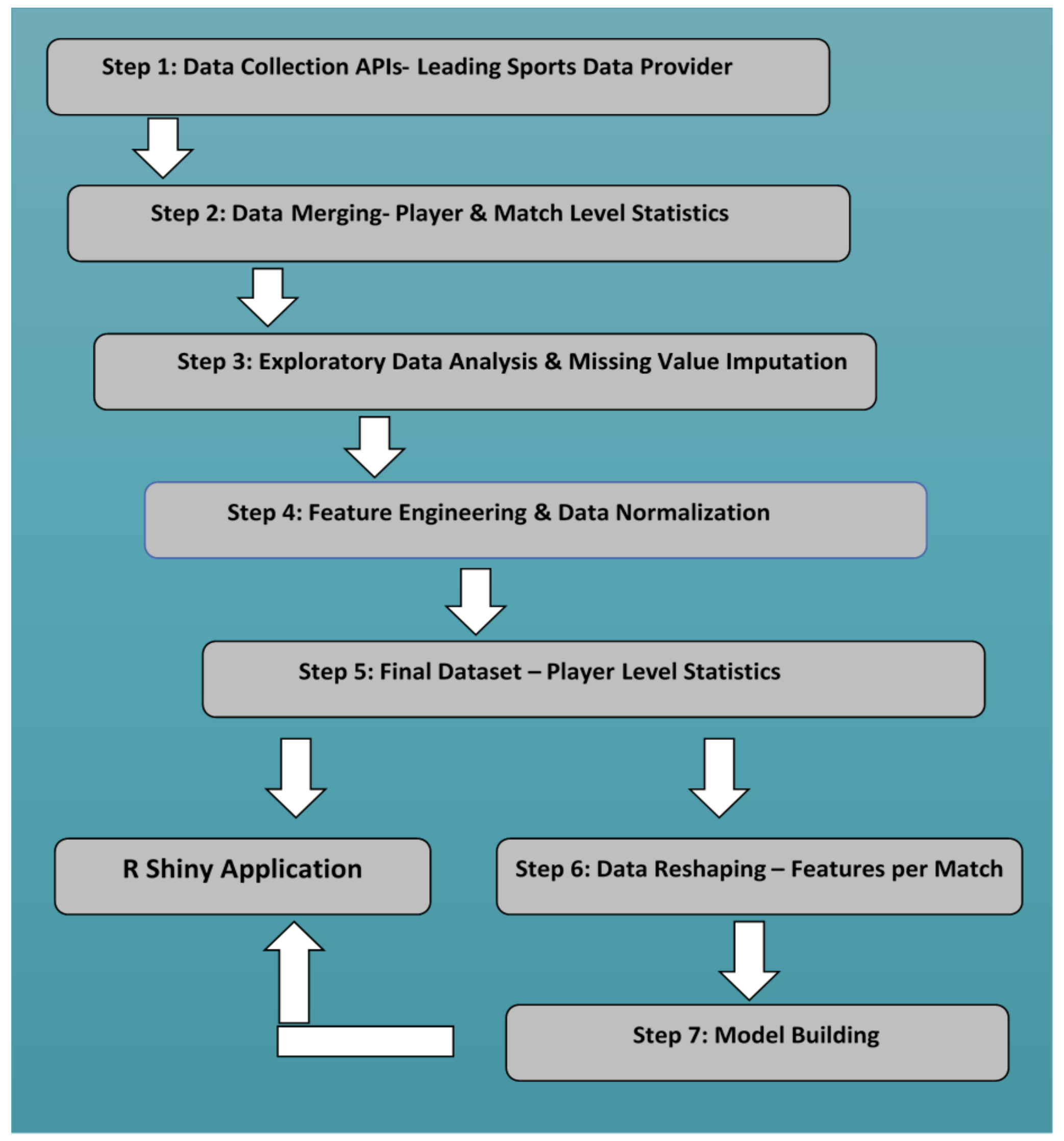
For demonstration, we developed an R-Shiny application which would enable one to visualize performance of a player and make comparisons among players based on their historical performance. This app could provide insights on the skill areas where improvement is required.

Data

The data used in our project was provided by Sportradar, a worldwide leader in the field of sports data and is thus proprietary in nature. The data consisted of ATP Ratings, WTA Ratings, Match Statistics, and Player Statistics and was scattered across different tables in the client database.

Methodology

The figure below summaries our overall analytics process. Data was mined, fused, explored, cleaned, and pre-processed for predictive modeling. Our application performs these tasks automatically.


Step 1: Data Collection APIs- Leading Sports Data Provider

Step 2: Data Merging- Player & Match Level Statistics

Step 3: Exploratory Data Analysis & Missing Value Imputation

Step 4: Feature Engineering & Data Normalization

Step 5: Final Dataset – Player Level Statistics

Step 6: Data Reshaping – Features per Match

Step 7: Model Building

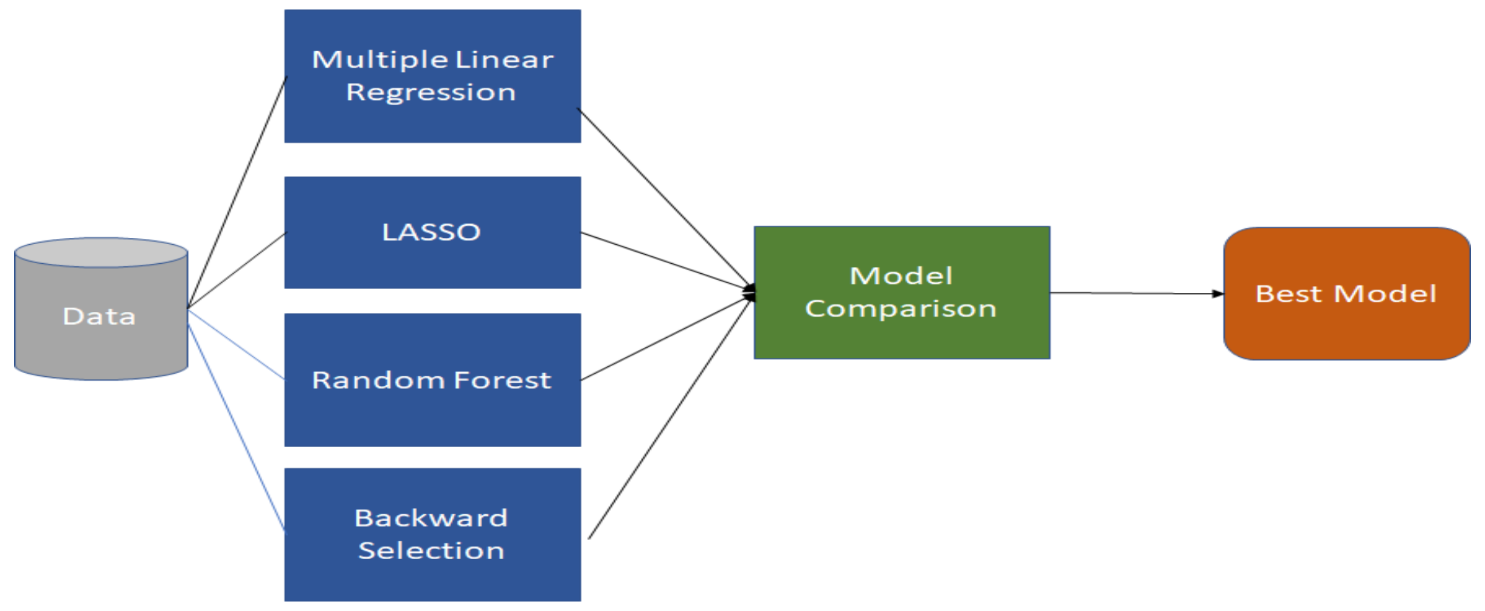
R Shiny Application

Data Cleaning & Pre-Processing

Data was merged and reshaped in a single data frame. It was then filtered for only the top 500 ATP and WTA ranked players. Data cleaning was performed to generate a new dataset with only the relevant variables identified from EDA. Only finished matches were considered so that there were no missing values present in the dataset. To remove the bias of players based on number of different matches and statistics, data was normalized by computing the per match level statistics of each player.

Model Building and Selection:

Four different supervised learning methods were to used to predict the number of games won per match for any player. Data was split into a 70:30 ratio and 3-fold cross-validated results for each model were obtained. From the model, a list of the most significant features were extracted and then models were re-trained to find the final accuracy scores. For example, from 17 features, nine features were extracted which were statistically significant. Adjusted R-square and RMSE statistics were used to evaluate the models.


Data

Multiple Linear Regression

LASSO

Random Forest

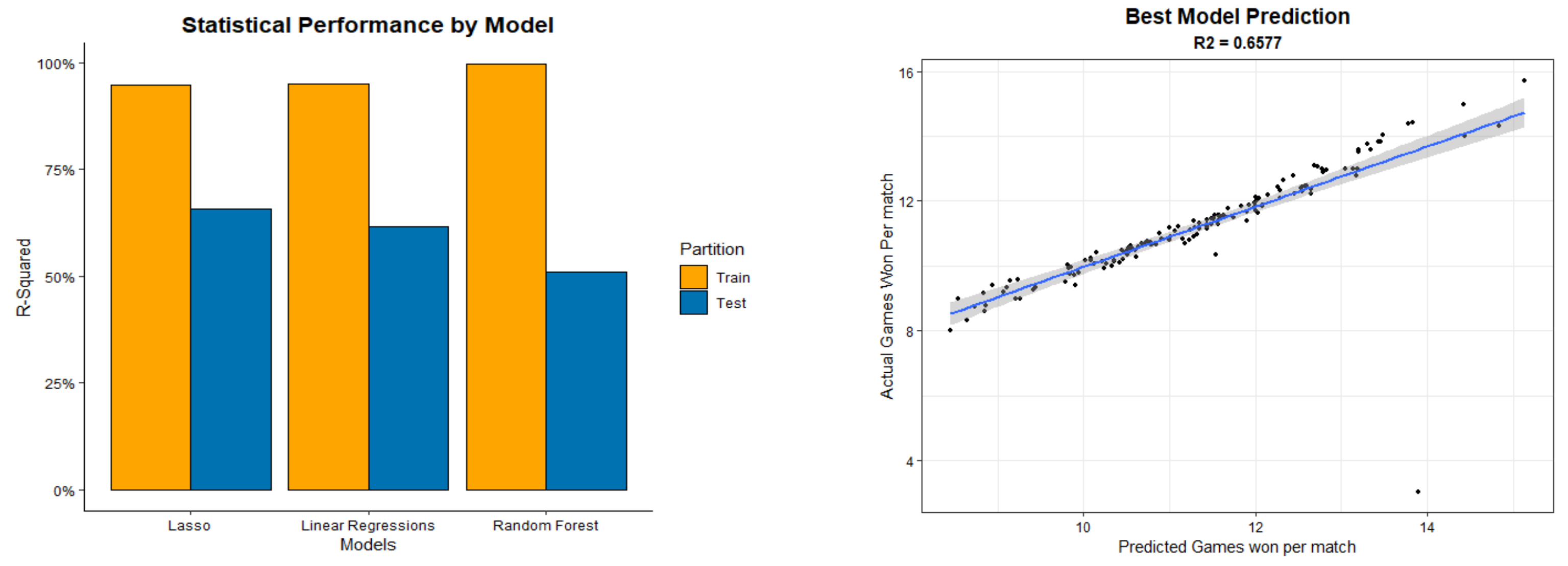
Backward Selection

Model Comparison

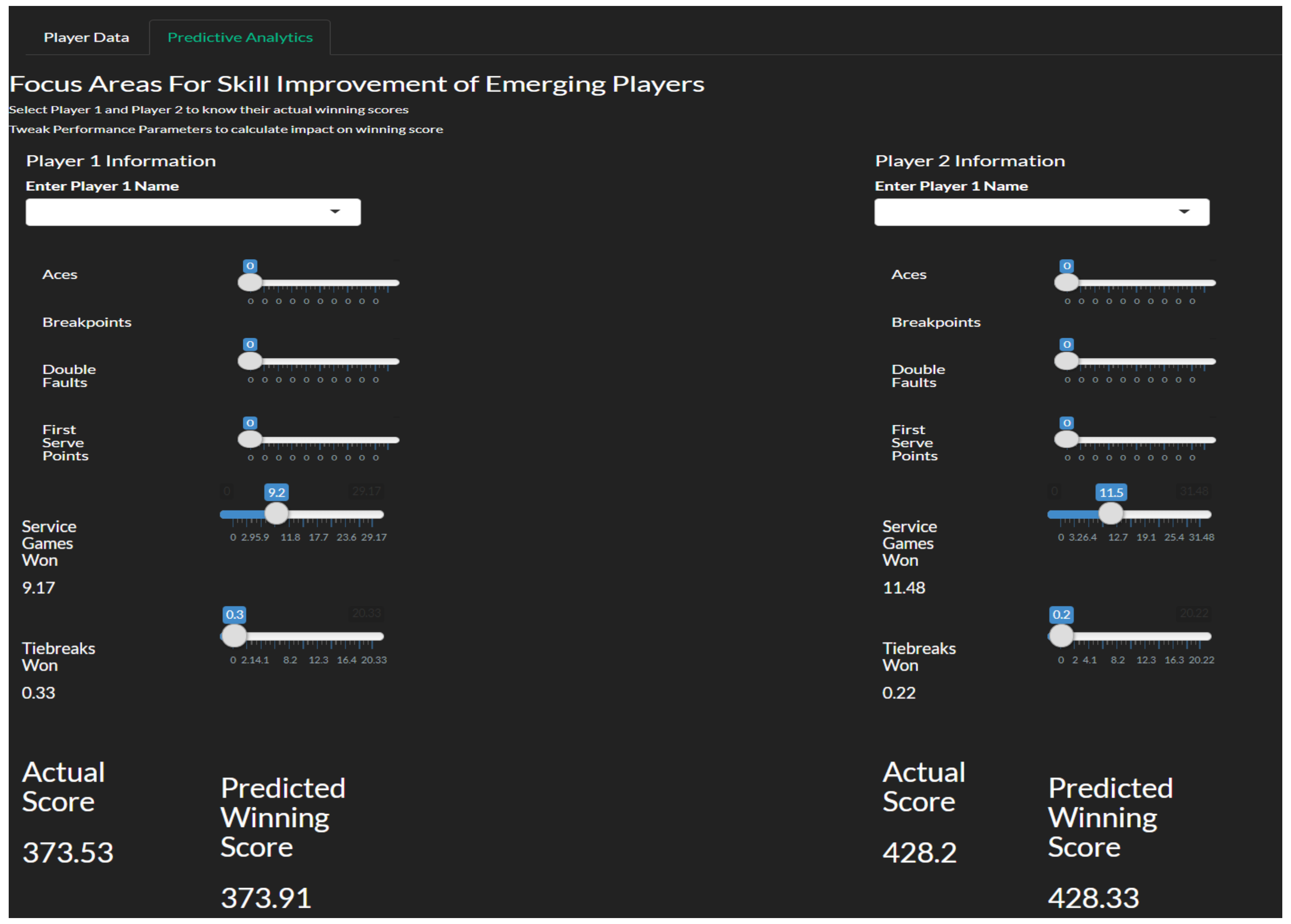
Best Model

Results

All three models we evaluated overfit to the training data as shown in the train versus test set statistics. We decided to use the Lasso model in this example where the test set performance was the best.



Model Evaluation



Focus Areas For Skill Improvement of Emerging Players

Select Player 1 and Player 2 to know their actual winning scores

Tweak Performance Parameters to calculate impact on winning score

Player 1 Information		Player 2 Information	
Enter Player 1 Name		Enter Player 1 Name	
Aces	0	Aces	0
Breakpoints		Breakpoints	
Double Faults	0	Double Faults	0
First Serve Points	0	First Serve Points	0
Service Games Won 9.17	9.2	Service Games Won 11.48	11.5
Tiebreaks Won 0.33	0.3	Tiebreaks Won 0.22	0.2
Actual Score 373.53	Predicted Winning Score 373.91	Actual Score 428.2	Predicted Winning Score 428.33

Shiny App Demonstration to evaluate performance of emerging players

Conclusions & Future Work

Based on our demonstration, the winning score prediction can help the emerging players to identify a few skill sets which they can focus on to increase the likelihood of winning against better ranked players. This could help the coaching staff plan trainings accordingly.

We are considering looking into weather conditions and court type to improve predictive performance. Also we are going to investigate Data Envelopment Analysis to benchmark performance among peer players.

Acknowledgements

We would like to thank our sports data provider, Sportradar, for allowing us to investigate this problem in Professor Lanham's MGMT 571 Data Mining