Identifying the most important variables influencing a team's match outcome in Rugby 7s

Capstone 1 Project
Rob Chudzik
3/20/18

Problem Statement

In sports team performance analysis, there are a myriad of performance variables that impact a team's performance and match outcome. The challenge for performance analysts and coaches is to determine which of these variables have the largest impact on a team's performance and match outcome, i.e., a win or a loss.

Client

The client for this project is the coaching staff of an International Men's Rugby 7s team ("the team"), and as such, the project will focus on identifying the most important variables for the this particular team.

Understanding the most impactful variables will allow the coaching staff (coaches, performance analyst and strength and conditioning trainers) to take action to either improve in these areas by allocating more resources (i.e., team training time; often the most valuable resource) to these identified areas or by adjusting the team's tactical and strategic game plans.

Data

The data to be used for this analysis is tournament and match data previously scraped from the 2016-17 HSBC World Rugby Sevens Series web site.

According to the official web site, "The HSBC World Rugby Sevens series consists of 10 tournaments held around the world, in which national sevens teams compete for series points at each round."

"An overall champion is crowned at the end of the season based on points accumulated throughout the 10 events, from the opening round in Dubai in December to the final round in Paris in June."

Each tournament features 16 teams, distributed into four pools of four teams, with 45 matches being played over a two- or three-day period, depending on the tournament. Each team plays every other team in it's pool, then the top two teams in

the pool standings progress to the Cup playoff round, the bottom two progress to the Challenge playoff round. Series points are awarded for the final finishing place in the tournament, and an overall World Rugby Sevens Series champion is determined based on the total number of points earned from all tournaments over the Series.

Solution Approach

The planned approach for solving this problem is to build a supervised machine-learning model to classify each of the team's match outcome as a win or a loss. The outcome variable will be derived from the point differential ('scores' variable) between the team and their opponent.

Predictor Variables
Predictor variables include the variables contained in the match data scraped from each of the team's the Single Match Summary report from World Rugby:

```
['Possession Time',
 'Possessions',
 'Ave Possession Time',
 'Scores',
 'Tries',
 'Conversions',
 'Goals',
 'Tries Conceded',
 'Points Conceded',
'Passes',
 'Tackle / Ruck / Mauls',
 'Retained',
 'Lost',
 'T/Overs Won',
 'Tackle Only (Defence)',
 'Mauls',
 'Kicks',
 'General Play T/Overs',
 'Lineouts',
 'Own Won',
 'Own Lost',
 'Scrums',
 'Own Won',
 'Own Lost',
 'Restarts',
 'Long',
 'Short',
 'Regained',
 'Errors',
```

```
'Pens / Frees Against',
'Ruck / Maul',
'Set Piece',
'General Play',
'Foul Play',
'Yellow / Red Cards']
```

Additionally, feature engineering will be used to create a list of players in the lineup for each tournament, to establish consistency and strength of the Eagles' team from tournament to tournament.

Data will be scraped to obtain each team's position in the World Series Standings (https://www.worldrugby.org/sevens-series/standings/mens) after each tournament in order to determine the team's and their match opponent's standing on the Standings table, to establish the relative strength of each opponent.

Feature engineering will also be used to develop stronger predictor variables, as needed.

After a classification model is built and trained, techniques such as feature importance scores can be used to identify the most important predictor variables.

Training data
Each 'core' team plays a minimum of four matches and a maximum of six matches in a tournament, over the course of 10 tournaments. This results in a range of 40-60 matches of the course of the series. In the 2016-17 Series, the team played a total of 57 matches.

The team's 57 matches from 2016-17 will be split into test and training sets.

Deliverables

A GitHub repository will be created, containing this project proposal, as well as the following project deliverables:

- Python Code
- 2016-17 HSBC World Rugby Sevens Series data
- Project Presentation