

Data Wrangling Report

Rob Chudzik
Capstone 1 Project

Data Sources

The source of the data used for this project is World Rugby, the global governing body for the sport of rugby. World Rugby compiles a statistical report at the completion of each tournament throughout the course of the Sevens World Series. For the purposes of this project, I planned to collect data from the current (2017-18) and previous two years (2015-16, 2016-17) Sevens World Series.

2017-18 Sevens World Series Data

The current season's match data was obtained from the World Rugby "Game Analysis Statistical Output," an Excel file that was published after each tournament in the World Series. The report featured several worksheets, including a "Raw Data" worksheet that contained match data for every match played in the tournament. A Python script was written to read a directory of reports, then clean and write the data for each of the USA's matches from each tournament to a CSV file.

2015-16, 2016-17 Sevens World Series Data

From the start, the existing data posed several challenges. First, the 2015-16 and 2016-17 data existed in PDF "Match Report" documents, with a separate PDF for each match played in a tournament. As there are 45 matches played in a tournament, over 10 tournaments in the World Series, the sheer number of PDF documents appeared daunting. Additionally, each PDF would need to be scraped to extract the match data. As a result, a decision was made to only extract data for matches in which the USA was playing (including the match data on their opponent), reducing the total number of PDFs to be processed from 900 to 117.

The [PDFTables product](#) was used to scrape the data from the Match Report PDFs, as the product featured a Python API. A Python script was created to read all PDFs from a directory, scrape the data from each PDF, and write the resulting data from each match to an Excel file. A Python script was written to read a directory of reports, then clean and write the data for each of the USA's matches from each tournament to a CSV file.

Data Wrangling

The second major challenge with the data sources was inconsistent features between the PDF and Excel reports. The Excel reports contained many more features than the PDF reports, so steps were taken to determine a common set of features that existed across both reports, or could be engineered from data that existed. Ultimately, two different data import and cleaning Python scripts were created – one for data from PDF files and one for data from the Excel files – as the cleaning process was quite different for each set of data.

The script to clean PDF data read a directory of Excel match files, performed cleaning and wrangling on each, and then wrote the results to a temporary dataframe, which would eventually be written to a CSV file.

As the data scraped from the PDFs resulted in extraneous data, and data represented in the wrong data type (i.e., time values represented as strings), much of the initial data wrangling steps consisted of cleaning these problem areas. Additional data wrangling included basic parsing of text strings to extract the Tournament Name and Match Number, as well as converting string representations of time values (i.e., 2:30 for possession time) to time values in seconds.

Feature engineering-type tasks included creating and calculating a new “Conversion %” feature from the existing “Tries” and “Total Points” features, and creating a new feature called “Avg Possession Time”, calculated from the 'Possession Time' / 'Possessions'.

Once these steps were complete, columns were renamed or dropped to match the common set of data features that would be used across both the Excel data and the PDF data.

Lastly, the data was transformed into what I called “diff data”, which was calculated for each feature as the difference between the match results for the USA and their opponent. For example, if the USA played Fiji and the USA’s “Contestable KO Win %” was 50% and Fiji’s was 25%, the diff result would be +25% for the USA.

The final dataframe was written to a CSV file created to store all of the match diff results from 2015-2018 (168 matches).

The Excel data wrangling consisted of much of the same types of tasks, specific to the Excel file data layout. Rows of summary statistics were dropped, data from merged or spanned data cells was split or concatenated, and any mismatched data types were cast to the correct type.

The exact same “diff data” transformation was created, and the final dataframe was appended to the CSV file written from the PDF data processing script.

Missing Data

Missing data was treated based on my domain knowledge. There was indeed missing data, but a closer examination revealed that the situation was not missing data, but the absence of a result or value for a particular feature. For example, if a value for the feature “Missed Tackles” was absent, the data (or lack thereof) represented the fact that there were zero missed tackles, not a missing value for “Missed Tackles”. This approach was validated by checking and comparing related features or aggregated features using zero for the value. This resolved all of the missing data issues.

Outliers

There were no steps taken to deal with outliers. In the course of exploratory data analysis, there were no outliers that caused concern for their impact on the predictive model.

Additional Feature Engineering

Lastly, additional feature engineering was done before the data was ready for model building.

First, as the feature “Contestable KO %” was identified in our hypothesis as an important feature in determining the outcome of a USA match, the “Contestable KO % diff” was binned into increments of 25%, so bins were created in the range of '-175 : -150' to '126 : 150'.

Lastly, a label feature was created to label the outcome of the match for the USA (0 = Loss, 1 = Win, 2 = Tie).

After feature engineering steps were completed, the data was ready for model building.