Capstone Project 1 - Final Report

Rob Chudzik

Project Objective

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

Problem Statement

In sports team performance analysis, there are myriad 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 the identified areas or by adjusting the team's tactical or strategic game plans.

Data Description and Transformation

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 collected data from the 2017-18, 2015-16, 2016-17 Sevens World Series.

2017-18 Sevens World Series Data

The most recent 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 each match played in the tournament. A Python script was written to read a directory of reports, and 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.

A commercial software product called PDFTables 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 a CSV file.

Data Wrangling

The second major challenge with the source data 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 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. The same tasks were conducted in the PDF version of the script.

Because the data scraped from the PDFs resulted in extraneous data, and some data features 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 for easier manipulation later.

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 were split or concatenated, and any mismatched data types were cast to the correct type.

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 total 'Possession Time' divided by the number of 'Possessions'.

Once these steps were complete, columns were renamed or dropped to match the common set of data features that would be used in the final set of data, drawn from both the Excel data and the PDF data files, from the 2015-16 Series through the 2017-18 Series.

Lastly, each feature in the data set was transformed into differential data ("diff data"), which was calculated for 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. Or if the USA missed 6 tackles and Fiji missed 3, the differential would be -3.

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

Missing Data

Decisions about the treatment of missing data were based on domain knowledge. There was what appeared to be 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 approach 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 potential 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'.

Ultimately, a decision was made during the building of the model to not use the binned KO% Diffs, as it resulted in more "noise" and diluted the impact of other features. The 12 binned "Contestable KO %" features were dropped in the model-building process.

Finally, a label feature was created to label the outcome of the match for the USA (0 = Loss, 1 = Win, 2 = Tie). A further decision was made during the model building, to drop matches resulting in a Tie from the dataframe, given that the classification models were binary classification models, and only accepted two values – i.e., a Win [1] or a Loss [0]. Only 3.2% of the USA's matches over the three seasons resulted in a Tie (5/156 matches).

After dropping the 12 binned "Contestable KO %" features, the final engineered data set contained 15 features:

```
'Opp', 'Tournament', 'Poss_Time_Diff', 'Score_Diff', 'Conv_Diff',        '
Tries_Diff', 'Passes_Diff', 'Contestable_KO_Win_pct_Diff','PenFK_Against_D
```

```
iff', 'RuckMaul_Diff', 'Ruck_Win_pct_Diff','Cards_diff', 'Lineout_Win_Pct_
Diff', 'Scrum_Win_Pct_Diff', 'Result'
```

Additional Data Sets Considered

There are several other data sources that could have potentially added useful information to the project and model, but after consulting with my mentor it was decided that obtaining, cleaning and transforming these additional data sources would be beyond the scope of the project. This is an area for future expansion or improvement of the model.

Player/Roster Data

The team's roster for each match could have an impact on the consistency of the team's performance across different matches or different tournaments. Additionally, adding match data for individual players would enable the scoring of individual players and their contribution or impact to the team, including the impact of not having top players in the team (due to injury, for example).

Tournament Location and Scheduling Info

An interesting area to explore is the impact of the tournament location and the distance a team has to travel to the tournament, as well as the scheduling information. Both pieces of data impact the physiological effects of travel and recovery on the players, which ultimately impacts their physical performance. The Sevens World Series is played over 5 different "legs", with each leg consisting of two tournaments on consecutive weekends, typically in the same general geographic region. For example, the first leg of World Series consists of the first tournament in Dubai, UAE, followed the next weekend by the second tournament in Cape Town, South Africa. Exploring the distance a team has to travel to get to the first tournament of a leg, as well as looking at how a team performs in the 2nd leg of a tournament would add previously unexplored relationships to the analysis.

EDA - Initial Findings

Exploratory Data Analysis (EDA) was conducted, primarily exploring the correlations between variables. The findings from the EDA analysis of the variables and correlations revealed that the strongest positive correlation is between the Possession Time Difference and Passes Difference (r=0.89). This makes intuitive sense, as typically the more time a team has possession of the ball, the more passes they will make.

Additionally, there was a moderately strong positive correlation between the Penalty-Free Kick Against Diff variable and the Ruck Maul Diff (number of rucks/mauls by a team) (r=0.77). As the number of rucks that a team has rises, so does the incidence of penalties - likely from ruck infringements/penalties. The ruck is often the most penalized area of the game, and it stands to reason that the more rucks you have in a match, the more potential for penalties.

There were two weak-moderately strength negative correlations. The first, between PenFK Against Diff and Ruck Win Pct Diff (r=-0.40), suggests that as the percentage of rucks won increases, the penalties against that team decreases.

The second, between Ruck Win Pct Diff and RuckMaul Diff (r=-0.42), may indicate that as the number of rucks in a game increases, the percentage of rucks won decreases.
It's worth noting that three of the four strongest correlations are related to rucks and penalties. As previously mentioned, the ruck contest is often the most heavily-penalized area of the game, so these correlations align with that feature of the game.

Interestingly, there is no single variable that has even a moderately strong correlation with Score Difference (a.k.a, a win or loss). The strongest correlation is with Possession Time Difference (r=0.33), and the next strongest is Contestable Restart Win % (r=0.18).

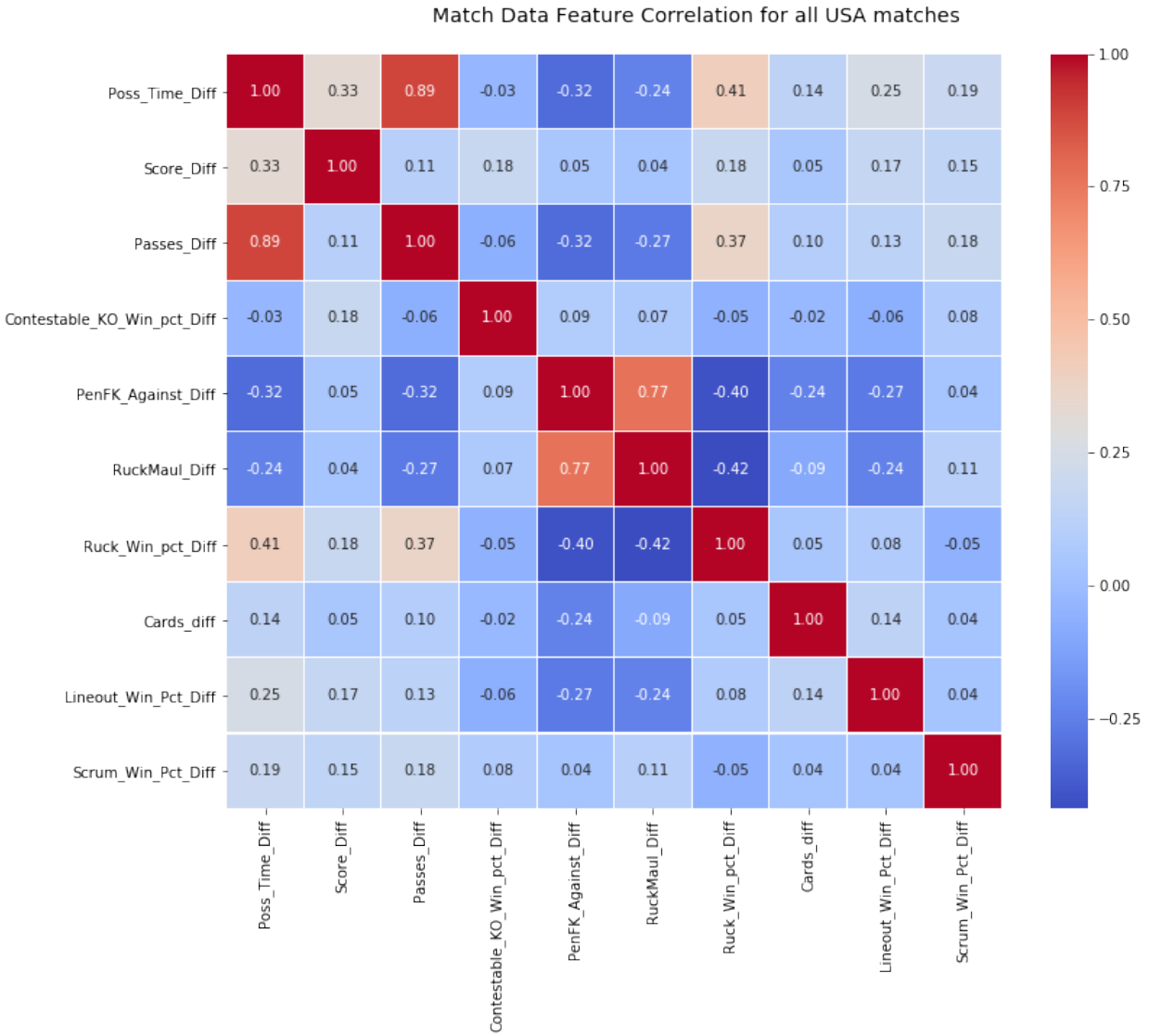

Figure 1 - Correlation Matrix

An area to explore further is examining the USA's performance in different features/variables against the top teams on the World Series circuit, to find areas of strengths or weakness against each of these top teams.

After EDA was completed, the data was pre-processed, before it was ready for model building.

Pre-Processing
Once the data was pre-processed and analyzed through EDA, several types of classification models were built to predict the result of the matches, and then extract the most important features from that model: Decision Tree, Logistic Regression, Random Forest, and an XGBoost (gradient boosted trees) model. Each model's performance was evaluated, and the best performing model was selected as the final model.

For each model, the same pre-processing steps were followed:

First, features that would bias the prediction, or that were non-numerical features were dropped. An example of a feature that would bias the prediction is the score differential ('Score_Diff'). It doesn't take much to imagine that if the model had the score differential as a feature, it wouldn't have much trouble predicting the outcome of the match! In addition to the binned Kickoff features, the following features were dropped: 'Opp', 'Score_Diff', 'Tries_Diff','Tournament', 'Conv_Diff'

Because the scale of each feature's values was so varied, the data needed to be scaled for better model performance, using the scikit-learn 'StandardScaler' library.

Model Building

Once the standard pre-processing was complete, each of the selected models was built, tuned, and evaluated for performance.

Train-Test Split
At the start of each model build, the data was split into training and test data, using the scikit-learn library 'train_test_split', which is part of the library's 'model_selection' library.

A 70% training and 30% test split was used, using the 'train_test_split' parameter 'test_size=0.3'. A random state seed parameter was used (random_state=77) to ensure reproducibility.

The train-test split resulted in counts* of:

Test = 46 records (30%)
Train = 105 records (70%)

**Recall that 5 matches that resulted in a Tie were dropped, prior to the train-test split.*

Validation Data
Additionally, after the model building was completed, the 2018-19 World Sevens Series started, providing an opportunity to utilize a set of validation data, on which to further test the predictive power of the model. At the time the project was completed, there were four tournaments played, so an additional 24 matches were added for validation.

Model Selection
As mentioned, four classification models were built, tuned and evaluated.

Each model was tuned using scikit-learn's GridSearchCV for finding the best hyperparameters.

Results

Model	Precision	Recall	F1 Score
Decision Tree	0.63	0.62	0.61
Logistic Regression	0.55	0.52	0.53/0.60*
XGBoost	0.50	0.50	0.49 / 0.50*
Random Forest	0.67	0.64	0.58/0.64*

Table 1 - Model Performance Scores

* sklearn.metrics f1_score

After evaluating the Precision, Recall and F1 score (using scikit-learn metrics.f1_score) of each model, the Random Forest model was chosen, given it's higher Precision, Recall and F1 score.

Interestingly, when run against an unseen validation data set from the most recent tournaments this year, the model performed even better than it did on the Test data.

Model	Precision	Recall	F1 Score
Random Forest	0.74	0.70	0.71/0.76*

Table 2 - Model Performance on Validation Data

Random Forest Model
After the Random Forest model was selected, further hyperparameter tuning was performed, using Randomized Search (from scikit-learn's RandomizedSearchCV) in addition to GridSearchCV.

First, Randomized Search was used to narrow down the parameter values. Then, the smaller range of hyperparameter values were used in GridSearch, to reduce the space that need

Decision Trees and the Random Forest Algorithm

The Random Forest algorithm is an ensemble method, based on decision trees. Decision trees take the input (X, or the features) and traverses down the "tree" of binary splits for each variable. Each node splits a group of observations according to a feature. The decision tree's

aim is to split the data into groupings so that every element in one group, or branch, belongs to the same classification.

Individually, decision trees may be “weak learners” – their predictive power is barely above chance. Ensemble algorithms like Random Forest use many instances of the weak learner, pooled together by bagging (bootstrap aggregation) the weak learners together (or boosting, in the case of XGBoost) to create a much stronger ensemble classifier.

A visualization of the decision tree for the 7s World Series data will help to illustrate how the decision tree works, splitting on each feature.

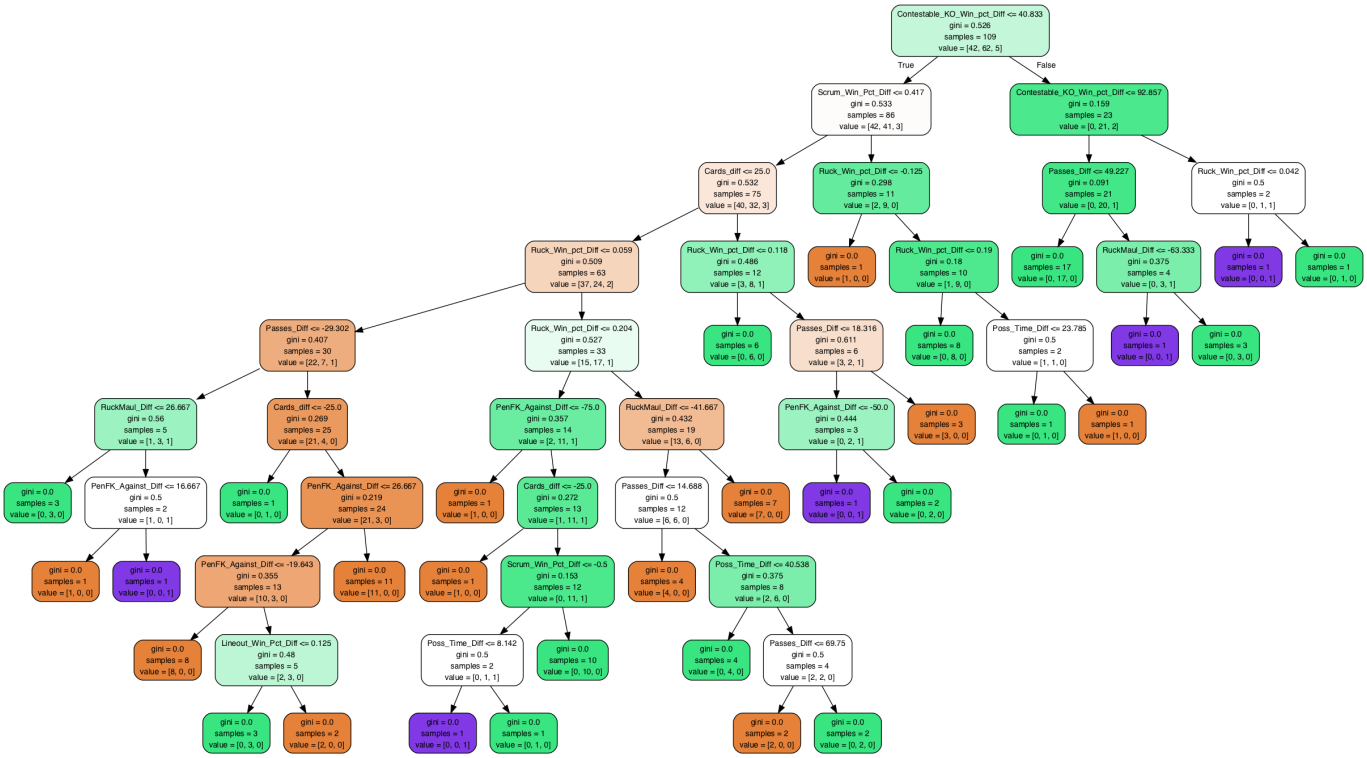

Figure 2 - Decision Tree for World Sevens Series 'Diff' Data

Evaluating the model performance scores, the single decision tree had an F1 score of 0.61. Using the Random Forest ensemble method to pool many weak decision trees together, we improved the F1 score to 0.71.

Feature Importance

If we take a step back to the objective of the project, recall that the objective is to not only predicting the outcome of a match, but also “identifying the most important variables influencing a team’s match outcome in Rugby 7s”.

In pursuit of that objective, a method was needed to identify the features that had the greatest influence on the predictive power of the model. Extrapolating this concept to the “real world”

insights to be delivered, the project's success is dependent upon providing actionable information that the client – coaching staff of a National 7s team – can use to inform their planning and preparation. As such, the important features are statistics representing areas of the game that have the greatest influence on the outcome (win/loss), on which coaching staff can focus and invest their resources.

The method used to identify these important features is the 'feature_importances_' attribute of the RandomForestClassifier algorithm, which provides us with the required method to extract the model's feature importance.

Node Impurity and Feature Importance

An important concept of decision tree-based algorithms, and their relation to feature importance, is the concept of node impurity. The measure on which the optimal split of a feature is chosen is referred to as impurity. The measure of impurity chosen for my Random Forest model is Gini impurity, which is "a measure of how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset." Classification models such as Random Forest can also use a measure of impurity called Entropy.
-- https://en.wikipedia.org/wiki/Decision_tree_learning#Gini_impurity

For example, in Figure 2, the example for the 'Contestable Kickoff Win Percentage Diff' node of the decision tree in Figure 1 has a gini impurity measure of 0.526.

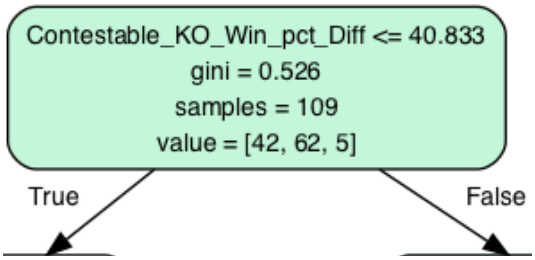

Figure 3 - Decision Tree Node for 'Contestable Kickoff Win Percentage Diff'

Feature importance, then, "is calculated as the decrease in node impurity weighted by the probability of reaching that node. The node probability can be calculated by the number of samples that reach the node, divided by the total number of samples."
- https://medium.com/@srnghn/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3

The result is intuitive - the higher the feature importance value, the more important the feature. We can extrapolate these concepts to say that the success of 'weak learners' depends on these important features for prediction.

Calling the feature_importances_ attribute of the Random Forest model, we were able to extract the list of Features and their feature importance.

Feature	Importance
Contestable_KO_Win_pct_Diff	0.190070
Poss_Time_Diff	0.184493
Ruck_Win_pct_Diff	0.150908
Passes_Diff	0.149123
PenFK_Against_Diff	0.105448
Lineout_Win_Pct_Diff	0.084172
RuckMaul_Diff	0.060571
Scrum_Win_Pct_Diff	0.052753
Cards_diff	0.022462

Table 3 - Feature Importances

Knowing the most important features of the model is a critical element of solving the "business problem" that the project was designed to solve.

However, if the insight is not actionable and useful for improving the team's performance, it is simply an intellectual exercise.

Actionable Insights

The Exploratory Data Analysis (EDA) and Data Storytelling reports delve deeper into the analysis of each feature contained in the feature importances, analyzing the top five features in the feature importance in order to focus on a manageable and actionable set of features.

The data analyzed in the Data Storytelling was from the 2017-18, the most recent season completed. Data referenced below is from the 2017-18, unless otherwise stated.

Contestable_KO_Win_pct_Diff

The 'Contestable_KO_Win_pct_Diff' feature represents the difference in the USA and their opponent in Contestable Restart Win Percentage, which is the proportion of USA kickoffs in which they caught and retained their own kickoff. The term 'Contestable' indicates that the kicking team was trying to contest the kickoff by kicking it just pas the required 10m, where they can compete to regain the kick. An Uncontestable kickoff would typically be a deep kickoff, where the objective is to pin the receiving team deep in their own end.

The descriptive statistics tell us that the overall mean Contestable Restart Win Percentage for 2017-18 matches the USA played is 33%. The mean Contestable Restart Win Percentage by the USA in wins is 52%, 33% in losses, and the mean Contestable Restart Win Percentage by the opposition across all USA matches is 22%.

The distribution of Contestable Restart Win Percentage in USA wins is more tightly clustered around the mean of 51.73%, between 33% and 66%.

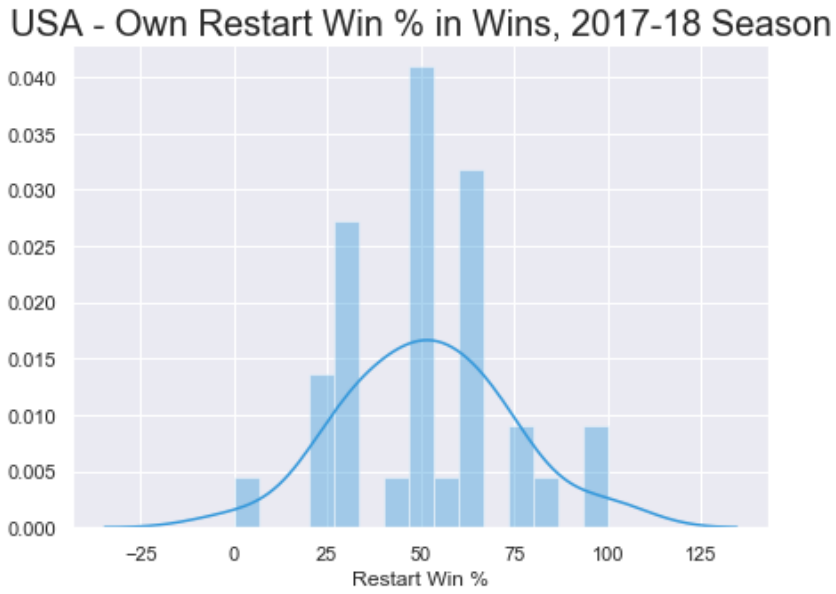

Figure 4 - USA Restart Win % in Wins

The distribution of Contestable Restart Win Percentage in USA losses skews towards the left (low) side of the distribution.

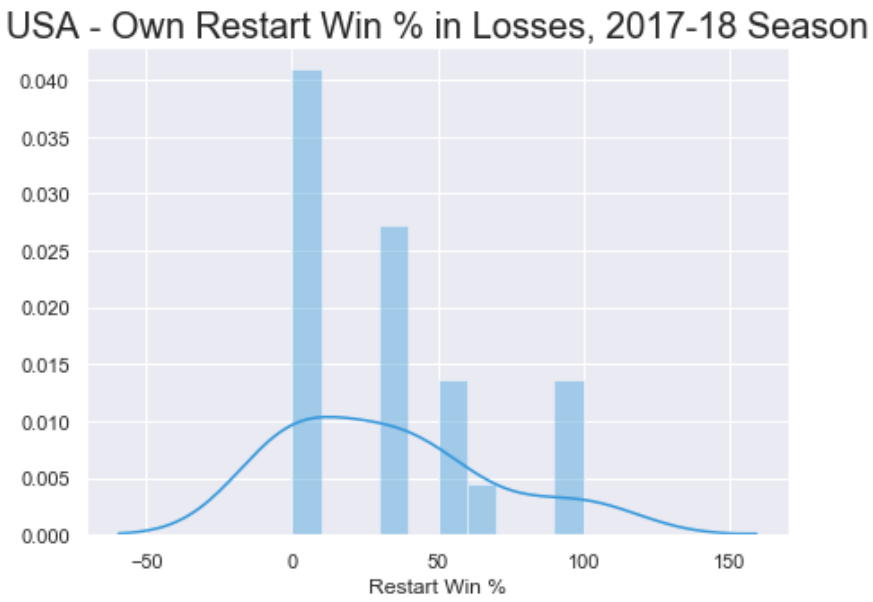

Figure 5 - USA Restart Win % in Losses

It appears that there is clearly a relationship between the USA winning their own restarts and winning matches.

The fact that the Contestable Restart Win Percentage appears to be a key factor in USA wins is not surprising, as winning their own restarts has become a hallmark of the USA game, often leading to higher time of possession per match. Kickoffs/restarts are how the game is restarted after a score, with the scoring team kicking. So, if a team has possession, scores, kicks off and regains their own kick, they will typically have long periods of possession, often translating into additional points.

While the relationship between winning their own restarts and winning matches may not be surprising, it may serve as solid confirmation of what was previously anecdotal evidence.

Steps that can be taken by the coaching staff to capitalize on these findings would be to allocate more training time to kickers as well as the "jumpers" who contest and retrieve the ball in the air on kickoffs. Additionally, recruitment should be focused on recruiting players who have the required kicking skills or skill potential, and players who fit the profile of successful jumpers.

Poss_Time_Diff and Passes_Diff

A strong correlation (r = .89) was found between the features 'Poss Time Diff' and 'Passes Diff', two of the top five features in importance.

The jointplot below shows the strength and direction of the relationship between the two features.

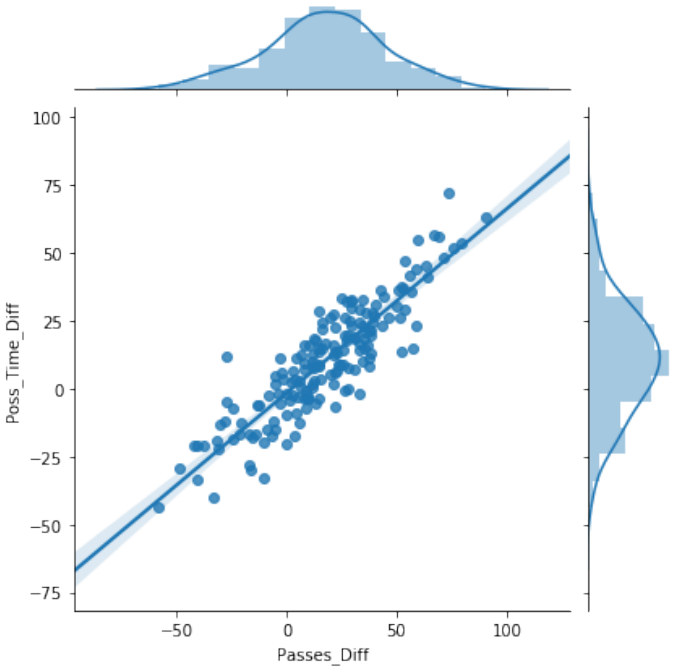

Figure 6 - Relationship between 'Poss Time Diff ' and 'Passes Diff'

To an observer with any familiarity with the game, the correlation between Possession Time and Passes makes intuitive sense - the longer a team has the ball, the more passes they are likely to make. One way a team can retain possession is by passing the ball to change the point of attack away from potential tacklers and defensive pressure.

When one observes the pattern of play the USA employs, it becomes clear why Possession Time, Passes and Ruck Win Percentage ('**Ruck_Win_pct_Diff**') are important features in predicting a win or loss for the USA. The USA plays a patient game, passing the ball from sideline to sideline, winning their rucks to retain possession until the defense is broken down, stretched out of their shape or a favorable matchup is created, and a break is made for a score.

In the USA's observed pattern of play, the team is successful when all of these features work together. More passes = more possession, and ball retention at the tackle (Ruck Win Percentage) ensures continued possession. And continued possession eventually is turned into points. When combined with winning their own restarts, their possession virtually starves the opposition of the ball, and as they saying goes, "you can't score if you don't have the ball".

Actionable insights in this area are twofold. First, the importance of these features and their relationship to each other can serve as data-based confirmation that their game plan is indeed effective and leads to wins when executed. Secondly, a training focus on the elements of the game that must be executed in this game plan – long and accurate passing, contact skills at the tackle, and rucking/cleanout skills to retain possession – will support this game plan.

One caveat that must be stated is that possession alone is not enough. The correlation between 'Poss_Time_Diff' and 'Score_Diff' (a proxy for a win) is r = 0.41, which is a moderately strength correlation, at best.

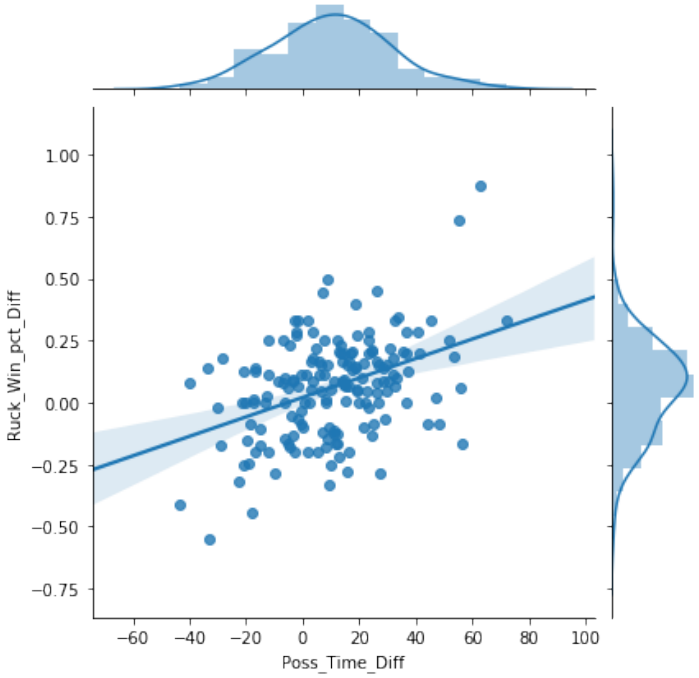

Figure 7 - Jointplot correlation of 'Poss_Time_Diff' and 'Score Diff

PenFK_Against_Diff

The last feature in the top five feature importances is 'PenFK_Against_Diff', which is the difference between teams in the Penalties or Free Kicks against them.

The obvious takeaway here is that discipline is another key component to a winning formula for the USA. However, considering that the majority of penalties in a rugby 7s match occur at the ruck, this feature also ties into the Passes, Possession, Ruck Win Percentage relationship. A strong focus on contact area and ruck technique will improve the Penalties or Free Kicks against the team, but will also improve the length of possession.

How these features all work together to impact the outcome of a match can best be illustrated by a snapshot of one match, against the reigning Olympic Gold Medalists Fiji, in the Cup Quarterfinal of the Dubai 7s in the 2018-19 Series.

In this match, the USA held the ball for 5:15 to Fiji's 2:12, made 63 passes to Fiji's 25, won 75% (3/4) of their own kickoffs, won 100% of their own rucks, and conceded only one penalty/free kick, en route to a 24-14 win. To further illustrate the point, through regaining their own kickoffs and holding possession through a patient ball-control attack, Fiji did not lay their hands on the ball in the second half until there was 45 seconds remaining in the game. That is "starving your opponent of possession" at it's finest.

Future Work

As mentioned earlier in "Additional Data Sets Considered", future work should consider the Player/Roster data and data on Tournament Location and Scheduling.

Exploring the relationship between the consistency of the team's playing roster and the impact on the team's performance across different matches or different tournaments. Additionally, adding match data for individual players scoring of individual players and their contribution or impact to the team, including the impact of not having top players in the team (due to injury, for example) would be an excellent line of further inquiry.

Analyzing the distance a team has to travel to the tournament and it's impact on how a team performs in both the first and second leg of a tournament would add previously unexplored relationships to the analysis.

Conclusion

In the course of this analysis, it has become clear that, while feature importance is a crucial tool in model evaluation, a model with 76% accuracy is not necessarily a reliable predictor of a match's outcome - not in individual matches, nor in macro-level metrics for the entire season.

The relationship between Passes, Possession, Ruck Win Percentage, and Contestable Restart Win Percentage does appear to be a strong indicator for the USA when looking at macro-level performance across a Series. But it does appear that the game of Rugby 7s is too volatile to be able to develop strong and accurate predictors at the match level, at least using the feature set that were used in this model.

This volatility is best illustrated in the USA's recent loss to New Zealand in Dubai, where they outperformed New Zealand in nearly every metric, but were still beaten soundly.

Preparing a team for the Sevens World Series cannot be done with data and statistics alone, as they only tell part for the story and offer no context. However, data and statistics can be used to analyze how a team matches up against opponents in certain areas of the game. Then, using this data analysis as a starting point to drill down into further video analysis inquiry into tactics and technique, can provide a tactical advantage.

When predictive models are used in tandem with video analysis to provide context, they can combine to form a powerful tool, and is an area that is undoubtedly beginning to be utilized by top teams looking for an advantage.