The Economic Impact of the World Cup

Jonah Witt, Michael Morales, and Taylor Moorman

IST 707

March 17, 2019

**Introduction**

The first team sport to be included in the Olympic games was soccer, in 1900. While the first two Olympic soccer tournaments were merely demonstrations with no medals awarded, the popularity grew so that by 1908 the International Olympic Committee (IOC) had conferred official status to the sport, with medals awarded to the top three teams. As with all Olympic events, the IOC established rules governing what constituted amateur status, and it was not long before controversy arose. As the 1928 games came to a close, a disagreement was forming between soccer’s international governing body, the Fédération Internationale de Football Association (known globally as FIFA), and the IOC. At issue was amateur status, which was the catalyst for FIFA to attempt to capitalize on the recent success of the Olympic tournaments.

The first FIFA World Cup was held in Montevideo, Uruguay in 1930. As Uruguay was celebrating its 100th birthday that year and had taken gold in the previous two Olympic tournaments, it was the ideal host country for the new international tournament. In keeping with Olympic tradition, FIFA designated that the tournament should be held every four years, and would decide on host countries by vote of the committee according, in part, to the host country’s fiscal ability to host. Fiscal concerns were at the forefront of many decisions from the very beginning. The early tournaments did not pay prize money, with potential profits shared among the associations and deficits being absorbed by the host country should profits fail to materialize. Participating national football associations played only for the glory of victory and a trophy named for FIFA’s director from 1921-1954, Jules Rimet.

World War II brought with it cancellations of both the World Cup and Olympic games as war efforts around the world consumed more resources than could be produced. This time was difficult for FIFA, as the association was subject to serious financial difficulties as a result of the war. Fortunately, pre-war problems with participation in the World Cup by European associations were resolved once the war ended, and money began to flow again. The next two tournaments were held in Europe to great fanfare and participation, ensuring the future of the “beautiful game’s” international contest.

**Analysis and Models**

*About the Data*

As with most sports, FIFA understood the inherent value of keeping game statistics long before analytics were ever part of the game. To the benefit of lovers of the game and its history, data are available for every match of every World Cup tournament played since the inception of the quadrennial contest in 1930. In order to try to understand whether there is value in analyzing individual matches, these data were included in this analysis. For the portion of this project concerned more with the financial impact of the World Cup, a much smaller dataset with information about each tournament was much more pertinent. This dataset, called World Cups, outlines host countries, winners and runners-up, some match statistics, total attendance, number of stadiums, and finances. Several other financial data were included as points of interest along the journey of this analysis, as total financial data is not nearly as complete or available as match data.

For the World Cups data, it is important to note that grossed-up numbers were added to normalize cost and revenue to 2018 values, as the most recent World Cup tournament took place in the summer of 2018. There are missing values in the dataset, as no prize money was awarded prior to 1982. For cost and revenue figures prior to 1986, averages of inflation-adjusted numbers were used for the 2018 normalized figures, then backward adjusted for each year. These figures were not available, but for the sake of comparison it was necessary to establish a benchmark.

Nonetheless, it is difficult even in the best circumstances to compare the cost of a tournament to host countries, as some countries have established infrastructure due to the popularity of soccer and large populations concentrated densely in major metropolitan areas, while other nations must build stadiums and other facilities necessary to host matches. It is equally important to note that the number of stadiums wherein the host country chooses to field teams is a major cost concern. The first World Cup was played entirely in the city of Montevideo in 2 stadiums, whereas the 2002 contest saw matches in 20 cities across South Korea and Japan. While the choice to include as many cities as possible spreads potential profit and an increase in soccer’s popularity across a much greater range, it may also increase cost since national associations and their support staff must be transported long distances and provided with hospitality. These considerations must be taken into account when financial comparisons are being made.

The matches data have been transformed in a few ways. Since there was no prize money officially awarded by FIFA prior to 1982, the match data from 1930 through 1978 were not included. Since one of the goals of the analysis is to assess economic and financial concerns, the matches that have a direct financial impact are the most important to consider. Other variables that were not important to the analysis were also deleted from the dataset. One of the variables included in the set was the nationality of each referee and assistant. This was removed. The names of officials and assistants were also converted to word type, since they were imported from the original Excel file as a generic data type. Some columns were incomplete. One of the columns has data about win conditions such as whether a win happened after full time has expired. Since this information only applies to wins that took place after the expiration of normal game time, it is both incomplete and unnecessary to the analysis. This information was removed.

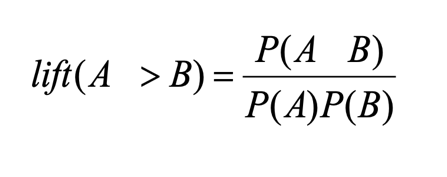
*Association Rule Analysis*

To attempt to gain information on the variables, it is important to look at the associations between them. This is accomplished using association rule analysis. Association rule analysis is used for discovering relationships in large data sets. In this case, an association rule is a set of attributes that frequently occur together. These rules can consist of two or more attributes. Each rule will have two sides (left-hand side and right-hand side). The association rule shows the relationship between the attributes on the left-hand side and the attribute on the right-hand side. These association rules are measured based on the following:

*Support*: How often the rule occurs in the entire data set (percentage).

*Confidence*: How often the value in the right-hand side appears in the data that includes the value in the right-hand side (Value from 0 to 1).

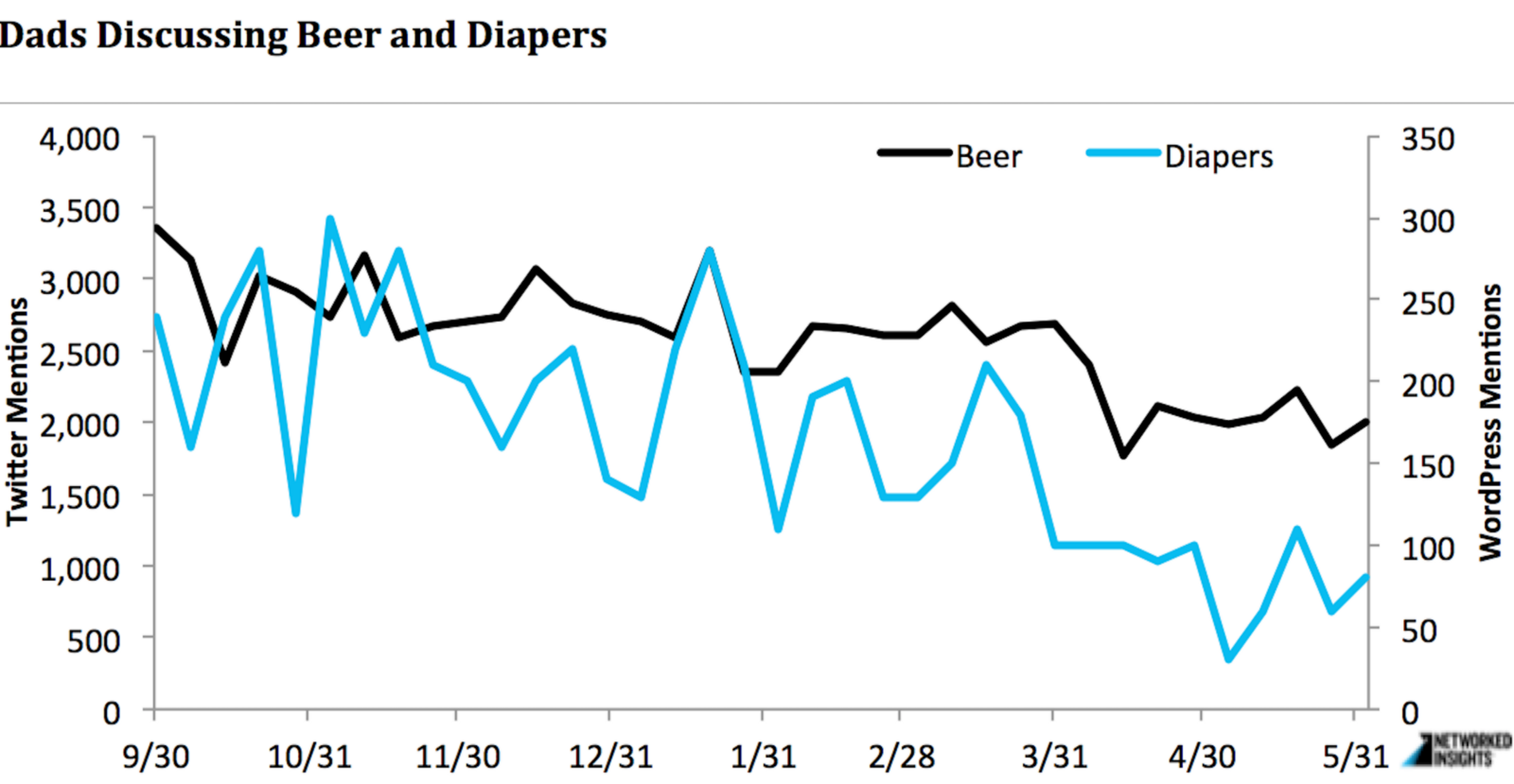
*Lift*: The ratio of the confidence of the rule against the expected confidence of the rule. The formula for lift is shown below:



A rule must have a value greater than one to show that the rule has any meaning. Support,

confidence, and lift must all be considered when evaluating an association rule.

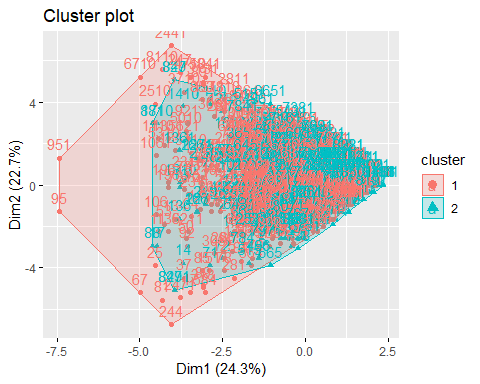
For the World Cup analysis, many associations were evident. Some are obvious, such as the association between winning the overall tournament and placing in the top three, or accomplishing neither. The match data were used in this project to attempt to associate such rules as winning with being the home team (which is surprisingly common). Overall, association rule mining was not very useful for this analysis. Association rules make the most sense in a marketing setting where retailers want to know which items in their inventory (or service offering) are frequently purchased together. Beer and diapers are famously associated, and the relationship between them even pops up online in social networking interactions, as seen below.



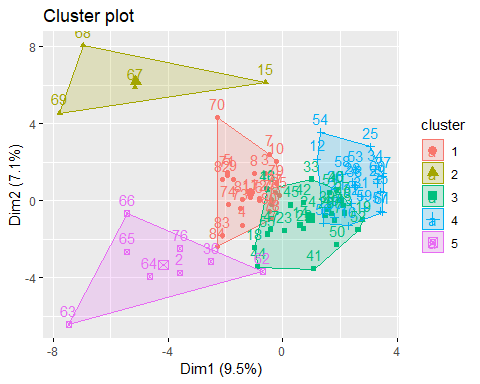
For retailers who want to know what products are frequently consumed during World Cup tournaments, association rule mining can produce results that will help them market successfully to a very specific set of consumers, but this is not useful for this analysis.

*k-Means Clustering*

K-Means clustering is a method used to group like data into “clusters” where the data points are most similar. The data are divided into a specified number (k) of clusters. There are k data points selected randomly (unless otherwise specified) as cluster centers. Afterwards, the remaining data points are grouped into the cluster where the point’s mean value is closest to the center of the cluster. If conducted properly, this will group the data points into clusters with similar characteristics.



The match data produced the cluster above, which has a large amount of overlap. It was not particularly helpful in this analysis, as it did not reveal new information that was helpful in predicting wins. The clusters that are most useful are the kind that separate features into distinct categories, allowing an analyst to see the distance between them.



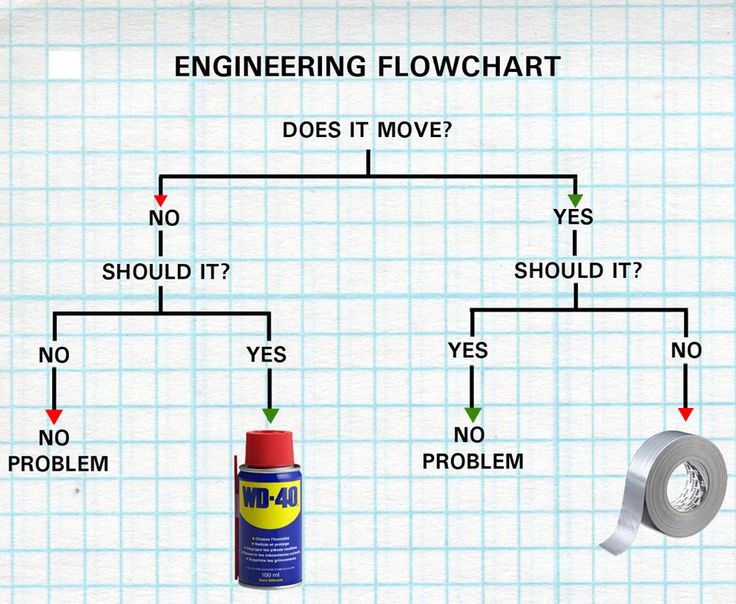
The cluster above was used to come to a determination of who authored several documents of disputed authorship. By clustering the authors according to features evident in their works, it was easy to see which writers the unknown documents clustered with. Again, this was not a useful method for predicting match wins in the World Cup.

*Decision Tree Analysis*

Decision Tree Analysis is a method of classifying data by splitting it into different label groups based on the values of various features. The data is broken off into “branches” from the main data set. These branches are determined by whether the data point meets certain criteria for a particular feature. For example, these decision nodes could represent whether the data record represents a male or female, a child or an adult, or any other split based on a single feature. After the data are split, the data points reach the “terminal node” which classifies the data point based on the label in question. These trees can be tuned using various control measures. The two measures that were used to tune the trees in this analysis were minimum size of each split and the complexity parameter. The minimum size of each split indicates the minimum number of data points that must be in each “leaf” of the tree. The complexity parameter indicates how much the data point must improve the fit of the tree for the split to be included.



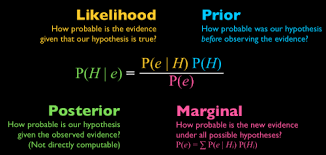
The image above is one of the decision trees for this analysis. It is complex and too small to really see in a way that helps. This one was used to test winning the World Cup against all the other variables. The figure below allows a much simpler and more applicable example of how decision trees work at their simplest.



Ultimately, decision trees act somewhat like flowcharts. Decision trees are not powerful, *per se*. They are powerful in their persistence. They take decisions one-by-one, working in an intuitive way to find the best conclusion. They work quickly and with efficiency to arrive at a prediction, and they can be “pruned” for even better results. In this analysis, pruned decision trees produced 94% accuracy in predicting the tournament winner, and 86% accuracy in predicting which teams would place.

*Naïve Bayes*

Naïve Bayes is a method of classification based on Bayes Theorem. This theorem assumes that each predictor in a dataset is independent of the others. The model is considered “naive” because of this assumption. Even if it is well known that the predictors are related to one another, the model still treats every predictor as entirely independent. These predictors are then used to predict the posterior probability. The assumption of independence allows the model to be run very quickly since it does not have to calculate the probability for every single combination of predictors. However, this speed is a tradeoff for a more accurate model that takes the relationships of predictors into account. Naïve Bayes works using the equation below.

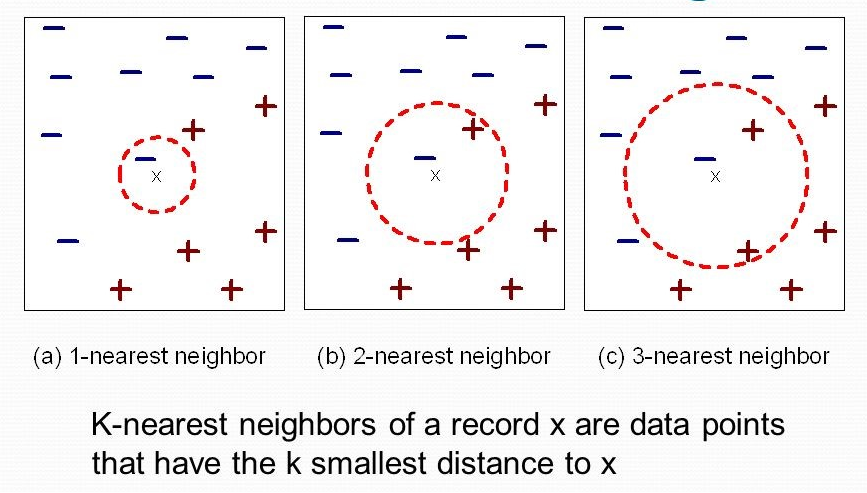


In epidemiology, Naïve Bayes is very useful. Finding the probabilities that a person has a disease that is spreading quickly can help in quarantining efforts, which can help to stop the epidemic. It can also be helpful in determining genetic risk scores for prediction of congenital diseases or cancer traits.

Naïve Bayes was employed in this analysis on the match data to attempt to predict which teams would win, as well as which teams would place. The win prediction accuracy came close to 90%, and the place prediction accuracy was just over 81%.

*k-Nearest Neighbor*

The fifth model employed was the KNN algorithm. KNN stands for K-Nearest Neighbors. It works in a lazy way, which is to say that it defers computation until the classification stage of the process, only approximating prior to reaching that stage. Another way to express this is to say that it does not train on the data, saving this step until it is ready to make predictions. Other algorithms train themselves on the training data, and then apply the training to the prediction portion of the analysis.

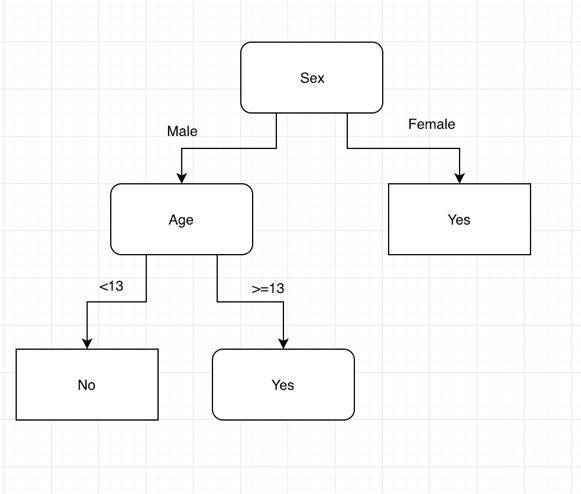


KNN works by finding the nearest neighbors of each classification and weighting them according to their distance to x, an unclassed variable. X can only belong to one class or the other, so the KNN algorithm will draw a circle that includes x and one or more of the other classes to determine how many “votes” are cast for each class. In the example above, the circle only includes the blue negative at first, then includes a vote each for the blue negative and red positive as it expands to the second level. When it reaches the third level of expansion, it includes two votes for red positive and one for blue negative. While one blue negative is closest to x, it seems that further expansion will class it as red negative, although that may not be true. This is KNN in a nutshell. It is not always perfect, but works well when the data are easy to classify.

This method is supervised, which means the label of the training data is known, and the model is looking to train the on the other features to classify the data on that label. A larger k value can reduce the effects of noisy data. However, using a k value that is too large can diminish the effects of smaller patterns in the data. Ideally, KNN is used for domains like recommendation engines, search engines, and anomaly detection. KNN was able to predict wins at over 86% accuracy and placing at over 69% accuracy.

*Random Forest*

A random forest is a collection of randomly generated decision trees. However, these trees are not generated in the same manner as in a Decision Tree analysis. When creating a single decision tree, the tree is created using all of the data and variables and then pruned to generate the best tree. A simple decision tree on a data set of Titanic Survivors is shown below.



In decision tree analysis, each tree is created using a randomly selected group of data and variables. The output from these trees is combined to create the final output of the model. Each tree in the forest gets a “vote” to classify the data point. The label with the highest percentage of votes will classify the data point. This random method leads to a less biased result and typically performs better than a single decision tree. A simplified example from an article on Medium.com is shown below (Koehrsen 2017).



RF is able to use features of the variables to differentiate them from each other. In a forest, one may observe Douglas firs, spruce, and pine growing together in a grove. It would be enormously difficult to differentiate the trees without entering the forest and looking closely at the features of each tree. Looking at bark patterns, needles, and cones would give a forester the necessary information to classify the trees, information that could potentially be extrapolated out to the forest as a whole.



Using the RF algorithm, the analysis can be tuned by generating features that differentiate variables, giving the model greater power to predict the target variable. For the World Cup match data, RF predicted wins with an error rate of only 5.66%. It was less successful in predicting which teams would place, with an error rate of 12.65%.

*Support Vector Machine*

The Support Vector Machine (SVM) is a supervised machine learning method used to classify binary data. This means that the SVM is used to split data into two groups with either “Label A” or “Label B”. This is achieved by creating a plane to separate the data. For data that is linearly separable, this plane is simply a line that maximizes the distance between both groups (the line that is evenly between the two groups).



While this seems simple for data that is linearly separable, this method becomes much more complicated when the data cannot be separated this way. In this case, the data will need to be transformed into a higher dimension. For example, consider the following example where the data appear to be separated by a circle.



In this example, the third-dimension *z* can be defined as z = x2 + y2 (This is the equation for a circle). A slice of this three-dimensional data can be shown below.



The linear plane can now be created to separate the data. When transformed back to the two-dimensional data, the plane of separation is shown below.



However, the method to define the third dimension is not often obvious. To accommodate this, different kernel functions can be used to train the model. These kernel functions can be linear, radial, polynomial, etc.



Ultimately, the key to SVM is maximizing margin between similar features in the data. This is important so that they are not confused for each other. In the photo above, the man on the left is an Iranian student named Reza Parastesh. The man on the right is Argentinian soccer superstar Leonel Messi. Parastesh has been threatened with arrest for disturbing the peace because of his astonishing resemblance to Messi. If Parastesh appeared at a World Cup venue during the tournament, he would likely be mobbed by crowds hoping for a selfie with Argentina’s famous footballer. SVM would take two such features and maximize margin so they could be clearly distinguished. Metaphorically, SVM would remove Parastesh’s jersey and beard so he could not be confused with Messi.

**Results**

*Association Rule Analysis*

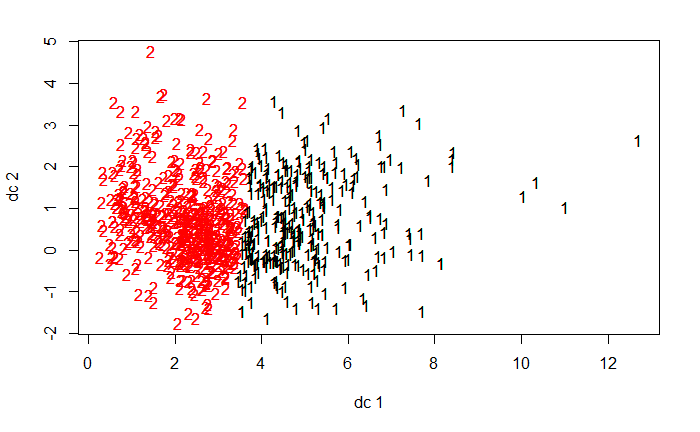
The association rule analysis produced a set of 114 rules. While this was not an ideal model for this analysis, it did reveal some associations which were interesting.



As seen in the figure above, there is an obvious relationship between winning the tournament and placing, as well as winning and number of goals against. But there also seems to be some correlation between winning and being the home team. This is interesting because home team advantage should not be an actual advantage in World Cup soccer. The only true home team is the host country’s own soccer association, yet the data suggests being the home team is a factor in wins. In fact, not a single “visiting” team won any match in the first three World Cups. Other than revealing some items of interest, association rule mining did not produce usable results for this project.

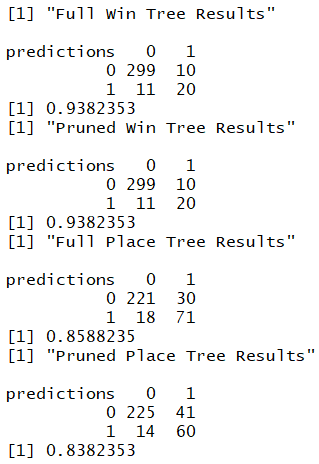
*k-Means Clustering*

The results of the k-Means clustering were equally useless in a real-world application for this analysis. K-Means simply did not reveal anything of use. While the figure below, which clustered the entire k-Means-specific data frame against the results of the clustering model, showed distinct clusters, it did not reveal anything novel or useful. Furthermore, the model required the data to be transformed even further than the initial data-cleaning portion of the project required. k-Means requires numeric values only as opposed to factors, integers, or characters.



*Decision Tree Analysis*

The decision tree model was much more successful than the others used so far. While pruning did not improve the tournament win prediction, it was effective in improving tournament place prediction. The data were randomly split 80% for training and 20% for testing.



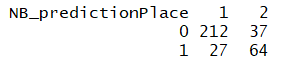
Predicting wins will always have high percentages since there have only ever been 8 winning national teams. This is likely why the pruning did not help the win prediction. Data on World Cup wins is limited. The limits are not a result of the unavailability of data, but the small number of World Cup tournaments (21). Pruning did increase the prediction of whether teams would place in the top 4 to over 85%.

*Naïve Bayes*

Naïve Bayes was able to predict wins with an accuracy of nearly 90%, as seen in the confusion matrix below. Again, this prediction is probably not very informed, as there is a dearth of available observations where wins are concerned. An 80/20 train/test split of the data was used.



The accuracy for predicting whether teams would place in the top 4 in the World Cup was a bit more than 81% using Naïve Bayes. Because many more teams place than win, Naïve Bayes has to work harder at this prediction, but the results are likely more meaningful than the win predictions.



*k-Nearest Neighbor*

KNN predicted wins with an accuracy of nearly 87%. This result came after transforming all the data in the data frame to numeric type. As previously stated, KNN is ideally used for domains like recommendation engines, search engines, and anomaly detection. It was not the best model for this analysis, and it seemed all the supervised models performed similarly, so this was not a surprise.

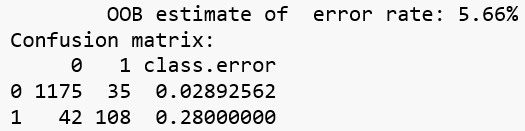


KNN place prediction accuracy came in at over 69%. As with the other models, an 80/20 train/test split was employed to arrive at this result.



*Random Forest*

For the Random Forest model, data was split 80% for training and 20% for testing using a random sample method. 500 trees were built for each prediction model, and a classification model was used. The other option would have been a linear regression model, which would not be ideal for this analysis. The win prediction performed as expected, with a 5.66% error rate. Again, this is a small set of data, so the results must be considered from that perspective.



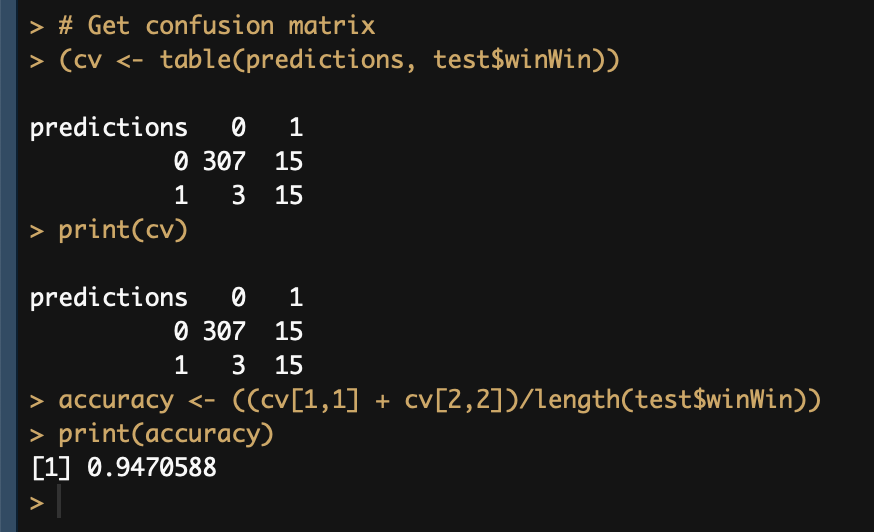
The place prediction also performed as expected, with an error rate of 12.65%.



When these models were tested against the known results in the test data, they returned the best results thus far, with a win prediction accuracy of more than 95%, and a place prediction accuracy of 87.6%. These results came from an analysis of 340 predictions.

*Support Vector Machine*

The SVM model was constructed in the same manner as the rest of the models, with the data split randomly into 80% training data and 20% testing data. Similar to the Random Forest, it predicted wins at around 95% accuracy. It predicted whether teams would place with an accuracy of around 80%.

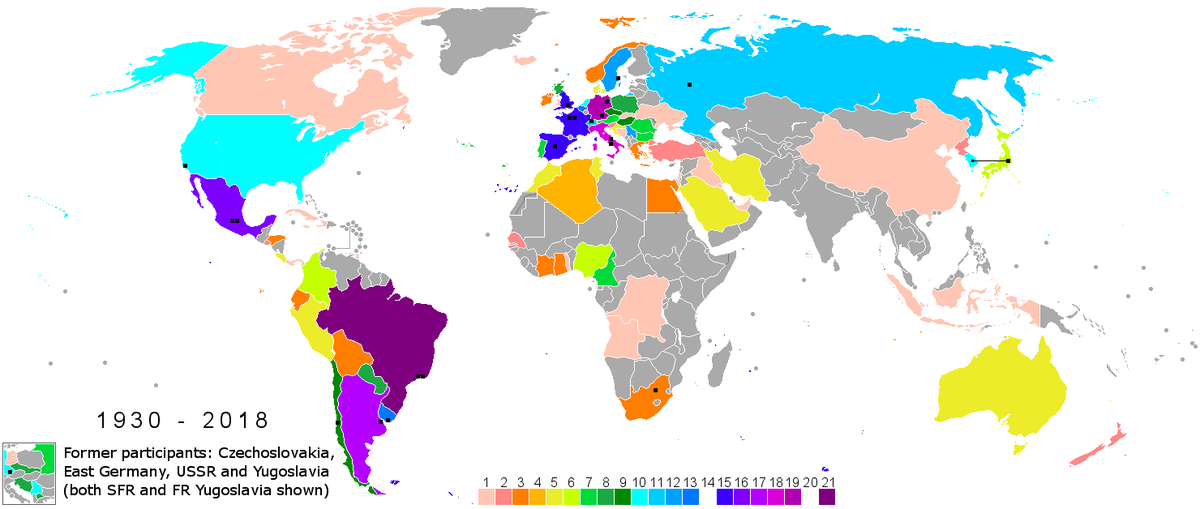


**Conclusions**

As the World Cup is only held every four years, the amount of data available is not very large. Since tournaments began in 1930 (and with two of them cancelled due to war), there have only been 21 of them. Only 8 teams have ever won, and Brazil is the only team to participate in every tournament since the inception of the World Cup.



What does this mean? It means that predicting the winner is too easy for complex machine learning algorithms. Of the 79 national teams that have ever participated in the World Cup, only around 10% have won it all. While there are many potential observations available, predicting wins depends on tiny portions of the data. The predictions for whether teams will place are more reliable, but they are also based on a rather small data set.



While the original intent of this analysis was to report on the financial benefits of the World Cup, research suggests that there is very little financial benefit to the host nation. In fact, hosting seems to more frequently result in great financial losses, as stadiums built at enormous cost sit latent or are sold on to the private sector. National teams can benefit from the very large prizes awarded to the top 4 placing teams, provided they are not the national teams of the host nation. That said, the best-case scenario for the host nation is to also win the tournament, as the prize money will then mitigate losses.



What the models have revealed is that predictions can be made, so long as the user is aware that the small data sets are going to impact the predictions. This caveat will help them interpret the models from a place of knowledge. In any case, the focus for countries should be winning or placing in the tournament, as opposed to hosting it.

**References**

Koehrsen, W., & Koehrsen, W. (2017, December 27). Random Forest Simple Explanation – Will Koehrsen – Medium. Retrieved March 5, 2019, from

https://medium.com/@williamkoehrsen/random-forest-simple-explanation-

377895a60d2d