An Econometric and Machine Learning Investigation: Examining Factors of Success in Cricket World Cups

Neil Menghani

ECON W4911: Sports Economics
Professor Sunil Gulati
22 December 2018

Abstract

With 2.5 billion fans around the world, cricket is second only to soccer in international popularity. Every four years, the world's premier cricketing nations come together to compete in the Cricket World Cup. Each nation participates in one or two group stages followed by a knockout stage, and each match follows the One Day International format. Most of the competing nations live and breathe cricket, and thus success in this tournament provides a great deal of nationalistic pride.

This paper will consist of an econometric analysis of the factors that play into success in this tournament, based on historical data from all 11 previous world cups from 1975-2015. The factors to be analyzed include weather, home field advantage, rest, historical performance, coaching experience, and makeup of squads (experience, chemistry, and skill).

The first stage of analysis will be to build a logistic regression model to quantify the impact of these factors on success in world cup matchups, which is best framed as a classification question. After this interpretable model is built and factors that play a significant role in match outcomes are determined, an appropriate statistical machine learning model will be built using different portions of the historical data as training, validation, and test sets. Once a reasonable accuracy is achieved on the validation set, the final objective will be to use the predictive model and the equivalent input variables for the upcoming 2019 Cricket World Cup in order to predict the likelihood of each team's success in 2019.

1. Introduction

Cricket is an exciting sport that garners a great deal of interest around the world. The game is played on a large circular ground between a home team and an away team, each with 11 players, and the primary objective is to score more runs than the opposing team. A match consists of either up to four innings in which one team bats – with a limit of 10 wickets and sometimes a limited number of balls – depending on the format being played. The form of cricket I will study in this paper is One Day International (ODI) cricket, as this has been the format played in the Cricket World Cup since the first of its kind in 1975.

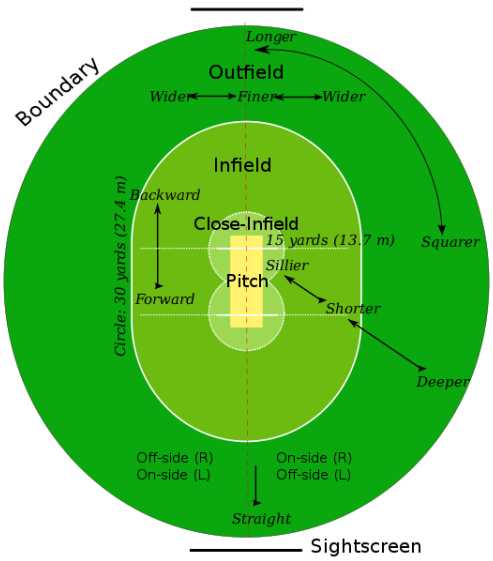

Figure 1: A cricket ground ("Cricket Rules," *Cricket Rules*).

One Day International Cricket matches consist of only two innings – each team batting once– and teams flip a coin to determine who can choose whether to bat first. The team that bats first has 50 overs (6 balls each) or 10 wickets (outs) – whichever of the two is reached first – to score as many runs as possible. Each over consists of 6 balls, and an individual bowler cannot bowl two consecutive overs or more than one-fifth of the overs in an inning (in this case 10 overs). After the first team bats, the number of runs they achieved is referred to as the target score. The second team to bat must attempt to score more runs than the target score. The match is over when the second team either surpasses the target score or reaches the over or wicket limit.

Runs can be scored in three primary ways: the batter hits the ball in the infield or outfield and runs to the opposing end of the pitch (and potentially back and forth) to score one run at a time, the batter hits the ball such that it reaches the boundary of the outfield to score four runs, or the batter hits the ball over the boundary without touching anywhere inside the field ("Scoring Runs in Cricket," *Wikipedia*). Losing a wicket, or getting out, essentially occurs when one of the two sets of wickets (consisting of three stumps

and two bails) at either end of the pitch is knocked down. A wicket usually occurs in one of the six most common ways: the bowler hits a legitimate ball that is then caught by a fielder or the bowler (caught); the bowler puts down the wicket directly by delivering a ball (bowled); a delivered ball would have struck the wicket, but the umpire rules that it was intercepted by any part of the batsman's body except the hand, preventing it from hitting the wicket (leg before wicket); the wicket is put down while a batter is running and has not reached the crease in front of the wicket (run out); the wicket-keeper standing behind the wicket puts it down when the batter has strayed too far in front of the crease (stumped); or the bowler has begun to deliver a ball, and the batter puts the wicket down with his bat or his person (hit wicket) ("Dismissal (Cricket)," *Wikipedia*). The four other less common ways of getting out include when the batsman intentionally obstructs the fielding side by word or action (obstructing the field); a batsman intentionally takes more than three minutes to be ready for the next ball (timed out); the batsman intentionally hits the ball a second time with any part of his body (hit the ball twice); or a batsman leaves the field without the consent of the umpire and opposing captain (retired out) ("Dismissal").

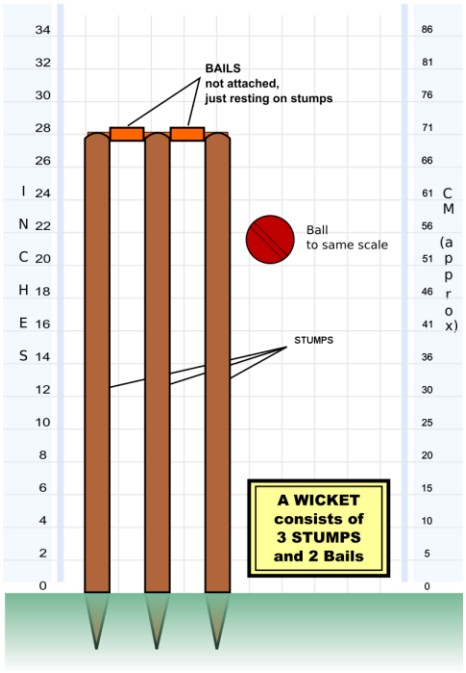

Figure 2: Diagram of a wicket (*Wikipedia*).

In some rare circumstances, fewer than 50 overs can be bowled and/or a tie can occur. A tie can occur if the over or wicket limit is reached by the second team and the number of runs scored by both teams is the same. A draw can also occur if weather or light permanently interrupts the match and play must be stopped (referred to as no result). If, however, weather interrupts the match and play may continue, the team batting second may have an adjusted target score to reach. A mathematical formula called the Duckworth-Lewis-Stern (DLS) method determines how many runs need to be achieved if fewer overs are bowled. This calculation is not as simple as taking a direct proportion on the number of overs and runs, since teams tend to bat more aggressively with fewer

overs left. Instead, the number of overs lost, the current stage of innings, and wickets remaining are all considered when setting the target score (Rajesh, 2017).

In addition to ODI Cricket, there are two other major forms of cricket played at the international level. Test cricket, which gets it name from the mentally and physically demanding nature of the format, consists of four innings (each team bats twice) and no over limit. Many consider Test cricket to be the best measure of a team's ability and endurance, as Test matches can last up to 5 days ("Test Cricket," *Wikipedia*). While this format is considered by many to be the purest form of cricket, spectators can easily lose interest due to its tedious nature. The other major format, Twenty20 (T20) cricket, is quite similar to the ODI format except that only 20 overs are bowled in each inning. Because such few overs are bowled with the same number of wickets allowed, players naturally play much more aggressively, going for big hits to score more runs rather than attempting to conserve wickets (Pobjie, 2014). While this makes the match more exciting to watch for newer, younger, and less invested fans, there also tends to be more luck and less pure cricket skill involved in this format, as players can more easily take a wicket or have a wicket taken through luck (Pobjie, 2014). One Day International strikes the balance between the above two formats in that the action taking place in a single day makes following along more manageable for newer fans, and the 50 overs requires utilization of pace and finesse by bowlers and batters (Pobjie, 2014).

2. Motivation

I have chosen to study factors of success in cricket for two main reasons: there are significant economic and nationalistic implications of determining how countries can improve their performance in the Cricket World Cup, and there has not been a great deal of analysis of in-depth analysis of factors of success in ODI cricket. With 2.5 billion estimated fans, cricket is the second most popular around the world. According to the largest ever market research project into cricket, conducted by the ICC, T20 cricket is the most popular format at 92% interest among fans with ODI a close second at 88% interest (*Rediff*, 2018). With two thirds of fans over 16 years old being interested in all three international cricket formats, the ICC

chief executive David Richardson believes that cricket to be in an "undoubtedly…exciting and strong position from which we can drive the sport forward" (*Rediff*, 2018).

The International Cricket Council (ICC) has 105 official member nations, with 12 being full members (Afghanistan. Australia, Bangladesh, England/Wales, India, Ireland, New Zealand, Pakistan, South Africa, Sri Lanka, West Indies, Zimbabwe), which means that these countries can participate in all three formats of international cricket. The remaining 93 participating nations are associate numbers with limited ability to participate in international cricket. A handful of associate members, such as United Arab Emirates, Scotland, and the Netherlands, qualify for ODI cricket based on their performance. Currently, members must earn their T20 cricket status, but starting in 2019, all members will participate in T20 cricket ("List of ICC Members," *Wikipedia*).

On the business side, domestic cricket rivals some of the most popular sports around the world, including major American sports. The Indian Premier League (IPL) is only in its 11th season and has already generated more sponsorship money than other major sports leagues (Geeter, 2018). According to the advertising media company GroupM, the IPL made $1 billion in sponsorships in 2017, which exceeds the $892 million made by the MLB in sponsorships in 2017 according to IEG (Geeter, 2018). In 2017, the IPL was valued at $5.3 billion, up 26% from its value in 2016 due to new broadcasting deals, the value of its title sponsor Viva, and an increase in brand value for all of its teams, according to financial advisory firm Duff & Phelps (Geeter, 2018). Considering that the league is so young, that there are only 8 teams currently in the IPL, and that an excess of talented players both within India and around the world looking for an opportunity in the league, the valuation and growth figures are quite promising.

Figure 3: IPL Logo (DIY Logo Designs).

Given the widespread popularity and financial success of the sport, success in the world cup clearly has positive economic implications. Beyond the $4 million in prize money awarded to first place (a mere

drop in the bucket), successful countries can garner interest and support from fans in the country, leading to an increase in their global revenue share. Currently, India brings in 70% of income in international cricket, and as such gets a share of $405 million from 2016-2023, which is three times more than England's $139 million, the second largest share (Wilson, 2017). If countries that currently bring in less revenue look to figure out how they can perform better in the world cup, perhaps they can increase their share of funding, thus having the resources to improve their squads, facilities, and domestic leagues. These can all lead to further improving performance, leading to a snowball effect.

With stronger performance in the world cup and fan interest comes not only a larger revenue share but also the potential opportunity to host a world cup. The 2015 Cricket World Cup in Australia and New Zealand was one of the biggest events in the history of both countries, and according to the ICC, hosting provided a significant positive boost to these countries' local economies (Sharma, 2015). An analysis of the economic impact by PricewaterhouseCoopers showed that across the two countries, the tournament generated over $1.1 billion in direct spending, created the equivalent of 8,320 full time jobs, and had 2 million bed nights in hotels (Sharma, 2015). There were around 145,000 international visitors to the two countries during the world cup, primarily from Asia, which provided a huge boost to tourism (Sharma, 2015). Currently, the opportunity to jointly host the Cricket World Cup unofficially rotates between England/Wales, Australia/New Zealand, and India/Pakistan/Sri-Lanka, and all of these teams share the fact that they perform well and have high fan interest. Full members that have hosted in the past, such as South Africa and West Indies, full members that have never hosted, and associate members that are not currently eligible to host all have a large economic incentive to improve their performance to the level of the top nations to attain the opportunity to host a world cup and profit similarly.

Aside from the financial aspect of performing well, the intangible aspect of pride from winning the world cup is important. Countries like India, Pakistan, and Australia take a great deal of nationalistic pride from beating their foes, so much so that countries essentially shut down for the day when they face their biggest rivals. For example, in the 2011 Cricket World Cup, India's billion-plus population called in sick to work, shut down shops early, and cancelled doctor's appointments to witness the hyped-up world cup

semi-final match against their arch rival Pakistan (Jamkhandikar, 2011). Both the Indian and Pakistani Prime Minister were scheduled to watch the encounter, as well as Bollywood celebrities and corporate leaders (Jamkhandikar, 2011). On a Facebook poll, over 100,000 people voted "yes" to a poll on whether the match date should be declared a national holiday, while another 40,000 said they would "bunk" work (Jamkhandikar, 2011). While this dip in productivity for the day may be seen as alarming, it clearly shows that a lot is on the line, not only financially but also in terms of nationalistic pride, when these teams represent their countries at the world cup. These significant economic and nationalistic implications of success in the world cup provide some context for why I have decided to pursue this topic

3. Problem Formulation

The two main questions I am looking to investigate in this study are the following: what primary factors have determined success in previous Cricket World Cups and whether we can use this information to predict likely outcomes in the 2019 Cricket World Cup. As I will discuss in the next section, there have been a few small studies that have explored success in ODIs based on factors such as home team advantage and playing conditions. I will look to validate the findings from these studies by including these variables. Further, I will look to build on these studies by incorporating new datasets and variables, such as player and team statistics, and testing new classification models. Using a selection of variables as predictors, I will build logistic regression models to determine how much these variables factor into individual world cup match outcomes. Using logistic regression will allow me to determine the statistical significance of this impact, which is key for any econometric analysis, as individual coefficients do not tell us much.

After using logistic regression to quantify what goes into success in the Cricket World Cup, my goal is to build an interpretable model that will predict outcomes for the 2019 Cricket World Cup. Using data from all world cups played between 1975-2011, I can work on a model until I attain a reasonable accuracy on the validation set. Once (and only after) the model is built, I will test this model for the 2015 Cricket World Cup, which of course has already occurred, by simulating the matches repeatedly using

predicted probabilities. This will at least give me a sense of how well the model is performing before I then perform simulations for the 2019 Cricket World Cup and make predictions that will go unvalidated until May-July 2019.

While achieving a reasonable accuracy and attaining predictions that make sense is quite important, my primary objective is to build a model that can be interpreted. Rather than dabbling in models like Neural Networks to fit a function on the validation set that we are trying to predict, I feel that using simpler yet more interpretable predictive models like Logistic Regression, Naïve Bayes, and Random Forest will prove more useful because I will be able to output which variables have played the largest roles in these predictions to interpret my results.

4. Previous Literature

Because there are so many different formats in which cricket is played on both the international and domestic level, there has not been a large number of studies into any one specific format. Nonetheless, a few studies of significance have been done in the area of One Day International matches, and three of these in particular stand out as precursors to my work.

In an article titled "Winning the coin toss and the home team advantage in one-day international cricket matches," de Silva and Swartz use a logistic regression model to investigate whether winning the coin toss or home team advantage have an impact on ODI outcomes, hypothesizing that these factors would play a role (de Silva, 1998). From their statistical analysis, they conclude that winning the coin toss at the beginning of the match provides no significant advantage and that playing on one's home field increases the log-odds of the probability of winning by approximately 0.5 (de Silva, 1998). Because this study uses all the data available and comes to the conclusion that coin tosses do not impact ODI outcomes, I decided not to use coin toss as a variable in my study of world cup ODIs. I also omitted the variable because we cannot use it to predict 2019 Cricket World Cup outcomes (as the coin tosses have not happened yet). Since world cup home team advantage can be more or less pronounced than home team

advantage in other ODIs, I decided to include this as a variable in my study, partially to validate de Silva and Swartz's claim that having home team advantage increases chances of success and partially to see if this effect manifested itself any differently in world cups.

Another study, "Predicting the match outcome in one day international cricket matches, while the game is in progress," attempts to demonstrate inefficiencies in betting markets through regression models (Bailey, 2006). In this study, Bailey and Clarke utilized variables such as home ground advantage, match experience, and current form to perform multivariate regressions, allowing them to predict run totals and match outcomes (Bailey, 2006). They even utilize the Duckworth-Lewis-Stern (DLS) method to update run predictions based on resources remaining (overs and wickets) (Bailey, 2006). Although their analyses and conclusions are compelling – they suggest that people tend to overreact to match events, creating brief inefficiencies in the wagering market – they do not provide interpretable results regarding specific variable (Bailey 2006). The model Bailey and Clarke build could prove useful in the massive industry of sports betting, but it does not serve the primary motivation for this paper, specifically the economic and nationalistic impact of success in world cups. Nevertheless, I will incorporate some of these new variables in my prediction of match outcome, such as experience and form, except I will take averages of individual players rather than using these variables for the team as a whole.

A third study, one of the only studies that uses machine learning on the game of cricket, "CricAI: A Classification Based Tool to Predict the Outcome in ODI Cricket," utilizes probabilistic classifiers to first determine winning factors in ODIs and then makes use of these to predict victory (Kaluarachchi, 2010). Some of the main factors examined in their study include home team advantage, day/night effect, winning the coin toss, and batting first. In the end, they find that a classification – and specifically Naïve Bayes – approach to the problem results in the best performance (lowest squared error), and they build a software tools around this model that takes in the above factors and outputs a match outcome (Kaluarachchi, 2010). After I have built my own model, I will not only have an idea of which factors are important in outcome but a user could also theoretically input values for the variables I am examining, and my model would make

a prediction. Thus, Kaluarachchi and Varde's interpretable AI approach to predicting match outcomes aligns well with my own approach.

5. Hypothesis

I have chosen to study eight categories of variables: round of play; relative temperature in home country compared to the match location; home team advantage; rest since last match; experience of the coach, various roles on the squad, and the team overall; team chemistry as measured by shared local teams; previous world cup success; and individual batting and bowling skill. I hypothesize that some of these will have a low impact on success, while others would have a moderate or high effect.

The variables I anticipate will have a low effect on match outcomes include the current round being played, as this should not skew toward either team in the matchup. The only way this variable could have predictive power is if it were to interact with a variable like previous success, as teams that have previously been in the playoff round may perform better than teams that have not. Regardless, I guess that this variable will show no significant impact. I also predict that temperature will have a low impact on match outcome because temperature is usually controlled for; matches can only be played within moderate temperature ranges. Even if there is some deviation, these players are professionals who play all around the world, so I do not feel temperature will have much of an impact on success.

There are four variables I believe will have a moderate effect on match outcome. I believe home team advantage would have a moderate effect based on the previous studies that concluded so and based on my intuition. I feel that rest could also play a medium role in determining outcome because fatigue plays a role in any sport. Further, I believe team chemistry will have a moderate impact on outcome because if more players are playing on the same domestic team throughout the cricket seasons, they have established synergy in place, which is key in a sport like cricket where communication plays a major role in fielding and batting. Players bat in partnerships, meaning two batsmen are on the field at once, and players often strategically hit 1s or 2s to ensure the stronger batsman is in batting position. Finally, I feel that world cup

and ODI experience of the coach, captain, wicket-keeper, and whole squad will have a moderate impact. Since the Cricket World Cup only takes place every four years and ODIs are not all that common either, most players will not have been to many world cups, so these variables could be easily skewed toward zero, preventing them from having a major impact on outcome. Further, even if captain experience were to show a correlation with match outcome, it would be difficult to conclude any causation, as stronger squads are more likely to be able to afford stronger coaches due to higher budgets and fan pressure for good performance. For these reasons, despite the reasonable intuition that more experience means better performance, I expect only a moderate impact.

The two factors that I hypothesize will have the most impact are previous outcome in world cups and batting/bowling averages/records of individual players in previous ODIs. Since I could not find previous studies that examined these variables, I feel they were the most compelling, as correlations with success have very direct and tangible explanations. Having success in previous world cups could suggest that the current squad is more likely to be high quality, and countries like Australia, India, Sri Lanka, and Pakistan have been consistently successful, which seems to continually build on their success. Further, having strong batting and bowling skill directly suggests that a squad is high quality and thus more likely to do well. While predicting that these factors will impact outcome does not seem groundbreaking, the important piece is that being able predict outcomes more effectively using these readily available data could prove quite useful and could help diagnose which specific areas nations should focus on.

For the 2015 Cricket World Cup, outcomes are already known, so I plan to examine the outcomes my model predicts for 2015 and use these matches as a test set. Then, I can use the same method to predict outcomes in 2019. Based on my intuition, knowledge of Australia's historical status as a cricket powerhouse, and knowledge that India has arguably the best player in the world in Virat Kohli, I predict that my model will give Australia the highest probability of winning the 2019 Cricket World Cup and will give India the second highest chance.

6. Data

In this section, I will explain each variable that I include in the classification models. I will discuss what each variable represents, how I acquired the data, and if applicable how I calculated the values. The dataset is made up of all 439 Cricket World Cup matchups ever played (acquired from Wikipedia), and in each row of this dataset there are two teams: Nation A and Nation B. Each row also contains the date, month, year, and location of the matchup along with a column for the outcome of the match. This outcome value is 0 if Nation A won, 1 if Nation B won, empty if the match is yet to be played in 2019, or the entire row is omitted if the match was tied. While I could have included ties and made the outcome a tertiary variable rather than binary, making this a multiclass problem would have greatly complicated the model, thus hurting the prediction accuracy. Also, since ties are much rarer in cricket than in other sports like soccer, attempting to predict them is not very useful. One important thing to consider is that Wikipedia may have a bias toward which team it includes first in its list of historical matchups (i.e. teams that bat first, teams at home, teams that end up winning, etc.). While no trend in the matchup order is visibly apparent, in order to definitively eliminate this potential bias, I randomize which team is Nation A and which team is Nation B. For most variables in the dataset, I take the difference of the value between Nation A and Nation B. Therefore, any selection of A or B by the model should be solely based on variables in the dataset (e.g. higher values of a certain variable seem to indicate team B winning more or vice versa).

Round: This variable represents the current round in which the matchup is being played, which comes from Wikipedia's matchup data ("Cricket World Cup," *Wikipedia*). The value 0 specifies group stage, 1 specifies "super six" (a second group stage sometimes played in the world cup before the knockout stage), and 2 specifies the playoff (an elimination round that either begins from the quarterfinals or semifinals).

Temperature: This variable represents, in one value, how far the participating nations' temperatures deviate from the temperature at the match location. The data for temperature (in Fahrenheit)

comes from Climate-Data.org ("Climate Data," *Climate Data*). For each matchup, I do the following for Nation A and Nation B: take the absolute value of the difference between the average temperatures in the participating nation's capital and in the city in which the matchup takes place (both within the month of the matchup). To calculate the final temperature statistic, subtract the value for Nation A from Nation B. This final value represents the difference in deviation from the host temperature between the nations (i.e. a higher value means that B deviates more and vice versa). I only use temperature to measure weather impacts because rainfall is already considered in the outcome when a match is cancelled or the DLS method is used, and it is not practical to get data of other game-day conditions that may affect play (e.g. sunlight, clouds, pitch condition, etc.).

Home: This variable represents whether a team has home team advantage with the data again coming from Wikipedia matchup data ("Cricket World Cup"). Each nation gets a 1 if it is playing at home and 0 if not, then A's value is subtracted from B. Therefore, a final value of -1 would mean Nation A is at home, 0 would mean neither team is at home, and 1 would mean Nation B is at home. Note: Matches hosted by Wales are considered to be home matches for England, as England and Wales share their cricket team.

Rest: This variable compares how long it has been between the teams' last matches. The data comes from Wikipedia matchup data ("Cricket World Cup"). If either team is playing their first match, the value is set to 0 because the amount of time since the teams' last match is unknown (sometime before the world cup began). If both teams have played at least one match in the world cup, the number of days since the last match is calculated for each team, and that of Nation A is subtracted from that of Nation B. Thus, a higher value for rest would signify that Nation B has had more rest than Nation A and vice versa.

Experience: This variable measures how much experience different roles on each team have in both world cups and ODIs. I use ODIs in addition to world cups to measure experience because there are so few world cups that the averages will likely be very close to 0, giving them low predictive power for match outcome. All basic player and coach information for the 133 unique world cup squads (uniquely identified by nation and year) comes from Wikipedia ("Cricket World Cup"). All detailed statistics on the

1,937 individual players (uniquely identified by name, DOB, and year) comes from the querying tool from ESPN Cricket Info ("Statsguru," *ESPN Cricinfo*). To quantify the experience of different roles on the team, I use eight different variables: both the average number of world cups and ODI's played by the coach, captain (leads the team), wicket-keeper (stands behind the batter to attempt to take a wicket), and starting 11. If a squad does not have a coach (which is the case surprisingly often), then coach experience variables are set to 0. As before, I take each of these eight values for Nation A and Nation B and subtract that of A from that of B to get the final value. I am left with eight variables to try to predict match outcome.

Chemistry: This variable represents how much chemistry each team has based on shared local teams. A squad can be thought of as an undirected graph, where a player is a node, and a shared squad is an edge between two teams. There are many ways we can quantify how "connected" this graph is and therefore how much chemistry a squad has. I decided to go with my own unique method. Starting with pairs (chemistry_2), I go through each player and if he is on a domestic team with at least one other player in the squad, I increment the variable. I repeat this method for triples (chemistry_3) and increment if the player is on a domestic team with at least two other players in the squad. I repeat this process up to chemistry_11. Each of these variables is on a scale from 0 (all players on unique teams) to 11 (all players on the same team). As with the other variables, I calculate each chemistry value for Nation A and Nation B and subtract that of A from that of B so that the chemistry variables demonstrate how much more chemistry Nation B has.

Individual Skill: To quantify individual skill, I utilize six variables measured in previous matches played by each player in the ODI format. For bowling statistics, I use average batting average (total runs scored / number of times out) of the starting 11, average batting average of the top 5 players on the squad, and total centuries (100-run games) achieved by players on the squad. For bowling statistics, I use average bowling average (total runs conceded / wickets taken) of the starting 11, average bowling average of the top 5 players on the squad, and total 5-wicket games bowled. I examine averages for the top 5 because that gives an idea of the best batters on the team, and that is the minimum number of players that

need to bowl in a match. I look at the starting 11 because depth of the squad may also be important (5 players alone are unlikely to win the match). While two other more advanced statistics, strike rate and economy rate, are thought to demonstrate batting and bowling skill a bit better than pure averages, these statistics are not available for querying on ESPN Cricket Info. However, plotting and statistical analysis shows correlations between batting average and strike rate as well as bowling average and economy rate (Morgan-Mar, 2006). Therefore, the use of batting and bowling averages as variables in my model is sufficient for my purposes. I retrieved individual statistics by using ESPN Cricket Info to query a player's stats before the start of the given world cup ("Statsguru"). I then scraped the data from the resulting webpage with Python and automated the process for all 1,937 players.

Previous Success: To measure prior success, I utilize five variables. The first four represent how many times the nation has appeared in a world cup, has been at least a semi-finalist, has been at least a runner-up, and has won the world cup. The last gives a score of 0 (did not appear) to 4 (won) based on how well the team did in the most recent world cup. As usual, I take the value of each of these five variables for Nation A and Nation B and subtract A's from B's to get the final value in each row.

Team \ Host	1975 (8)	1979 (8)	1983 (8)	1987 (8)	1992 (9)	1996 (12)	1999 (12)	2003 (14)	2007 (16)	2011 (14)	2015 (14)	2019 (10)	2023 (10)
Afghanistan											GP	Q	
Australia	RU	GP	GP	W	GP	RU	W	W	W	QF	W	Q	
Bangladesh							GP	GP	S8	GP	QF	Q	
Bermuda									GP				
Canada		GP						GP	GP	GP			
East Africa†	GP												
England	SF	RU	SF	RU	RU	QF	GP	GP	S8	QF	GP	Q	
India	GP	GP	W	SF	GP	SF	S6	RU	GP	W	SF	Q	Q
Ireland									S8	GP	GP		
Kenya						GP	GP	SF	GP	GP			
Namibia								GP					
Netherlands						GP		GP	GP	GP			
New Zealand	SF	SF	GP	GP	SF	QF	SF	S6	SF	SF	RU	Q	
Pakistan	GP	SF	SF	SF	W	QF	RU	GP	GP	SF	QF	Q	
Scotland							GP		GP		GP		
South Africa					SF	QF	SF	GP	SF	QF	SF	Q	
Sri Lanka	GP	GP	GP	GP	GP	W	GP	SF	RU	RU	QF	Q	
United Arab Emirates						GP					GP		
West Indies	W	W	RU	GP	GP	SF	GP	GP	S8	QF	QF	Q	
Zimbabwe			GP	GP	GP	GP	S6	S6	GP	GP	GP		

Figure 4: Table showing historical success in the Cricket World Cup ("Cricket World Cup")

For the 2019 Cricket World Cup projection model, I collect the same or equivalent data for each of the variables above. Luckily, we know all of the matchups that will happen, as the 2019 Cricket World Cup will feature only 10 teams playing each other in one large group, followed by a

playoff with the top four teams from the group stage. Data for the round, temperature, home team advantage (hosted by England/Wales in 2019), rest, and previous success are all calculated in the same way as above model. For coaching and squad information, such as experience, chemistry, and individual skill, I use data for each country's current national team. While these are subject to change by May 2019, they can serve as a good proxy for the coach and squad to be used in the upcoming tournament.

7. Analysis

a) Logistic Regression

I decided to begin my analysis with the logistic regression because it is the simplest and most interpretable classification model. I first ran individual regressions for each variable in addition to three separate regressions (shown in Figure 5) to examine the how these factors impact match outcome. The results (shown in Figure 6) suggest which variables have a significant impact on outcome, as I highlight below.

Matchup Model

$$\text{outcome} = \beta_0 + \beta_1 Round + \beta_2 Home + \beta_3 Temperature + \beta_4 Rest$$

Player/Team Stats Model

$$\text{outcome} = \beta_0 + < \beta_1, ..., \beta_8 > \cdot < Coach_{wc}, Captain_{wc}, ..., Wicketkeeper_{odi}, Overall_{odi} > + < \beta_9, ..., \beta_{18} > \cdot < Chem_2, ..., Chem_{11} > + < \beta_{19}, ..., \beta_{22} > \cdot < BattingAvg_{11}, ..., BowlingAvg_5 > + \beta_{23} Centuries + \beta_{24} Fives + < \beta_{25}, ..., \beta_{29} > < Appeared, SemiFinalist, RunnerUp, Winner, Prev >$$

Hybrid Model

$$\text{outcome} = \beta_0 + \beta_1 Round + \beta_2 Home + \beta_3 Temperature + \beta_4 Rest + < \beta_5, ..., \beta_{12} > \cdot < Coach_{wc}, Captain_{wc}, ..., Wicketkeeper_{odi}, Overall_{odi} > + < \beta_{13}, ..., \beta_{22} > \cdot < Chem_2, ..., Chem_{11} > + < \beta_{23}, ..., \beta_{26} > \cdot < BattingAvg_{11}, ..., BowlingAvg_5 > + \beta_{27} Centuries + \beta_{28} Fives + < \beta_{29}, ..., \beta_{33} > < Appeared, SemiFinalist, RunnerUp, Winner, Prev >$$

Figure 5: Three logistic regression model equations.

Round: As expected, the round of the matchup showed no statistically significant impact on match outcome.

Temperature: As I predicted, temperature did not appear to have a statistically significant impact on match outcome.

Home: As I anticipated, home team advantage had a positive effect on match outcome at the 0.1% significance level, which corroborates the findings of previous studies.

Rest: Contradicting my hypothesis, differences in rest had no significant impact on match outcome.

Experience: The amount of world cup and ODI experience of the captain had a significant positive impact on outcome, but the coach and overall squad experience did not appear to have a significant impact. This may at first seem counterintuitive but does actually make sense. The coach actually does not play a central role in cricket, unlike most other sports, and instead the captain makes all important decisions such as choosing the starting 11, strategically setting the batting and bowling order, telling the fielders where to position themselves, etc. As mentioned earlier, many players are playing in their first world cup, and the number of ODIs played will be skewed toward zero as well. Because of the insignificant role of the coach and minimal ODI experience of squads as a whole, it makes sense that the experience of specific team roles like captain and wicket-keeper would be more predictive of outcome than that of the coach and overall team.

	Individual	p	Matchup	p	Player/Team	p	Hybrid	p
(Intercept)	2.28057	<2e-16 ***	0.470697	0.000835 ***	-0.122679	0.2879	-0.162702	0.2208
round	-0.1345	0.573	-0.001525	0.993521			0.058306	0.7265
temp	0.002927	0.767	-0.007973	0.337285			-0.005905	0.4981
home	0.6446	0.000668 ***	0.423114	0.097914 `			0.353507	0.1465
appear	0.11268	0.021 *			0.0410352	0.5885	-0.052482	0.0294 *
semi_finalist	0.11857	0.0694 `			-0.056354	0.5752	0.038648	0.6148
runner_up	0.16144	0.0451 *			0.1630994	0.2572	-0.06754	0.5113
winner	0.1818	0.17			-0.37481	0.0734 `	0.217363	0.1412
prev	0.2848	0.00628 **			0.0189816	0.8575	-0.444097	0.0379 *
coach_exp	0.07423	0.461			0.0588911	0.4533	0.005986	0.9559
rest	-0.06774	0.227	-0.034117	0.184021			0.094199	0.2501
c_odi	0.004256	0.00676 **			0.0036767	0.1967	0.003736	0.1956
wk_odi	0.003664	0.0612 `			0.0035213	0.2246	0.002916	0.3173
squad_odi	0.002097	0.479			-0.005416	0.065 `	-0.005964	0.0463 *
c_wc	0.52683	9.71e-09 ***			0.057092	0.7524	0.046793	0.798
wk_wc	0.2092	0.0535 `			-0.271457	0.3992	-0.137765	0.6758
squad_wc	0.14904	0.122			0.0935698	0.6858	-0.040152	0.8682
bat_avg_11	0.03499	0.0218 *			0.1068681	0.0379 *	0.099692	0.0577 `
bat_avg_5	0.02452	0.0312 *			-0.052666	0.1343	-0.048369	0.1783
centuries	0.00996	0.142			0.0159593	0.0524 `	0.015094	0.0712 `
bowl_avg_11	0.003594	0.701			-0.004663	0.7454	-0.004266	0.7679
bowl_avg_5	0.008925	0.441			0.0193287	0.2641	0.016954	0.3349
fives	0.03313	0.181			-0.000193	0.9946	0.007833	0.7878
chem_2	0.0819	0.000123 ***			0.0510358	0.065 `	0.05141	0.0677 `
chem_3	0.04345	0.0139 *						
chem_4	0.02094	0.244						
chem_5	-0.01106	0.612						
chem_6	-0.05278	0.0411 *						
chem_7	-0.07414	0.0224 *						
chem_8	-0.08228	0.0184 *						
chem_9	-0.0737	0.063 `						
chem_10	-0.0737	0.063 `						
chem_11	-0.0022	0.964						
Signif:	'***' 0.001	'**' 0.01	'*'0.05	'`' 0.1				

Figure 6: Logistic regression results (coefficients and p-values.

Chemistry: The chemistry_2 and chemistry_3 variables showed a statistically significant impact on outcome, but the high-order chemistry variables either showed no significance or showed the opposite trend. This could potentially be explained by the fact that countries with smaller leagues tended to have all of their best players concentrated on a few domestic teams, so chemistry_5 and chemistry_6 tended to be higher than more cricket-focused nations despite poorer squad strength. Because of the poor predictive

power of higher-order chemistry metrics and the correlation between chemistry_2 and chemistry_3, I only included chemistry_2 in the combined models.

Individual Skill: Out of the individual statistics I examined, batting average of the starting 11 and top 5 were the only variables to show a statistically significant (positive) impact on outcome throughout the models. The insignificance of records like centuries and 5-wicket games could be due to the rarity of these achievements (a low value for these records does not reliably signify a poorer team). The statistical insignificance of the impact of bowling statistics could be because of the way averaging was done. If data were more readily available, economy rate could provide a better measure of bowling skill or averaging could have been done on aggregate rather than averaging individual players' averages, which are prone to being inflated or deflated due to small sample size.

Previous Success: The variables for previous success, except for the variable for total championships, all showed some degree of statistical significance in their impact on outcome. This mostly aligns with my prediction that historical success would be a strong indicator of outcome. The reason championships may not have shown significance is because there are such a limited number of titles to go around that most values in this column will be relatively close to 0, limiting its ability to predict match outcomes.

In the three combined models – the Matchup, Player/Team, and Hybrid Models – variables that were significant by themselves sometimes showed less significance. This observation has two potential explanations: in a multivariate regression, there are more degrees of freedom, so t-statistics will be lower (less statistically significant), and variables that are correlated by multicollinearity may have a smaller coefficient and less significance when used together in a standard logistic regression. For these reasons, despite improving explanatory power of the output value, the combined models are less useful in demonstrating significance of variables that we are examining.

b) Random Forest

After building logistic regression models to show significance of the impact of the variables of interest, I decided to focus my efforts on improving the accuracy without sacrificing too much interpretability. I trained models on data from 1975-2011 (with rows randomized and an 80%/20% split between training/validation data). A few models I tried were K-Nearest Neighbors, which finds the Euclidean distance between an observed point and training points to make a prediction; Naïve Bayes, which takes a probabilistic approach to generating a model to fit training data onto a function (similarly to Kaluarachchi's approach in CricAI); and Decision Tree, which builds a tree using inequality rules on the input variables to fit the input data to the outputs. None of these models performed as well as logistic regression, which achieved ~76% validation accuracy, but when I tried a model called Random Forest and tuned some parameters, this model outperformed all previous models by achieving a validation accuracy of ~84%. Given that flipping a coin would achieve 50%, a validation accuracy (which represents accuracy for a set on which the model has not been directly trained) of above 80% is quite strong. The way that Random Forest, a powerful bagging tool, generally works is by aggregating multiple "learners" – in this case many individual Decision Trees – to make one larger model (see figure 7). The individual Decision Trees may not perform well by themselves and may be very different from one another, but when we average many of these models together, the result is strong.

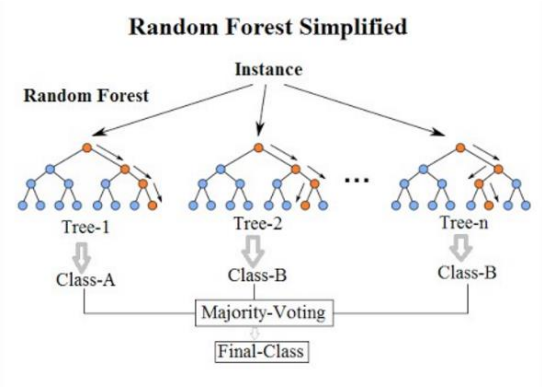

Figure 7: Random Forest model structure.

The Random Forest Model can output weights representing how important a role each variable played in forming the Decision Trees. The feature importance weights did not tell a very different story than the logistic regression significance values except that they seemed to emphasize bowling averages and batting records more than expected (see Figure 8). Perhaps the fluid nature of the Random Forest model – multiple conditions can be placed on a single variable within a tree and between trees – helps these variables predict match outcomes (e.g. outliers may be removed by specifying bowl_avg_11 > 5 and bowl_avg_11

< 100 at the same time, while logistic regression simply cares how high or low the value is). The ability of Random Forests to hide some of the flaws that may exist in the statistics I used or metrics I calculated can explain why it performs better than logistic regression when predicting on a validation set.

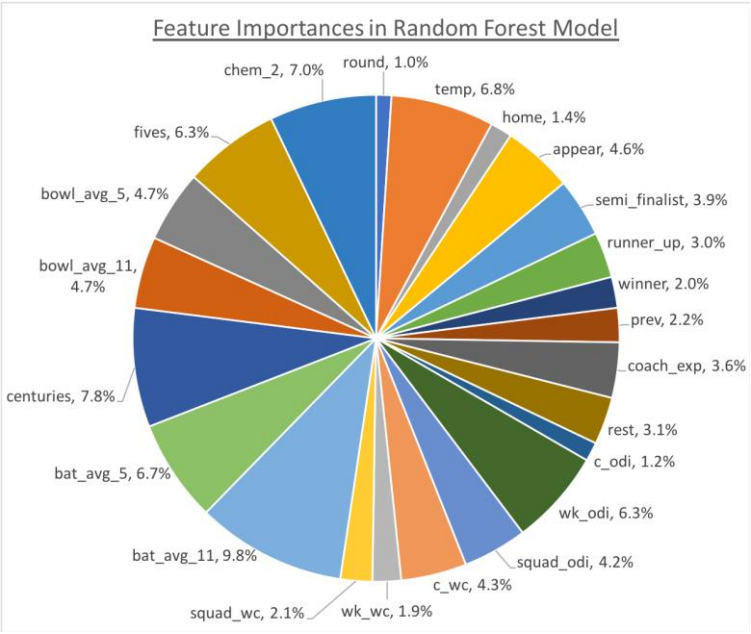

Figure 8: Feature importance weights in Random Forest model.

c) Outcome Prediction

After training a Random Forest model and achieving a suitable validation accuracy for the split 1975-2011 set, I then used 2015 as a "test" set by running 10,000 simulations of the entire 2015 Cricket World Cup and comparing my results to what actually occurred. I performed random simulations of the match outcomes using probability outputs from the model for each matchup in 2015. In doing so, I was able to

	Probability of Achieving:		
	semi-finalist	finalist	champion
afghanistan	0.2%	0.0%	0.0%
australia	64.1%	33.0%	16.1%
bangladesh	3.4%	0.4%	0.1%
england	7.3%	0.8%	0.1%
india	83.1%	55.3%	28.2%
ireland	4.9%	0.8%	0.2%
new_zealand	76.5%	52.4%	**33.8%**
pakistan	25.2%	6.2%	1.7%
scotland	1.8%	0.2%	0.1%
south_africa	55.5%	21.5%	8.4%
sri_lanka	56.4%	25.1%	10.4%
uae	3.9%	0.4%	0.1%
west_indies	13.6%	3.4%	0.8%
zimbabwe	4.0%	0.6%	0.2%

Figure 9: Probability predictions for teams to reach various stages of the 2015 Cricket World Cup.

Figure 10: Actual playoff results of the 2015 Cricket World Cup.

keep track of how many times each team reached different stages of the tournament and thus predict the probability of these outcomes (see Figure 9).

My predictions implied that New Zealand had the best chance of winning the tournament with India in second and Australia, Sri Lanka, and South Africa having the next three best chances. As Figure 10 demonstrates, of the five teams my model gave the best chances of success, four actually advanced to the playoffs. Thus, my model performed well on this 2015 test set.

After making outcome predictions for 2015, I felt my model was ready to tackle 2019. Using the same method of simulating matches, I used my model to predict probabilities of different outcomes for each team (see Figure 11). The predictions show that India has the overwhelmingly best odds of winning the tournament, with South Africa, New Zealand, England, and Pakistan following behind. What contradicts my prediction and comes as a surprise is the fact that my model gives Australia only a 1.9% chance of winning the 2019 Cricket World Cup (and a 17.9% chance of making the semi-finals). While these low odds may seem unrealistic considering Australia's historical domination of world cups, the current Australian squad is not exactly impressive. Ever since a ball-tampering scandal that occurred in 2018 that led to coaching and captaincy changes as well as blows to player contracts and sponsorships, the Australian national team has not performed as well as they have in years and decades past ("2018 Australian Ball-Tampering Scandal," *Wikipedia*). In fact, the ICC releases rankings of each team within each format, and the current ODI Rankings have England, India, New Zealand, South Africa, and Pakistan all ahead of Australia ("Cricket/ICC Rankings," *Wikipedia*). These rankings suggest that perhaps the low chance my model gives to Australia is not an anomaly and that there are simply other stronger squads that have a better chance than Australia in this upcoming world cup. It remains to be

	Probability of Achieving:		
	semi-finalist	finalist	champion
afghanistan	5.9%	1.2%	0.4%
australia	17.9%	5.3%	1.9%
bangladesh	7.3%	1.8%	0.5%
england	56.2%	26.1%	11.8%
india	82.9%	60.2%	**40.2%**
new_zealand	61.9%	30.5%	12.1%
pakistan	45.7%	17.6%	7.4%
south_africa	72.1%	39.4%	18.2%
sri_lanka	39.0%	15.1%	6.7%
west_indies	11.1%	2.7%	0.9%

Figure 11: Probability predictions for teams to reach various stages of the 2019 Cricket World Cup.

seen who will actually take the throne in the 2019 Cricket World Cup, but I am confident in the predictions my model assigns to each country based on the reasonable and interpretable feature importance weights, the strong validation accuracy, the similarity between projections and reality in 2015, and the alignment with current ICC rankings (without having consulted them before building the model).

d) Technologies

I used the following technologies for data collection, cleaning, and analysis:

- Python (pandas, scikit-learn, BeautifulSoup, etc.)
- R (glm)
- Google Colaboratory (iPython Notebooks)

8. Conclusions

My motivation behind exploring factors of success in cricket world cups was the economic and nationalistic implications behind such success. Given these motivations, I focused on two primary objectives: building an interpretable model and attaining a reasonable accuracy. Using econometric tools to build logistic regression models allowed me to achieve the first goal of interpretable results. Most trends I observed in the input variables aligned quite well with what I hypothesized, but of course some did not. In particular, batting averages, captain experience, home team advantage, general team success, and chemistry met and even surpassed my expectation in their predictive power. On the other hand, some factors that were underwhelming in their ability to affect outcome compared to my expectations were rest, coach experience, overall squad experience, bowling averages, and previous championships. In building models that could be interpreted, achieving a reasonable validation accuracy of ~84%, and making predictions for the 2015 and 2019 Cricket World Cups (which seem to fit nicely with reality), I achieved all of the objectives I set for myself in this project. Of course, there are always some limitations and some future work that can be done to build on these insights and answer new questions.

There are a few limitations of this study worth mentioning. As mentioned earlier in this paper, ties could be predicted in addition to a win or loss. However, ties not only complicate the model by making it multiclass, which logistic regression does not support, but they also do not serve very much practical use, as most ties are match cancellations rather than pure run ties, which are very rare. Another limitation is the fact that batting and bowling average could be skewed upward or downward by low sample size or by factors like wickets remaining and overs remaining, as players tend to play more aggressive if they are later in the batting order. While the Duckworth-Lewis-Stern (DLS) method could help address this issue, information about batting order and resources remaining in a match when batting is not readily available on an individual basis. Finally, my methods of incorporating chemistry and team success into my model were not ideal, as they added too many excess variables. The ratio of observations to variables should ideally be high when building a model, and if I had built metrics to reliably quantify these factors in one variable, this ratio would have increased thus improving performance of the model.

The most important limitation of my work is the issue of overfitting, which future work could seek to address. Overfitting is a common problem when using strong predictive models like Random Forest on samples as low as 438 rows. When adjusting the model to improve validation accuracy, a function is being fit specifically for that set, which may impede performance on a test set or when making predictions. In order to reduce overfitting, methods like boosting (trimming parts of the decision tree in the model), cross-validation (taking repeated slices of the training set to validate on instead of one single validation set), and regularization (introducing a parameter improves precision at the cost of accuracy) could all be employed to reduce overfitting. More advanced models, such as neural networks, could be employed in future work; however, these would not have served my objective of building an interpretable model, as they operate as a black box. Finally, increasing the amount of data – say, by increasing the scope of the problem to include all ODIs rather than just world cup ODIs – to improve the observation-to-predictor ratio could also help prevent overfitting and improve the model. While such future work would be compelling, the interpretable and reasonably accurate tools I used and the scope I set for this project allowed me to fulfill my primary objectives and motivations.

GitHub Repository

To explore the code and statistical models, please check out the following repository: https://github.com/neil-menghani/cricket_wc.

Acknowledgements

I would like to thank Professor Sunil Gulati for his feedback on this project throughout the semester and for the wonderful experience I have had in his Sports Economics seminar.

Works Cited

"2018 Australian Ball-Tampering Scandal." *Wikipedia*, Wikimedia Foundation, 6 Dec. 2018, en.wikipedia.org/wiki/2018_Australian_ball-tampering_scandal.

Bailey, Michael, and Stephen R Clarke. "Predicting the Match Outcome in One Day International Cricket Matches, While the Game Is in Progress." *Journal of Sports Science & Medicine*, vol. 5, no. 4, 2006, pp. 480–487.

"Climate Data for Cities Worldwide." *Climate Data*, Climate Data, en.climate-data.org/.

"Cricket Has over One 1 Billion Global Fans." *Rediff India News*, 27 June 2018, www.rediff.com/cricket/report/cricket-has-over-one-1-billion-global-fans-t20-most-popular/20180627.htm.

"Cricket Rules." *Cricket Rules*, cricket-rules.com/.

"Cricket World Cup." *Wikipedia*, Wikimedia Foundation, 26 Nov. 2018, en.wikipedia.org/wiki/Cricket_World_Cup.

De Silva, Basil M, and Tim B Swartz. "Winning the Coin Toss and the Home Team Advantage in One-Day International Cricket Matches." 1998, pp. 1–15.

"Dismissal (Cricket)." *Wikipedia*, Wikimedia Foundation, 12 Oct. 2018, en.wikipedia.org/wiki/Dismissal_(cricket).

DIY Logo Designs. "Indian Premier League - IPL 2018 - All Teams Logos." *DIY Logo Designs*, 18 Sept. 2018, diylogodesigns.com/blog/indian-premier-league-ipl-2018-all-teams-logos-png-transparent-background-download/.

Geeter, Darren. "The Business of Cricket ." *CNBC*, CNBC, 1 Aug. 2018, www.cnbc.com/2018/07/03/cricket-ipl-india-sports-mlb-baseball.html.

Jamkhandikar, Shilpa. "India Shuts down for India-Pakistan Showdown." *Reuters*, Thomson Reuters, 29 Mar. 2011, uk.reuters.com/article/uk-cricket-world-india-shutdown-idUKTRE72S1TI20110329?feedType=RSS&feedName=everything&virtualBrandChannel=11708.

Kaluarachchi, Amal, and Aparna S Varde. “CricAI: A Classification Based Tool to Predict the Outcome in ODI Cricket.” *ICIAfs*, 17 Dec. 2010, pp. 250–255.

Koehrsen, Will. “Random Forest Simple Explanation.” *Medium*, Medium, 27 Dec. 2017, medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d.

“List of International Cricket Council Members.” *Wikipedia*, Wikimedia Foundation, 21 Dec. 2018, en.wikipedia.org/wiki/List_of_International_Cricket_Council_members.

Morgan-Mar, David. “Statistical Analysis of Cricket.” *DM*, 16 Feb. 2006, www.dangermouse.net/cricket/statanalysis.html.

“One Day International.” *Wikipedia*, Wikimedia Foundation, 16 Dec. 2018, en.wikipedia.org/wiki/One_Day_International.

Pobjie, Ben. “What Cricket Is the Best Cricket?” *The Roar*, The Roar, 21 Nov. 2014, www.theroar.com.au/2014/11/21/what-cricket-is-the-best-cricket/.

“Cricket/ICC Rankings.” *Wikipedia*, Wikimedia Foundation, 8 Aug. 2018, en.wikipedia.org/wiki/Portal:Cricket/ICC_Rankings.

Rajesh, S. “How the Duckworth-Lewis-Stern Method Works.” *ESPN Cricinfo*, 8 June 2017, www.espncricinfo.com/story/_/id/19577040/how-duckworth-lewis-stern-method-works.

“Scoring Runs in Cricket.” *Wikipedia*, Wikimedia Foundation, 29 Aug. 2018, en.wikipedia.org/wiki/Scoring_runs_in_cricket.

Sharma, Rajiv Teja. “Cricket World Cup 2015 Boosted Local Economies of Australia and New Zealand: Study.” *The Economic Times*, Economic Times, 30 June 2015, economictimes.indiatimes.com/news/sports/cricket-world-cup-2015-boosted-local-economies-of-australia-and-new-zealand-study/articleshow/47875879.cms.

“Statsguru.” *ESPN Cricinfo*, stats.espncricinfo.com/ci/engine/stats/index.html.

“Test Cricket.” *Wikipedia*, Wikimedia Foundation, 19 Dec. 2018, en.wikipedia.org/wiki/Test_cricket.

Wilson, Joe. “India's Share of ICC Global Revenues Adjusted after Initial Vote - BBC Sport.” *BBC News*, BBC, 22 June 2017, www.bbc.com/sport/cricket/40374596.