A Machine Learning Study into Batting Strategies in the Indian Premier League

2106625, 2106259, 2105574

## Background and Project Aim

The bat and ball game of cricket is widely considered to have originated in England in the 16th century and as a result of the British empire has since become one of the most played and watched sports in the world. According to the International Cricket Council (2018), there are over 1 billion cricket supporters worldwide and nearly 90% of them come from the Indian subcontinent. It is the most popular sport in Pakistan, Bangladesh, Sri Lanka, Afghanistan and India with support for the game particularly fervent in the latter.

Twenty20 (T20) cricket is a form of the game that was first played professionally in 2003, when the England and Wales Cricket Board introduced the Twenty20 Cup to the domestic schedule, to be played by the 18 First Class counties. T20 is a shortened form of the game where two teams of 11 players play one innings each. Each teams bowls 20 overs, each consisting of 6 deliveries, at the other team and the aim is to score the most runs. Two members of the batting team bat at any one time and runs are scored by either running from one end of the wicket, a 22 yard by 10 foot strip of grass in the middle of the pitch, to the other, or when the ball crosses the boundary of the pitch either after hitting the ground (4 runs) or without hitting the ground after making contact with the bat (6 runs). The bowling team attempt to prevent the batting team from scoring runs and also attempt to get the batters out (take wickets), principally by bowling the ball at the stumps, two sets of three wooden poles at each end of the wicket, catching the ball off the bat before it bounces or hitting the stumps with the ball while the batters are in the process of completing a run. Since two batters have to be in at a time, when a team loses its 10th wicket the innings is completed.

In 2007 India beat their biggest rivals, Pakistan, in the final of the inaugural Twenty20 World Cup. Prior to this, Indian supporters had been sceptical of the newest format of the game, but the international team’s success led to a burgeoning interest and the advent of the Indian Premier League (IPL), which was announced by the Board of Control for Cricket in India (BCCI) in September 2007 and would start in 2008. The inaugural edition consisted of 8 teams belonging to franchises that were sold for an initial $723.59m (ESPN Cricinfo, 2008) playing a round robin, followed by semi-finals and a final. The format has varied slightly from 2008 to the present day, but has always been contested by 8, 9 or 10 franchise teams. The sum of money paid for the franchises before the 2008 tournament is a reflection of the status of the owners of the franchises, including Bollywood stars and multinational conglomerates. The wealth of the franchise owners meant they were able to entice many of the world’s best cricketers to play in the IPL and this, along with the marketability of and excitement provided by the short format of the game has led to it being one of the most popular sporting events in the world, with an average of 116 million unique viewers watching 14 matches in the first two weeks of the 2020 competition (Tewari, 2021).

Some sports, like football, are continuous in nature and can be difficult to analyse statistically. However, cricket is a sport that consists of a number of discrete events (balls, or deliveries), the outcome of which can be measured in terms of the number of runs (typically 0 to 6) and wickets (0 or 1) that result from it. Such ball-by-ball data is widely available on the internet and can be used to gain insights into the game (Davis, Perera & Swartz, 2015). More complex data has also been collected in recent years, including ball tracking data that can illustrate where the ball was released from, where it hit the ground, how fast it was travelling and where it passed the batter and the stumps. This revolution in data collection has enriched the simple ball-by-ball data, but is not publicly available due to the costs associated with collecting it.

Baseball is another game that consists of a number of discrete events and, although the two are very different tactically, can be used as a blueprint for the use of data analysis in cricket. Sabermetrics is a term invented by Bill James and encompasses the use of statistical analysis in baseball in an attempt to empirically compare the performance of different baseball players (Lewis, 2003). Sabermetrics was brought to public attention by the book Moneyball, published in 2003, which chronicles the statistical approach of the Oakland Athletics’ player recruitment and how it enabled them to reach the 2002 Major League Baseball playoffs despite having the third lowest team salary in the competition. While complex data analysis has been used successfully in baseball for nearly two decades, only recently have professional cricket teams started enlisting the help of individuals with expertise in data analysis and data science. Companies like CricViz, founded in 2014, have brought analytics to the fore in cricket, working with teams in the IPL, Australia’s Big Bash League and the Pakistan Super League to improve results on the pitch, as well as with broadcasters and media organisations including Sky Sports and The Telegraph to improve the understanding of the cricket following public (CricViz).

The aims of this research are to use data mining techniques to evaluate batting tactics and strategies in the IPL. We aim to investigate how they differ between teams and players. For example, it could be that some teams look to attack throughout the 20 overs and accept the higher risk of losing wickets that goes along with that, whereas others may look to accumulate and keep wickets in hand for the final (“death”) overs, when they will look to score very quickly. According to Nathan Leamon, the England cricket team’s Lead Analyst and co-founder of CricViz, the role of analysis in cricket is to provide a clearer view for captains, coaches and players and to reveal patterns that are not otherwise apparent (Sky Sports, 2021). Given the number of matches each team in the IPL play, understanding the tactics and strategies of opposition teams is not straightforward without detailed analysis and as a result this could provide useful insight for the aforementioned captains, coaches and players.

## Literature Review

The relationship between cricket and academic statistics dates back over two decades, and perhaps the most famous example is the Duckworth-Lewis method (Duckworth & Lewis, 1998). Cricket is a game that is frequently interrupted by adverse weather conditions and the Duckworth-Lewis method (now Duckworth-Lewis-Stern), first used in international cricket in 1997, was designed to set revised run targets in the second innings of limited overs matches that had been interrupted, in a fairer way than this had previously been done.

More recently, with the advent of more advanced techniques and programs, there has been a concerted effort not just within Cricket but all competitive team sports to make use of the vast amounts of data available on the internet. Despite this increased effort in recent years, literature searching relating to grouping players or teams directly has brought back few research projects. The majority of these articles look to predict the outcome of a specific league and season or even individual matches. For example, Clarke (1988) took an overarching view of batting performance by comparing the first and second innings of a one-day (one innings per team) match. This was to provide an estimation of the best possible scoring rates to chase a total and the chances of each team winning throughout the second innings.

Prakash, Patvardhan and Lakshmi (2017) have made various efforts to produce an automated selection of a playing lineup for the IPL league, using various bowling and batting metrics that are common in cricketing statistical analysis, ending up with an 11 man team incorporating an even balance of various different playing styles commonly found in cricket. Kapadia et al. (2019) looked at using historic IPL data, much like the analysis you will see in the following paper, with an aim to predict match outcomes and scores. A varied suite of machine learning techniques undertaken to achieve this found that Random Forest performed the best over probablistic models, however none of the techniques used produced an accurate enough model. This is typical of sports-focused predictive projects since there are so many uncontrollable variables at play.

Clarke and Norman (1999) explored a strategy in cricket that is specific to certain situations: when a batter should turn down runs to keep themselves at the striker’s end and keep a weaker batter off strike, therefore giving them a better chance of preserving wickets. Similarly, Swartz et al. (2006) and Norman and Clarke (2010) used complex statistical techniques including Markov chains and Monte Carlo simulations to explore batting orders in cricket. Swartz et al. found that some batting orders that contradict accepted wisdom in traditional cricket circles could be more successful than those currently used and Norman and Clarke found that, while batting orders traditionally do not change over the couse of an innings, greater rewards can be produced by varying batting orders based on the game situation. Joshi (2020) used network analysis to investigate effective batting partnerships and the worth of batters based on metrics that go beyond the runs they score and balls they face as an individual. This analysis builds on Swartz et al.’s findings and can be used to plan a more productive batting order. These studies demonstrate that the use of cricket statistics to improve a team’s performance is a growing field and the use of formal analysis of an evidence base, rather than the tactical hunch of people involved in the game, to improve performance could be seen as a precursor to the research carried out in this study.

Kampakis and Thomas (2015) approached with a seemingly monetary motivation, used key cricketing indexes such as strike rates, economy rates and averages of individual players to produce Naive Bayes and Random Forest models based upon rank based Pearson Correlation Coefficient scores. The results actually produced a more sensitive model when compared to the most popular gambling models, in predicting individual games. Caveated of course with a statement explaining “the overall level of accuracy is lower than that observed in many other sports” (Kampakis and Thomas, 2005).

Prakash et al.’s research revolves around outputting an even mix of various cricket roles such as openers, middles, finishers, pace bowlers, spinners and team roles such as captains. Clustering using K-means and evaulating number of clusters using the Elbow method, found 6 and 7 clusters of varying sizes for batters and bowlers, respectively. ReliefF algorithm was then taken to determine weighting of each feature within a cluster, with an aim of evaluating the key cricketing metric Most Valuable Player Index (MVPI), which is used both for batters and bowling and is calculated using the below functions:

Batting = (Player’s Batting Average / Tournament Batting Average) \* Runs Scored + (Player’s Batting Strike Rate / Tournament Batting Strike Rate) ^2 \* Runs Scored by the Player

Bowling = (Tournament Bowling Average / Player’s Bowling Average) + (Tournament Economy Rate / Player’s Economy Rate) \* 2 \* Wickets taken by the player

ReliefF works iteratively, taking a random vector (X) in turn and normalising feature values within each X, this is then compared to those in the same classes with the weighting being reduced if it is further away from a shared feature. Each cluster corresponds to a specific type of player, distinct for batters and bowlers and according to strict IPL team roster rules (see below), a player in a specific cluster would get preference in order of selection within a team.

* There cannot be more than 4 foreign players in the playing eleven.
* There must be one captain and one wicketkeeper in the playing eleven.
* There must be 2 openers, 3 middle order batters and 2 all-rounders.
* There must be at least one uncapped (player who hasn’t played for the country ) Indian player.

Using the results of the cluster analysis, a novel Cluster Based Index (CBI) is defined by Prakash et al. to calculate the actual ranks of a player within a cluster using feature value and feature weights calculated earlier. Using iterative choice, the 8 IPL teams were then scored according to correct choices of individual players. In 73.33% of matches, the team that more closely matched the team selected by the model won. Since T20 is known to be quite erratic and hard to predict, this is taken to validate the model produced in the study. There are uncontrollable factors such as injury whereby the results may have been skewed in one direction, however a deep learning approach using more features and not reduced to the MVPI metric would have been interesting to see. Prakash et al. does mention an approach whereby a Genetic Algorithm approach be used in tandem for team selection. This would look to overcome problems relating to these aforementioned uncontrollable factors. Non-playing factors are also glazed over here, much like other team prediction models, in contrast to sabermetrics approaches in Baseball whose feature repertoire is significantly wider and selection is more robust. These other, often ignored factors such as salary have a key impact on a team’s ability to change their lineup at will.

Despite these drawbacks, the model described here can be seen as a step toward greater possibility of teams to incorporate a more data based decision making when it comes to signing/releasing players. Despite statistical-based team selection techniques not being a new concept within competitive sports, there still remains a lot of work to be done in the field especially a highly volatile sport as Cricket.

## Business Understanding and Data Understanding

A need to determine and implement and effective team is imperative and inherent in nature to success of any sports team. This need branches into the strategic plays a coach may want to implement, depending on any opposing team that may be faced, i.e. – is a smart batting strategy and maximising total runs over each innings favoured in place of a more defensive strategy (Clarke and Norman, 1999) and taking more wickets.

Analysis of a team’s potential, not just as individuals but through batting partnerships is important to assess weaknesses and strengths. This can be useful not only for a well-established team to employ, it may be useful when scrutinising the exact placement of a new player that has been recruited within a team.

As has been discussed, cricket is a complex sport with an incredibly high ceiling not only for skill within the game itself but the statistical analysis possible from a single ball, over, innings or match. In fact the Duckworth-Lewis-Stern (Duckworth & Lewis, 1998) method is an internationally recognised mathematical function that is actively used in individual games of cricket, something which is very rare for many other team sports. With this in mind, it seems reasonable to develop upon the mathematical foundations that exist within the sport.

High-value competitions such as the Australian Big Bash League or the Indian Premier League have a great deal of potential for statistical-based decision making. The games are short, fast and intense, which contrasts a great deal against the five day long, sleep-inducing stereotypical perception that many have of cricket. The IPL alone has an estimated brand value of nearly $7 billion USD, and creates hundreds of millions in value to the GDP of India every single year (Duff and Phelps, 2019). Boasting billionaire owners, a global audience and the second highest average salary of all sports leagues in the world, the IPL presents a perfect opportunity for teams to exploit the vast statistical analysis within the sport.

Player salaries range from $28,000 for young, uncapped Indian players, through $740,000 for England white-ball captain Eoin Morgan to $2.4m for India captain and megastar Virat Kohli (Sportekz, 2021) and the wealth of the franchise owners means there are plenty of resources available to spend on data analysis. The difference in cost between some bespoke data analysis work and the players who could be brought in to try to change the fortunes of a franchise makes this a worthwhile avenue for franchises to explore in the interest of aggregating marginal gains. There is evidence that IPL franchises are already starting to tap into this market, with the England team’s Lead Analyst, Nathan Leamon, working closely with Kolkata Knight Riders (Roller, 2021) and Rajstahan Royals employing the services of cricket analytics company CricViz (CricViz, 2020).

Here we set out a clear but fairly open objective, which is to test whether publically available in-game data can be used to identify trends and clusters within players and teams. Since the data that is used here is hosted on an open website, there are no concerns regarding data security or data protection. To abide by the scope and scale of the project outlines, ball-by-ball data has been aggregated into individual innings and individual players. As stated previously the vast amount of wealth involved within the IPL makes this a particularly hot topic and further research that is triggered by the work presented here and others like it could prove incredibly useful to teams and the coaches making decisions.

We must digest that since this data is free to access, and no doubt individuals have attempted similar research efforts previously, why is this not standard practise within teams? Complex analysis is of course used daily in competitive sports teams and in recent years has formed the backbone of decision making in all aspects of a club. Our aims therefore do not seek to predict the outcome of a single game, or an entire season but more to understand the tactics of IPL teams and how these tactics differ between both teams and individuals. Batting strategies can vary greatly between teams in T20 cricket. For example, some teams have a top-order batter whose job is to bat for as much of the innings as possible, ‘anchoring’ the innings while other batters look to take advantage of this stability by scoring lots of 4s and 6s, meaning they often get out for fewer runs but score runs more quickly, whereas other teams may give their top order batters licence to score as quickly as possible and have middle order batters who provide more stability as an insurance policy. Any objective insight that can be gained into the batting strategies of rival franchises will be of interest to the coaches and captains of IPL teams, who will then be able to tailor their own team selection, bowling plans and field settings to counter specific oppositions.

Success will be measured not by any particular accuracy measure but by interesting trends/clusters. We hope to find whether it is possible to see differences between teams/players in their playing style. To keep the work we do succinct we will only focus upon batters, leaving space for future work to develop a model that studies bowlers, fielders and even all-rounders. Questions are limitless, is the reason the Mumbai Indians are so dominant due to their aggressive batting style? What makes Virat Kohli one of the best batters in the league?

There is a risk that we are unable to find any patterns or significant differences between the batting strategies of the franchises, whether because of the fairly basic nature of the data or the techniques that we employ. However, the relatively low cost of the work compared to the money spent by the owners every year on the wages of players and coaches make this a risk worth taking.

Ball-by-ball data for every IPL match since the competition’s advent in 2008 is available from https://cricsheet.org/. The entire dataset is split into individual games over the life course of the IPL, the most recent being in May of this year when the IPL was suspended indefinitely due to the COVID-19 pandemic for the second successive time. This constitutes 846 individual games, each with its own csv file. The data here is presented ball-by-ball for a total of 177,559 balls, which whilst useful is far beyond the scope of this project. Therefore innings data will form the basis of the model shown here, which results in a total of 11,076 innings across 33 variables including those present in the raw data and those calculated.

Luckily the data that is used in the following report is incredibly robust in terms of data quality. There are no null possibilities, no fields with multiple erroneous values and all possible data that is available is used. The nature of cricket makes the data almost perfect to work with, especially when dealing with ball-by-ball data; each run, the method said run(s) were scored and who they were scored by are all measured with an exact accuracy.

## Data Exploration and Pre-processing

The table below contains the fields included in the data.

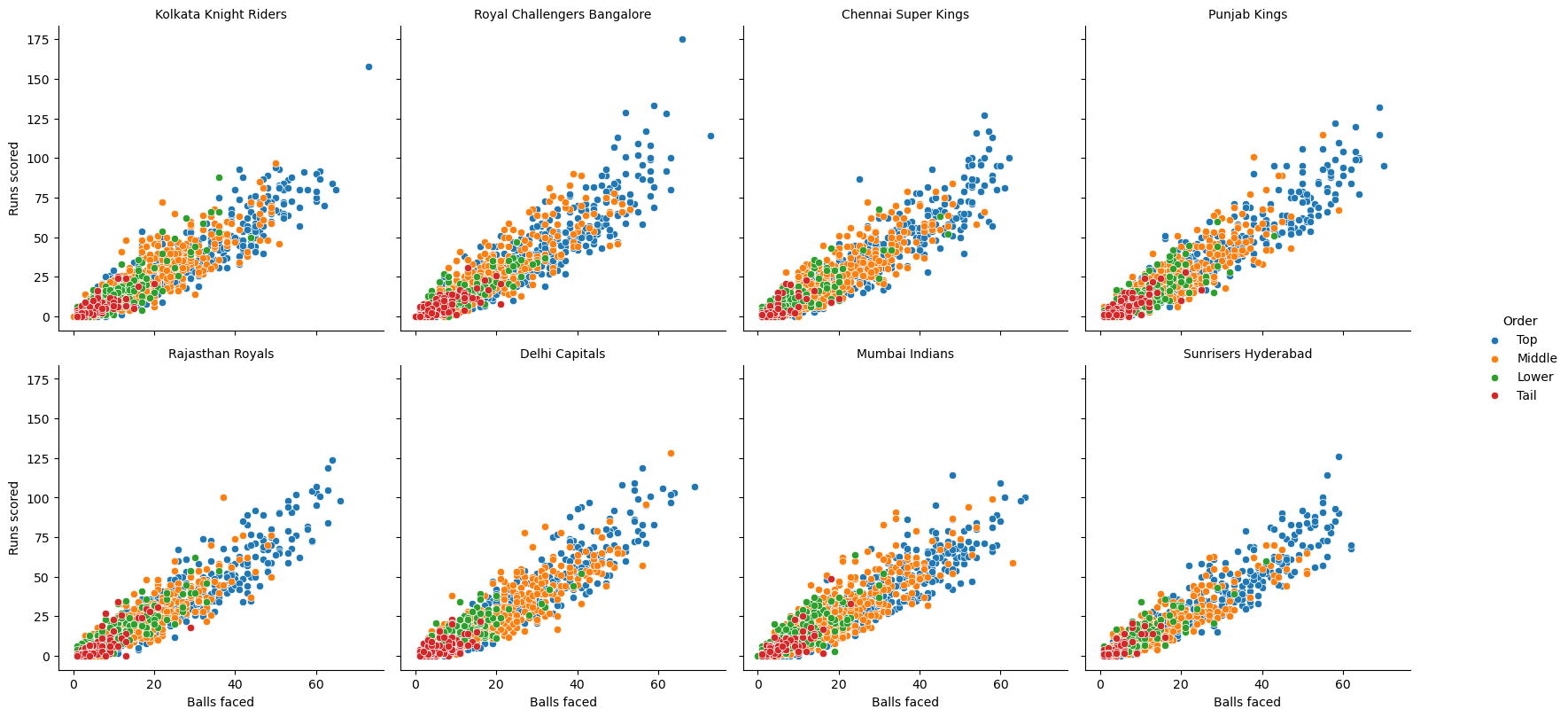
|  |  |
| --- | --- |
| Field | Description |
| match\_id | ID field, unique for each match |
| season | IPL season during which the match was played |
| start\_date | Date on which the match was played |
| venue | Venue at which the match was played |
| innings | Innings of the match |
| ball | Ball number. This is written in base 6, since cricket convention states that, for example, the third ball of the first over is known as ball 0.3. If a wide or no ball is bowled, an extra ball has to be bowled, meaning the ball number can go to e.g. 1.7 or above if a wide or no ball is bowled in the 2nd over. |
| batting\_team | Name of the batting team |
| bowling\_team | Name of the bowling team |
| striker | Name of the batter at the striker’s end |
| non\_striker | Name of the batter at the non-striker’s end |
| bowler | Name of the bowler |
| runs\_off\_bat | Number of runs scored by the batter from the ball in question |
| extras | Number of extras (wides, no-balls, byes, leg byes and penalty runs) scored from the ball in question |
| wides | Number of wides scored from the ball in question. A wide is given when a ball is deemed too wide for the batter to hit. |
| noballs | Number of no-balls scored from the ball in question. A no-ball is given if the bowler bowls from in front of the popping crease or the ball reaches the batter above waist height without bouncing. |
| byes | Number of byes scored from the ball in question. A bye is a run scored when the ball does not touch any part of the batter’s bat or body and the ball is not a wide. |
| legbyes | Number of leg byes scored from the ball in question. A leg bye is a run scored when the ball hits any part of the batter’s body apart from their hands. |
| penalty | Number of penalty runs scored from the ball in question. |
| wicket\_type | Method of dismissal (e.g. bowled, caught), if a wicket is taken from the ball in question. |
| player\_dismissed | Name of the batter who has been dismissed, if a wicket is taken from the ball in question. |
| other\_wicket\_type | This is always empty |
| other\_player\_dismissed | This is always empty |

In order to investigate individual innings, some pre-processing was required to clean the data and convert the ball-by-ball data described above into innings-by-innings data. First, innings played by teams other than the eight who currently compete in the IPL were removed. This is because a) some of these teams competed in the IPL for very short periods and so the data collected may not have been sufficient from which to draw any conclusions and b) the batting strategies of now defunct teams is of little interest to current coaches and players. The eight current teams are Chennai Super Kings (CSK), Delhi Capitals (DC), Kolkata Knight Riders (KKR), Mumbai Indians (MI), Punjab Kings (PK), Rajasthan Royals (RR), Royal Challengers Bangalore (RCB) and Sunrisers Hyderabad (SH). Punjab Kings used to be called Kings XI Punjab and Delhi Capitals used to be called Delhi Daredevils. Instances where the data contained the old names for these teams were fixed to include the new names.

If the two teams in an IPL match score the same number of runs (a tied match), a super-over is played. This is where the two teams play a single over each and the team who scores the most runs wins the match. If the first super-over is tied, a second super-over is played. No IPL match has needed more than two super-overs to decide the result. Super-overs are recorded in the raw data as innings 3, 4, 5 and 6 and since they are isolated events with distinct tactics, instances with *innings* greater than 2 have been removed. The pre-processing included aggregating the *runs\_off\_bat* field to get a rolling total of runs scored by a batter in an individual innings, as well as creating a rolling count of the number of balls faced and using the *runs\_off\_bat* field to create a rolling count of the number of 0s (dot balls) faced and 1s, 2s, 3s, 4s, 5s and 6s scored by a batter in an innings. The idea being that this will give us an idea of the way in which a batter goes about scoring their runs: whether they are happy to face dot balls and wait for a ball they can hit for 4 or 6 or whether they look to score from every ball they face but hit fewer boundaries.

The pre-processing described above allowed us to group the data by *match\_id* and *striker* and visualise the data as distinct innings played by individual batters. The scatter plot below shows the runs scored and balls faced in every individual innings by each of the eight current teams. The colours correspond to the position of the batter in the batting order, where 1 to 3 is classed as ‘Top Order’, 4 to 6 is classed as ‘Middle Order’, 7 to 9 is classed as ‘Lower Order’ and 10 and 11 are classed as the ‘Tail’.

#### Scatter plots showing runs scored and balls faced by batters in each IPL team



The scatter plots unsurpisingly show that the top order batters play the longest innings for all teams, since they have the opportunity to bat for the longest time. For some teams there appears to be less of a difference between the type of innings played by top- and middle-order batters. For example, DC and MI have had some longer innings played middle-order batters, which could be an indication that their top-order batters are more likely to be dismissed early, allowing the middle-order batters to bat for longer. Whereas top-order batters from other teams seem to score at a similar rate regardless of the length of their innings, RCB and CSK top-order batters appear to accelerate the scoring rate when they are able to play long innings.

A batter’s strike rate is the number of runs scored divided by the number of balls faced, multiplied by 100 and, as a rate of run scoring, can be used as a proxy for batting aggression. The line graphs below show how the strike rates of top-, middle-, lower-order and tailend batters have changed for each team over the history of the IPL.

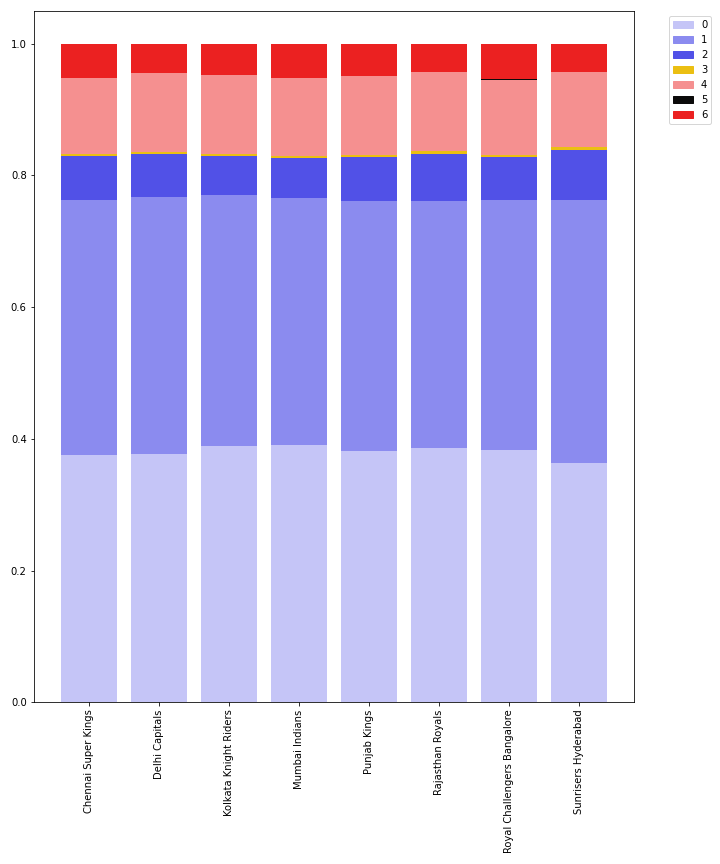
#### Line plots showing strike rates of different type of batters in each IPL since the advent of the competition



The patterns vary greatly over the history of the competition, in particular the strike rates of tailend batters. This is an indicator of the unpredictability of tailend batters, largely due to their relative lack of batting ability and the fact they are often given licence to try to score very quickly without too much concern for preserving their wicket. The relationship between the strike rate of top- and middle-order batters is consistent across time for some teams, such as MI and RCB, whose middle-orders score more quickly than their top-orders, RR, whose top-order batters score more quickly than their middle-order batters and CSK, whose top- and middle-order batters have always scored at similar rates.

Runs scored, balls faced and strike rates can only tell us so much about tactics. For example, it is impossible to tell whether a batter dismissed for 20 runs off 15 balls started slowly and had just started to accelerate when they were dismissed or whether they had scored at a constant rate throughout, or indeed what their strategy would have been had they not been dismissed when they were. The way in which an innings was compiled, the number of dot balls faced and the way a batter scored their runs, whether by hitting 1s and 2s or 4s and 6s, may help to further differentiate between different types of innings. The stacked bar plot below shows the proportion of balls faced by each team from which they scored 0, 1, 2, 3, 4, 5 and 6 runs off the bat.

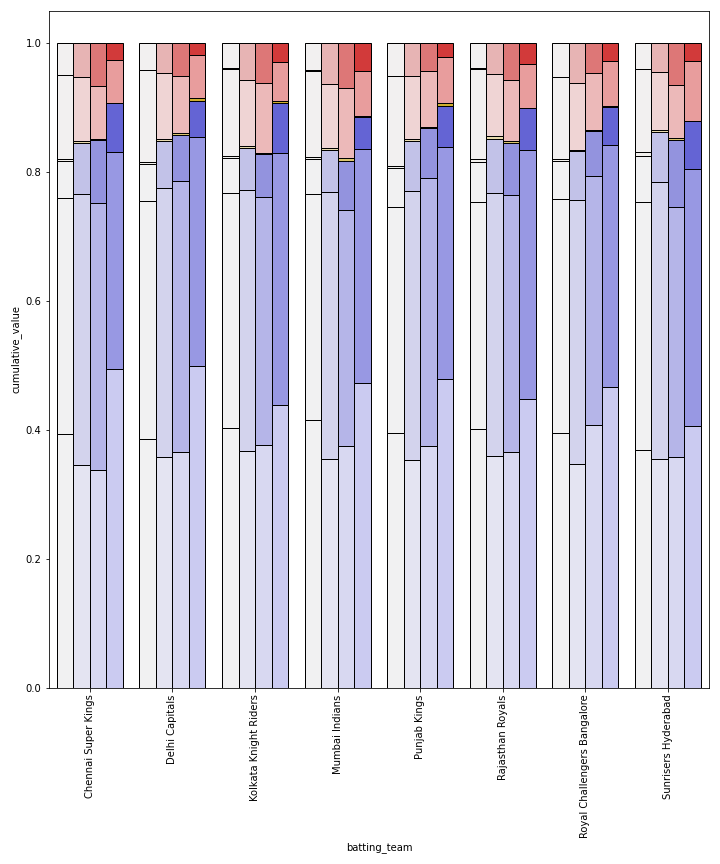
#### Stacked bar plot showing types of scoring shots by batters in each IPL team



It is evident from the above bar plot that MI face more dot balls but hit more balls for 6, whereas SH face fewer dot balls and run more 2s, but hit fewer balls for 6. More than anything, this graph shows that 3s and 5s are so rare in IPL cricket that they should not be included in the analysis as they are not likely to feature in the batting strategy of any of the teams.

The bar plot below shows how innings are compiled differently by top-, middle- and lower-order batters and tailenders.

#### Stacked bar plot showing types of scoring shots by batters in each IPL team, separated by batting order category



Top-order batters face more dot balls than middle-order batters in all teams and more than lower-order batters in all teams apart from RCB. The obvious explanation for this is that the first 6 overs of an IPL match are known as the PowerPlay, where only two fielders are allowed more than 30 yards from the wicket. This provides more opportunity for hitting the ball to the boundary but also fewer opportunities to run 1s and 2s. The top-order batters of PK hit a higher proportion of the balls they face for 4 or 6 than the other teams, most notably SH.

## Data Modelling and Model Evaluation

### Player-team batting order: classification

To first gain understanding in the general order of batsman, and how this affects the style of play by individual batsman as described in the data exploration stage, the first analysis done revolved around using various supervised learning techniques of classification. This was undertaken to cement the general trends seen in the above visualisations by using a number of calculated metrics that are commonly used in cricketing analysis. The aim of this classification section is to see whether or not batting statistical measures can be used to estimate the batsman order (top, middle, lower, tail).

The data at this stage was per innings, however for the remit of this section these innings data were summarised into per player/team partnerships. The metrics calculated from the data are shown in the table below.

|  |  |  |
| --- | --- | --- |
| Metric | Description | Calculation |
| strike\_rate | average number of runs scored per 100 balls | (total runs / total balls faced) \* 100 |
| ran\_runs | number of runs physically ran | total runs - (runs from 4s + runs from 6s) |
| activity\_rate | how often batsman physically ran, without boundaries | ran\_runs / (total balls faced - (number of 4s + number of 6s)) |
| boundary\_rate | proportion of boundaries hit | (number of 4s + number of 6s) / total balls faced |
| average | a batsman average score per innings | total runs / number of innings |
| consistency | how consistent a batsman stays in the crease | total runs / number of times out\* |

\* This metric induced both *NaN* and *Inf* errors when batsman had only ducks (out without scoring) or were never taken as a wicket, so these were managed by taking the number of runs scored for consistency.

Both 20% and 30% test sizes were taken from the data, and models performed on both, deciding on a 20% test size. Inititally, five different classification methods were used; k-nearest neighbour (knn), support vector classifier (svc), stochastic gradient descent (sgd), random forest (rf) and decision tree classifier (dt). All models were cross validated using K-fold strategy, which splits the training dataset into consecutive k-folds, allowing for pr. The results are tabulated below.

|  |  |  |
| --- | --- | --- |
| Classifier | Train Accuracy | Test Accuracy |
| knn | 56.0% | 57.4% |
| svc | 61.4% | 59.9% |
| sgd | 56.4% | 58.6% |
| rf | 61.7% | 60.4% |
| dt | 51.7% | 53.0% |

As can be seen, random forest is (very slightly) the most accurate model across the entire dataset. The results above show for 20% test sizes, with 30% being slightly lower in all cases except stochastic gradient descent. Random forest contains the highest testing accuracy of 60.4%, which admittedly is slightly dissapointing at first glace, support vector classifier was one of the most accurate techniques. The lowest accuracy, unsurprisingly was decision tree, since the classification was multi-class output.

Moving on, it was decided that since random forest had the highest accuracy on the test set, further tuning of the parameters within rf would be undertaken. To tune the rf model, RandomisedSearchCV was chosen, which is implemented through a randomised search over selected parameters, and each possible parameter pairing is sampled through a wide distribution testing - producing a fit method output. The best performing parameters are listed below.

|  |  |
| --- | --- |
| Parameter | Value |
| n\_estimators | 400 |
| max\_features | ‘sqrt’ |
| max\_depth | 300 |

The rf model was then ran with the above found parameters, fit and scored against out output prediction variable. This was further cross validated for accuracy score which resulted in a slight increase to 64.8%, which whilst is an improvement isn’t the best for the model overall. However upon further inspection, once the predicted values were appended back to the original dataframe further accuracy scoring could be investigated. Filtering the dataset to those batsman/team partnerships who have faced 50 or more balls, the accuracy of the model increases to just over 70% (70.8%). Further investigation found that for the unfiltered dataframe, the accuracy scoring was 93.8% for both correct values and those deemed ‘close’ (i.e. the model predicted a batting order category 1 away from the actual order), and 100% for the filtered over 50 balls faced dataset. Meaning that the predictions were within a small margin of error.

#### Mosaic plot for unfiltered dataframe (test split): random forest classification

Chart, treemap chart

Description automatically generated

#### Mosaic plot for filtered dataframe of those faced 50 balls or over (test split): random forest classification

Chart, treemap chart

Description automatically generated

It is clear from the above mosaic crosstab plots that the easiest batting order category ot predict was the ‘Top’, with the accuracy decreasing the further down the order you go, evening it out between lower and tail. This is very likely down to the specific batting style that a top order batsman might have, there are similar themes within both of the dataframes visualised. However it is very apparent that for the tail order batsman, there is a large increase in both the ‘incorrect’ and ‘exact’ results, which might point to a random assignment being correct. This is because there were less balls and therefore less data to go on through a partnership between batsman and team. Some of the metrics calculated therefore would be less sensitive to actual batting ability. However this also might be due to the a tail batsman having a very distinct batting style, orders from the team and a very different situation; perhaps a more frantic one. It’s also worth noting that the lack of tail batsman values in the filtered dataframe is due to these batsman having less chance to bat and therefore lower number of balls faced.

This classification modelling leads nicely into the following section regarding unsupervised clustering of individual batsman, with the knowledge that it is possible to some degree to predict a batsman order category.

### 

### Team batting strategies: k-means clustering

K-means clustering is an unsupervised learning method that uses Euclidean distances to group individual observations into a number, k, of clusters. We will perform k-means clustering on the innings-by-innings data in an attempt to reveal more information about the batting strategies used by different IPL teams.

The parameters used in the k-means algorithm, as defined in the scikit-learn documentation (Scikit-learn, 2020), are described in the table below.

|  |  |  |
| --- | --- | --- |
| Parameter | Description | Default value |
| n\_init | Number of times the k-means algorithm will be run with different centroid seeds. | 10 |
| max\_iter | Maximum number of iterations of the k-means algorithm for a single run. | 300 |
| max\_nclus | Maximum number of clusters the k-means solution will have. | 20 |

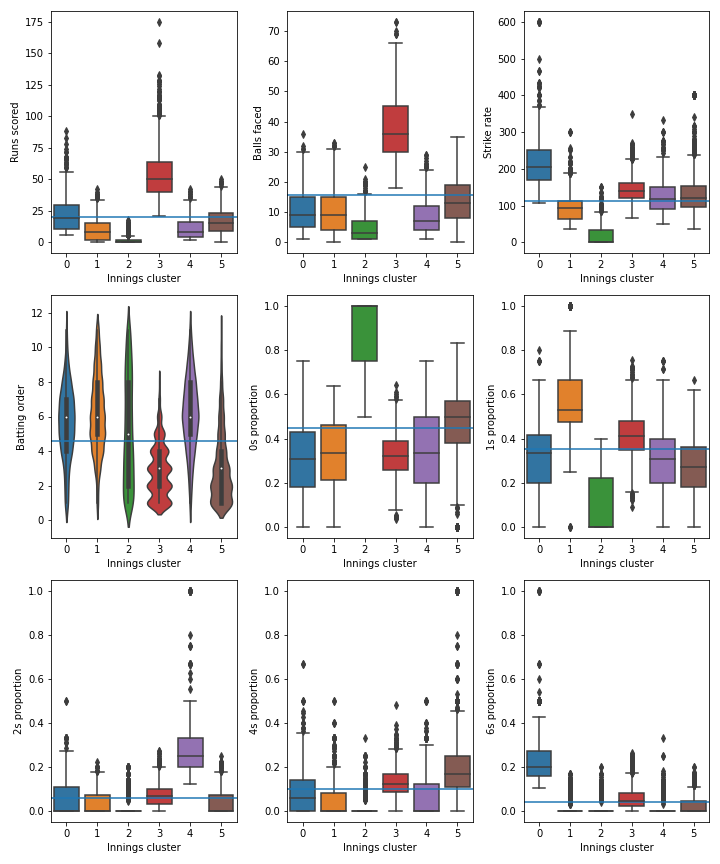
The k-means method used calculates the sum of squared error, that is the sum of the Euclidean distance from each point to its cluster centroid, for solutions with a range of *k*s, from 1 to the user-defined maximum, *max\_nclus*. The ‘knee-method’ was then used to find the number of clusters that gives the optimum compromise between the number of clusters and the discrimination between clusters. For the first k-means solution, the default values of all parameters were used.

The variables used to create the first k-means clustering solution are listed in the table below.

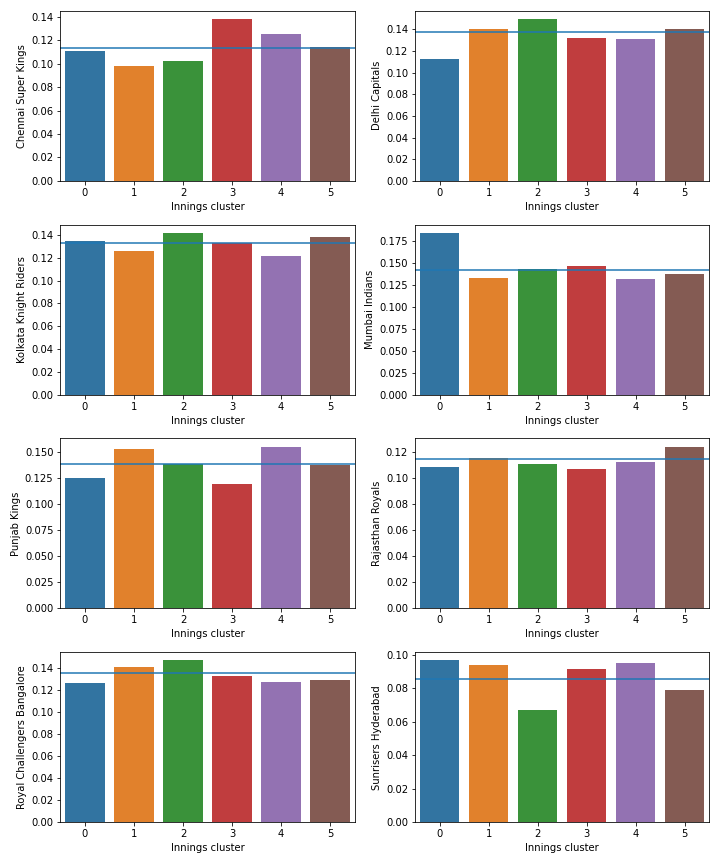
|  |  |
| --- | --- |
| Variable | Description |
| bat\_innings\_runs | Runs scored by batter in innings |
| bat\_innings\_balls\_faced | Balls faced by batter in innings |
| bat\_innings\_0s\_prop | Proportion of balls faced by batter in innings that were dot balls |
| bat\_innings\_1s\_prop | Proportion of balls faced by batter in innings from which 1 run was scored off the bat |
| bat\_innings\_2s\_prop | Proportion of balls faced by batter in innings from which 2 runs were scored off the bat |
| bat\_innings\_4s\_prop | Proportion of balls faced by batter in innings from which 4 runs were scored off the bat |
| bat\_innings\_6s\_prop | Proportion of balls faced by batter in innings from which 6 runs were scored off the bat |
| bat\_order\_striker | Batting position of the batter who played the innings (1-11) |
| bat\_order\_striker\_cat | Batting position category of the batter who played the innings (Top/Middle/Lower/Tail) |

The resulting clusters are then profiled against the numeric input variables, the results of which are shown in box/violin plots below, and against the eight IPL teams, the results of which are show in the bar plots. The horizontal lines on the box/violin plots show the mean value for each variable across all innings and the horizontal lines on the bar plots show the proportion of the total innings used to create the clustering solution that were played by the corresponding team.

#### Box and violin plots of numeric variables by cluster: solution 1



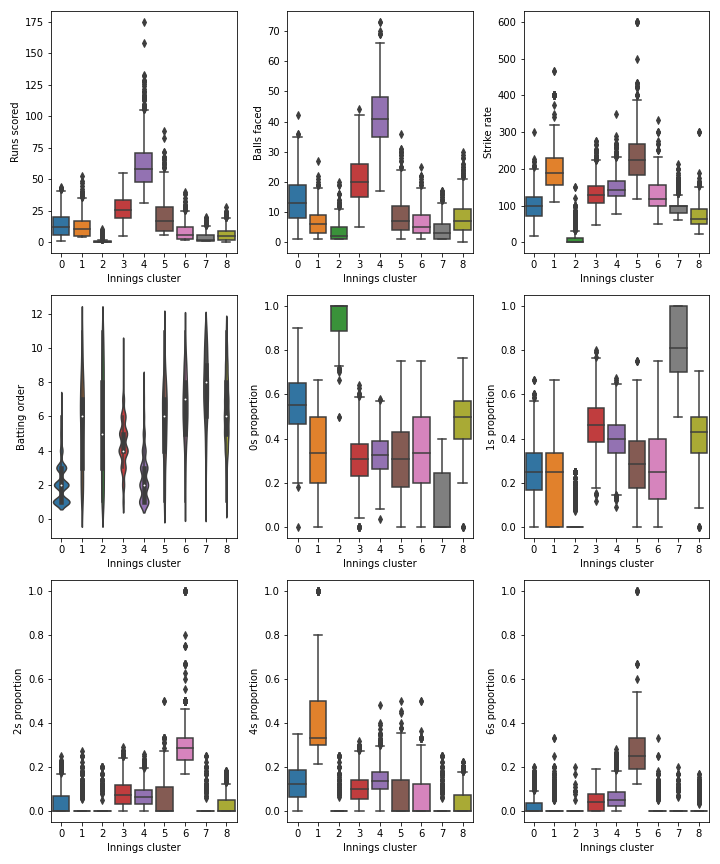
#### Bar plots of IPL teams by cluster: solution 1



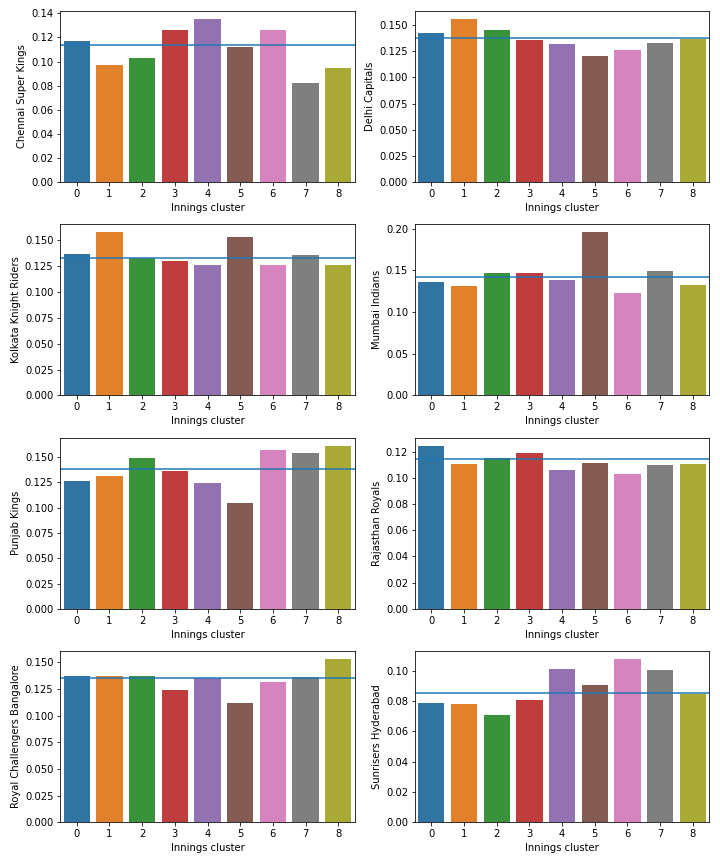
The first clustering solution contains 6 clusters and while the results are promising since the clusters appear to discriminate well between certain types of innings, they do not discriminate well enough between the teams.

For the second clustering solution, the same input variables are used, *max\_nclus* is increased to 40, *n\_init* is increased to 20 and *max\_iter* is increased to 600.

#### Box and violin plots of numeric variables by cluster: solution 2



#### Bar plots of IPL teams by cluster: solution 2



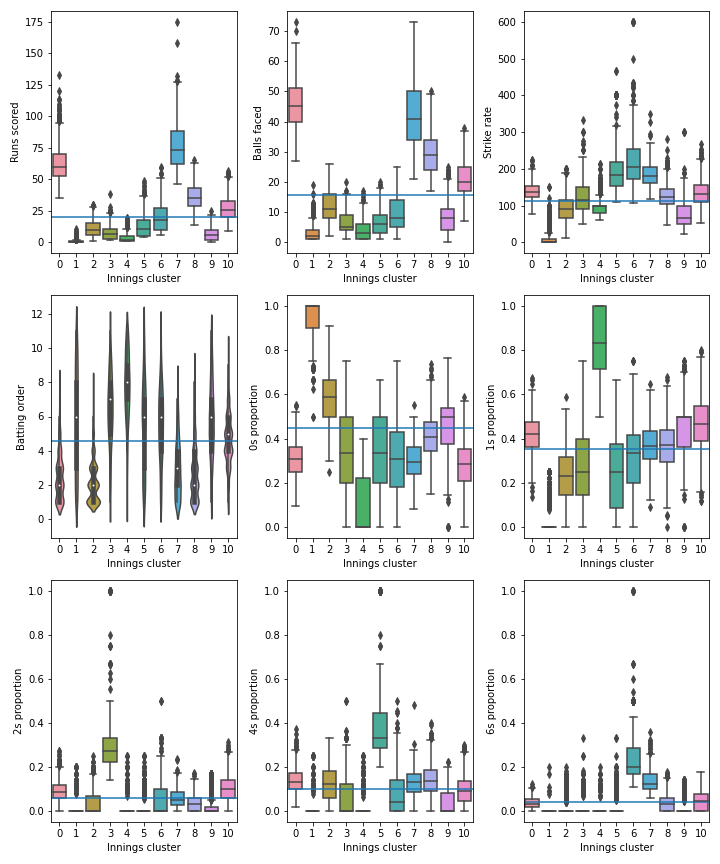
The clusters are described in the table below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Cluster | Average runs (average balls) | Innings makeup | Batter category | Teams |
| 0 | 14(14) | Over-indexed on dots and 4s | Top-order | RR |
| 1 | 13(7) | Over-indexed on 4s | Lower-order/tail | DC, KKR |
| 2 | 1(3) | Dot balls | Tail | CSK and SH good at avoiding these innings |
| 3 | 27(21) | 1s, 2s, 6s | Middle-order | CSK |
| 4 | 61(42) | 1s, 2s, 4s, 6s | Top-order | CSK, SH |
| 5 | 21(9) | 6s | Middle-/lower-order | MI, KKR |
| 6 | 8(7) | 2s | Middle-/lower-order/tail | SH, PK, CSK |
| 7 | 4(4) | 1s | Lower-order/tail | SH, PK |
| 8 | 6(8) | Dot balls, 1s | Middle-/lower-order/tail | PK, RCB |

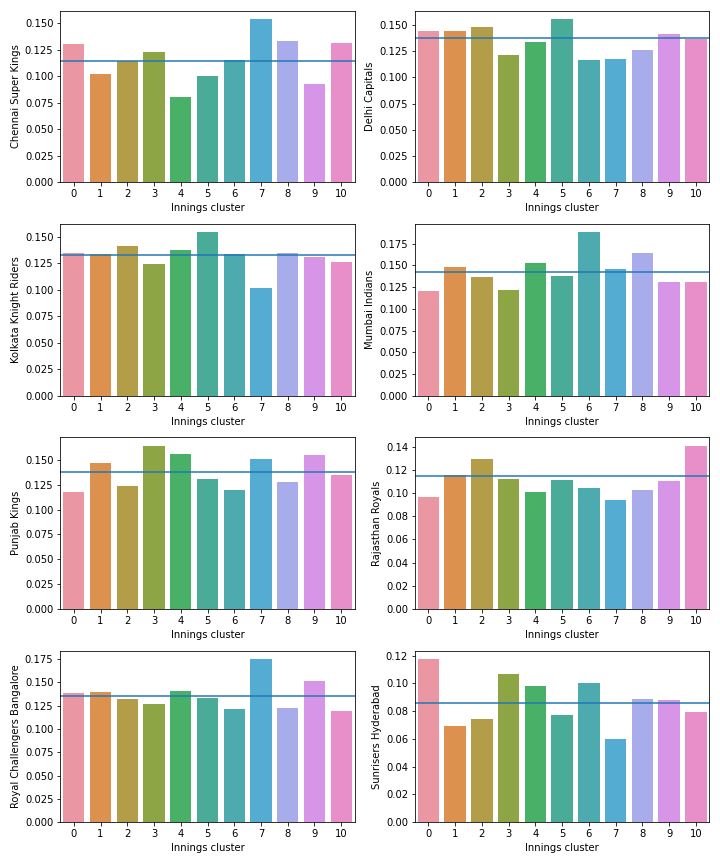
This solution offers a bit more discrimination between teams and therefore enables us to tell a bit more about the batting tactics employed by different teams. However, the clusters are still defined slightly too much by the number of runs scored and balls faced, rather than which scoring shots have been used to put together the innings. Scoring a lot of runs from a small number of balls is a universal goal of batting in T20 cricket and this therefore tells us more about who has been successful than how teams try to go about achieving this goal. For the third clustering solution, *max\_nclus* has been increased to 60, *n\_init* has been increased to 40, *max\_iter* has been increased to 1000 and the following variables have been added.

|  |  |
| --- | --- |
| Variable | Description |
| bat\_innings\_0s | Number of dot balls in an innings |
| bat\_innings\_1s | Number of balls from which 1 run was scored off the bat |
| bat\_innings\_2s | Number of balls from which 2 runs were scored off the bat |
| bat\_innings\_4s | Number of balls from which 4 runs were scored off the bat |
| bat\_innings\_6s | Number of balls from which 6 runs were scored off the bat |

#### Box and violin plots of numeric variables by cluster: solution 3



#### Bar plots of IPL teams by cluster: solution 3



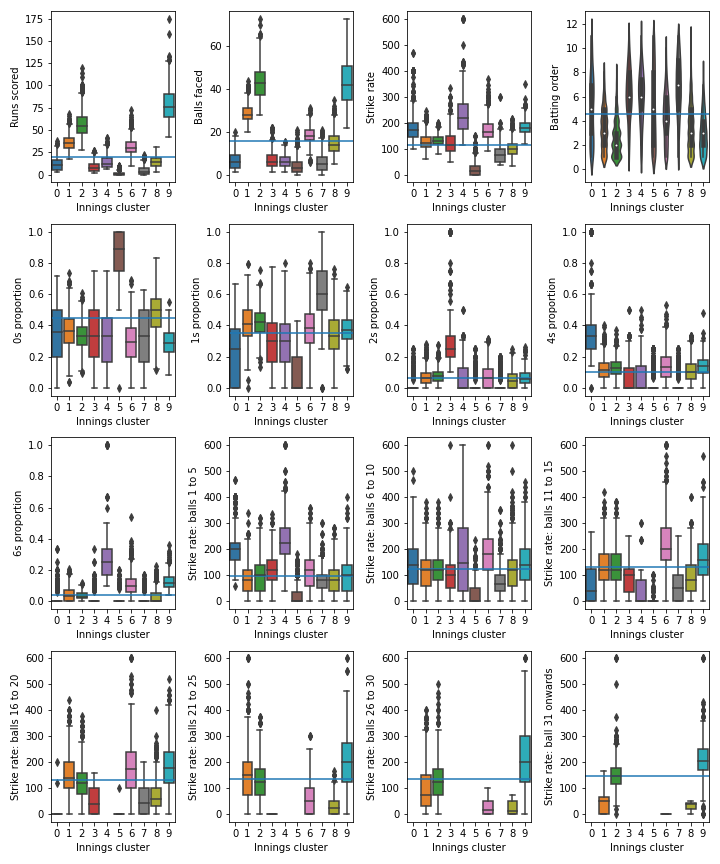
The clusters are described in the table below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Cluster | Average runs (average balls) | Innings makeup | Batter category | Teams |
| 0 | 63(46) | 1s, 2s, 4s | Top-order | SH, CSK |
| 1 | 1(3) | Dots | Lower-order/tail | CSK and SH good at avoiding these innings |
| 2 | 11(12) | Dots, 4s | Top-order | RR, avoided by SH and PK |
| 3 | 8(6) | 2s | Middle-/lower-order/tail | SH, PK |
| 4 | 4(4) | 1s | Lower-order/tail | SH, PK |
| 5 | 13(7) | 4s | Lower-order/tail | DC, KKR |
| 6 | 20(10) | 6s | Middle-/lower-order | MI, SH |
| 7 | 76(42) | 4s, 6s | Top-order | CSK, RCB |
| 8 | 36(29) | 4s | Top-order | CSK, MI |
| 9 | 7(8) | 1s | Middle-/lower-order/tail | PK, RCB |
| 10 | 27(21) | 1s, 2s | Middle-order | CSK, RR |

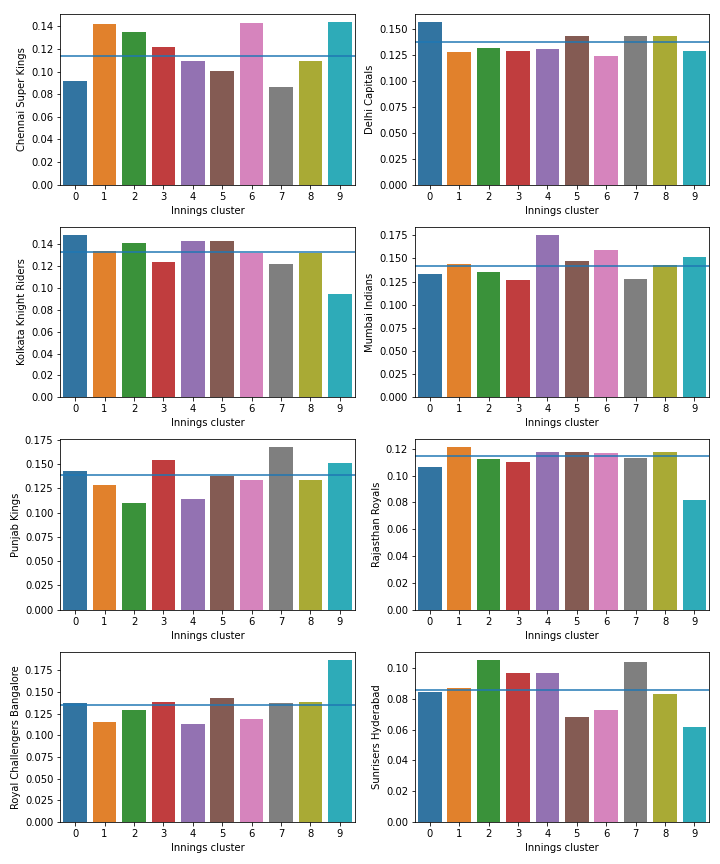
This solution is a big improvement, in that it there are more clusters that define what could be classed as ‘good’ innings and therefore that could be part of a genuine batting strategy. The solution contains two clusters with mean scores of greater than 50 but that are compiled in different ways: cluster 0 contains more 1s and 2s as well as some 4s, whereas the innings in cluster 7 contain the highest proportion of 4s and 6s. In addition to this, there are three clusters with mean scores between 20 and 40, again compiled in different ways and by different types of batter. Cluster 6 is defined by its high strike rate and relatively astronomical rate of 6s and the innings are played predominantly by middle-/lower-order batters, whereas the innings in cluster 8 are played mainly by top-order batters and contain more 4s and fewer 2s than the average innings, probably because of the powerplay fielding restrictions that are more often in place when top-order batters are batting. Cluster 10 is defined mainly by its low dot ball proportion and high proportion of 1s and 2s and the innings are played by hard-running middle-order batters. The fact that CSK over-index on five of these six ‘good’ innings is in part a reflection of the fact that they have reached eight of the 13 IPL finals, more than any other team, but also gives an insight into a type of innings that their batters do not often play, in cluster 6. One thing this solution does not tell us is how the strike rate progresses through the innings. To differentiate on this basis, for the next solution the following features are added.

|  |  |
| --- | --- |
| Variable | Description |
| bat\_innings\_1\_5\_strike\_rate | Strike rate for balls 1 to 5 of a batter’s innings |
| bat\_innings\_6\_10\_strike\_rate | Strike rate for balls 6 to 10 |
| bat\_innings\_11\_15\_strike\_rate | Strike rate for balls 11 to 15 |
| bat\_innings\_16\_20\_strike\_rate | Strike rate for balls 16 to 20 |
| bat\_innings\_21\_25\_strike\_rate | Strike rate for balls 21 to 25 |
| bat\_innings\_26\_30\_strike\_rate | Strike rate for balls 26 to 30 |
| bat\_innings\_31plus\_strike\_rate | Strike rate for ball 31 onwards |

#### Box and violin plots of numeric variables by cluster: solution 4



#### Bar plots of IPL teams by cluster: solution 4



The clusters are described in the table below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Cluster | Average runs (average balls) | Innings makeup | Strike rate pattern | Batter category | Teams |
| 0 | 12(6) | 4s | Very fast start (first 5 balls SR: 200) | Lower-order | DC, KKR |
| 1 | 36(28) | 1s, 4s | Slow start, accelerate through balls 11 to 25 | Top-/middle-order | CSK |
| 2 | 57(43) | 1s, 2s, 4s | Slow start, accelerate throughout innings (31+ balls SR: 152) | Top-order | CSK, SH |
| 3 | 8(7) | 2s | Relatively fast start (first 5 balls SR: 126) | Lower-order/tail | PK, SH |
| 4 | 14(6) | 6s | Very fast start (first 5 balls SR: 243, next 5 balls SR: 168 | Lower-order/tail | MI |
| 5 | 1(4) | Dot balls | N/A | Tail | Fairly universal |
| 6 | 31(18) | 2s, 4s, 6s | Fairly quick start, accelerate rapidly (balls 11 to 15 SR: 220) | Middle-order | CSK |
| 7 | 5(6) | 1s | N/A | Lower-order/tail | PK, SH |
| 8 | 15(15) | Not conclusive, slightly more dots and 4s than average | Below average throughout innings | Top-order | Fairly universal |
| 9 | 79(43) | 4s, 6s | Relatively fast start, accelerate steadily throughout long innings | Top-order | RCB, CSK |

There are some clear similarities between the clusters in solutions 3 and 4, as illustrated by the table below, which contains the percentages of the solution 4 clusters that are in each of the solution 3 clusters.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Solution 3 Cluster | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Solution 4 Cluster |  |  |  |  |  |  |  |  |  |  |  |
| 0 | 0 | 0 | 13 | 0 | 0 | 82 | 0 | 0 | 0 | 3 | 2 |
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 61 | 0 | 35 |
| 2 | 69 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 26 | 0 | 3 |
| 3 | 0 | 0 | 2 | 82 | 1 | 0 | 0 | 0 | 0 | 7 | 8 |
| 4 | 0 | 0 | 2 | 0 | 1 | 1 | 91 | 0 | 0 | 4 | 1 |
| 5 | 0 | 66 | 20 | 0 | 0 | 0 | 0 | 0 | 0 | 14 | 0 |
| 6 | 0 | 0 | 3 | 1 | 0 | 5 | 23 | 2 | 18 | 1 | 47 |
| 7 | 0 | 0 | 0 | 0 | 36 | 0 | 0 | 0 | 0 | 62 | 1 |
| 8 | 0 | 0 | 49 | 0 | 0 | 2 | 2 | 0 | 7 | 24 | 17 |
| 9 | 14 | 0 | 0 | 0 | 0 | 0 | 0 | 82 | 3 | 0 | 1 |

For example, 91% of the innings in solution 4, cluster 4 are also in solution 3, cluster 6 and 82% of the innings in solution 4, cluster 9 are also in solution 3, cluster 7. However, the addition of the features that capture the strike rates throughout the innings has had some impact on the solution. The majority of innings in solution 4, cluster 2 are also in solution 3, cluster 0, but 26% of them are in solution 3, cluster 8. The innings in solution 3, cluster 0 are longer in terms of runs scored and balls faced than in solution 3, cluster 8, but the two have similar profiles in terms of their strike rates, how the runs are scored and by which type of batters. When the strike rate progression is also taken into account, it makes sense that some of the innings from these two clusters should be grouped together.

This research has uncovered some results that could be significant for a captain or coach of an IPL franchise team and some that are already well known. For example, the fact that Chennai Super Kings are the team most likely to have top-order batters play a long innings, made up of 1s, 2s and 4s and with a strike rate in the 130s, and the second most likely to have top-order batters play a long, explosive innings with a strike rate in the 180s and lots of 6s, is not a surprise as they have been the most consistently successful team in the competition since its advent. However, that Sunrisers Hyderabad are the second most likely team to have batters play the former of the aforementioned types of innings and the Royal Challengers Bangalore are the most likely to have batters play the latter is a real insight. This information could be useful in helping opposing captains and coaches to know to prioritise dismissing the top-order batters of Sunrisers Hyderabad and Royal Challengers Bangalore to prevent them from causing damage and giving their own team the best possible chance of winning. Likewise, the fact that the lower-order batters of Delhi Capitals and Kolkata Knight Riders (solution 4, cluster 0) and Mumbai Indians (solution 4, cluster 4) often play similar innings in terms of runs scored and balls faced, but that the former two hit more 4s and the latter hit more 6s can help opposing teams to ensure they a) do not become complacent when bowling to the lower-order batters of these three teams and b) adjust their bowling and fielding tactics accordingly to counter their opposition’s favoured method of scoring runs (4s or 6s). The clusters mentioned above describe innings that could be broadly described as being ‘good’ for the batting team, but the research has also uncovered some useful insights about what could be described as ‘poor’ innings for the batting team. For example, the fact that Rajasthan Royals’ top-order batters are the most likely to play slow innings that waste the advantageous conditions of the first 6 overs (solution 3, cluster 2) could lead opposing captains to target this stage of the innings with their best bowlers and aggressive fielding to gain an early advantage.

## Summary

The modelling classification of batsman order categories aimed to introduce the possibilities with cricketing analysis. Despite using a number of different supervised learning methods, the overall accuracies were at first somewhat dissapointing, around 50%s to early 60%s. However during further inspection of the results, a trend was seen within the lower batsman category. A learned suggestion that lower batsman would have generally faced a lower number of balls, by using the more analytically robust data it was found that the accuracy did increase. This supports (albeit not fully) the hypothesis that it is possible to use cricketing metrics (strike rate, boundary rate etc.) to predict batsman order. A change in batsman order also brings with it a change in batting styles and tactics. Therefore the aim of the clustering section which seeks to see if tactics between teams (and thus individual batsman) changes, starts off with a solid foundation.

The section of the research dedicated to using k-means clustering to investigate batting strategies favoured by certain teams uncovered some interesting information about the different types of innings played by batters for different IPL teams, as described in the data modelling and model evaluation section. However, despite the addition of the variables that capture the strike rates at different stages of an innings, the time series element of different innings has not been captured as effectively as it perhaps could be if more advanced techniques than k-means were applied to the problem. One such technique is dynamic time warping (DTW) which can be used to calculate differences and similarities between two time series. An opportunity for future research is to apply DTW to the raw, ball-by-ball data above to pick out similarities between innings of very different lengths, in terms of runs scored and balls faced. An innings in cricket is a time series that can be brought to an end for different reasons (the dismissal of the batter in question, the dismissal of all 10 of the batters teammates, the end of the 20 overs or the successful pursuit of the target), so this could prove valuable in grouping innings played in the pursuit of similar strategies, even if they are not completely successful, and in differentiating between innings that look similar at the end (e.g. in terms of runs scored and balls faced) but that are composed at different speeds.

Another improvement that could be made is the addition of ball tracking data. As discussed in the Background and Project Aim section, ball tracking data is not widely available for free, so was not within the scope of this project. However, it is something that all IPL franchises would be able to afford to enrich the analysis used to help improve their results on the pitch. The addition of this data would lead to many more possibilities, the most obvious of which in this context is the addition of ball tracking variables to the k-means clustering solutions discussed above. This would enable captains and coaches to evaluate which kind of deliveries are particularly successful in taking wickets or vulnerable to being hit for 4s and 6s against certain teams. Any such greater insight into the batting strategies and strengths of opposition teams would help captains and coaches to counter these by targeting certain batters with types of bowlers that they are less successful against. This tactic is known as exploiting positive matchups and can be very successful, particularly in T20 cricket (CricViz, 2020). Throughout this report - research has been presented on batting only, which is a portion of the game of cricket; the surface has been scratched on the realms of possibilty within the game itself.

## References

Clarke, S. R. (1988). “Dynamic programming in one-day cricket- optimal scoring rates” *Journal of the Operational Research Society 39*, 331-337

Clarke, S. R., Norman, J. M. (1999). “To run or not?: Some dynamic programming models in cricket” *Journal of the Operational Research Society 50*, 536-545

CricViz. Cricket Intelligence at the Next Level. https://www.cricviz.com/

CricViz (2020, September 17). How Mumbai Indians Mastered The Art Of The Match-Up In T20 Cricket. https://www.cricviz.com/how-mumbai-indians-mastered-the-art-of-the-match-up-in-t20-cricket/

Davis, J., Perera, H. & Swartz, T.B. (2015) “Player evaluation in Twenty20 cricket” *Journal of Sports Analytics I*, 19-31.

Duckworth, F., Lewis, A.J. (1998) “A fair method for resetting the target in interrupted one-day cricket matches” *Journal of the Operational Research Society*, 220-227

ESPN Cricinfo (2008, January 24). Big business and Bollywood grab stakes in IPL. https://www.espncricinfo.com/story/ipl-announces-franchise-owners-333193

Gupta, V. (2019, September 20). Duff & Phelps Launches IPL Brand Valuation Report 2019. Duff and Phelps. https://www.duffandphelps.com/insights/publications/valuation/ipl-brand-valuation-report-2019

International Cricket Council (2018, June 27). First global market research project unveils more than one billion cricket fans. https://www.icc-cricket.com/media-releases/759733

Kampakis, S., Thomas, W. (2015) “Using Machine Learning to Predict the Outcome of English County twenty over Cricket Matches” Available: https://arxiv.org/abs/1511.05837

Kapadia, K., Abdel-Jaber, H., Thabtah, F., Hadi, W. (2019) “Sport analytics for cricket game results using machine learning: An experimental study” *Applied Computing and Informatics*

Lewis, M. (2003). *Moneyball: The Art of Winning an Unfair Game* W.W. Norton & Co.

Norman, J. M., Clarke, S. R. (2010). “Optimal batting order in cricket” *Journal of the Operational Research Society 61*, 980-986

Prakash, C.D., Patvardhan, C., Lakshmi, C.V. (2017) “AI Methodology for Automated Selection of Playing XI in IPL Cricket” *International Journal of Engineering Technology Science and Research 4*(6), 419-432

Roller, M. (2021, February 01). England analyst Nathan Leamon to join Kolkata Knight Riders. https://www.espncricinfo.com/story/ipl-2021-england-analyst-nathan-leamon-to-join-kolkata-knight-riders-1249726

Scikit-learn (2020). sklearn.cluster.KMeans. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

Sky Sports (2021, June 04). England’s lead analyst Nathan Leamon explains the role of data and analytics in cricket. YouTube. https://www.youtube.com/watch?v=UXPpsBmA-gE

Sportekz (2021, April 09). IPL 2021 Players Salaries. https://www.sportekz.com/cricket/ipl-2021-players-salaries/

Swartz, T.B., Gill, P.S., Beaudoin, D., deSilva, B.M. (2006). “Optimal batting orders in one-day cricket” *Computers and Operations Research 33*, 1939-1950.

Tewari, S. (2021, April 29). IPL viewership sees double digit decline in first two weeks over 2020. LiveMint. https://www.livemint.com/industry/media/ipl-viewership-sees-double-digit-decline-in-first-two-weeks-over-2020-11619709505228.html