International Men’s Cricket – Predicting the Market Value of Individual Cricketers

Problem statement

Cricket is a major international sporting business. Identifying players or teams with the highest chance of success is a highly lucrative business, both for traditional cricketing formats, but particularly for national or international cricketing franchises. Although a global sport, the largest international cricket market is in India, based upon its own domestic cricket product, but also through its leading “cricketing product”, the India Premier League (IPL), an annual international cricket franchise in which players from all the over the world participate in an auction where each franchise/team purchases its players for each season – see here for more information - https://www.iplt20.com . The most recent market value for the IPL brand as a whole put it at US$4.7bn for the 2021 season.

Choosing the right players for your team (whether for international competition or commercial competition) is big business. Being able to predict the right value for the best players and knowing which ones to avoid is crucial to both sporting and economic success.

Using data from the IPL’s player auction and linking it with individual cricketer’s performance data from previous leagues, this project aims to develop a model that can learn from previous auctions and player performance to predict a likely or realistic value or price for new or experienced/upcoming players. Overpaid upstarts or worth their weight in gold? What makes the most valuable IPL player and are they worth it? That’s what the project aims to find out.

Context

Although cricket may be seen as a niche sport, and is only played in a limited number of countries, it has genuine claims to being a global sport and a major international business. Stemming from Great Britain, and to some extents spread by the former British Empire, cricket is played domestically and internationally by a growing number of countries, although there are clear distinctions between top tier playing nations and those in other tiers.

However, the professionalism and commercialisation of the sport is breaking down those distinctions and barriers faster than at any other time in the history of the sport. In part this is due to the changing formats of the game. Historically, cricket matches were played over a number of days until they finished (with no set time for completion). This could result in games of significantly varying lengths, from a couple of days to a couple of weeks (games being based on two innings per side (and 11 players by side), with the winning side being the team with the highest overall score in those two innings). After the end of the Second World War, this practice stopped and the role of broadcasting rights (and associated advertising revenue) has played an increasing role in setting the parameters and formats for the game. So, from the end of the Second World War, international cricket became limited to five days, with an acceptance that there would be an increase in the number of unconcluded, hence drawn games. Over time, pressure to adapt the game at international level, and grow the business of the game has seen different formats emerge, from one-day limited-overs cricket (eg, 40, 50 or 60 overs per side, which is usually manageable within one day) to shorter formats such as T20 (high-intensity, fast-paced 20-over cricket, that is finished within a matter of hours, a form of heresy for older cricket fans brought up on longer, slower and more strategic forms of the game).

At the same time, the power base of cricket has moved from England/Great Britain to South Asia, in particular India, where the appeal, popularity, market and revenue for all forms of cricket is massive. Globalisation, the internet, cheaper/easier travel have also created international cricket competitions and formats outside of the historical/traditional competitions where one country competed against another in a Test (5-day cricket match) series. Now, most senior cricket playing nations will host their own domestic competitions, which are also genuinely international competitions open to cricketers and teams from around the world. Nowhere is this better exemplified than in the Indian Premier League, where a number of team franchises bid for Indian and international players at an auction to build a team that will compete in April/May of each year to win the annual trophy and lucrative prices. The audiences in the huge India stadia are massive, but are dwarfed by the national and international television/online audiences and the betting market associated with the games. Players from around the world flock to the IPL for the money and the fame, confounding the traditional domestic route to cricketing fame and glory - many players prefer to compete in the IPL than to be available for international selection in their own countries. It is a fact of life that national cricketing authorities have had to accept the financial reality of events such as the IPL or the Australian Big Bash and adapt their own local/national cricket leagues and competitions to fit in with the international markets. And the situation continues to shift, as participation in the IPL is now seen as beneficial to the development of certain or different types of cricketers at international level. There is always something to learn in both directions it seems.

Data wrangling

First, auction data was collected from the IPL website for the two most recent auctions:

- December 2022 for the current, 2023, season, and
- December 2021 auction for the previous 2022 season.

This represented auction price data for 1,005 players (600 from the 2021 auction and 405 from the 2022 auction), both sold and unsold.

Secondly, performance data on player performances in the 2021 and 2022 seasons was scraped from the IPL website, categorised in individual spreadsheets for batting, bowling and Most Valuable Player (MVP) metrics. When collated, cleaned and merged this resulted in performance data for just over 252 players.

Consideration was also given to finding additional performance data from outside the IPL, both in relation to pricing/value and to player performance. The majority of international players are also centrally contracted to their own countries cricket boards, with varying rates of contractual pay – a summary of this data is provided here for the top tier cricket playing nations - https://www.totalsportal.com/cricket/central-contracts-in-cricket/ . Complete data on all players' performance stats and averages up to 2019 can be found here on Kaggle, in all different formats: https://www.kaggle.com/datasets/mahendran1/icc-cricket. Given that there a number of different cricketing formats (including Test cricket (ie, long form played over five days), ODI cricket (ie, one-day internationals played with 50 overs per side) and T20 (ie, short-form, explosive, 20-over per side cricket, as found in the IPL), there's potentially a vast amount of data here, but given the different skills and player profiles required, it would be wise to use only T20 player performance data in relation to any IPL data models.

For the auction price data, the sold and unsold players' data for each of the 2021 and 2022 years was analysed. For 2021, c400 of 600 players were unsold, as were 325 of the c400 auctioned players in 2022. This immediately presented problems as it was highly unlikely that there would be performance data available for the majority of unsold players. Particular data wrangling issues included currency conversion from Indian rupees (and a numbering system that includes Crores and lakhs, which are not used in Western currency systems) into British sterling/pounds. A number of duplicates also had to be identified and removed from the merging of the 2021 and 2022 data, as well as corrections to the mislabelling of players' nationality.

For the player performance data, the data wrangling took the form of reviewing and cleaning the individual batting, bowling and MVP statistics for each of the 2021 and 2022 years, before merging them into a full dataset covering 2021 and 2022. The batting and bowling data for the two IPL seasons highlighted attributes that included:

- for batting:
    - number of matches/innings played,
    - highest scores, averages,
    - strike rates,
    - sixes and fours hit, and
- for bowling:
    - number of matches/innings played,
    - overs,
    - runs conceded,
    - wickets,
    - averages,
    - economy rate,
    - strike rate.

In addition, existing "best player" ratings (aka, MVP or most valuable player) for current season players were available.

The data was checked and corrected regarding data type, as many columns were not numeric, when they needed to be. Review also discovered a number of data inaccuracies or inconsistencies, which had to be addressed and corrected. Rows with missing values were dropped, and duplicate data in the MVP datasets were dropped, so that the MVP data ended up covering only dot balls, catches and stumpings (which weren't covered in either the batting or bowling data).

When the two years of data were initially merged, there were some missing data, which were then replaced with zeroes, as these represented data gaps where a player hadn't played in that season. Because there was data for 2021 and 2022, the two years had to be amalgamated, with a number of different arithmetical calculations (given that some represented simple totals, while others represented averages or rates), so that there was a single set of data for the different features, rather than one for each year. This reduced the overall number of features from around 50 to 28 for each player.

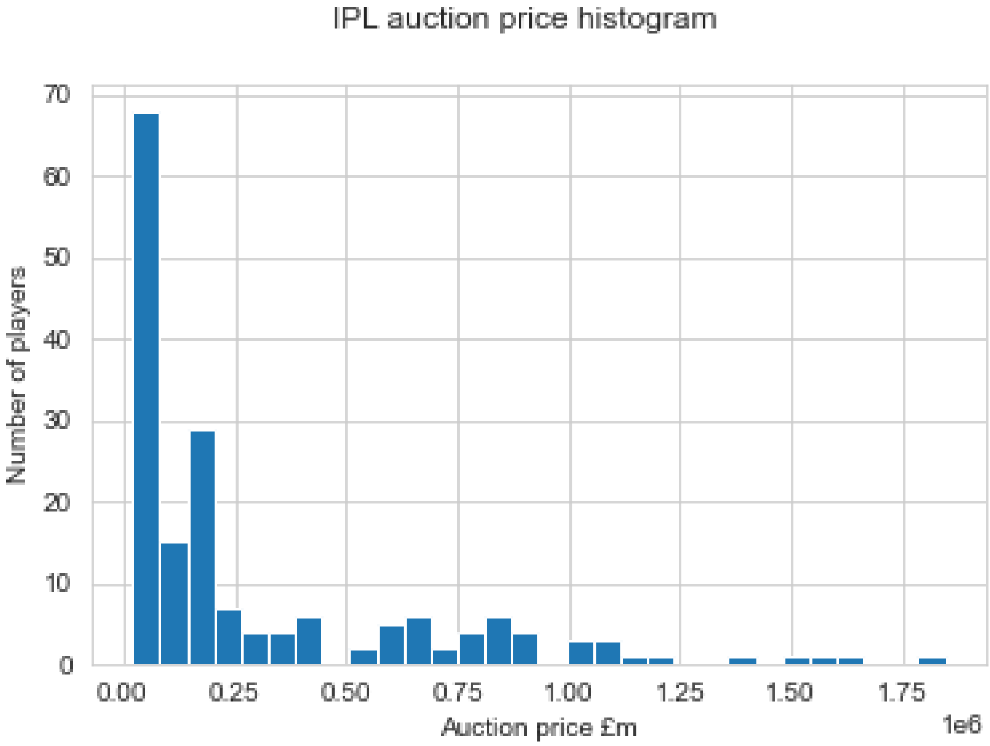
Work was also done on separate data regarding International Cricket Salaries, based on national cricket associations' approaches to remunerating their own players for international duties and participation. Ultimately, it was decided not to use this data, as it wasn't fully comparable across different countries and economies.

Finally, the pricing and performance data were merged, which resulted in only pricing and performance data being available for 175 players, even though the dataset included a total of 782 players (but the majority of these were unsold players with no performance data). Consideration was given to using the non-IPL international T20 data to fill in and increase the amount of player data that could be available, but it was decided against doing this due to the state of the data, the difficulties in working with some aspects of it and the time involved.

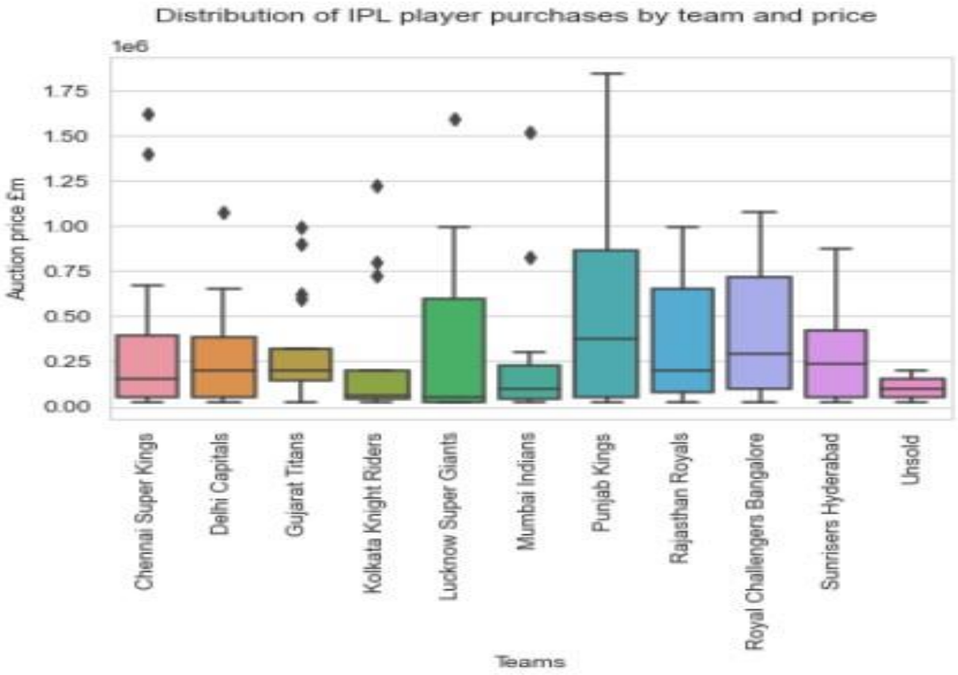
Exploratory Data Analysis

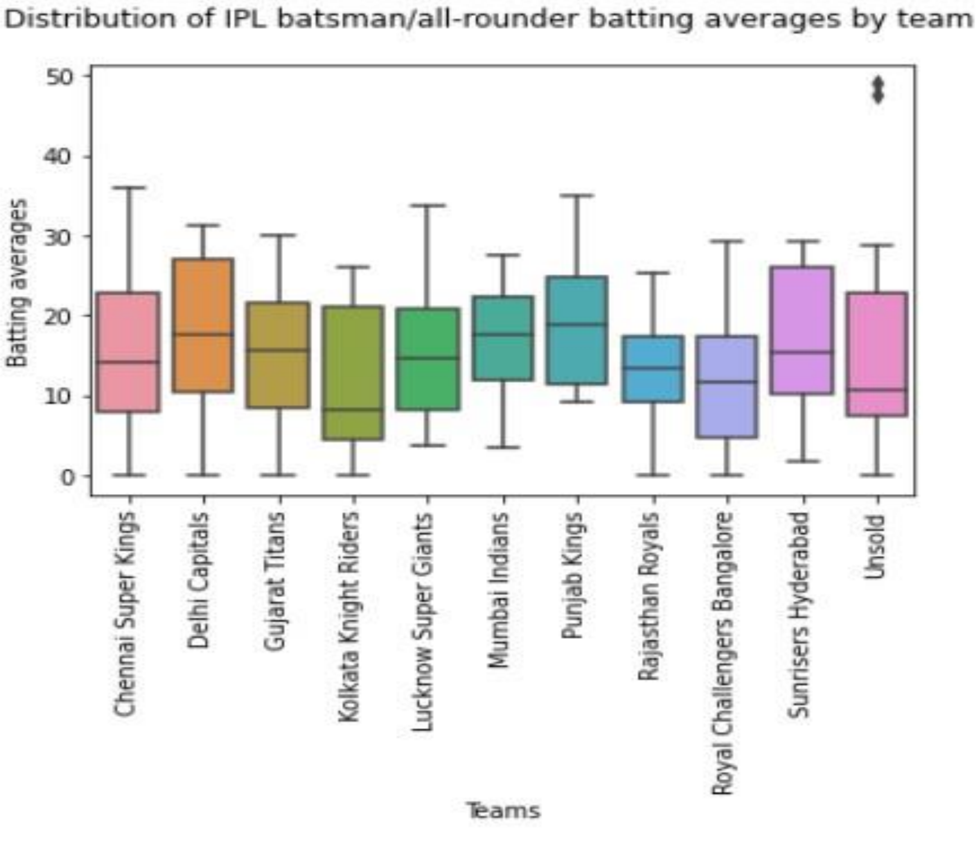
With the level, volume and very nature of cricket, you could easily get lost in data analysis with cricket. It's a statistician's dream. Given the different specialist roles within a cricket team (primarily batter, bowler and all-rounder, but also wicket-keeper), initial analysis looked at these different player categories, including stats such (for batsmen) batting averages, highest scores and strike rates (ie, number of runs scored per 100 balls), or (for bowlers) total wickets taken, bowling average, economy rates (ie, how few runs conceded per over), strike rate (ie, how few balls on average to take a wicket). Most of this analysis tended to show a handful of players at the top or extreme end of the ranking tables (effectively outliers).

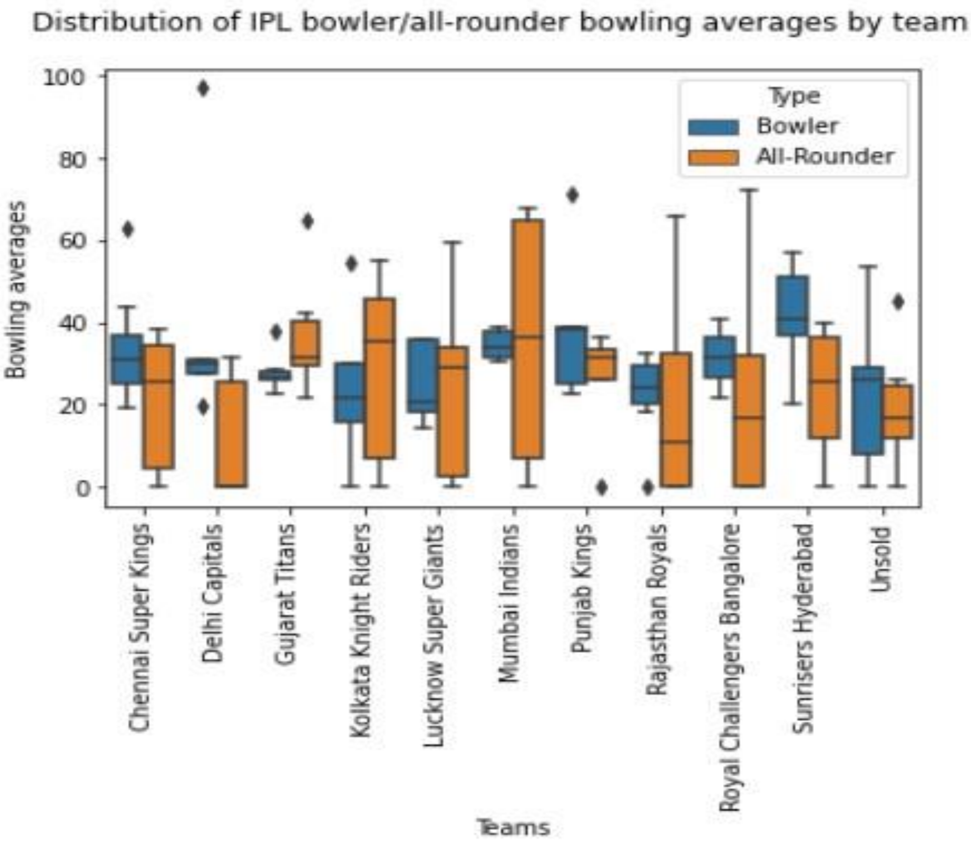
Auction prices themselves were analysed and it was interesting to see how few players went for very high prices (see table below)



Analysis was also carried out into purchasing distribution by team/tranchise, and how this then fitted with player performances, which showed some interesting approaches and differences – and confirmed there would definitely be an opportunity to analyse this in more detail to determine or model the success or otherwise of different team purchasing strategies. See example boxplots below. Don't forget: batting averages are better when they're higher (more runs) and bowling averages are better when they're lower (fewer runs conceded per wicket taken).

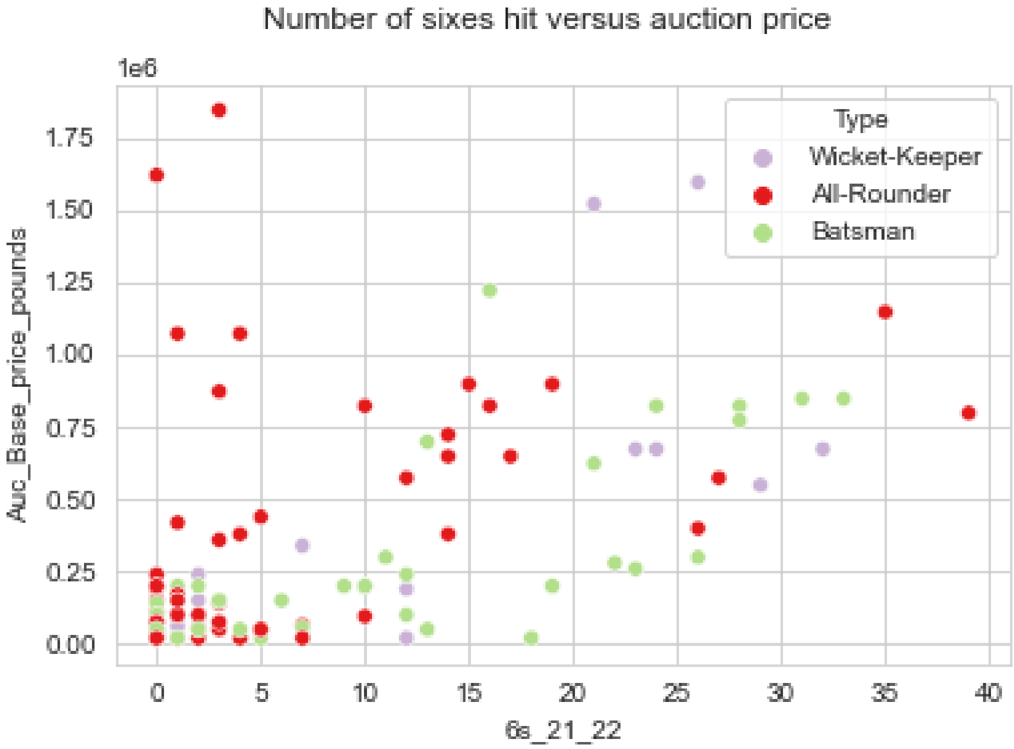

Distribution of IPL player purchases by team and price
1e6
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Auction price £m
Chennai Super Kings
Delhi Capitals
Gujarat Titans
Kolkata Knight Riders
Lucknow Super Giants
Mumbai Indians
Punjab Kings
Rajasthan Royals
Royal Challengers Bangalore
Sunrisers Hyderabad
Unsold
Teams

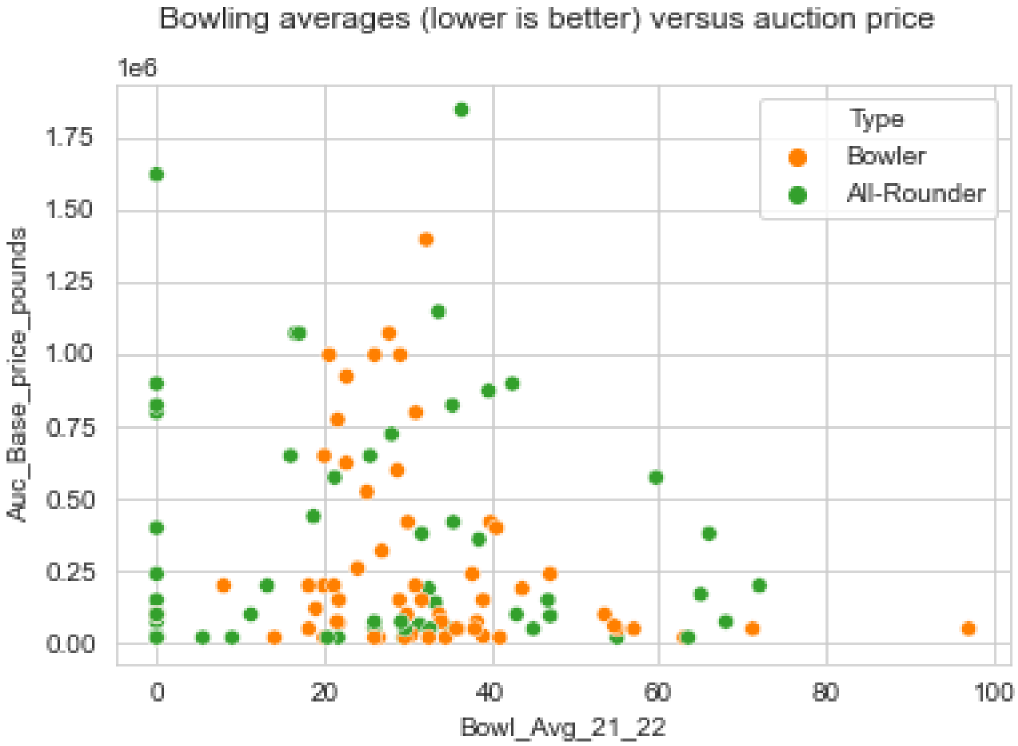

Distribution of IPL batsman/all-rounder batting averages by team
50
40
30
20
10
0
Batting averages
Chennai Super Kings
Delhi Capitals
Gujarat Titans
Kolkata Knight Riders
Lucknow Super Giants
Mumbai Indians
Punjab Kings
Rajasthan Royals
Royal Challengers Bangalore
Sunrisers Hyderabad
Unsold
Teams



Histograms were also plotted of various batting and bowling statistics, which highlighted the skewed and variable nature of most measures of cricket performance, and in the case of some of the bowling statistics highlighted some inaccuracies and unreliability with the bowling data, which were in part addressed later on. This concerned the fact that a significant proportion of bowlers were credited with “superhuman” bowling averages of 0, when in fact this was just an indication that they hadn’t taken any wickets. In normal cricket stats, this would be seen as an average of infinity (because the number of runs conceded would be divided by 0 and equated to infinity to indicate a meaninglessly high and thus unsuccessful bowling statistic, event though this is a mathematically incorrect approach). An example of this can be seen in the scatter plots later in this report.

Scatter plots were also created between key performance stats and auction prices, with further analysis by player type, which highlighted the variability between performance and price, as well as showing the performance value of many cheaper players and the great success of wicketkeepers as batsmen, as well as the valuable commodities represented by the majority of allrounders, which wasn’t always reflected in their pricing. See example scatter plots below.





Correlation heatmaps and all-feature scatterplots were produced for the data also. It proved hard to identify lots of strong correlations from the different plots - but there were positive relationships between pricing and batting qualities such as total runs (but not necessarily average), number of 50s scored, number of 6s, as you might expect. For bowlers, it was more about total wickets taken, bowling average, strike rate and also dot balls (so the number of times you bowl a ball that isn't scored off - certainly a valuable commodity in this form of cricket).

Pre-processing

At the pre-processing stage, a little further work was done on tidying up and removing some of the incorrect or distorting data found at the exploratory data analysis stage, namely re bowlers with misleading good averages, as indicated above. The feature data was also standardised. The "player type" (categorical data) was one-hot encoded, while a distinction was then made between continuous and discrete data types. Continuous data, such as total number of runs scored, batting and bowling averages, strike rates, etc, was standardised using StandardScaler, while discrete data, such as number of not-outs, 100s, 50s, 5-wicket-hauls was standardised using MinMaxScaler.

Model selection

The final dataset ended up with 26 features for the players, excluding their auction price. Before proceeding to model trial and selection, a principal component analysis was carried out, which identified that the majority of the data variance could be explained by the first four or five components. By constructing a loadings table and then plotting them, and looking at the weightings, it cold be seen that the following features stood out:

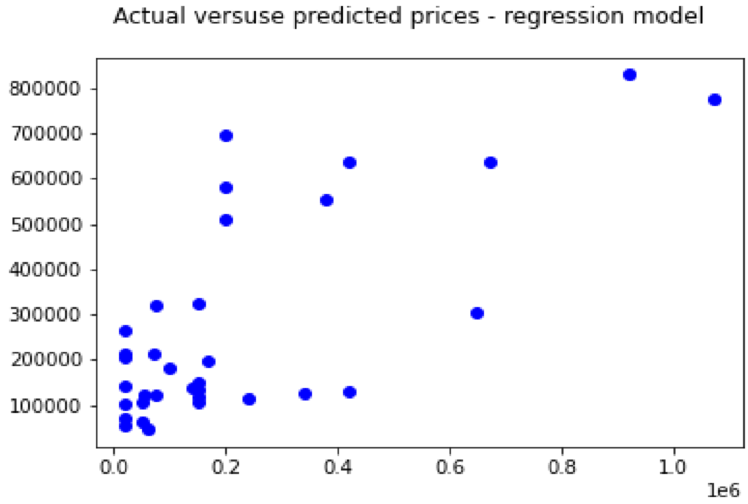
- Bowling economy
- Bowling strike rate
- Batting average.
- Batting runs total
- Bowling average.

As this was a regression problem, the following models were tried, with data split into training and test sets:

- Linear Regression
- Ridge Regression
- Lasso Regression
- RandomForestRegressor.

R-squared, the regression coefficient, was chosen as the key metric.

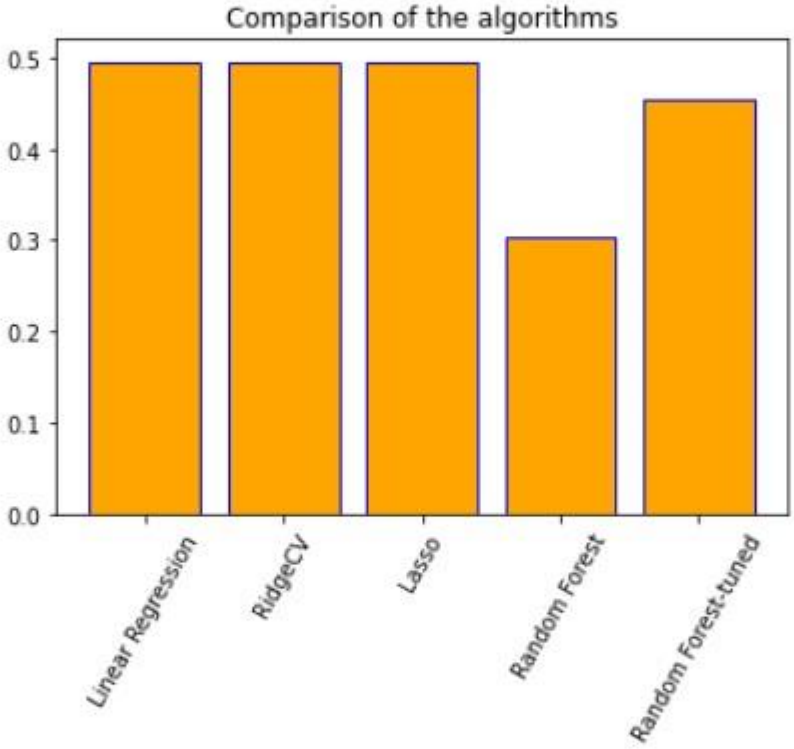
The performance of all models was poor – see linear regression example plot below, with a regression coefficient of about 0.49.



Hyperparameter tuning was tried out on the respective models, including:

- alpha and solver for Ridge Regression (involving the use of GridSearchCV),
- the use of an adjusted alpha for Lasso Regression (found from cross-validation on the training data), and
- a full parameter grid for the Random Forest Regressor (using GridSearchCV again).

The model results were as follows:



Conclusions and next steps

There may be a number of reasons why the models weren't successful enough:

- The dataset is too small - even though two years of pricing auction data were collected, the overall dataset ended up being cut down significantly due to a range of inadequacies (eg, missing data, inaccuracies).
- Distortion within some of the data, eg, bowlers having a bowling average of zero (which would be impossible, and so creates a mistaken understanding, given that it's really indicating that they, maybe, didn't bowl at all for some reason).
- The reality that players underperform or overperform in relation to their sale price, thus distorting the correlations.
- Limited engineering of the existing features within the dataset.
- Highly priced players not playing, due to injury or unavailability.
- The player categories may be unhelpful (especially for those who are bowlers only, and so tend not to have much data on other aspects of their cricket, eg, batting, while allrounders have a something of everything).
- Lack of corresponding data on team success – after all this is a team game, although it is driven by highly individualised performances (one of the many quirks or contradictions of cricket – it is often described as “an individual sport disguised as a team game”).

Some of these problems could be addressed by the following:

- Adding additional data from previous years, ideally by accessing an API from IPL or possibly purchasing fuller data from, eg, Statista.
- Adding data on team performance (eg, the game/match results), which may be an additional or better measure of success, given the team nature of the sport.
- Reworking the data to ensure inaccuracies and distortions are corrected, removed or properly addressed.
- Creating additional features to support different analysis, including across.
- Trying additional regression models.

Next steps

Although this work has been completed for a Capstone project within the Springboard Data Science, I intend to spend further time collecting additional data on previous years to see if I can build an improved model. I also plan to extend the analysis and potentially the modelling to look at team performance and how this fits with the different pricing strategies that seem to be apparent.