PREDICTION OF PLAYER PRICE IN IPL AUCTION USING MACHINE LEARNING REGRESSION ALGORITHMS

Dr. Jhansi Rani P
Department of Computer Science & Engineering,
CMR Institute of Technology,
Bangalore, India
jhansirani.p@gmail.com

Apurva Kulkarni
Department of Computer Science & Engineering,
CMR Institute of Technology,
Bangalore, India
apurvak407@gmail.com

Aditya Vidyadhar Kamath
Department of Computer Science & Engineering,
CMR Institute of Technology,
Bangalore, India
akamath996@gmail.com

Aadith Menon
Department of Computer Science & Engineering,
CMR Institute of Technology,
Bangalore, India
aadith.menon97@gmail.com

Prajwal Dhatwalia
Department of Computer Science & Engineering,
CMR Institute of Technology,
Bangalore, India
dh.prajwal@gmail.com

Rishabh D
Department of Computer Science & Engineering,
CMR Institute of Technology,
Bangalore, India
rishijain9999@gmail.com

***Abstract*— In this work, we have applied machine learning-based algorithms that predicts the cost at which a player can be sold in the Indian Premier League Auction. We estimated the players' selling price using their past performance parameters like runs, balls, innings, wickets and matches played. Tests were carried out in various machine learning models like Decision Tree Regressor, K-Nearest Neighbors (KNN), Linear Regression, Stochastic Logistic Regression, Random Forest Regressor and Support Vector Regression (SVR). Among these SVR and Linear Regression gave best results for predicting batsman and bowlers respectively. These algorithms can produce fast and accurate results within 3 seconds, helping auctioneers make quick decisions. We have also considered inflation factor and mapping of the same to the budget during the training of the model.**

Keywords*— *Decision tree, Machine Learning, SVR, Linear Regression, Random Forest Regressor, K-Nearest Neighbor, Stochastic Logistic Regression.

I. INTRODUCTION

The Indian Premier League is an annual cricket tournament administered by the Board for Control of Cricket in India. It is a 20 over format of playing cricket which was introduced in 2003 as a need for shorter version of the game in order to counter the falling attendance of spectators [1]. Currently it is the 6th most attended league in the world and the worlds most attended league as of 2016[2].

The participating franchises form their teams by conducting an auction before the tournament from a pool of national and international cricket players of different countries. The auction starts with the 'marquee' list, where 16 players in two batches of eight each, come up for bidding. The marquee lot is followed by capped and uncapped players: batsmen, all-rounders, wicket-keepers, fast bowlers and spinners. After the players have been presented for bidding, the accelerated process will enable franchises to nominate a set number of players from the remaining ones. The list is submitted on the first day and the nominated players go up for bidding the next day. Once these players have been presented, the franchises will be asked to submit their wish-list from the full list of players. Those unsold will come up for auction after all the players have been called once. The list of unsold players, however, is drawn up subject to the franchises' request.

As per the new IPL player policy, each franchise can spend a maximum of INR 80 crore (US $12.5m) on their squad salaries. The new rules also dictate that franchises spend 75% of the salary cap, which amounts to INR 60 crore (US $9.4m approx.). The player retention rules stipulate that a team can retain a maximum of five players through a combination of pre-auction retention and Right-To-Match (RTM). A franchise can retain a maximum of three capped India players and two capped overseas players. Having each retained three players, Super Kings, Daredevils, Mumbai Indians and Royal Challengers have only two RTMs remaining. The other four franchises have three RTMs, each having retained no more than two players.

To predict the price of a batsman there are several important parameters to consider. Runs is an important parameter which tells us the best performance a player will give. This in-turn decides the player's popularity among fans. Batting average is the total number of runs a batsman has scored divided by the number of times they have been out [3]. This tells us how consistent a batsman is, which is important because the team must score within 10 wickets of which not all are specialized in batting. Batting strike rate is a measure of how frequently a batsman achieves the primary goal of batting, namely scoring runs [4]. This is important because the speed at which the runs are scored are important in high scoring games. Balls faced is the number of balls that the batsman has played in the entire season. The batsman can bat in the first Inning or the second Inning which is decided by the captain of the team which wins the toss. Price is the output of our model.

To predict the price of a bowler there are several important parameters to consider. Runs is an important parameter that tells us how much many runs a bowler has given in the past. The lesser runs he/she gives, the more the bowler is difficult to score against. Bowling average is the ratio of runs conceded per wickets taken, meaning that the

978-1-7281-6828-9/20/$31.00 ©2020 IEEE

lower the bowling average is, the better the bowler is performing [5]. This tells us how much the bowler will allow the opponent to make runs. Lesser number of runs is preferred. Bowling strike rate is the balls bowled by wickets taken. This tells how quickly a bowler can take wickets [4]. Economy is runs per over. This tells if the bowler can bowl with lesser runs given. Balls bowled is the number of balls that the bowler has played in the entire season. The bowler can ball in the first Inning or the second Inning which is decided by the captain of the team which wins the toss. Price is the output of our model.

A. *Proposed solution for the problems facing with the existing method*

At present, the decision of choosing to bid for a player is done by 8 people from each team. Each member has different expertise and contributes to the final decision making. This makes it difficult to make decisions. Furthermore, pressure situations and hurried decisions which need to be made at the auction can lead to costly mistakes. The proposed model makes the decision making more accurate, fast and less prone to errors. We are able to predict the price of a player based on his past performance.

Another problem that the present method suffers from is that opponents are often aware of the team's favorite player due to regional popularity. This knowledge has often been abused by the opponents to inflate the price of the desired players. Our model helps the bidder realize that the exact range of the price the player is expected to be sold at, thereby making them aware of the same to adjust their budget plans.

Thirdly, at present the accuracy with which auctioneers are able to valuate a player is dependent on the experience in valuating that the 8 members of the bidding team have. This is also being taken care by our model as it involves use of the past performance to predict the cost.

II. Literature Survey

There have been several studies on players' compensation in various sports. A study had predicted the bowling performance in cricket. However, batsmen's performance and price prediction was not considered [11]. Another study performed a comparative analysis of the proposed team [12]. Hedonic Price Analysis is based on the hypothesis that a good/service can be treated as a collection of attributes that differentiates it from other goods/services [7]. Waugh (1928) propounded this concept based on his observation of different prices for different lots of vegetables. Waugh sought to identify the quality traits influencing daily market prices. Later, Rosen (1974) made his model of product differentiation based on the hypothesis that goods are valued for their utility generating attributes. According to him, while making a purchase decision, consumers evaluate product quality attributes, and pay the sum of implicit prices for each quality attribute, which is reflected in observed market price. Hence, price of a product is nothing but summation of the shadow prices of all quality attributes.

Shapiro (1983) presented a theoretical framework to examine halo effect on prices. Developing an equilibrium price-quality schedule for high-quality products, assuming competitive markets and imperfect information, he showed that reputation facilitates a price premium; hence, reputation building can be considered as an investment good. Weemaes and Riethmuller (2001) studied the role of quality attributes on preferences for fruit juices. The study involved market valuation of various attributes of fruit juice. The study did not consider consumers' preferences per se but generated quality attributes from the product label. The study revealed that consumers paid a premium for nutrition, convenience, and information. In a similar study on tea, Deodhar and Intodia (2004) showed that color and aroma were the two important attributes of a prepared tea [6].

However, for Hedonic Price Analysis, the amount of data that needs to be collected and worked with is very large [7]. We have split the batsman and bowler data into national and international players which is lesser number of records for this model to learn and predict. The availability and accessibility of data directly affects the amount of time and the expense that will be undertaken to carry out an application of the model.

Moreover, this method estimates people's willingness to pay for the supposed variation in environmental qualities and their consequences. However, if the people are unaware of the relation between the environmental qualities and their benefits to them [7].

Additionally, this model assumes that, given their income, people have the opportunity to choose the combination of attributes they prefer. It fails to see is that the real estate market can also be affected by external factors such as inflation etc. Furthermore, it has been assumed that, the prices in the market will automatically adjust to any changes in the attributes. In reality, there is a lag especially in localities where purchase and sale of real estate is limited. The model is relatively complex to interpret and requires a high level of statistical knowledge and expertise [7].

In our work we have made our own dataset that considers the inflation rates and the important attributes that describe a player of a certain category. We have made a careful and a detailed study of every attribute and its dependence on the price. A cricket player sells his cricketing services for the IPL tournament. The franchisee team owners bid for the player services, for team owners would like to maximize their utility (chances of winning and maximizing profit), and, player performance is an important argument of their utility function. In equilibrium, the final bid price of a player must be a function of the valuation of winning attributes of a player. Therefore, given the data on values of various attributes of cricket players and their final bid prices, one can estimate the price equation.

$$P_i = g(z_{i1} \dots, z_{ij} \dots, z_{in})$$

Equation 1 Price Equation

where P_i is the final bid price paid to a cricketer i for the IPL tournament and z_{ij} is the value of the attribute j of the cricket player i. The linear regression and SVR model maximize the prices given the right set of attributes and sufficient amount of data.

A. *Technicalities of the solution*

Machine learning is a branch of artificial intelligence which trains the machine to predict future outcomes based on past data. Our problem is with regression, because a player 's price is a numerical value.

978-1-7281-6828-9/20/$31.00 ©2020 IEEE

Regression is a branch of machine learning in which a y(a player's price) is predicted by calculating an equation, given n attributes.

Support Vector Machine can also be used as a regression m ethod, maintaining all the main characteristics (maximum margi n) that characterise the algorithm. The Help Vector Regression (SVR) applies the same classification rules as the SVM, with ju st a few slight differences[8]. First of all, because output is a rea l number, predicting the information at hand becomes very diffi cult.

In the case of regression, a tolerance margin (epsilon) is set in approximation to the SVM that would have already been deman ded from the problem. But besides this fact, there is also a more complicated reason to consider, therefore the algorithm is more complicated.The main idea, however, is always the same: to mi nimise error, to individualise the hyperplane that maximises the margin, keeping in mind that some of the error is tolerated[6].

The main idea, however, is always the same: to minimise er ror, to individualise the hyperplane that maximises the margin, keeping in mind that some of the error is tolerated[6].

$$y = \sum_{i=1}^{N}(a_i - a_i^*).(x_i, x) + b$$

Equation 2. Linear SVR [6]

$$f(x) = \sum m_i.x_i + c, where\ i = 0\ to\ n - 1\ records$$

Equation 3. Linear Regression

where n is the number of players, m_i is the slope of the line (coefficient of x), x_i = value of the i^{th} attribute and c is the y intercept.

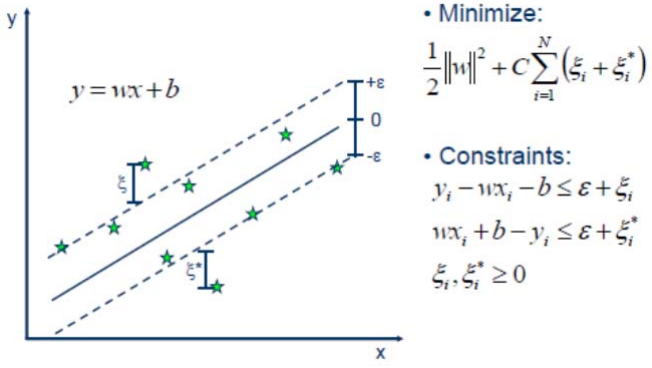

Figure 1. SVR Problem [6]

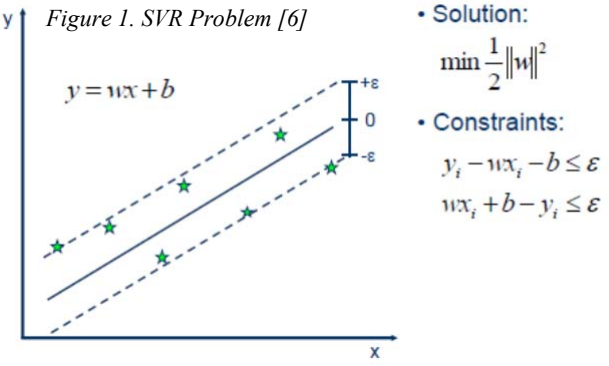

Figure 2. SVR Solution [6]

B. Innovative Content

The data-set for our problem was formed explicitly after merging the player's performance with the price of the player for both bowlers and batsmen to predict the future price of the player. The predictive analysis is performed by training our model with the most important features of the players along with the player price.

1) Problem Formulation

Data Collection

The performance parameters available for batsmen were – Matches, Innings, runs, Strike rate, average, 4s, 6s, 50s, 100s and price. The performance parameters available for bowlers were – Matches, Innings, runs, Strike rate, average, Wickets, Overs, Economy and price. The data of performance of the players [9] and the price of the players [10] were merge in order to prepare the data set. Some of players' names were not written correctly. Their price for that specific year had to be manually.

Player	Mat	Inns	Ov	Runs	Wkts	Avg	Eco	SR	4w	Price
Andrew Tye	14	14	56	448	24	19	8	14	3	1255K
Rashid Khan	17	17	68	458	21	22	7	19	0	1406K
Bhuvneshwary	12	12	46	354	9	39	8	31	0	1328K
S. Kaul	17	17	66	547	21	26	8	19	0	350K
Tim Southee	11	11	43	329	9	36	8	29	0	1156K
Ryan Harris	5	5	18	130	8	16	7	14	1	325K
Umesh Yadav	14	14	48	410	17	24	8	17	1	656K

Figure 4. Bowlers data after merging

Furthermore, in a dynamic economic environment, the value of money is ever fluctuating. This fluctuation in the value of money reduces purchasing power i.e. the ability to buy any commodity in the same quantity and quality at one or more different time periods. This fluctuation in the value of money is controlled by economic forces of inflation and deflation. In order to cut the effect of inflation on the purchasing power of money, economists use the Cost of Living Index. Through indexation process the economists ascertain the real value of money as on given date in relation to any prior or future time period. Since the earlier IPL auction of the players was done in US Dollar, it had to be converted to Indian National Rupee, for unifying the unit of price. The conversion was done taking a normalized value of $1 = ₹62 in 2007 using the indexation process.

2) Problems with data

National and International players

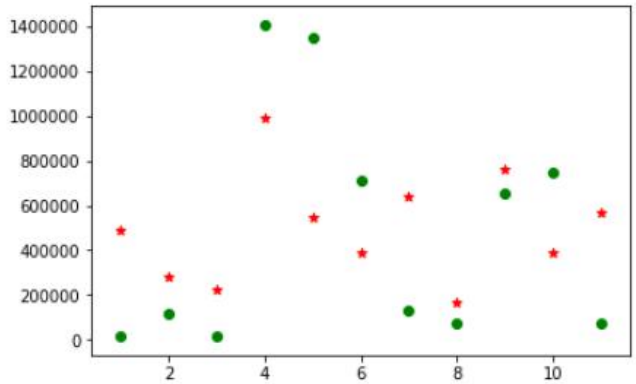

Figure 5. Bowler prices: actual and predicted (linear regression) Here x axis corresponds to bowler (in no particular order) and y axis corresponds to the bowler's price prediction (in INR).

Green represents the predicted price and red represents the actual price.

978-1-7281-6828-9/20/$31.00 ©2020 IEEE

Note: In Figure 5, player no. 5, which belongs to Lasith Malinga, has a large gap between target and predicted output. This is because he is an international cricket player. Green represents the predicted price and red represents the actual price. In Figure 6 the player no. 5, which belongs to Travis Head, has a large gap between target and predicted output. This is because he is also an international cricket player.

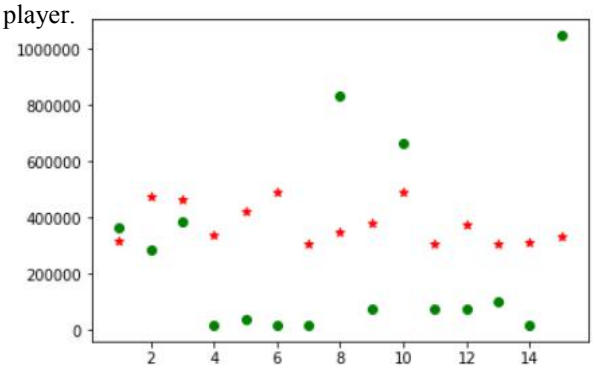

Figure 6. Batsman prices: actual and predicted (linear regression) Here x axis corresponds to batsman (in no particular order) and y axis corresponds to batsman's price prediction (in INR).

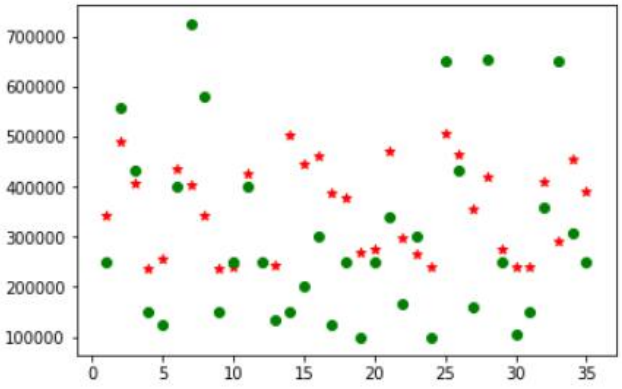

Figure 7. Bowler price prediction after separation of international and national bowlers(linear regression) Here x axis corresponds to bowler (in no particular order) and y axis corresponds to bowler's price prediction (in INR).

Note: In Figure 7 and Figure 8, the difference between predicted and actual price has been reduced.

In order to deal with these large gaps, we need to separate national (Indian) and international players. By doing so, the gap between predicted and actual values reduces. This is because the auctioneers are willing to spend more money on international players. Hence national and international players need to be trained in separate models.

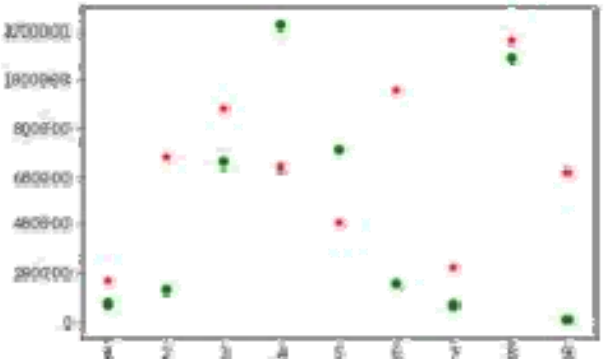

Figure 8. Batsman price prediction after separation of international and national batsmen (linear regression) Here x axis corresponds to bowler (in no particular order) and y axis corresponds to bowler's price prediction (in INR).

3) Outlier data

The second problem that price prediction faces is that the data of most of the well performing players. If we consider batting average, which is a ratio of runs to matches played, most players have them between 20 and 40 as shown below. This leads to difficulty in training the model for players who are not well performing in ratio-based parameters like strike rate. As a consequence of this the model will memorize the data, instead of learning. This can be observed from the fact that regressor over-fits due to large number of players in a range. However, when outlier data as seen in Fig. 5.5 is given to the model, the predictions are having huge error in price prediction.

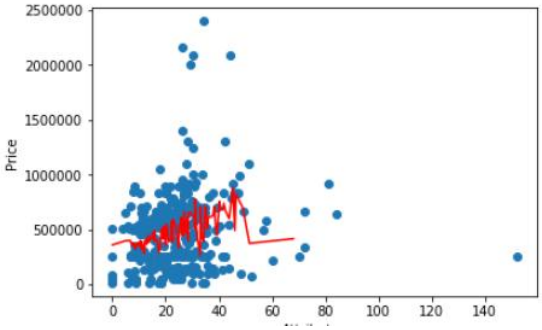

Figure 9: Average vs Price plot

The over-fitting of data when training the model can be seen in Figure 9 from the regression line.

IV. Proposed Methodology

A. *Performance parameters to consider*

Case 1: If we observe batting performance parameters- runs, average, batting strike rate and balls faced- we see that they are dependent on runs, wickets and balls faced. Therefore, we considered wickets, runs, balls faced, innings and price for batsman price prediction.

Case 2: If we observe bowling performance parameters- runs, average, bowling strike rate and balls bowled- we see that they are dependent on runs, wickets and balls bowled. Therefore, we considered wickets, runs, balls bowled, innings and price for bowler price prediction.

Another advantage considering only these parameters is that the machine will be able to distinguish between the performances of two players. This is because average and strike rate for both bowler and batsman are in the form of ratios. Therefore, the model tends to gives less weight age to these parameters for price prediction. Furthermore, each parameter versus the price of player gives us positive slope. This behavior of slope being same is the advantage to the model.

B. *Algorithms used*

Our problem is a regression problem because the price of players increases linearly. We applied Decision Tree based regression, KNN, Stochastic Logistic and Linear regression, Random Forest Regressor, and SVM to predict player price.

978-1-7281-6828-9/20/$31.00 ©2020 IEEE

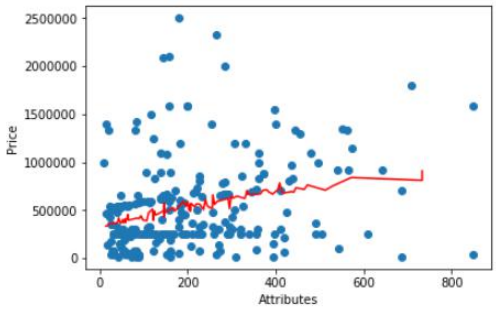

Figure 10. SVR on national batsmen

Support Vector Regressor (SVR) on national batsmen gives best results because the data is linearly separable. This linear separability is due to runs scored being dependent on balls faced and number of innings which in turn is directly proportional to the price of the player. The kernel used is linear.

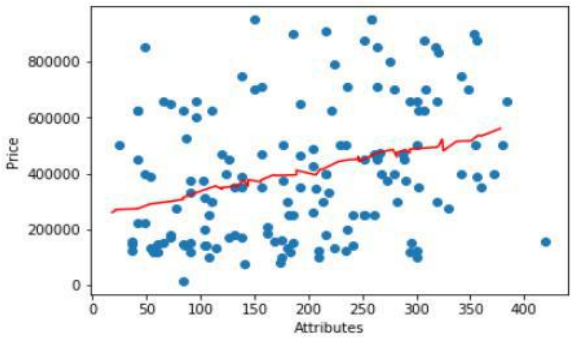

Figure 11. SVR on international batsmen

Support Vector Regressor (SVR) is suitable for international batsmen as it is neither over-fitting not under-fitting too much. We have considered runs scored, innings and balls faced(batting). These gave us stable (not fluctuating) positive slopes.

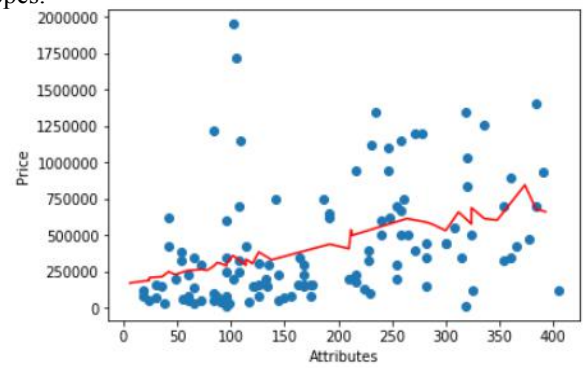

Figure 12. Linear Regression on national bowler

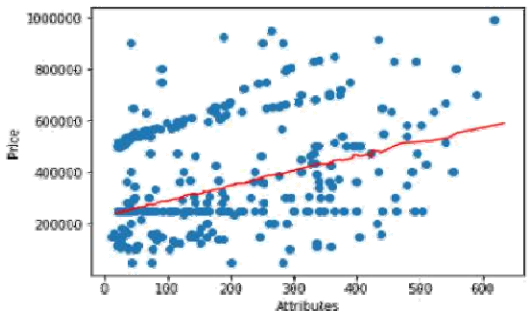

Figure 13. Linear Regression on international bowler

Linear Regression works for national bowler and international bowler again as it is neither over-fitting nor under-fitting too much. We have considered runs given, innings, best fast(bowling) and wickets taken. These gave us stable (not fluctuating) positive slopes.

V. RESULTS

Support Vector Regressor (SVR) cannot be applied to predict the price of bowlers because the range of wickets taken is usually not linear with respect to price. Therefore, we get a vector which is parallel to X axis, which implies that the prices of all the players are same. Thus, with SVR learning does not occur with bowlers' data.

Table 1. Mean Absolute Error in lakh (INR)

Models used	Batsman		Bowlers	
	National	International	National	International
Decision Tree Regressor	54	70	73	75
K-Nearest Neighbor (KNN)	58	55	73	62
Stochastic Logistic Regression	70	180	78	80
Linear Regression	45	52	48	55
Random Forest Regressor	53	65	54	63
Support Vector Regressor (SVR)	42	53	N/A	N/A

A. *Comparison of Results*

Decision Tree Regressor gives high Mean Squared Error (MSE) for Indian and international players because the model has over-fitted the data. The model is memorizing the data instead of learning which is confirmed by poor test results despite good training.

Linear Regression gives us the second lowest MSE. It has further showed that IPL auctioneers are willing to spend more on obtaining foreign players. This is because the team must contain 7 national players. This constraint means that the remaining players must be high quality players. This supply-demand gap leads to inflation in their price. Since the slope of all the curves, formed by plotting the price against each attribute, are positive. Linear Regression is also suitable for this problem.

Stochastic Logistic Regression gives us good MSE, though slightly poorer than Linear Regression. This is because the data is not logistic with respect to price of the players.

Random Forest Regressor is polling for the price from multiple independent trees within the forest of size 100. This leads to a lesser risk of result being influenced by noise. Furthermore, this model also confirms that foreign players are going for higher price for the same performance as Indian players. But the problems of tree exist even here.

K-Nearest Neighbor (KNN) is giving poor MSE similar to Decision Tree Regression. This is due to over-fitting. The model is memorizing the data instead of learning which is confirmed by poor test results despite good training.

Support Vector Regressor (SVR) fails to work for our bowler's data. This is because it finds vectors parallel to

978-1-7281-6828-9/20/$31.00 ©2020 IEEE

each attribute and makes learning task redundant. Furthermore, it indicates the same price for all kernels. For the batsman, since the data is linearly separable, it is able to get the vectors in proper directions and thus gives the best results for national players. Hence, as seen in Table 1, SVR gives the lowest Mean Absolute Error (MAE) for batsmen and Linear regression gives the lowest MAE for bowlers.

In order to cross validate our models, we had applied K Fold method. We had kept 20% of data in the validation set. For obtaining an optimal cost and variance, we have used K=5 and obtained the results in Table 2.

Table 2. Mean absolute errors using K Fold method in lakhs (INR)

No. of folds	Batsman		Bowlers	
	National	International	National	International
1	48	58	51	60
2	60	59	42	55
3	62	75	58	51
4	72	60	48	62
5	42	52	45	75

We have also trained our model until 2017 data and tested the model using 2018 data for which the predicted price and mean absolute error are shown in Table 3 and Table 4.

Table 3. Actual price, predicted price using 5-Folds method and the mean absolute error in lakh rupees (INR) for 2018 IPL auction for bowlers

Bowler Name	Actual	Model	Error
Jaydev Unadkat	11.5	9.2	-2.3
Imran Tahir	1	0.8	-0.2
Sandeep Sharma	3	1.1	-1.9
Umesh Yadav	4.2	4.1	-0.1
Chris Woakes	7.4	5	-2.4
Siddharth Kaul	3.8	4.1	0.3

Table 4. Actual price, predicted price using 5-Folds method and the mean absolute error in lakh rupees (INR) for 2018 IPL auction for batsmen

Hence, as seen in Table 3, SVR gives a low Mean Absolute Error (MAE) for all the batsmen. Similarly, as seen in Table 4, Linear regression gives low MAE for all the bowlers. Thus, SVR is the best model for batsmen and Linear regression is the best model for bowlers.

Player	Actual	Model	Error
Gautam Gambhir	2.8	2.9	-0.1
Shikhar Dhawan	5.2	4.2	1
Chris Gayle	2	2.1	-0.1
Glenn Maxwell	9	5.5	3.5
Lokesh Rahul	11	5.4	5.6
Robin Uthappa	6.4	6	0.4

B. Justification of Results

The players marked in the red are the popular players. The model has been able to capture all the parameter results and predict the prices as shown in Table 3 and Table 4 accurately for most players. The players on whom the fluctuation is a lot is because the popularity could not be captured. The player prices are also dependent on their popularity. The intercepts indicate the average base price of those players sold at their base prices. The other parameters- runs, innings, balls faced for batsman; runs given, balls given, wickets, innings for bowlers- contribute to the price linearly which are signs that a good player with higher values of the attributes considered will be sold at a higher price.

VI. Conclusion

The SVR and linear regression model has captured all the important features to predict the price of both batsman and bowlers respectively. The price of high performing player's increases as price of those whose performance is falling reduces. The average error is ±₹4,50,000 per bowler and ±₹5,20,000 per batsman. In worst-case scenario, the auctioneer will bid for 22 bowlers, which amounts to ₹6,00,00,000 error in the budget, which is only 8% of the ₹80,00,00,000 budget. This is acceptable as not always does an auctioneer spend all his/her money and they will never buy only bowlers. Additionally, our model might slightly over price some players and slightly under price some players. Therefore, the budget will not be affected much.

VII. Future Work

Age effects on performance and price isn't considered. The popularity of players also affects their price. For some players it significantly affects their price. This popularity varies across regions in India. Popularity based on recent records created such as recently made centuries, hat-tricks affect the player price in Auction and are to be captured.

References

[1] A. Gupta, "India and the IPL: Cricket's Globalized Empire. The Round Table," 2009.

[2] C. Barrett, "Big Bash League jumps into top 10 of most attended sports leagues in the world," The Sydney Morning Herald, 10 January 2016. [Online]. Available: https://www.smh.com.au/sport/cricket/big-bash-league-jumps-into-top-10-of-most-attended-sports-leagues-in-the-world-20160110-gm2w8z.html. [Accessed 7 January 2018].

[3] A. C. Kimber, " A Statistical Analysis of Batting in Cricket," 1993.

[4] H. H. Lemmer, "The Single Match Approach to Strike Rate Adjustments in Batting Performance Measures in Cricket ," 2011.

[5] V. Staden, "Comparison of cricketers' bowling and batting performances using graphical displays," 2009.

[6] S. K. R. a. S. Y. Deodhar, "Player Pricing and Valuation of Cricketing Attributes: Exploring the IPL Twenty20 Vision," *VIKALPA,* vol. 34, no. 2, pp. 15-23, 2009.

[7] M. A. S. Hagan, "Factors Driving Farm Gate Price of Tomatoes in Ghana: An Application of Hedonic Model," 2020.

[8] D. S. Sayad, "Support Vector Regression," 23 August 2018. [Online]. Available: http://www.saedsayad.com/support_vector_machine_reg.htm.

[9] "Indian Premier League Official Website," IPL20.COM, 27 May 2018. [Online]. Available: https://www.iplt20.com/stats/2018/most-wickets.

[10] "IPL 2017 player salary," CricMetric, [Online]. Available: http://www.cricmetric.com/ipl/salary.py?year=2017. [Accessed 15 July 2018].

[11] A. A. A. Rupai "Predicting Bowling Performance in Cricket from Publicly Available Data," 2020.

[12] A. Adhikari "An innovative super-efficiency data envelopment analysis, semi-variance, and Shannon-entropy-based methodology for player selection: evidence from cricket," 2020.

978-1-7281-6828-9/20/$31.00 ©2020 IEEE