Abstract

In this project, we will apply some machine learning-based algorithm that predicts the cost at which a player can be sold in the Indian premier league auction.

Here we will try to find out estimate the player selling price using their past performance parameters like runs, wicket, balls, innings, etc.

We will use some machine learning models to carry out a few tests like decision Tree Regressor, K-nearest-Neighbors(KNN), Linear Regression, Stochastic Logistic Regression, Random Forest Regressor, and Support Vector Regression (SVR).

Here we will try to find out which algorithm gives us the fastest and accurate results on predicting the cost of the player.

It helps the auctioneers to make quick decisions.

Table of Contents

1. INTRODUCTION ---------- 01-09

1.1. Technology (Language and Tools) ---------- 03

1.1.1. Python ---------- 03

1.1.2. Jupyter Notebook ---------- 03

1.1.3. Machine Learning Techniques ---------- 03

1.1.3.1. Linear Regression ---------- 04-05

1.1.3.2. Decision Trees ---------- 05-06

1.1.3.3. Support Vector Machines (SVM) ---------- 06-07

1.1.3.4. Random Forest ---------- 07-08

1.1.3.5. Logistic Regression ---------- 08

1.1.3.6. K-Nearest Neighbors ---------- 08-09

2. LITERATURE REVIEW ---------- 10-12

2.1. Data Collection ---------- 11-12

3. METHODOLOGY & APPROACH ---------- 13-18

3.1. Methodology ---------- 13

3.1.1. Performance parameter to consider ---------- 13

3.1.2. Algorithms used ---------- 15

- Linear Regression ---------- 15
- Decision Tree Regressor ---------- 15-16
- The random Forest Regressor ---------- 16
- K-Nearest Neighbors Regressor ---------- 16-17
- Support Vector Machine ---------- 17
- Logistic Regressor ---------- 17-18

3.2. Work Approach ---------- 19-20

4. RESULT & DISCUSSION ---------- 21-24

5. CONCLUSION ---------- 25

6. REFERENCE ---------- 26-27

Table of fig.

1. Linear Regression
   1.1 Linear Regression Correlation between dependent and independent variable-----05

2. Machine Learning Techniques
   2.1 Decision Tree----------05
   2.2 Decision Boundary in support vector Machines ---------- 06
   2.3 Random Forest----------07
   2.4 Sigmoid Function----------08

3. Work Approach
   3.1 Modules of the System----------20

Table of Screenshots

1. K-Nearest neighbors
1.1 Elbow curve---------- 09

2. Data Collection
2.1 IPL Dataset founded on IPL official site, Manipulated and Filtered ---------- 12

3. Methodology
3.1 Features selection through variance inflation factor & heatmap ---------- 13-14

4. Linear Regression
4.1 Linear Regression Model by Sklearn Library ---------- 15

5. Decision Tree
5.1. Decision tree by Sklearn library---------- 16

6. The Random Forest Regressor
6.1 Random Forest Model by Sklearn library---------- 16

7. K-Nearest Neighbour Regressor
7.1 K-Nearest Neighbor by Sklearn library ---------- 17

8. Support Vector Machine
8.1 Support vector machine model by Sklearn library---------- 17

9. Logistic Regression
9.1. Logistic Regression model by Sklearn library---------- 18

10. Result & Discussion
10.1 Predicted Values of Linear Regression Model of Indian Bowler ---------- 21
10.2 Actual & Predicted Value of the player salaries ---------- 22
10.3 Indian bowler knn predicted vs real ---------- 22
10.4 Overseas bowler SVM predicted vs real ---------- 23
10.5 Overseas Batsman knn elbow curve ---------- 23
10.6 Overseas Batsman knn predicted vs real ---------- 24

1. Introduction

The Board for Control of Cricket in India organizes an annual cricket tournament known as the Indian Premier League. It is a 20-over cricket format that was established in 2003 to address the demand for a shorter version of the game to combat declining spectator attendance.

As of 2016, it is the world's 6th most popular league and the most popular league in the globe.

Before the event, the participating franchisees hold an auction to select their teams from a pool of national and international cricket players from various countries.

The auction begins with the 'marquee' list, which features 16 players divided into two groups of eight. Capped and uncapped players, including batsmen, all-rounders, wicketkeepers, fast bowlers, and spinners, follow the marquee lot. The faster process will allow franchisees to nominate a set number of players from the remaining players once the players have been presented forbidding. On the first day, the list is submitted, and the nominated players are put up for auction the next day. Following the presentation of these players, the franchises will be requested to submit a wish list of players from the entire list. After all of the players have been named once, the unsold tickets will be auctioned off. 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 on their squad salaries. The new rules also dictate that franchises spend 75% of the salary cap, which amounts to INR 60 crore.

A team can retain a maximum of five players through a mix of pre-auction retention and Right-To-Match, according to the player retention rules (RTM). A franchise can keep a maximum of three India-capped players and two overseas-capped players. Super Kings, Daredevils, Mumbai Indians, and Royal Challengers have only two RTMs left after each retaining three players. The other four franchises each have three RTMs, with no more than two players kept in each.

There are various significant criteria to consider while predicting the price of a batsman. Runs are an important metric that tells us how well a player will perform. This, in turn, determines the player's fan base. The batting average is calculated by dividing the total number of runs scored by the number of times a batsman has been out. This indicates a batsman's consistency, which is vital because the team must score inside 10 wickets, not all of whom are batting specialists. Batting strike rate is a metric that indicates how often a batter achieves his primary goal of batting. This is significant because in high-scoring games, the rate at which runs are scored is critical. The number of balls faced by the batsman during the course of the season is calculated. The batsman can bat in either the first or second inning, depending on the captain of the team that wins the toss.

There are various essential criteria to consider while predicting the price of a bowler. The number of runs a bowler has given up in the past is an important parameter to consider. The bowler becomes more difficult to score against the less runs he gives up. Bowling average is the ratio of runs conceded per wickets taken, with the lower the bowler's average, the better. This indicates how many runs the bowler will allow the opponent to score. It is preferable to have fewer runs. The balls bowled by wickets taken make up the bowling strike rate. This metric indicates how fast a bowler can take wickets. The economy is running at a breakneck pace. This indicates whether the bowler is capable of bowling with less runs. The number of balls bowled in a season is the total number of balls bowled by a bowler. The bowler can bat in either the first or second inning, depending on which team's captain wins the toss. The outcome of our model will be price.

Proposed solution for the problems facing with the existing method

At the moment, each team's decision to bid on a player is made by eight persons. Each participant brings a unique set of skills to the table and contributes to the ultimate choice. This makes making decisions tough. Furthermore, high-pressure conditions and rushed decisions made at the auction can result in costly errors. The proposed paradigm improves decision-making accuracy, speed, and mistake resistance. We can forecast a player's price based on his previous performance.

Another issue with the current technique is that, due to regional popularity, opponents are typically aware of the team's favourite player. Opponents have frequently utilised this information to increase the price of wanted players. Our methodology informs the bidder of the exact price range at which the player is projected to be sold, allowing them to adapt their budget plans accordingly.

1.1 Technology (Language and Tools)

1.1.1 Python

Python is a dynamically semantic, interpreted, object-oriented high-level programming language. Its high-level built-in data structures, together with dynamic typing and dynamic binding, making it ideal for Rapid Application Development and as a scripting or glue language for connecting existing components. Python's basic, easy-to-learn syntax prioritises readability, lowering software maintenance costs. Python facilitates programme flexibility and code reuse by supporting modules and packages. The Python interpreter and its substantial standard library are free to download and distribute in source or binary form for all major platforms.

1.1.2 Jupyter Notebook

JupyterLab is a web-based interactive development environment for Jupyter notebooks, code, and data. JupyterLab is flexible: configure and arrange the user interface to support a wide range of workflows in data science, scientific computing, and machine learning.

1.1.3 Machine Learning Techniques

Machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of Computer Programs that can change when exposed to new data.

Machine Learning is making the computer learn from studying data and statistics.

Machine Learning is a step in the direction of artificial intelligence (AI).

Machine Learning is a program that analyses data and learns to predict the outcome.

It is the concept of programming the machine in such a way that it learns from its experiences and different examples, without being programmed explicitly. It is an application of AI that allows machines to learn on their own. Machine learning algorithms are a combination of math and logic that adjust themselves to perform more progressively once the input data varies. Being a general-purpose, easy-to-learn and understand language, Python can be used for a large variety of development tasks. It is capable of doing several machine learning tasks, which is why most algorithms are written in Python.

The process of creating machine learning algorithms is divided into 2 parts – The training and Testing Phase. Even though there are a large variety of machine learning algorithms, they are grouped into these categories: **Supervised Learning, Unsupervised Learning, and Reinforcement Learning.**

Some of the machine learning algorithms are the following:

1.1.3.1 Linear Regression

It is one of the most popular Supervised Machine Learning algorithms in Python that maintains an observation of continuous features and based on it, predicts an outcome. It establishes a relationship between dependent and independent variables by fitting a best line.

This **best fit line is represented by a linear equation Y=a*X+b,** commonly called the regression line.

Where:

Y – Dependent variable

a - Slope

X – Independent variable

b- Intercept

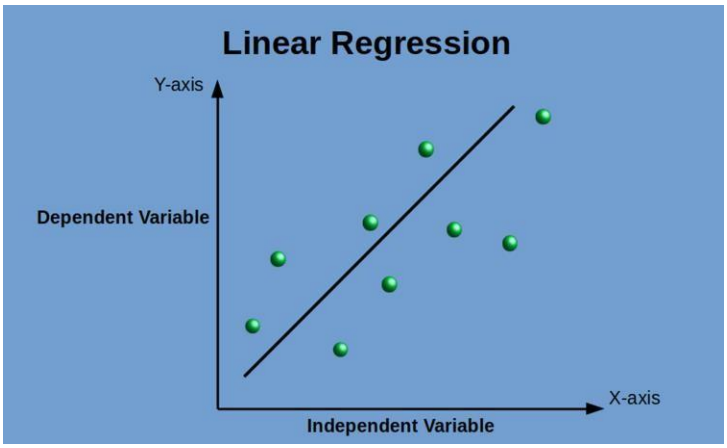


Fig1.1: Linear Regression Correlation between dependent and independent variable

The regression line is the line that **fits best in the equation to supply a relationship between the dependent and independent variables**. When it runs on a single variable or feature, we call it **simple linear regression** and when it runs on different variables, we call it **multiple linear regression.** This is often used to estimate the cost of houses, total sales, or the total number of calls based on continuous variables.

1.1.3.2 Decision Trees

A decision tree is built by repeatedly asking questions to the partition data. The decision tree algorithm aims to increase the predictiveness at each level of partitioning so that the model is always updated with information about the dataset.

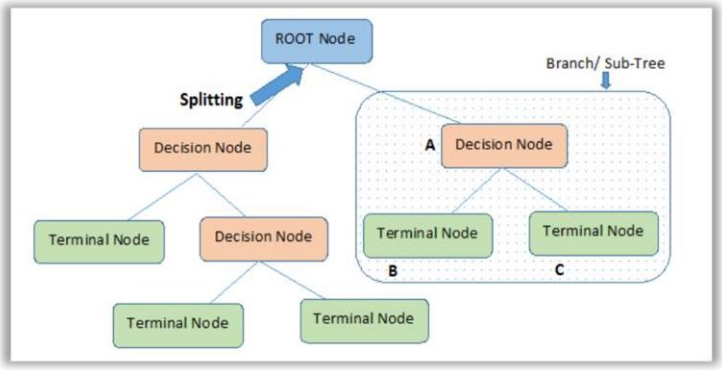


Fig 2.1: Decision Tree

Even though it is a **Supervised Machine Learning algorithm**, it is used mainly for **classification rather than regression**. In a nutshell, the model takes a particular instance, traverses the decision tree by comparing important features with a conditional statement. As it descends to the left child branch or right child branch of the tree, depending on the result, the more important features are closer to the root. The good part about this machine learning algorithm is that **it works on both continuous dependent and categorical variables.**

1.1.3.3 Support Vector Machines (SVM)

This is one of the most important machine learning algorithms in Python which is mainly used for **classification but can also be used for regression tasks**. In this algorithm, each data item is plotted as a point in n-dimensional space, where **n denotes the number of features you have, with the value of each feature as the value of a particular coordinate.**

SVM does the **distinction of these classes by a decision boundary.**

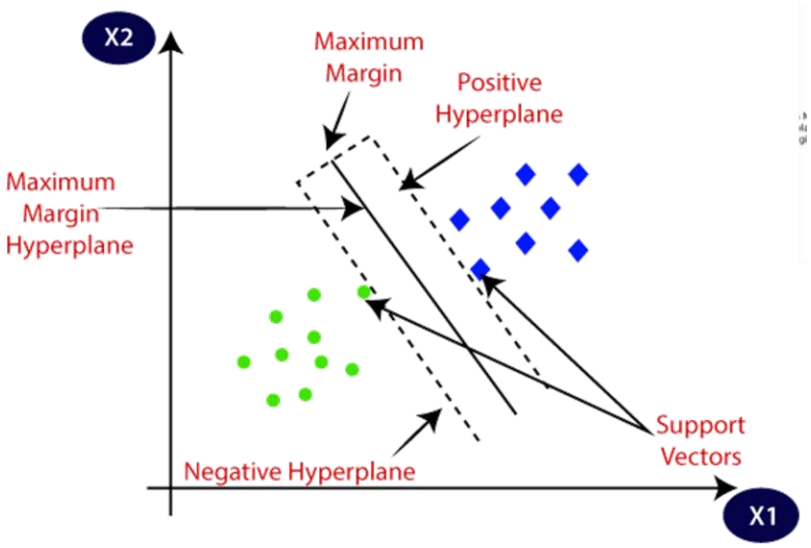


Fig2.2: Decision Boundary in Support Vector Machine

For e.g: If length and width are used to classify different cells, their observations are plotted in a 2D space and a line serves the purpose of a decision boundary. If you use 3 features, your decision boundary is a plane in a 3D space. SVM is highly effective in cases where the number of dimensions exceeds the number of samples.

1.1.3.4 Random Forest

Random Forest Regression is a supervised learning algorithm that uses **ensemble learning** method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.

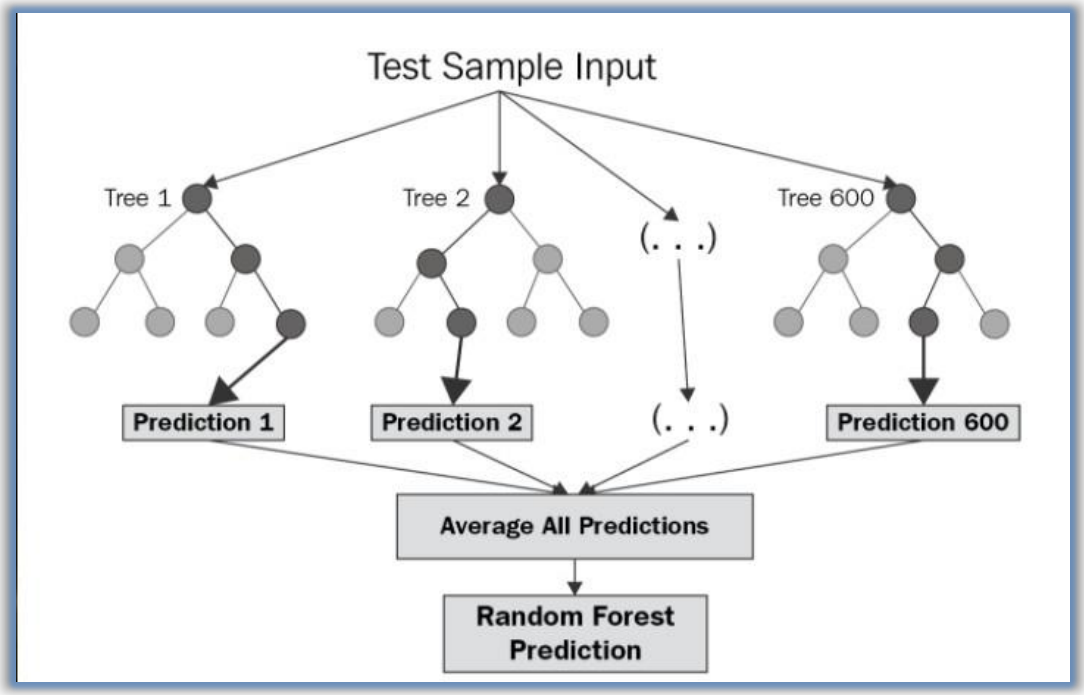


Fig2.3: Random Forest

You can notice that the trees run in parallel with no interaction amongst them. A Random Forest operates by constructing several decision trees during training time and outputting the mean of the classes as the prediction of all the trees.

A Random Forest Regression model is powerful and accurate. It usually performs great on many problems, including features with non-linear relationships.

Disadvantages, however, include the following: there is no interpretability, over fitting may easily occur, we must choose the number of trees to include in the model.

1.1.3.5 Logistic Regression

Logistic regression is a process of modeling the probability of a discrete outcome given an input variable. The most common logistic regression models a binary outcome; something that can take two values such as true/false, yes/no, and so on.

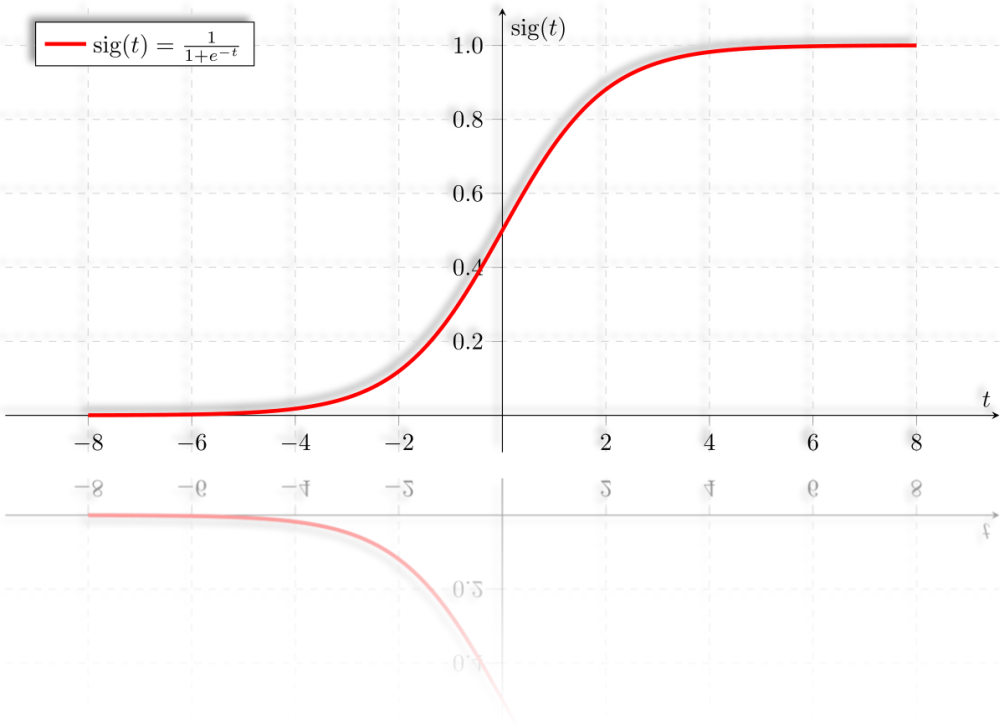


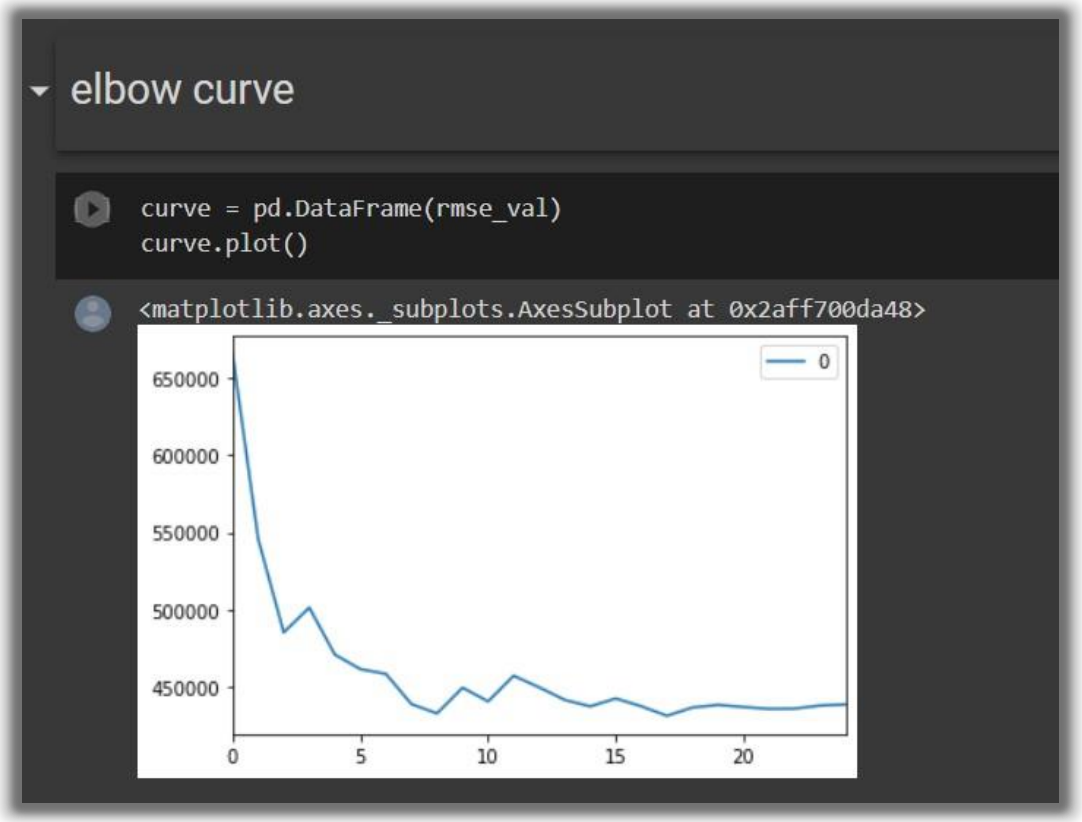
Fig 2.4: Sigmoid Function

1.1.3.6 K-nearest neighbors

K-nearest neighbors is probably the simplest widely used model conceptually. KNN models are really just technical implementations of a common intuition, that things that share similar features tend to be, well, similar.

Elbow Curve:

The most critical stage in supervised machine learning with k-Nearest Neighborhood is determining the ideal value of K, or how many clusters your data should be divided into. The best value of k decreases the effect of noise on classification while making class boundaries less apparent. The elbow approach assists data scientists in determining the best number of clusters for KNN clustering.



Sc 1.1: Elbow Curve

A simple implementation of KNN regression is to calculate the average of the numerical target of the K nearest neighbors. Another approach uses an inverse distance weighted average of the K nearest neighbors. KNN regression uses the same distance functions as KNN classification.

2. LITERATURE REVIEW

Several studies on player remuneration in various sports have been conducted. In cricket, a study predicted bowling performance. The performance of batting and price predictions, however, were not taken into account. A comparative analysis of the projected team was conducted in another study. The concept behind hedonic price analysis is that a good/service can be considered as a set of features that distinguish it from other goods/services.

Based on his observations of varied pricing for different kinds of vegetables, Waugh (1928) proposed this hypothesis. Waugh wanted to figure out what factors influence everyday market pricing. Rosen (1974) built his product differentiation model on the premise that items are valued for their utility-generating features. According to him, buyers analyse product quality qualities and pay the sum of implicit pricing for each quality trait, which is represented in observed market price, while making a purchase decision. As a result, a product's price is nothing more than the sum of all quality attributes' shadow prices.

A theoretical framework for exploring the price halo effect was given by Shapiro (1983). He established an equilibrium price-quality schedule for high-quality products, assuming competitive markets and imperfect knowledge, to show that reputation justifies a price premium. As a result, reputation management might be considered a worthwhile investment. The impact of quality parameters on fruit juice preferences was explored by Weemaes and Riethmuller (2001). The study evaluates a variety of elements of fruit juice on the market. In this study, consumer preferences were ignored, but quality attributes were determined from the product label. According to the report, consumers paid a premium for nutrition, convenience, and information. Deodhar and Intodia (2004) found that colour and scent were the two most important characteristics of a prepared tea in a comparable study.

The amount of data that needs to be collected and handled with for Hedonic Price Analysis, on the other hand, is enormous. We divided the batting and bowler data into national and international players so that the machine could learn and forecast with fewer records. The

amount of time and money required to carry out a model application is directly proportional to the availability and accessibility of data.

In our research, we created our own dataset that takes into account inflation rates as well as key features that define a player in a specific category. Every attribute and its relationship to the price has been thoroughly researched. For the IPL event, a cricket player sells his cricketing services. The franchisee team owners compete for player services because they want to optimise their utility (winning chances and profit), and player performance is a key argument in their utility function. In order for a player's ultimate offer price to be in equilibrium, the player's winning traits must be valued. Therefore, given the data on values of various attributes of cricket players and their final bid prices, one can estimate the price equation.

For both bowlers and batsmen, the data set for our challenge was deliberately constructed after merging the player's performance with the player's price to anticipate the future. the player's price Our model is trained with the most relevant aspects of the players as well as the player pricing to do the prediction analysis.

2.1 Data Collection

Batsmen's performance metrics included matches, innings, runs, strike rate, average, 4s, 6s, 50s, 100s, and price. Matches, Innings, runs, Strike rate, average, Wickets, Overs, Economy, and price were the performance parameters accessible to bowlers. In order to build the data set, the data on the players' performance and their prices were combined. Some of the players' names were misspelt. Their pricing for that year had to be carefully calculated.

1	Player	5w	Avg	Color	Econ	Inns	Mat	Ov	4w	Runs	SR	Salary	Team.1	Wkts	Year	Fsalary
2	Shane Warne	0	21.26	450,000	7.76	15	15	52	0	404	16.42	$450,000	Rajasthan Royals	19	2008	450,000
3	Shane Watson	0	22.52	125,000	7.07	15	15	54.1	0	383	19.11	$125,000	Rajasthan Royals	17	2008	125,000
4	Piyush Chawla	0	22.88	400,000	8.3	15	15	46.5	0	389	16.52	$400,000	Kings XI Punjab	17	2008	400,000
5	Irfan Pathan	0	23.33	925,000	6.6	14	14	53	0	350	21.2	$925,000	Kings XI Punjab	15	2008	925,000
6	Munaf Patel	0	30	275,000	7.63	15	15	55	0	420	23.57	$275,000	Rajasthan Royals	14	2008	275,000
7	Zaheer Khan	0	27.46	450,000	8.5	11	11	42	0	357	19.38	$450,000	Royal Challengers Bangalore	13	2008	450,000
8	Glenn McGrath	0	29.75	350,000	6.61	14	14	54	1	357	27	$350,000	Delhi Daredevils	12	2008	0
9	Umar Gul	0	15.33	150,000	8.17	6	6	22.3	1	184	11.25	$150,000	Kolkata Knight Riders	12	2008	0
10	Shaun Pollock	0	27.36	550,000	6.54	13	13	46	0	301	25.09	$550,000	Mumbai Indians	11	2008	0
11	Dale Steyn	0	25.2	325,000	6.63	10	10	38	0	252	22.8	$325,000	Royal Challengers Bangalore	10	2008	325,000
12	Dilhara Fernando	0	16	150,000	8	5	5	20	1	160	12	$150,000	Mumbai Indians	10	2008	150,000
13	Shahid Afridi	0	25	675,000	7.5	10	10	30	0	225	20	$675,000	Deccan Chargers	9	2008	0
14	Ishant Sharma	0	41.37	950,000	7.66	13	13	43.1	0	331	32.37	$950,000	Kolkata Knight Riders	8	2008	950,000
15	Ajit Agarkar	0	25.87	350,000	7.96	9	9	26	0	207	19.5	$350,000	Kolkata Knight Riders	8	2008	350,000
16	Yusuf Pathan	0	28.75	475,000	8.16	13	16	28.1	0	230	21.12	$475,000	Rajasthan Royals	8	2008	475,000
17	Mohammad Asif	0	37	650,000	9.25	8	8	32	0	296	24	$650,000	Delhi Daredevils	8	2008	0
18	Anil Kumble	0	43.42	500,000	7.93	10	10	38.2	0	304	32.85	$500,000	Royal Challengers Bangalore	7	2008	500,000
19	Shoaib Akhtar	0	10.8	425,000	7.71	3	3	7	1	54	8.4	$425,000	Kolkata Knight Riders	5	2008	0
20	Harbhajan Singh	0	16.4	850,000	8.2	3	3	10	0	82	12	$850,000	Mumbai Indians	5	2008	850,000
21	Brett Lee	0	28	900,000	7	4	4	16	0	112	24	$900,000	Kings XI Punjab	4	2008	900,000
22	Sanath Jayasuriya	0	39.75	975,000	7.57	8	14	21	0	159	31.5	$975,000	Mumbai Indians	4	2008	975,000
23	Jacques Kallis	0	77.75	900,000	9.05	11	11	34.2	0	311	51.5	$900,000	Royal Challengers Bangalore	4	2008	900,000
24	Scott Styris	0	66.33	175,000	7.37	8	8	27	0	199	54	$175,000	Deccan Chargers	3	2008	175,000
25	Ramesh Powar	0	30.33	170,000	7.58	5	5	12	0	91	24	$170,000	Kings XI Punjab	3	2008	170,000

Sc 2.1: IPL Dataset founded on IPL official site, Manipulated and Filtered

We must distinguish between domestic (Indian) and international players. As a result, the difference between projected and actual values narrows. The auctioneers are willing to spend greater money on international players because of this. As a result, various training approaches are required for national and international players.

3. Methodology & Approach

3.1 Methodology :-

3.1.1 Performance parameter to consider

Case 1: Batting performance factors including as runs, average, batting strike rate, and balls faced are all influenced by runs, wickets, and balls faced. For batsman price prediction, we will examine wickets, runs, balls faced, innings, and price.

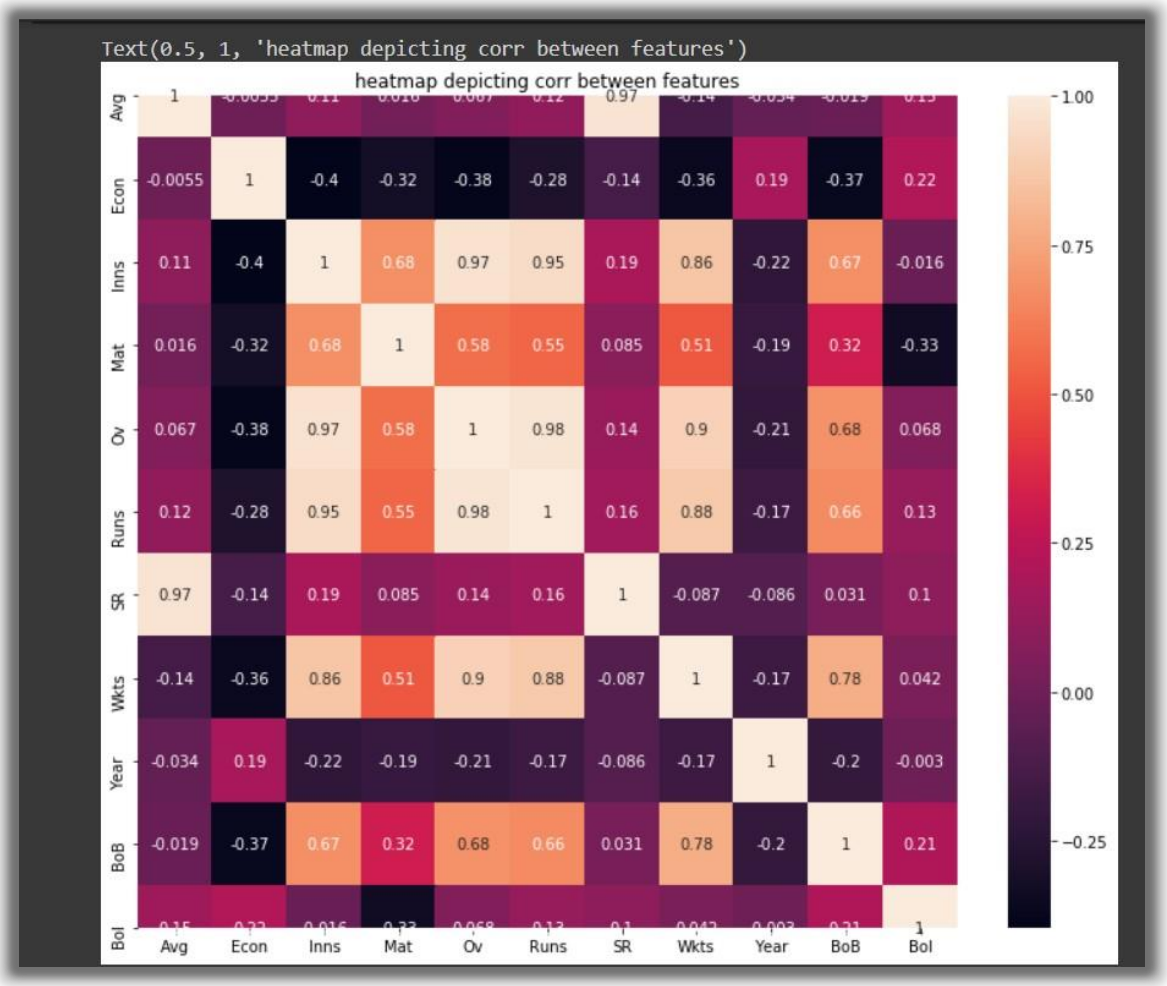
Case 2: Runs, average, bowling strike rate, and balls bowled are all dependent on runs, wickets, and balls bowled when looking at bowling performance characteristics. For bowler price prediction, we will examine wickets, runs, balls bowled, and price. Another benefit of using only these characteristics is that the machine will be able to tell the difference between two players' performances. This is because both the bowler's and batsman's average and strike rate are expressed as ratios. As a result, the model gives these parameters less weight in price prediction. Furthermore, each parameter has a positive slope when compared to the player's price. The model's advantage is that the slope behaves consistently.

Variance Inflation Factor

```
[ ] from statsmodels.stats.outliers_influence import variance_inflation_factor
    def get_vif_factors(x):
        x_matrix=x.to_numpy()
        vif=[variance_inflation_factor(x_matrix,i) for i in range (x_matrix.shape[1])]
        vif_factors=pd.DataFrame()
        vif_factors['column']=x.columns
        vif_factors['VIF']=vif
        return vif_factors
```

```
vif_factors=get_vif_factors(x[x_features])
vif_factors
```

	column	VIF
0	4w	1.899720
1	5w	1.388344
2	Avg	91.676530
3	Color	3.697675
4	Econ	34.948458
5	Inns	92.516282



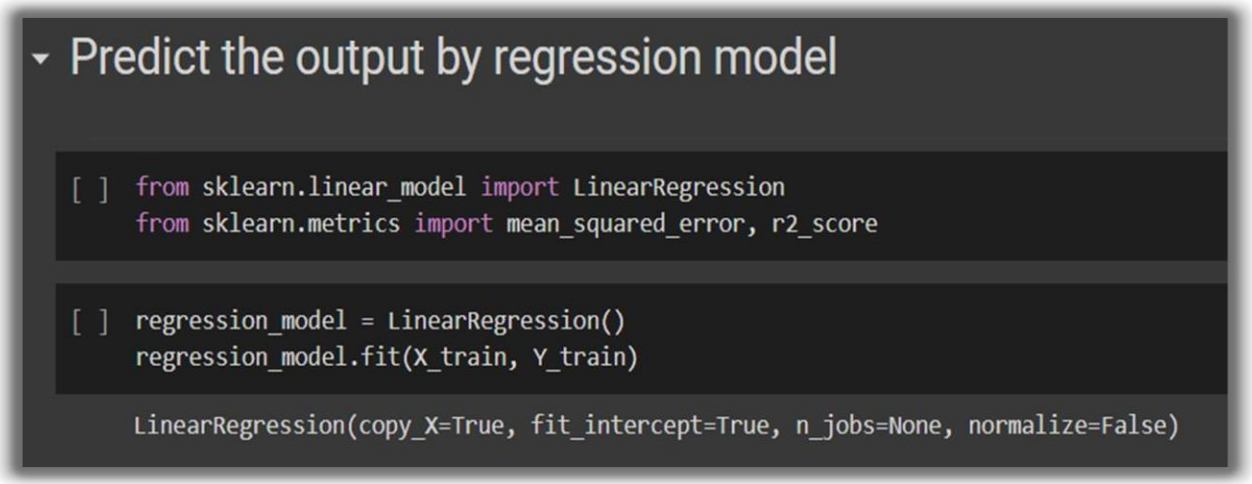
Sc 3.1: Feature Selection through Variance Inflation Factor and Heatmap

3.1.2 Algorithms used

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

- **Linear Regression**

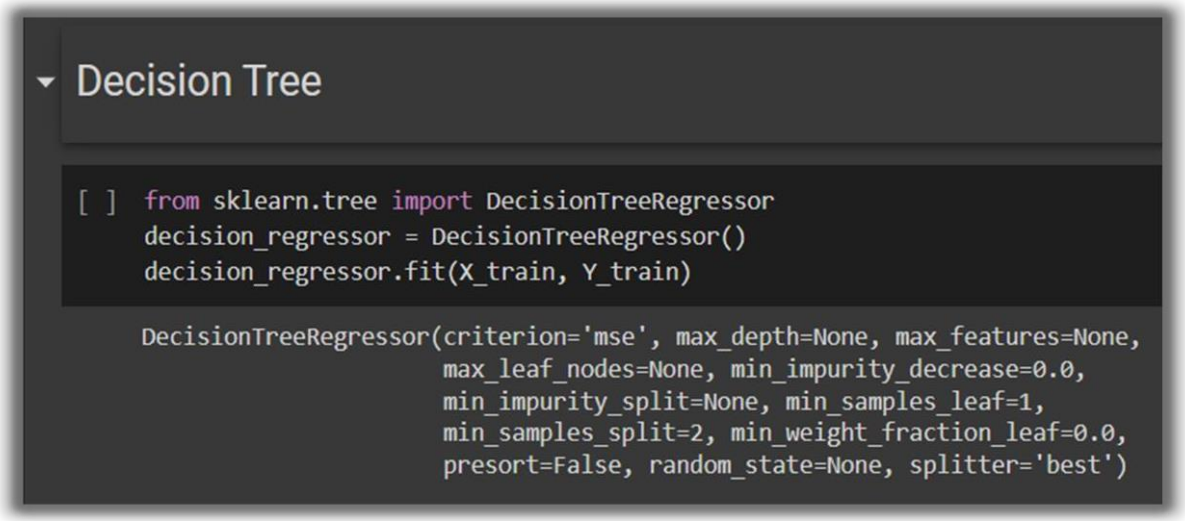
It is one of the most popular Supervised Machine Learning algorithms in Python that maintains an observation of continuous features and based on it, predicts an outcome. It establishes a relationship between dependent and independent variables by fitting a best line.



Sc 4.1: Linear Regression Model by sklearn library

- **Decision Trees**

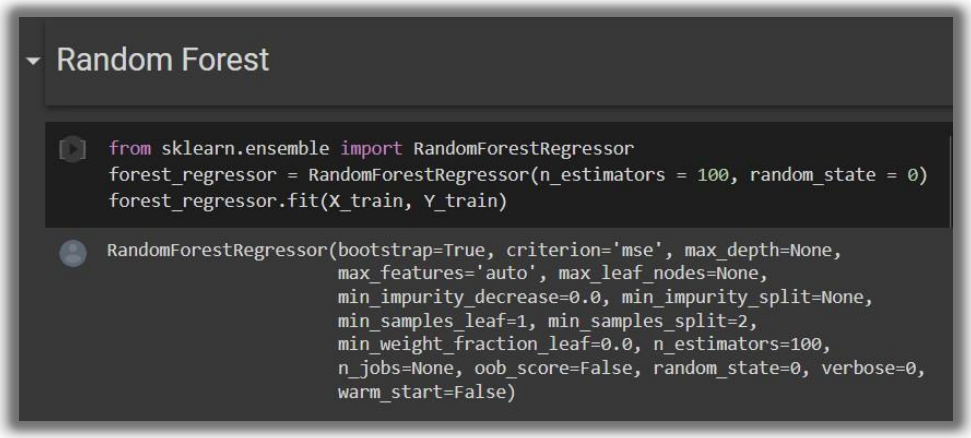
A decision tree is built by repeatedly asking questions to the partition data. The decision tree algorithm aims to **increase the predictiveness at each level of partitioning so that the model is always updated with information about the dataset.**



Sc 5.1: Decision Trees Model by sklearn library

- **The Random Forest Regressor**

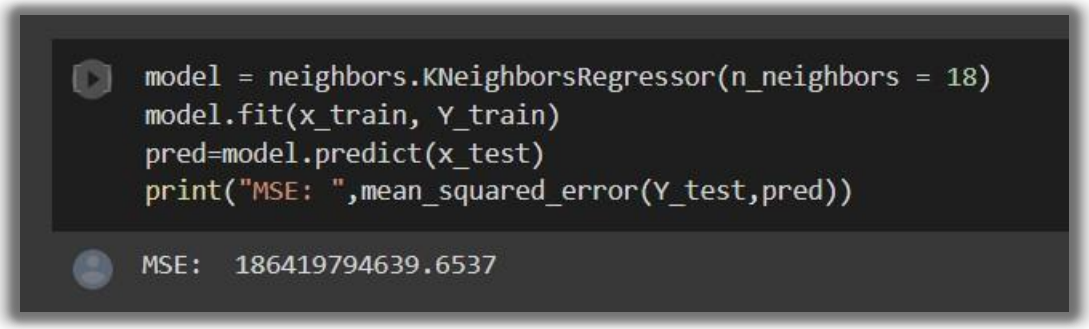
Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction.



Sc 6.1: Random Forest Regressor Model by sklearn library

- **K-nearest neighbor Regressor**

K-nearest neighbors is probably the simplest widely used model conceptually. KNN models are really just technical implementations of a common intuition, that things that share similar features tend to be, well, similar.

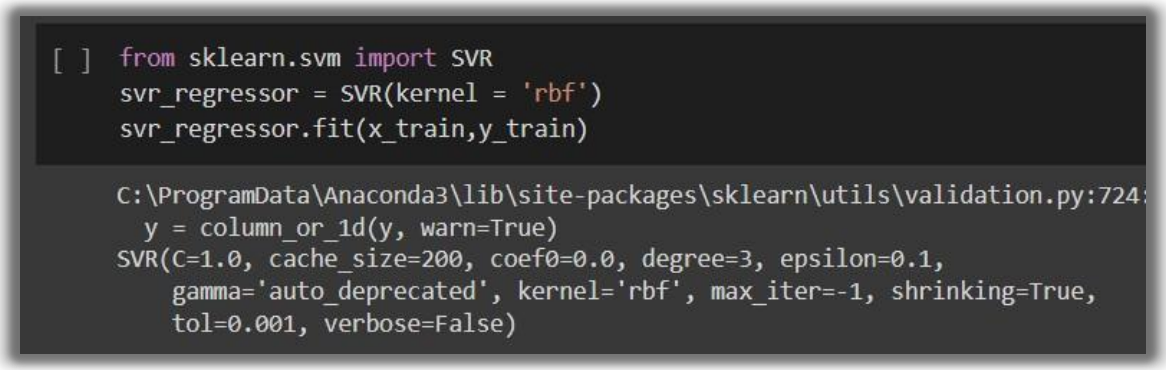


Sc 7.1: K-Nearest Neighbor Model by sklearn library

Elbow method helps data scientists to select the optimal number of clusters for KNN clustering.

- **Support Vector Machines (SVM)**

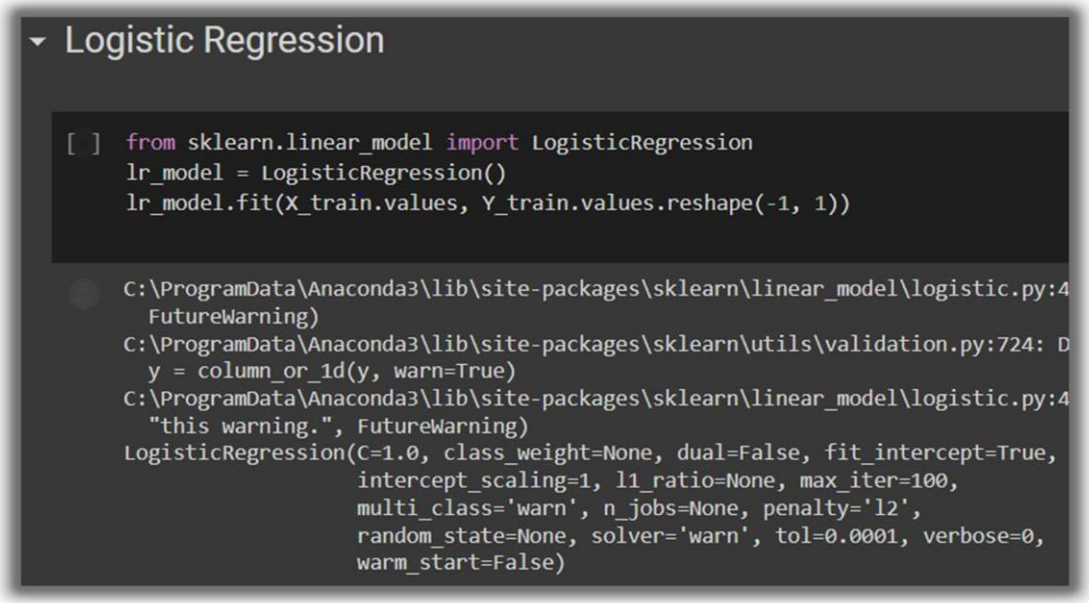
This is one of the most important machine learning algorithms in Python which is mainly used for **classification but can also be used for regression tasks**. In this algorithm, each data item is plotted as a point in n-dimensional space, where **n denotes the number of features you have, with the value of each feature as the value of a particular coordinate.**



Sc 8.1: Support Vector Machines Model by sklearn library

- **Logistic Regression**

Logistic regression is a process of modeling the probability of a discrete outcome given an input variable. The most common logistic regression models a binary outcome; something that can take two values such as true/false, yes/no, and so on.



Sc 9.1: Logistic Regression Model by sklearn library

3.2 Work Approach:-

The complete work done has compactly organized into this architecture. It will begin with the processing of datasets and loading them in the backend. Then user interface is provided with different functionalities, which can be performed on the player/match. It can also be used to perform prediction.

We will implement the following modules for analysis, prediction, ranking, and visualization.

– Processing of datasets

– Batsmen performance analysis

– Bowler performance analysis

– Match analysis

– Head-on-head analysis of teams

– Team overall performance analysis

– Ranking of teams

– Match prediction

– User interface creation

The below diagram illustrates the various modules of the proposed system. Modules of our proposed system are demonstrated in Fig.

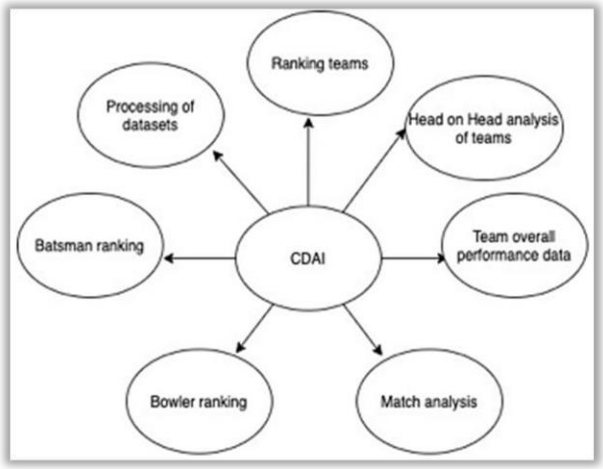


Fig 3.1: Modules of the system

Linear regression is effective for both national and international bowlers since it does neither over-fit or under-fit. We took into account the number of runs scored, the innings played, the best fast bowling, and the number of wickets taken. Positive slopes that were steady (rather than changing).

4. Result & Discussion

Because the model has over-fitted the data, the Mean Squared Error (MSE) for Indian and international players is very high. The model memorizes the data rather than learns it, as evidenced by low test scores despite adequate training.

The MSE of Linear Regression is the second lowest. It also demonstrated that IPL auctioneers are willing to spend more on overseas players. This is due to the fact that the team must include seven national players. Because of this limitation, the remaining players must be of excellent quality. This supply-demand imbalance causes price inflation. Because all of the curves created by graphing the price against each attribute have a positive slope. This problem can also be solved using linear regression.

10	4.226850e+08	42	1.996686e+09	145	7.592263e+10	16	2.736436e+10
368	1.411720e+11	206	6.412597e+07	487	1.831038e+10	171	8.405196e+07
351	6.728580e+09	391	2.074174e+10	49	1.042451e+10	236	8.683217e+08
387	1.519561e+10	235	1.761561e+09	458	3.552000e+10	300	2.232107e+09
302	2.517116e+08	251	4.474923e+09	19	4.691583e+09	533	6.011014e+10
336	8.785968e+08	142	1.341549e+10	194	7.741803e+10	231	8.488625e+09
366	4.977302e+09	48	3.071802e+10	336	1.133827e+09	141	2.985021e+08
16	5.319563e+09	104	6.431792e+06	349	7.748563e+10	212	1.061560e+12
209	1.129951e+10	83	6.254978e+11	125	9.023461e+08	388	1.459476e+09
79	1.813687e+12	122	1.940886e+08	228	1.141078e+10	413	4.222807e+03
350	7.212861e+11	182	1.125440e+11	431	2.432959e+10	114	3.961883e+09
371	1.741067e+11	347	2.276891e+09	147	1.488933e+09	35	6.550422e+09
352	3.357532e+08	49	8.450979e+08	315	1.225059e+11	451	7.607843e+06
153	4.663225e+07	377	4.835799e+11	333	8.313033e+09	51	1.910083e+10
256	5.146265e+09	186	2.874627e+10	387	3.332767e+10	60	1.214856e+09
319	2.940001e+10	204	1.967155e+11	167	1.817553e+10	99	1.087484e+08

Sc5.1: Predicted Values of Linear Regression Model of Indian Bowler

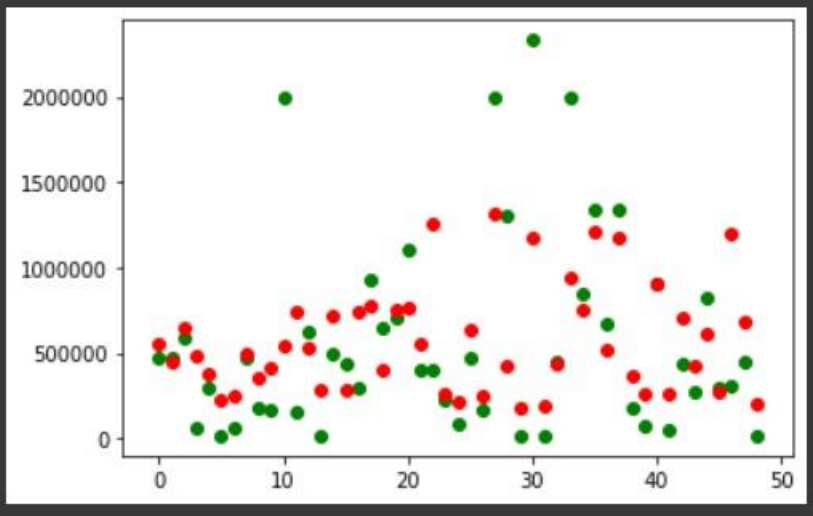
In order to cross validate our models, we had applied K-Fold method. We had kept 20% of data in the validation set. For obtaining an optimal cost and variance

Player Name	Actual	Predicted
Yusuf Pathan	475000	4.96E+05
Scott Styris	175000	2.48E+05
Rohit Sharma	2000000	6.53E+05
Ramesh Powar	170000	1.77E+05
Piyush Chawla	400000	4.91E+05

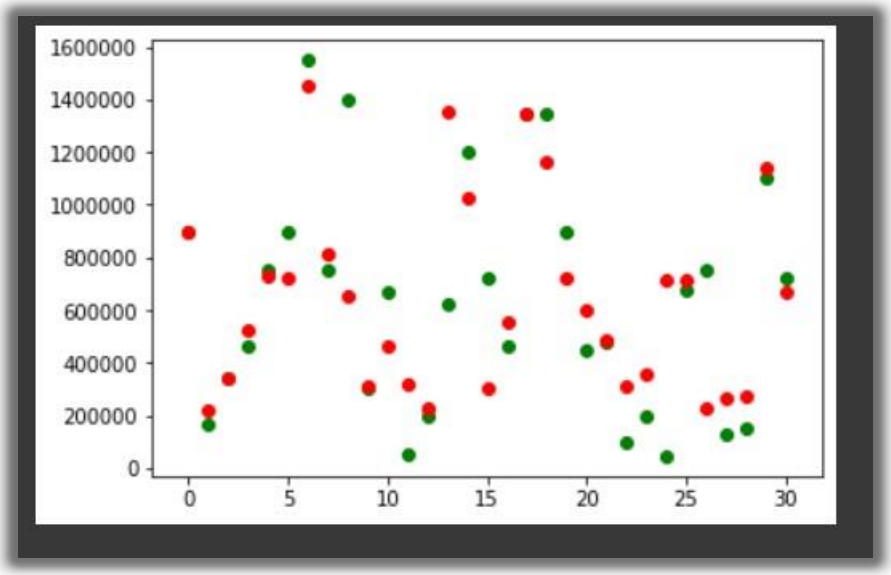
Sc5.2: Actual & Predicted Value of the player salaries

We can see the result difference by using the scatter plot too. Where the green points are real values and the red points are Predicted values. Some of the scatter plots are the following:

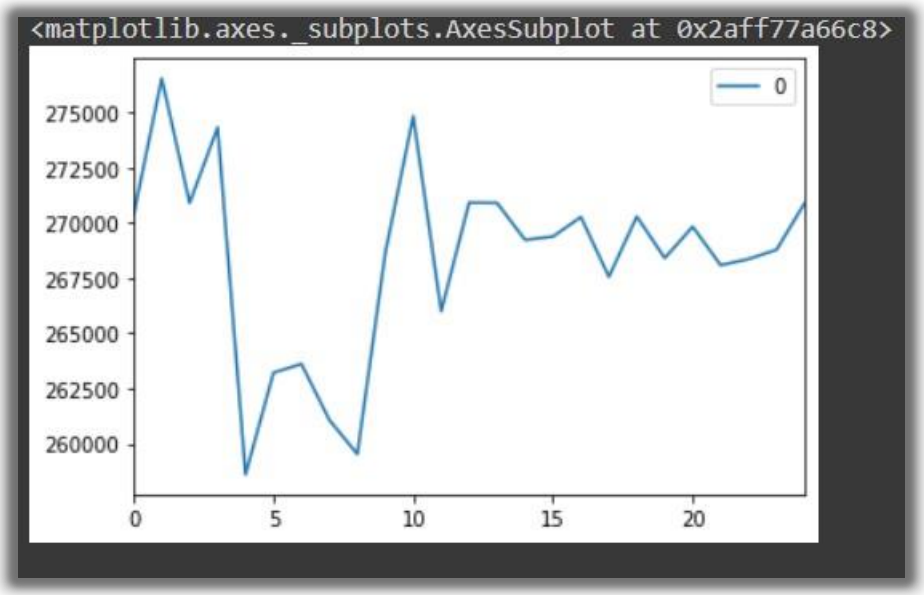
```
[ ]  plt.scatter(Y_test.index, Y_test, c='green')
     plt.scatter(x_test.index, pred, c='red')
     plt.show()
```



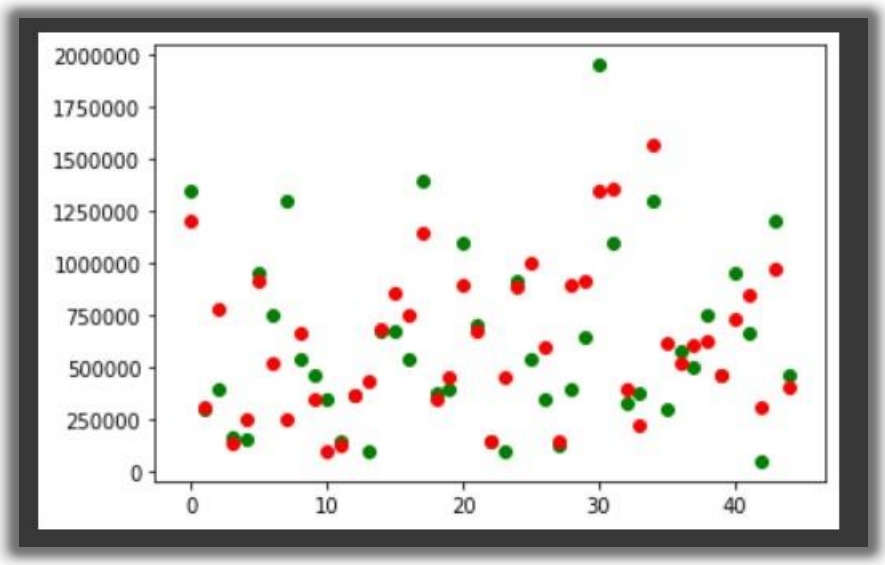
Sc10.1:-Indian Bowlers Knn Predicted Vs Real



Sc10.2:- Overseas Bowler SVM Predicted Vs Real



Sc10.3:- Overseas Batsman Knn elbow curve



Sc10.4:- Overseas Batsman knn predicted vs real

5. Conclusion

The linear regression model has caught all of the relevant characteristics for predicting batsman and bowler prices. The price of high-performing players rises, while the price of low-performing players falls. The average error per bowler is 4,50,000, and the average error per batsman is 5,20,000. In the worst-case scenario, the auctioneer will bid for 22 bowlers, resulting in a budget miscalculation of $6,00,00,000, or 8% of the $80,00,000 budget. This is okay because an auctioneer does not always spend all of his or her money and will never buy only bowlers. Furthermore, our approach may somewhat overprice some players while slightly underprice others. As a result, the budget will be minimally affected.

6. 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-leaguejumps-into-top-10-of-most-attended-sports-leagues-in-theworld-20160110-gm2w8z.html.

[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

[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-entropybased methodology for player selection: evidence from cricket,‖ 2020.

[13]. What is Python? Executive Summary,PYTHON.ORG,2021[Online].
Available: https://www.python.org/doc/essays/blurb/

[14]. 5 Most Used Machine Learning Algorithms in Python, GREAT LEARNING.COM ,2021. [Online]
Available:https://www.mygreatlearning.com/blog/most-used-machine-learning-algorithms-in-python/

[15]. Logistic Regression — Detailed Overview, TOWARDSDATASCIENCE.COM, 2021. [Online]
Available:https://towardsdatascience.com/logistic-regression-detailed-overview- 46c4da4303bc

[16]. Random Forest, CORPORATEFINANCEINSTITUTE.COM,2021,[Online]
Available:https://corporatefinanceinstitute.com/resources/knowledge/other/random-forest/

[17]. D. S. Sayad, "Support Vector Regression," 23 August 2021. [Online].
Available: http://www.saedsayad.com/support_vector_machine_reg.htm.