***IPL auction prediction system based on player performance.***

***-SOWMYA BALLARI***

**Abstract:**

The Indian Premier League (IPL) is the most entertaining league in cricket. Everyone who is interested in cricket has their eyes on IPL games. Every year each business in IPL comes up with strategic plans and decisions to buy players to play for their teams. With this paper, we aim to develop a model which can help business stakeholders in taking better decisions using key machine learning algorithms. Our model helps in finding the best player on whom a team can invest and buy during auctions. It also tells how much revenue can be earned with the invested money in the auction. Our model uses historical data about players' performances like style, statistics and players form in the previous league matches and auction data. We perform a player worth analysis for the price sold and predict high demand players in the previous leagues who can have strong form.  In our work, we use various machine learning models like Regression, Decision Trees, and Random Forest algorithms to identify the best model with accuracy. We are going to use Python language for implementation and provide reports on results. With this model, we help IPL team owners to take calculated decisions and improve their revenue.

**Keywords:**

Indian Premier League (IPL), One Day International (ODI), Twenty20 International (T20I), Machine Learning (ML), Mean Absolute Error (MAE), Mean Square Error (MSE), Root Mean Square Error (RMSE), Indian Rupee (INR).

**Introduction/Background:**

The IPL is a professional Twenty20 cricket league in India, which has gained immense popularity over the years. The league shows some of the biggest names in world cricket, and its auction is one of the most highly famous events in the cricketing calendar. The auction is a crucial aspect of the league, as it determines the composition of each team and can have a significant impact on their chances of success in the tournament. IPL opens auction table every year (mega auction session once in 3 years) for team franchise to sell the cricket players to the IPL teams like Chennai Super Kings, Mumbai Indians, Kolkata Knight riders, Sunrisers, Punjab Kings (formerly known as Kings XI Punjab), Delhi Capitals (formerly Delhi Daredevils), Gujarat Titans, Lucknow Super Giants and Royal Challengers Bengaluru.

With this paper, we are targeting to build a model which is not developed by any one when we are writing this paper. Our system is used to predict the IPL auction price worth of player based on his historical cricket statistics in IPL. We will be doing multiple modelling analysis and performance evaluation about batsmen and bowlers. We find out best predictive model out of all for our data. Our process includes all the steps required for any Machine Learning development model. We finally implement a system which can be used to predict auction price worth of a player for given performance statistics of the player. This system can be used as

Machine learning is defined as a system that involves the development of algorithms and statistical models which facilitate computers to learn and improve from experience, without being unequivocally programmed. It involves the use of statistical techniques to automatically recognize patterns in data and make accurate predictions or decisions based on that data. Machine learning algorithms can be used to solve a wide variety of tasks, such as image and speech recognition, natural language processing, anomaly detection, and predictive modeling. The goal of machine learning is to enable computers to perform tasks that would otherwise require human intelligence or expertise, and to do so more efficiently and accurately. Machine learning is classified as 3 types – Supervised, Unsupervised and Reinforcement learning. Supervised learning is an algorithm is trained on a labeled data set, where the input data is combined by a desired output or response. The algorithm learns to make predictions by finding patterns in the data that link the input to the output. The aim of supervised learning is to generalize from the training data and make true predictions on new, unseen data. Unsupervised learning is on which algorithm is trained on a dataset without having labels on what the output should be. Instead, the algorithm tries to find outlines and structure in the input data on its own, without being given any specific intentions.

All businessmen have strategies to grow the Return on Investment (ROI). Decision making plays crucial role in their day-to-day life. In the IPL auctioning session, quick decisions are essential for team owners/franchise to select best player who can perform well in their team for the season. Our model works as a catalyst which can give prediction about maximum price that can be invested on a player for his performance in the historical games.

**Project Objective:**

In this paper, we are aiming to achieve whether there is a relationship between IPL player statistics and the price at which there are sold in the auctioning. We are trying to identify the association by competing multiple algorithms using data we have. We are trying to build a proposal for IPL team franchise to predict their player buying prices using Machine Learning and plan for their auction budget accordingly with keeping higher profits.

We are developing this model for the first time which can benefit businessmen and stakeholders to take optimistic decisions on player purchases. We will be using Python language and Jupyter notebooks.

**Data/Problem Analytics:**

To achieve our end goal, suitable data is very critical. At first, we tried to find out data which can have player statistics and price sold in the historical auctioning sessions. After navigating the internet, we did not find what we expected. So, we later created our own data set for analysis and futuristic predictions by combining multiple data set like IPL player performance statistics and auctioning data. We have downloaded two different data set from Kaggle website, brainstormed required attributes and variables that are important for analysis.

**Data1 – player (batsman and bowler) performance statistics data:**

This data constitutes multiple cricket characteristics of the player from the year of 2016 to 2022. Data of batsmen and bowler statistics is previous as separately in different files based on years. The challenging part is the data that we found is not available in file, but it is divided into multiple files based on years. So, our one of the preprocessing steps include combining data of all years into one file of batsmen and bowler stats respectively.

Below table describes each attribute present in the batsmen dataset

|  |  |
| --- | --- |
| **Column Name** | **Description** |
| POS | Serial Number |
| Player | Name of the Player |
| Mat | Number of Matches the batsman played in the IPL for that year |
| Inns | Number of Innings the batsman played in the IPL for that year |
| NO | Not Out Count |
| Runs | Number of runs he scored |
| HS | Highest Score the batsman made |
| Avg | Average number of runs the batsman maintained against all matches |
| BF | Number of balls faced by batsman |
| SR | Strike Rate of the batsman |
| 100 | Number of 100s |
| 50 | Number of 50s |
| 4s | Number of Fours |
| 6s | Number Sixes |

In cricket, one individual game played between two teams called match and innings is number of matches does the player got the chance to play. Not Out represents the batsman did not get out in the complete match. Highest score made by the batsman in the year. Batsmen attributes like average, strike rate and number of fifties plays crucial role in determining batsman strength in a game.

Below table shows columns present in bowler data set

|  |  |
| --- | --- |
| **Column Name** | **Description** |
| POS | Serial number in the data |
| Player | Name of the Player |
| Mat | Number of Matches played by the bowler |
| Inns | Number of innings bowler have bowled |
| Ov | Number of Overs the bowler bowled |
| Runs | Number of runs conceded in the year |
| Wkts | Number of wickets |
| BBI | Best bowling in an innings |
| Avg | Number of runs bowler conceded per wicket taken |
| Econ | Average number of runs conceded in an over |
| SR | Average number of balls bowled per wicket |
| 4w | Number of four-wicket hauls |
| 5w | Number of five-wicket hauls |

In cricket, for any bowler the crucial attributes are wickets, average and economy which are required in determining the strength and agility of the bowler.

**Data2 – Auction data set:**

The auction data consists of data about players, the team who bought, year and sold price in the auction from the year of 2013 to 2022. This data has important attribute called sold\_price which is the money invested by the team franchise in a IPL season.

Below table shows columns present in Auction data set

|  |  |
| --- | --- |
| **Column Name** | **Description** |
| player\_name | Name of the Player |
| nationality | Nationality of the player |
| type | Type of player (batsman, bowler, wicketkeeper) |
| teams | Name of the team in the IPL |
| year | year of auction |
| sponsored\_by | Sponsorship of IPL |
| sold\_price | Amount at which player got sold |

There are a few redundant attributes as well in the auction data which we are going use a strategy to resolve them. Further we are going to use this data in most important step in this paper.

**Joining Data sets:**

As we illustrated in above section that our data is not readily available. We are creating our own dataset by joining two different datasets. When we look at the attributes present in datasets, we can see player name is common in both datasets. Hence, attribute player name can be used to join these two datasets.

Before joining these datasets, two significant things we notice is data in auctions dataset have player auctioned price the years from 2013-2022. Firstly, as we considered two different IPL data sets, the naming convention might be different. Hence we removed case-sensitivity on player name column. Secondly, there is a chance of same player got auctioned in two or more years, to resolve this problem, we combined each player present multiple times in the dataset and taken average of the sold price to maintain parity. Once we overcome the problem, we joined two datasets and processed further to the next steps in the process.

**Data cleaning:**

Data cleaning is imperative in machine learning. Data cleaning is the process of cleanse the data as required for developing machine learning model. Our dataset has couple of data cleaning steps included like removing unwanted columns, finding duplicates and removing them etc.

**Removing unwanted columns:**

In our combined dataset, we have few unnecessary columns which are redundant in our machine learning data flow. Hence, we dropped them before developing model. The columns that we dropped from batsmen data set are: "SR", "Avg","50", "Player",'year', "Unnamed: 0", '100', '4s', '6s', 'player\_name' and from bowler data set are: "POS","BBI","Mat","Avg", "Econ", "SR","4w","Unnamed: 0\_x", "5w", "Unnamed: 0\_y", "year", "Player", "player\_name" , "type"

**Removing duplicates:**

Another crucial step in the data cleaning process is removing duplicates. Keeping duplicates in the data set will lead to ambiguous results. So, we will find if there are any duplicate observations and remove them.

**Data Pre-processing:**

We found correlation matrix before applying ML algorithms to identify the correlation between attributes present in batsmen data set and bowler data set.

Below image shows correlation matrix on batsmen data set.

Chart, treemap chart

Description automatically generated

Below image shows correlation matrix on bowler data set.

Chart, bar chart

Description automatically generated

Secondly, target variable (sold\_price) units are in INR with values of bigger scale. This can lead to abnormal results in the prediction. Hence we divided sold\_price attribute with 1,000,000 (1 million) to convert values in small scale.

A picture containing table

Description automatically generated

Same applied for bowler data set as well

A picture containing table

Description automatically generated

Machine learning development include training the model and evaluating the model about performance. Therefore, we divide player stats auction data set into training subset and test subset in the percentage of 90% and 10% respectively. Now we have batsmen and bowler training and test data sets for both predictors and target attributes.

**Developing the model:**

Once we make our data ready, next step is to start developing the best model for our datasets. As we can see our consists of complete numerical attributes hence Regression is the best modelling technique which can be used. Regression is a statistical technique used to analyze and model the relationship between a target variable and one or more predictors. In other words, regression helps us to understand how changes in one variable are related to changes in another variable. The target variable is the outcome variable that we want to predict, while the independent variables are the variables that we use to predict or explain the dependent variable.

Each step we do analysis on batsmen data and bowler data separately since the attributes of batsmen and bowler are different in real. We will be studying different regression models on data and identify best suitable model for our application.

**Linear Regression:**

Linear regression is a method used to model the association between a dependent variable and one or more independent variables by fitting a linear equation to the data. The linear equation takes the form of a straight line, with the dependent variable on the y-axis and the independent variable on the x-axis.

Y = b0 + b1xz + b2x2 +……. + bnxn

Where, Y = target variable

x = predictors

b = coefficients

Modelling batsmen data using Linear Regression

Graphical user interface, text, application, email

Description automatically generated

Modelling bowler data using Linear Regression

Graphical user interface, text, application, chat or text message

Description automatically generated

**Decision Tree Regression:**

A decision tree is used as a predictive model to independent variables to outputs dependent variable in the non-parametric regression approach known as decision tree regression. Each internal node in a decision tree represents a test on an input variable, and each branch denotes the test's result. The values of the dependent variable are represented by leaves of the tree.

Modelling batsmen data using Decision Tree Regression.

Graphical user interface, text, application, email

Description automatically generated

Modelling bowler data using Decision Tree Regression

Graphical user interface, text, application, email

Description automatically generated

**Random Forest Regression:**

Random Forest regression is a popular ensemble learning model for regression problems. It combines multiple decision trees to improve the accuracy and stability of the prediction. In Random Forest regression, each decision tree is constructed using a random subset of the training data and a random subset of the features.

Modelling batsmen data using Random Forest Regression

Graphical user interface, text, application, email

Description automatically generated

Modelling bowler data using Random Forest Regression

Graphical user interface, text

Description automatically generated

**Adaptive boost Regression:**

AdaBoost (Adaptive Boosting) regression is an ensemble learning technique for regression problems that combines multiple weak regression models to create a stronger model. The basic idea of AdaBoost regression is to iteratively train a sequence of weak regression models on the data, and to adjust the weights of the training samples based on the errors of the previous models.

Modelling batsmen data using Ada boost Regression.

Graphical user interface, text, application, email

Description automatically generated

Modelling bowler data using Ada boost Regression.

Graphical user interface, text, application

Description automatically generated

**Model Evaluation:**

We will compare each model based on widely used regression evaluation techniques like Mean Absolute Error (MAE), Mean Squared Error (MSE) and Root Mean Square Error (RMSE).

**Mean Absolute Error:**

MAE is a measure of the average magnitude of the errors between the predicted values and the actual values in a regression model. It is calculated as the average absolute difference between the predicted values and the actual values.

The formula for calculating MAE is:

MAE = 1/n \*

where n is the number of observations, yi is the actual value of the dependent variable for the ith observation, and ŷi is the predicted value of the dependent variable for the ith observation.

The MAE is a useful indicator of a regression model's accuracy since it shows how distant, on average, the predicted values are from the actual values. The model does better at forecasting actual values when the MAE is smaller.

**Mean Squared Error:**

MSE is a measure of the average squared difference between the predicted and actual values in a regression model. It is calculated as the average of the squared differences between the predicted and actual values.

The formula for calculating MSE is:

MSE =

where n is the number of observations, yi is the actual value of the dependent variable for the ith observation, and ŷi is the predicted value of the dependent variable for the ith observation.

MSE is a helpful indicator of a regression model's accuracy since it shows how well the model fits the data. The MSE indicates how well the model predicts real values; the lower the MSE, the better.

**Root Mean Square Error:**

RMSE is a measure of the average magnitude of the errors between the predicted values and the actual values in a regression model. It is calculated as the square root of the average of the squared differences between the predicted and actual values.

The formula is:

RMSE =

where n is the number of observations, yi is the actual value of the dependent variable for the i-th observation, and ŷi is the predicted value of the dependent variable for the i-th observation.

The RMSE is a valued indicator of a regression model since it provides information on the average distance between the predicted and actual values in the same units as the dependent variable.

Results from Linear Regression model on batsmen data

---- Linear Regression - Model Evaluation ----

Mean Absolute Error (MAE): 17.942817283407635

Mean Squared Error (MSE): 600.2206669166255

Root Mean Squared Error (RMSE): 24.499401358331706

Results from Linear Regression model on bowler data

---- linear Regression - Model Evaluation ----

Mean Absolute Error (MAE): 15.13738547730593

Mean Squared Error (MSE): 287.9894676082754

Root Mean Squared Error (RMSE): 16.970252432072876

Results from Decision Tree Regression model on batsmen data

---- Decision Tree Regression - Model Evaluation ----

Mean Absolute Error (MAE): 18.548924731182794

Mean Squared Error (MSE): 764.8703853046595

Root Mean Squared Error (RMSE): 27.656290158021186

Results from Decision Tree Regression model on bowler data

---- Decision Tree Regression - Model Evaluation ----

Mean Absolute Error (MAE): 25.24621212121212

Mean Squared Error (MSE): 1436.683869949495

Root Mean Squared Error (RMSE): 37.903612887817104

Results from Random Forest Regression model on batsmen data

---- Random Forest Regression - Model Evaluation ----

Mean Absolute Error (MAE): 18.504492661684193

Mean Squared Error (MSE): 671.2131050686381

Root Mean Squared Error (RMSE): 25.907780782395047

Results from Random Forest Regression model on bowler data

---- Random Forest Regression - Model Evaluation ----

Mean Absolute Error (MAE): 11.427830086580087

Mean Squared Error (MSE): 177.66316780458922

Root Mean Squared Error (RMSE): 13.329034766425858

Results from Ada boost Regression model on batsmen data

---- Ada Boost Regression - Model Evaluation ----

Mean Absolute Error (MAE): 18.874732514437156

Mean Squared Error (MSE): 557.9591596173399

Root Mean Squared Error (RMSE): 23.621159150586575

Results from Ada boost Regression model on bowler data

---- ADA Regression - Model Evaluation ----

Mean Absolute Error (MAE): 15.30151574103798

Mean Squared Error (MSE): 311.91767914045016

Root Mean Squared Error (RMSE): 17.66119132845942

**Findings/Conclusions:**

As we clearly analyzed various types of regression techniques in the previous section. This is the time for decisioning based on the outcomes we got while modelling. We have developed different regression models using Linear, Decision Tree, Random Forest and Ada Boost techniques. We are evaluating performance of these models based on MAE, MSE and RMSE. We will first interpret results from batsmen data models then we interpret bowler data regression models.

**Batsmen models evaluation:**

Summary of the results obtained on batsmen data set.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Linear** | **DT** | **RF** | **ADAB** |
| **MAE** | **17.94** | 18.54 | 18.5 | 18.8 |
| **MSE** | **600.22** | 764.87 | 671.21 | 557.95 |

When we compare results of different models on batsmen data. we can see the Linear Regression performance is better. MAE value is less when compared other and MSE is also not so large. Hence, we decided to go with Linear model for batsmen data.

Below screenshot shows that we are using linear regression for predicting sample input data. We have taken 3 different set of IPL batsmen stats inputs taken randomly for testing the model.

Input 1: Innings: 228, Runs: 4163

Output 1: 127.6 million Rupee

Input 2: Innings: 134, Runs: 2838

Output 2: 76.31 million Rupee

Input 3: Innings: 1, Runs: 10

Output 3: 16.39 million Rupee

Graphical user interface, text, application, email

Description automatically generated

**Bowler models evaluation:**

Summary of results on bowlers data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Linear** | **DT** | **RF** | **ADAB** |
| **MAE** | 15.13 | 25.24 | **11.42** | 18.87 |
| **MSE** | 287.98 | 1436.68 | **177.66** | 311.91 |

When we compare results of different models on batsmen data. we can see the Random Forest Regression performance is better. MAE value is less when compared other and MSE is also not so large. Hence, we decided to go with Linear model for batsmen data.

Below screenshot shows our bowler being sold price prediction system using Random Forest regression prediction for predicting sample input data. We have taken 3 different set of IPL bowler stats inputs taken randomly from real auction data for testing the model.

Input 1: Innings: 20, Overs:80, Wickets: 20, Runs: 600

Output 1: 25.82 million Rupee

Input 2: Innings: 40, Overs:120, Wickets: 60, Runs: 1229

Output 2: 22.96 million Rupee

Input 3: Innings: 60, Overs: 240, Wickets: 142, Runs: 1692

Output 3: 15.97 million Rupee

Graphical user interface, text, application, email

Description automatically generated

**Managerial Implications:**

We have developed a model which can predict IPL player being sold on the auction hammer based on his IPL career. We found out best model suitable for doing this job. This model or a system can be highly beneficial to the IPL team franchise owners while making decisions. Business owner can now decide which player can be bought at what price. They can further make convictions amount that can be invested on player. Further, they can use this system for budgeting and money purse during the auction session.

**Idea Sharing:**

The main challenge we faced during the work is unavailability of data. We have collected our own data for analysis and prediction. Later, we merged multiple data sets from internet and solved the problem. We have used correlation matrix to identify usefulness of the attributes that are available in the data set that we created. Next, we have run multiple models to finally come up with these illustrated models in our research work. Finally, we understood the results from the models and evaluated based on core parameters of regression analysis to identify best suitable model for our data.

This model can be used as base platform for any kind of future work in this area. To name a few, a) the same model can be further redesigned get more accurate results by training the model based on other formats of the game like T20I and ODIs. b) we can develop model which can integrate business auction purse with investment on players and predict the best player to whom can buy with the remaining purse amount. c) Unsold player analysis: what is the reason for player left unsold & A model to predict better player out of unsold players.

**References:**

1. Player statistics data set downloaded from Kaggle - https://www.kaggle.com/datasets/anandkumarsahu09/ipl-player-stats-20162022
2. Auction data downloaded from Kaggle - https://www.kaggle.com/datasets/kuchhbhi/ipl-auction-20132022-data
3. Rastogi, S.K., & Deodhar, S.Y. (2009). Player Pricing and Valuation of Cricketing Attributes: Exploring the IPL Twenty20 Vision. Vikalpa: The Journal for Decision Makers, 34, 15 - 24.
4. Malhotra, Gaurav. ‘A Comprehensive Approach to Predict Auction Prices and Economic Value Creation of Cricketers in the Indian Premier League (IPL)’. 1 Jan. 2022: 149 – 170.

**Appendix:**

We have used Jupyter Notebooks the code is present in 3 notebook files (Batsmen Stats 16-2022.pdf is for merging batsmen data, Bowler Stats 16-2022.pdf is for merging bowler data and IPL auction prediction system based on player performance.pdf is for models).