Prediction of players prices in IPL using machine learning algorithms

SHAMIR ROY and RUDRA SAHA*, BRAC University, Bangladesh

In this work, we have applied machine learning algorithms to predict the price of players in the IPL league using machine learning algorithms. Here the price of the players were predicted using certain parameters such as previous historical data of the players as well as their nationality and more. Using these parameters, the model was trained which later on predicted the prices of the players accordingly. The machine learning algorithms that were used in this work are Linear regression and Random forest regression. Among these algorithms Random Forest Regressor gave the maximum accuracy result for predicting player prices. These algorithms in this work can produce outputs within a few seconds which can help the auctioneers to make best decisions.

Additional Key Words and Phrases: Machine Learning, Linear Regressor, Random Forest Regressor

ACM Reference Format:
Shamir Roy and Rudra Saha. 2024. Prediction of players prices in IPL using machine learning algorithms. 1, 1 (May 2024), 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The IPL or Indian Premier League is an annual Indian T-20 league where 10 teams contest with each other to win the trophy. The business of sports is changing with time, so are the techniques of the team. Evaluation of the players has been declared as so important, anyone can practically hear these words from the decision-makers, and their accurate selection is always a game-changer for the teams' strategies, economic decisions as well as fan engagement. It is common knowledge whether it's the player trade market in football, the player drafts in basketball, or the cricket player auctions, the ability to predict precisely the players' prices very often leads to success for all parties involved: the teams, the agents, and the sports analysts.

Basically, player valuation used to be carried out in an old-timey way that mostly consisted of personal opinions provided by the scouts, coaches, and the sports experts along with some of the performance statistics of the player in the past. However, the introduction of ML techniques has brought about significant changes to the process because of the possibility to predict the prices through the usage of the data which is capable of being requested by the very things about the players that have an impact on the performance and the bigger picture.

Winning auctions and getting the best players for a particular team is always the top priority of a team's auctioneers. But it's not easy to win auctions. Because the IPL auction is always unpredictable and no one knows what the sold price of a player will be. So here it becomes hard for the team owners to make a precise choice that will help their team to have the best possible players. However, this paper deals with this issue.

This research paper aims to explore the machine learning algorithms such as Linear Regression and Random Forest Regression to predict the IPL player prices by leveraging from the vast dataset of the players' previous auction records with the history of their performances.

Authors' address: Shamir Roy, shamir.roy@g.bracu.ac.bd; Rudra Saha, rudra.saha@g.bracu.ac.bd, BRAC University, Merul Badda, Dhaka, Bangladesh.

The importance of this paper is this study has potential to provide accurate actionable insights for the stakeholders in the IPL. By harnessing the power of machine learning algorithms the teams in the IPL can make the best decisions regarding player prices which will help optimize their competitive advantage on the field.

Moreover, winning the auction is very crucial to have good players in a team for IPL stakeholders as having better players increases the possibility of winning the Cup and thus increases the total revenue of a team. So having this kind of ML driven work really helps a lot to make the best decisions during IPL auctions. So in this paper we have worked with a machine learning approach to have accurate predictions of the player prices.

Apart from that, very little research have been done in this field. There was one such a research which was according to the study by Jhansi Rani P. et al. [?], the player prices of IPL can be predicted using previous Ipl history data as well as IPL auction records. Besides, according to the study by Das et al. [?], the modified hedonic based model aims to predict the IPL player prices by estimating the hedonic price equations for the players. This paper compares with modified hedonic regression and eXreme Gradient boosting based models, this paper states that hedonice based model provides better prediction accuracy. According to another study by Sukanya Chakrabarty [?] the Ipl player prices could be predicted using machine learning approach and they created a model to predict IPL player prices.

2 BACKGROUND

Here we discuss the dataset, Linear Regression and Random Forest Regression that were implemented in this work.

2.1 Dataset

The dataset that was used in this work consists of 26 columns and 130 rows. The dataset holds the records of the previous IPL auctions. The dataset contained some columns like player name, age, country, total runs, strike rate, total sixes, ODI runs , total wickets and more. The dataset contained some categorical data features which were preprocessed to produce some columns that could be used to train and test the model. The dataset was relatively small. However, a bigger dataset would be more helpful to give more accurate predictions results. Because training the model with more data would really increase its effectiveness as well as its accuracy.

2.2 Linear Regression

Linear Regression is a statistical model which estimates between a dependent variable and one or more independent variables. Modeling is assumed by the fact that a linear relationship exists between the variables. The job here is to find a linear equation that best fits the data. Having a single independent variable means it is simple linear regression. If there are more than one independent variable then it is called multiple linear regression. The line that's determined is used to make future predictions. Linear regression is used in machine learning. In this work this model was used to make predictions of the prices of players.

2.3 Random Forest Regression

The random forest regression is a machine learning algorithm that is used to do regression tasks by plotting the data points. In random forest regression, instead of a single decision tree multiple decision trees are built while training the model. For each bag, outputs of many should be used to predict the final output which is used in the regression tasks. To clarify, during making each tree, the subsets of both training data and features are chosen based on the principles of randomness which may reduce overfitting and increase the accuracy. Random forest is well known for dealing with overfitting issues. Thus it gives better output results and also have robust performance.

2.4 Input

The input of this project was the dataset that consisted of the auction records of IPL league. The dataset had 26 columns and 130 rows which is pretty low. However, using a bigger dataset consisting of more records of IPL auctions will significantly increase the efficiency, efficacy and accuracy of the model. Training the model with more data rows will make it better and more accurate. There were some categorical data columns in the dataset which were preprocessed and converted into some other meaningful columns that could be used to train the model. The dataset columns contained mostly the previous track records of the players pointing to their past performances. Besides, their names, nationality were also there. However, their name column was dropped as it was not important at all. So the model's input was the dataset that contained the track records of the players and their past auction details. Here two major machine learning algorithms were implemented in this work which were Linear regression model and Random Forest regression model. Using these two ML models the data was trained. Here some low correlated features columns were dropped from the preprocessed dataset before that was trained to the model.

2.5 Output

The output of the model was the predicted sold price of the players. Here the dataset was preprocessed and then the preprocessed dataset was trained to the machine learning models for output. However, some less important columns with very low correlations columns were dropped for more consistent results. The Linear regression and Random Forest models were used here which then predicted the sold price of the players in IPL.

3 METHODOLOGY

In this work we used the machine learning approach to predict the prices of IPL players in auction to give more insights to stakeholders of IPL teams so that they can make better decisions in the auction. Here in this work the dataset was used which had previous track records of previous IPL auction data along with the history records of IPL players. The dataset was preprocessed and many more columns emerged and later on the low correlated features columns were dropped to give more consistent output. Thus the preprocessed dataset was trained into the model. Here two main machine learning algorithms were used. One is Linear Regression and another one is Random Forest Regression. Using these two models the data was trained to predict the sold price of the players.

3.1 Objectives

This work deals with machine learning algorithms such as Linear Regression and Random Forest Regression to predict the player prices in the IPL. Such an approach can be used for not only the IPL but also for other sports leagues. The primary objectives of this work is as follows:

- To predict the prices of the players in IPL
- Helping the IPL team stakeholders to make more efficient and accurate decisions.
- Helping the IPL team stakeholders to have the best players.
- Helping the IPL team stakeholders to form a better team so that their team comes on the top, increasing the profit and popularity of their particular team.

3.2 Study selection

In this review of the project , two main machine learning algorithms were used that are Linear Regression and Random Forest Regression. Using these two algorithms the model was implemented and the preprocessed dataset was trained to predict the prices of IPL players.

3.3 Linear Regression

In machine learning, linear regression is a supervised learning algorithm which is used for predictive analysis. It's used when there is a target variable and one or more independent variables. In the Linear regression model, a linear equation is determined between the target variable and independent variables. Using that the predictions are made. So the main objective is to find the best fitted relationship between the independent variable, X and the target variable, Y. There are two types of linear regression model. Simple linear regression and multiple linear regression.

3.4 Simple linear Regression

In simple linear regression, there is only one independent variable. The relationship between X and Y can be expressed as follows:

$\mathrm{Y} = \beta_0 + \beta_1 X + \epsilon$

- Y is the dependent variable
- X is the independent variable
- β_0 is the y-intercept
- β_1 is the slope of the line
- ϵ is the error term

3.5 Multiple linear Regression

In multiple linear regression, there are multiple independent variables. The relationship between the independent variables and independent variable can be expressed as follows:

$\mathrm{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n$

- $X_1, X_2, X_3, \ldots, X_n$ are the independent variables.
- β_0, β_1, β_2, β_n are the coefficients associated with the independent variables.
- ϵ is the error term.

3.6 Random Forest Regression

Random forest regression is a powerful machine learning algorithm which is used for regression tasks that are concerned with predicting continuous values. It is a part of an ensemble learning family that combines multiple individual models to produce more accurate predictions.

The working principles of Random forest regressor -

- Ensemble of Decision trees: Random forest Regression builds and ensemble of decision trees while the training phase runs. Each decision tree is constructed based on randomly selected subsets of the training data and a random subset of features.
- Random feature selection: At each node of decision tree, a random subset of features is taken for splitting. This randomness helps to get rid of overfitting.

- Bootstrap Aggregation: Random forest has a technique called bagging where multiple decision trees are trained on different subsets of training data. This helps to introduce diversity among the trees and this improves the overall generalization performance.
- Voting for prediction: While at the time of prediction, each decision tree in the forest predicts the target variable. For regression tasks, the final prediction is often the mean or median of the predictions made by all the trees in the forest.

Advantage of Random forest regression -

- High accuracy: Random forest regression trees provide high prediction accuracy.
- Resistance to overfitting: The randomness in this model helps to reduce overfitting.
- Non-linear relationships: Random forests can capture complex non-linear relationships between the target variable and the independent variables which makes it more suitable for a wide range of regression tasks.

3.7 Data Preprocessing

In this work, the dataset that was used contained previous auction history of players in the IPL as well as the history of player performances in the previous IPL franchises. The dataset contained a total of 130 rows and 26 columns. In the data preprocessing phase, the features columns such as Team, Playing role and Country were divided into categorical columns using getdummies() method from pandas. Then the columns such as Team, Playing role and Country were dropped. After this the Player name was dropped. After this the target column named Sold price was dropped as well. Finally, 25 columns with highest correlations were taken and WI, KXIP, ENG and Batsman columns were dropped before the dataset was splitted for training.

3.8 Implementing models

For this work, two regression models were used. One is Linear regression and another one is Random forest regression. The final preprocessed dataset was split. 10% of the data was for testing and 90% of the data was for model training. After this, the data was scaled and was given to the models for training. Thus, the final predictions were made.

4 RESULTS

After running the models perfectly, the results that it showed were fairly good with the dataset considering how small it was.

The results are given below -

- The accuracy rate of random forest regression model was 77.018%.
- The accuracy rate of linear regression model was 63.098%.

5 DISCUSSION

The overall accuracy rates of the models were good depending on how small the dataset was. It had only 130 rows and 26 columns to begin with. However, with a better and bigger dataset, the accuracy rates of the models in this project will significantly be higher.

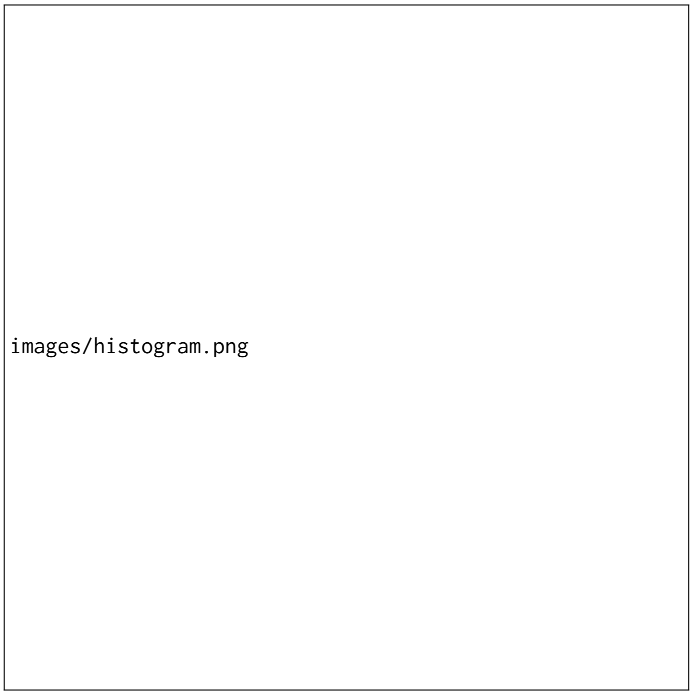

Fig. 1. Histogram of the dataset

6 LIMITATIONS

The only limitation of this work is the small dataset. The dataset of such kind was rare which created an issue to produce a much higher prediction score. However, this limitation can be solved by using a bigger dataset which can significantly improve the efficacy and accuracy of the models.

7 CONCLUSION

This project dealt with the machine learning approach to predict the prices of IPL players. The most important objective of this paper is that this project can really be helpful when it comes to making the best decision in the IPL Such kind of project can also be incorporated to predict the prices of players in other famous sport leagues that can help the stakeholders of a team to make the best decisions to win the auctions and this can significantly help them to take their team at the top of the league. For example, this kind of model can be used in famous football leagues such as LaLiga which can significantly help the stakeholders of a team to take the best decisions in the auctions. Here in this work linear regression and Random forest regression models were used. Thus with the power of machine learning algorithms it was possible to predict the prices of players in the IPL.

REFERENCES

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009

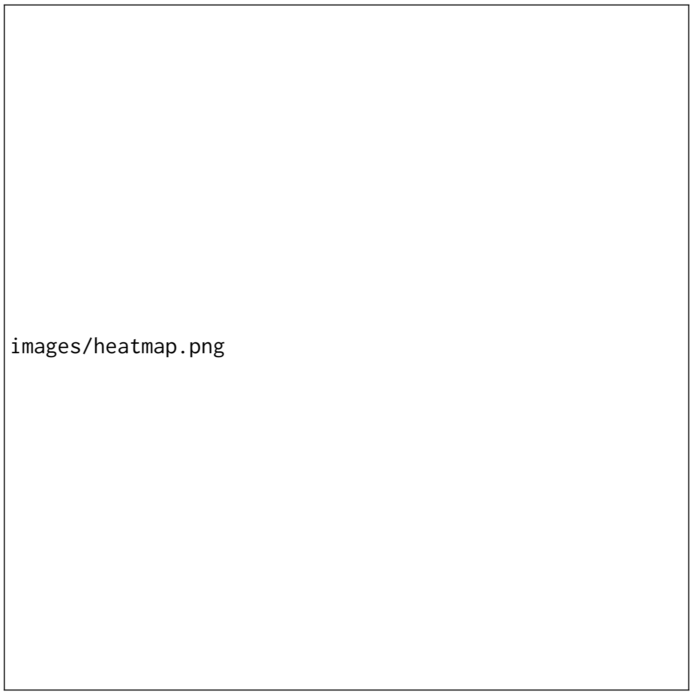

Fig. 2. Heatmap of the preprocessed dataset

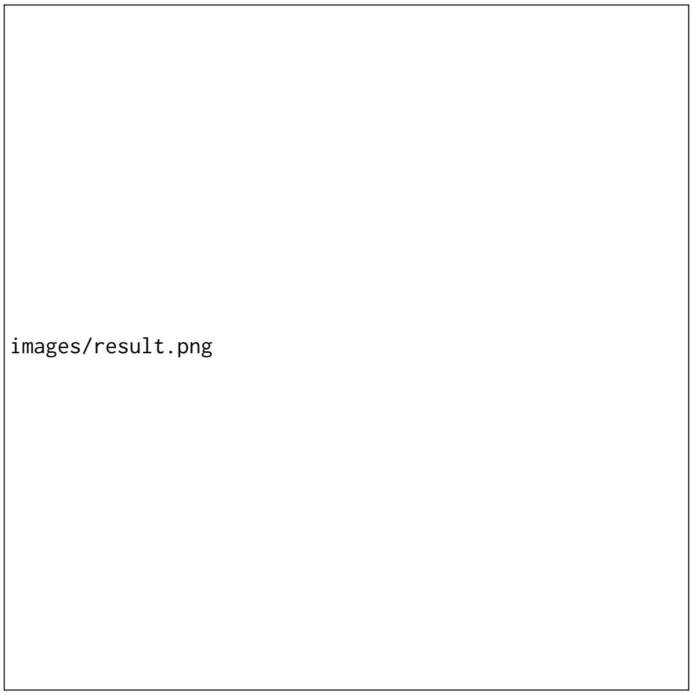

Fig. 3. Result of model

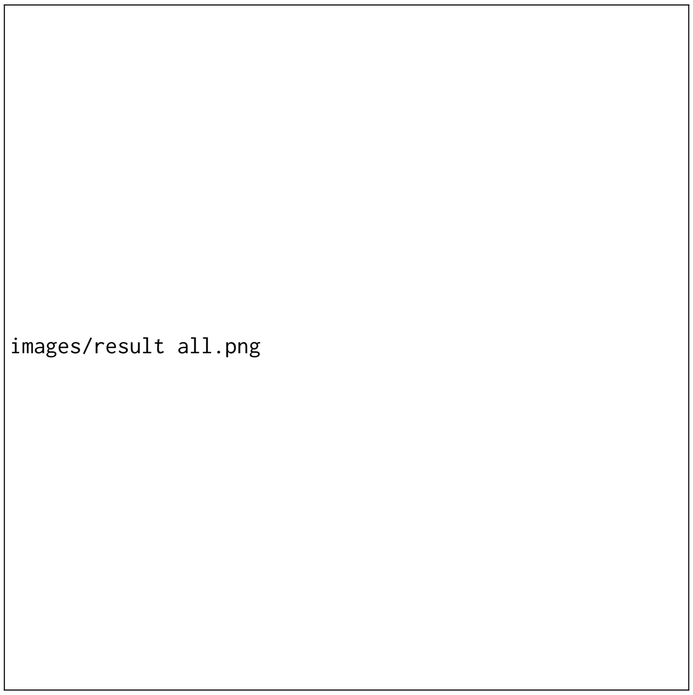

Fig. 4. Accuracy rate of the project