Machine Learning Based Pricing Model For International Football Players

Manas Swami, Tanishk Gangrade

Abstract: At international level players are picked up by clubs at various prices ranging from very low amounts to very high prices.Datasets of recent past on vital statistics and playing skills of international soccer players along with their price worth are available on various portals for analysis. Authors have built a decision tree regressor based upon past data to predict price of player based upon player’s attributes. Fitted Decision tree regressor is tested to produce results with high accuracy. Authors have also reduced dimensionality by finding correlation among independent variables. Dataset obtained after dimensionality reduction has been used to construct again a decision tree regressor, that has been tested and found to produce high accuracy without loss due to dimensionality reduction. Authors have used Football analytics dataset of FIFA 2019 and machine learning using Python to develop the regressor.

Keywords: Machine Learning, Football, Regression

I. INTRODUCTION

Football is extremely popular game worldwide, at the same time football game involves huge money. Money is involved in the form of prices of international players paid by clubs and profit made by teams. At international level there are domestic clubs. However there is no region based restriction on players and they can join clubs not located in the native region of players. As an example, Chelsea - a London based club had 2 players each from Brazil, Spain, France, Belgium and 1 each from Nigeria, England and Serbia [1]. With the advent of increased computing powers, internet, data collection sensors and devices, lot of players’ data is being generated and collected. Automated analytics have become inevitable in games like Basketball, Baseball and Football. Automated analysis has helped in many ways such as improving players’ performance, team performance, estimating value of players and in many other areas. Recent research in machine learning and ease of programming with modern languages has expanded the field of application of machine learning multi folds. Machine learning can be broadly divided into supervised learning, unsupervised learning and reinforced learning. Supervised learning has broad range of applications, that has been applied in this paper. Supervised learning requires past samples preferably in large numbers, encompassing independent variable and response variable so that models can be constructed on independent variables to predict the value of response variable. If response variable is continuous type, a regression model can be constructed where as for categorical values of response variable a classifier model construction model is preferred. There are various approaches of model construction. Decision tree, Artificial Neural Network, Support Vector Machines and Genetic Algorithms based classifier model construction approaches are few of them.

In this paper a Decision tree regression model has been constructed to predict football players’ price based upon past data of FIFA 19. Also Decision tree regression model has been constructed with high accuracy after significant dimensionality reduction in independent variables. In this paper Section 2 covers some of the related work, Methodology adapted is presented in section 3, and section 4 covers details of experiment conducted, results obtained and discussion thereon. Section 5 concludes this paper.

II. RELATED WORK

A lot of money is involved with the transfers of top players in big European Leagues. Estimation of player’s price throughout the year is an important task. Various resources on the web provide data on parameters of the players that may contribute to estimation of player’s price [2]. With mammoth fan following and tremendous popularity of football, teams generate significant revenue for team owners. Team Owners have shaped up as companies with share holders. Choice of players is utmost important to make their teams win. Hence evaluating player’s economic value for transfer is of utmost importance. Constructing an underlying model based upon machine learning is an important area. Players’ characteristics like age, height, position etc., player’s performance (like Goals, passing, fouls etc.) and players’ popularity are important indicators for estimating market value of players [3]. Attributes have been compared on merit to find out which of the attributes are more suitable for the prediction of value of a player. Different machine learning algorithms have been evaluated for their prediction performance [4]. Authors in [5] have attempted to identify a data driven consolidated metric for the evaluation of performance of players. Authors have demonstrated in [2] how market value and performance of the Spanish League players can be estimated using several data sources available on public domain. An artificial neural network based approach has been followed in [1] for pricing of Football players.

III. METHODOLOGY

A csv file of popular benchmarked dataset of players attributes with their price value has been down loaded. This dataset includes latest edition FIFA 2019 players attributes like Value, Wage, Preferred Foot, crossing, Finishing, Heading and many other such attributes relevant for analysis [6]. Upon observation we found 41 relevant attributes which may contribute to evaluate response variable after
deprecating attributes like name of the player, jersey number and other such attributes. Using Python machine learning libraries, a decision tree regressor has been fitted and tested. Further correlation among attributes have been calculated and using heat map and correlation table, highly correlated attributes have been identified. Based upon high correlation around 0.9 value of correlation coefficient, dimensionality reduction has been performed. With reduced dimensionality dataset again a decision tree regression model has been constructed and tested.

IV. USING EXPERIMENT, RESULTS AND DISCUSSION

An interesting dataset on international football players has been found at [6]. After removing non contributing columns in price determination, dataset has been left with 40 independent variables. Response variable ‘value’ in this data set has been named as ‘value’. Attribute names of independent variables are as follows:


On available date set it would be apt to fit a decision tree regression model as target variable is a continuous variable. If target variable or response variable is non-numeric or categorical, a decision tree classification model can be constructed. Scikit-learn is a very popular open source machine learning library with vast documentation available [7]. This has been written in Python which also is an open source. For all our experiments we have opted Scikit-learn. Considering 80% data for training and 20% for testing samples, a decision tree regression model has been fitted with following parameters:

```
DecisionTreeRegressor
criterion='mse', max_depth=5, max_features=None, max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, presort=False, random_state=0, splitter='best')
```

Resulting predictions by fitted regression model produced following statistics:

```
coefficient of determination R^2 of the prediction.:  0.9625131791342137
Mean squared error: 1.94
Test Variance score: 0.94
```

The key parameter ‘coefficient R^2’ is defined as (1 - u/v), where u is the residual sum of squares ((true values – predicted values) and v is the total sum of squares (true values - mean of true values). The best possible score is 1.0 and it can be negative if the fitted model is arbitrarily worse. In our case value of coefficient R^2 is 0.9625131791342137 which is very near to ideal value 1. Hence fitted model can be used with high accuracy to predict price value of a player if independent attribute values are known. Accuracy of the model is also evident from Figure 1 which is a snapshot of our program results.

Further we explored the possibility of dimensionality reduction of independent attributes and fitting regression model with acceptable accuracy. We attempted to reduce its dimensionality by preparing a correction table and a heat map among independent variables. Heat Map provides a visual depiction of correlated variable. Darker the intersection of a block of independent variable along two different axes, higher is the correlation. We prepared a heat map of correlation of independent variables and observed the correlation. For brevity and understanding Figure 2 is presented to depict correlation heatmap of a subset of independent variables.

Table I presents a subset of correlation table showing highest correlation in bold. Since total numbers of independent attributes considered for model construction are 40, hence, correlation table has been of size 40x40 matrix. For paucity of
space only a few of those correlated values are shown here which have highest correlation.

**TABLE I. CORRELATION TABLE**

<table>
<thead>
<tr>
<th>Attribute</th>
<th>BallControl</th>
<th>SprintSpeed</th>
<th>Marking</th>
<th>SlidingTackle</th>
<th>GKHandling</th>
<th>GKKicking</th>
<th>GKPositioning</th>
<th>GKReflexes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Finishing</td>
<td>0.788</td>
<td>0.594</td>
<td>0.024</td>
<td>-0.072</td>
<td>-0.587</td>
<td>-0.583</td>
<td>0.585</td>
<td>-0.587</td>
</tr>
<tr>
<td>ShortPassing</td>
<td>0.911</td>
<td>0.555</td>
<td>0.56</td>
<td>0.509</td>
<td>-0.728</td>
<td>-0.724</td>
<td>-0.724</td>
<td>-0.729</td>
</tr>
<tr>
<td>Curve</td>
<td>0.83</td>
<td>0.579</td>
<td>0.29</td>
<td>0.233</td>
<td>-0.603</td>
<td>-0.6</td>
<td>-0.604</td>
<td>-0.605</td>
</tr>
<tr>
<td>Acceleration</td>
<td>0.676</td>
<td>0.922</td>
<td>0.195</td>
<td>0.158</td>
<td>-0.595</td>
<td>-0.592</td>
<td>-0.592</td>
<td>-0.593</td>
</tr>
<tr>
<td>StandingTackle</td>
<td>0.418</td>
<td>0.178</td>
<td>0.907</td>
<td>0.975</td>
<td>-0.532</td>
<td>-0.531</td>
<td>-0.528</td>
<td>-0.531</td>
</tr>
<tr>
<td>GKDiving</td>
<td>0.788</td>
<td>-0.598</td>
<td>-0.551</td>
<td>-0.509</td>
<td>0.97</td>
<td>0.966</td>
<td>0.97</td>
<td>0.973</td>
</tr>
</tbody>
</table>

In all 17 attributes in top row of Table I (which has subset of those 17 attributes for paucity of space) have high correlation around 0.9 with at least one attribute present in first column of Table 1. It is expected that highly correlated attributes are redundant for the purpose of decision tree regression. Hence out of 40 attributes, 17 attributes have been removed from dataset of players. Attribute names of 17 independent variables being deleted are as follows

Crossing, Volleys, Dribbling, FKAccuracy, LongPassing, BallControl, SprintSpeed, Agility, ShotPower, LongShots, Penalties, Marking, SlidingTackle, GKHandling, GKKicking, GKPositioning, GKReflexes

Now with reduced data set with 23 attributes in place of originally 40 attributes, again considering 80% data for training and 20% for testing samples, a decision tree regression has been fitted with the same parameters as were used in previous experiment on original data set with 40 attributes. Resulting predictions by fitted regression model produced following statistics

- **Coefficient of determination** $R^2$ of the prediction: 0.9613929658176731
- Mean squared error: 1.86
- Test Variance score: 0.93

Thus it is evident that the result are still with almost same accuracy as has been in case of data set used before dimensionality reduction. The key parameter ‘coefficient $R^2$’ value has been declined by (0.9625131791342137 - 0.9613929658176731) = 0.0011202133165406 which is insignificant. So the Regression model constructed still holds good for prediction of player’s price. Fig. 3. shows pictorial depiction of fitted regression model after dimensionality reduction.

V. CONCLUSIONS

Sports analysis is highly applied field of study. Football players’ price valuation is an integral part of international football game. Various parameters of Football players’ abilities are identified and recorded for practical purposes. Many such datasets are made available on public domain. In this paper a decision tree regression model has been successfully implemented with 0.96 value of parameter ‘coefficient $R^2$’ with the proximity of ideal value of 1 to predict the price of the players. Furthermore, with dimensionality reduction of 17 attributes in existing dataset, fitted decision tree still holds good with 0.96 value of parameter ‘coefficient $R^2$’. This paves the way for prediction of price even if any or all of these 17 attribute data values are missing in the database.

References