Predicting Soccer Players' Market Value With Regression Models

Yiyi Zhang
Machine Learning Technologies,
Department of Data Analytics
San Jose State University
SID Last 4 digit: 4852
Email:yiyi.zhang@sjsu.edu

Hieu Tran
Machine Learning Technologies,
Department of Data Analytics
San Jose State University
SID Last 4 digit: 3773
Email: hieu.tran@sjsu.edu

Lu Yang
Machine Learning Technologies,
Department of Data Analytics
San Jose State University
SID Last 4 digit: 7164
Email: Lu.yang@sjsu.edu

Sajit Valiya Kizhakke
Machine Learning Technologies,
Department of Data Analytics
San Jose State University
SID Last 4 digit: 0059
Email: sajit.valiyakizhakke@sjsu.edu

***Abstract*—Transfers of top soccer players typically require a lot of money to be invested, especially in the big leagues. For various reasons, obtaining an excellent assessment of soccer players all year long is valuable, rather than just when the player has recently transferred. Unfortunately, all the player evaluations, including the performance metrics and transfer prices, were done manually for a long time. This paper substitutes manual analysis with regression models for predicting soccer players' market value. We implement a practical application of machine learning in the field of sports analytics. We aim to achieve this by introducing several regression models, including lasso regression, ridge regression, polynomial regression, and random forest regression, which will predict soccer player's market value, also compare the accuracy and performance of these models, and find top features for a club to consider when and more crucially which players should be recommended to be a part of the club. Furthermore, because the market value estimate is based on key performance parameters, it may even help prevent subjective player evaluation and criticism by sports channels, critics, and newspapers.**

***Index Terms*—Soccer Player, market value, lasso regression, ridge regression, polynomial regression, random forest**

I. Introduction

Soccer has remained the most renowned sport for a long time now. But when compared to its contemporaries like Basketball or Motor racing, soccer has yet to gain prominence in sports analytics. Owing to the numerous leagues in existence, complexities which arise from the distribution of countless players within these leagues, and the lack of an efficient approach to perform this has been the primary motivation to pursue this project. Clubs dispatch their scout teams to evaluate players, which is a prolonged and costly process. We intend to assess the proposed methodology's quality and demonstrate that the data-driven evaluation will tackle the issues faced by manual scouting. We also like to explore the relatively fewer applications of machine learning-based studies by deploying the data retrieved from Kaggle and transfermarkt, the website that holds authority in evaluating the player's value, into the regression models.

II. Literature Survey

We explored couples of literature surveys that discuss the soccer player's market value prediction. The following literature are helpful in understanding the features selection and model selection, which our group intended to highlight finding solutions with our approach.

Soccer players have a significant impact on the club. A soccer player transferring to a new club not only increases the club's value but also affects the club's finance and fans. By predicting the market value of a player, the club management can have a good understanding and decide if they want to sign that player. Shuangxian has predicted soccer player value using multiple regression analysis. [1] She has chosen 15 independent variables, such as age, height, nationality, goals, assists, etc., and one dependent variable for the experiment. First, she studied the correlation between variables. Second, she went through the data wrangling process. Then, she applied a mathematical formula to analyze the variables with descriptive statistics. The analysis included count, mean, standard deviation, minimum, maximum, and quartile. The result from multiple regression analysis concluded that "APT1, AG will have a significant positive effect on player" [1] and "Age, CR will have a significant negative effect on player value. However, height, weight, position, APT2, AS, ACS, PG, PS, RC, YC, PR does not have an impact on Player value" [1]. Since the dataset included player data in 2018-2019 Premier League, this could be a reason that the accuracy was not high, only 68.4% [1].

Li spoke that the Key Performance Indicators (KPI's) of players in most sports were examined by sports analysts or even coaches and experts who employ a notional approach [2], i.e., analysis of video footage which was used to compile

a statistical summary of events in addition to the number of goals scored, i.e., in soccer. However, human-based scouting has several drawbacks, including ineffective scaling to a large group of active players, the high-cost factor, the inability to scale to a large number of active players, and the presence of individual biases. The authors implement machine learning based methodologies, chiefly multiple linear regression and decision tree methods to overcome these inefficiencies. They executed machine learning algorithms to recognize meaningful patterns and establish specific structures. Various features of players were evaluated, such as the salary of players and the market value of players. In addition, it will use other features to establish and train the ML model for predicting the reasonable pay for a large number of players. The motivation is to explore the relations between different features of soccer players and their wages - mainly, how each feature influences their salaries or which features are most crucial in determining wages. While there are a variety of standards that can be used to evaluate the value of soccer players, the players' compensation provides one of the most intuitive and crucial indicators, which is why this study uses salary as a proxy to measure the value of players. Moreover, many Players' characteristics can affect their valuation. Still, players' value is primarily determined by their performance, divided into three factors: basic features, court performance, and club achievements.

The value of soccer players affects not only the level of athletic sports but also the decisions managers make. Mustafa performed a comparative study of an objective quantitative method to estimate the market value of soccer players. [3]The project worked with the performance data of soccer players collected from sofifa.com, such as general information, shooting scores, passing scores, and other features showing the player's skills. The authors used four regression models, including linear regression, multiple linear regression, decision trees, and random forests, to predict the soccer players' value. They trained all four regression models and used the result of linear regression as a baseline to evaluate the performance of the models. In addition, the Mean Absolute Errors (MEA), the Mean Square Errors (MSE), and the Root Mean Square errors (RMSE) are used to compare the performance of these models as evaluation metrics. As a result, random forest regression achieved better performance and higher accuracy with the lowest MSE and RMSE. They believed that the model they built was promising for negotiations in the soccer player transfer market, which can provide an essential reference instead of traditional expert estimation.

In a Machine Learning Ensembling Approach to Predicting Transfer Values, Aydemir use in-game performance, popularity, and transfer values to predict future values. [4] The model captures data/features from the real dynamic market, which comes from TransferMarkt, Wyscout, and Google trends. The model considers two ways, log, and root, for long-tailed transfer fee, and distribution to reduce the effect of the outlier to normal prediction. For those just on the longtail data (superstar) prediction, without log or root transformation. And lightGBM as a machine learning model, RMSE (Root Mean Squared Error) as an optimized target for its sensitivity to those data located on the longtail.In this paper, for optimization, they use bayesian hyperparameter + 4-fold cross-validation for log/root/no transformation lightGBM model. Compared with lasso regression and non-regularised regression models, they recommend lasso for this dataset with many features.

III. Methodology

A. Experiment Design

This project uses a dataset from open resouce Kaggle. [5] This dataset contains soccer data from Transfermarkt, a sports information website. The data include games, clubs, players, player market valuations, player appearance records, etc. The dataset we use has six CSV files and a total of 83 columns combined. A GDP dataset from Kaggle, which contains the country's GDP in US dollars from 1960 to 2021, is used as a timeline and to indicate the inflation over the year. We use the Amazon S3 bucket to store the data, and Jupyter Notebook to write, run, and test code using Python.

The ML pipeline is as follows: First, we upload the dataset downloaded from Kaggle on the AWS S3 bucket. Second, we connect the S3 bucket to Jupyter Notebook to ingest and pre-process data. The data wrangling process includes data exploration, finding the correlation between columns and files, extracting important features, joining tables, cleaning data, and aggregation. Third, we split the dataset into training and testing data. Then we build models using the algorithms that we choose with training data and validate the models using the K-folder validation method with testing data. Finally, we evaluate and compare each model.

B. Algorithms

In this project, we will predict soccer players' market value by using four algorithms: Lasso Regression, Ridge Regression, Polynomial Regression, and Random Forest. We will then compare all four models to evaluate which algorithm is the best fit for the prediction.

- Lasso Regression is known as Least Absolute Shrinkage and Selection Operator Regression, a regularized regression model. Its cost function is based on linear regression, which minimizes the sum of squared error and one more term with a penalty factor on the absolute value of the regression coefficient. So, it has a trade-off between minimized error and the number of features.
- Ridge regression is one of the linear regression models derived from the same primary regression equation $y = mx + b$, where y is a dependent variable, x is an independent variable, m is slope or weight, and b is bias error. TThe cost function of ridge regression is $Min(||Y - X(theta)||^2 + ||theta||^2)$, where is a penalty term that shrinks the coefficient of features towards zero but not exactly zero. Ridge regression, also known as L2 regularization, is used to prevent multicollinearity and discourages large weights. The advantage of ridge

regression over linear regression is that ridge regression avoids the overfitting of the model by reducing weights and biases. Ridge regression is useful when we do not want to overfit the training data and want the model to be more accurate on unseen data.

- Polynomial regression is used to address the absence of a linear relationship between the predictor variable and the target variable. Polynomial regression refers to a type of multiple linear regression in which a polynomial equation is fit on the data by obtaining a curvilinear relationship between the dependent and independent variables, i.e., the initial features will be modified into polynomial features of a certain degree, which may be 2,3,...n. Polynomial regression obtains a better accuracy than linear regression as it can evaluate the nonlinear relationship in the data. $y = b0 + b1x1 + b2x12 + b2x13 + ...bnx1n$ where b0 is the bias, b1, b2 ... bn are the weights, and n indicates the degree of the polynomial.
- Random forest is one ensemble method consisting of multiple independent decision trees. Each tree works with different sample data and different features generated by sampling with replacements from the dataset. In addition, each tree gets the classification prediction. The most voting for the classification of test data will result from prediction. As the advantage of random forest, higher accuracy makes the model competitive and powerful since it is collective intelligence. But less interpretation is still the disadvantage of random forest.

C. Evaluation Methods

We will compare the prediction and the market value to get the RMSE (Root Mean Squared Error). We will also consider the MAE (Mean Absolute Error) for there are a lot of outliers in the dataset. In addition, we are going to use the coefficient of determination (R^2) to evaluate the models with the formula. The R-squared score represents the goodness of model fitting. Combining these three, we can know whether categorizing the data would increase or decrease prediction performance and which model performs best for players' value prediction.

D. Technical Difficulties

One of the difficulties that we have encountered is feature selection. There are so many columns/features that we need to consider because choosing fewer features or non-important features could affect the accuracy of the models. Therefore, we plan to use multiple feature selection methods to solve this problem. There are three kinds of ways to select features; filter method, wrapper method, and embedded method. For filter methods, we summarize the operations which will be used on our project, which would be like ranking the feature by correlation, setting the gain, and selecting the features. For instance, we will use a correlation matrix to show the relationship between each column/feature. For wrapper methods, we summarized it as running the model and comparing prediction scores by selecting different features, in this case, we consider it as an optimized method for model optimization, we would consider it after we get the first predicted results of our models. For embedded methods, we need trained models which can rank the features, get the rank of features, and then select the features by rank. Another difficulty we have encountered is encoding the categorical features such as country name, player's position, and player's dominant foot. We have tried to use label encoding, for example, the United States as number one and France as number two, but this seems quite wrong because if we encode like that, it means that we are assuming that the United States is less/smaller than France - one is less/smaller than two. Therefore, to solve this problem, we plan to use One-Hot Encoding which creates dummy variables for categorical features. For example, instead of one feature of a player's dominant foot, by using One-Hot Encoding, it converts into three features; left-foot, right-foot, and both-foot. Given a player with the left foot as dominant, it will be encoded as 1-0-0 where one is left-foot and the other zeros are right-foot and both-foot.

IV. Implementation

A. Data Preparation

Since we collected the dataset from two different resources, there is a total of six tables which are appearances, players, clubs, competitions, player_valuations, uk_gdp_usd needed to be joined. As the first step, for data cleaning, we cleaned all six tables by dropping columns having no impact on the market value based on the literature research and self-knowledge. Moreover, we renamed some columns to clarify the similar meaning of different tables like years, and values. Then, we filled N/A or minimum numbers (1) into missing values instead of directly dropping instances in order to avoid producing biased estimates. We also calculated the mean of the market value for each year and the sum of the players' appearances by grouping players_id. After that, for data transformation, we convert the date_of_birth column into year, convert the market value to million scales, and round 2 decimal place. In addition, we convert categorical variables to 1/0 by using One Hot Encoding. For example, we picked the top 20 countries and set the rest as others, then marked the values for each column. Finally, we combined all preprocessed data into one table through primary keys. The final processed data is saved in 'proccessed_data.csv' with 63975 rows and 67 columns, shown in Fig 1,2.

The CSV file containing processed data is loaded into a dataframe for modeling. We extract the market value column, which is the label, and features (all other columns) from the processed dataframe and convert each of them to a NumPy array. We then standardize the features using StandardScaler from the Scikit-Learn library. Finally, we split data into training and test sets with a ratio of 0.8 and 0.2 respectively.

B. Models

1) Lasso Regression: For Lasso Regression, we use AIC(Akaike information criterion) and BIC (Bayesian Information Criterion) to tune the model with the hyperparameter alpha. Both methods are included in the Scikit-Learn library.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 63975 entries, 0 to 63974
Data columns (total 67 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   height_in_cm                          63975 non-null  int64
 1   position_Attack                       63975 non-null  uint8
 2   position_Defender                     63975 non-null  uint8
 3   position_Goalkeeper                   63975 non-null  uint8
 4   position_Midfield                     63975 non-null  uint8
 5   position_N/A                          63975 non-null  uint8
 6   sub_position_Attacking Midfield       63975 non-null  uint8
 7   sub_position_Central Midfield         63975 non-null  uint8
 8   sub_position_Centre-Back              63975 non-null  uint8
 9   sub_position_Centre-Forward           63975 non-null  uint8
 10  sub_position_Defensive Midfield       63975 non-null  uint8
 11  sub_position_Left Midfield            63975 non-null  uint8
 12  sub_position_Left Winger              63975 non-null  uint8
 13  sub_position_Left-Back                63975 non-null  uint8
 14  sub_position_N/A                      63975 non-null  uint8
 15  sub_position_Right Midfield           63975 non-null  uint8
 16  sub_position_Right Winger             63975 non-null  uint8
 17  sub_position_Right-Back               63975 non-null  uint8
 18  sub_position_Second Striker           63975 non-null  uint8
 19  foot_Both                             63975 non-null  uint8
 20  foot_Left                             63975 non-null  uint8
 21  foot_N/A                              63975 non-null  uint8
 22  foot_Right                            63975 non-null  uint8
 23  domestic_league_code_BE1              63975 non-null  uint8
 24  domestic_league_code_DK1              63975 non-null  uint8
 25  domestic_league_code_ES1              63975 non-null  uint8
 26  domestic_league_code_FR1              63975 non-null  uint8
 27  domestic_league_code_GB1              63975 non-null  uint8
 28  domestic_league_code_GR1              63975 non-null  uint8
 29  domestic_league_code_IT1              63975 non-null  uint8
 30  domestic_league_code_L1               63975 non-null  uint8
 31  domestic_league_code_N/A              63975 non-null  uint8
 32  domestic_league_code_NL1              63975 non-null  uint8
 33  domestic_league_code_PO1              63975 non-null  uint8
 34  domestic_league_code_RU1              63975 non-null  uint8
 35  domestic_league_code_SC1              63975 non-null  uint8
```

Fig. 1. Processed Data

```
 36  domestic_league_code_TR1              63975 non-null  uint8
 37  domestic_league_code_UKR1             63975 non-null  uint8
 38  country_of_birth_Argentina            63975 non-null  uint8
 39  country_of_birth_Belgium              63975 non-null  uint8
 40  country_of_birth_Brazil               63975 non-null  uint8
 41  country_of_birth_Denmark              63975 non-null  uint8
 42  country_of_birth_England              63975 non-null  uint8
 43  country_of_birth_France               63975 non-null  uint8
 44  country_of_birth_Germany              63975 non-null  uint8
 45  country_of_birth_Greece               63975 non-null  uint8
 46  country_of_birth_Italy                63975 non-null  uint8
 47  country_of_birth_Jugoslawien (SFR)    63975 non-null  uint8
 48  country_of_birth_N/A                  63975 non-null  uint8
 49  country_of_birth_Netherlands          63975 non-null  uint8
 50  country_of_birth_Portugal             63975 non-null  uint8
 51  country_of_birth_Russia               63975 non-null  uint8
 52  country_of_birth_Scotland             63975 non-null  uint8
 53  country_of_birth_Spain                63975 non-null  uint8
 54  country_of_birth_Sweden               63975 non-null  uint8
 55  country_of_birth_Turkey               63975 non-null  uint8
 56  country_of_birth_UdSSR                63975 non-null  uint8
 57  country_of_birth_Ukraine              63975 non-null  uint8
 58  club_total_market_value_in_million    63975 non-null  float64
 59  market_value_in_million               63975 non-null  float64
 60  gdp_usd_in_trillion                   63975 non-null  float64
 61  evaluation_age                        63975 non-null  float64
 62  goals                                 63975 non-null  int64
 63  assists                               63975 non-null  int64
 64  minutes_played                        63975 non-null  int64
 65  yellow_cards                          63975 non-null  int64
 66  red_cards                             63975 non-null  int64
dtypes: float64(4), int64(6), uint8(57)
memory usage: 8.8 MB
```

Fig. 2. Processed Data

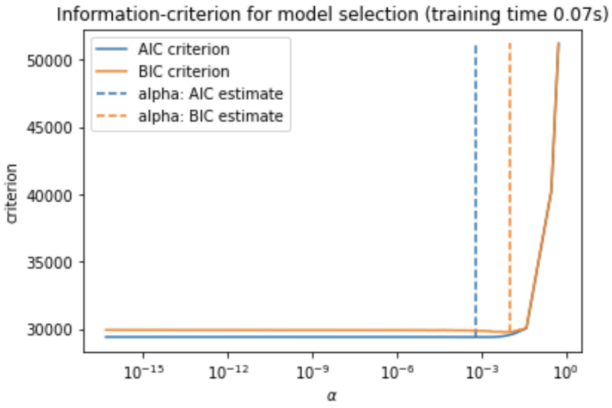


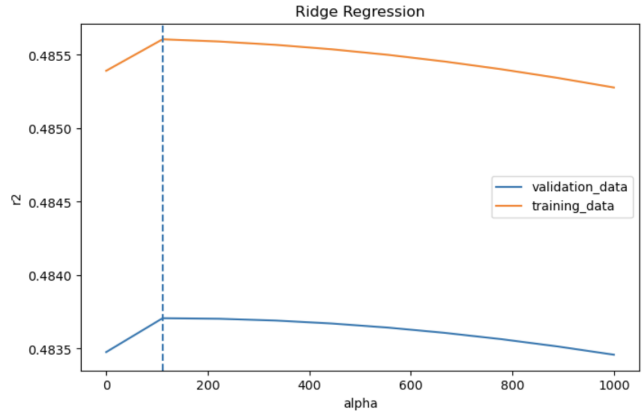
Fig. 3. Lasso Regression

$AIC = 2k2ln(L)$, where k is the feature number and L is maximized value of likelihood function of Lasso Regression. $BIC = ln(n)k–2ln(L)$, where n is the observation number, k and L are the same as AIC.

For the AIC is better on prediction and BIC is better on fitting, we focus more on prediction, so choose the alpha from AIC, get the alpha is 0.0006164712757976496, which has the lowest criterion. And in this condition, the R^2 score grows from 0.3938 with a random select alpha 0.1 to 0.4275 with this best alpha, shown in Fig 3.

2) Ridge Regression: The Ridge regression model is implemented by using the Scikit-Learn library. We tune the model by hyperparameter alpha. We use the k-fold cross-validation method to choose the best alpha parameter which generates the highest r2 score of the validation set. For instance, with three folds and a range of alphas from zero to 1000, the ridge regression model has the highest validation r2 score at alpha 111.111. With this alpha, the model has a training r2 score of 0.48561 and a validation r2 score of 0.48371. Based on the result of validation in Fig 4, we conclude that the mode is underfitting because both the training and validation r2 score are low. Underfitting is when training error and validation error are very low. We plan to solve this underfitting problem by increasing the complexity of the model which decreases the bias and increases the variance. There are several ways to increase the complexity of the model including reducing regularization, and adding more features.

3) Polynomial Regression: We analyze the prediction accuracy of the model for various degrees of polynomials. Hence, a Multivariate Polynomial Regression analysis is implemented. We've first selected the features that have the highest impact on the player's market value and trained the regression model accordingly. We evaluated the model performance by finding out several evaluation metrics like root mean squared, r squared or co-efficient of determination and also the mean absolute error. The best suited polynomial degree for the dataset is




```
Chosen alpha: 111.11111
Training r2 score: 0.48561
Validation r2 score: 0.48371
```

Fig. 4. Finding The Best Alpha Parameter for Ridge Regression Model

plotted against the mean squared error. From Fig 5, we see that model performed best when the degree is 5, meaning a 'quintic' function or the fifth degree polynomial helped us predict the player's market value accurately when compared with the rest.

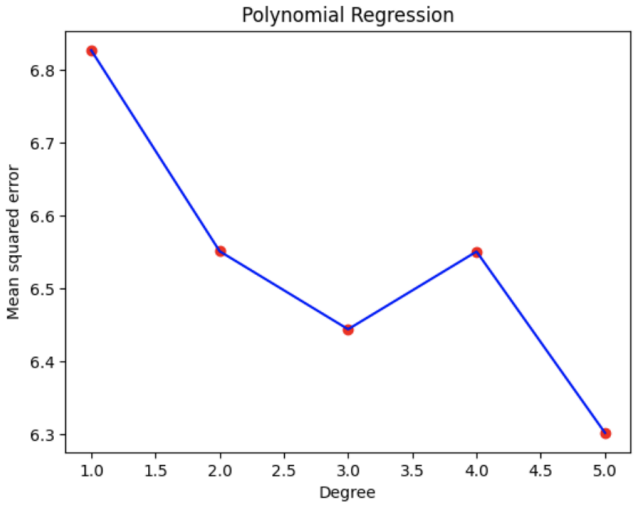


Fig. 5. Polynomial Regression Accuracy

4) Random Forest Regression: Random Forest Regression was implemented by using Scikit-Learn Library. We normalized the features of training data in order to enhance the performance and metrics before starting training. We improved the model by using hyperparameter tuning including the number of decision trees, the number of features at each split, the max depth of each tree, and so on. We set up the grid instead of using the default value provided by the Scikit-Learn package, and randomly searched parameters across 100 different combinations by using 5-Fold cross validation, shown in Fig 6. Finally, we applied the best parameters to the random forest model.

```
rf = RandomForestRegressor()
rf_random = RandomizedSearchCV(estimator = rf,param_distributions = random_grid,
                               n_iter = 100, cv = 5, verbose=2, random_state=35, n_jobs = -1)
rf_random.fit(input_train, output_train)
```

```
Fitting 5 folds for each of 100 candidates, totalling 500 fits
RandomizedSearchCV(cv=5, estimator=RandomForestRegressor(), n_iter=100,
                   n_jobs=-1,
                   param_distributions={'bootstrap': [True, False],
                                        'max_depth': [10, 20, 30, 40, 50, 60,
                                                      70, 80, 90, 100, 110,
                                                      120],
                                        'max_features': ['auto', 'sqrt'],
                                        'min_samples_leaf': [1, 3, 4],
                                        'min_samples_split': [2, 6, 10],
                                        'n_estimators': [5, 20, 50, 100]},
                   random_state=35, verbose=2)
```

&use best parameters

```
print ('Best Parameters: ', rf_random.best_params_, ' \n')
```

```
Best Parameters:  {'n_estimators': 100, 'min_samples_split': 2, 'min_samples_leaf': 1, 'max_features': 'auto', 'max_depth': 90, 'bootstrap': True}
```

Fig. 6. Search Best Paramaters

C. Evaluation Metrics

To evaluate the performance of models, we use multiple metrics from the Scikit-Learn library including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R2).

MAE estimates the mean error size in predictions in spite of directions, with formula:

$$MAE = \frac{1}{n}\sum_{n}^{i=1}|x_i - y_i| \tag{1}$$

where y_i is the prediction value of market value, n is the number of market values observation, x_i is one of the market values for players. [6]

RMSE is the square root of MSE which is the square difference between the truth and prediction output and calculates the mean of the values for each data point. The formula is

$$RMSE = \sqrt{\frac{\sum_{N}^{i=1}(x_i - \hat{x})^2}{N}} \tag{2}$$

where i is a player_id, N is the number of market values observed, x_i is the actual transfer market value, $\hat{x}$ is a prediction value. [7]

The coefficient of determination is related to MSE with formula

$$R^2 = \frac{1 - MSE(model)}{MSE(baseline)} \tag{3}$$

In addition, the mean and standard deviation of coefficient determination i.e. r-squared is also found in polynomial regression along with the standard deviation of mean absolute error.

V. Results

A. Data Analysis

To understand the data inside the datasets used in this project, we perform exploratory data analysis (EDA). It helps us to find hidden patterns, trends, or insights from the data. It also points out outliers, identifies missing values, and checks for duplicates.

The first step of the EDA is to load the data. Datasets are loaded using Pandas dataframe. To have a glance at the

structure of the dataset, we use the function df.info() to show the information of the dataframe including the number of records, number of columns, column name, column type, and memory usage.

Next, we count the number of null values for each column by using the function df.isnull().sum(). This step is very important because it reveals which column is needed to handle missing values. Then, we check for duplicate records using the function df.duplicated().any(). Fortunately, there are no duplicate values in the datasets.

For visualization, seaborn library is used to plot the charts. First, we perform a univariate analysis. We plot the distribution of players' ages at the year of market valuation in the Fig 7. It looks like a normal distribution.

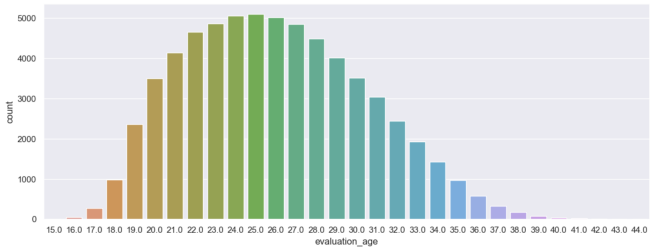


Fig. 7. Player's age distribution

We use a bar chart to show the count of the player's position shown in Fig 8.

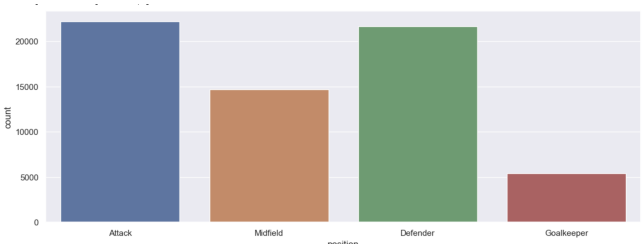


Fig. 8. Player's position distribution

We use a boxplot to show outliers in the market value column shown in Fig 9. The reason there are so many outliers in market value is that there are a few players who are outstanding and evaluated at a higher value.

Next, we perform bivariate analysis using scatter plots. Fig 10 shows the relationship between the club's total market value and the player's market value. Players' market value tends to increase as their current club's total market value increases. In other words, players from popular and wealthy clubs will be more likely evaluated at a higher value. Fig 11 shows the relationship between a player's position and goals. The attack position tends to have more goals than other positions.

To show the relationship between each column, we use a heatmap and correlation matrix shown in Fig 12. Based

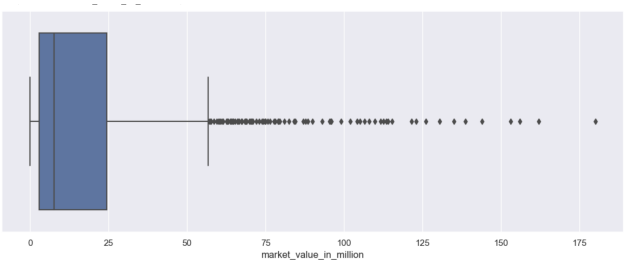


Fig. 9. Market value outliers

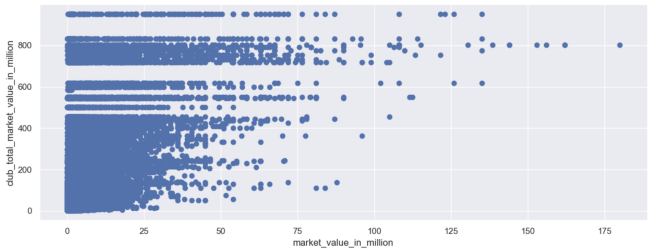


Fig. 10. Bivariate analysis

on the heatmap, we can easily see that goals, assists, and minutes_played have a higher correlation coefficient which indicates that they somewhat relate to the market value which is the label.

B. Compare Models

The Table 1 below shows the mean absolute error, Root mean square error, and the coefficient of determination for all four regression models.

Metric	Lasso	Ridge	Polynomial	RF
MAE	3.0484	2.8327	2.1922	1.8002
RMSE	6.0334	5.3487	5.0058	4.0307
R2	0.4343	0.5041	0.5656	0.7184

TABLE I
COMPARING MODELS BY EVALUATION METRICS

We can see that the random forest gets the best results compared to other models. It demonstrates that Random Forest Regression does a better job of evaluating the market value of the football players. We rank features with both Lasso model and Random Forest, the top 10 features in Lasso model like below Fig 13, And top 10 features in Random Forest model like below Fig 16:

The most important feature is the same, and the latter 3 features are also the same but with different rank.

For the polynomial regression, we have implemented the correlation matrix to understand the top features shown in Fig 14.

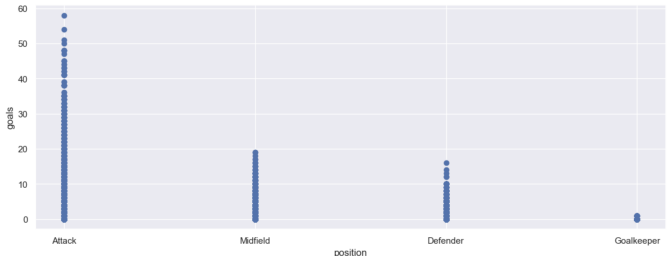


Fig. 11. Player Position and goals comparison

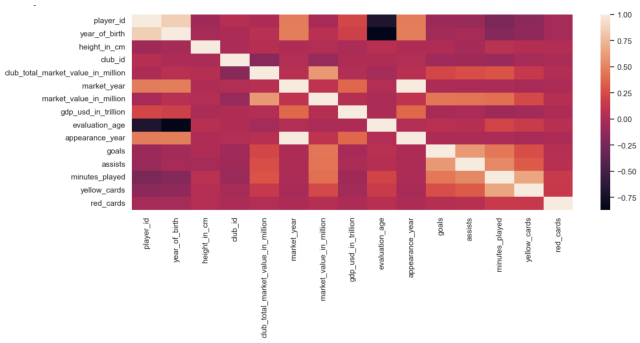


Fig. 12. Correlation heatmap

C. Discussion

We analyzed the market value data, found the dataset has many outliers and we had to mitigate this, shown in Fig 15. We've performed the necessary data preprocessing steps to get a clean dataset.

And we tried to categorize the player by market value, bigger than 5 million as one category and others as another category, the scores getting worse. But after we logged the market value, the score increased by about 40%, and the top rated feature changes from club market value to players minutes played, shown in Table 2.

Metric	Original	Category1	Category2	market val
MAE	3.0484	7.6354	0.6675	0.7427
RMSE	6.0334	11.4389	0.9334	0.9541
R2	0.4343	0.3671	0.3444	0.5907
Top feature	1	1	1	2

TABLE II
LASSO REGRESSION METRIC COMPARISON
1: CLUB_TOTAL_MARKET_VALUE_IN_MILLION
2: MINUTES_PLAYED

Moreover, we calculate the importance of each feature for random forest regression. Fig 16 shows the top 10 importance of predictors according to the input training features and model of random forest. From the chart, we can see that 'club_total_market_value in million', 'minutes_played', and 'goals' determine players' value in the market mostly.

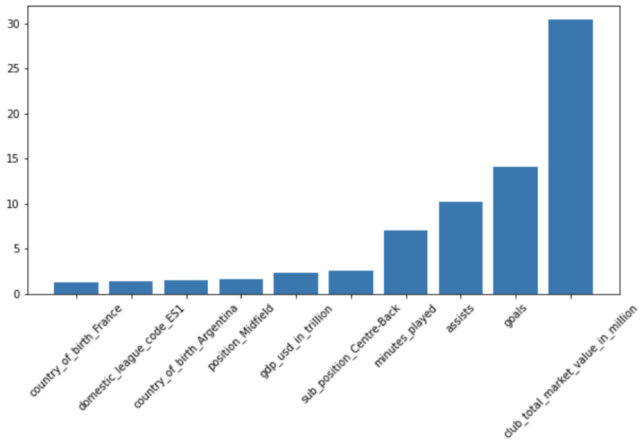


Fig. 13. Top 10 features- Lasso regression mode

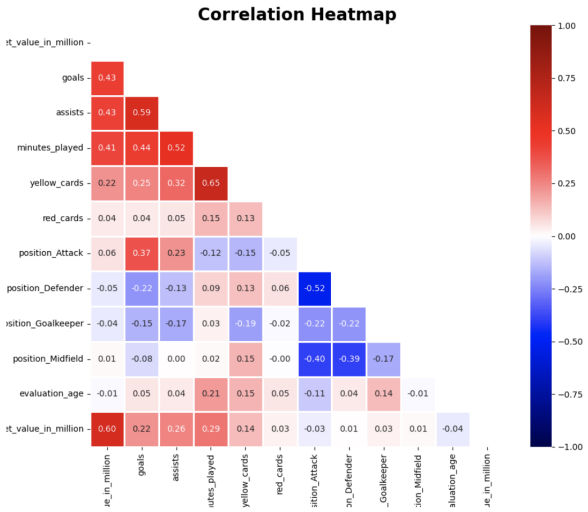


Fig.14 Feature selection-Polynomial regression

Fig. 14. Feature selection-Polynomial regression

VI. Conclustion and Reconmmendations

A. Summary and Conclusions

We have utilized regression analysis to predict the market value of the players. We divided the project workflow into four stages. The first being the ingestion stage wherein we procured a dataset and stored it in the local system for further use. The second stage is data wrangling/exploration stage. In the penultimate phase we prepared the chosen data for regression modeling. And final stage is the modeling stage in which we executed and measured the performance of all the models.

Multiple datasets have been chosen to fulfill our goal of predicting the player's transfer market value. Performed the necessary data wrangling on the cleaned dataset. We then build

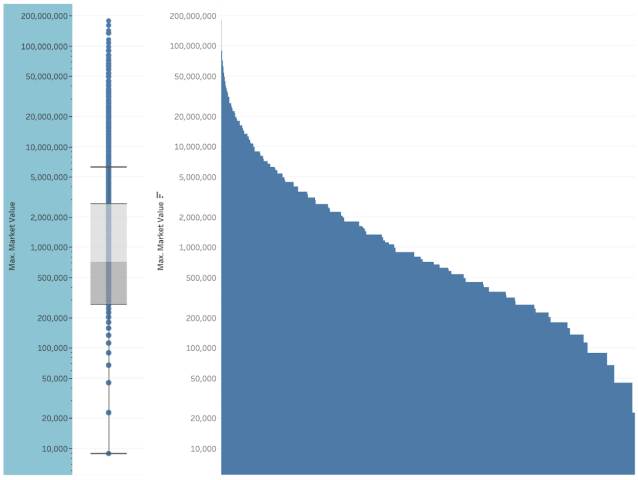


Fig. 15. Outliers evaluation

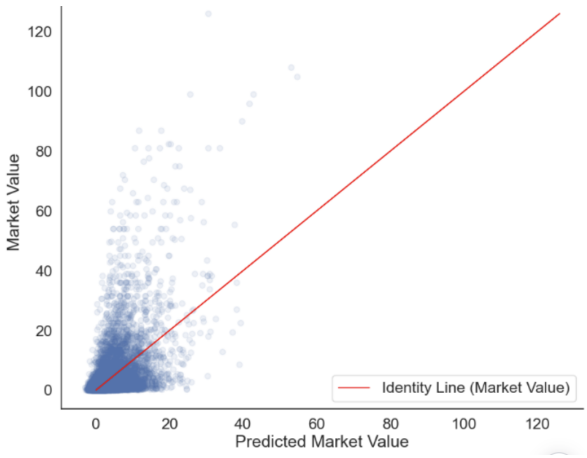


Fig. 17. Comparison of actual v/s predicted market value

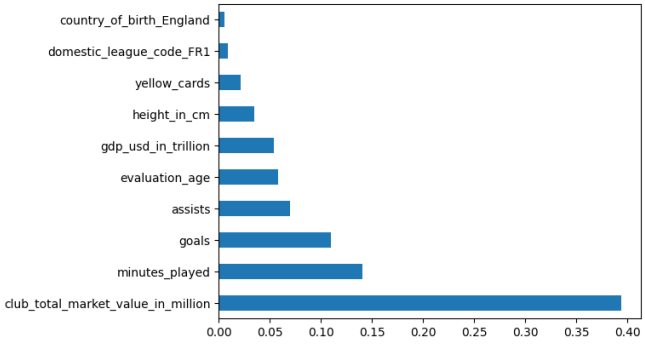


Fig. 16. Top 10 features in random forest regression

the four algorithms to compare and find out the best suited model for the dataset at hand.

We have implemented several novel methods in our project such as one-hot encoding, random selection of hyperparameters for feature selection.

The top 3 important features related to players' market value are 'club_total_market_value in million', 'minutes_played', and 'goals', which means the market value of the club that the player is a part of and the time a player plays in soccer games and the overall goals scored by the player. When a club considers new players, these information could affect the players market value most.

Fig.17 provides a comparison between the predicted market value and the actual market value. We can evaluate the model performance based on this visualization. In the figure the identity line classifies the actual and predicted values.

As most of the data points close to identity line we can conclude that models perform significantly well. Meaning the models predicted the market values accurately with little or no errors. .

B. Recommendations for Future Works

As we mentioned above in chapter 6, we may focus on clustering for soccer players and use random forest regression to do predictions for different clusters. Also, a method to preprocess this kind of distribution of players' market value can be considered in the future.

Employing devices which can handle higher RAM usage for model processing can also be taken into consideration as modelling have encountered memory usage complexities.

With the advent of sports analytics, not just soccer but various other sports can utilize machine learning to analyze player and team performances. With regards to this project, as the data sources keep piling up, several crucial factors such as media coefficient, the real time analysis of on-field statistics can be included in predicting the market value too.

.

REFERENCES

[1] Shuangxian Li. (2020). Multiple regression model for predicting soccer player value in English Primer League. The Frontiers of Society, Science and Technology, Vol. 2 Issue 15:132-143. https://doi.org/10.25236/FSST.2020.021516

[2] Li, C., Kampakis, S., Treleaven, P.C. (2022). Machine Learning Modeling to Evaluate the Value of Football Players. https://doi.org/10.48550/arXiv.2207.11361

[3] M. A. Al-Asadi and S. Tasdemır, "Predict the Value of Football Players Using FIFA Video Game Data and Machine Learning Techniques," in IEEE Access, vol. 10, pp. 22631-22645, 2022, https://doi.org/10.1109/ACCESS.2022.3154767

[4] Aydemir, A.E., Taskaya Temizel, T. Temizel, A. (2022). A Machine Learning Ensembling Approach to Predicting Transfer Values. SN COMPUT. SCI. 3, 201. https://doi.org/10.1007/s42979-022-01095-z

[5] [Online]. Available: https://www.kaggle.com/datasets/davidcariboo/player-scores?select=appearances.csv.

[6] [Online]. Available: https://en.wikipedia.org/wiki/Mean_absolute_error.

[7] [Online]. Available: https://en.wikipedia.org/wiki/Root_mean_square_deviation.

APPENDIX

Term Project Rubric

Criteria	How	Pts
Visualization Includes exploratory analysis (heat maps and other visuals)	Included in the report at chapter 6	5 pts
Presentation Skills Includes time management	Will be included in ppt and presentation	5 pts
Significance to the real world	Included in the report at chapter 7	3 pts
Saving the model for quick demo See https://www.kaggle.com/prmohanty/python-how-to-save-and-load-ml-models Links to an external site.	**Will be included in assignment Term Project Elevator Pitch Video**	5 pts
Code Walkthrough	Will be included in the presentation	5 pts
Report Format, completeness, language, plagiarism, whether turnItIn could process it (no unnecessary screenshots), etc	**Yes, report is generated in Latex in IEEE format.**	7 pts
Version Control Use of Git / GitHub or equivalent; must be publicly accessible	Yes, we used Github to contribute together: https://github.com/hieutransjsu/data245	3 pts
Discussion / Q&A	Included in the project chp 5 and slides	5 pts
Lessons learned Included in the report and presentation?	Yes, including in the report and presentation	5 pts
Prospects of winning competition / publication	Yes, interesting topic and consider to improve for publication	3 pts
Innovation	Yes, try different ways to preprocess data (One Hot Encoding) and try different regression models	5 pts
Evaluation of performance	Included in the report in chapter 6	5 pts
Teamwork	All work done by team members together	5 pts
Technical difficulty	Included in the report in chapter 4	7 pts
Practiced pair programming? See: https://en.wikipedia.org/wiki/Pair_programmingLinksLinks Links to an external site.	Yes, all of us worked together by using Github:	2 pts
to an external site. to an external site. Use GitHub Copilot, if you can and describe the experience using screenshots	https://github.com/hieutransjsu/data245	
Practiced agile / scrum (1-week sprints)? Submit evidence on Canvas - meeting minutes, sprint backlog, and any other artifacts. Use tools such as https://trello.com/en-US/pricing Links to an external site.(Free license available)	Yes, we used Notion as our Agile tool to track the project process shown in the slides	3 pts
Used Grammarly / other tools for language? Grammarly free version is sufficient; can use other tools as well. Submit report screenshot on Canvas.	Spelling and grammar tools from google doc	2 pts
Slides	Included in the assignment	3 pts
Demo	Will be included in the assignment	5 pts
Used LaTeX. Upload .tex file (it should indicate that the IEEE LaTeX template was used and not generated from doc or other format). Using editors such as Lyx is fine. Also checkout https://www.overleaf.com/	Our report is generated by overleaf in LaTex format	2 pts
Used creative presentation techniques animation, effects, newer features such as those offered by prezi, etc	Our presentation document was made by Google Slides with animation and template.	2 pts
Literature Survey 1. Did not miss out on any important existing work that is relevant to the project. 2. Literature survey is organized into meaningful subsections 3. All references are cited and follow standard notation used in the template	Meet this criterion, and included in the report in chapter 3.	7 pts
Total Points: 94		