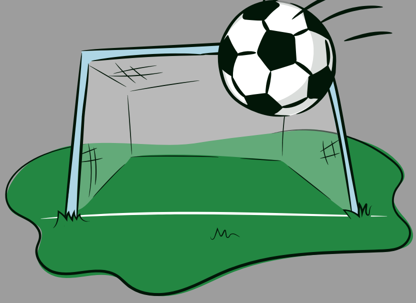
PREPARED FOR DR.SAFA ALSFRI

NATIONAL FOOTBALLER VALUE PREDICTING REPORT

Group memmber

Name	ID
Alaa Yusuf Mohammed	1905458
Raneem Alomari	2006352
Bedoor Ayad	2005961

CONTENTS



Introduction

If you are the CEO of some investing company and you know that the world cup starts in a few months and how important it is for commerce, then, as a wise person, you want to start investing in the field, especially in football players. You may start searching for the most famous player, you may look for their potential, and you may start collecting other criteria and comparing them, but you know for sure you want to maximize your profit, which means you are looking for the player with the highest value. So you ask yourself, "Can we really predict how much money a player will earn?"

we'll walk you through how to train a model for the task of national footballer value prediction with machine learning using Python.

Problem Definition

By predicting player value, an investor company can select the most potential players who are most likely to succeed.

We will be analyzing their acceleration, balance, and other best relative factors that lead us to a true prediction of player value in euros.
This task is important for every investor's agency.

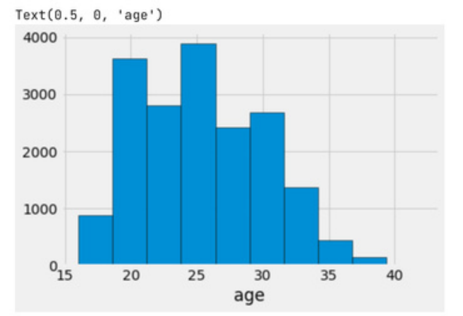
Data Description

players_20.csv is attached with the submission code; it has been scraped from the publicly available website sofifa.com.
We have 106 columns and 391 rows.

- URL of the scraped players
- URL of the uploaded player faces, club, and nation logos
- Player positions, with the role in the club and in the national team
- Player attributes with statistics such as attacking, skills, defense, mentality, GK skills, etc.
- Player personal data like nationality, club, date of birth, wage, salary, etc.

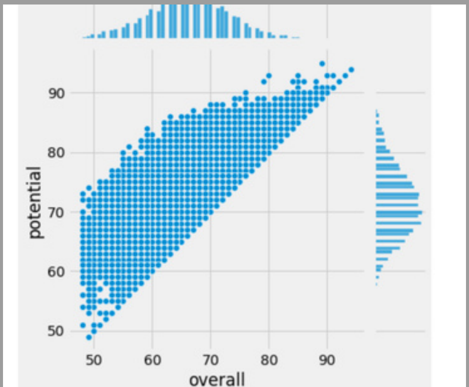
Our target vector y is the player value in euros, while x is the featuers definition.

- histogram of the main features (age)

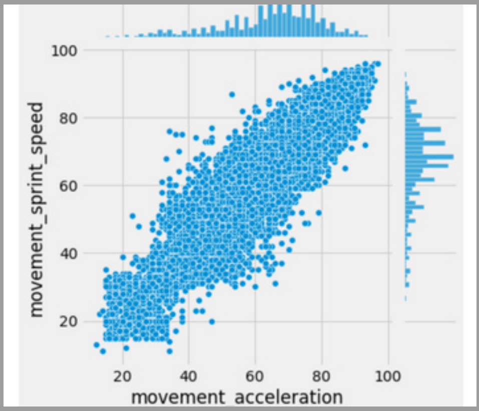


```
Text(0.5, 0, 'age')
```

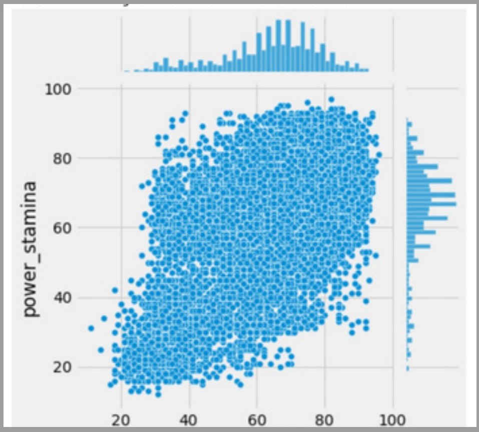
- Some unusual behavior in Potential vs Overall



- Accelaration and SprintSpeed follow a proper linear relationship



- Agility vs Stamina have linear relationship



Method

For Modeling we use:

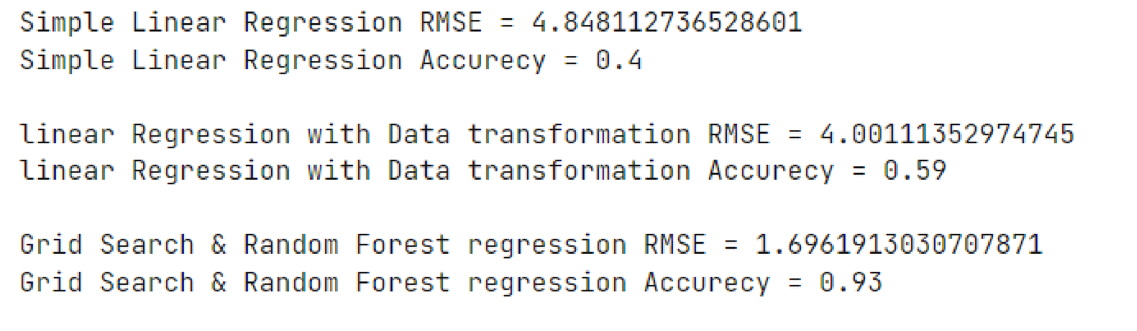
- Simple Linear Regression.
- linear Regression with Data transformation.
- Grid Search & Random Forest regression technique.

other libraries

numpy, Pandas, matplotlib.pyplot, seaborn

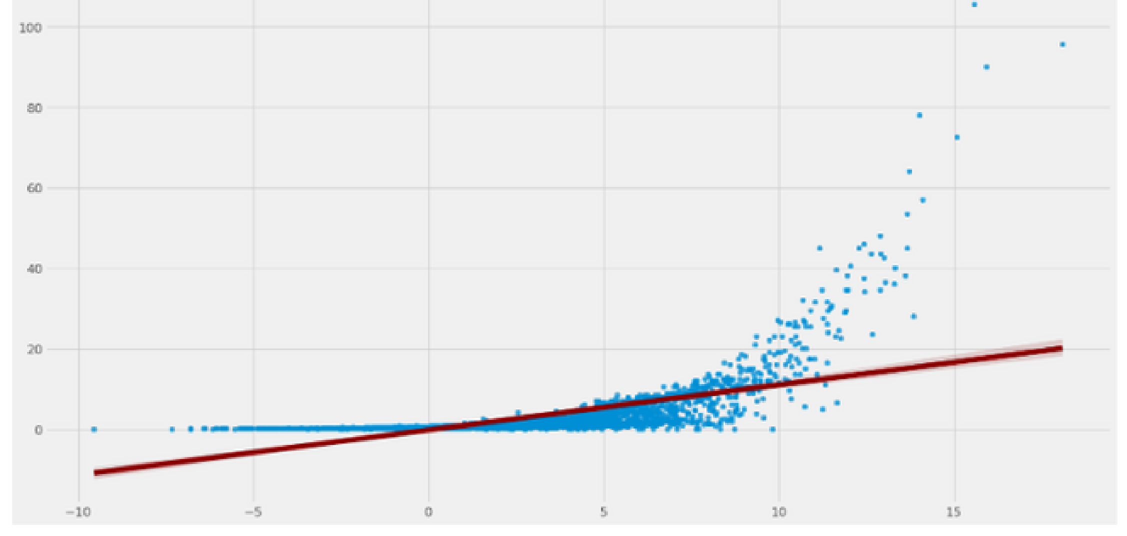
Experemental Results

We managed to decrease RMSE from 4.8 to 4 at first, then ended up with 1.6, and increase Accurecy from 0.4 to 0.93 using the models mentioned above.

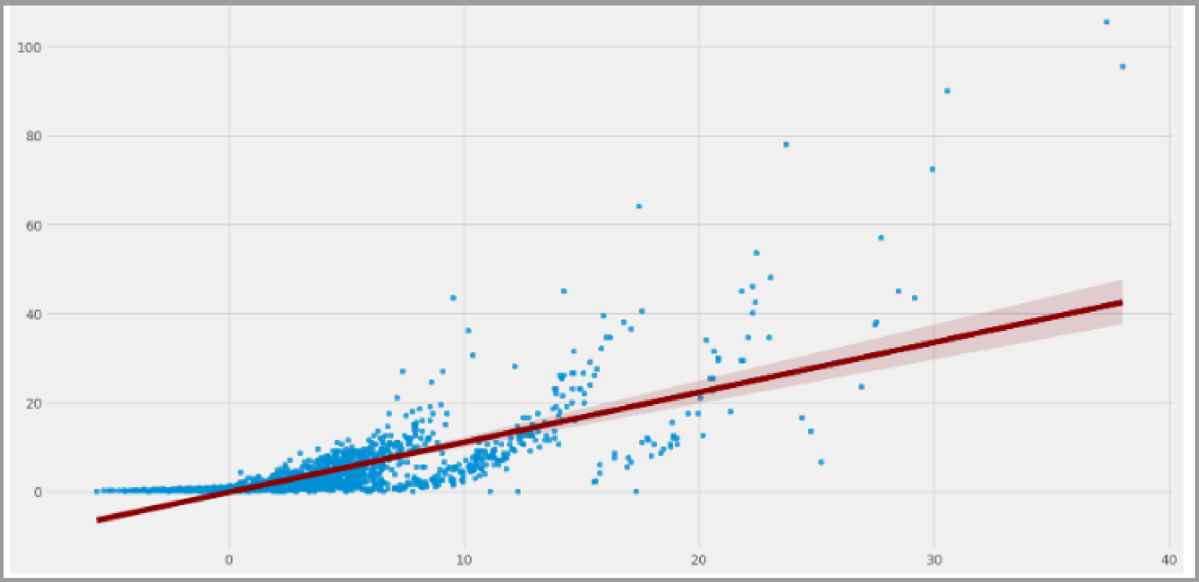


```
Simple Linear Regression RMSE = 4.848112736528601
Simple Linear Regression Accurecy = 0.4

linear Regression with Data transformation RMSE = 4.00111352974745
linear Regression with Data transformation Accurecy = 0.59

Grid Search & Random Forest regression RMSE = 1.6961913030707871
Grid Search & Random Forest regression Accurecy = 0.93
```

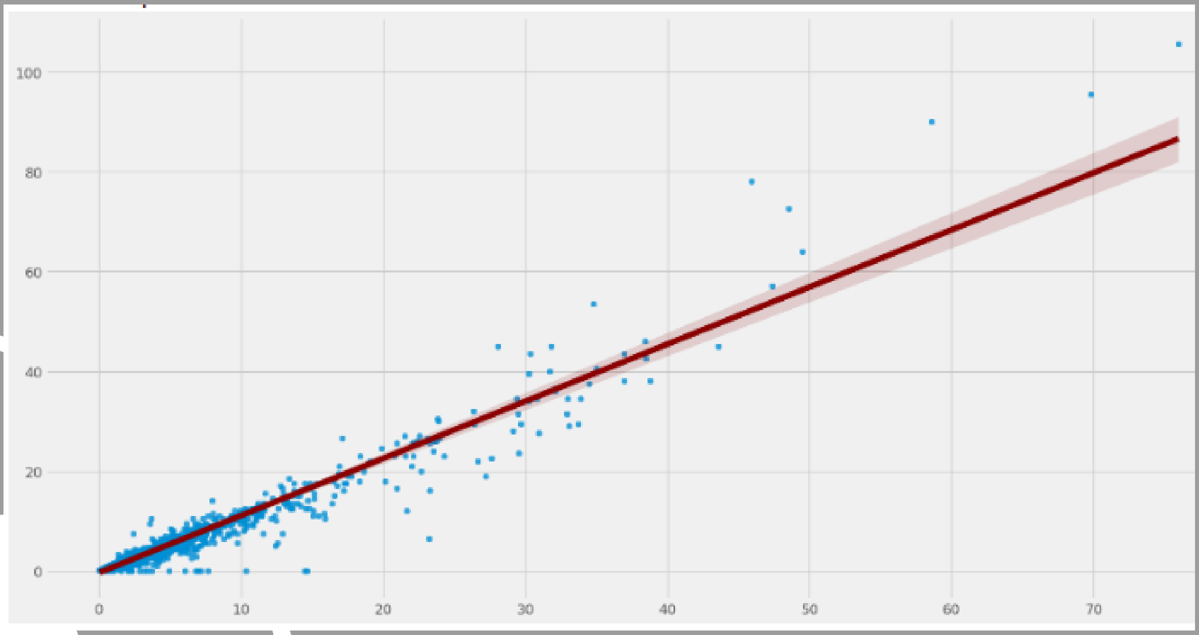
- Plot the fitted Simple Linear Regression Model



- Plot the fitted Linear Regression with Data transformation Model



- Plot the fitted Grid Search & Random Forest regression Model



Discution

Firstly, We apply simple Linear Regression on selected features to find out their effect on the target, which is players' current market values. but we gain disapointed result upon that we decided to

- first, we remove goalkeepers cause they have statics vary from other players which effect the model accuracy.

- secondarily, we abstract 42 relative features from 106 instead of 7 as before.

- thirdly, we will go through some data transformation functions that add a lot to improve our model.

after that we managed to decreas the previous rmse from 4.8 and to become 4.0 that is a good thing but not good enough.
Third phase for improvment we apply Grid Search with Random Forest Regressor to reach the best results, grid search finds the optimal combination of one or more hyperparameters that gives the most optimal model complying with a bias-variance tradeoff, in other words finding the optimal set of parameters by evaluation of all possible parameter combinations. inaddition to Random Forest witch has multiple decision trees as a base for learning models. we approach the Random Forest regression technique like any other machine learning technique.

Conclusion

In conclusion, we used a machine-learning-based prediction model with excellent performance to predict player value. We went through linear regression with data transformation outperformed simple linear regression, but it performed poorly compared to grid search with Random Forest Regression Technique, which had the fewest errors and better predicting ability with 0.93 accuracy.

Refrence

- sklearn.pipeline.Pipeline — scikit-learn 1.1.3 documentation
- sklearn.preprocessing.StandardScaler — scikit-learn 1.1.3 documentation
- Sklearn SimpleImputer Example - Impute Missing Data - Data Analytics (vitalflux.com)
- Grid Search Explained - Python Sklearn Examples - Data Analytics (vitalflux.com)
- Random Forest Regression in Python - GeeksforGeeks

NATIONAL FOOTBALLER VALUE PREDICTING REPORT