**Preprocessing Techniques**

* **Missing Values:** #'club\_team''club\_position','club\_join\_date','contract\_end\_year','tags', 'traits''LS','ST','RS',

'LW','LF','CF','RF','RW','LAM','CAM','RAM','LM','LCM','CM','RCM','RM','LWB','LDM','CDM','RDM','RWB','LB','LCB','CB','RCB','RB': got the mean/median of the col and refill the null values with the mean.

1. **Mean values**: 'club\_rating','club\_jersey\_number'

**Implemented:** using data[i].fillna(data[i].mean())

1. **Median values**: 'LS','ST','RS','LW','LF','CF','RF','RW','LAM','CAM','RAM','LM','LCM','CM','RCM','RM','LWB','LDM','CDM','RDM','RWB','LB','LCB','CB','RCB','RB' , 'club\_join\_date','contract\_end\_year'

**Implemented:** using data[i].replace(0,data[i].median())

* **Transformation**

1. **Feature encoder**: the process of turning categorical data in a dataset into numerical data

Col: nationality, club team, traits

**Implemented :**using function Feature\_Encoder(data, column list)

1. **One hot encoding**: the essential process of converting the categorical data variables to be provided to machine and deep learning algorithms which in turn improve predictions as well as classification accuracy of a model.

Col: 'preferred\_foot','work\_rate','body\_type','club\_position','pos1', 'pos2', 'pos3','pos4','tags'

**Implemented :**using function encode\_and\_bind(data, column list)

**Feature scaling:** method used to normalize the range of independent variables or features of data

Col: overall rating, potential, wage, international reputation , release clause euro , club rating , reactions

**Implemented :**using function featureScaling(X ,value1,value2)

* **Outliers :**

Wage->1592

release\_clause\_euro -> 1847

value->1917

Position col -> split to 4 cols

**Implemented :**using function find\_outliers\_IQR(data[i])

* **Splitting:**

Position col -> split to 4 cols (pos1,pos2,pos3,pos4)

**Implemented :**using data['positions'].str.split(',', expand=True)

* **Sum**

'LS','ST','RS','LW','LF','CF','RF','RW','LAM','CAM','RAM','LM','LCM','CM','RCM','RM','LWB','LDM','CDM','RDM','RWB','LB','LCB','CB','RCB','RB' -> Sum the value in the cell

**Implemented :**using for loop to split each cell two number and sum

* **Convert**

contract\_end\_year, club\_join\_date -> to year

**Implemented :**using for loop to extract year from date

**Analysis**

Here we analyze the coloration of the Value col with the top features(above 0.5)

* Overall rating:0.62
* Potential: 0.57
* Wage : 0.86
* international reputation:0.64
* Release clause euro: 0.97
* Club rating: 0.53
* Reactions: 0.54

**Regression Techniques**

**Multi regression**: statistical technique that can be used to analyze the relationship between a single dependent variable and several independent variables

**Polynomial**: form of Linear regression known as a special case of Multiple linear regression which estimates the relationship as an nth degree polynomial

**Some Differences**

|  |  |  |
| --- | --- | --- |
|  | Model 1 (multi) | Model 2(poly) |
| Accuracy (Test) | 0.9656857017996736 | 0.9829983298409519 |
| Error | 1284228335241.1196 | 391793094071.735 |
| Training Time | 0.008407354354858398 | 0.08975076675415039 |

**Features**

* **Features used:** overall rating, potential, wage, international reputation , release clause euro , club rating , reactions (all rows)
* **Dropped cols**: 'id', 'name','full\_name','birth\_date','national\_team',

'national\_rating','national\_team\_position','national\_jersey\_number','positions'

**Sizes**

Training size: model 1->70 , model 2-> 70

Test size: model 1->30 , model 2-> 30

**Screenshots**

**Chart

Description automatically generated with low confidence**

**Conclusion**

We handled the data by a lot of preprocessing steps (filling the missing values with their median/ mean and detecting the cols of the outliers) and then used two regression techniques to determine the value of different players with respect to his different attributes that affects his value.