**Pattern Recognition Project**

**(Player-value-prediction)**

**Computer Science department**

**Milestone one report**

**Phase 1: Preprocessing**

* First the process from id column to preferred foot column:
* By checking [id , name , full name , birth date, age, height, weight ,positions ,nationality, overall rating, potential and preferred foot] don’t have null values
* Wage has null value so change it with the column mean value
* Give values from 1 to 5 starting from GK give him 1 then all the defense line 2 , any midfield then 3 and any attack by 4 as the price of attackers is always high
* Also preferred foot if it is right then give it 1 else then 2 as the left foot footballers are high priced
* Second the process from international reputation to national team jersey number:
* International reputation, weak foot, skill moves don’t need preprocess as the are numeric
* Used label encoder at work rate and body type as they don’t have null values
* Release clause by put it zero as it means there is no release clause
* Club team by writing no team then use label encoder
* Club rating by writing zero
* Club position by taking first position in player position in position column and use label encoder
* Club jersey number by set equal to zero
* Club join date and end year if one of them is missing by set by making calculations to present year
* National team take value from nationality
* National rating from the overall rating
* National team position from club position
* National team jersey by set it to zero
* Third from crossing column to RB column expect Tags and Traits columns:
* By putting mean values in all fields
* For tags use label encoder if any null values found write unknown
* For traits give it zero if no trait found or take away 20 from trait if ‘injury prone found’ other wise

**Phase 2: feature selection**

['id','name','full\_name','birth\_date','nationality','wage','tags', 'release\_clause\_euro','club\_jersey\_number','club\_join\_date',

'contract\_end\_year', 'national\_team','national\_team\_position',

'national\_jersey\_number']

* These features were dropped as the don’t affect in the price of a player and are extra information
* Other features we dropped according to the correlation

Calendar

Description automatically generated with medium confidence

**Phase 3 : modeling**

**Using Polynomial regression and multivariate regression**

* Used the (80/20) split 80 %for training and 20% for test
* First model By using polynomial regression at degree = 3 :
* Mean Square Error 560301278647.9445
* Accuracy: 0.978569051178942
* Training time: 0.3993546962738037s
* true player value: 475000.0
* predicted player value : 431177.3195286151
* Second model by using Multivariate regression:
* Mean Square Error 560301278647.9445
* Accuracy: 0.978569051178942
* Training time: 0.0019943714141845703s
* true player value: 475000.0
* predicted player value : 431177.3195286151
* Conclusion:
* There is no difference between multivariate or polynomial as the get same result
* Big difference between features and Value is high result to high value of mean square error