* Preprocessing

1. Replace the nan values in the data with zeros
2. Drop columns that almost have null values (national\_team, national\_rating, national\_team\_position, national\_jersey\_number , tags)
3. Split (work\_rate ) column in two columns and convert them into numeric values using label encoder
4. Split (positions ) column in 4 columns and convert them into numeric values using label encoder
5. Convert all columns that have string values to numeric values using feature encoder
6. For columns that have (+2) in its value , split them into one column with the value before (+) , then fill nans with zeros , then convert their type to integer
7. feature scaling for the top features after selecting them

* Data analysis

using the correlation to know the dependency between the features

we use the features that the Y depend on it with 0.3 correlation , and remove the features that other features depend on it

* Models

1. Multiple linear regression model
2. Multiple polynomial regression model

* Difference between models

|  |  |  |
| --- | --- | --- |
|  | Multiple linear model | Multiple polynomial model |
| Train error | 1661452090219.2021 | 299535546651.7732 |
| Test error | 1680895119279.0374 | 849824167000.3585 |
| Accuracy | 0.9613153652624946 | 0.9804418865195977 |
| Train time | 0.004988908767700195 s | 0.16513681411743164 s |

* Feature selection

We use the features that the value column depends on it with more than 0.3 correlation (overall\_rating, potential, wage, international\_reputation(1-5), skill\_moves(1-5), release\_clause\_euro, club\_rating, short\_passing, long\_passing, ball\_control, reactions, vision, composure) , and discard (ball\_control , long\_passing ) because other features depend on them

* Size of data

80% train , 20% test

* Conclusion

The multi polynomial model gives error less than multi linear model and accuracy greater than