Course Work-1 Report

Predicting the Price of a Football Player

A Course Work Report Submitted

in Partial Fulfillment of the Requirements

for the Degree of

Bachelor of Technology

In

Computer Science and Engineering Department

By

Palak Singh
190138

CSE-I
3rd year

BML MUNJAL UNIVERSITY™

SCHOOL OF ENGINEERING AND TECHNOLOGY

BML MUNJAL UNIVERSITY GURGAON

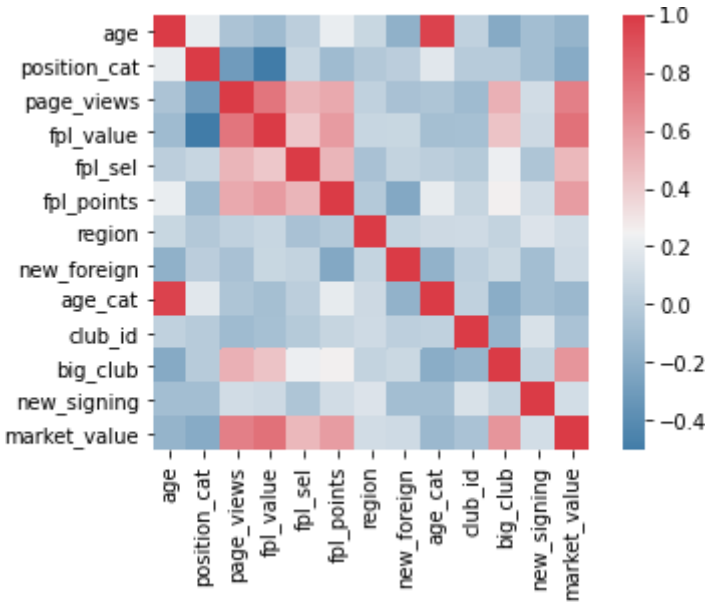
December,2021

1. Pearson's Correlation between each variable and output variable:

position_cat	-0.202518
age	-0.144592
age_cat	-0.116853
club_id	-0.052287
new_foreign	0.097173
region	0.110158
new_signing	0.115376
fpl_sel	0.484164
fpl_points	0.595919
big_club	0.624354
page_views	0.716096
fpl_value	0.771985
market_value	1.000000

The Correlation between each variable in data set with market value is found. The Correlation value indicates the dependence of the variables on each other. The correlation value lies between 0 and 1 (-1 to 1 if the variables are dependent on each other inversely). Higher the value of correlation, the higher dependency and the variables can be merged or removed according to the value of correlation. The Heat map below shows the interdependency of each variable on other. If the intensity of color is more, then it can be assumed that the variables are more correlated. Dark red represents higher dependency on each other positively and dark blue represents higher dependency on each other negatively.

Heat Map showing correlation among all the variables:



2. Comparison of Different Regression Models:

Linear Regression

MSE score: 2253.1826765864375
R2 score: -11.382032075377651

Lasso Regression

MSE score: 44.635748023994644
R2 score: -9.948664696519803

Ridge Regression

MSE score: 47.42696824958404
R2 score: -11.360790565498204

Kth Nearest Neighbour Regression

RMSE value for k= 20 is: 8.774911812480529

SVR Regression

MSE score: 10.47592874913882
R2 score: 0.39691213061159725

Tree Regression:

MSE score: 9.841880295889636
R2 score: 0.46770586839901285

Random Forest

MSE score: 6.158698043109998
R2 score: 0.7915637053996277

Gradient Boost Regression

MSE score: 6.180443705944269
R2 score: 0.790089177057778

3. Tune the hyperparameters and build the most accurate model

Linear Regression

Best Score: -4.319281419264125
Best Hyperparameters: {'copy_X': True, 'fit_intercept': True, 'normalize': True}
LinearRegression GridSearch Accuracy: -11.382032075377651
RMSE score: 54.74359131021873
R2 score: -15.468805129349892

Lasso Regression

Best Score: -4.240566475479193
Best Hyperparameters: {'alpha': 0.01, 'copy_X': False, 'fit_intercept': True, 'max_iter': 10, 'normalize': True, 'precompute': True, 'selection': 'random', 'warm_start': True}

Lasso GridSearch Accuracy: -10.808741351789669
RMSE score: 44.991174613855385
R2 score: -10.123723485983499

Ridge Regression

Best Score: -4.227049326675834
Best Hyperparameters: {'alpha': 0.1, 'copy_X': True, 'fit_intercept': True, 'max_iter': 10, 'normalize': True, 'solver': 'sag'}
Ridge GridSearch Accuracy: -16.53391516942623
RMSE score: 7.3203682790345175
R2 score: 0.7055162179805238

Kth Nearest Neighbour Regression

Best Score: 0.5200726289993008
Best Hyperparameters: {'leaf_size': 1, 'n_neighbors': 24, 'p': 1}
Knn GridSearch Score: 0.5680141582840241
RMSE value : 8.622021473671152
R2 score: 0.5914796253878369

SVR Regression

Best Score: -6.099489733412793
Best Hyperparameters: {'gamma': 0.0001, 'kernel': 'rbf', 'max_iter': 100}
SVR GridSearch Accuracy: 0.2185178234939857
MSE score: 10.47592874913882
R2 score: 0.39691213061159725

Decision Trees

Best Score: -4.537185920582436
Best Hyperparameters: {'criterion': 'friedman_mse', 'max_depth': 5, 'max_features': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'splitter': 'best'}
RMSE score: 7.11986429441369
R2 score: 0.7214270436370289

Random Forest

Best Score: 0.7774042844927579
Best Hyperparameters: {'n_estimators': 1600, 'min_samples_split': 2, 'min_samples_leaf': 1, 'max_features': 'log2', 'max_depth': 1000}
Random Forest GridSearch Accuracy: 0.7597747041311681
RMSE score: 6.679054568133203
R2 score: 0.75485360871305

Gradient Boost Regression

Best Score: -3.45074930135845

Best Hyperparameters: {'subsample': 1, 'n_estimators': 500, 'max_features': 'log2', 'max_depth': 4, 'loss': 'lad', 'alpha': 0.1}

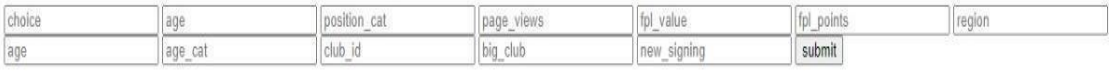
RMSE score: 6.906251422901736
R2 score: 0.7378919913863964

	Before Tuning(R2)	After Tuning(R2)	Best Scores
Linear regression	-11.382032075377651	-15.468805129349892	-4.319281419264125
Lasso Regression	-9.948664696519803	-10.123723485983499	-4.240566475479193
Ridge Regression	-11.360790565498204	0.7055162179805238	-4.227049326675834
Near Neighbor Regression	0.685247698439583	0.5914796253878369	0.5200726289993008
Support vector Machine	0.39691213061159725	0.39691213061159725	-6.099489733412793
Decision Trees	0.46770586839901285	0.7214270436370289	-4.537185920582436
Random Forest	0.7915637053996277	0.75485360871305	0.7774042844927579
Gradient Boost Regression	0.790089177057778	0.7378919913863964	-3.45074930135845

From all the above models we can observe that the **SUPPORT VECTOR REGRESSION,** reports us the best results as the R2 value of the SVR is lowest among all other regressions. The Best Score of SVR is -6.099489733412793, Best Hyperparameters: {'gamma': 0.0001, 'kernel': 'rbf', 'max_iter': 100}, Accuracy: 0.2185178234939857, MSE score: 10.47592874913882 and R2 score: 0.39691213061159725

So, Support vector Machine is the best algorithm among all others in this scenario. So, we deployed Support Vector Machine Regression in Restful API.

4. Model Deployment as RESTFUL API service



Attaching ipynb file containing all the tests and api model.