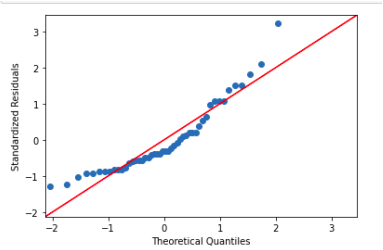
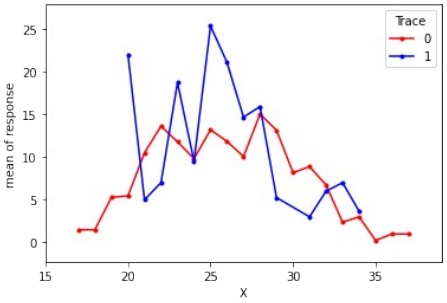
Use Seaborn to investigate the data and present your findings



Analysis of variance is used to compare the means of more than 2 groups. In our dataset we found the p value to be 0.007 ($p < 0.05$) therefore, this suggests that there is significant difference in the market values based on region.

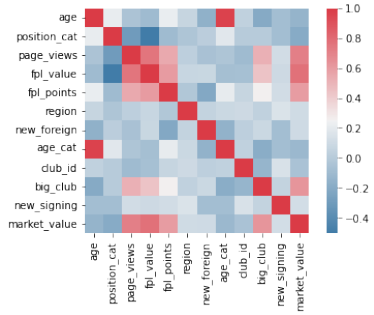


The following graph shows that players from ages 22 to 30 tend to have the highest market values. (given that age_cat is an encoding for all ages).

```
age                 -0.144592
position_cat        -0.202518
page_views           0.716096
fpl_value            0.771985
fpl_points           0.595919
region               0.104927
new_foreign          0.097173
age_cat             -0.116853
club_id             -0.052287
big_club             0.624354
new_signing          0.115376
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:



Comparison of Different Regression Models

Linear Regression

MSE score: 5.6799098501072445

R2 score: 0.6852867918063055

Lasso Regression

MSE score: 7.1098557199162284

R2 score: -0.5068788157152118

Ridge Regression

MSE score: 5.632324409323887

R2 score: 0.6905379435856579

R Nearest Neighbour Regression

MSE score: 5.541150406376546

R2 score: 0.7004757689754371

SVR Regression

MSE score: 6.0510688117404285

R2 score: 0.6428124868817404

Tree Regression

MSE score: 6.880975663883468

R2 score: 0.5381168400088514

Random Forest

MSE score: 5.178360912405627

R2 score: 0.7384126631998229

Gradient Boost Regression

MSE score: 4.9483375849729905

R2 score: 0.7611359864236373

Tune the hyperparameters and build the most accurate model

Linear Regression

Best Score: -8.32833200962021

Best Hyperparameters: {'positive': True, 'normalize': True, 'n_jobs': 100, 'fit_intercept': True, 'copy_X': True}

RMSE score: 9.88198067761516

R2 score: -0.051431689912723844

Lasso Regression

Best Score: -8.170894515743631

Best Hyperparameters: {'warm_start': False, 'selection': 'cyclic', 'precompute': False, 'normalize': True, 'max_iter': 10, 'fit_intercept': True, 'copy_X': True, 'alpha': 0.1}

RMSE score: 9.88538268570681

R2 score: -0.05215575423662866

Ridge Regression

Best Score: -8.16765189597041

Best Hyperparameters: {'solver': 'lsqr', 'normalize': False, 'max_iter': 100, 'fit_intercept': True, 'copy_X': True, 'alpha': 0.001}

RMSE score: 9.870075322244723

R2 score: -0.04889978300675679

R Nearest Neighbour Regression

Best Score: -8.355698924731183

Best Hyperparameters: {'weights': 'uniform', 'n_neighbors': 15, 'algorithm': 'brute'}

RMSE score: 9.389403010249847

R2 score:0.050775318711754225

SVR Regression

Best Score: -8.136444659829587

Best Hyperparameters: {'max_iter': 50, 'kernel': 'rbf', 'gamma': 0.001}

RMSE score: 11.116250655819774

R2 score: -0.3304842063010145

Tree Regression

Best Score: -8.170894515743631

Best Hyperparameters: {'splitter': 'best', 'min_samples_split': 2, 'min_samples_leaf': 0.5, 'max_features': 'sqrt', 'max_depth': 1, 'criterion': 'mse'}

RMSE score: 10.004669877354209

R2 score: -0.07770174828763499

Random Forest

Best Score: -8.683754286229043

Best Hyperparameters: {'warm_start': False, 'n_estimators': 200, 'min_samples_split': 4, 'min_samples_leaf': 4, 'bootstrap': True}

RMSE score: 9.603221977316444

R2 score: 0.007050876093161795

Gradient Boost Regression

Best Score: -3.324361173434046

Best Hyperparameters: {'n_estimators': 500, 'max_features': 'sqrt', 'max_depth': 4, 'loss': 'ls', 'alpha': 0.99}

RMSE score: 4.7247762037698795

R2 score: 0.782231745907084

From the above R2 values of all the models we can conclude that "GradientBoostingRegressor" gives us the most accurate model.The Best Score of "GradientBoostingRegressor" is -3.324361173434046, Best Hyperparameters: {'n_estimators': 500, 'max_features': 'sqrt', 'max_depth': 4, 'loss': 'ls', 'alpha': 0.99}, MSE score: 4.7247762037698795 and R2 score: 0.782231745907084

Implement a Genetic Algorithm for learning attribute weights for the Nearest Neighbour Algorithm. Implement at least one mechanism for maintaining Diversity within the Population

Here the "child" is produced using the methods like crossover and mutation, wherein the child is known to share characterstics with its parents.These algorithms help maintain the diversity in the population.In our code we implmented both crossover and mutation.

Here the score we obtained was : 45% with

Deploy your model as a RESTful Web Service

We deployed the RESTful Web Service and took care of all the extra encoded values to make the user experience better.

Here is our website: