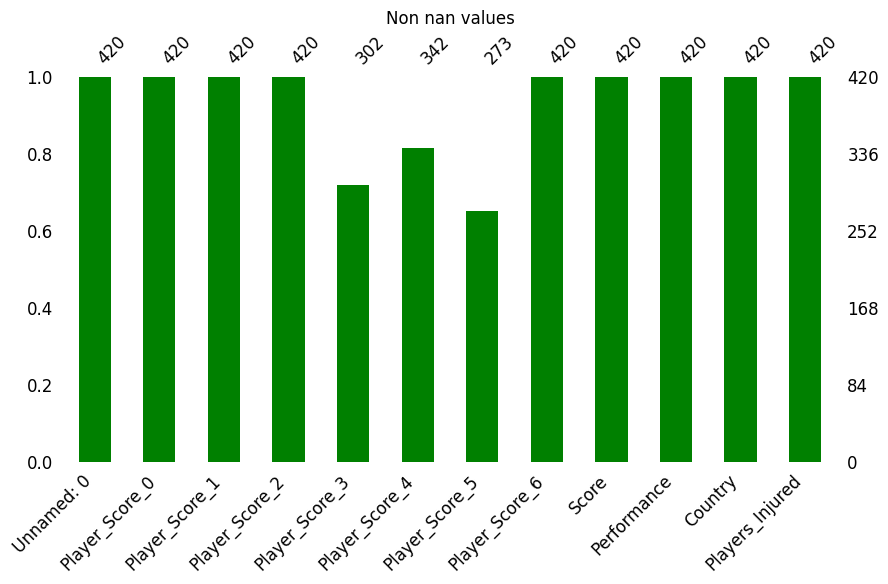
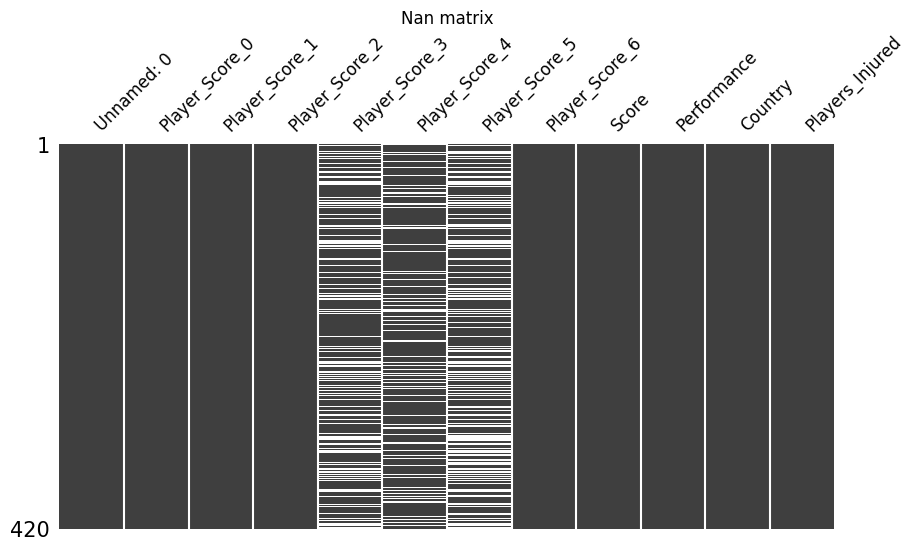
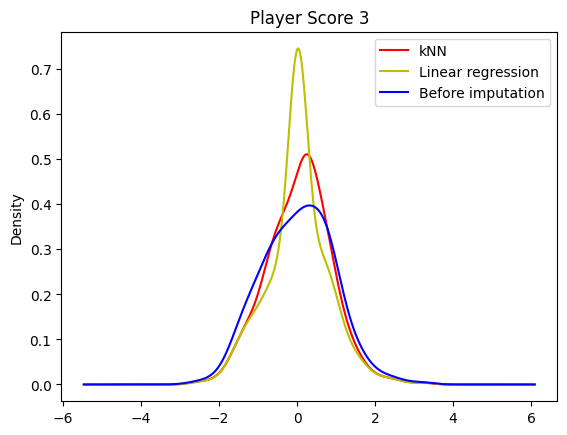
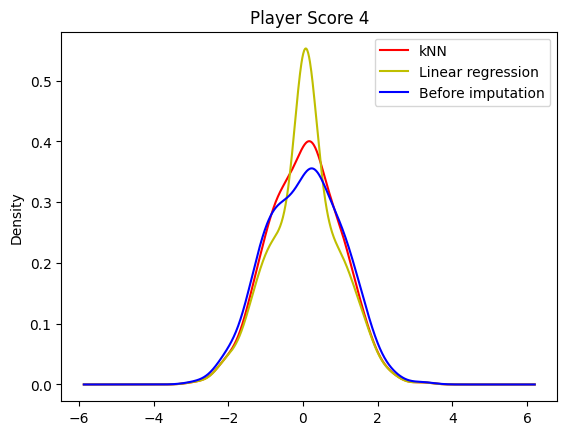
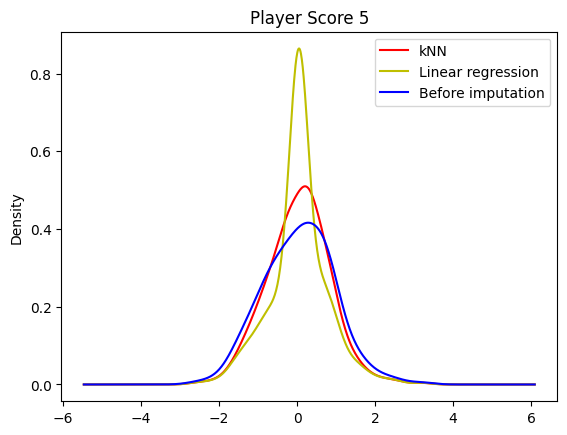
Dataset Analysis

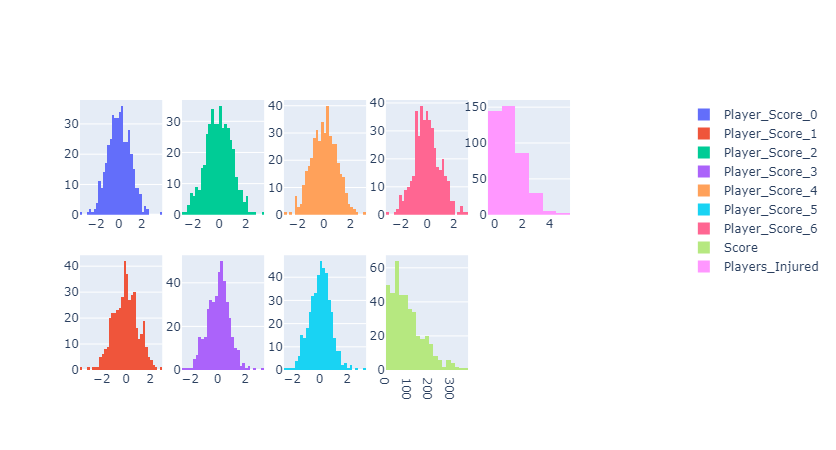
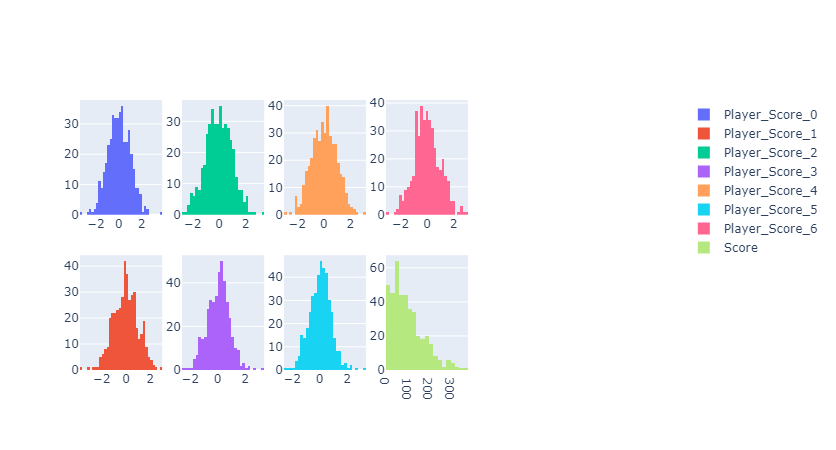
Supervised Learning by Oleg Lastocichin

27/11/2022

Questions and Tasks

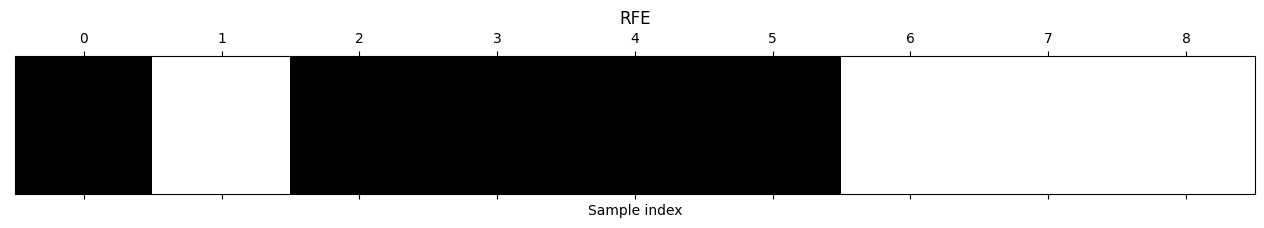
1. Load and explore the dataset
   1. Immagine che contiene testo

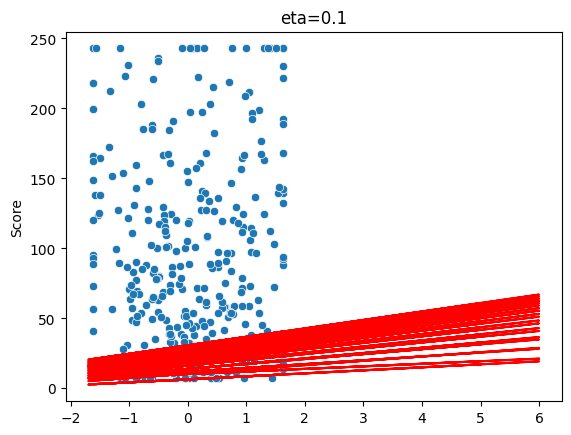
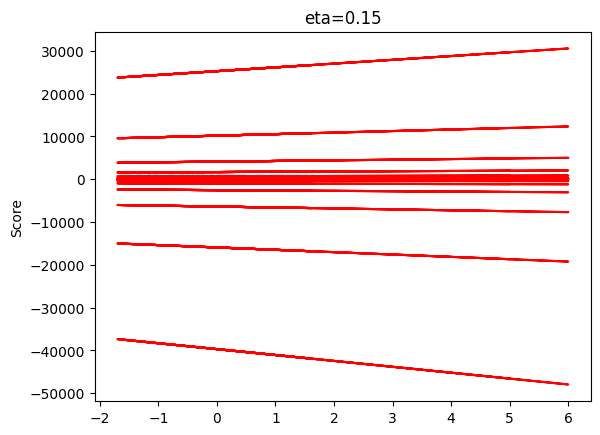
      Descrizione generata automaticamente There are 9 numerical features (Player Score 0 to 6, Score, Players Injured), there are 2 categorical features (Performance, Country)
   2. There are missing values in the dataset  
        
        
      Handling the missing values in the dataset:  
      Since the missing values don’t go over 70%, I won’t drop the columns instead I will try to impute them with Linear Regression or kNN   
        
        
        
      Immagine che contiene testo

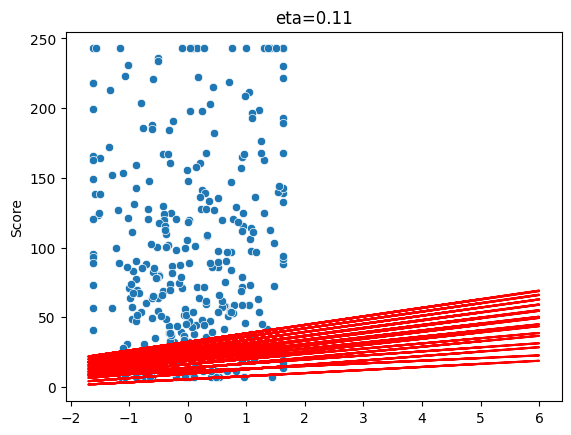
      Descrizione generata automaticamente
   3. Since the Shannon distance is smaller with kNN I will use that to fill my dataset’s Nan values and it’s clearly visible from the plots too.
2. Prepare the dataset for a Linear Regression task.
   1. 
   2. I will use a Robust scaler such that I can manage the outliers too, and I can represent the data for a better Linear Regression processdistribution after scaling
   3. Since robust scaling handles outliers too this process should be skipped but I did a IQR-score outlier detection to be sure, and as expected nothing changed  
      Immagine che contiene testo

      Descrizione generata automaticamente
   4. I think it’s necessary if I want to use the dataset for a Linear Regression and use this features too, then I used label encoding using dataframe.replace from pandas  
      Immagine che contiene testo

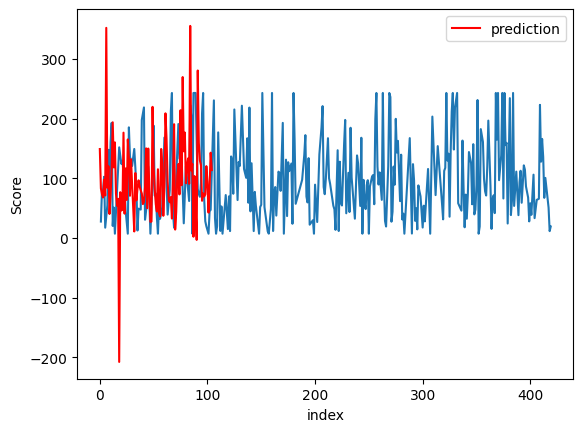
      Descrizione generata automaticamente
   5. I increased the dimensionality of the dataset using PolynomialFeatuers from sklearn
   6. I think I did all the useful/necessary transformation for the dataset so I will proceed with the next task.
3. Features Selection
   1. I performed an ANOVA test between Country and Score and ended up accepting H0 meaning that it doesn’t exist variance between the groups, I won't include this feature for model training.
   2. I tried with all the methodologies for feature selection, whereas some were not made for this dataset  
      I ended up making a choice between Model selection with random forest classifier and RFE, since I think those features would be more helpful for the Linear Regression task, my choice in the end was Model Selection  
      Immagine che contiene testo

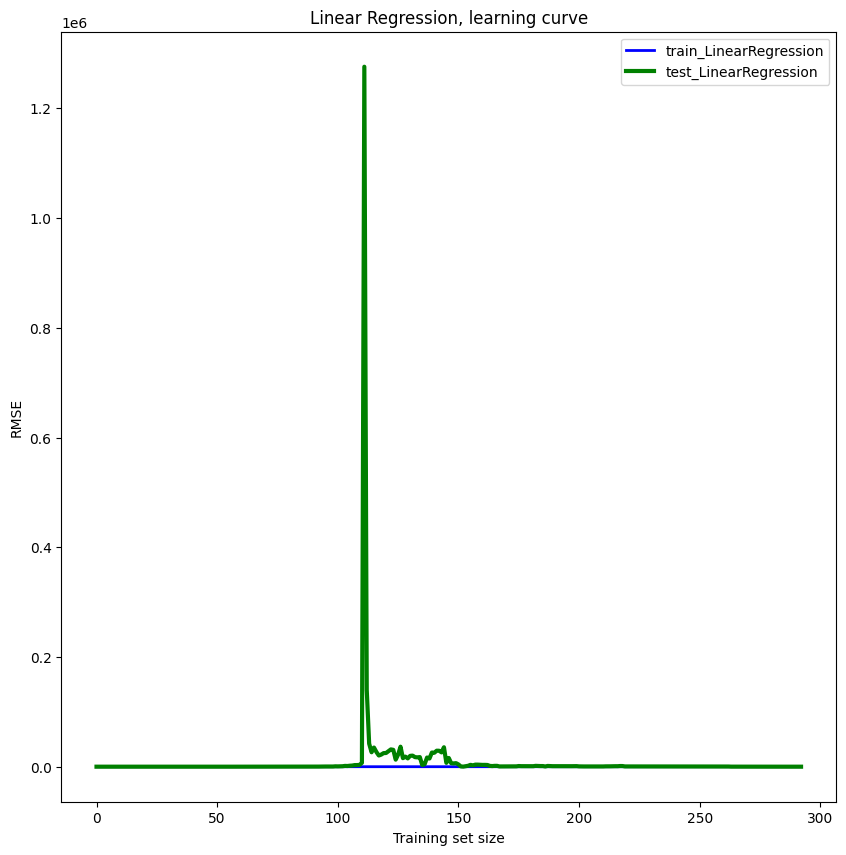
      Descrizione generata automaticamente  
      
4. Linear Regression
   1. I trained the model using the transformed features including Performance and Country and not using the polynomials  
      test RMSE=66.78679659791713
   2. test R2=-0.033550616190128
   3. train RMSE=66.4131367390314
   4. train R2=0.04970599069771997  
      Immagine che contiene testo

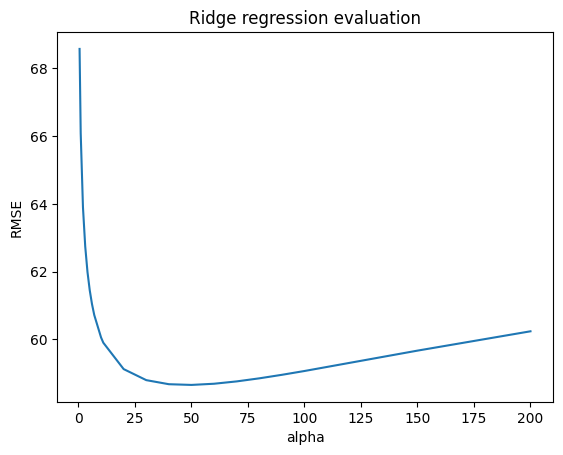
      Descrizione generata automaticamente
   5. I used eta 0.1, 0.15 and 0.11 where the best one that wasn’t overshooting was 0.11  
      



* 1. This linear regression felt more accurate from the predictions only



And in fact from the learning curve we can see that the RMSE is almost from the start 0 except from 1 strange spike at around 100 to 150 for the test set. 

* 1. I performed the Ridge regression (I used all the features and the polynomial features) and evaluated the RMSE and I noticed that it needed a really large alpha to reduce the RMSE 

And the RMSE didn’t shift that drastically either (while using big differences of alpha) Immagine che contiene testo

Descrizione generata automaticamente

* 1. Performing the Lasso regression felt better since I didn’t need such large alpha numbers (Even though I think Alpha = 3 is still kind of big) and RMSE was smallest at alpha = 3

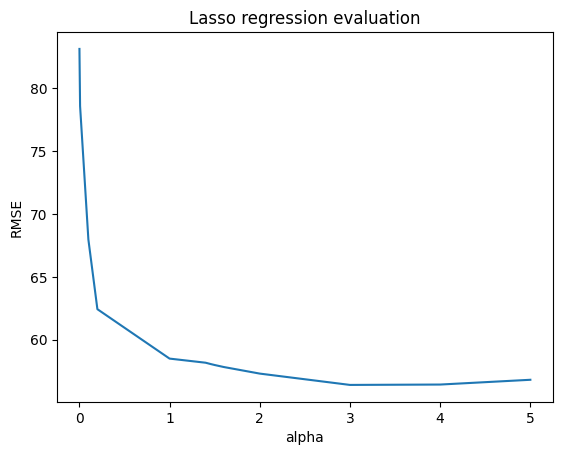


Immagine che contiene testo

Descrizione generata automaticamente

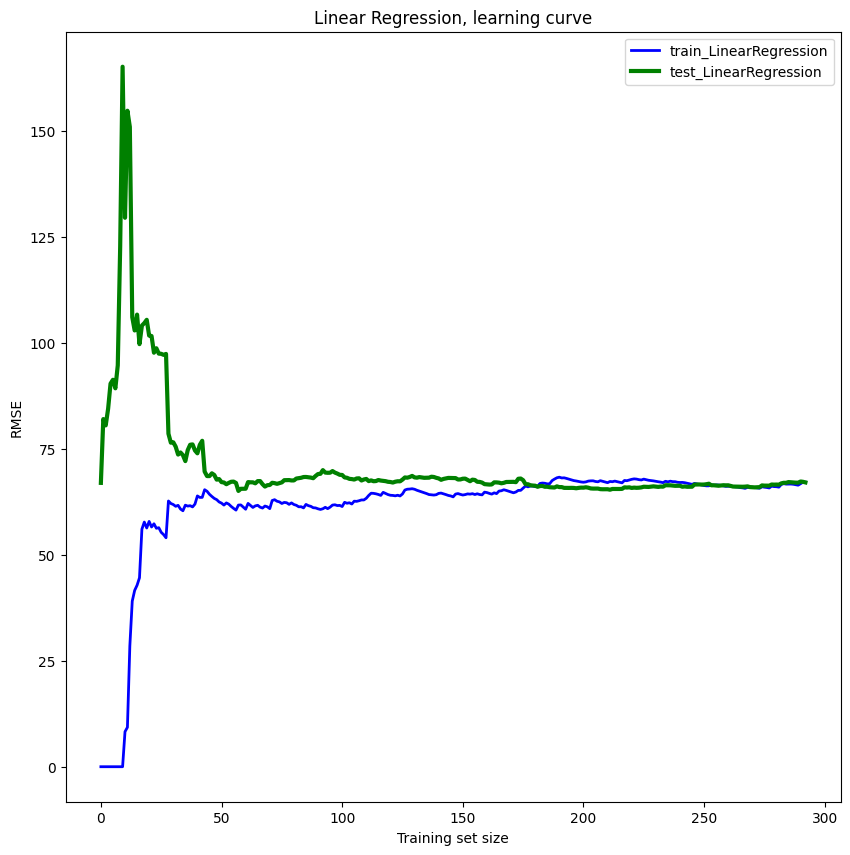
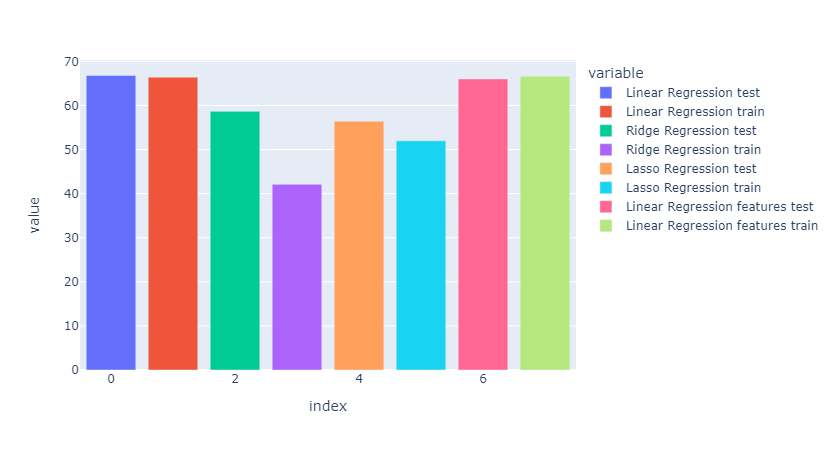
* 1. After the feature selection and using only the features I thought were going to represent a better regression I ended up with similar RMSE results just slightly better ones I suppose

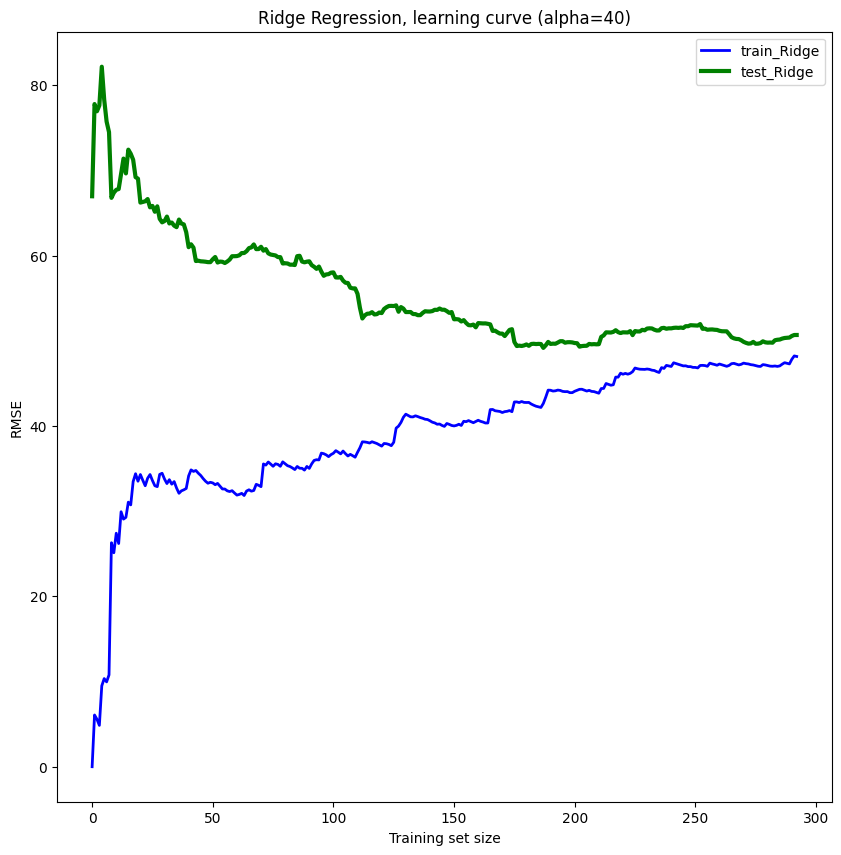
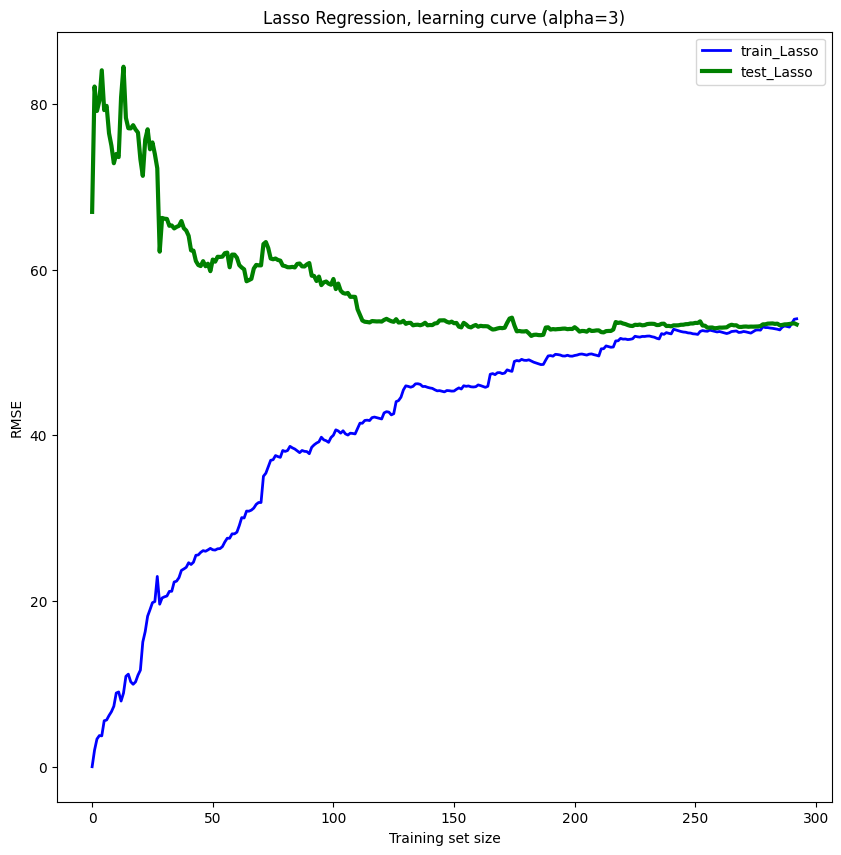
test RMSE=66.0050336150932

test R2=-0.009496080255134487

train RMSE=66.61741320391856

train R2=0.04385108740699217

g. I computed the learning curves and a table for the models where evaluation was requested, I noticed that the Linear Regressions always converge quickly but usually have a larger RMSE, instead the Lasso converges slightly slower but provides a smaller RMSE, and lastly the Ridge regression I think is the one who converges the slowest but has a smaller RMSE overall on the train set and Lasso had the best RMSE for the test set.



h.

Since the data was synthetic kNN imputation worked better instead of Linear Regression for this kind of dataset even though realistically Linear Regression should perform better since we shouldn’t be able to evaluate a team’s capabilities from the “Neighbouring” teams.

From the results I obtained above I think the best regression is Lasso Regression since it delivered the smallest RMSE for the test adding the fact that it converged faster than the Ridge regression, instead I think that the Batch Gradient could converge with eta = 0.1 and it was the best suited since it was converging to the linear regression results.

The dataset was obviously underfitting across all the models since I had a real high value for RMSE and R2 meaning that the complexity was too low to fit the data.

I don’t think any model is good to evaluate a team’s Score and a better job on fitting the data could’ve been done.