# Assignment 1

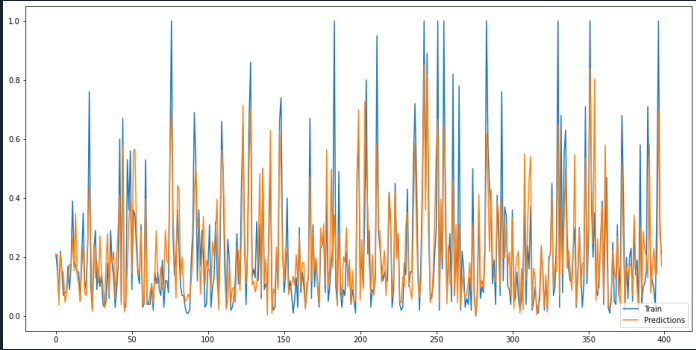
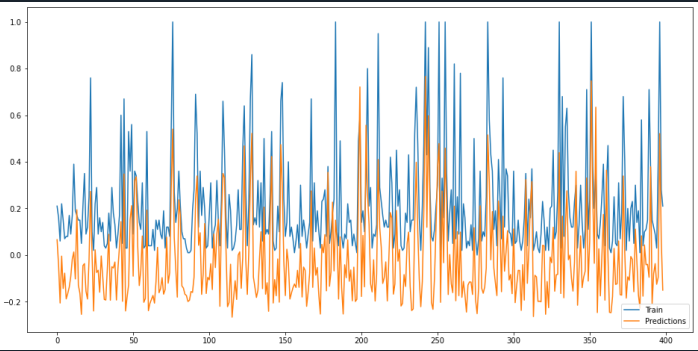
## Problem 1. Regression

### Data handling

The dataset used for this question came in a .csv file with 1994 rows and 128 columns of data (rows being each sample and columns having related data). In the image below there is an array of values which show the number of missing data per column. As seen below some columns have over 50% of missing values for their total sample space. Columns that had 1174 or more missing values were removed as training inputs. That left only one sample left that had a missing value. The sample/row was removed from the dataset. The columns relating to location or the first 5 columns where also removed as indicated by the question. This left 101 columns and 1993 rows left or 19993 samples with 100 input parameters for each sample.

### 

The data was not standardised as this seemed to cause the base and lasso linear regressor models to produce negative R2 values. Visually when the data was plotted the models looked to clearly be performing worse. An example of this can be seen below. In the 2 images left being with standardised data and right without any standardising for the data for the lasso model. The base linear regressor model had a similar change while the ridge model seemed unaffected.



The dataset was then shuffled randomly and split into 3 sets. The first set was the data that each of the models would be train on consisting of 60% percent of the total dataset. The remaining 40% of the dataset was then split equally (20% of total data) between a validation set and testing set. The validation set was used to determine hyperparameter of each of the models, training was used to train models to predict accuracy of the model on the remaining training set.

### Model details

3 models where trained on the dataset. These being linear, lasso regressor and ridge regressor. Hyperparameters where chosen for the linear regressor model by using a grid search method with cross validation used to determine accuracy with 4 folds. 3 parameters were chosen to be optimised. The same 3 hyperparameters where used for the lasso and ridge regressors however a 4th parameter was chosen to be optimised being λ (Alpha in sklearn api).

### Model comparison

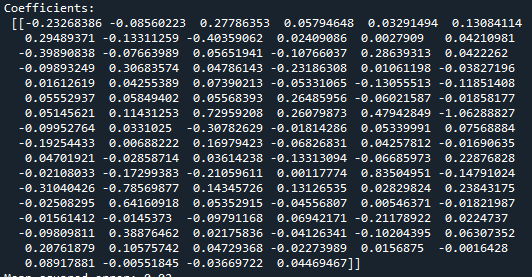
* An evaluation comparing the three models, considering model accuracy and model validity.

### Evaluation metrics

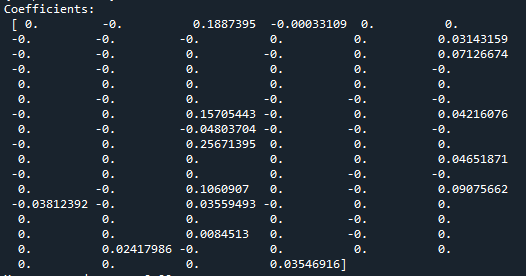
|  |  |  |
| --- | --- | --- |
| Model | Mean Square Error | R2 |
| Linear | 0.02 | 0.66 |
| Lasso | 0.02 | 0.66 |
| Ridge | 0.02 | 0.67 |

### Parameter Coefficients

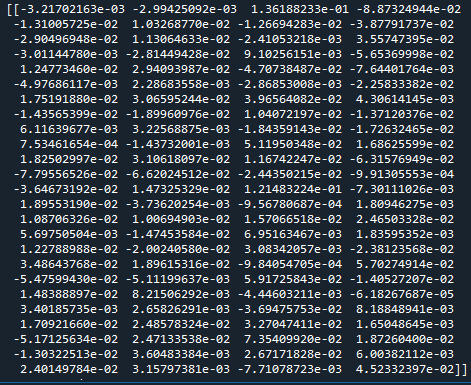
#### Linear



#### Lasso



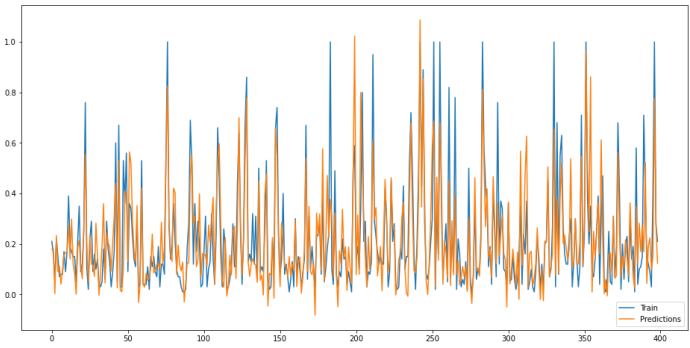
#### Ridge

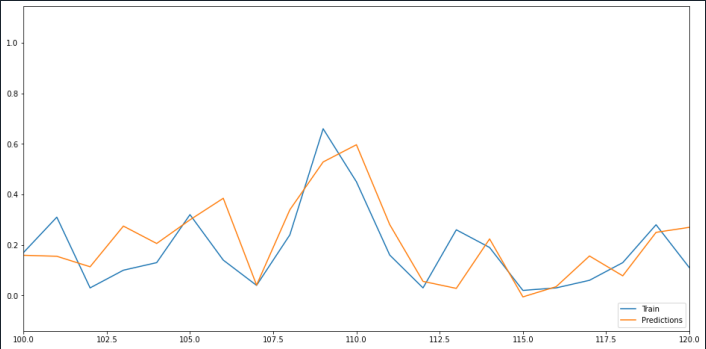


### Plots

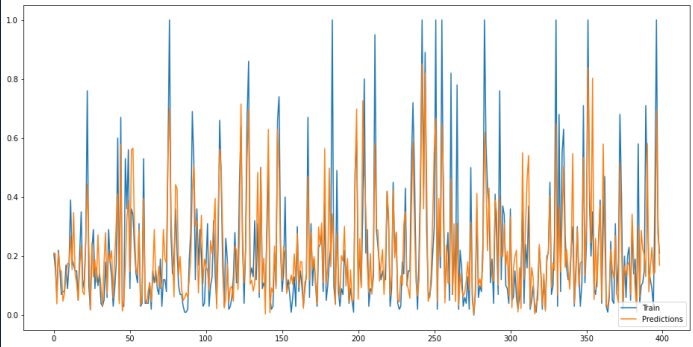
In the below plots the y axis is the predicted values and the x axis is an individual test sample. Blue represents the actual value a sample should produce and orange is the predicted value of a model.

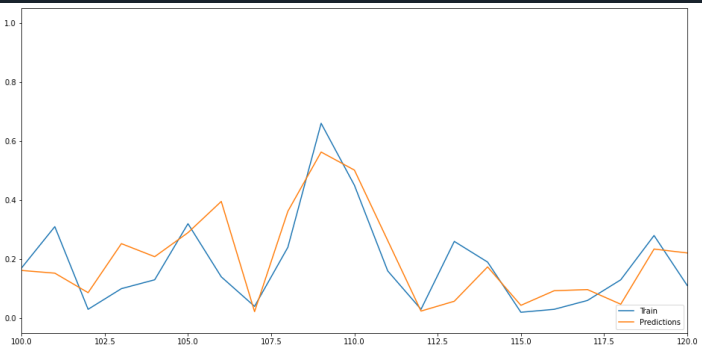
#### Linear plots



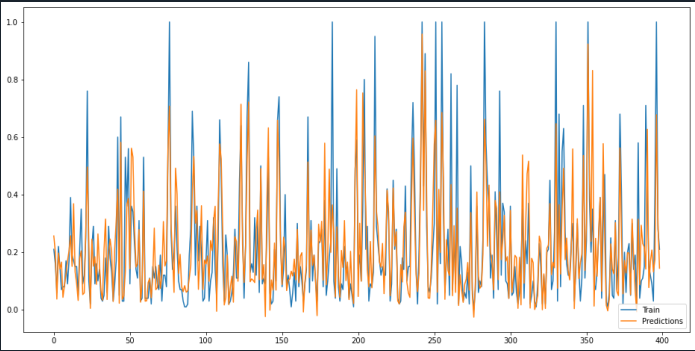


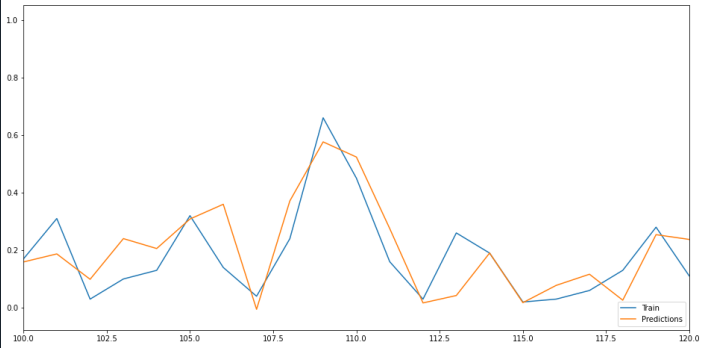
#### Lasso plots





#### Ridge plots





All three models results are significant, after applying grid search methodology. In lasso model there are 84 coefficients become 0 and eliminated from the model, only 16 coefficients left. In Ridge regression, after best lambada = 5.5 applied, it performs similar with the others two models, many features weight near to 0 after penalty lambada applied.

Your final response should include:

• An evaluation comparing the three models, considering model accuracy and model validity.

## Problem 2. classification

### Data handling

As the dataset given with no missing data there no cleaning required. The dataset was already split into training and testing set. The testing set was split in half into a new testing and validation set.

### Hyperparameter selection

### Model comparison

Your answer to this question should include:

• Details on how the data was split into training, validation and testing sets.

• Details of hyper-parameter selection, including justification for the approach taken and any intermediate results that led to the final models.

• An evaluation and comparison of the final three models, including a discussion exploring any difference in performance between the models. This should also highlight any major differences in the models in terms of classification performance or capabilities.