Course Name: Statistical learning for big data
Course Code: MVE440
Project 2

Devosmita Chatterjee
Personnummer: 910812-7748

June 11, 2021

Contents

Question 2: Feature selection

Firstly, the required libraries are imported. Then, the input data csv files are loaded and imported it into dataframe.

1 Methods and Results

1.1 Exploratory data analysis

An exploratory data analysis is performed with data in order to understand the dataset by summarizing their main characteristics, either statistically or visually. Data size, data type, missing data, duplicate data are listed in table 1.

Table 1: *The table presents the exploratory data analysis.*

Exploratory Data Analysis							
Data	**Row size** n	**Column size** p	**Data type**	**Missing values**	**Duplicate rows**	**Duplicate columns**	**Constant columns**
y	209	1	Numeric	0	-	-	-
X_1	209	160	Numeric	0	0	0	0
$X = (X_1, X_2)$	209	260	Numeric	0	0	0	0

1.2 Feature selection by Lasso regression using (A) only the features in X_1

1.2.1 Bootstrapping

In this case, only the features in X_1 are used.

Bootstrapping is used to Train Test splits of the given dataset to learn a Classifier (in this case, fitting a Lasso Regression model) in Machine Learning. The idea behind bootstrap is to use the data of a sample for approximating the sampling distribution of a statistic.

1.2.2 Building a pipeline by data standardization and Lasso model

Data standardization of a feature is given by the formula

$$Z = \frac{X - mean(X)}{sd(X)}$$

The idea of standardization is to scale the observations of the feature with mean 0 and standard deviation 1.

Lasso regression is a linear model that optimizes the cost function

$$\frac{1}{N_{training}} \sum_{i=1}^{N_{training}} (y_{real}^{(i)} - y_{pred}^{(i)})^2 + \alpha \sum_{j=1}^{n} |a_j|$$

where a_j is the coefficient of the jth feature and α is the hyperparameter that tunes the intensity of this penalty term.

The idea of Lasso regression is to optimize the cost function by minimizing the absolute values of the coefficients.

1.2.3 Optimizing the hyperparameter α of Lasso regression

The hyperparameter α of Lasso regression is optimized using the GridSearchCV object in the following way.

1. Several α values are tested from 0.1 to 10 with step 0.1.

2. For each α value, the average value of the mean squared error is calculated in a 5-folds cross-validation.

3. Select the value of α that minimizes such performance metrics.

Fitting the grid search to the training set, the best value of the hyperparameter α is found in table 2.

Table 2: *The table presents the methods' best hyperparameter using grid search algorithm.*

Method	α
lasso regression	0.1

1.2.4 Selecting the important features by Lasso regression

By minimizing the cost function, Lasso regression will automatically select the features that are useful or important, discarding the useless or redundant features. In Lasso regression, the importance of a feature is the absolute value of its coefficient. The absolute coefficient of a feature equal to 0, implies that the feature is redundant and the absolute coefficient of a feature greater than 0, implies that the feature is useful.

In this case, there are 25 selected important features by Lasso regression. They are the following:
'D4' 'D25' 'D33' 'D34' 'D35' 'D47' 'D51' 'D60' 'D67' 'D69' 'D78' 'D80' 'D89' 'D90' 'D93' 'D96' 'D99' 'D100' 'D104' 'D105' 'D128' 'D129' 'D140' 'D146' 'D157'

Lasso regression selected 25 features which is represented in table 3.

Table 3

Total features	**Selected features**	**Features with coefficients shrank to zero**
160	25	135

1.2.5 Model Evaluation

Regression metrics like the mean absolute error, the mean squared error, and the R-squared before and after feature selection are computed in table 4. The Actual versus Predicted graphs before and after feature selection are plotted in figures 1 and 2.

Table 4: *The table presents the comparison of Regression Metrics before and after feature selection.*

Regression Metrics			
Data	**Mean absolute error**	**Mean squared error**	**R^2 score**
All Features	45.17	9122.17	-0.62
Selected Features	14.71	1165.99	0.79

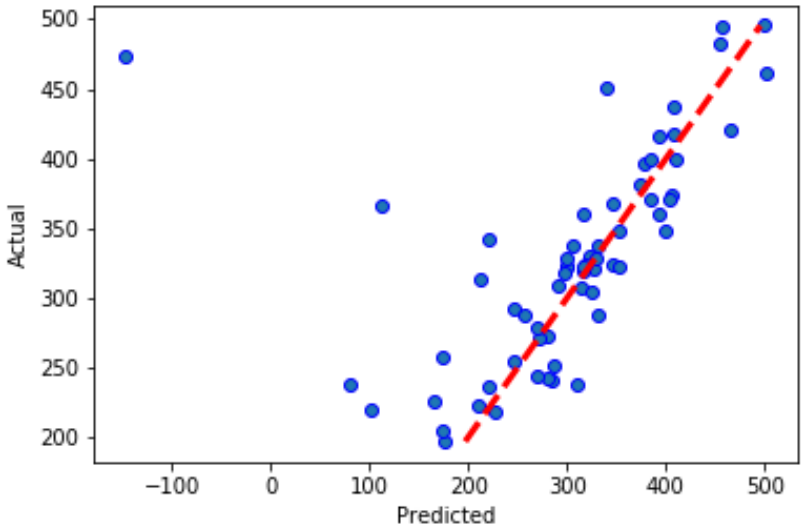

Figure 1: *The figure shows the Actual vs Predicted graph before feature selection.*

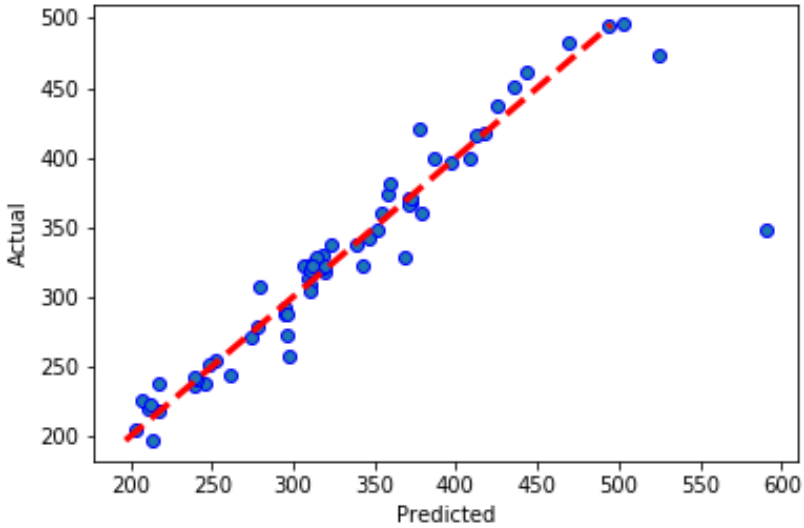

Figure 2: *The figure shows the Actual vs Predicted graph after feature selection.*

1.3 Feature selection by Lasso regression using (B) only the features in X_1 and X_2 together

1.3.1 Bootstrapping

In this case, the features in X_1 and X_2 together are used.

Bootstrapping is used to Train Test splits of the given dataset to learn a Classifier (in this case, fitting a Lasso Regression model) in Machine Learning. The idea behind bootstrap is to use the data of a sample for approximating the sampling distribution of a statistic.

1.3.2 Building a pipeline by data standardization and Lasso model

Data standardization of a feature is given by the formula

$$Z = \frac{X - mean(X)}{sd(X)}$$

The idea of standardization is to scale the observations of the feature with mean 0 and standard deviation 1.

Lasso regression is a linear model that optimizes the cos function

$$\frac{1}{N_{training}} \sum_{i=1}^{N_{training}} (y_{real}^{(i)} - y_{pred}^{(i)})^2 + \alpha \sum_{j=1}^{n} |a_j|$$

where a_j is the coefficient of the jth feature and α is the hyperparameter that tunes the intensity of this penalty term.

The idea of Lasso regression is to optimize the cost function by minimizing the absolute values of the coefficients.

1.3.3 Optimizing the hyperparameter α of Lasso regression

The hyperparameter α of Lasso regression is optimized using the GridSearchCV object in the following way.

1. Several α values are tested from 0.1 to 10 with step 0.1.

2. For each α value, the average value of the mean squared error is calculated in a 5-folds cross-validation.

3. Select the value of α that minimizes such performance metrics.

Fitting the grid search to the training set, the best value of the hyperparameter α is found in table 5.

Table 5: *The table presents the methods' best hyperparameter using grid search algorithm.*

Method	α
lasso regression	0.1

1.3.4 Selecting the important features by Lasso regression

Minimizing the cost function, Lasso regression will automatically select the features that are useful or important, discarding the useless or redundant features. In Lasso regression, absolute coefficient of a feature equal to 0, implies discarding the feature and absolute coefficient of a feature greater than 0, implies that the feature is useful.

In this case, there are 34 selected important features by Lasso regression. They are the following:
'D25' 'D33' 'D51' 'D60' 'D67' 'D68' 'D69' 'D78' 'D89' 'D90' 'D92' 'D96' 'D99' 'D100' 'D105' 'D106' 'D121' 'D124' 'D128' 'D134' 'D140' 'D149' 'D157' 'D159' 'D169' 'D170' 'D171' 'D182' 'D189' 'D196' 'D202' 'D203' 'D214' 'D220'

Lasso regression selected 34 features which is represented in table 6.

Table 6

Total features	**Selected features**	**Features with coefficients shrank to zero**
260	34	226

1.3.5 Model Evaluation

Regression metrics like the mean absolute error, the mean squared error, and the R-squared before and after feature selection are computed in table 7. The Actual versus Predicted graphs before and after feature selection are plotted in figures 3 and 4.

Table 7: *The table presents the comparison of Regression Metrics before and after feature selection.*

Regression Metrics			
Data	**Mean absolute error**	**Mean squared error**	**R^2 score**
All Features	40.13	6864.94	-0.22
Selected Features	12.04	237.70	0.96

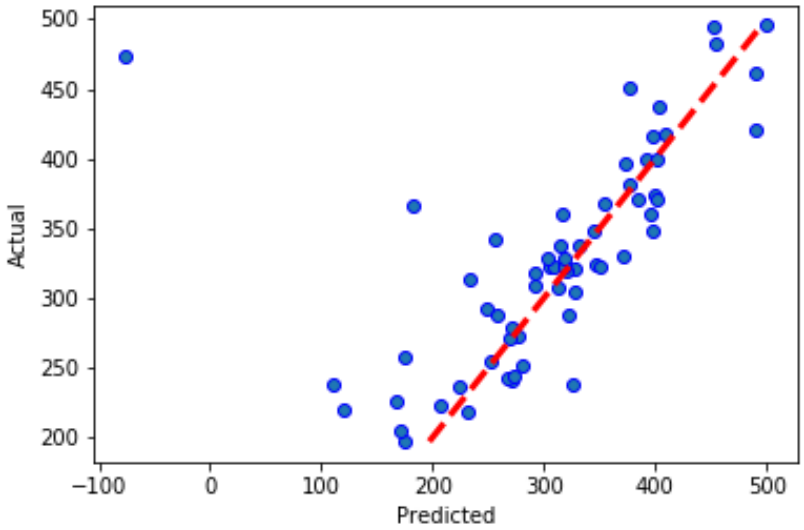

Figure 3: *The figure shows the Actual vs Predicted graph before feature selection.*

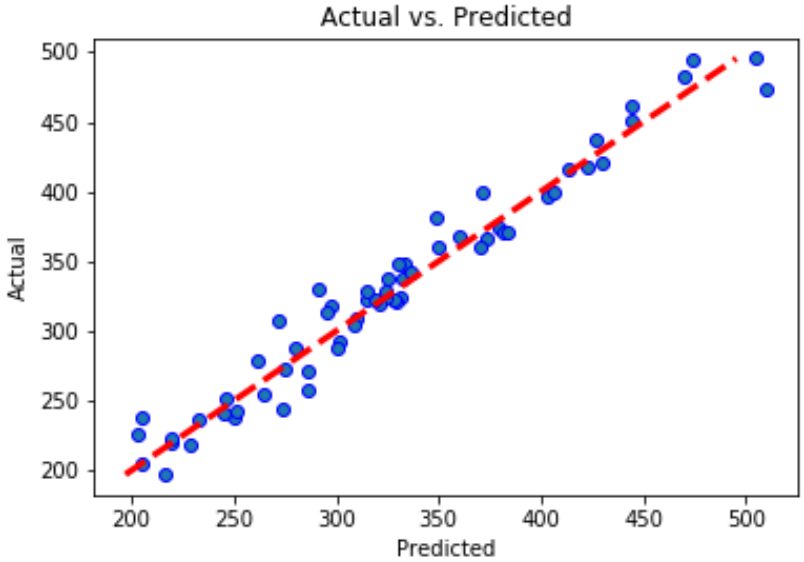

Figure 4: *The figure shows the Actual vs Predicted graph after feature selection.*

2 Discussion

In the section, I discuss about the following results:-

1. From exploratory data analysis of the given datasets, t is found that X=(X_1,X_2) is high dimensional dataset that is, number of features greater than number of observations.

2. Lasso regression is a popular feature selection technique for high dimensional dataset. Lasso uses a penalization method which reduces the chance of overestimating a regression coefficient.

3. Considering the computed regression metrics like the mean absolute error, the mean squared error, and the R-squared, it is concluded that the linear regression model is improved after feature selection in both the cases.