Innopolis University – Course of Machine Learning – Academic Year 2021/2022

Report Assignment 1

Mauricio Salguero Gaspar

https://github.com/m-salguerogaspar/ML-Assigment_1

1 Introduction

Flight delay is an inevitable issue and it plays an important role in profits of the airlines. Prior prediction of flight arrival delays is necessary for travelers and airlines companies because delays in flights is not just only about profits but also about losing the prestige that was built for many years and passengers losing their valuable time. The results of a precise estimation of flight delay is crucial for airlines because can be applied to increase customer satisfaction and incomes of airline agencies.

The main goal of the present research was explore different regression algorithms with different hyperparameters to find the best possible model to predict flight delays and help to solve the problematic mentioned above. You will find in this document the description of the dataset used, a detailed analysis of the proposed models, advantages and drawbacks, explanation of some techniques used to overcome the problems found, the metrics choosen to measure the performance and a final model choice with the best efficiency.

2 Dataset

The Dataset comes from Innopolis University partner company analyzing flights delays. There are 675.513 samples in the flight_delay.csv file and each sample in the dataset corresponds to a flight, the data was recorded over a period of 4 years, from 2015 to 2018, it is important to mention that the model will be created and trained with the first three years data and evaluated with the last year data. These flights are described according to 5 variables. A sneck peek of the dataset can be seen in the table below:

	Depature Airport	Scheduled depature time	Destination Airport	Scheduled arrival time	Delay
0	SVO	2015-10-27 07:40:00	HAV	2015-10-27 20:45:00	0.0
1	SVO	2015-10-27 09:50:00	JFK	2015-10-27 20:35:00	2.0
2	SVO	2015-10-27 10:45:00	MIA	2015-10-27 23:35:00	0.0
3	SVO	2015-10-27 12:30:00	LAX	2015-10-28 01:20:00	0.0
4	OTP	2015-10-27 14:15:00	SVO	2015-10-27 16:40:00	9.0
5	HAM	2015-10-27 14:30:00	SVO	2015-10-27 17:15:00	0.0
6	SVO	2015-10-27 14:35:00	JFK	2015-10-28 01:25:00	0.0
7	DXB	2015-10-27 15:40:00	SVO	2015-10-27 21:20:00	1.0
8	SVO	2015-10-27 16:10:00	VVO	2015-10-28 00:35:00	0.0
9	TLV	2015-10-27 16:45:00	SVO	2015-10-27 20:55:00	0.0

Figure 1: Data Description.

The description of the 5 variables describing each flight are:

Variable name	Description
Departure Airport	Name of the airport where the flight departed. The name is given as airport international code
Scheduled departure time	Time scheduled for the flight take-off from origin airport
Destination Airport	Flight destination airport. The name is given as airport international code
Scheduled arrival time	Time scheduled for the flight touch-down at the destination airport
Delay (in minutes)	Flight delay in minutes

Figure 2: Variables Description.

Although during the first process analysis a new column will be added with the name 'Flight duration' that will represent the difference between Scheduled arrival time and Scheduled departure time, this is done to get a closer understanding of the problem and to create a readable graphical representation of the task. It is important to mention that there were not missing values for that reason the imputation process was skipped.

3 Preprocessing Steps

3.1 Visualization

First of all, to understand the problem is always important visualize it in a plot, the dependent variable will be Delay and the other feature that could be meaningful to reduce the chart to two dimensions would be the Flight duration:

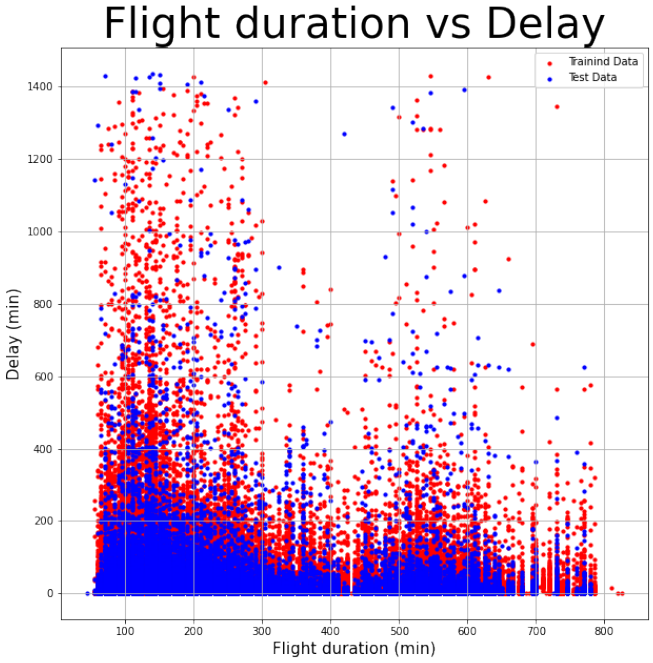

Figure 3: Visual Representation of the task considering just two variables.

As you can see it is not easy try to identify the type of relationship between these two variables and it does not look like a linear relationship, even so this representation is useful as well to identify the outliers but that is a topic dealt later.

3.2 Outlier Detection

By definition an outlier is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to the us to decide what will be considered abnormal. As you can see in the following box plot , the IQR region is almost imperceptible due to the big amount of outlier in our case.

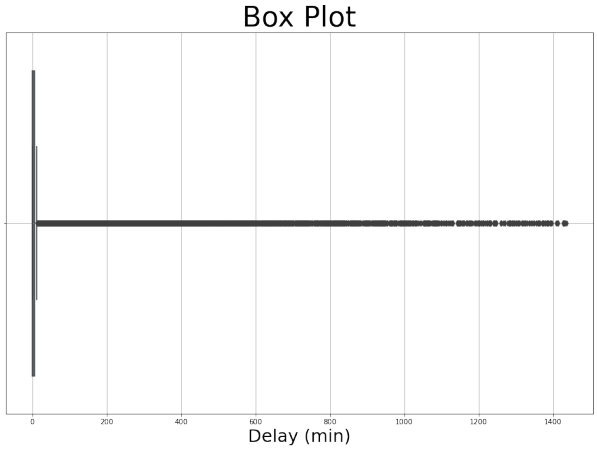

Figure 4: Representation of outliers through box plot.

The criteria used to determine extreme outliers is any point greater than The upper outer fence:

$$\text{Upper Outer Fence:} \quad Q_3 + Criteria * Iq$$

The reason is because the distribution is not normal to use Z-score criteria, it is too right skewed even trying to centered around the mean as it is shown in the figure, for that reason it was decided to use the Iq criteria.

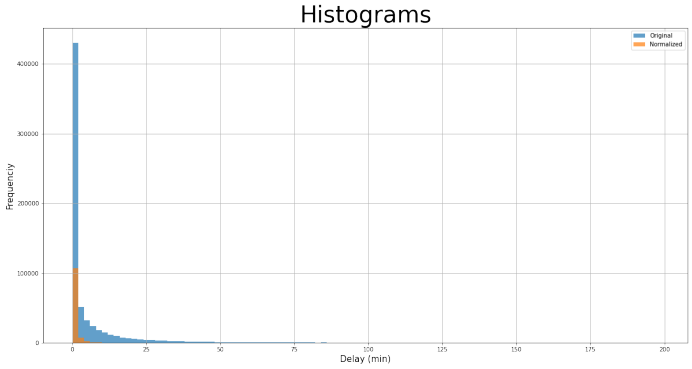

Figure 5: Comparison of histograms to the original data vs centered around the mean data.

Two values for criteria are going to be used, 1.5 acepted and standard value, and 6.0, this with the goal of not lossing too many data, 15% and 5% respectively.

3.3 Encoder

In this research the target is the prediction of delay in minutes and the features variable are: the new column created Flight duration, and four categorical variables: Departure and Destination airport, Scheduled departure and arrival time. The uniques values found in the airport categories were 72 and 69 respectively so it was used the OneHotEncoder tool to create the dummy variables because they are nominal categories. For the departure and arrival time, the variables extracted are:

- The month, treated as numerical values.
- The day of the month, in this case a normal month was divided in four weeks, and each day was mapped to one of these regions and after that the OneHotEncoder was used.
- The hour of the day, in this case a normal day was divided in three regions: morning, day and night, each hour was mapped to one of these regions and after that the OneHotEncoder was used.

This mapping, dates and times treated as categorical variables, was done to reduce the number of dimensions and not make way bigger the feature space.

4 Models and Results

First of all, before talking about the models it is important to define the metrics used to measure sucess. This is a regression problem and the metrics choosen are MSE, MAE and MRSE, the most common ones, due to the following advantages:

- The MAE you get is in the same unit as the output variable.
- The MAE is most Robust to outliers.
- The graph of MSE is differentiable, so you can easily use it as a loss function.
- The MRSE output value you get is in the same unit as the required output variable which makes interpretation of loss easy.

Secondly, many models were created looking for the best one, linear regression, with regularization, and polynomial regression, with regularization, these models will be compared to choose the best one.

4.1 Linear Regression

The first model proposed was simple linear regression with some percentage of dimensions reducted, and comparing the two values of criteria, the values obtained are found in the next table:

Table 1: Linear Rregression performance (Outlier criteria = 6).

	PCA Reduction	0%	10%	20%	30%	40%	50%
Train	MSE	44.83	44.84	44.86	44.98	45.26	45.41
Test	MSE	30.29	30.29	30.37	30.31	30.12	29.98
Train	RMSE	6.69	6.69	6.69	6.70	6.72	6.73
Test	RMSE	5.50	5.50	5.51	5.50	5.48	5.47
Train	MAE	4.60	4.60	4.60	4.61	4.63	4.65
Test	MAE	4.16	4.16	4.17	4.17	4.16	4.16

In the results there seems to be no evidence of overfitting and underfitting, the best results obtained are related with 0% of dimensions reducted and it looks in the multidimensional space the problem could be really good treated as linear, although the results in MSE are not low like in the other two metrics.

Table 2: Linear Regression performance (Outlier criteria = 1.5).

	PCA Reduction	0%	10%	20%	30%	40%	50%
Train	MSE	8.66	8.66	8.67	8.70	8.70	8.77
Test	MSE	5.64	5.64	5.64	5.58	5.55	5.52
Train	RMSE	2.94	2.94	2.94	2.95	2.95	2.96
Test	RMSE	2.37	2.37	2.37	2.36	2.35	2.35
Train	MAE	2.19	2.19	2.19	2.19	2.20	2.21
Test	MAE	1.94	1.94	1.94	1.93	1.93	1.93

There are two main conclusions in that last table, the first one, the behaviour of the metrics are the same, but better results are obtained for the metrics with more outliers removed and second, the percentage of dimensions reducted seems not affect the performance of the model which is excelent for the running time of the predictor model created. The following analysis will be done just with 1.5 as outlier criteria due to this better results.

4.2 Linear Regression with Regularization

Looking for a better model regularization is applied through Lasso and Ridge regression, 15 differents values of alphas are tested including the same reduction of dimensions as before. In each image there are the three metrics used, for train and test stages, the following image shows the best model with the respective errors found, the 10 remaining images could be found in the Appendix A.

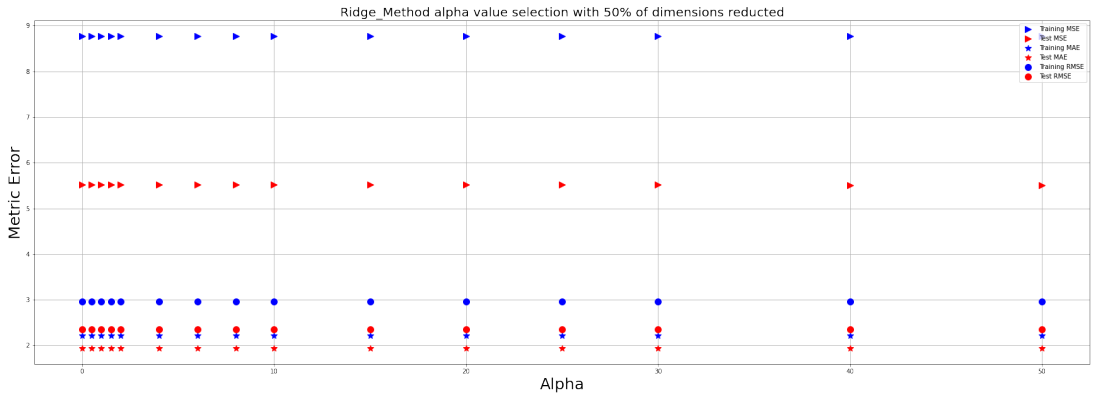

Figure 6: Ridge model selected with small regularization with PCA = 50%.

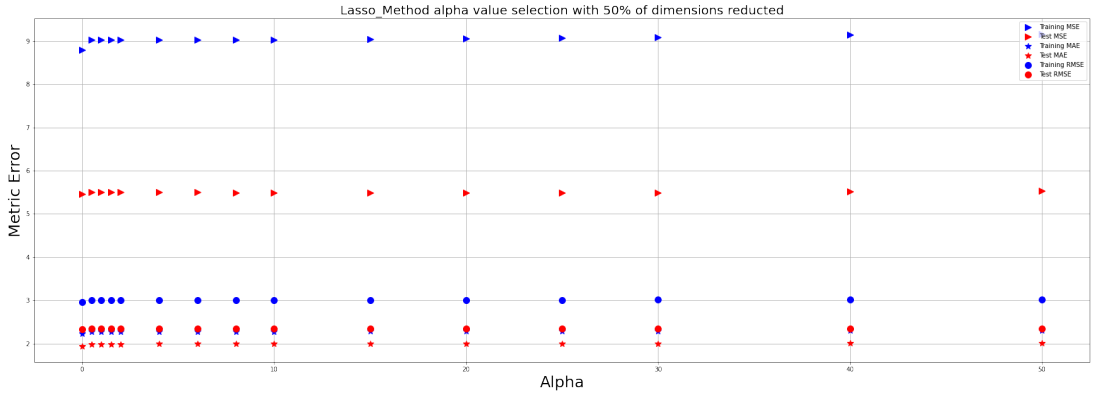

Figure 7: Lasso model selected with small regularization with PCA = 50%.

For both models, Lasso and Ridge, you will notice in the images in Appendix A, there is no a notable difference between the models results, the advantages of the model selected is the reductions of features, and the best value of alpha, no just the model selected but all models, is $\alpha = 0.001$, so it looks that the models do not need a strong regularization, and again, tbe behaviour of the metrics are similar being the ”worst” behavior MSE for both models.

No overfitting or underfitting phenomenon was observed in the previous models

4.3 Polynomial Regression with Regularization

So far, the problem has been treated as linear, in these last models we are going to create polynomial features, unfortunately to do this we have to reduce 90% and 95% of dimensions, due to the large amount of new features created, and also just degree = 2 for the same reason mentioned before. Here is the best model found with the respective errors, the remaining 2 images can be found in Appendix B.

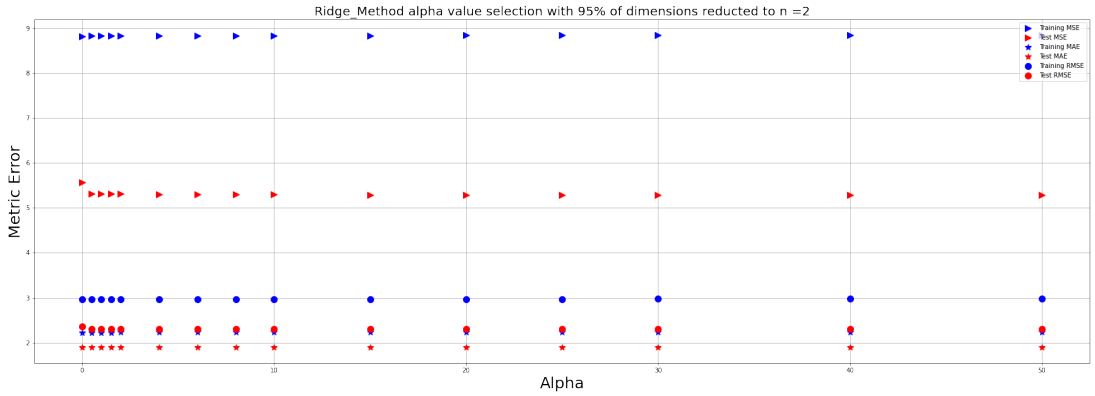

Figure 8: Ridge model selected with small regularization with PCA = 95%.

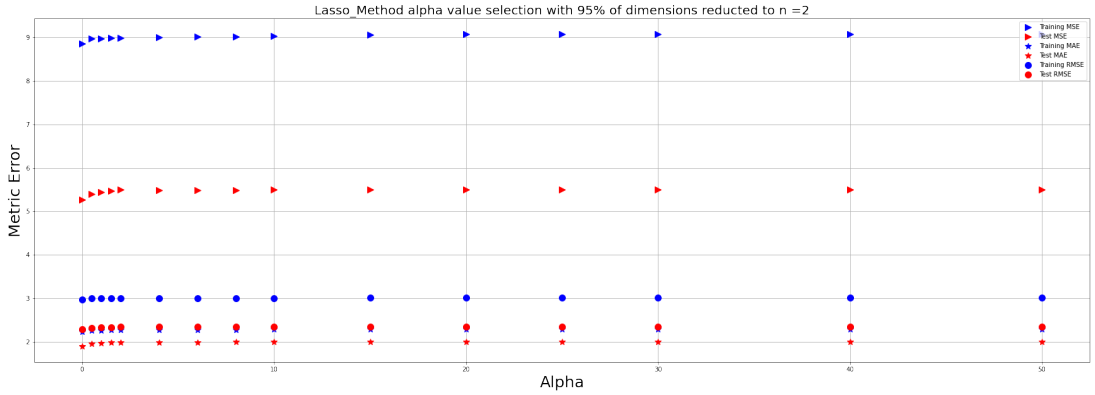

Figure 9: Lasso model selected with small regularization with PCA = 95%.

There are slight differences between the models, fabulously the models have a really good behaviour with this polynomial approximation, although degree two is still a quite low degree, and more impressive, the extreme reduction of dimensions seems not affect the model.

Lasso ans Ridge methods have the same behaviour of the error in metrics with similar values, the best behaviours are obtained for small regularization as well. No overfitting or underfitting phenomenon was observed in the previous models.

In conlusion, taking into consideration the three metrics and the behaviour over reduction of dimensions, for small differences in the values of the MSE metric, the best model found is Ridge regression with $\alpha = 0.001$ and PCA = 50%. Although, the behaviour of the three different metrics are similar, the error of MAE was always the smallest one which is a good indicator of robustness against outliers, in this model the performance in test set was better than the performance in training set, as in every model.

Table 3: Best Model Errors.

	PCA Reduction	50%
Train	MSE	8.76
Test	MSE	5.51
Train	RMSE	2.96
Test	RMSE	2.34
Train	MAE	2.21
Test	MAE	1.93

5 Appendix

5.1 Linear Regression with Regularization, Images:

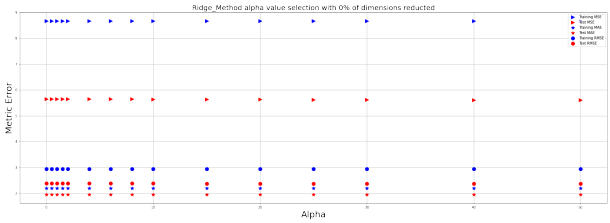

Figure 10: Ridge Model with PCA = 0%.

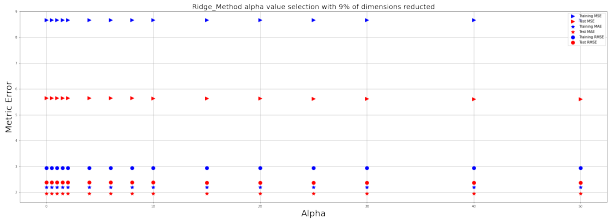

Figure 11: Ridge Model with PCA = 10%.

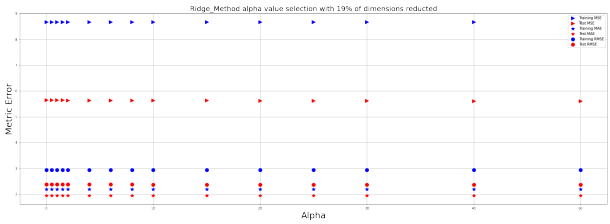

Figure 12: Ridge Model with PCA = 20%.

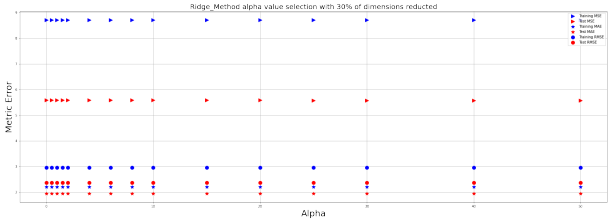

Figure 13: Ridge Model with PCA = 30%.

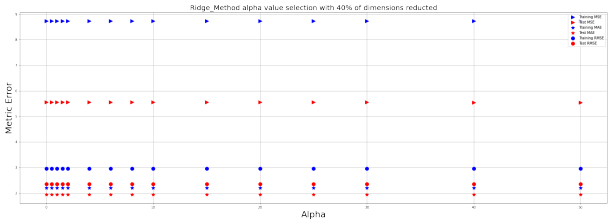

Figure 14: Ridge Model with PCA = 40%.

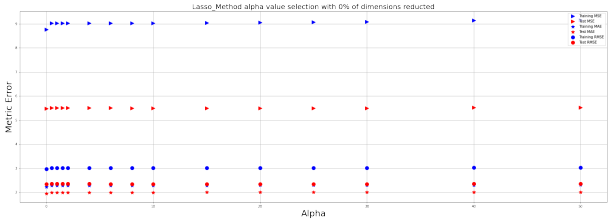

Figure 15: Lasso Model with PCA = 0%.

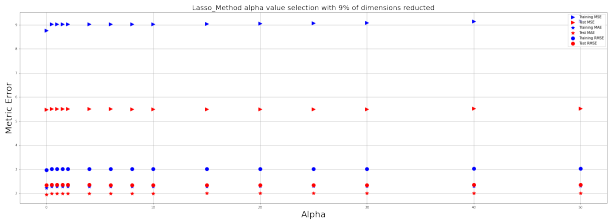

Figure 16: Lasso Model with PCA = 10%.

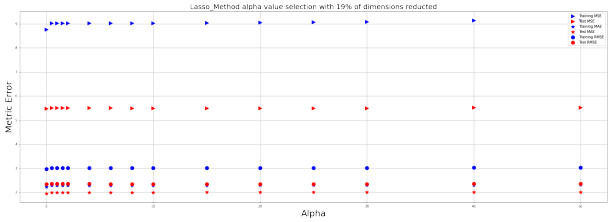

Figure 17: Lasso Model with PCA = 20%.

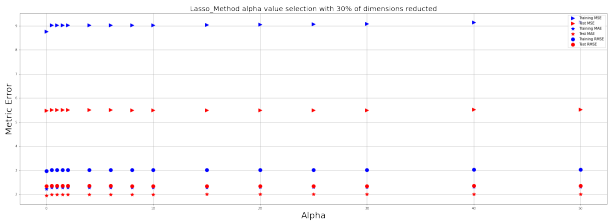

Figure 18: Lasso Model with PCA = 30%.

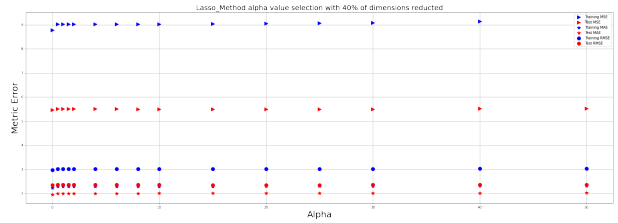

Figure 19: Lasso Model with PCA = 40%.

5.2 Polynomial Regression with Regularization, Images:

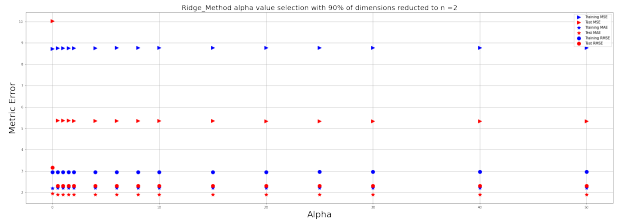

Figure 20: Polynomial Ridge Model with PCA = 90%.

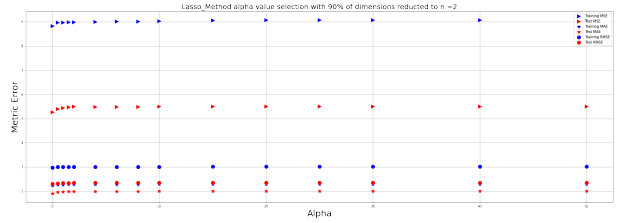

Figure 21: Polynomial Lasso Model with PCA = 90%.