**Machine Learning Lab (CS-353) Report**

**Lasso Regression**

**Team Members:**

1. Shashank D 181CO248
2. Tarun S Anur 181CO255

## **Lasso Regression - Introduction**

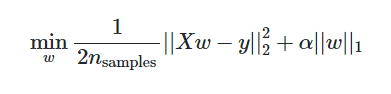
The Lasso is a linear model that estimates sparse coefficients. It is useful in some contexts due to its tendency to prefer solutions with fewer non-zero coefficients of features, effectively reducing the number of features upon which the given solution is dependent. Hence lasso can be used as a feature selection technique as well. Lasso regression is a type of linear regression that uses shrinkage where slope coefficient values are shrunk towards zero or large values. The lasso procedure encourages simple, sparse models (i.e., models with fewer parameters). This particular type of regression is well-suited for models showing high levels of multicollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination. Lasso was introduced in order to improve the prediction accuracy and interpretability of regression models. It selects a reduced set of the known covariates for use in a model.

The acronym “LASSO” stands for Least Absolute Shrinkage and Selection Operator.

Lasso regression performs L1 regularization, which adds a penalty equal to the absolute value of the magnitude of coefficients (slope terms). This type of regularization can result in sparse models with few coefficients because some coefficients can become zero which can be eliminated from the model. Larger penalties result in coefficient values closer to zero, which is the ideal for producing simpler models. On the other hand, L2 regularization (e.g., Ridge regression where squares of coefficients are added as penalty) doesn’t result in elimination of coefficients or sparse models. This makes Lasso far easier to interpret than the Ridge as under certain conditions, it can recover the exact set of non-zero coefficients.

## **Lasso Regression - Algorithm**

Mathematically, it consists of a linear model with an added regularization term. The objective function to minimize is:



The lasso estimate thus solves the minimization of the least-squares penalty with added where is a constant and 1 is the 1 -norm of the coefficient vector.

A tuning parameter, controls the strength of the L1 penalty. is basically the amount of shrinkage.

When = 0, no parameters are eliminated. The estimate is equal to the one found with linear regression (least squares method).

As *α* increases, more and more coefficients are set to zero and eliminated (theoretically, when *α* = ∞, all coefficients are eliminated).

As *α* increases, bias with the training set increases.

As *α* decreases, variance of testing set output increases.

If an intercept is included in the model, it is usually left unchanged.

## **Pros and Cons of Lasso Regression**

**Pros:**

* Select features, by shrinking co-efficient towards zero.
* It is fast in terms of inference and fitting.
* Avoids overfitting. It can also be applied where number of features is larger than the amount of data.

**Cons:**

* Selected features will be highly biased.
* For n<<p (n-number of data points, p-number of features), LASSO selects at most n features.
* LASSO will select only one feature from a group of correlated features, the selection is arbitrary in nature.
* For different bootstrapped data, the feature selected can be very different.
* Prediction performance is worse than Ridge regression.

## **Applications of Lasso Regression**

* This method can be used in business platforms to estimate revenue and other financial related statistics.
* It can be used as a feature selection method as well.
* It can be used in Medical field to diagnose the blood level and other related parameters and in turn predict any major ailments.
* In the Economics field to predict high growth firms.
* In the Financial department to predict corporate bankruptcy.

## **Lasso Regression - Numerical Analysis**

1. Dataset description

salary\_data.xls, a spreadsheet which contains total number of years of experience and the present salary of 30 employees in a company.

1. Numerical Solution

* We shall first find out the mean of the independent variable x (years of experience in this example) and the mean of the dependent variable y (salary in this example).
* We then calculate (xi-mean(xi)) and (yi-mean(yi)) for all the xi and yi in the dataset.
* Then we calculate 1. (sum((xi-mean(xi))) \* sum((yi-mean(yi)))) and 2. Sum((xi-mean(xi))) ^2.
* We calculate (1) / (2) to find the slope of the best fitting line (let’s say w1)
* Calculate w0 using the formula mean(yi) – w1\*mean(xi).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Years of Experience (xi) | salary (yi) | xi - xbar | yi - ybar | (xi - xbar)^2 | (xi - xbar)\*(yi - ybar) |
| 1.1 | 39343 | -4.2133 | -36326.7 | 17.75189689 | 153055.2851 |
| 1.3 | 46205 | -4.0133 | -29464.7 | 16.10657689 | 118250.6805 |
| 1.5 | 37731 | -3.8133 | -37938.7 | 14.54125689 | 144671.6447 |
| 2 | 43525 | -3.3133 | -32144.7 | 10.97795689 | 106505.0345 |
| 2.2 | 39891 | -3.1133 | -35778.7 | 9.69263689 | 111389.8267 |
| 2.9 | 56642 | -2.4133 | -19027.7 | 5.82401689 | 45919.54841 |
| 3 | 60150 | -2.3133 | -15519.7 | 5.35135689 | 35901.72201 |
| 3.2 | 54445 | -2.1133 | -21224.7 | 4.46603689 | 44854.15851 |
| 3.2 | 64445 | -2.1133 | -11224.7 | 4.46603689 | 23721.15851 |
| 3.7 | 57189 | -1.6133 | -18480.7 | 2.60273689 | 29814.91331 |
| 3.9 | 63218 | -1.4133 | -12451.7 | 1.99741689 | 17597.98761 |
| 4 | 55794 | -1.3133 | -19875.7 | 1.72475689 | 26102.75681 |
| 4 | 56957 | -1.3133 | -18712.7 | 1.72475689 | 24575.38891 |
| 4.1 | 57081 | -1.2133 | -18588.7 | 1.47209689 | 22553.66971 |
| 4.5 | 61111 | -0.8133 | -14558.7 | 0.66145689 | 11840.59071 |
| 4.9 | 67938 | -0.4133 | -7731.7 | 0.17081689 | 3195.51161 |
| 5.1 | 66029 | -0.2133 | -9640.7 | 0.04549689 | 2056.36131 |
| 5.3 | 83088 | -0.0133 | 7418.3 | 0.00017689 | -98.66339 |
| 5.9 | 81363 | 0.5867 | 5693.3 | 0.34421689 | 3340.25911 |
| 6 | 93940 | 0.6867 | 18270.3 | 0.47155689 | 12546.21501 |
| 6.8 | 91738 | 1.4867 | 16068.3 | 2.21027689 | 23888.74161 |
| 7.1 | 98273 | 1.7867 | 22603.3 | 3.19229689 | 40385.31611 |
| 7.9 | 101302 | 2.5867 | 25632.3 | 6.69101689 | 66303.07041 |
| 8.2 | 113812 | 2.8867 | 38142.3 | 8.33303689 | 110105.3774 |
| 8.7 | 109431 | 3.3867 | 33761.3 | 11.46973689 | 114339.3947 |
| 9 | 105582 | 3.6867 | 29912.3 | 13.59175689 | 110277.6764 |
| 9.5 | 116969 | 4.1867 | 41299.3 | 17.52845689 | 172907.7793 |
| 9.6 | 112635 | 4.2867 | 36965.3 | 18.37579689 | 158459.1515 |
| 10.3 | 112391 | 4.9867 | 36721.3 | 24.86717689 | 183118.1067 |
| 10.5 | 121872 | 5.1867 | 46202.3 | 26.90185689 | 239637.4694 |
|  |  |  |  |  |  |
| MEAN (xi) = 5.3133 | MEAN (yi) = 75669.67 |  |  | SUM = 223.55 | SUM = 2157216.13 |

w1 = (2157216.13)/ (223.55) = 9236.45

w0 = 75669.67 – (9236.451 \* 5.3133) = 26593.33

Final Equation => **salary = 26593.33 + 9236.45 \* years of experience**.

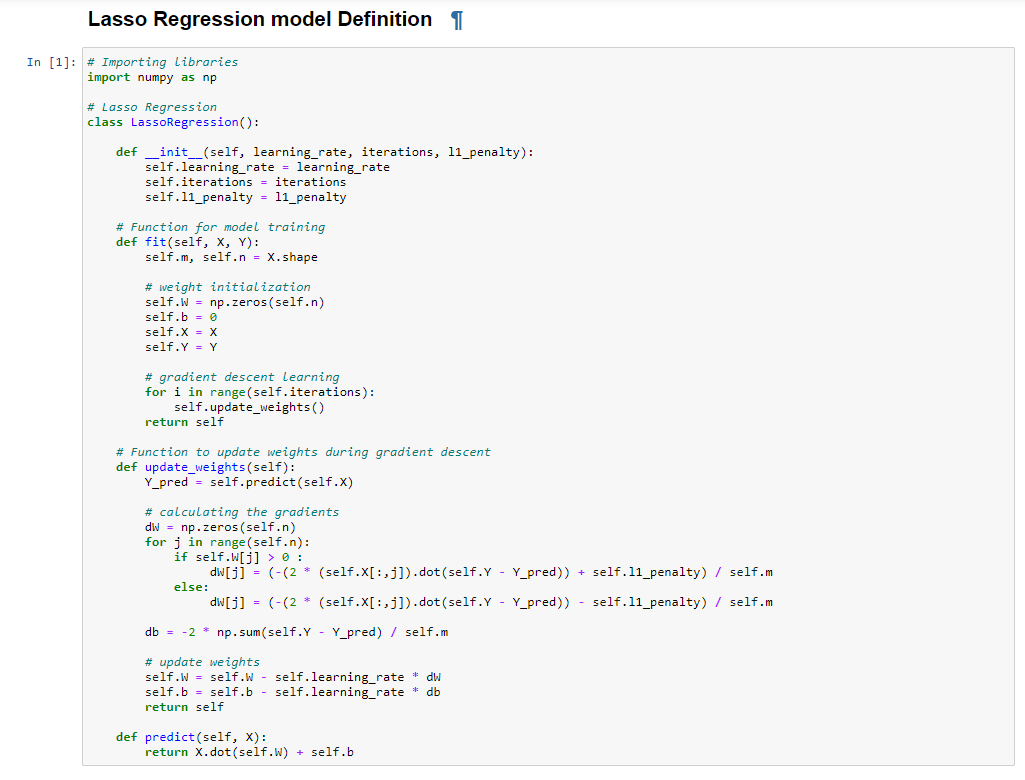
## **Implementation**

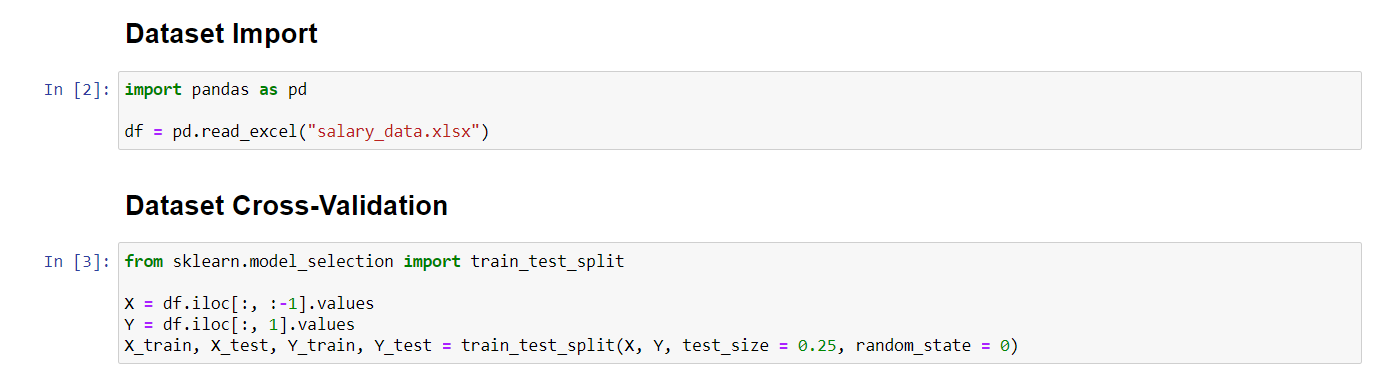
1. Python - Lasso Regression Implementation using user defined classes and functions.

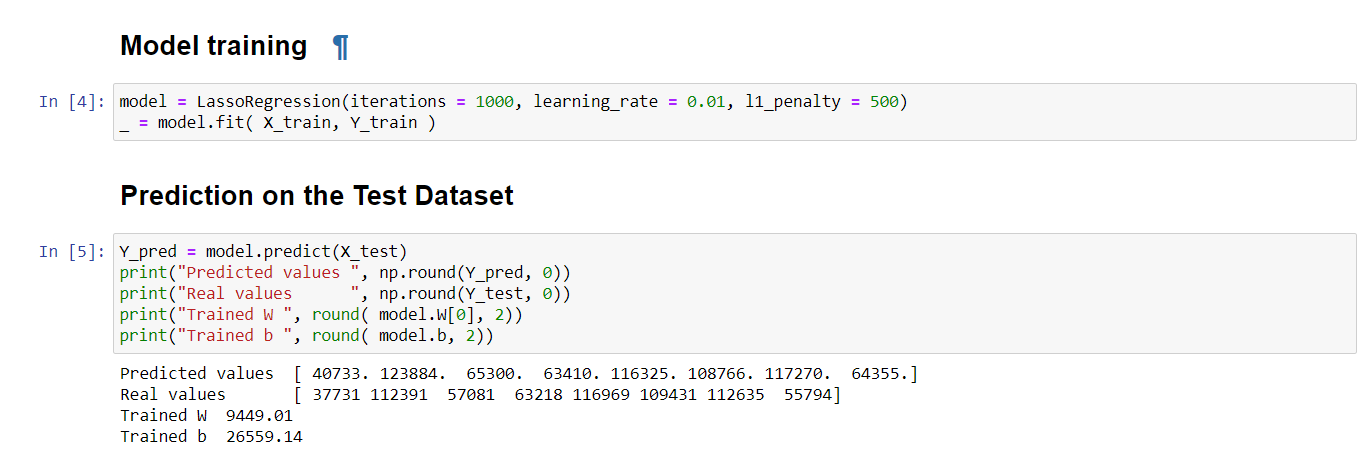
**Input:** salary\_data.xlsx, a spreadsheet which contains total numbers of years of experience and the present salary of 30 employees in a company.

|  |  |
| --- | --- |
| Years of Experience (xi) | salary (yi) |
| 1.1 | 39343 |
| 1.3 | 46205 |
| 1.5 | 37731 |
| 2 | 43525 |
| 2.2 | 39891 |
| 2.9 | 56642 |
| 3 | 60150 |
| 3.2 | 54445 |
| 3.2 | 64445 |
| 3.7 | 57189 |
| 3.9 | 63218 |
| 4 | 55794 |
| 4 | 56957 |
| 4.1 | 57081 |
| 4.5 | 61111 |
| 4.9 | 67938 |
| 5.1 | 66029 |
| 5.3 | 83088 |
| 5.9 | 81363 |
| 6 | 93940 |
| 6.8 | 91738 |
| 7.1 | 98273 |
| 7.9 | 101302 |
| 8.2 | 113812 |
| 8.7 | 109431 |
| 9 | 105582 |
| 9.5 | 116969 |
| 9.6 | 112635 |
| 10.3 | 112391 |
| 10.5 | 121872 |

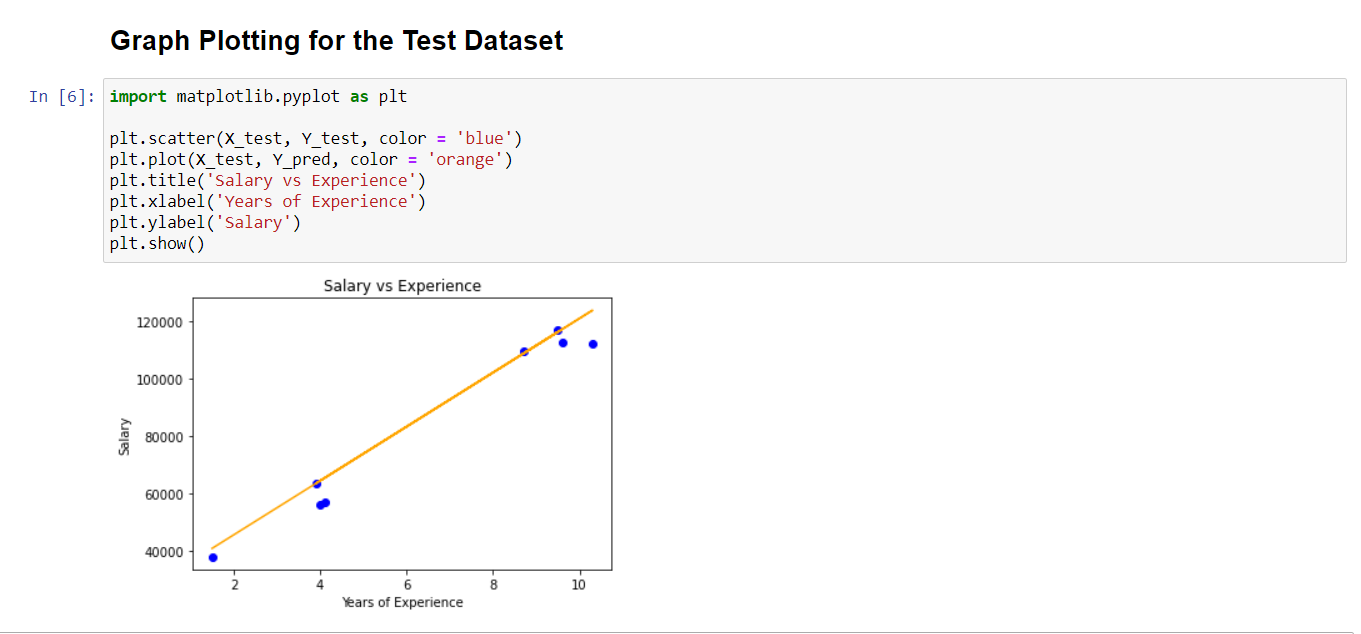
**Python code:**

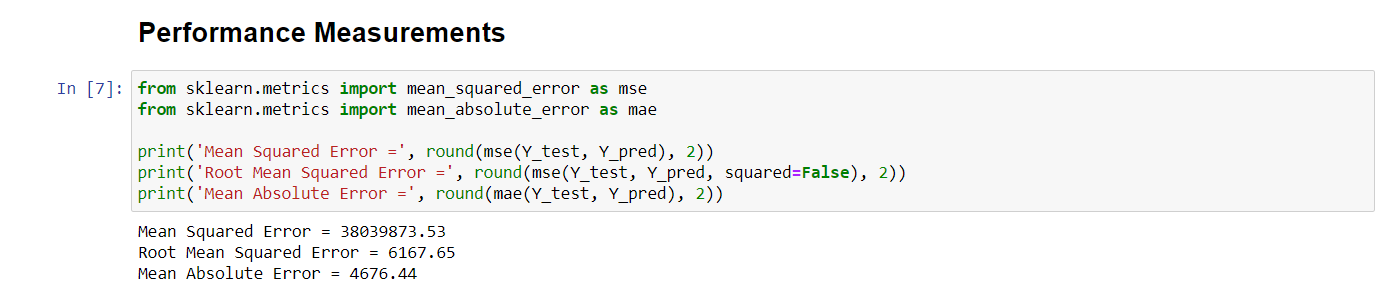
****

****

****

**Output:** The model plots the best fitting line for the given training dataset and predicts values for the test dataset. Performance metrics such as Mean Square Error, Root Mean Square Error, Mean Absolute Error are also calculated to give an idea of the performance of the Lasso Regression Model.

****

****

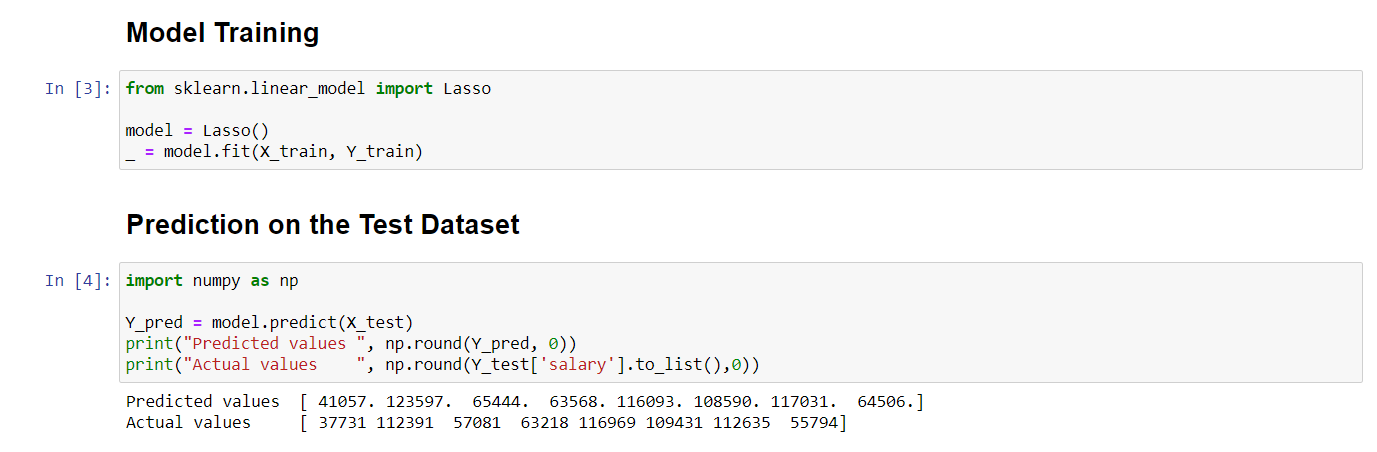
2. Python - Lasso Regression Implementation with scikit-libraries

**Input:** salary\_data.xlsx, a spreadsheet which contains total numbers of years of experience and the present salary of 30 employees in a company.

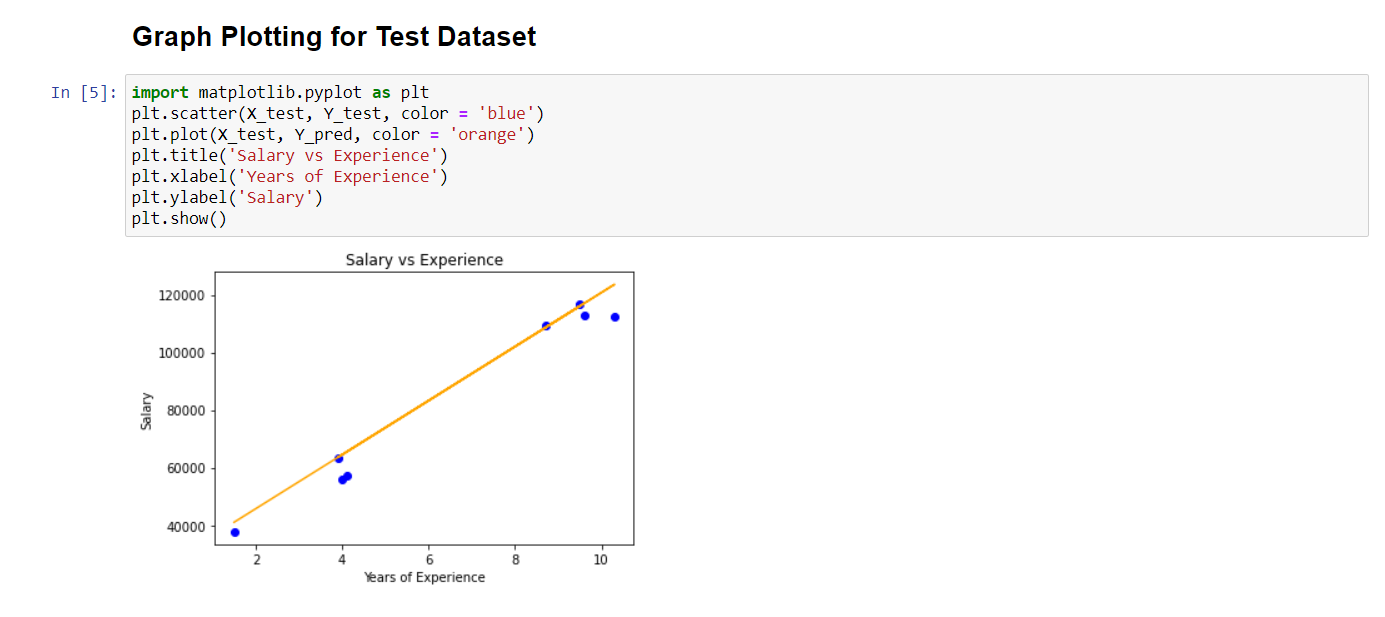
|  |  |
| --- | --- |
| Years of Experience (xi) | salary (yi) |
| 1.1 | 39343 |
| 1.3 | 46205 |
| 1.5 | 37731 |
| 2 | 43525 |
| 2.2 | 39891 |
| 2.9 | 56642 |
| 3 | 60150 |
| 3.2 | 54445 |
| 3.2 | 64445 |
| 3.7 | 57189 |
| 3.9 | 63218 |
| 4 | 55794 |
| 4 | 56957 |
| 4.1 | 57081 |
| 4.5 | 61111 |
| 4.9 | 67938 |
| 5.1 | 66029 |
| 5.3 | 83088 |
| 5.9 | 81363 |
| 6 | 93940 |
| 6.8 | 91738 |
| 7.1 | 98273 |
| 7.9 | 101302 |
| 8.2 | 113812 |
| 8.7 | 109431 |
| 9 | 105582 |
| 9.5 | 116969 |
| 9.6 | 112635 |
| 10.3 | 112391 |
| 10.5 | 121872 |

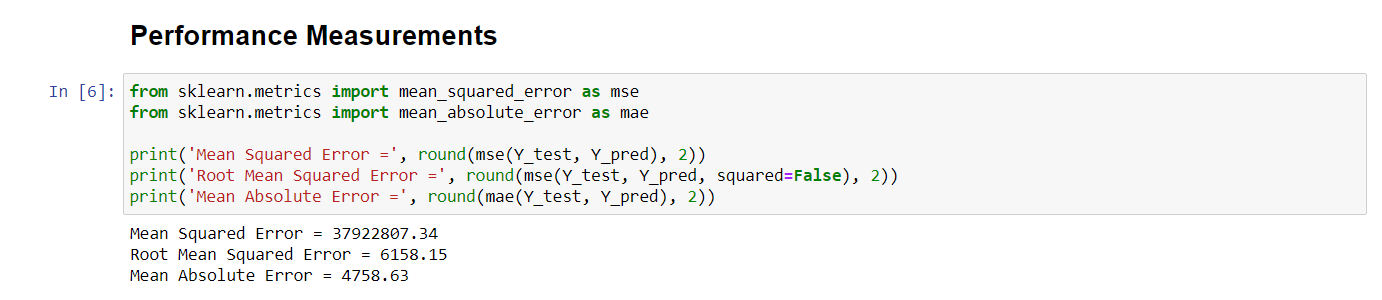
**Python code:**

****

****

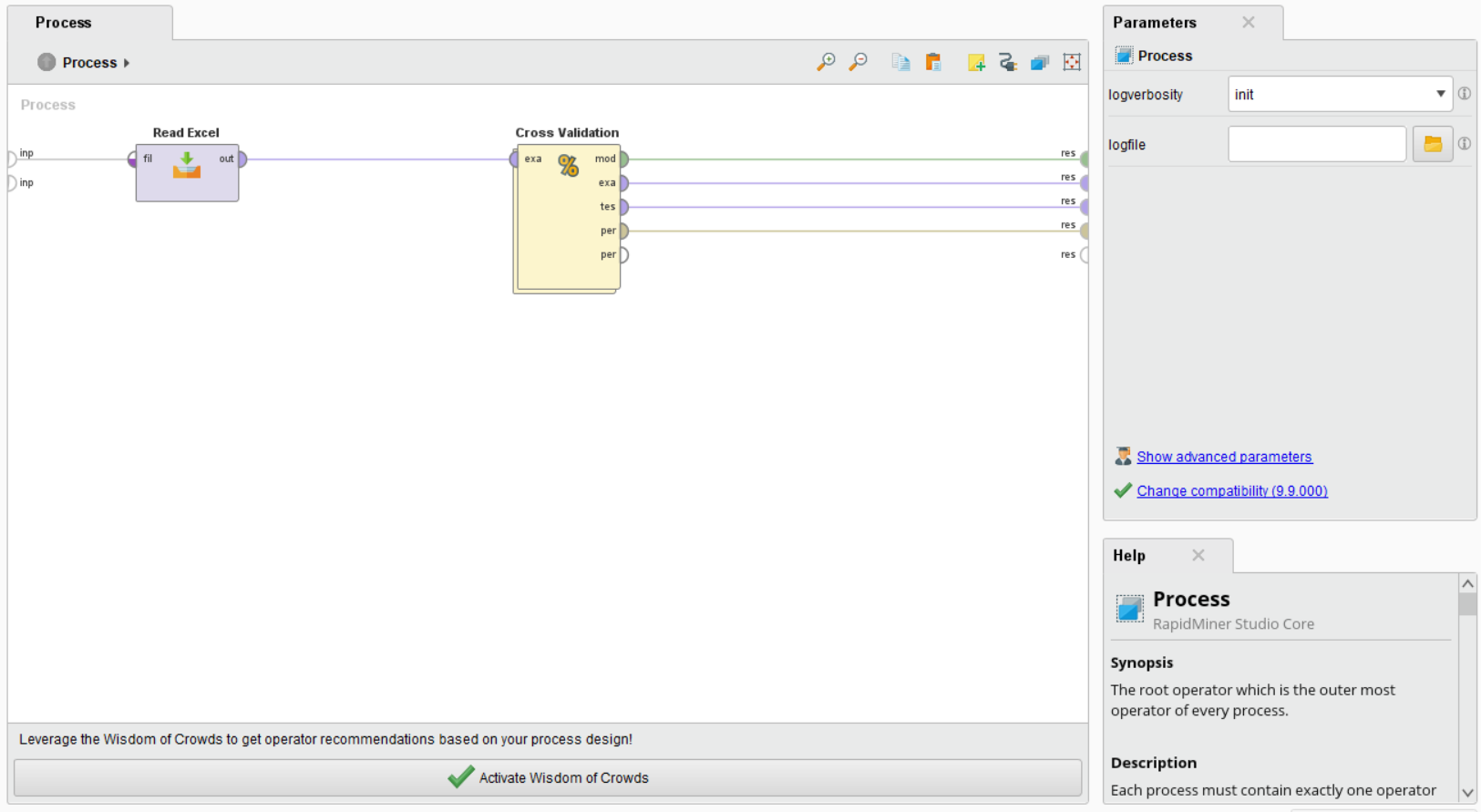
**Output:** The model plots the best fitting line for the given training dataset and predicts values for the test dataset. Performance metrics such as Mean Square Error, Root Mean Square Error, Mean Absolute Error are also calculated to give an idea of the performance of the Lasso Regression Model.

****

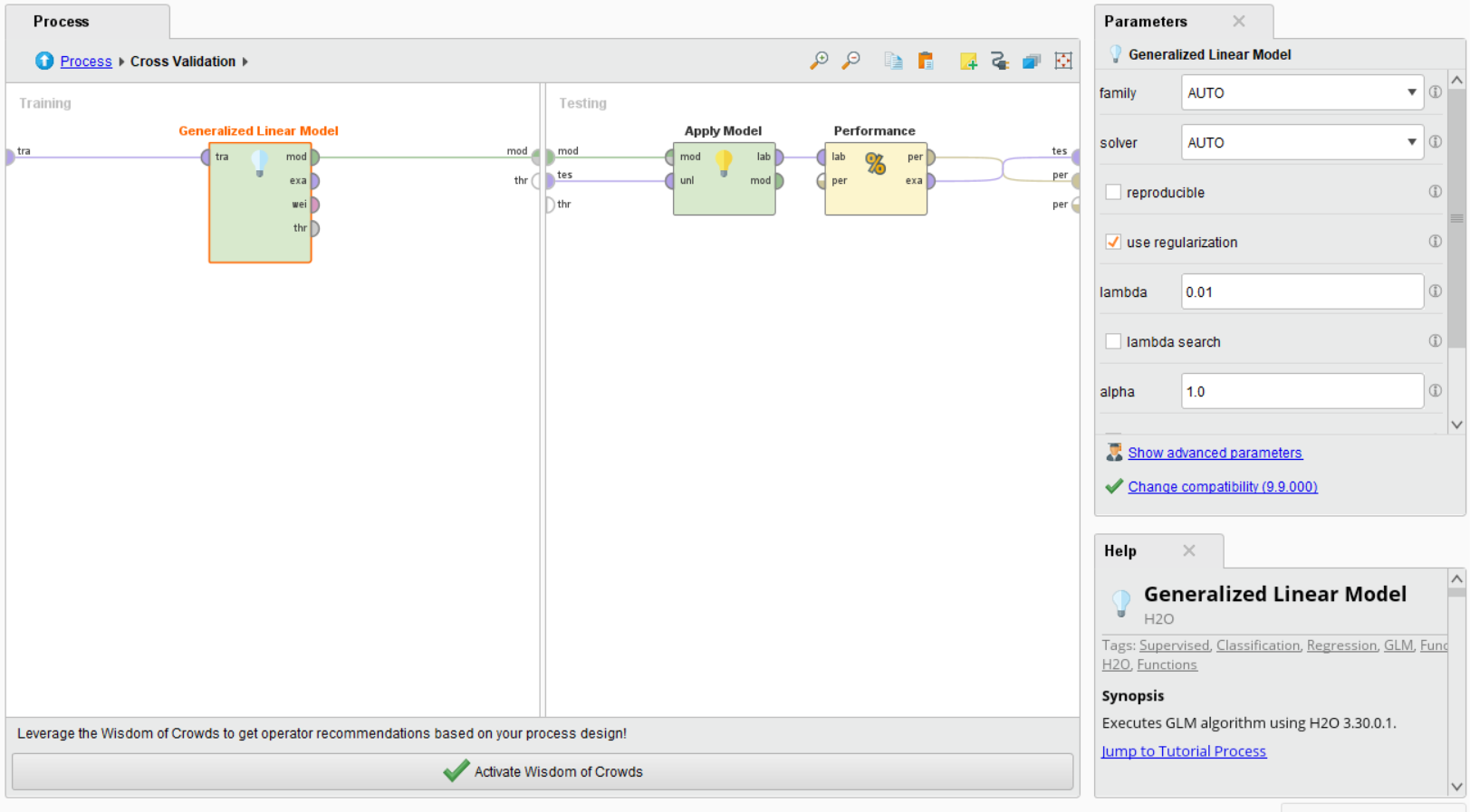
****

1. RapidMiner Implementation (Screenshots)

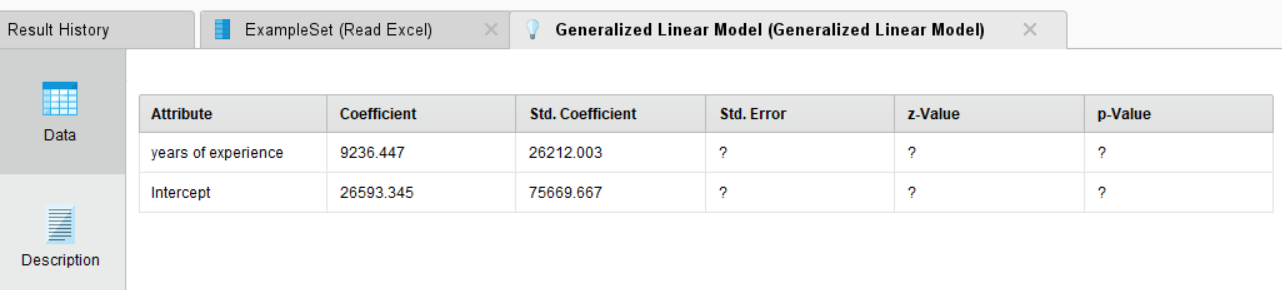
**Overall Model:**



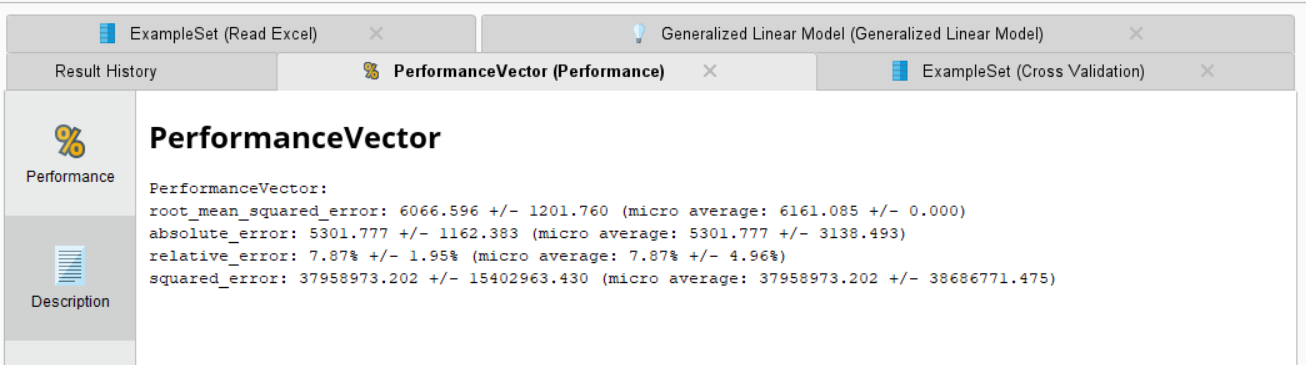
**The Lasso Regression model after Cross-Validation of Data:**

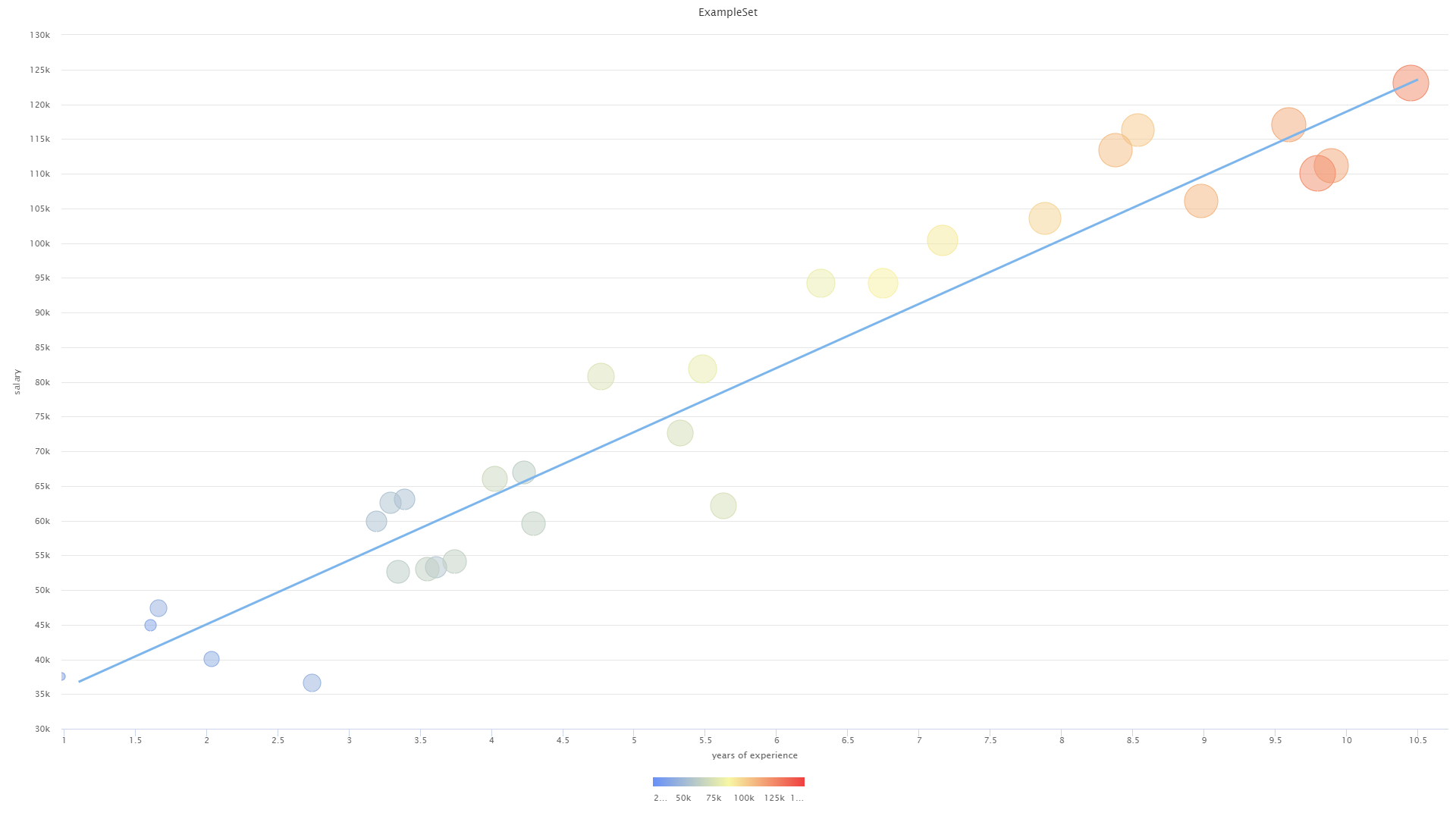


**Stats from the Generalized Linear Model (Lasso Regression):**



**Performance Measurements from the Lasso Regression Model:**



**Graph Plotting for the Generalized Linear Model (Lasso Regression):**

## **References**

* Beyer, W. H. CRC Standard Mathematical Tables, 31st ed. Boca Raton, FL: CRC Press, 2002.
* Agresti A. (1990) Categorical Data Analysis. John Wiley and Sons, New York.
* Katz, S.; et al., eds. (2006), Encyclopedia of Statistical Sciences, Wiley.
* Wheelan, C. (2014). Naked Statistics. W. W. Norton & Company.