Intro to ML Project 1 Report

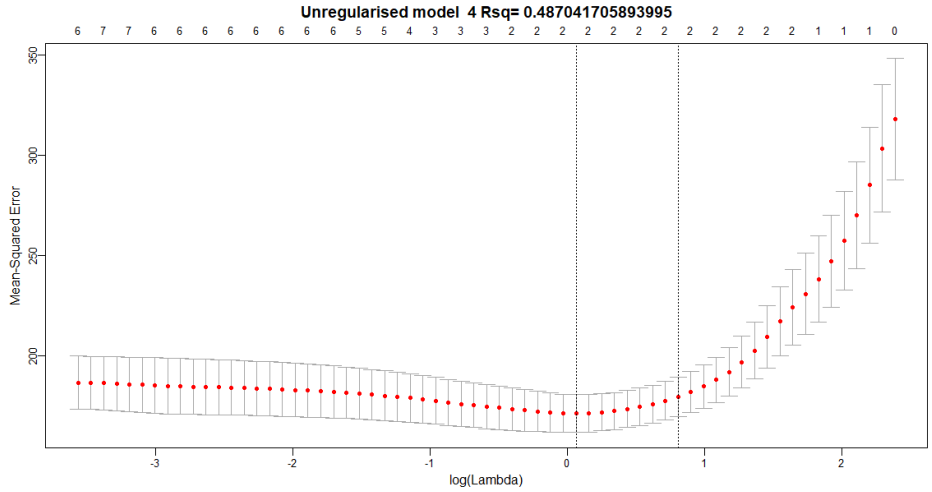
Task-0

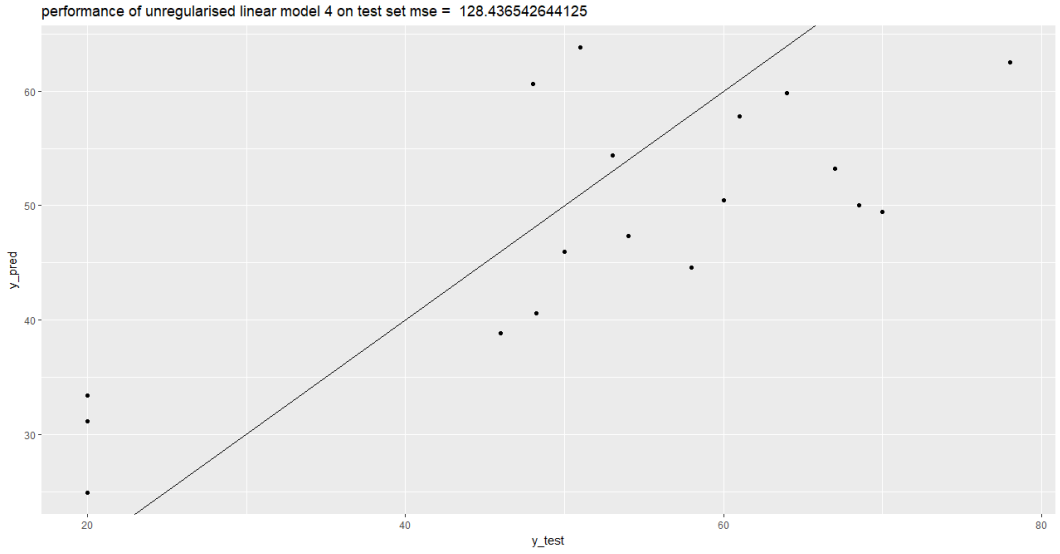
- Environment: Windows 10
- Time Spent: 1 Week
- Programming Language: R
- Libraries Used: Glmnet, caTools, ggplot2
- Collaboration: Ankesh N .Bhoi(ankeshni@buffalo.edu), Shashank Raghunath(sraghuna@buffalo.edu)

Task-1

Unregularized Regression

- The mean **R-squared** for **10** iterations is **0.517** with **SD** of **0.029** and the mean **MSE** is **209** with **SD** of **51.**
- MSE of models 1 to 10 on test set:
  *274.9655, 196.7950, 284.0492, **128.4365**, 173.1720, 187.8367, 233.6112, 251.6670, 206.4925, 157.6003*
- R-square values of models 1 to 10 on training set:
  0.5778609, 0.5271576, 0.5377060, 0.4870417, 0.4867348, 0.5039519, 0.5271875, 0.5341908, 0.5058729, 0.4850152
- Best Unregularised Model is model 4 as it has min MSE on the test set.





Regularized Regression: Ridge

- The mean **R-squared** for 10 iterations is **0.477** with **SD** of **0.048** and the mean **MSE** is **206** with **SD** of **50.**
- MSE of models 1 to 10 on test set:

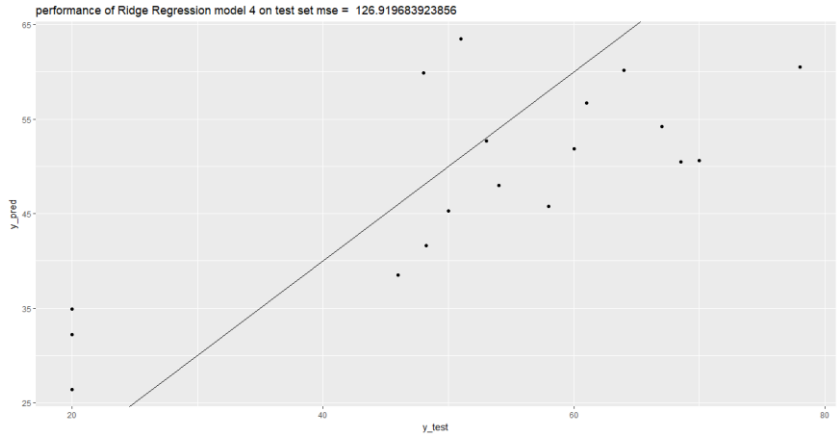
  *260.1241, 198.7340, 287.2600, **126.9197**, 172.0712, 183.9118, 221.1980, 257.4454, 202.4738, 153.1231*

- R-square values of models 1 to 10 on training set:

  0.5464303, 0.4927091, 0.5184776, 0.4706722, 0.4545269, 0.4817757, 0.4720269, 0.5019988, 0.4785214, 0.362 7046
- Best regularized Ridge Model is model 4 as it has min MSE on the test set with regularization parameter 0.97.

```
> cvfit_lasso$lambda.min
[1] 0.9707253
```

- The performance of Ridge *regression* is better than Unregularised regression as evident by the lower average mean squared error of 206.



Regularized Regression: Lasso

- The mean **R-squared** for 10 iterations is **0.5** with **SD** of **0.031** and the mean **MSE** is **195** with **SD** of **47**
- MSE of models 1 to 10 on test set:

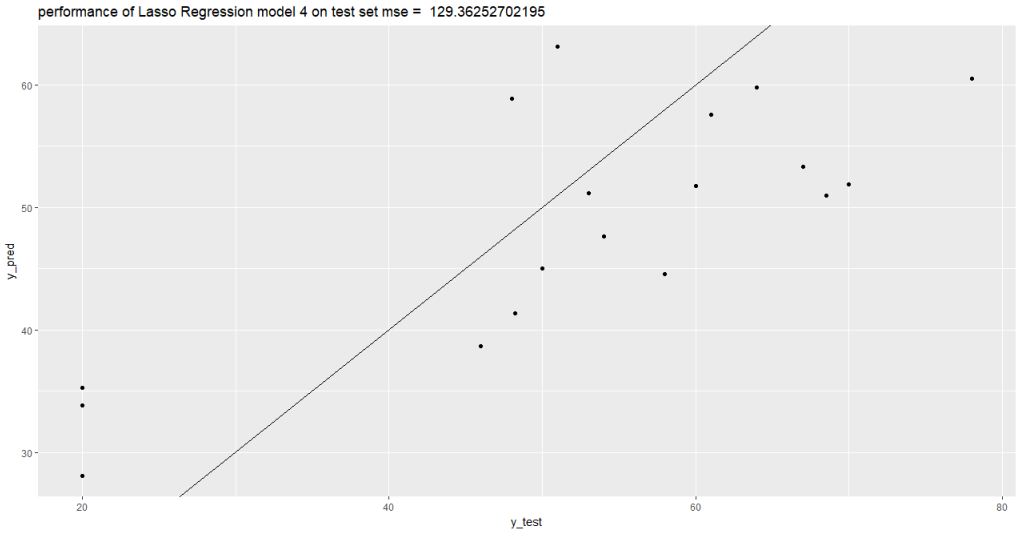
  *244.3346, 195.0800, 271.7668, **129.3625**, 169.9484, 163.4074, 198.0553, 252.1878, 180.0836, 148.2110*
- R-square values of models 1 to 10 on training set:

  0.5595370, 0.5164467, 0.5314832, 0.4755914, 0.4790928, 0.4860902, 0.5025553, 0.5208867, 0.4827065, 0.455 5893
- Best regularized Lasso Model is model 4 as it has min MSE of 129 on the test set with regularization parameter 1.3.

```
 > cvfit_ridge$lambda.min
[1] 1.313438
```

- The average performance of Lasso regression is better than Unregularised regression and Ridge Regression as evident by the lower average mean squared error of 195.
- Number of coefficients used by lasso are: 2, Slag and Water as evident below

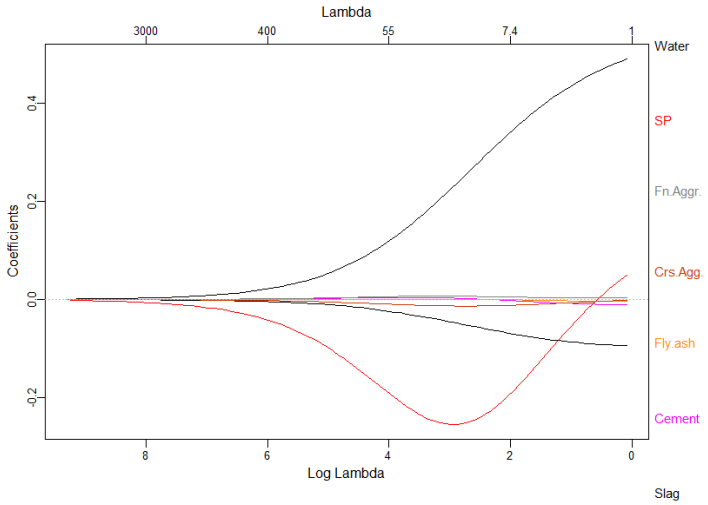
```
> coefficients(cvfit_lasso)
8 x 1 sparse Matrix of class "dgCMatrix"
1
(Intercept)  -22.22958289
Cement          .
Slag           -0.04544001
Fly.ash         .
Water           0.37935727
SP              .
Coarse.Aggr.    .
Fine.Aggr.      .
```



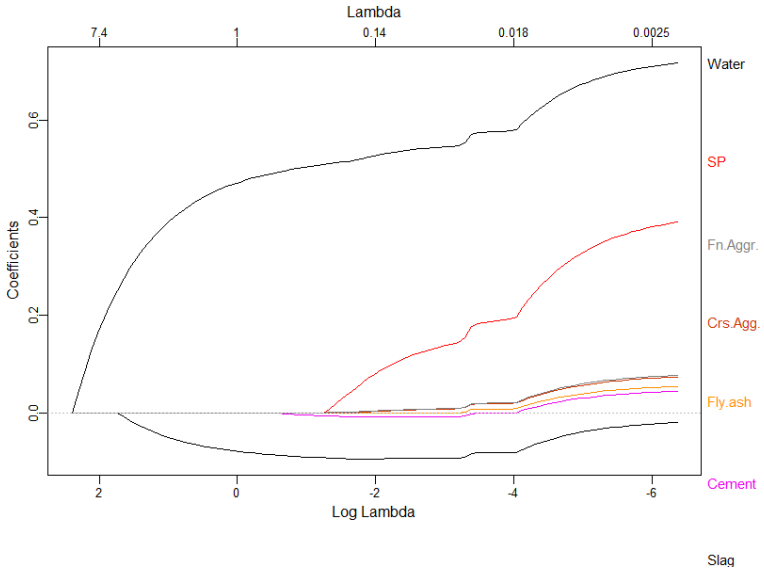
Task-2

Regularization Paths

Ridge Model-4



Lasso Model-4



Extra Task

	X	X^2	X^3	Log X
mean.mse_unregularised	209.4626022	196.3988727	202.1088	214.6087
mean.mse_ridge	206.3261042	190.9561916	195.0809	208.3848
mean.mse_lasso	195.2437458	189.918766612805	191.7158	217.849
sd.mse_unregularised	51.01966845	28.59639913	34.42844	32.31776
sd.mse_ridge	50.73141443	31.72755623	30.38875	57.19006
sd.mse_lasso	47.08372171	28.46161358	28.48551	41.07474
mean.rsq_unregularised	0.517271919	0.56080336	0.543499	0.575996
mean.rsq_ridge	0.477984359	0.505799992	0.456478	0.470362
mean.rsq_lasso	0.500997907	0.538383639	0.521571	0.573876
sd.rsq_unregularised	0.029373273	0.017027908	0.020552	0.01809
sd.rsq_ridge	0.477984359	0.505799992	0.456478	0.470362
sd.rsq_lasso	0.500997907	0.538383639	0.521571	0.573876

Analysis

1. Regularized Lasso model on an average gives minimum mean square error when X^2 is used as matrix of features instead of X.
2. The R-squared value of Unregularised model increases significantly when X^2 and Log X are used.
3. Standard Deviation in R-squared values in Unregularised regression is minimum in all cases.