CE 475 Machine Learning Final Project Report

Created by Ögeday ÖZTOPRAK

20150602028

Identification of the Problem

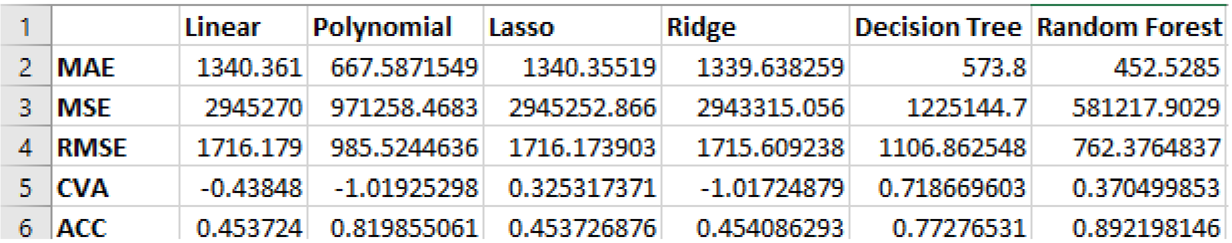
We have a dataset that include 120 data points on the columns named x1,x2,x3,x4,x5,x6 and also we have another column named Y and it only contains 100 data points. The mission was about deciding which regression model is better for estimating last 20 rows of the Y. I have used X columns as my train data.

Methodology

I have believed that deciding best model for this problem is trying several estimation methods and getting some result from them.Then comparing the results of the regression models and selecting the best model would be the best for estimating last 20 rows of Y.I had necessary researched about this calculations then I have realized that I need to use a selection algorithm before starting the process.I took a look at the sklearn library and another sources then found few selection algorithms named Pipeline,Tree Based Feature Selection,L1 Based Feature Selection and Recursive Feature Elimination,Feature Selection, so I decided to use Feature Elimination for my calculations.According to results of the algorithm my most efficient first four columns for these calculations are; **x1,x2,x3 and x5.**I have used first 100 rows of the columns till to find the best estimation model.I have tried to use **Linear Regression,Polynomial Regression,Lasso Regression,Decision Tree,Ridge Regression and Random Forest Regression**. I collected their **Mean Absolute Error, Mean Squared Error,Root Mean Squared Error, Cross Validation Accuracy and Accuracy** results by using metrics library of the Sklearn.According the comparison of these results especially about the mean absolute error, Random Forest Regression is the best regression model to estimate 20 values for the Y.

Results

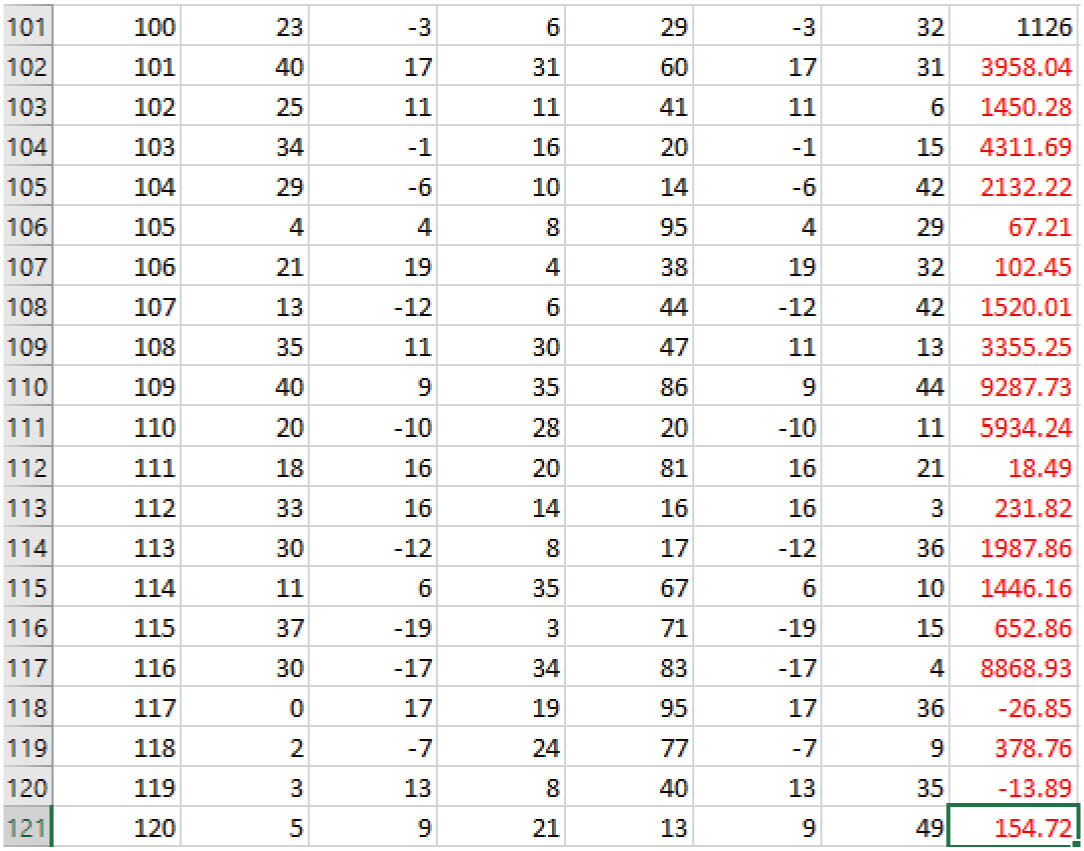
(Original version of the results attached to the report file.)



1		Linear	Polynomial	Lasso	Ridge	Decision Tree	Random Forest
2	MAE	1340.361	667.5871549	1340.35519	1339.638259	573.8	452.5285
3	MSE	2945270	971258.4683	2945252.866	2943315.056	1225144.7	581217.9029
4	RMSE	1716.179	985.5244636	1716.173903	1715.609238	1106.862548	762.3764837
5	CVA	-0.43848	-1.01925298	0.325317371	-1.01724879	0.718669603	0.370499853
6	ACC	0.453724	0.819855061	0.453726876	0.454086293	0.77276531	0.892198146

According to these calculations it is obvious that using Random Forest Regression to estimate the last 20 Y values is the best model because Random Forest has smaller error values mostly especially Mean Absolut Error has big difference and also accuracy of the Random Forest Regression is a way better than the other models that I have used.

Estimations via Random Forest Regression



101	100	23	-3	6	29	-3	32	1126
102	101	40	17	31	60	17	31	3958.04
103	102	25	11	11	41	11	6	1450.28
104	103	34	-1	16	20	-1	15	4311.69
105	104	29	-6	10	14	-6	42	2132.22
106	105	4	4	8	95	4	29	67.21
107	106	21	19	4	38	19	32	102.45
108	107	13	-12	6	44	-12	42	1520.01
109	108	35	11	30	47	11	13	3355.25
110	109	40	9	35	86	9	44	9287.73
111	110	20	-10	28	20	-10	11	5934.24
112	111	18	16	20	81	16	21	18.49
113	112	33	16	14	16	16	3	231.82
114	113	30	-12	8	17	-12	36	1987.86
115	114	11	6	35	67	6	10	1446.16
116	115	37	-19	3	71	-19	15	652.86
117	116	30	-17	34	83	-17	4	8868.93
118	117	0	17	19	95	17	36	-26.85
119	118	2	-7	24	77	-7	9	378.76
120	119	3	13	8	40	13	35	-13.89
121	120	5	9	21	13	9	49	154.72

The numbers which has red as a color are the estimated values between 101-121.

Conclusion

To sum up,I have tried to use several regression methods as much as possible to me.I have used our lessons and laboratory sections as a guide and source. Even if I can not learn every concept about Machine Learning , this project encourage me about the Python language and digital way of the algebra. I also got helped from Sklearn libraries but I have reviewed ' my course notes and short description about the regressions that I have used before using the libraries.I hope I could be able to find best model for the estimations. I hope I found the best model for the estimations.