Project 1

Manager: Brendan McDonnel
Masum Billah
Madhu Chencharapu
Gabriel Loos
Roshan Ravichandran

February 2026

1 Introduction

2 Auto MPG

2.1 EDA

2.2 Quality of Fit for each model

Table 1: Auto MPG In-Sample QoF Comparison

Metric	Regression	Ridge	Lasso	Sqrt	log1p
rSq	0.809255	0.776580	0.809163	0.835138	0.849102
rSqBar	0.806283	0.772507	0.806189	0.832569	0.846751
sst	23819.0	23819.0	23819.0	23819.0	23819.0
sse	4543.35	5321.63	4545.54	3926.85	3594.23
sde	3.40878	3.68883	3.40888	3.16757	3.02514
mse0	11.5902	13.5756	11.5958	10.0175	9.16895
rmse	3.40443	3.68451	3.40526	3.16504	3.02803
mae	2.61826	2.79509	2.61703	2.34675	2.18422
smape	12.0589	65.4181	11.9861	10.2046	9.32001
m	392.000	392.000	392.000	392.000	392.000
dfr	6.00000	7.00000	6.00000	6.00000	6.00000
df	385.000	384.000	385.000	385.000	385.000
fStat	272.234	190.677	272.072	325.048	361.067
aic	-1022.45	-1051.45	-1022.55	-993.873	-976.527
bic	-994.656	-1019.68	-994.750	-966.074	-948.728

Table 2: Auto MPG Out-of-Sample QoF Comparison

Metric	Regression	Ridge	Lasso	Sqrt	log1p
rSq	0.822842	0.797491	0.822903	0.846480	0.852864
rSqBar	0.820081	0.793799	0.820143	0.844088	0.850571
sst	4731.23	4731.23	4731.23	4731.23	4731.23
sse	838.174	958.118	837.889	726.337	696.133
sde	3.29026	3.51709	3.28969	3.05724	2.97969
mse0	10.7458	12.2836	10.7422	9.31202	8.92478
rmse	3.27808	3.50479	3.27752	3.05156	2.98744
mae	2.48735	2.62052	2.48643	2.11846	1.98665
smape	11.8858	61.9272	11.8808	9.00068	8.28831
m	78.0000	78.0000	78.0000	78.0000	78.0000
dfr	6.00000	7.00000	6.00000	6.00000	6.00000
df	385.000	384.000	385.000	385.000	385.000
fStat	298.034	216.030	298.158	353.804	371.939
aic	-189.284	-192.500	-189.271	-183.700	-182.048
bic	-172.787	-173.646	-172.774	-167.203	-165.551

Table 3: Scalation - Auto MPG Linear Regression CV

Name	num	min	max	mean	stdev	interval
rSq	5	0.788	0.823	0.798	0.014	0.018
rSqBar	5	0.785	0.820	0.795	0.014	0.018
sst	5	3962.818	5671.580	4700.481	620.767	770.935
sse	5	824.554	1176.435	950.494	142.696	177.215
sde	5	3.177	3.738	3.431	0.226	0.281
mse0	5	10.571	15.083	12.186	1.829	2.272
rmse	5	3.251	3.884	3.483	0.256	0.318
mae	5	2.487	2.850	2.689	0.151	0.188
smape	5	11.886	12.905	12.372	0.427	0.530
m	5	78.000	78.000	78.000	0.000	0.000
dfr	5	6.000	6.000	6.000	0.000	0.000
df	5	385.000	385.000	385.000	0.000	0.000
fStat	5	239.054	298.034	254.430	24.517	30.448
aic	5	-205.110	-188.647	-194.539	6.676	8.291
bic	5	-188.613	-172.150	-178.042	6.676	8.291

2.3 Regression

2.4 Ridge

2.5 Lasso

2.6 Sqrt

2.7 log1p

3 Housing Prices

3.1 EDA

3.2 Quality of Fit for each model

3.3 Regression

3.4 Ridge

3.5 Lasso

3.6 Sqrt

3.7 log1p

Table 4: Statsmodels - Auto MPG Linear Regression CV

Name	In-num folds	min	max	mean	stdev
rSq	5	0.7654	0.8282	0.8010	0.0216
rSqBar	5	0.7458	0.8139	0.7845	0.0234
sst	5	4032.2061	5792.1365	4724.2751	617.8288
sse	5	745.1709	1359.0577	947.8346	213.7052
sde	5	3.2171	4.3446	3.5980	0.3912
mse0	5	51.0406	74.2582	60.2918	8.1328
rmse	5	3.2171	4.3446	3.5980	0.3912
mae	5	2.5039	3.1904	2.6786	0.2601
smape	5	11.2379	14.1325	12.3805	0.9795

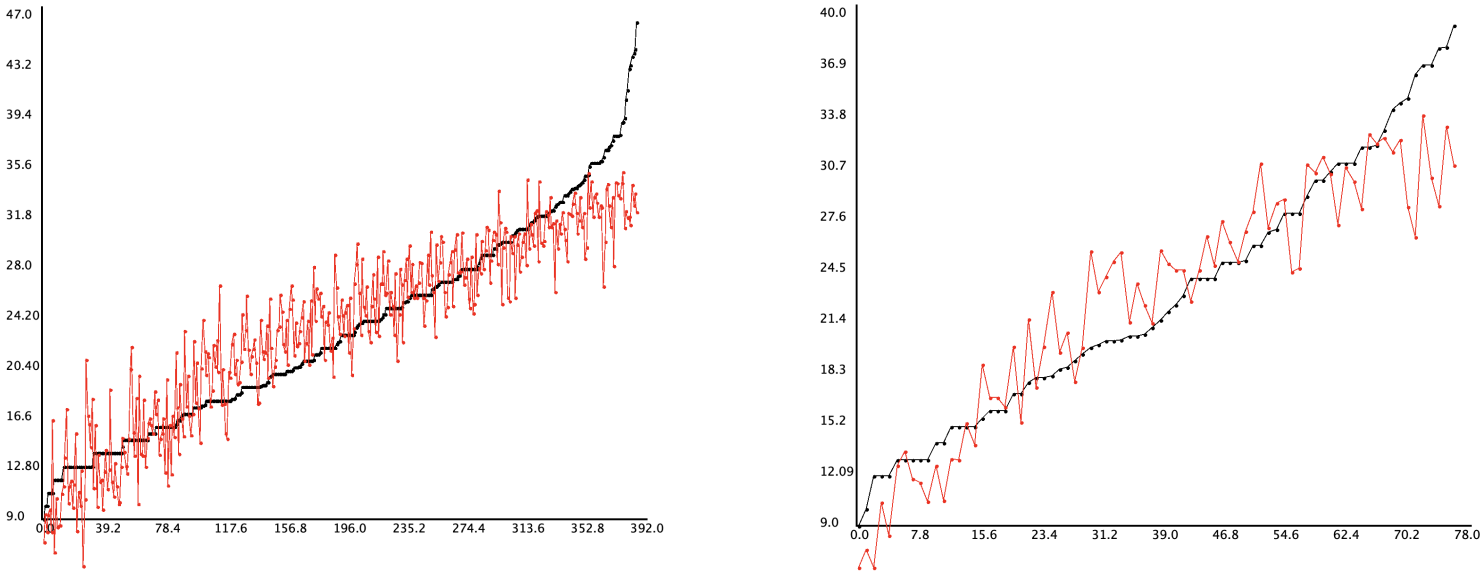


Figure 1: Scalation - Auto MPG Regression
Left: In Sample Predictions
Right: 80-20 Out of Sample Predictions
yy black/actual vs. yp red/predicted

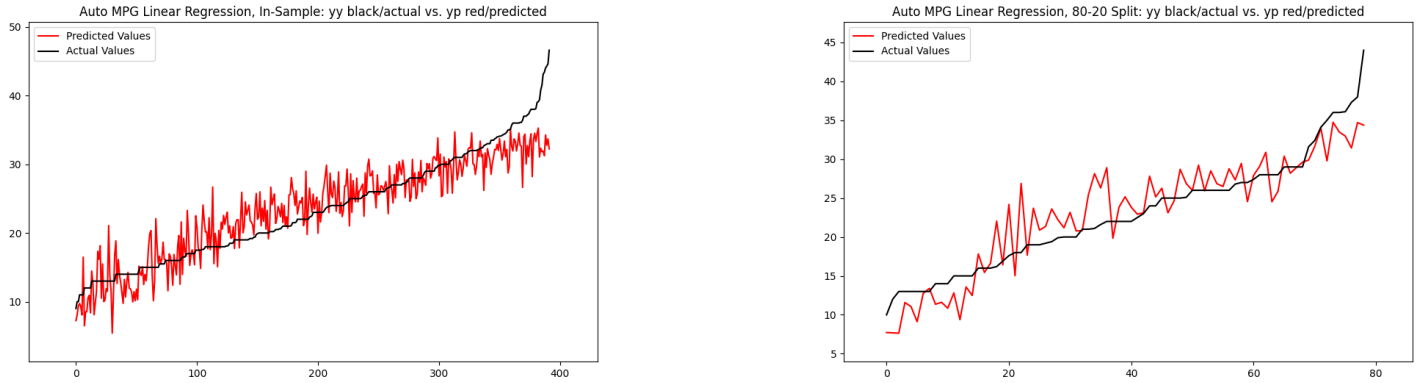


Figure 2: Statsmodels - Auto MPG Regression
Left: In Sample Predictions
Right: 80-20 Out of Sample Predictions
yy black/actual vs. yp red/predicted

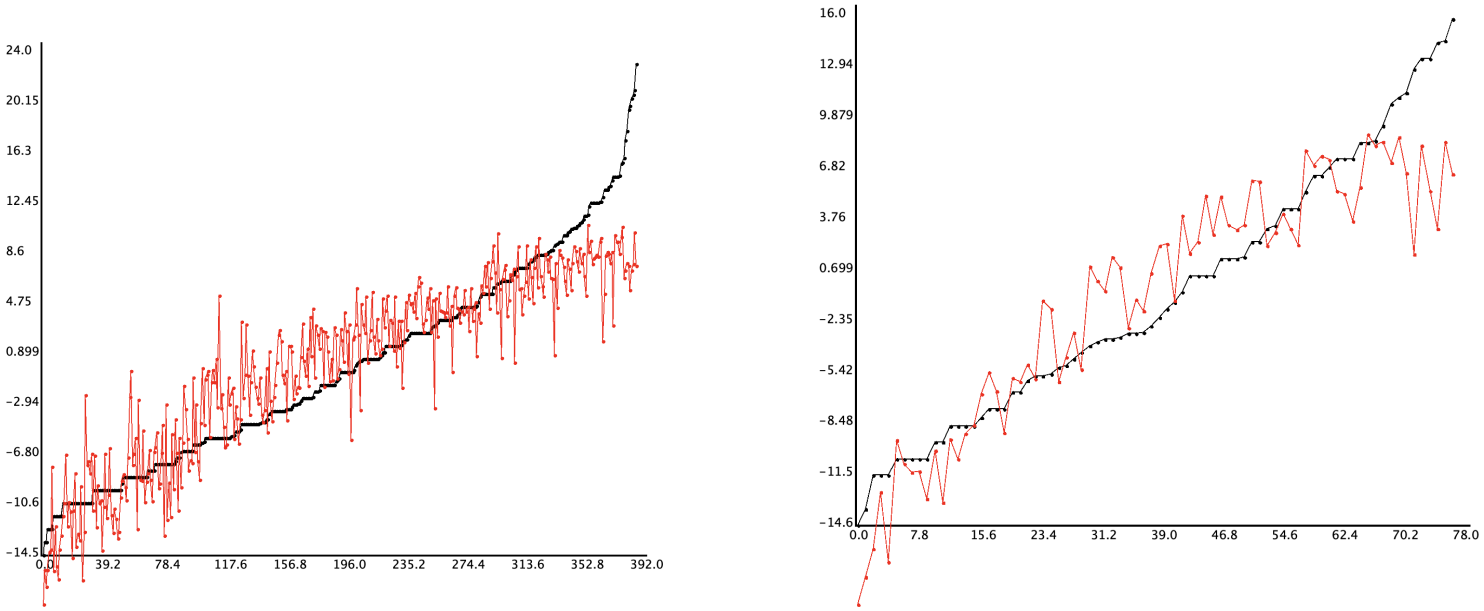


Figure 3: Scalation - Auto MPG Ridge
Left: In Sample Predictions
Right: 80-20 Out of Sample Predictions
yy black/actual vs. yp red/predicted

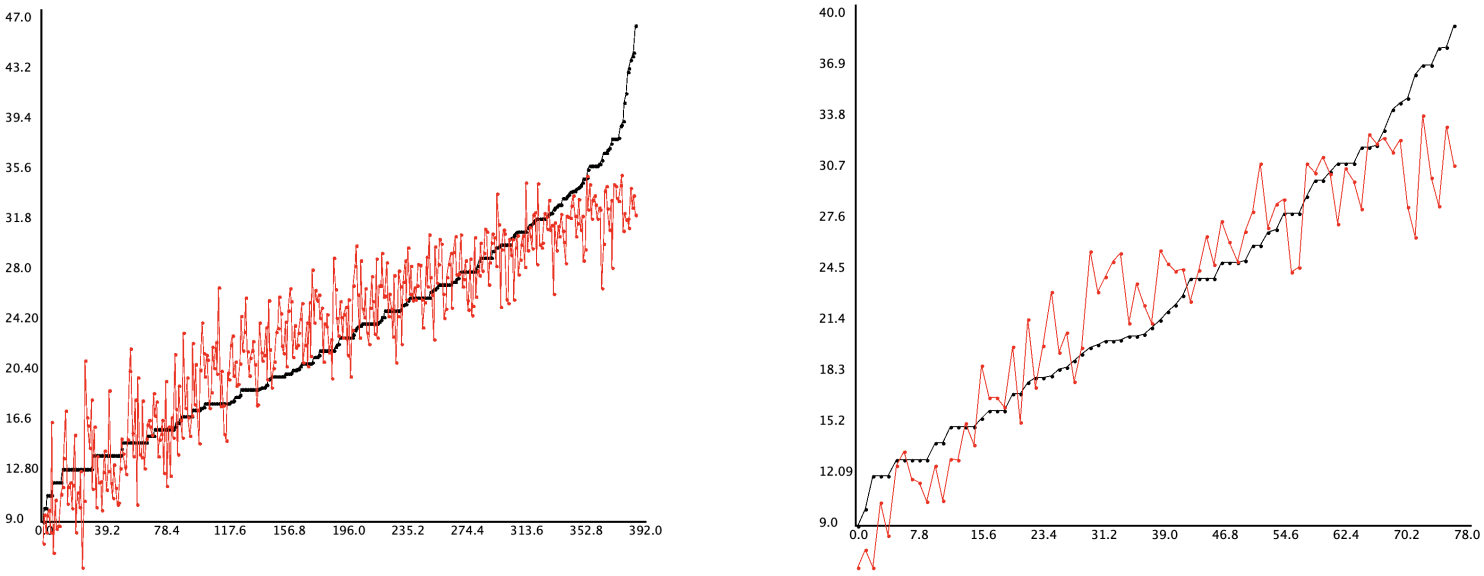


Figure 4: Scalation - Auto MPG Lasso
Left: In Sample Predictions
Right: 80-20 Out of Sample Predictions
yy black/actual vs. yp red/predicted

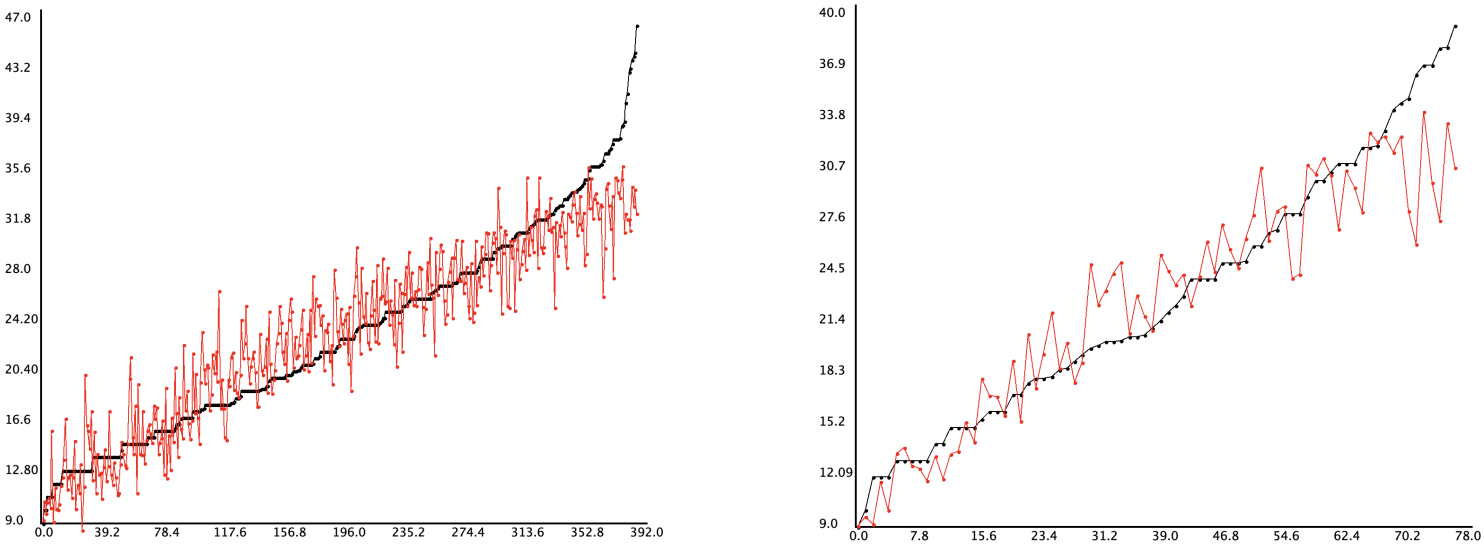


Figure 5: Scalation - Auto MPG Sqrt
Left: In Sample Predictions
Right: 80-20 Out of Sample Predictions
yy black/actual vs. yp red/predicted

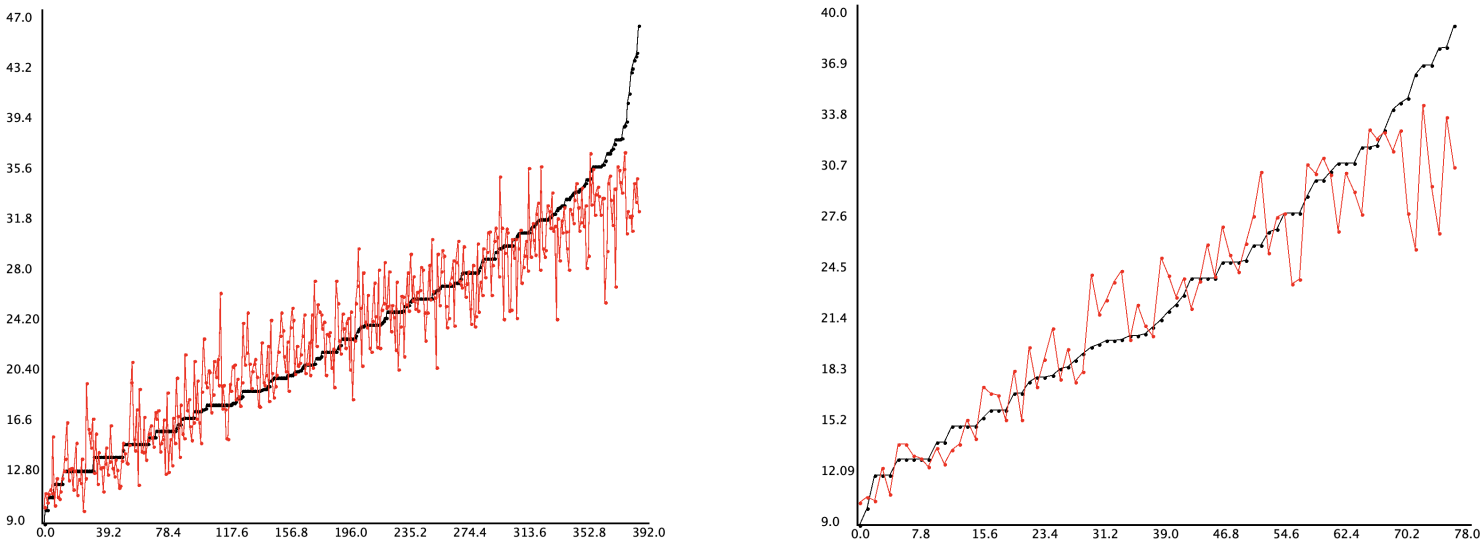


Figure 6: Scalation - Auto MPG log1p
Left: In Sample Predictions
Right: 80-20 Out of Sample Predictions
yy black/actual vs. yp red/predicted

Table 5: Housing Prices In-Sample QoF Comparison

Metric	Regression	Ridge	Lasso	Sqrt	log1p
rSq	0.998516	0.987398	0.998516	0.986152	0.922307
rSqBar	0.998507	0.987309	0.998507	0.986068	0.921838
sst	6.42325e+13	6.42325e+13	6.42325e+13	6.42325e+13	6.42325e+13
sse	9.53030e+10	8.09438e+11	9.53030e+10	8.89504e+11	4.99039e+12
sde	9767.21	28463.5	9767.21	29836.4	70604.5
mse0	9.53030e+07	8.09438e+08	9.53030e+07	8.89504e+08	4.99039e+09
rmse	9762.32	28450.6	9762.32	29824.5	70642.7
mae	7747.66	24147.7	7747.66	24309.3	53070.5
smape	1.57791	23.5576	1.57791	4.86164	9.22319
m	1000.00	1000.00	1000.00	1000.00	1000.00
dfr	6.00000	7.00000	6.00000	6.00000	6.00000
df	993.000	992.000	993.000	993.000	993.000
fStat	111378	11103.9	111378	11785.5	1964.69
aic	-10591.2	-11658.9	-10591.2	-11708.0	-12570.3
bic	-10556.9	-11619.6	-10556.9	-11673.7	-12536.0

Table 6: Housing Prices Out-of-Sample QoF Comparison

Metric	Regression	Ridge	Lasso	Sqrt	log1p
rSq	0.998649	0.989193	0.998649	0.984018	0.906860
rSqBar	0.998641	0.989117	0.998641	0.983921	0.906297
sst	1.33700e+13	1.33700e+13	1.33700e+13	1.33700e+13	1.33700e+13
sse	1.80638e+10	1.44485e+11	1.80638e+10	2.13678e+11	1.24528e+12
sde	9501.56	26880.4	9501.56	32757.9	78673.0
mse0	9.03190e+07	7.22424e+08	9.03189e+07	1.06839e+09	6.22641e+09
rmse	9503.63	26877.9	9503.63	32686.3	78907.6
mae	7484.48	22419.4	7484.48	26186.8	58148.8
smape	1.60847	19.6544	1.60847	5.19788	9.91629
m	200.000	200.000	200.000	200.000	200.000
dfr	6.00000	7.00000	6.00000	6.00000	6.00000
df	993.000	992.000	993.000	993.000	993.000
fStat	122330	12971.9	122330	10189.9	1611.39
aic	-2101.67	-2307.60	-2101.67	-2348.73	-2525.00
bic	-2078.59	-2281.21	-2078.59	-2325.64	-2501.91

Table 7: Scalation - House Price Linear Regression CV

Name	num folds	min	max	mean	stdev	interval
rSq	5	0.998	0.999	0.998	0.000	0.000
rSqBar	5	0.998	0.999	0.998	0.000	0.000
sst	5	11805100457351.200	13369989427686.710	12829460198984.865	606159535374.424	752793908917.133
sse	5	17790794374.705	22663619750.593	19394433560.564	2117948592.845	2630295668.134
sde	5	9365.942	10608.530	9818.348	533.068	662.021
mse0	5	88953971.874	113318098.753	96972167.803	10589742.964	13151478.341
rmse	5	9431.541	10645.097	9836.093	528.486	656.330
mae	5	7418.536	8609.308	7817.189	534.033	663.219
smape	5	1.408	1.804	1.592	0.141	0.176
m	5	200.000	200.000	200.000	0.000	0.000
dfr	5	6.000	6.000	6.000	0.000	0.000
df	5	993.000	993.000	993.000	0.000	0.000
fStat	5	96001.466	122329.999	110142.703	10843.190	13466.236
aic	5	-2127.278	-2100.155	-2109.082	11.789	14.641
bic	5	-2104.190	-2077.067	-2085.993	11.789	14.641

Table 8: Statsmodels - House Price Linear Regression CV

Name	In-num folds	min	max	mean	stdev
rSq	5	0.9984	0.9986	0.9985	0.0001
rSqBar	5	0.9984	0.9986	0.9984	0.0001
sst	5	12384264865084.2500	13564576490393.4668	12842056251568.3359	395911885307.9127
sse	5	17238729843.0851	20996608313.5931	19277864710.9665	1296174615.0581
sde	5	9450.9176	10430.2788	9988.5569	337.6978
mse0	5	61921324325.4212	67822882451.9673	64210281257.8417	1979559426.5396
rmse	5	9450.9176	10430.2788	9988.5569	337.6978
mae	5	7516.9137	8174.5836	7794.6342	224.6310
smape	5	1.5104	1.6620	1.5879	0.0540

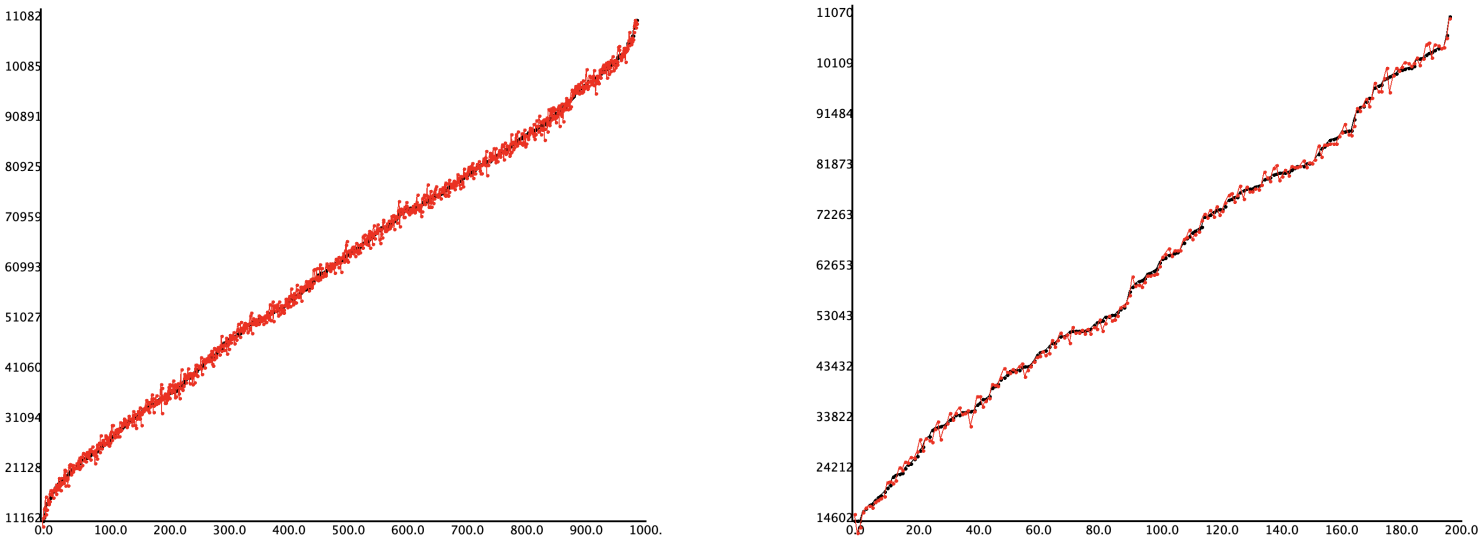
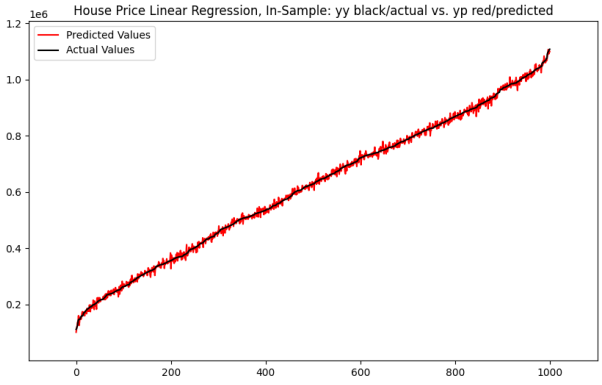


Figure 7: Scalation - House Price Regression
Left: In Sample Predictions
Right: 80-20 Out of Sample Predictions
yy black/actual vs. yp red/predicted



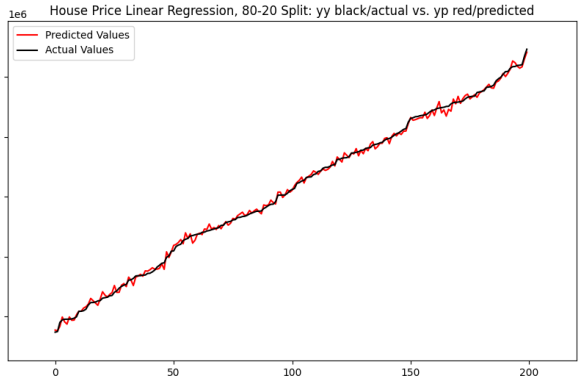
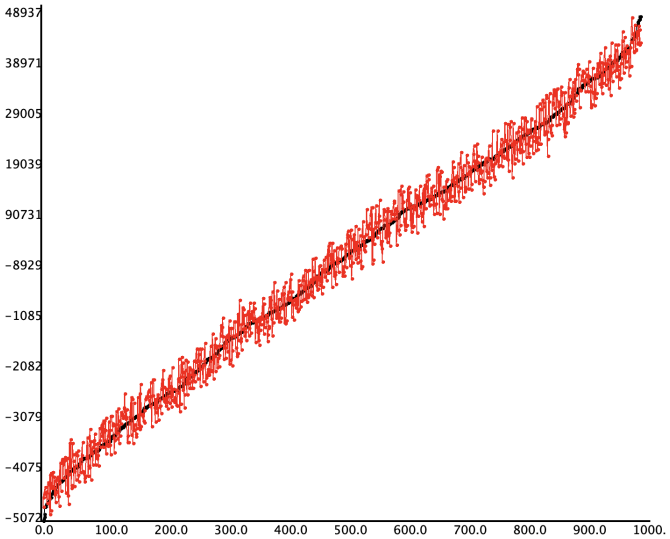


Figure 8: Statsmodels - House Price Regression
Left: In Sample Predictions
Right: 80-20 Out of Sample Predictions
yy black/actual vs. yp red/predicted



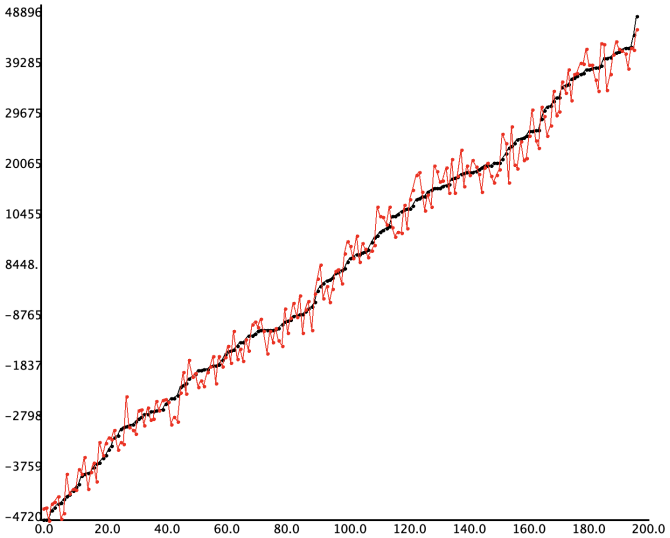


Figure 9: Scalation - House Price Ridge
Left: In Sample Predictions
Right: 80-20 Out of Sample Predictions
yy black/actual vs. yp red/predicted

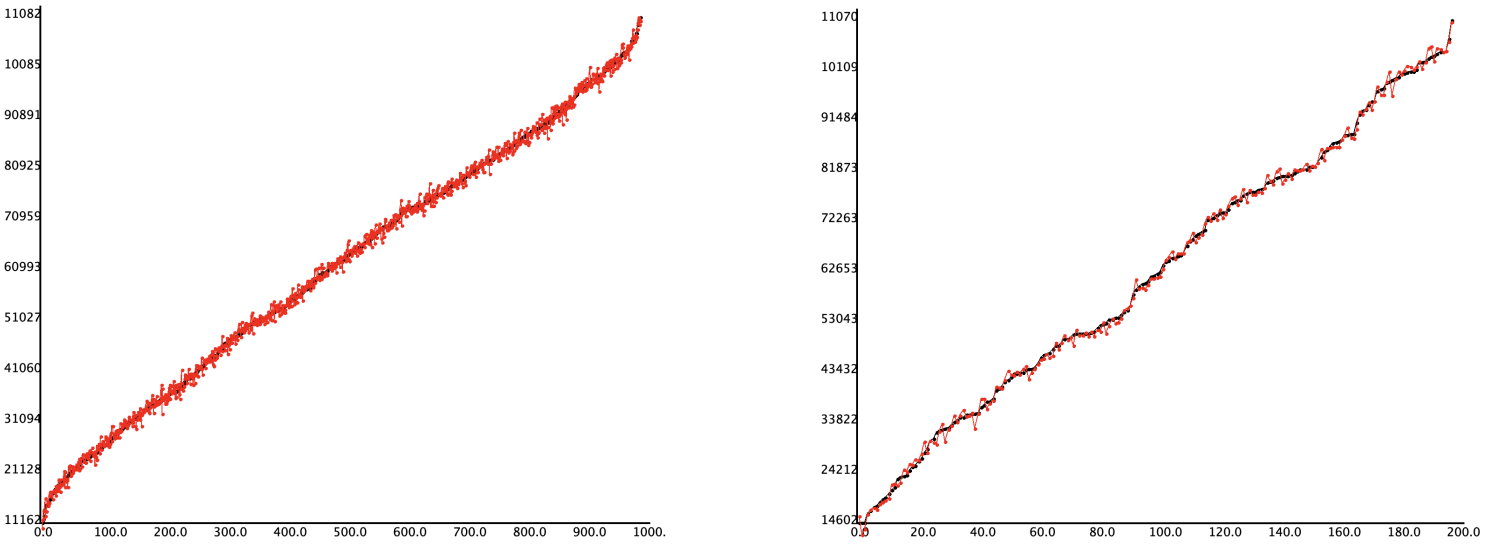


Figure 10: Scalation - House Price Lasso
Left: In Sample Predictions
Right: 80-20 Out of Sample Predictions
yy black/actual vs. yp red/predicted

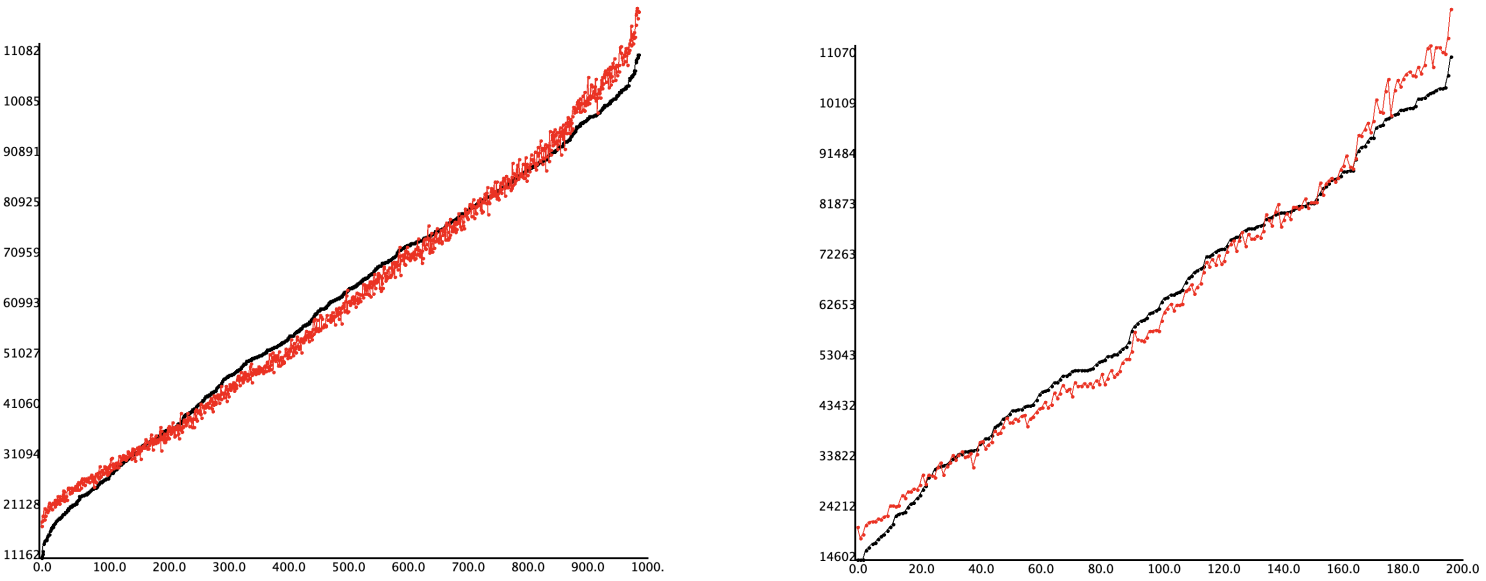


Figure 11: Scalation - House Price Sqrt
Left: In Sample Predictions
Right: 80-20 Out of Sample Predictions
yy black/actual vs. yp red/predicted

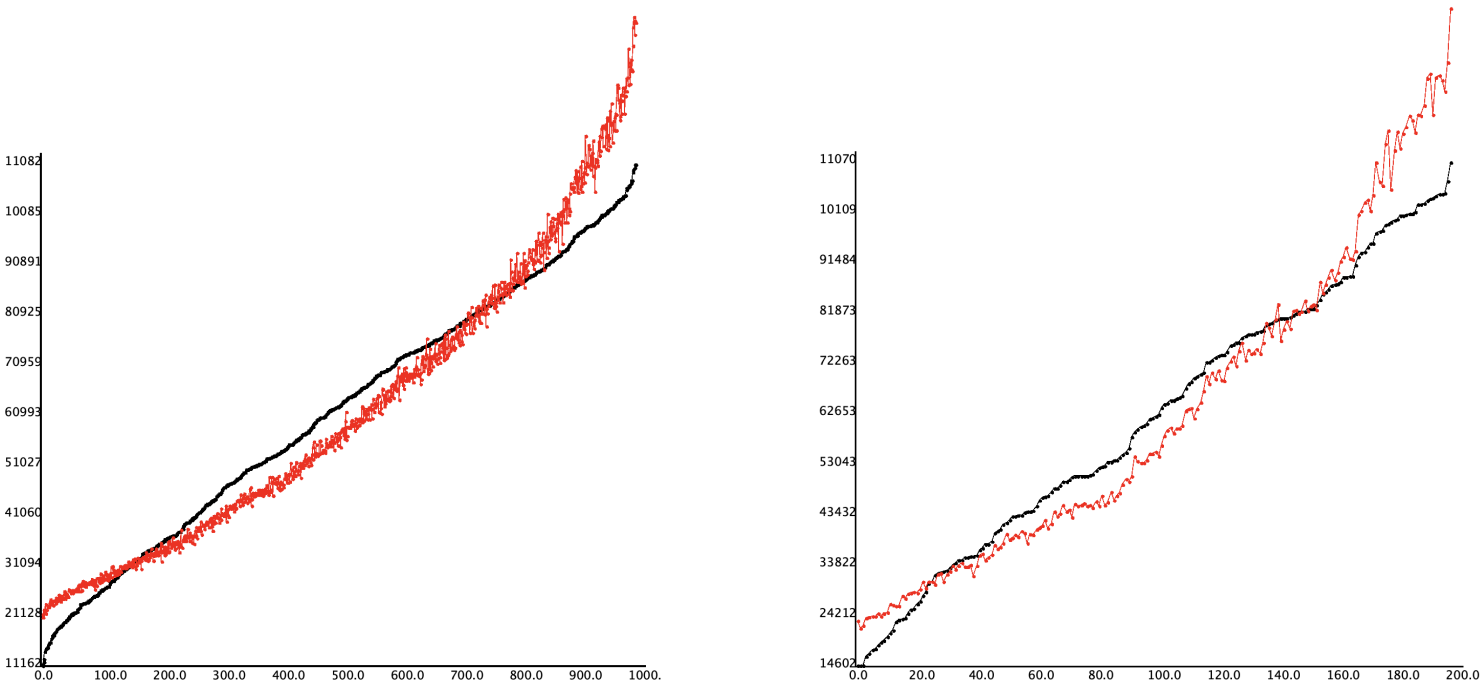


Figure 12: Scalation - House Price log1p
Left: In Sample Predictions
Right: 80-20 Out of Sample Predictions
yy black/actual vs. yp red/predicted

Table 9: Insurance Charges In-Sample QoF Comparison

Metric	Regression	Ridge	Lasso	Sqrt	log1p
rSq	0.750157	0.749962	0.750157	0.752656	0.527431
rSqBar	0.748842	0.748457	0.748842	0.751354	0.524943
sst	1.96074e+11	1.96074e+11	1.96074e+11	1.96074e+11	1.96074e+11
sse	4.89878e+10	4.90260e+10	4.89878e+10	4.84977e+10	9.26587e+10
sde	6053.11	6055.42	6053.11	6001.45	8316.89
mse0	3.66127e+07	3.66413e+07	3.66127e+07	3.62464e+07	6.92516e+07
rmse	6050.84	6053.20	6050.84	6020.50	8321.76
mae	4179.54	4150.46	4179.54	3623.20	4215.04
smape	37.9722	67.8472	37.9722	27.9341	26.5034
m	1338.00	1338.00	1338.00	1338.00	1338.00
dfr	7.00000	8.00000	7.00000	7.00000	7.00000
df	1330.00	1329.00	1330.00	1330.00	1330.00
fStat	570.477	498.274	570.477	578.161	212.057
aic	-13533.8	-13532.3	-13533.8	-13527.1	-13960.2
bic	-13492.2	-13485.5	-13492.2	-13485.5	-13918.6

Table 10: Insurance Charges Out-of-Sample QoF Comparison

Metric	Regression	Ridge	Lasso	Sqrt	log1p
rSq	0.720005	0.719714	0.720005	0.730926	0.583107
rSqBar	0.718531	0.718027	0.718531	0.729510	0.580913
sst	4.06432e+10	4.06432e+10	4.06432e+10	4.06432e+10	4.06432e+10
sse	1.13799e+10	1.13917e+10	1.13799e+10	1.09360e+10	1.69438e+10
sde	6540.46	6543.82	6540.46	6396.98	7972.86
mse0	4.26213e+07	4.26655e+07	4.26213e+07	4.09588e+07	6.34601e+07
rmse	6528.50	6531.89	6528.50	6399.91	7966.18
mae	4430.66	4429.38	4430.66	3868.65	4087.97
smape	40.0602	62.9178	40.0602	30.5895	27.5909
m	267.000	267.000	267.000	267.000	267.000
dfr	7.00000	8.00000	7.00000	7.00000	7.00000
df	1330.00	1329.00	1330.00	1330.00	1330.00
fStat	488.584	426.574	488.584	516.126	265.753
aic	-2708.17	-2706.31	-2708.17	-2702.86	-2761.31
bic	-2679.47	-2674.02	-2679.47	-2674.16	-2732.61

Table 11: Scalation - Insurance Charges Linear Regression

Regression	In-Sample	80-20 Split
rSq	0.750157	0.720005
rSqBar	0.748842	0.718531
sst	1.96074e+11	4.06432e+10
sse	4.89878e+10	1.13799e+10
sde	6053.11	6540.46
mse0	3.66127e+07	4.26213e+07
rmse	6050.84	6528.50
mae	4179.54	4430.66
smape	37.9722	40.0602
m	1338.00	267.000
dfr	7.00000	7.00000
df	1330.00	1330.00
fStat	570.477	488.584
aic	-13533.8	-2708.17
bic	-13492.2	-2679.47

Table 12: Scalation - Insurance Charges Linear Regression CV

Name	num	min	max	mean	stdev	interval
rSq	5	0.701	0.814	0.743	0.046	0.057
rSqBar	5	0.699	0.813	0.742	0.046	0.057
sst	5	31902777173.848	43430486230.657	38949749086.422	4343029725.651	5393640012.108
sse	5	7539480548.420	11454966761.392	9918152392.401	1670500799.560	2074607019.047
sde	5	5320.957	6562.018	6084.654	526.821	654.263
mse0	5	28237754.863	42902497.234	37146638.174	6256557.302	7770063.742
rmse	5	5313.921	6550.000	6076.688	525.006	652.009
mae	5	3810.754	4511.498	4216.703	292.688	363.491
smape	5	36.128	40.060	38.143	1.833	2.277
m	5	267.000	267.000	267.000	0.000	0.000
dfr	5	7.000	7.000	7.000	0.000	0.000
df	5	1330.000	1330.000	1330.000	0.000	0.000
fStat	5	444.882	830.519	573.015	157.534	195.643
aic	5	-2709.047	-2663.110	-2691.017	19.599	24.340
bic	5	-2680.349	-2634.412	-2662.319	19.599	24.340

Table 13: Statsmodels - Insurance Charges Linear Regression CV

Name	In-num folds	min	max	mean	stdev
rSq	5	0.6324	0.7956	0.7402	0.0578
rSqBar	5	0.6211	0.7893	0.7322	0.0596
sst	5	30189024179.7055	43857198758.4016	39154186092.5516	4823002859.8713
sse	5	8965018845.2774	11096336332.7613	9899546225.2180	821555419.0107
sde	5	5872.0390	6545.4562	6170.1628	260.9758
mse0	5	113067506.2910	163646264.0239	146298959.6759	17905426.2621
rmse	5	5872.0390	6545.4562	6170.1628	260.9758
mae	5	4054.1099	4427.9335	4203.4121	129.0554
smape	5	35.6194	40.0220	38.1279	1.5723

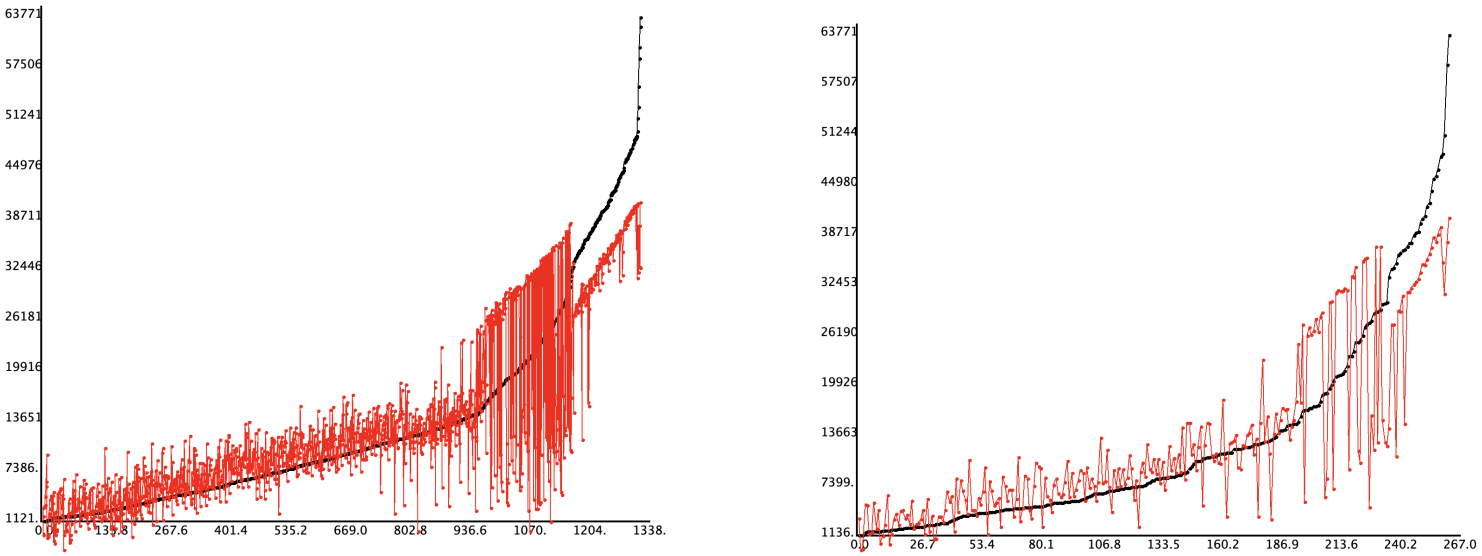


Figure 13: Scalation - Insurance Charges Regression
Left: In Sample Predictions
Right: 80-20 Out of Sample Predictions
yy black/actual vs. yp red/predicted

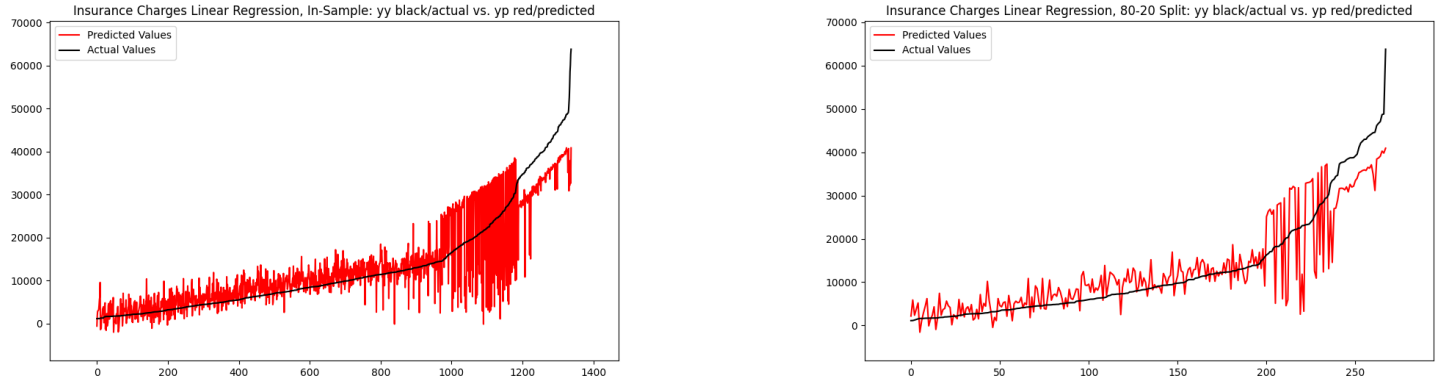


Figure 14: Statsmodels - Insurance Charges Regression
Left: In Sample Predictions
Right: 80-20 Out of Sample Predictions
yy black/actual vs. yp red/predicted

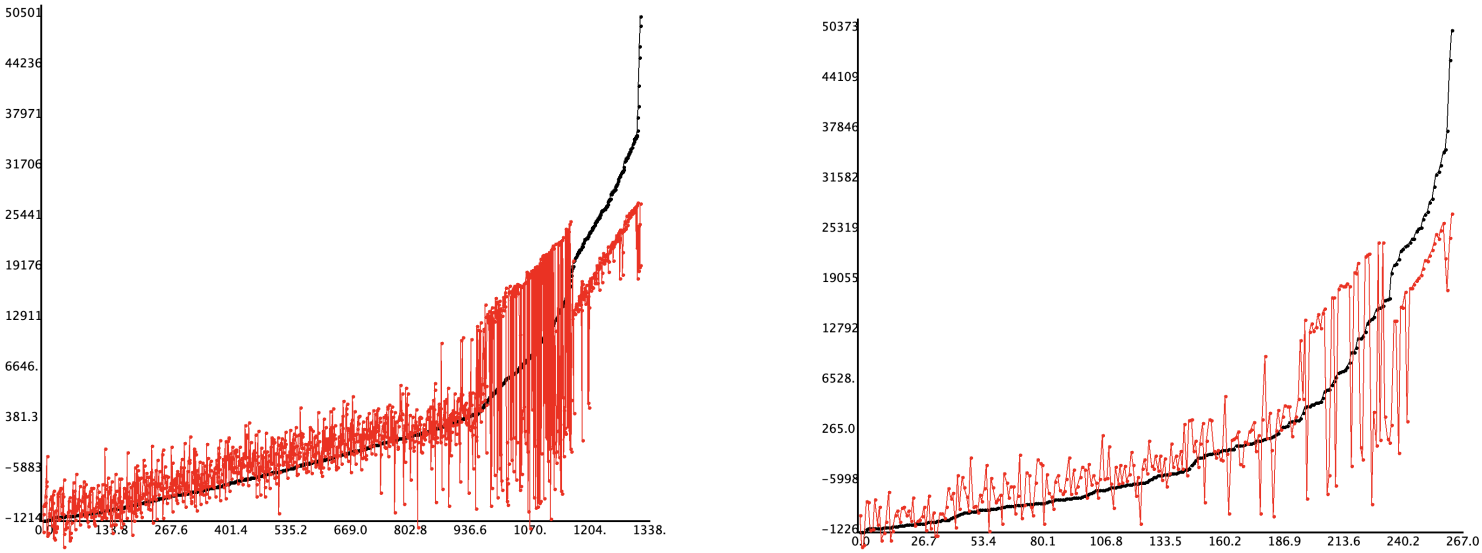


Figure 15: Scalation - Insurance Charges Ridge
Left: In Sample Predictions
Right: 80-20 Out of Sample Predictions
yy black/actual vs. yp red/predicted

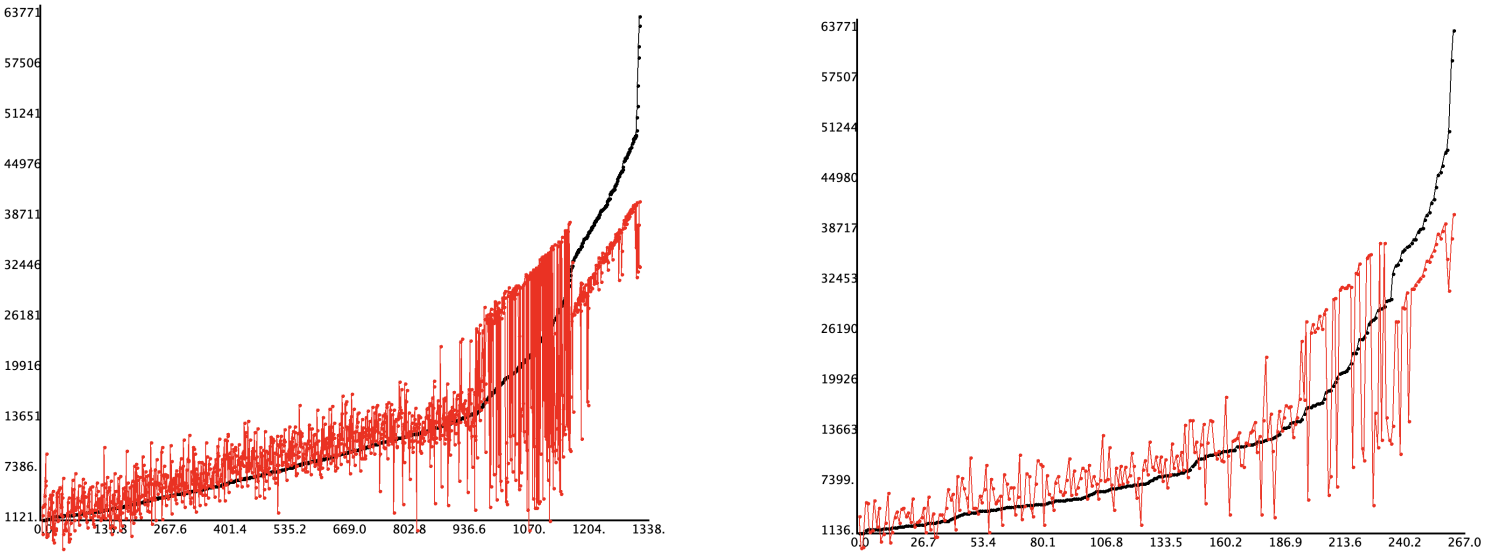


Figure 16: Scalation - Insurance Charges Lasso
Left: In Sample Predictions
Right: 80-20 Out of Sample Predictions
yy black/actual vs. yp red/predicted

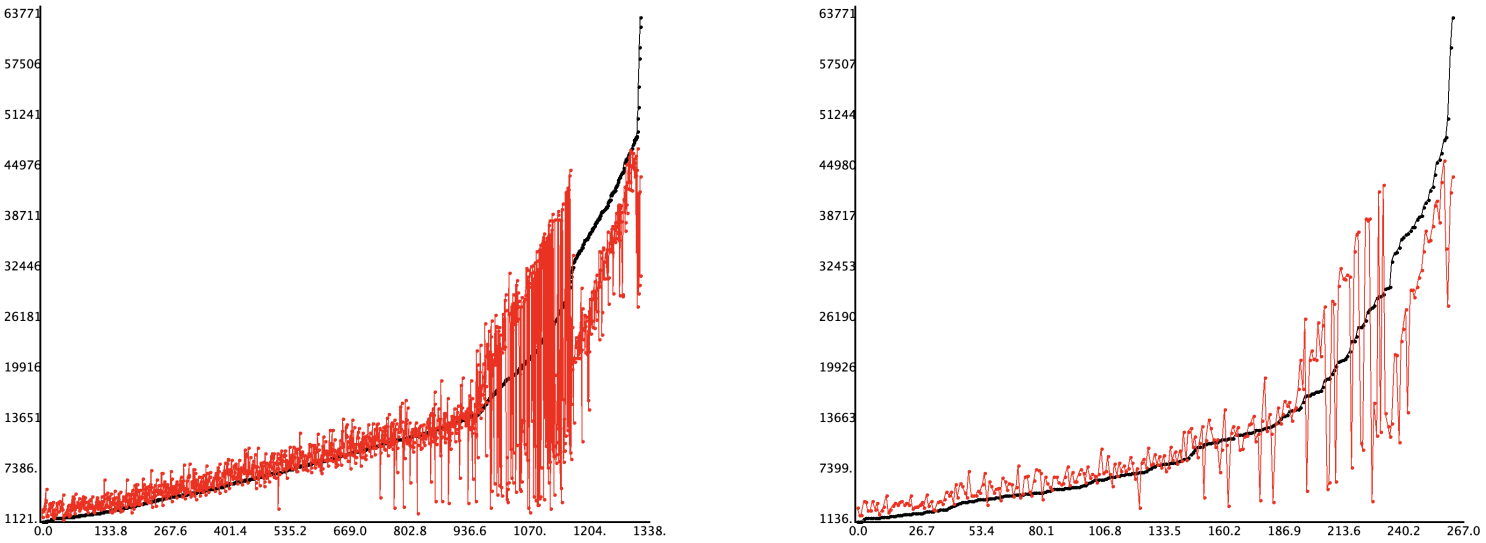


Figure 17: Scalation - Insurance Charges Sqrt
Left: In Sample Predictions
Right: 80-20 Out of Sample Predictions
yy black/actual vs. yp red/predicted

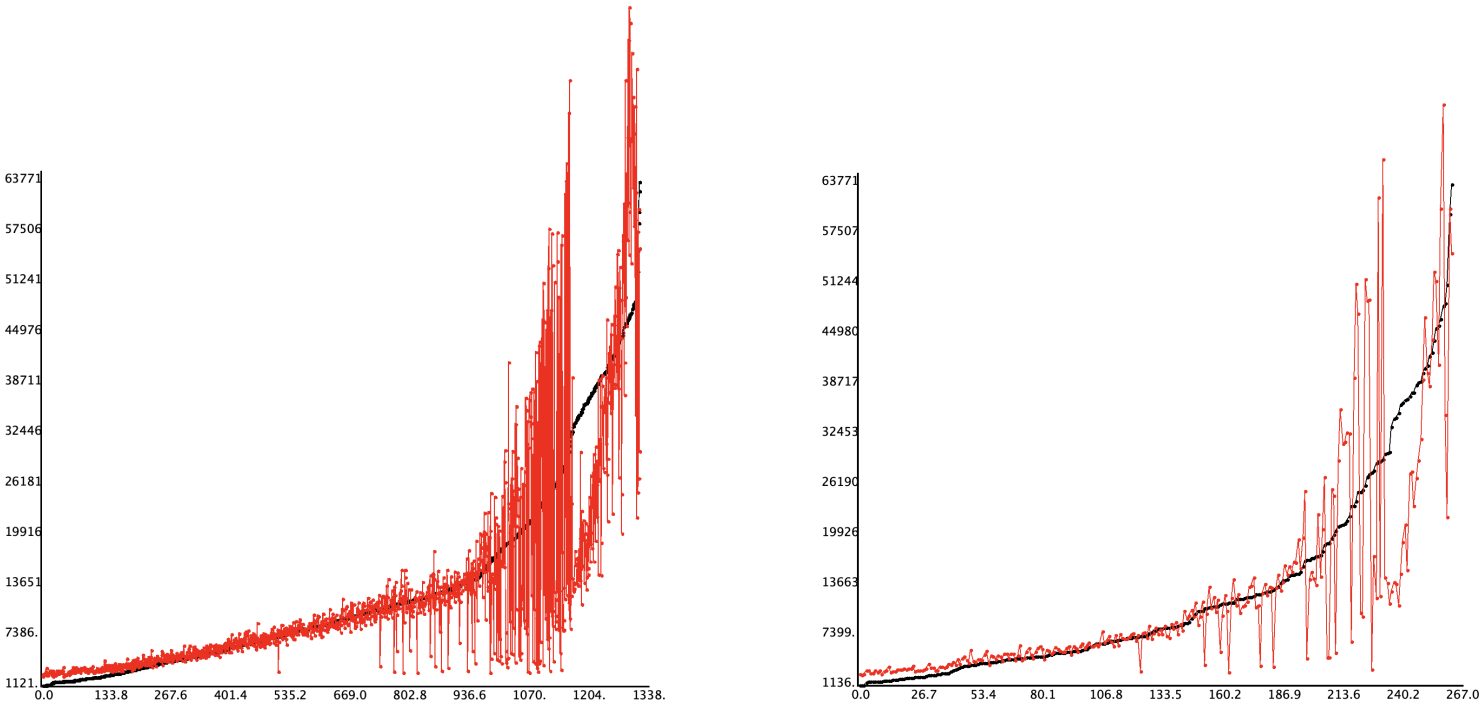


Figure 18: Scalation - Insurance Charges log1p
Left: In Sample Predictions
Right: 80-20 Out of Sample Predictions
yy black/actual vs. yp red/predicted