Code book

1. Name of the data file: organics - 迴歸

2. Data preprocessing:

Number	Event
1	將「Customer.Loyalty.ID」欄位刪除
2	刪除含缺失值的資料，共 7303 筆

3. Data overview:

Total sample size	14920
Independent variables (X)	
Total of categorical variables	6
Total of numeric variables	4
Sum	10
Dependent variables (Y)	
Total of categorical variables	
Total of numeric variables	1

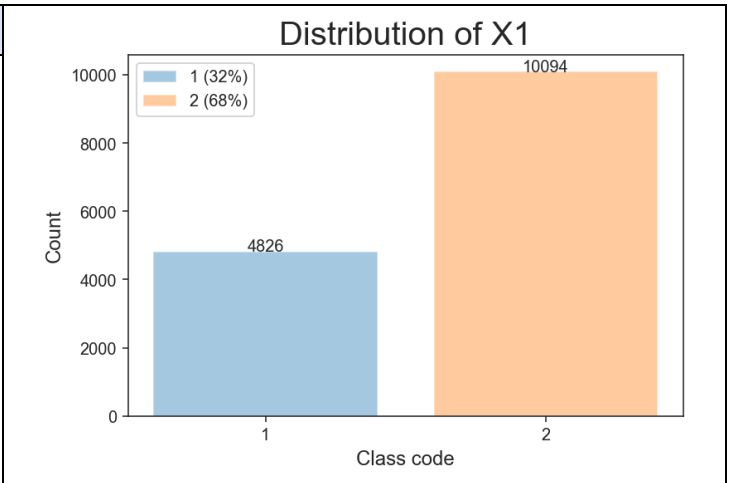
4. Variables overview:

Index	Variable	Description / Unit	Miss
X1	Gender	1:M 2:F	4259
X2	Geographic.Region	1:North 2:Scottish 3:Midlands 4:South East 5:South West	465
X3	Loyalty.Status	1:Platinum 2:Gold 3:Silver 4:Tin	0
X4	Neigborhood.Cluster.55.Level	Continuous	674
X5	Neighborhood.Cluster.7.Level	1:A 2:B 3:C 4:D	695

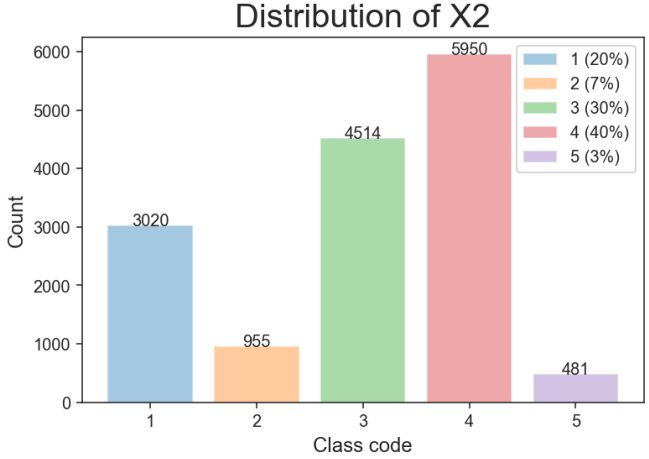
		5:E 6:F 7:U	
X6	Television.Region	1:Border 2:C Scotland 3:East 4:London 5:Midlands 6:N East 7:N Scot 8:N West 9:S & S East 10:S West 11:Wales & West 12:Yorkshire	465
X7	Affluence.Grade	Continuous	1085
X8	Age	Continuous	1508
X9	Loyalty.Card.Tenure	Continuous	281
X10	Organics.Purchase.Indicator	0:0 1:1	0
Y	Total.Spend	Continuous	0

5. Variable description:

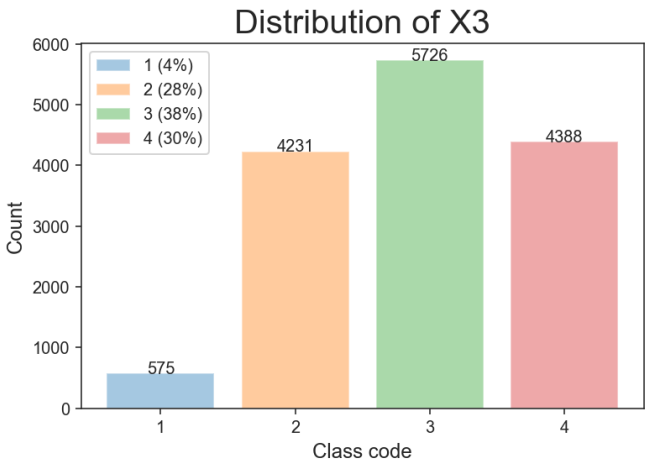
Independent variable (known as "X" or input)

Index	Variable	
X1	Gender	

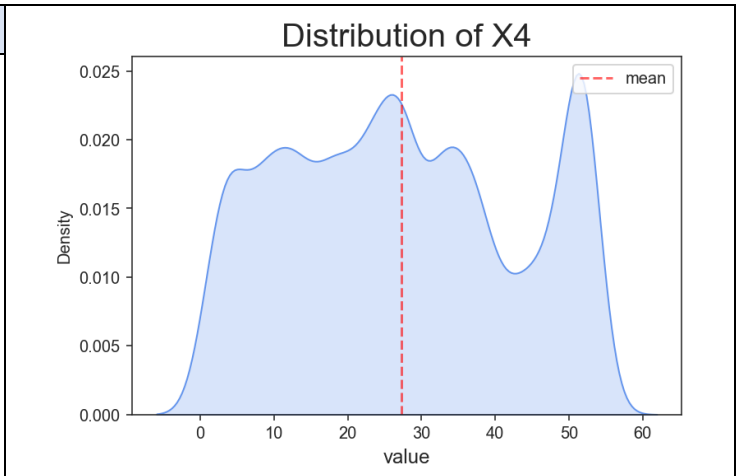
Description	Class code	Count	Percentage
F	1	4826	32%
M	2	10094	68%

Index	Variable	
X2	Geographic.Region	

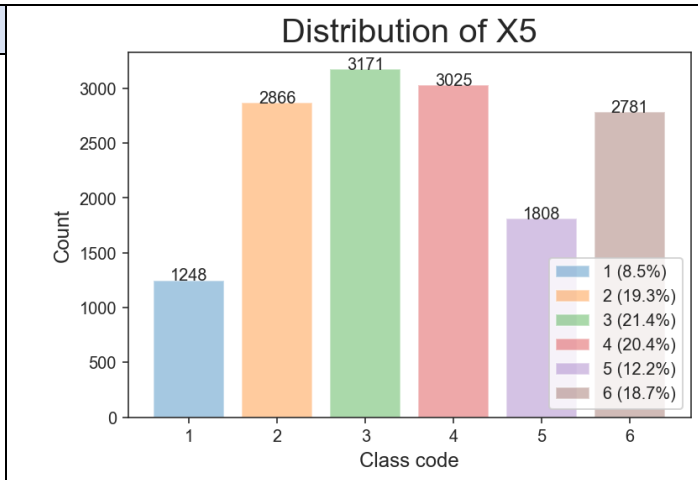
Description	Class code	Count	Percentage
North	1	3020	20%
Scottish	2	955	7%
Midlands	3	4514	30%
South East	4	5950	40%
South West	5	481	3%

Index	Variable	
X3	Loyalty.Status	

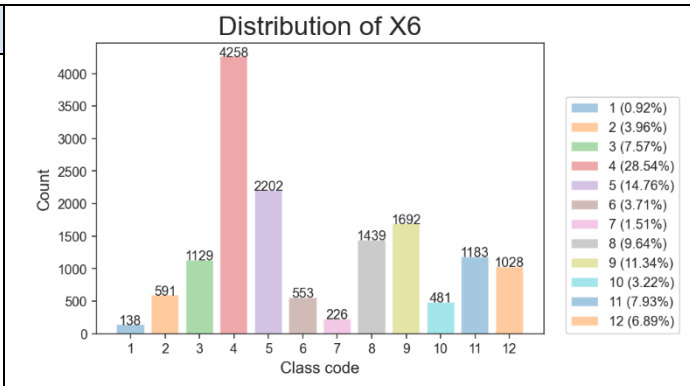
Description	Class code	Count	Percentage
Platinum	1	575	4%
Gold	2	4231	28%
Silver	3	5726	38%
Tin	4	4388	30%

Index	Variable	
X4	Neigborhood .Cluster.55. Level	

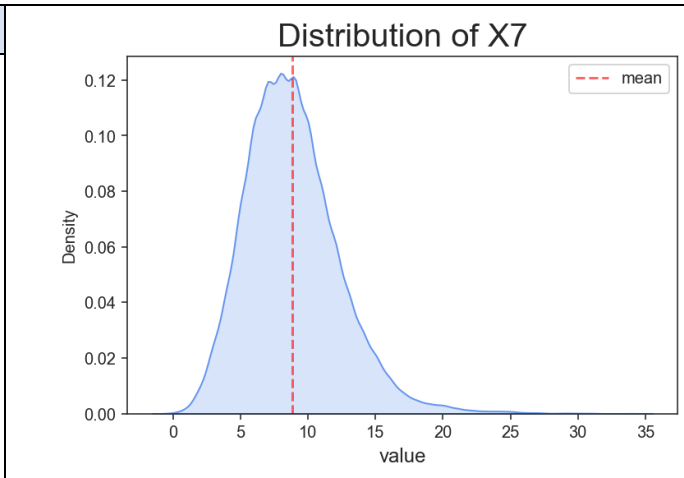
Mean	SD	Median	Mode
27.29	15.72	27	52
Min	**Max**	**Range**	**Variance**
1	55	54	247.26
Q1(25%)	**Q2(50%)**	**Q3(75%)**	**IQR**
14	27	39	25

Index	Variable	
X5	Neighborhood. Cluster.7.Level	

Description	Class code	Count	Percentage
A	1	1248	8.5%
B	2	2866	19.3%
C	3	3171	21.4%
D	4	3025	20.4%
E	5	1808	12.2%
F	6	2781	18.7%
U	7	21	1.5%

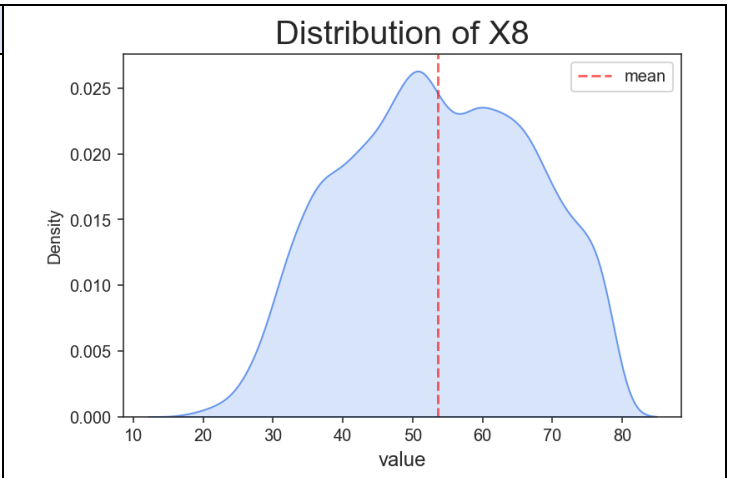
Index	Variable	
X6	Television.Region	

Description	Class code	Count	Percentage
Border	1	138	0.92%
C Scotland	2	591	3.96%
East	3	1129	7.57%
London	4	4258	28.54%
Midlands	5	2202	14.76
N East	6	553	3.71%
N Scot	7	226	1.51%
N West	8	1439	9.64%
S & S East	9	1692	11.34%
S West	10	481	3.22%
Wales & West	11	1183	7.93%
Yorkshire	12	1028	6.89%

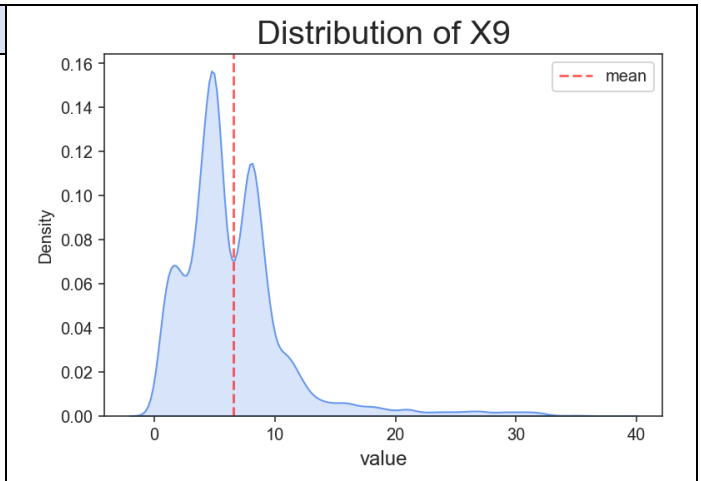
Index	Variable	
X7	Affluence. Grade	

Mean	SD	Median	Mode
8.83	3.49	8	9
Min	**Max**	**Range**	**Variance**
0	34	34	12.15

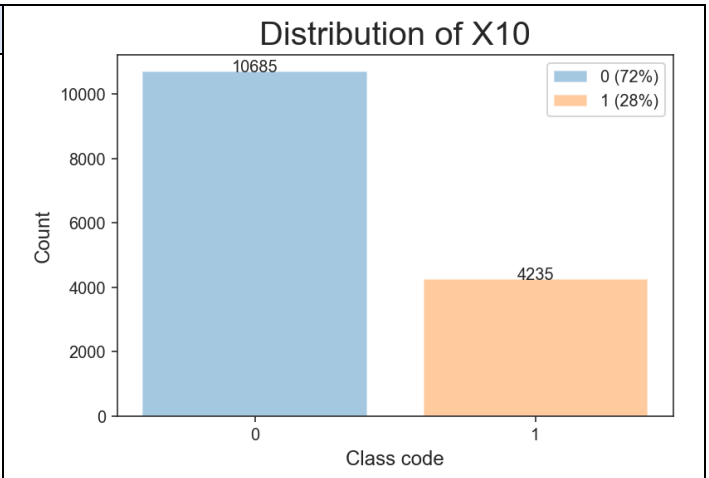
Q1(25%)	Q2(50%)	Q3(75%)	IQR
6	8	11	5

Index	Variable	
X8	Age	

Mean	SD	Median	Mode
53.54	13.31	53	51
Min	**Max**	**Range**	**Variance**
18	79	61	177.29
Q1(25%)	**Q2(50%)**	**Q3(75%)**	**IQR**
43	53	64	21

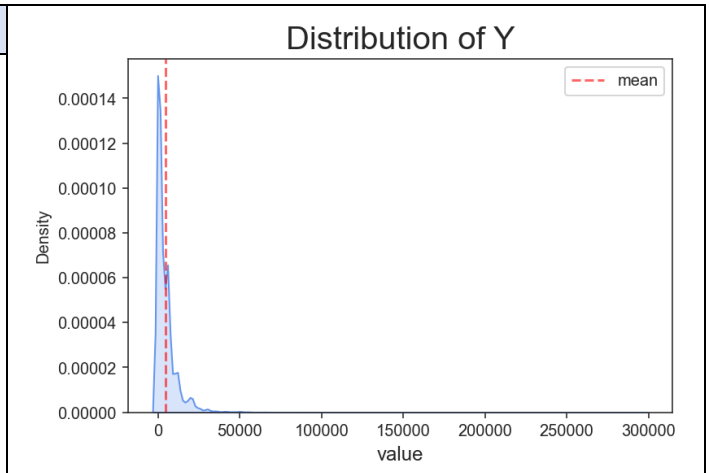
Index	Variable	
X9	Loyalty.Card.Tenure	

Mean	SD	Median	Mode
6.56	4.61	5	5
Min	**Max**	**Range**	**Variance**
0	38	38	21.27
Q1(25%)	**Q2(50%)**	**Q3(75%)**	**IQR**
4	5	8	4

Index	Variable	
X10	Organics.Purchase. Indicator	

Description	Class code	Count	Percentage
0	0	10685	72%
1	1	4235	28%

Dependent variable (known as "Y" or output)

Index	Variable	
Y	Total.Spend	

Mean	SD	Median	Mode
4429.82	7574.4	2000	0.01
Min	**Max**	**Range**	**Variance**
0.01	296313.85	296313.84	57375347
Q1(25%)	**Q2(50%)**	**Q3(75%)**	**IQR**
0.01	2000	6000	5999.99

6. Hyper parameter tuning:

Linear Regression

Round	Hyperparameter	MAE	MAPE	MSE	RMSE
Seed=1		1775.8,	9083.81	10872115.94	3297.29
Seed=2		1788.51	7909.1	9025024.19	**3004.17**
Seed=3		1816.17	14363.58	10086408.77	3175.91
Seed=4		1759.88	8233.15	32167860.28	5671.67
Seed=5		1742.65	8355.73	12961565.53	3600.22
Seed=6		1802.53	9248.59	10902356.25	3301.87
Seed=7		1841.51	8175.02	12653819.65	3557.22
Seed=8		1794.83	8380.42	11501833.21	3391.44
Seed=9		1834.75	8419.07	33884385.47	5821.03
Seed=10		1919.81	8355.05	28051554.34	5296.37
Mean		1807.64	9052.35	17210692.36	4011.72
SD		47.55	1811.64	9425485.02	1056.79

Lasso

Round	Hyperparameter	MAE	MAPE	MSE	RMSE
Seed=1	alpha=0.9977000638225533	1766.46	8532.75	10852298.87	3294.28
Seed=2	alpha=0.9977000638225533	1783.11	7536.28	9006994.42	**3001.17**
Seed=3	alpha=0.9977000638225533	1741.17	8676.57	9801252.1	3130.7
Seed=4	alpha=0.9977000638225533	1753.46	7751.62	32175311.08	5672.33
Seed=5	alpha=0.9977000638225533	1732.92	7986.56	12977752.23	3602.46
Seed=6	alpha=0.9977000638225533	1795.59	8489.42	10882628.81	3298.88
Seed=7	alpha=0.9977000638225533	1833.25	7666.88	12639459.49	3555.2
Seed=8	alpha=0.9977000638225533	1779.77	8094.97	11452661.46	3384.18
Seed=9	alpha=0.9977000638225533	1827.16	7955.12	33875158.98	5820.24
Seed=10	alpha=0.9977000638225533	1916.96	8139.34	28071691.91	5298.27
Mean		1792.98	8082.95	17173520.93	4005.77
SD		51.96	364.78	9456398.09	1061.75

CART

Round	Hyperparameter	MAE	MAPE	MSE	RMSE
Seed=1	max_depth=2, max_leaf_nodes=4, min_samples_leaf=1,	2035.88	32191.63	11532884.55	3396.01
Seed=2	max_depth=2, max_leaf_nodes=4, min_samples_leaf=1,	2086.45	33531.96	9700032.79	**3114.49**
Seed=3	max_depth=3, max_leaf_nodes=4, min_samples_leaf=1,	1676.18	924.41	9780900.03	3127.44
Seed=4	max_depth=3, max_leaf_nodes=4, min_samples_leaf=1,	1712.13	1244.09	32317151.25	5684.82
Seed=5	max_depth=3, max_leaf_nodes=4, min_samples_leaf=1,	1687.83	675.85	13026510.52	3609.23
Seed=6	max_depth=3, max_leaf_nodes=4, min_samples_leaf=1,	1745.15	688.14	10935422.76	3306.88
Seed=7	max_depth=3, max_leaf_nodes=4, min_samples_leaf=1,	1786.18	669.42	12738348.21	3569.08
Seed=8	max_depth=3, max_leaf_nodes=4, min_samples_leaf=1,	1727.57	753.93	11537998.08	3396.76
Seed=9	max_depth=3, max_leaf_nodes=4, min_samples_leaf=1,	1771	899.02	33919960.12	5824.08
Seed=10	max_depth=3, max_leaf_nodes=4, min_samples_leaf=1,	1859.81	934.25	28088805.29	5299.89
Mean		1808.82	7251.27	17357801.36	4032.87
SD		136.17	12809.82	9372506.02	1045.84

KNN

Round	Hyperparameter	MAE	MAPE	MSE	RMSE
Seed=1	n_neighbors=6	1937.29	2930.53	20096460.68	4482.91
Seed=2	n_neighbors=5	1908.6	1797.15	15681960.79	**3960.05**
Seed=3	n_neighbors=5	1877.53	1978.05	16102932.64	4012.85
Seed=4	n_neighbors=4	1856.36	1947.01	36609460.12	6050.58
Seed=5	n_neighbors=5	1942.1	1513.87	24924146.54	4992.41
Seed=6	n_neighbors=5	1942.69	2915.51	20601833.03	4538.92
Seed=7	n_neighbors=6	1982.09	1788.27	21403810.82	4626.43
Seed=8	n_neighbors=5	1917.22	2016.24	19543175.17	4420.77
Seed=9	n_neighbors=4	2008.93	1871.76	43133807.26	6567.63
Seed=10	n_neighbors=5	2127.19	2126.89	39901477.98	6316.76
Mean		1950	2088.53	25799906.5	4996.93
SD		72.79	445.43	9650686.88	911.37

RF

Round	Hyperparameter	MAE	MAPE	MSE	RMSE
Seed=1	max_depth=3, max_leaf_nodes=4, min_samples_leaf=5, n_estimators=50	1920.74	22226.24	10806864.96	3287.38
Seed=2	max_depth=3, max_leaf_nodes=4, min_samples_leaf=1, n_estimators=10	1902.99	17215.38	9312724.87	**3051.68**
Seed=3	max_depth=3, max_leaf_nodes=4, min_samples_leaf=5, n_estimators=100	1850.45	18370.99	9839978.39	3136.87
Seed=4	max_depth=3, max_leaf_nodes=4, min_samples_leaf=5, n_estimators=50	1809.79	12153.8	32301416.38	5683.43
Seed=5	max_depth=3, max_leaf_nodes=4, min_samples_leaf=5, n_estimators=50	1857.08	18078.77	13164719.93	3628.32
Seed=6	max_depth=3,	1891.42	15800.22	10756811.17	3279.76

	max_leaf_nodes=4, min_samples_leaf=5, n_estimators=100				
Seed=7	max_depth=3, max_leaf_nodes=4, min_samples_leaf=5, n_estimators=100	1915.88	13864.58	12433568.66	3526.13
Seed=8	max_depth=3, max_leaf_nodes=4, min_samples_leaf=5, n_estimators=100	1842.95	13418.61	11526187.04	3395.02
Seed=9	max_depth=3, max_leaf_nodes=4, min_samples_leaf=5, n_estimators=100	1913.58	14767.31	33973988.75	5828.72
Seed=10	max_depth=3, max_leaf_nodes=4, min_samples_leaf=5, n_estimators=100	1876.03	2904.21	28106886.71	5301.59
Mean		1878.09	14880.01	17222314.69	4011.89
SD		35.23	4865.37	9478053.94	1061.62

SVR

Round	Hyperparameter	MAE	MAPE	MSE	RMSE
Seed=1	C=10	2146.8	1315.49	29939929.14	5471.74
Seed=2	C=10	2174.6	1064.21	26789600.03	**5175.87**
Seed=3	C=10	2066.79	1390.61	29111245.66	5395.48
Seed=4	C=10	2017.1	1524.55	51108924.79	7149.05
Seed=5	C=10	2077.75	1002.55	31502777.23	5612.73
Seed=6	C=10	2129.43	1184.38	29152463.6	5399.3
Seed=7	C=10	2204.2	1068.62	33984480.25	5829.62
Seed=8	C=10	2112.52	1122.02	30393131.25	5513
Seed=9	C=10	2208.55	1273.88	57653759.4	7593.01
Seed=10	C=10	2348.97	1277.28	55366316.72	7440.85
Mean		2148.67	1222.36	37500262.81	6058.06
SD		88.42	155.71	11494934.41	894.49

Summary of results

	MAE	MAPE	MSE	RMSE
Linear Regression	1807.64 (±47.55)	9052.35 (±1811.64)	17210692.36 (±9425485.02)	4011.72 (±1056.79)
Lasso	1792.98 (±51.96)	8082.95 (±364.78)	17173520.93 (±9456398.09)	**4005.77 (±1061.75)**
CART	1808.82 (±136.17)	7251.27 (±12809.82)	17357801.36 (±9372506.02)	4032.87 (±1045.84)
KNN	1950 (±72.79)	2088.53 (±445.43)	25799906.5 (±9650686.88)	4996.93 (±911.37)
RF	1878.09 (±35.23)	14880.01 (±4865.37)	17222314.69 (±9478053.94)	4011.89 (±1061.62)
SVM	2148.67 (±88.42)	1222.36 (±155.71)	37500262.81 (±11494934.41)	6058.06 (±894.49)

7. Variables importance and ranking:

Feature importance

Variable	Linear Regression	Lasso	CART	RF
X1_1	4.5E+15	79.14941	0	0
X1_2	4.5E+15	0	0	0
X2_1	5.4E+15	12.9438	0	0
X2_2	1.04E+15	186.6391	0	0
X2_3	1.9E+15	116.855	0	0
X2_4	7E+15	0	0	0
X2_5	7.29E+14	0	0	0
X3_1	1.6E+16	28531.31	0.68652	0.67565
X3_2	1.6E+16	6700.405	0.29739	0.29442
X3_3	1.6E+16	0	0	0
X3_4	1.6E+16	1842.05	0	0.00847
X4	1511	580.84773	0	0
X5_1	2.2E+15	220.557	0	0
X5_2	2.2E+15	281.438	0	0
X5_3	2.2E+15	63.37076	0	0
X5_4	2.2E+15	0	0	0
X5_5	2.2E+15	187.127	0	0

X5_6	2.2E+15	249.1586	0	0
X5_7	2.2E+15	0	0	0
X6_1	2.3E+15	580.0151	0	0
X6_2	2.3E+15	234.89	0	0
X6_3	6.13E+14	170.1432	0	0
X6_4	5.71E+15	99.77139	0	0
X6_5	6.13E+14	13.1373	0	0
X6_6	4.09E+15	54.9414	0	0
X6_7	2.3E+15	6.04142	0	0
X6_8	4.09E+15	204.93	0	0
X6_9	5.71E+15	18.6406	0	0
X6_10	2E+15	0	0	0
X6_11	6.13E+14	0	0	0
X6_12	4.09E+15	0	0	0
X7	727	660.23382	0	0.01672
X8	1235	1222.19361	0	0
X9	1155	1111.64453	0.01609	0.00475
X10_0	3.86E+15	19.6229	0	0
X10_1	3.86E+15	0	0	0

Feature ranking

Variable	Linear Regression	Lasso	CART	RF	Average
X1_1	9	19	36	36	25
X1_2	9	36	36	36	29.3
X2_1	8	25	36	36	26.3
X2_2	28	15	36	36	28.8
X2_3	27	17	36	36	29
X2_4	5	36	36	36	28.3
X2_5	29	36	36	36	34.3
X3_1	1	1	1	1	1
X3_2	1	2	2	2	1.8
X3_3	1	36	36	36	27.3
X3_4	1	3	36	4	11
X4	33	7	36	36	28

X5_1	19	12	36	36	25.8
X5_2	19	9	36	36	25
X5_3	19	20	36	36	27.8
X5_4	19	36	36	36	31.8
X5_5	19	14	36	36	26.3
X5_6	19	10	36	36	25.3
X5_7	19	36	36	36	31.8
X6_1	16	8	36	36	24
X6_2	16	11	36	36	24.8
X6_3	30	16	36	36	29.5
X6_4	6	18	36	36	24
X6_5	30	24	36	36	31.5
X6_6	11	21	36	36	26
X6_7	16	26	36	36	28.5
X6_8	11	13	36	36	24
X6_9	6	23	36	36	25.3
X6_10	26	36	36	36	33.5
X6_11	30	36	36	36	34.5
X6_12	11	36	36	36	29.8
X7	36	6	36	3	20.3
X8	34	4	36	36	27.5
X9	35	5	3	5	12
X10_0	14	22	36	36	27
X10_1	14	36	36	36	30.5

Final Feature ranking

Rank	Variable
1	X3_1
2	X3_2
3	X3_4
4	X9
5	X7
6	X6_1
6	X6_4
6	X6_8
9	X6_2
10	X1_1

10	X5_2
12	X5_6
12	X6_9
14	X5_1
15	X6_6
16	X2_1
16	X5_5
18	X10_0
19	X3_3
20	X8
21	X5_3
22	X4
23	X2_4
24	X6_7
25	X2_2
26	X2_3
27	X1_2
28	X6_3
29	X6_12
30	X10_1
31	X6_5
32	X5_4
32	X5_7
34	X6_10
35	X2_5
36	X6_11