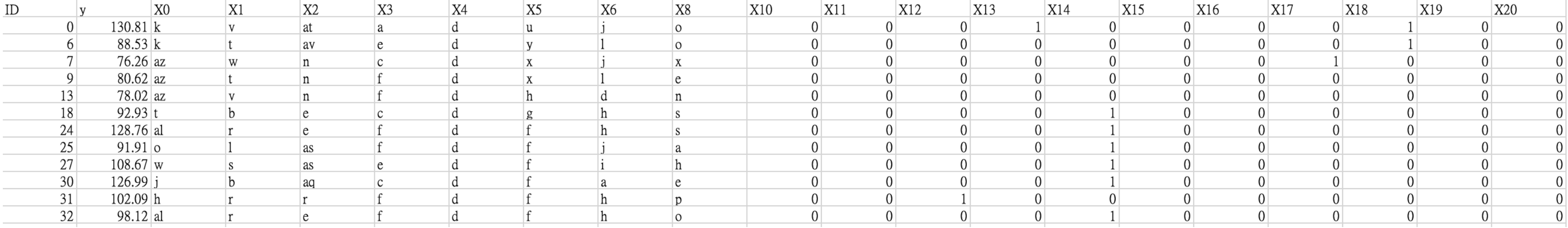
Cathay Assignment

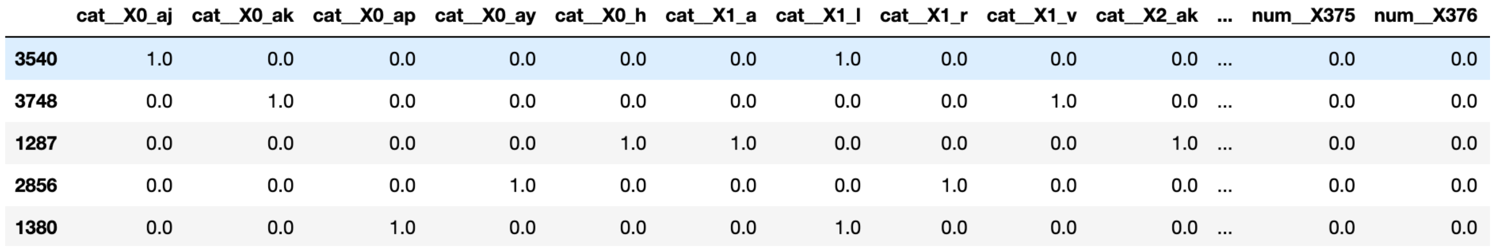
黃貞棠

Data preprocessing

- Split of data sets and preprocessing targets and features
- Data definition
  - Categorical data (Nominal or Ordinal)
  - Numerical data
- feature engineering
  - One hot encoding



ID	y	X0	X1	X2	X3	X4	X5	X6	X8	X10	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20
0	130.81	k	v	at	a	d	u	j	o	0	0	0	1	0	0	0	0	1	0	0
6	88.53	k	t	av	e	d	y	l	o	0	0	0	0	0	0	0	0	1	0	0
7	76.26	az	w	n	c	d	x	j	x	0	0	0	0	0	0	0	1	0	0	0
9	80.62	az	t	n	f	d	x	l	e	0	0	0	0	0	0	0	0	0	0	0
13	78.02	az	v	n	f	d	h	d	n	0	0	0	0	0	0	0	0	0	0	0
18	92.93	t	b	e	c	d	g	h	s	0	0	0	0	1	0	0	0	0	0	0
24	128.76	al	r	e	f	d	f	h	s	0	0	0	0	1	0	0	0	0	0	0
25	91.91	o	l	as	f	d	f	j	a	0	0	0	0	1	0	0	0	0	0	0
27	108.67	w	s	as	e	d	f	i	h	0	0	0	0	1	0	0	0	0	0	0
30	126.99	j	b	aq	c	d	f	a	e	0	0	0	0	1	0	0	0	0	0	0
31	102.09	h	r	r	f	d	f	h	p	0	0	1	0	0	0	0	0	0	0	0
32	98.12	al	r	e	f	d	f	h	o	0	0	0	0	1	0	0	0	0	0	0



	cat__X0_aj	cat__X0_ak	cat__X0_ap	cat__X0_ay	cat__X0_h	cat__X1_a	cat__X1_l	cat__X1_r	cat__X1_v	cat__X2_ak	...	num__X375	num__X376
3540	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	...	0.0	0.0
3748	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	...	0.0	0.0
1287	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	1.0	...	0.0	0.0
2856	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	...	0.0	0.0
1380	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	...	0.0	0.0

Model Selection

- Lasso (L1 norm)
  - Lasso sets the coefficients of certain features to zero, achieving feature selection.

$$\min_\beta \frac{1}{2n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \alpha \sum_{j=1}^{p} |\beta_j|$$

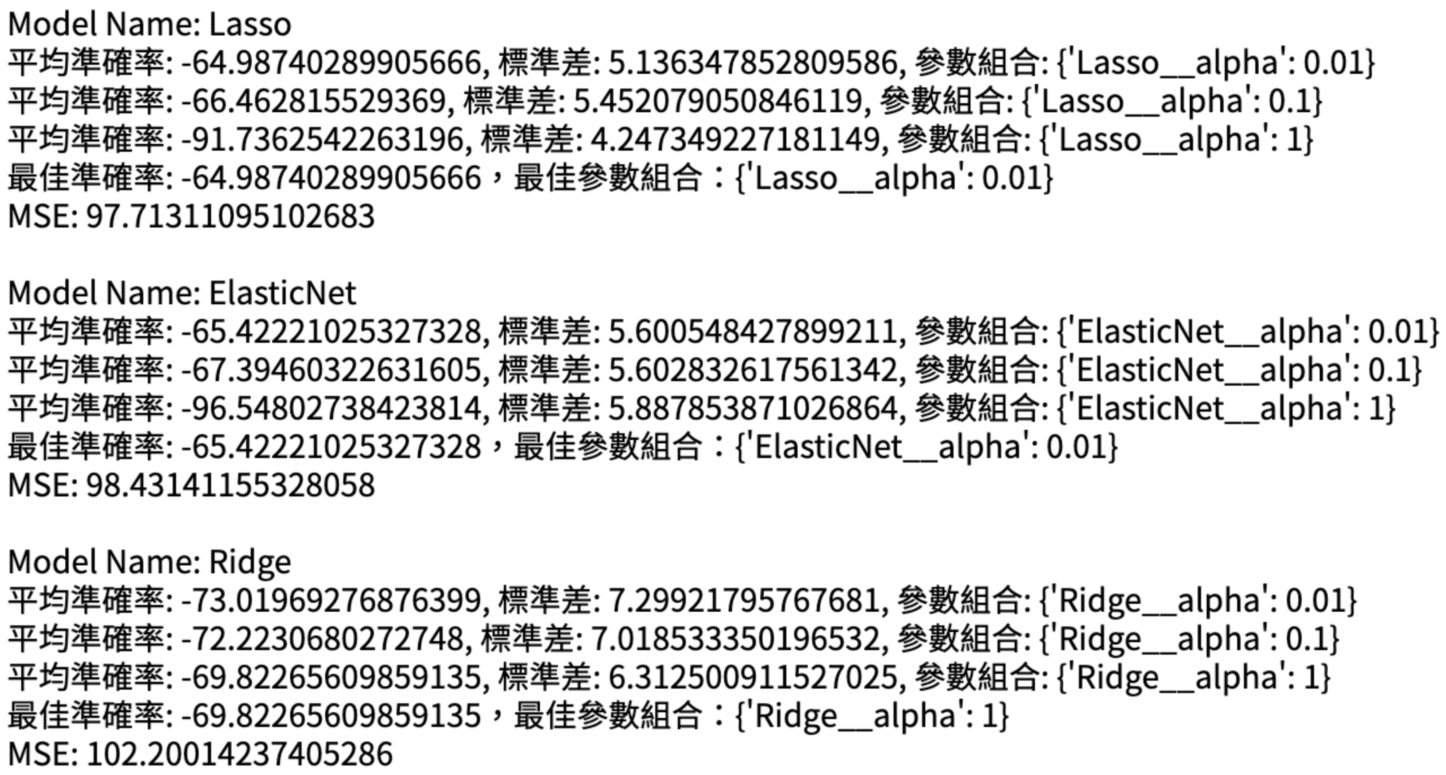
- Ridge (L2 norm)
  - Ridge retains all features but shrinks the coefficients.

$$\min_\beta \frac{1}{2n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \alpha \sum_{j=1}^{p} \beta_j^2$$

- Elastic Net (L1 norm +L2 norm)
  - Elastic Net combines both L1 and L2 regularization and allows for some level of feature selection.

$$\min_\beta \frac{1}{2n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \alpha \left(\rho \sum_{j=1}^{p} |\beta_j| + \frac{1}{2}(1 - \rho) \sum_{j=1}^{p} \beta_j^2 \right)$$

Grid Search and Result



```
Model Name: Lasso
平均準確率: -64.98740289905666, 標準差: 5.136347852809586, 參數組合: {'Lasso__alpha': 0.01}
平均準確率: -66.462815529369, 標準差: 5.452079050846119, 參數組合: {'Lasso__alpha': 0.1}
平均準確率: -91.7362542263196, 標準差: 4.247349227181149, 參數組合: {'Lasso__alpha': 1}
最佳準確率: -64.98740289905666，最佳參數組合：{'Lasso__alpha': 0.01}
MSE: 97.71311095102683

Model Name: ElasticNet
平均準確率: -65.42221025327328, 標準差: 5.600548427899211, 參數組合: {'ElasticNet__alpha': 0.01}
平均準確率: -67.39460322631605, 標準差: 5.602832617561342, 參數組合: {'ElasticNet__alpha': 0.1}
平均準確率: -96.54802738423814, 標準差: 5.887853871026864, 參數組合: {'ElasticNet__alpha': 1}
最佳準確率: -65.42221025327328，最佳參數組合：{'ElasticNet__alpha': 0.01}
MSE: 98.43141155328058

Model Name: Ridge
平均準確率: -73.01969276876399, 標準差: 7.29921795767681, 參數組合: {'Ridge__alpha': 0.01}
平均準確率: -72.2230680272748, 標準差: 7.018533350196532, 參數組合: {'Ridge__alpha': 0.1}
平均準確率: -69.82265609859135, 標準差: 6.312500911527025, 參數組合: {'Ridge__alpha': 1}
最佳準確率: -69.82265609859135，最佳參數組合：{'Ridge__alpha': 1}
MSE: 102.20014237405286
```

Conclusion

- Understanding the data and defining the problem, and then performing data preprocessing to deal with features is very important and takes the most time.
- Understand the goals and choose the appropriate one from different models.