

實驗結果__劉翊安__不分系112__F64081020

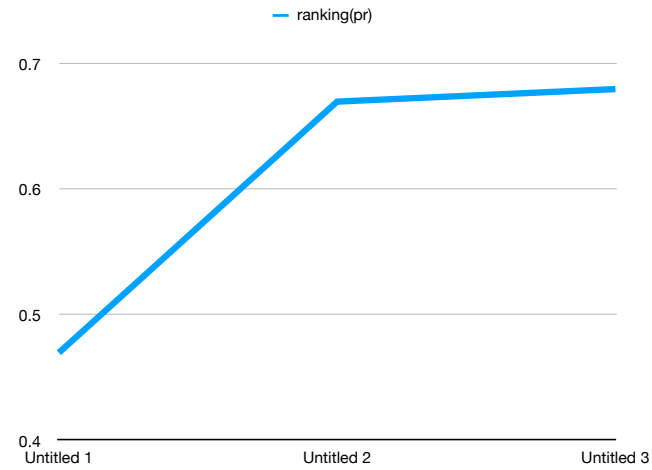
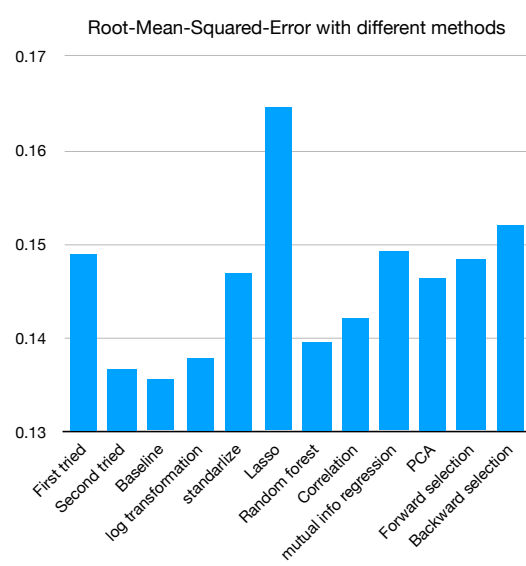
test	train_x	train_y
(1459, 204)	(1460, 204)	(1460,)

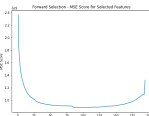
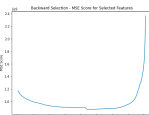
My best score

Best score : 0.13573

Filling missing value		
	Filling missing value	Drop
Method 1	Int -> Mean Object -> 0	More than 70% 5 features
Method 2	Decide with observation ex: Int -> categorical ex: fireplace -> FireplaceQu	More than 75% 4 features

	Missing value	One hot encoding	Data processing	Feature selection	ML algorithm	result	ranking(pr)
First tried	M0	V	X	X	XGBoost	0.14906	0.47
Second tried	M1	V	X	X	XGBoost	0.13672	0.67
Baseline	m1	V	X	X	XGBoost+randomize CV	0.13573	0.68
log transformation	m1	V	Log transformation	X	XGBoost	0.13791	
standarlize	m1	V	standarlize	X	XGBoost	0.14687	
Lasso	m1	V	X	Lasso	XGBoost	0.16443	
Random forest	m1	V	X	Random forest	XGBoost	0.1397	
Correlation	m1	V	X	Correlation	XGBoost	0.14215	
mutual info regression	m1	V	X	mutual info regression	XGBoost	0.14925	
PCA	m1	V	X	PCA	XGBoost	0.14652	
Forward selection	m1	V	X	Forward selection	XGBoost	0.14843	
Backward selection	m1	V	X	Backward selection	XGBoost	0.1522	



	選到第幾個獲得最好MSE	CV	圖片	選擇的features	Result
Forward selection	92	3	去除最後10個值 (太大) 	OverallQual GrLivArea BsmtFinSF1 GarageCars NridgHt 20 No StoneBr NoRidge YearRemodAdd Fireplaces WdShngl New Norm 160 Somerst	0.14843
Backward selection	115(被選擇的features是後88個)	3		OverallQual 1stFirSF 2ndFirSF YearBuilt NridgHt BsmtFullBath OverallCond StoneBr NoRidge TA Gd WdShngl Twnhs TwnhsE GarageCars Norm	0.1522

Random forest	
Output value	(1460, 17)
Keeped feature	['LotFrontage', 'LotArea', 'OverallQual', 'OverallCond', 'YearBuilt', 'YearRemodAdd', 'BsmtFinSF1', 'BsmtUnfSF', 'TotalBsmtSF', '1stFlrSF', '2ndFlrSF', 'GrLivArea', 'Fireplaces', 'GarageYrBlt', 'GarageCars', 'GarageArea', 'Y']
Result(trained by XGBoost)	0.1397

correlation	
Output dataframe	(1460, 189)
Deleted feature	{ '1.5Unf', '2Story', '2fmCon', 'CmentBd', 'Duplex', 'GarageArea', 'Hip', 'None', 'Partial', 'RM', 'SLvl', 'Somerst', 'TA', 'TotRmsAbvGrd', 'Unf' }
Result(trained by XGBoost)	

Set	mutual information > 0.0001
Output dataframe	(1460, 163)
Deleted feature	{'10', '11', '12', '150', '1Story', '2010', '4', '40', '6', '75', 'Alloca', 'Basment', 'Blueste', 'BrkFace', 'CBlock', 'CWD', 'CompShg', 'ConLI', 'ConLw', 'FR2', 'Family', 'FuseP', 'Gambrel', 'GasW', 'Mansard', 'NoSeWa', 'Norm', 'Oth', 'Other', 'Pave', 'Po', 'RRAn', 'RRNe', 'RRNn', 'Roll', 'Sev', 'Shed', 'Stone', 'Stucco', 'Wd Shng', 'WdShing'}
Note	Set mutual information > 0.0001 CUZ hope to keep 80% feature
Result(trained by XGBoost)	0.14925

Lasso					
Alpha	3E-05	5E-05	7E-05	9E-05	0.0001
Lasso avg	0.876836177927236	0.883983384435943	0.888728993506497	0.892541187672200	0.893544144114925
Try	alpha=0.00003	alpha = 0.00003, threshold = 60, no log transformation, no standardize			
Output shape	(1460, 182)				
Deleted feature	{'150', '190', '2.5Fin', '2008', '2fmCon', '40', 'AsphShn', 'ConLw', 'Duplex', 'Fa', 'FuseF', 'FuseP', 'GasA', 'IR3', 'Mix', 'Partial', 'RRNe', 'SWISU', 'Shed', 'Stone', 'Timber', 'TwnhsE'}	{'150', '1stFlrSF', '2ndFlrSF', '3SsnPorch', 'BsmtFinSF1', 'BsmtFinSF2', 'BsmtUnfSF', 'EnclosedPorch', 'GarageArea', 'GarageYrBlt', 'GrLivArea', 'LotArea', 'LowQualFinSF', 'MasVnrArea', 'MetalSd', 'MiscVal', 'OpenPorchSF', 'ScreenPorch', 'TotalBsmtSF', 'WoodDeckSF'}			
Result(trained by XGBoost)	0.44575				