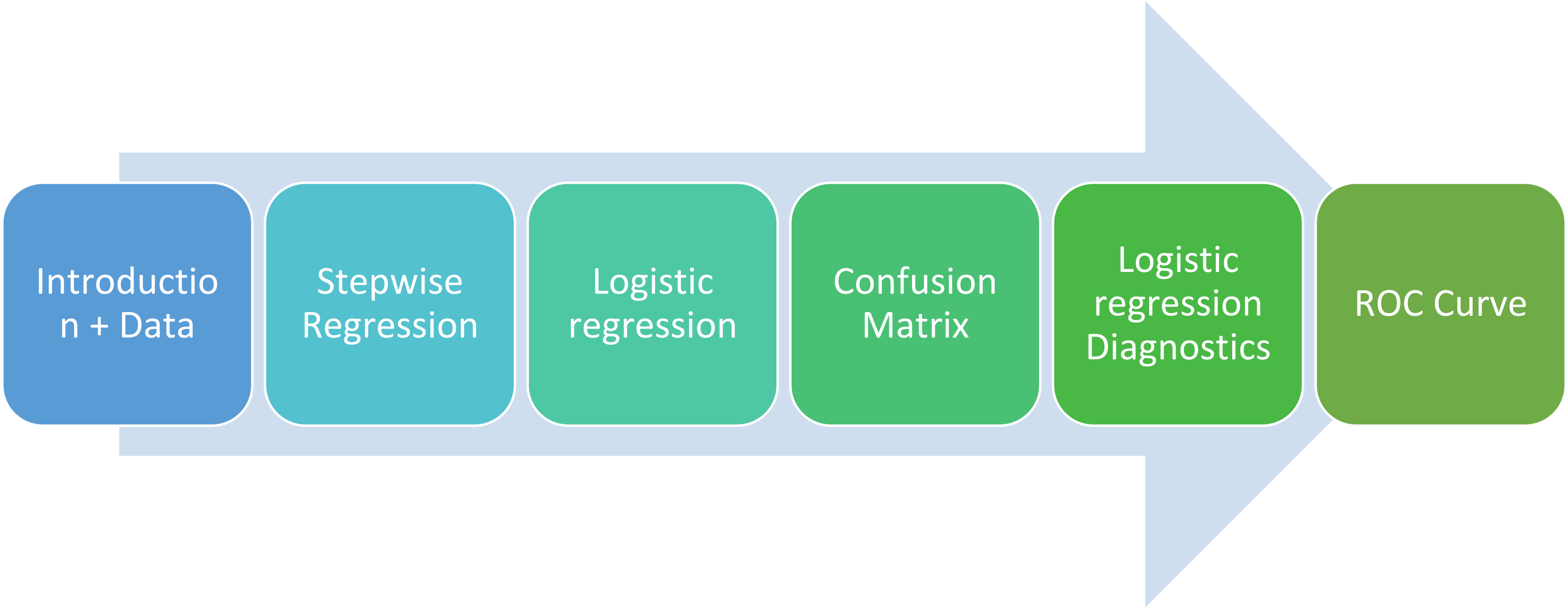
Predicting Foreign Exchange Rate Movement

Ke Cao

Agenda


Introductio
n + Data
Stepwise Regression
Logistic regression
Confusion Matrix
Logistic regression Diagnostics
ROC Curve

Introduction + Data

The SAS System

Obs	rsi1	rsi2	rsi3	rsi4	rsi5	rsi6	stoch1	stoch2	stoch3	stoch4	stoch5	stoch6	ema20Slope1	ema20Slope2	ema20Slope3	ema20Slope4
1	28.9	31.07	40.01	40.51	39.95	41.98	13.53	29.27	46.8	43.52	41.03	36.07	-0.0005	-0.00045	-0.00015	-0.00015
2	27.39	28.9	31.07	40.01	40.51	39.95	3.93	13.53	29.27	46.8	43.52	41.03	-0.00052	-0.0005	-0.00045	-0.00015
3	28.41	27.39	28.9	31.07	40.01	40.51	4.27	3.93	13.53	29.27	46.8	43.52	-0.00046	-0.00052	-0.0005	-0.00045
4	34.48	28.41	27.39	28.9	31.07	40.01	12.99	4.27	3.93	13.53	29.27	46.8	-0.0003	-0.00046	-0.00052	-0.0005
5	33.35	34.48	28.41	27.39	28.9	31.07	24.48	12.99	4.27	3.93	13.53	29.27	-0.00031	-0.0003	-0.00046	-0.00052
6	31.96	33.35	34.48	28.41	27.39	28.9	36.23	24.48	12.99	4.27	3.93	13.53	-0.00034	-0.00031	-0.0003	-0.00046
7	31.59	31.96	33.35	34.48	28.41	27.39	42.59	36.23	24.48	12.99	4.27	3.93	-0.00032	-0.00034	-0.00031	-0.0003
8	29.34	31.59	31.96	33.35	34.48	28.41	25.31	42.59	36.23	24.48	12.99	4.27	-0.00037	-0.00032	-0.00034	-0.00031
9	30.66	29.34	31.59	31.96	33.35	34.48	21.31	25.31	42.59	36.23	24.48	12.99	-0.00031	-0.00037	-0.00032	-0.00034
10	36.37	28.16	30.66	29.34	31.59	31.96	29.81	12.9	21.31	25.31	42.59	36.23	-0.00021	-0.00037	-0.00031	-0.00037
11	38.37	36.37	28.16	30.66	29.34	31.59	45.11	29.81	12.9	21.31	25.31	42.59	-0.00015	-0.00021	-0.00037	-0.00031
12	40.43	38.37	36.37	28.16	30.66	29.34	71.66	45.11	29.81	12.9	21.31	25.31	-0.0001	-0.00015	-0.00021	-0.00037
13	39.43	40.43	38.37	36.37	28.16	30.66	77.53	71.66	45.11	29.81	12.9	21.31	-0.00012	-0.0001	-0.00015	-0.00021

- The purpose of the project is to predict the foreign exchange movement
- FX (EUD to USD) from Kaggle
- Dataset contains 4479 transactions from 2014 to 2017.
- The dataset includes different technical indicators such as EMAs, RSI, MOM

Stepwise regression

Summary of Stepwise Selection								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	close1		1	0.0023	0.0023	10.6340	10.32	0.0013
2	force3		2	0.0012	0.0035	7.3295	5.30	0.0214
3	bearsPower4		3	0.0013	0.0048	3.4418	5.89	0.0153
4	WPR5		4	0.0013	0.0061	-0.2722	5.72	0.0168
5	BB_up_percen5		5	0.0008	0.0069	-1.9221	3.66	0.0559
6	dayOfWeek		6	0.0007	0.0075	-2.8998	2.98	0.0841
7	bullsPower4		7	0.0006	0.0081	-3.3781	2.48	0.1150
8	mom1		8	0.0006	0.0087	-4.0498	2.68	0.1017

- is the model by adding or removing variables based on the t-statistics of the estimated coefficient
- In this dataset, the indicators EMAS, BB, and RSI are independent variables
- dependent variable tipo consists of two values, 0 and 1, indicating whether to buy or sell.
- Initially, we intend to perform the feature selection in both directions
- Then we find the best model through 8 steps: close1, force3, bearpower4, WPR5, BB Up Percen5, day of week, bullspower4, and mom1 by adding one by one

- Specify an Alpha-to-Enter significance level. This will typically be greater than the usual 0.05 level -set this significance level by default to $\alpha_E = 0.15$.
- All variables are significant at the 0.15 level

Logistic Regression

- we will perform the logistic regression. Here, is the assumption of logistic regression
- Logistic regression is a predictive analysis, which conducts when the dependent variable is binary (like tipo 0, 1) and explain the relationship between on tipo 0,1 and more independent variables (emas, rsi…)
- The assumptions of logistic regression are: 1. The model should be fitted correctly. It is only for meaningful variables, so no important variables are omitted, no extraneous variables are included, independent variables are measured without error. 2. The error terms need to be independent and each observations to be independent. 3. The independent variables are not linear combination of each other. Perfect multicollinearity makes estimation inaccurate.

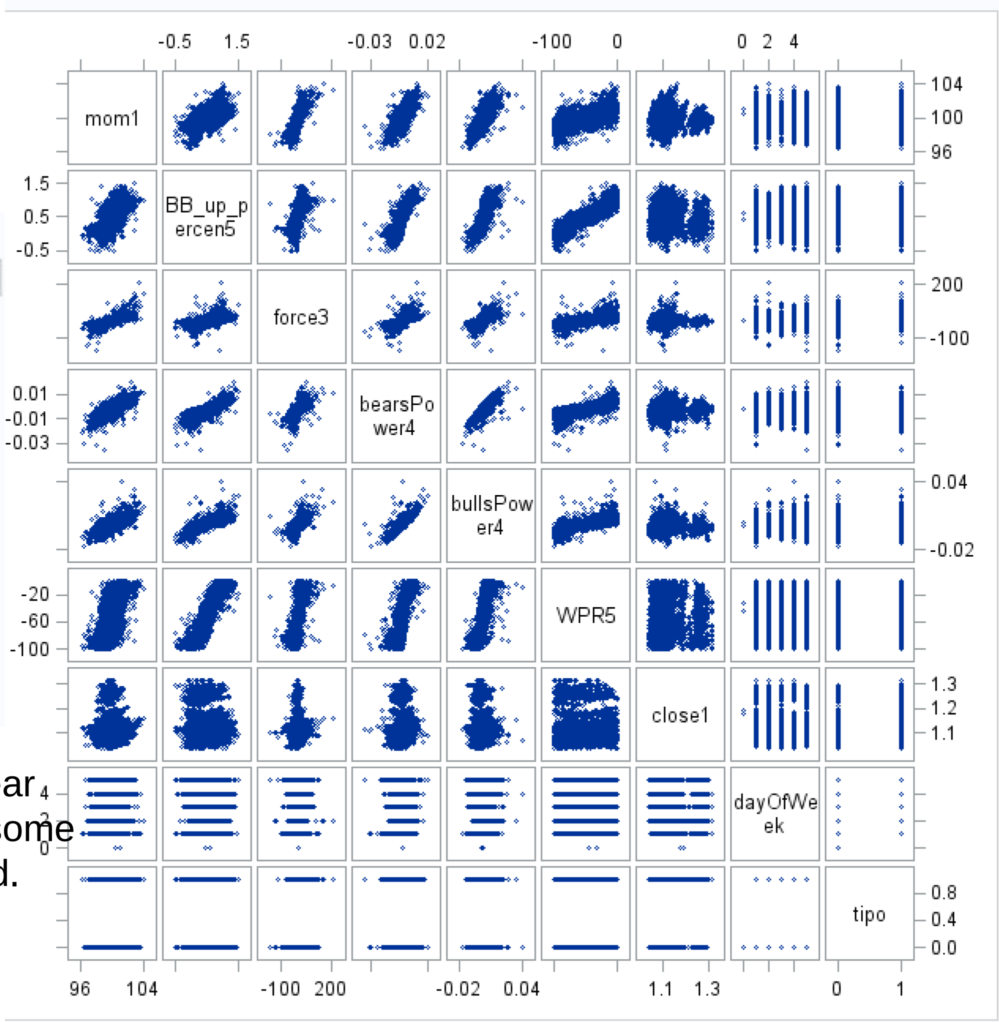
Correlation Matrix

The CORR Procedure

9 Variables:	mom1 BB_up_percen5 force3 bearsPower4 bullsPower4 WPR5 close1 dayOfWeek tipo

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
mom1	4479	99.97845	0.90014	447803	96.31000	104.10000
BB_up_percen5	4479	0.49022	0.33902	2196	-0.52000	1.51000
force3	4479	0.09733	18.77816	435.95000	-169.04000	213.44000
bearsPower4	4479	-0.00198	0.00431	-8.84723	-0.03596	0.01966
bullsPower4	4479	0.00177	0.00446	7.90682	-0.01702	0.04071
WPR5	4479	-52.21570	29.22139	-233874	-100.00000	0
close1	4479	1.12441	0.05751	5036	1.03690	1.31490
dayOfWeek	4479	2.99353	1.41380	13408	0	5.00000
tipo	4479	0.49297	0.50001	2208	0	1.00000



Based on the chart, we can state it exists strong linear Relationship between variables, which implies that some Assumption of logistic regression cannot be satisfied. We can check how such multilinearity impacts our logistic regression

Data

Using the equal-scale stratified sampling,we divided our data into two parts, training data and test data.

All evaluation indicators are calculated in the test data. The dependent variable that we are interested in is tipo, which is 0 for buy, 1 for sell.

The training data account for 80% of the total data, while the test data take up 20%.

Train Data			
tipo	Frequency	Percent	Total
0(Buy)	1817	0.506975	3584
1(Sell)	1767	0.493025	
Test Data			
tipo	Frequency	Percent	Total
0(Buy)	450	0.502793	895
1(Sell)	445	0.497207	

Logistic Regression Model

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-6.2040	6.7821	0.8368	0.3603
mom1	1	0.0717	0.0673	1.1362	0.2865
BB_up_percen5	1	0.6915	0.2783	6.1752	0.0130
force3	1	-0.00642	0.00276	5.3956	0.0202
bearsPower4	1	24.8352	15.2461	2.6535	0.1033
bullsPower4	1	-18.8840	14.7430	1.6407	0.2002
WPR5	1	-0.00815	0.00289	7.9756	0.0047
close1	1	-1.3145	0.5844	5.0600	0.0245
dayOfWeek	1	-0.0467	0.0239	3.8209	0.0506

Regression Function:

logit_sell=-3.2803+mom1*0.0460+BB_up_percen5*0.3875+force3*(-0.00548)+bearsPower4*31.8045+bullsPower4 *(-14.0762)

+WPR5*(-0.00633)+close1*(-1.4405)+dayOfWeek*(-0.0351)

Here is the summary of trained logistic regression. We can say some intercepts are not statistically significant.

Prediction

```
p=exp(logit_sell)/(exp(logit_sell)+1);
if p<0.5 then sell_buy_predicted='buy ';
else sell_buy_predicted='sell';
```

The formula is to calculate the probability value. How to classify each operation according to the probability value? We need a threshold value. Here we provide that when the probability exceeds 0.5, it is classified as sell and the rest is buy.

Results of Prediction

Obs	tipo	p	sell_buy_predicted
1	0	0.48879	buy
2	0	0.50067	sell
3	0	0.47700	buy
4	0	0.41775	buy
5	0	0.43432	buy
6	0	0.45120	buy
7	0	0.44119	buy
8	0	0.44660	buy
9	0	0.44808	buy
10	0	0.41986	buy
11	0	0.43180	buy
12	0	0.43327	buy
13	0	0.46106	buy
14	0	0.48621	buy
15	0	0.48576	buy
16	0	0.41737	buy
17	0	0.46394	buy
18	0	0.50443	sell
19	0	0.45190	buy

Put the previous formula into the test data. We get the results.

From the local data test_p, we can see that some value for tipo which are actually 0(buy), according to our model (threshold p takes 0.5), predict it as sell (Type I Error). While for some value which were originally 1(sell), but it is predicted to buy (Type II Error).

Confusion Matrix

A perfect classification model is that if a operation actually belongs to 0(buy), it is also predicted to be buy, while in the 1(sell), it is predicted to be sell. However, from the above we have seen that some operation, which are actually buy, based on our model, predicted that they are sell. For some of the operations which were originally sell, they are expected to buy. We need to know that this model predicts exactly how many the predictions are correct and how many the prediction are wrong. The confusion matrix puts all this information into a single table:

Table of tipo by sell_buy_predicted

tipo	sell_buy_predicted		
	buy	sell	Total
0	d 158	c 296	d+c 454
1	b 196	a 245	a+b 441
Total	b+d 354	a+c 541	895

1. a is the number of negative cases correctly predicted, True Negative(TN,0->0)
2. b is the number of negative cases predicted to be positive, False Positive(FP, 0->1)
3. c is the number of positive cases predicted to be negative, False Negative(FN, 1->0)
4. d is the number of correctly predicted positive examples, True Positive (TP, 1->1)
5. a+b is the actual number of negative cases, Actual Negative
6. c+d is the actual number of positive examples, Actual Positive
7. a+c is the number of negative cases predicted, Predicted Negative
8. b+d is the number of positive cases predicted, Predicted Positive

Evaluation

According to the table, there are several commonly used evaluation indicators:

Accuracy=true positive and true negative/total cases= a+d/a+b+c+d=17.65%+27.37%=45.02%

Error rate=false positive and false negative/total cases=b+c/a+b+c+d=1-Accuracy=54.98%

Recall=true positive/total actual positive=d/c+d=34.80%

Precision=true positive/ total predicted positive=d/b+d=44.63%

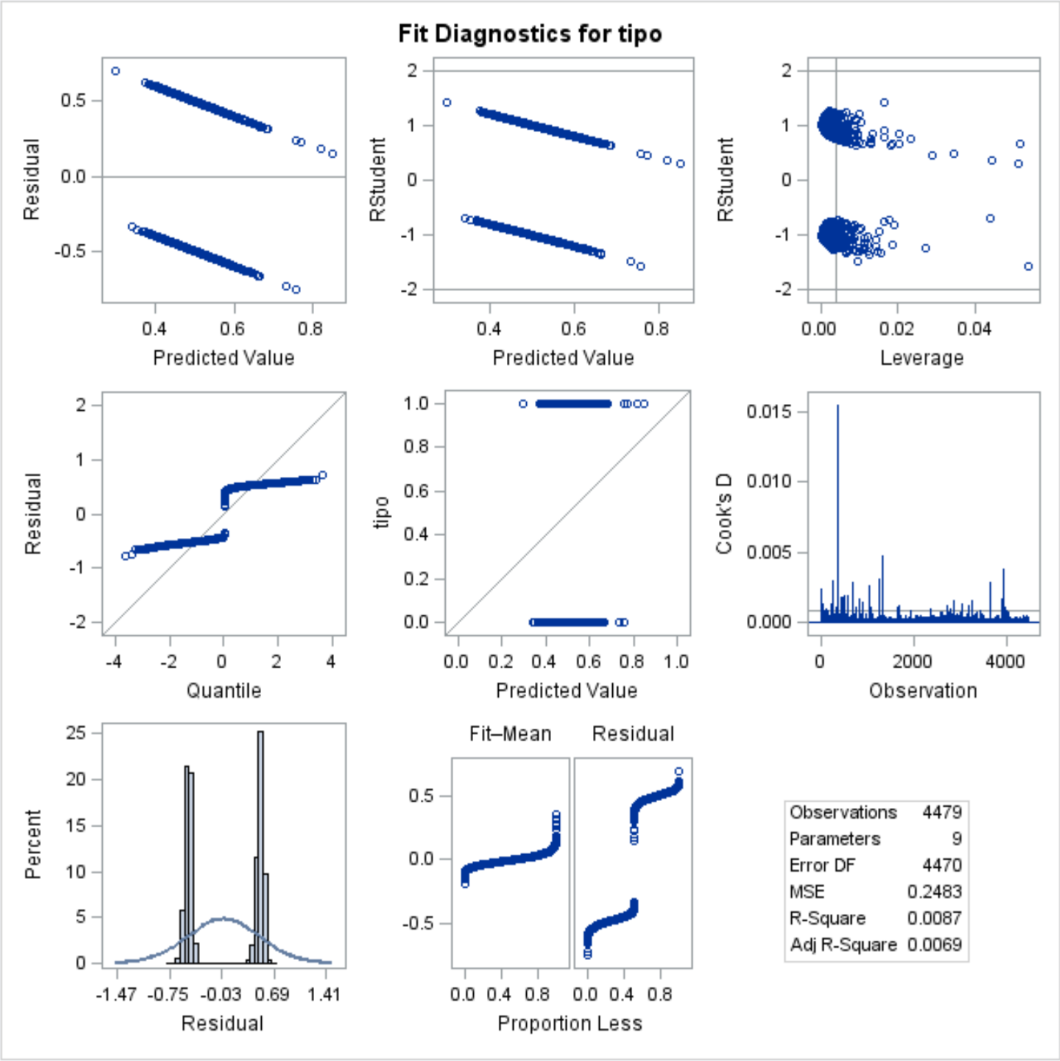
Frequency
Percent
Row Pct
Col Pct

Table of tipo by sell_buy_predicted

tipo	sell_buy_predicted: buy	sell	Total
0	158 17.65 34.80 44.63	296 33.07 65.20 54.71	454 50.73
1	196 21.90 44.44 55.37	245 27.37 55.56 45.29	441 49.27
Total	354 39.55	541 60.45	895 100.00

Logistic Regression Diagnostics

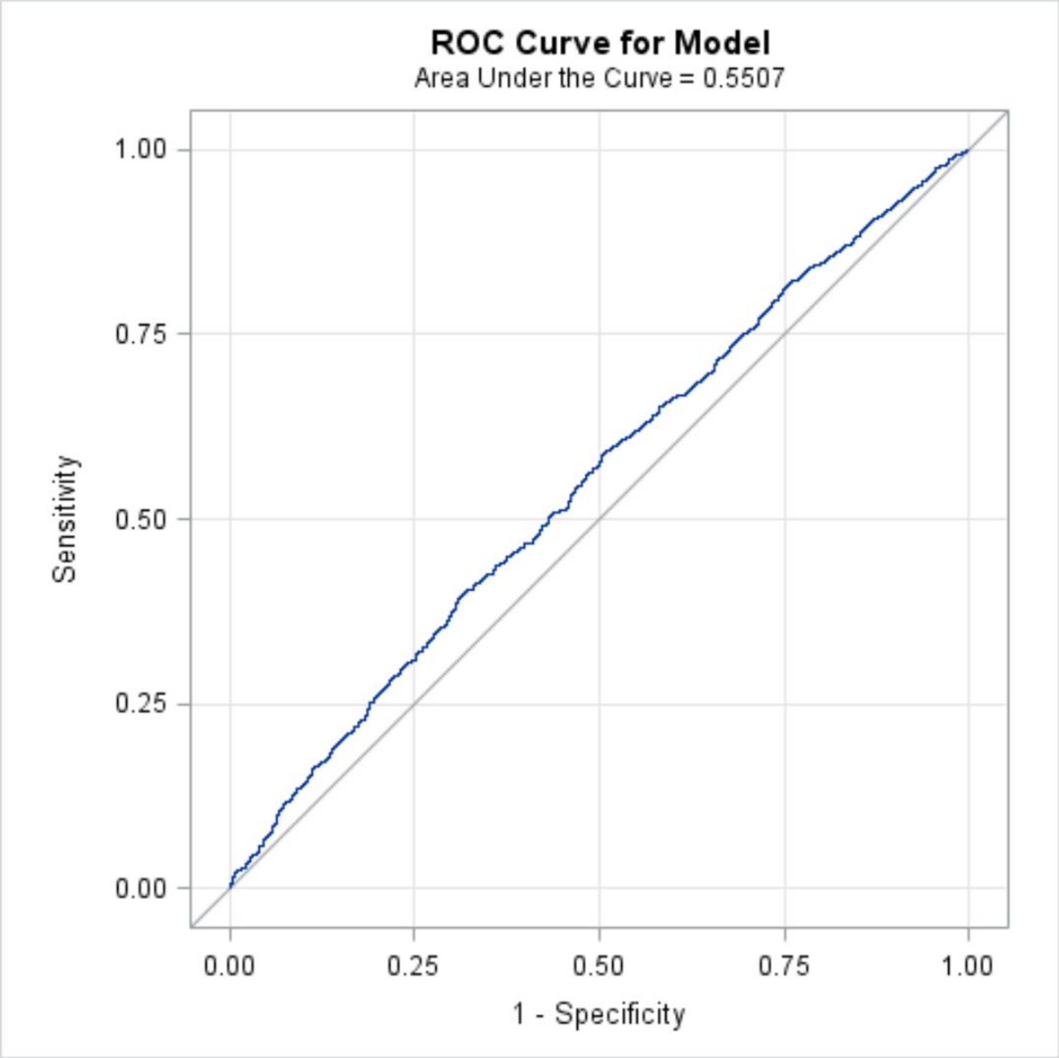
- Residuals are certainly less informative for Logistic regression then Linear Regression.
- Normality is not required.
- Outliers exist in our data.



ROC Curve

The area under the curve is 0.5507 and the cut probability we might choose is around 0.5. Since, the true positive (max = 0.5) and false positive (min = 0.3).

In this, there might be a case of multicollinearity. So, we can use PCA.



PCA

- PCA stands for Principal Component Analysis
- Principal Component Analysis is a variable reduction procedure. It is useful when you have obtained data on a number of variables (possibly a large number of variables), and believe that there is some correlation among those variables.
- We can apply PCA in our analysis as in Logistic Regression we have found the multicollinearity in variables.

PCA

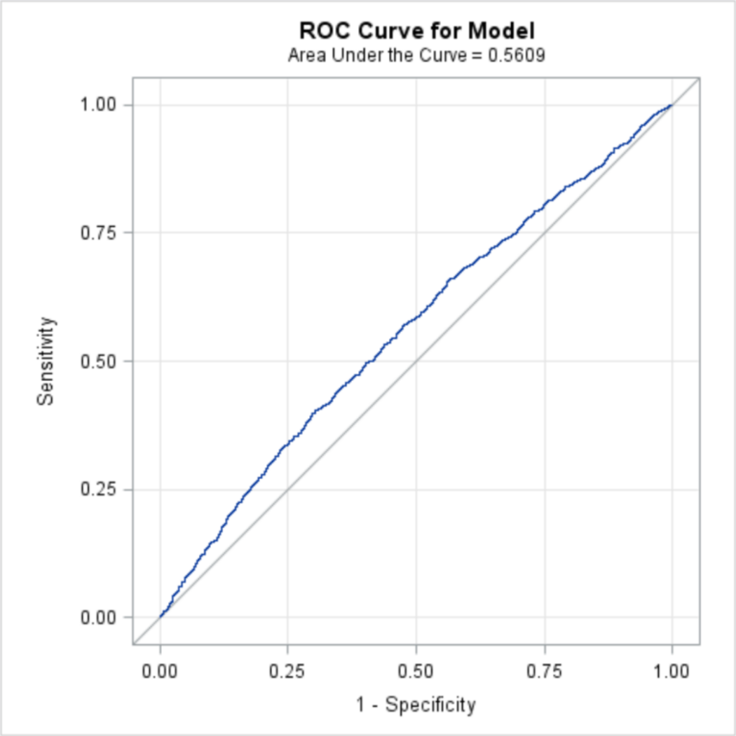
- According to Kaiser criterion, Eigenvalues should be greater than 1. Hence, we can observe that the top 3 variables have high Eigenvalues implies these are good.
- Approximately 82% of the data is explained by the top 3 principal component variables.

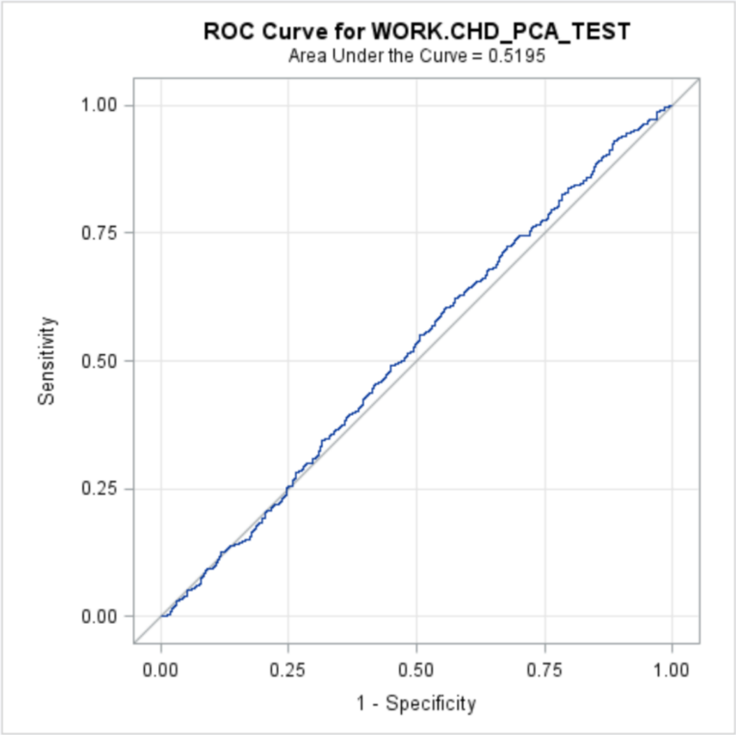
Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	4.59668414	3.55845812	0.5746	0.5746
2	1.03822602	0.07183728	0.1298	0.7044
3	0.96638874	0.38748470	0.1208	0.8252
4	0.57890405	0.30274849	0.0724	0.8975
5	0.27615555	0.04112457	0.0345	0.9320
6	0.23503098	0.01279843	0.0294	0.9614
7	0.22223255	0.13585458	0.0278	0.9892
8	0.08637797		0.0108	1.0000

PCA

- The eigenvectors indicate the relative importance of each variable within the individual.
- First three principal component variables indicates the majority of the variability among these eight variables (as Eigenvalues for these variables were good).
- We will select all the variables which have principal component greater than 0.4

Eigenvectors								
	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8
mom1	0.397152	0.005869	-.045199	0.459634	-.337079	-.622941	-.354586	0.038984
BB_up_percen5	0.424787	-.039216	0.045747	-.412820	0.232455	-.122882	0.012538	0.759086
force3	0.372078	0.004622	-.093106	0.648964	0.520360	0.281790	0.283726	0.032145
bearsPower4	0.416017	-.014888	-.011603	-.117427	-.567028	0.088794	0.684815	-.120032
bullsPower4	0.418897	-.018047	0.017714	-.077244	-.240198	0.655973	-.567405	-.089307
WPR5	0.417160	-.002104	0.052777	-.402194	0.427009	-.281225	-.042118	-.631054
close1	0.013841	0.723031	-.682453	-.103186	-.000730	0.001808	-.020073	0.015468
dayOfWeek	0.025000	0.689261	0.719875	0.070702	-.003862	0.012763	0.022694	0.019558





We can notice from the ROC curves for the model and the test dataset. The area under the curve is almost the same. Hence the PCA result does not satisfy our demand.

Thank You!