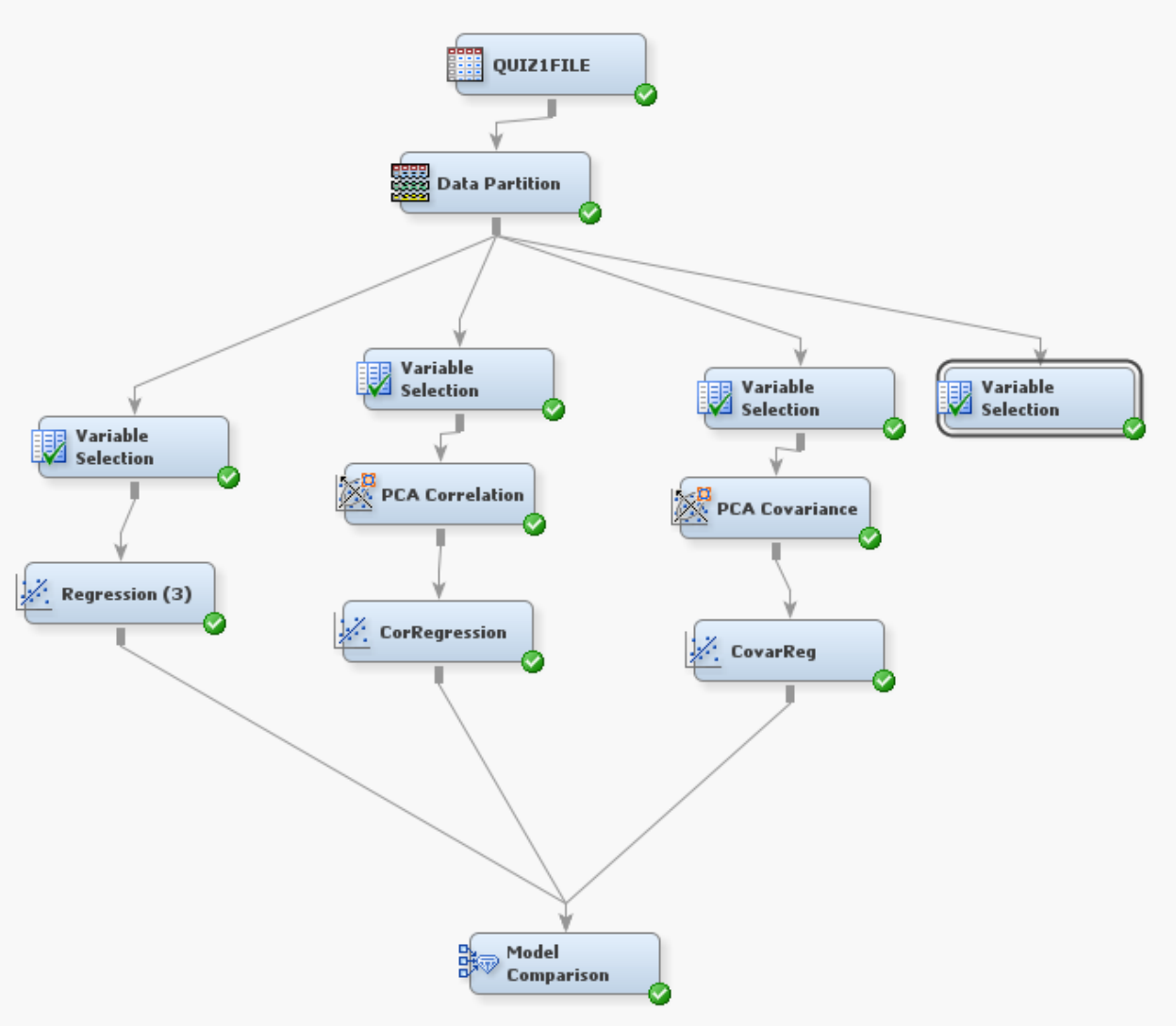
Diane Nguyen
Quiz1

Use the given SAS file to finish the following. The file contains donor data from the last campaign. The objective is to use the predictive model you have created to target the most likely donors.

- Create a SAS Enterprise Miner project to predict probability of donation given the donor data. Use Target B and ignore/reject Target D.



- Determine the appropriate level of each variable given the description and the data profile.

Name	Role	Level	Report	Order	Drop
CARD_PROM_12	Input	Interval	No		No
CLUSTER_CODE	Input	Nominal	No		No
CONTROL_NUMBER	Input	Nominal	No		No
DONOR_AGE	Input	Interval	No		No
DONOR_GENDER	Input	Nominal	No		No
FILE_AVG_GIFT	Input	Interval	No		No
FILE_CARD_GIFT	Input	Interval	No		No
FREQUENCY_STATUS_97NK	Input	Interval	No		No
HOME_OWNER	Input	Binary	No		No
INCOME_GROUP	Input	Interval	No		No
IN_HOUSE	Input	Binary	No		No
LAST_GIFT_AMT	Input	Interval	No		No
LIFETIME_AVG_GIFT_AMT	Input	Interval	No		No
LIFETIME_CARD_PROM	Input	Interval	No		No
LIFETIME_GIFT_AMOUNT	Input	Interval	No		No
LIFETIME_GIFT_COUNT	Input	Interval	No		No
LIFETIME_GIFT_RANGE	Input	Interval	No		No
LIFETIME_MAX_GIFT_AMT	Input	Interval	No		No
LIFETIME_MIN_GIFT_AMT	Input	Interval	No		No
LIFETIME_PROM	Input	Interval	No		No
MEDIAN_HOME_VALUE	Input	Interval	No		No
MEDIAN_HOUSEHOLD_INCOME	Input	Interval	No		No
MONTHS_SINCE_FIRST_GIFT	Input	Interval	No		No
MONTHS_SINCE_LAST_GIFT	Input	Interval	No		No
MONTHS_SINCE_LAST_PROM_RESP	Input	Interval	No		No
MONTHS_SINCE_ORIGIN	Input	Interval	No		No
MOR_HIT_RATE	Input	Interval	No		No
NUMBER_PROM_12	Input	Interval	No		No
OVERLAY_SOURCE	Input	Nominal	No		No
PCT_ATTRIBUTE1	Input	Interval	No		No
PCT_ATTRIBUTE2	Input	Interval	No		No
PCT_ATTRIBUTE3	Input	Interval	No		No
PCT_ATTRIBUTE4	Input	Interval	No		No
PCT_OWNER_OCCUPIED	Input	Interval	No		No
PEP_STAR	Input	Binary	No		No
PER_CAPITA_INCOME	Input	Interval	No		No
PUBLISHED_PHONE	Input	Binary	No		No
RECENCY_STATUS_96NK	Input	Nominal	No		No
RECENT_AVG_CARD_GIFT_AMT	Input	Interval	No		No
RECENT_AVG_GIFT_AMT	Input	Interval	No		No
RECENT_CARD_RESPONSE_COUNT	Input	Interval	No		No
RECENT_CARD_RESPONSE_PROP	Input	Interval	No		No
RECENT_RESPONSE_COUNT	Input	Interval	No		No
RECENT_RESPONSE_PROP	Input	Interval	No		No
RECENT_STAR_STATUS	Input	Binary	No		No
SES	Input	Nominal	No		No
TARGET_B	Target	Binary	No		No
TARGET_D	Input	Interval	No		No
URBANICITY	Input	Nominal	No		No
WEALTH_RATING	Input	Interval	No		No

- You must use techniques we have covered so far, including variable selection and principal components, to try to improve the performance of prediction. If you reject or recommend a particular technique, you must support your rejection/recommendation with at least two appropriate metrics.

    - First off you must do variable selection in order to get your results to run. As seen in the data, there are over 500 variables. This exceed the maximum limit for Regression. The variable selection node reduced your number of input variables to six. The rest were eliminated for reasons such as an r-square value not meeting a threshold value

Variable Name	Role ▲	Measurement Level	Type	Label	Reasons for Rejection
FREQUENCY STATUS 97NK	Input	Interval	Numeric		
G CLUSTER CODE	Input	Nominal	Numeric	Grouped Levels for CLU...	
G RECENCY STATUS 96NK	Input	Nominal	Numeric	Grouped Levels for RE...	
G RECENT STAR STATUS	Input	Nominal	Numeric	Grouped Levels for RE...	
MONTHS SINCE LAST GIFT	Input	Interval	Numeric		
PEP STAR	Input	Binary	Numeric		
RECENT CARD RESPONSE CO...	Input	Interval	Numeric		

    - According to SAS the correlation PCA with variable selection is the selected model to use based on the r-square value that was above a certain threshold and the misclassification rate. I would recommend the PCA correlation that utilized variable selection. It has a low root average squared error and it has the lowest misclassification rate out of the three regressions.

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Valid: Root Average Squared Error	Valid: Root Mean Square Error	Valid: Misclassification Rate	Valid: Sum of Square Errors	Selection Criterion: Valid: Misclassification Rate
	Selected Model									
Y	Reg	Reg	CorRegression	TARGET...	Target Variable Indi...	0.427585	0.427585	0.249161	2833.85	0.249161
	Reg3	Reg3	Regression (3)	TARGET...	Target Variable Indi...	0.42741	0.42741	0.24929	2831.526	0.24929
	Reg2	Reg2	CovarReg	TARGET...	Target Variable Indi...	0.42839	0.42839	0.249935	2844.535	0.249935

    - I would reject regular regression without principle component analysis. PCA is a method of extracting important variables (in form of components) from a large set of variables available in a data set. It extracts low dimensional set of features from a high dimensional data set in an attempt to capture as much information as it can. PCA is more useful when dealing with 3 or higher dimensional data. It is always performed on a symmetric correlation or covariance matrix. This means the matrix should be numeric and have standardized data. In terms of standardization, all your data should be standardized to measure in the same units. For example you may want to standardize all of your data in terms of a z-score distribution. The purpose of PCA is to try to reduce the number of variables you have, in the first run of the covariance PCA you can see that it accounts for over 99% of your data. Instead you will want to use PCA for correlation it normalizes your data for you.
    - I would reject PCA for covariance, as you can see from the eigenvalues below, we have more of our values explained over multiple runs. Also from the results above the correlation regression using PCA has a lower root average squared error and a lower misclassification rate than covariance PCA regression.

Eigenvalues of the Covariance Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	16.437502	13.384194	0.7449	0.7449
2	3.053308	2.440311	0.1384	0.8833
3	0.612997	0.082576	0.0278	0.9111
4	0.530421	0.205638	0.0240	0.9351
5	0.324783	0.137003	0.0147	0.9499
6	0.187780	0.016743	0.0085	0.9584
7	0.171037	0.008177	0.0078	0.9661
8	0.162860	0.038958	0.0074	0.9735
9	0.123902	0.011286	0.0056	0.9791
10	0.112616	0.007131	0.0051	0.9842
11	0.105485	0.038207	0.0048	0.9890
12	0.067278	0.000855	0.0030	0.9920
13	0.066423	0.018481	0.0030	0.9951
14	0.047943	0.015430	0.0022	0.9972
15	0.032513	0.015935	0.0015	0.9987
16	0.016579	0.005319	0.0008	0.9995
17	0.011260	0.010513	0.0005	1.0000
18	0.000747	0.000747	0.0000	1.0000
19	0	0	0.0000	1.0000
20	0	0	0.0000	1.0000
21	0	0	0.0000	1.0000
22	0		0.0000	1.0000

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	4.815062	2.982969	0.2189	0.2189
2	1.832093	0.435019	0.0833	0.3021
3	1.397074	0.166057	0.0635	0.3656
4	1.231017	0.020612	0.0560	0.4216
5	1.210405	0.015338	0.0550	0.4766
6	1.195067	0.037383	0.0543	0.5309
7	1.157684	0.026266	0.0526	0.5836
8	1.131418	0.039318	0.0514	0.6350
9	1.092100	0.018990	0.0496	0.6846
10	1.073111	0.036981	0.0488	0.7334
11	1.036129	0.017591	0.0471	0.7805
12	1.018538	0.020531	0.0463	0.8268
13	0.998007	0.041772	0.0454	0.8722
14	0.956235	0.215614	0.0435	0.9156
15	0.740621	0.034786	0.0337	0.9493
16	0.705834	0.381277	0.0321	0.9814
17	0.324558	0.239513	0.0148	0.9961
18	0.085044	0.085044	0.0039	1.0000
19	0	0	0.0000	1.0000
20	0	0	0.0000	1.0000
21	0	0	0.0000	1.0000
22	0		0.0000	1.0000

- Discuss a scenario in which the LIFT chart can be used to demonstrate the benefit of a predictive model as opposed to randomly sending out promotions.

Lift is a measure of the effectiveness of a predictive model calculated as the ratio between the results obtained with and without the predictive model. The Cumulative gains and lift charts are visual aids for measuring model performance. Both charts consist of a lift curve and a baseline. The greater the area between the lift curve and the baseline, the better the model. With a predictive model, a company can target their audience rather than blindly sending out mailers to everyone. For example, assume that the company has information on 100,000 customers and that they have a 20% response rate (20,000 positive responses). A response model will predict who will actually respond to the mailing campaign. For example we can use the response model to assign a score to each of the 100,000 customers and predict the results of contacting only the top 20,000 customers. Using the predictions of the response model, we can calculate the percentage of positive responses for the percent of customers contacted and map these points to create the lift curve. For contacting 10% of customers using no model we would only get 10% of responders. But using a lift model we should get about 30% of responders. (30/10=3).