

Predictive Modeling with Claim Fraud Dataset/ SAS Enterprise Miner

Didem B. Aykurt

Colorado State University Global

MIS530; Predictive Analytics

Dr.Jennifer Catalano

January 12, 2023

Claim Fraud Regression Analysis in SAS Enterprise Miner

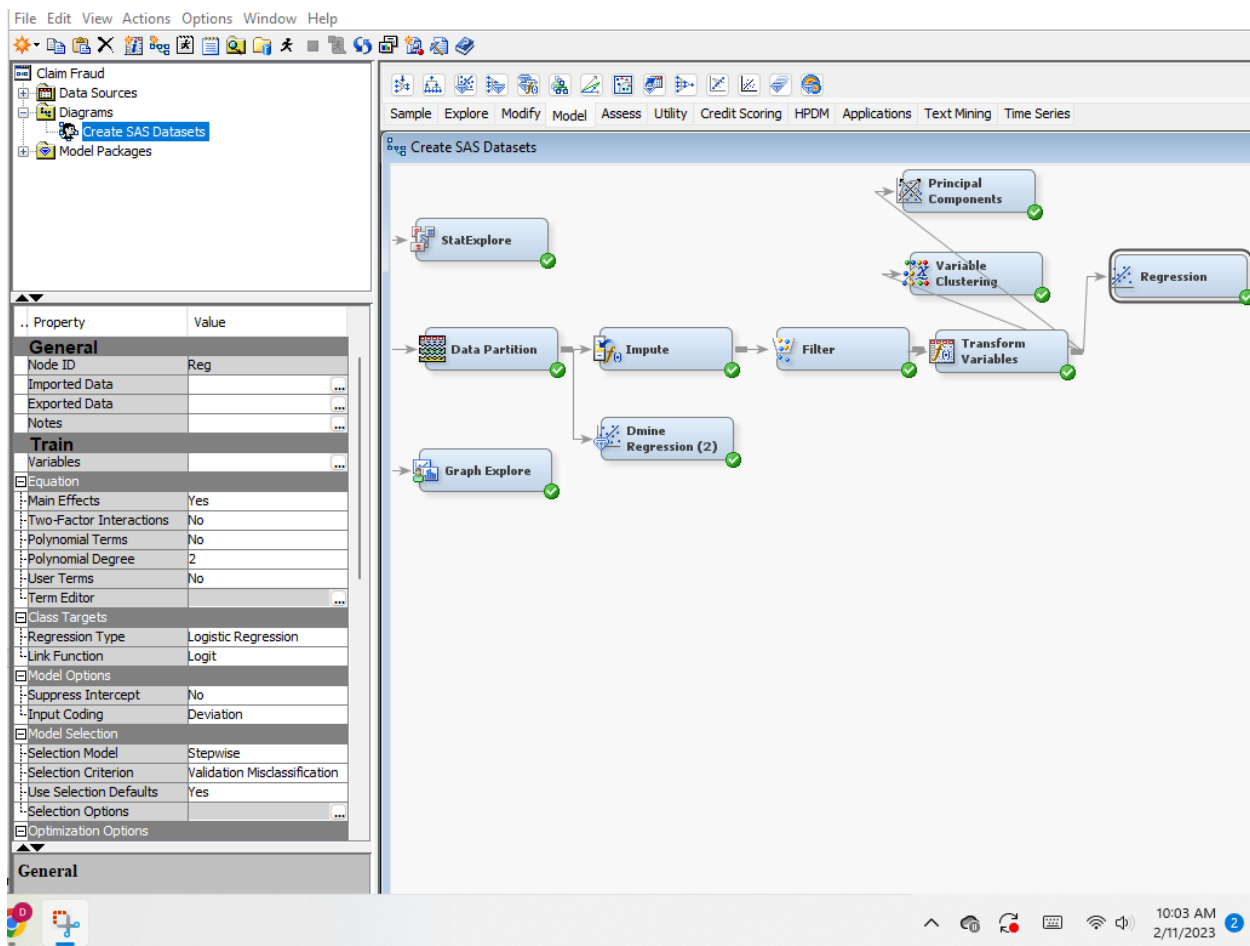
After achieving descriptive statistics and data preparation, the next step is building the predictive model. The most popular model is **linear regression** allows us to measure the strength of the relationship between the response and predictor variable, also known as line fitting and curve fitting; **logistic regression** helps to measure the relationship between a categorical target variable and one or more input variables; **principal component regression** mostly used another option to multiple linear regression model when applying the significant variables or when the variables are most relative; **partial least squares** is mainly used when the data set contains high observation than input variables, and there is multicollinearity. Few statistical results show how the model fits data, so R^2 is the square of the correlation coefficient, and the result range is 0% -100%. Hence, the best result, the higher means best fit the model, adjusted- R^2 help when adding more variables and using multiple regression because r-square increases when adding a variable, and the p-value is used when applying hypothesis testing or tells us that this model is a perfectly supported dataset. In this project, I built predictive analysis with the Claim Fraud dataset into SAS Enterprise Miner.

Regression Analyst

The Regression node is performed for both the linear and logistic regression models. First, I will work on the automobile insurance claim fraud dataset and build logistic regression. I will set logistic regression default when the target variable is binary targets. The linear regression applies a continuous target. Drag and drop the Regression node in the Model tab on the diagram workplace and set the -Regression Type in the Class Target group as Logistic Regression.

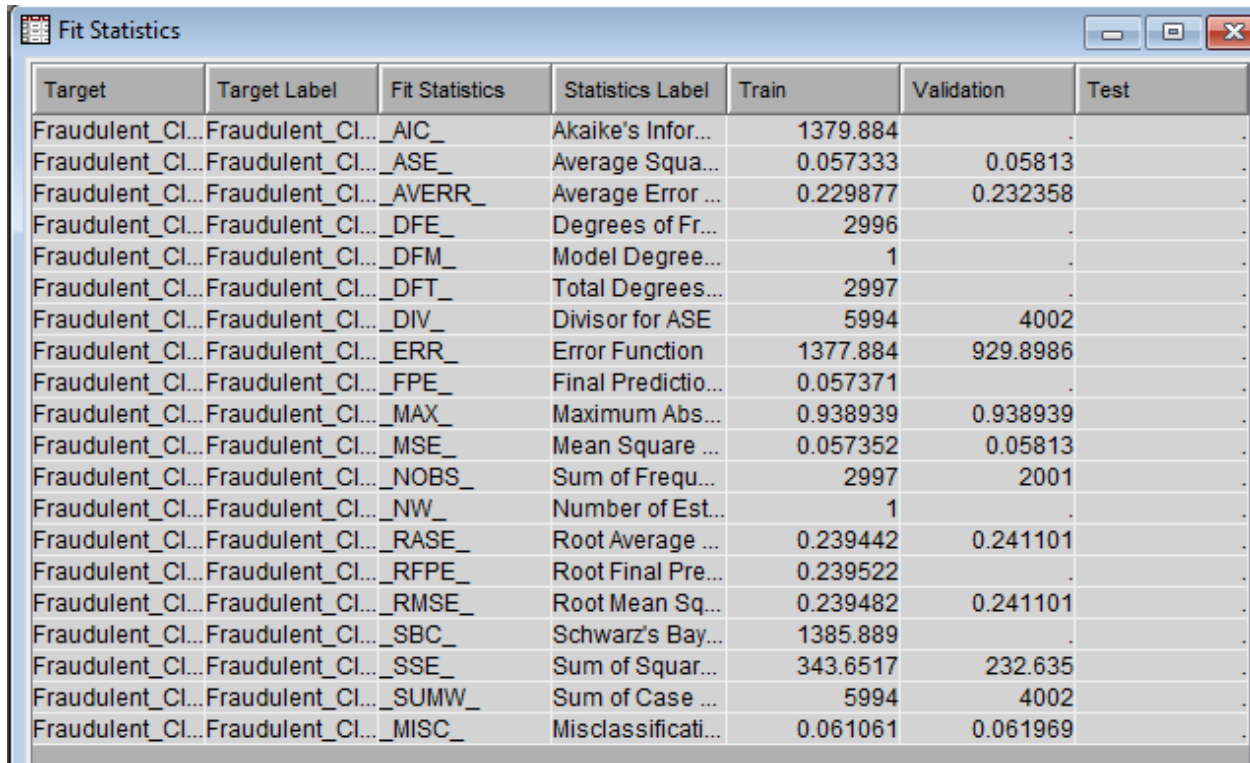
- Link function as Logit (if the linear regression sets the Link function Probit, Cloglog is the change of the cumulative extreme-value function).
- Selection Criterion in the Model Selection in the Train group as Stepwise
- Selection Criteria in the Model Selection in the Train group as Validation Misclassification than the model run.

Figure 1: Regression node and properties



Step two is the average square error of 0.05813 on the Validation partition. Suppose the high difference between the Train and Validation partition is the average squared error is insignificant. In that case, the contrast is 0.000784 for the claim fraud train and validation partition. The model does not appear to be overfitting.

Figure 2: Results of logistic regression-Fit Statistic window.



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Fraudulent_Cl...	Fraudulent_Cl...	_AIC_	Akaike's Infor...	1379.884		.
Fraudulent_Cl...	Fraudulent_Cl...	_ASE_	Average Squa...	0.057333	0.05813	.
Fraudulent_Cl...	Fraudulent_Cl...	_AVERR_	Average Error ...	0.229877	0.232358	.
Fraudulent_Cl...	Fraudulent_Cl...	_DFE_	Degrees of Fr...	2996		.
Fraudulent_Cl...	Fraudulent_Cl...	_DFM_	Model Degree...	1		.
Fraudulent_Cl...	Fraudulent_Cl...	_DFT_	Total Degrees...	2997		.
Fraudulent_Cl...	Fraudulent_Cl...	_DIV_	Divisor for ASE	5994	4002	.
Fraudulent_Cl...	Fraudulent_Cl...	_ERR_	Error Function	1377.884	929.8986	.
Fraudulent_Cl...	Fraudulent_Cl...	_FPE_	Final Predictio...	0.057371		.
Fraudulent_Cl...	Fraudulent_Cl...	_MAX_	Maximum Abs...	0.938939	0.938939	.
Fraudulent_Cl...	Fraudulent_Cl...	_MSE_	Mean Square ...	0.057352	0.05813	.
Fraudulent_Cl...	Fraudulent_Cl...	_NOBS_	Sum of Frequ...	2997	2001	.
Fraudulent_Cl...	Fraudulent_Cl...	_NW_	Number of Est...	1		.
Fraudulent_Cl...	Fraudulent_Cl...	_RASE_	Root Average ...	0.239442	0.241101	.
Fraudulent_Cl...	Fraudulent_Cl...	_RFPE_	Root Final Pre...	0.239522		.
Fraudulent_Cl...	Fraudulent_Cl...	_RMSE_	Root Mean Sq...	0.239482	0.241101	.
Fraudulent_Cl...	Fraudulent_Cl...	_SBC_	Schwarz's Bay...	1385.889		.
Fraudulent_Cl...	Fraudulent_Cl...	_SSE_	Sum of Squar...	343.6517	232.635	.
Fraudulent_Cl...	Fraudulent_Cl...	_SUMW_	Sum of Case ...	5994	4002	.
Fraudulent_Cl...	Fraudulent_Cl...	_MISC_	Misclassificati...	0.061061	0.061969	.

The Stepwise regression model includes the hypothesis test in Step 0 and Step 1. The most significant result is close to zero, and Figure 3 shows that $Pr > \text{ChiSq}$ is < 0.001 means significant input; if the result is near 1, the insignificant input variables can be removed as the input variable is Gender significant, as if the input statistically substantial should be included in further analysis.

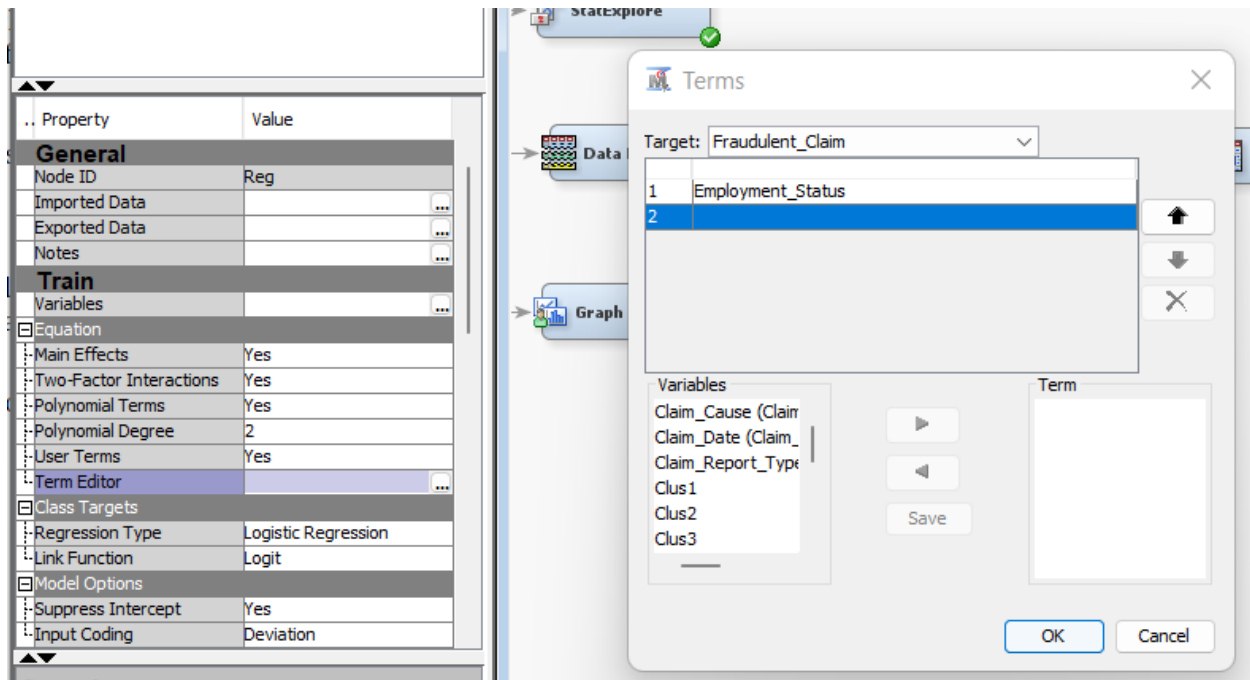
Figure 3: Automobile insurance claim fraud regression output

Likelihood Ratio Test for Global Null Hypothesis: BETA=0							
-2 Log Likelihood		Likelihood					
Intercept Only	Intercept & Covariates	Ratio Chi-Square	DF	Pr > ChiSq			
1377.884	1308.115	69.7686	1	<.0001			
Type 3 Analysis of Effects							
Effect	DF	Wald Chi-Square	Pr > ChiSq				
Gender	1	56.6360	<.0001				
Analysis of Maximum Likelihood Estimates							
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept	1	-2.9517	0.0939	987.33	<.0001		0.052
Gender	F	-0.7070	0.0939	56.64	<.0001		0.493
Odds Ratio Estimates							
Effect		Point Estimate					
Gender	F vs M	0.243					

Two-Factor Interaction and Polynomial Terms

The regression node can execute linear and logistic regression with two or more input variables with the target variable. The claim fraud can be extended to include a two-factor interaction between Gender and Employment Status. I will use the Regression node Train properties Equation group set all Yes and click the Term Editor ellipse to choose “Employment_Status” and OK.

Figure 4: Two-factor interaction



Two-factor interaction results show each level of the categorical variables by target variable. The employed level in Employee_Status and female level in Gender input are statistically significant, although between female and retired, they are not.

Figure 5: Two-factor interaction term sample output

Only	Covariates	Chi-Square	DF	Pr > ChiSq
178	1377.884	1283.662	94.2216	4
179				<.0001
180				
181				
182				
183				
184				
185				
186				
187				
188				
189				
190				
191				
192				
193				
194				
195				
196				
197				
198				
199				
200				
201				
202				
203				

Effect	DF	Wald Chi-Square	Pr > ChiSq
Employment_Status*Gender	4	76.0170	<.0001

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept	1	-3.0343	0.1000	920.55	<.0001		0.048
Employment_Status*Gender Disabled	1	0.2625	0.3381	0.60	0.4375		1.300
Employment_Status*Gender Employed	1	-1.0261	0.1183	75.24	<.0001		0.358
Employment_Status*Gender Medical Leave	1	0.7221	0.3333	4.70	0.0302		2.059
Employment_Status*Gender Retired	1	0.0771	0.3585	0.05	0.8298		1.080

DMINE Regression in SAS Enterprise Miner

A DMINE node helps to identify the input variables that are the most valuable predictors of a target variable. If the categorical input dataset has missing values, it converts a new category; if

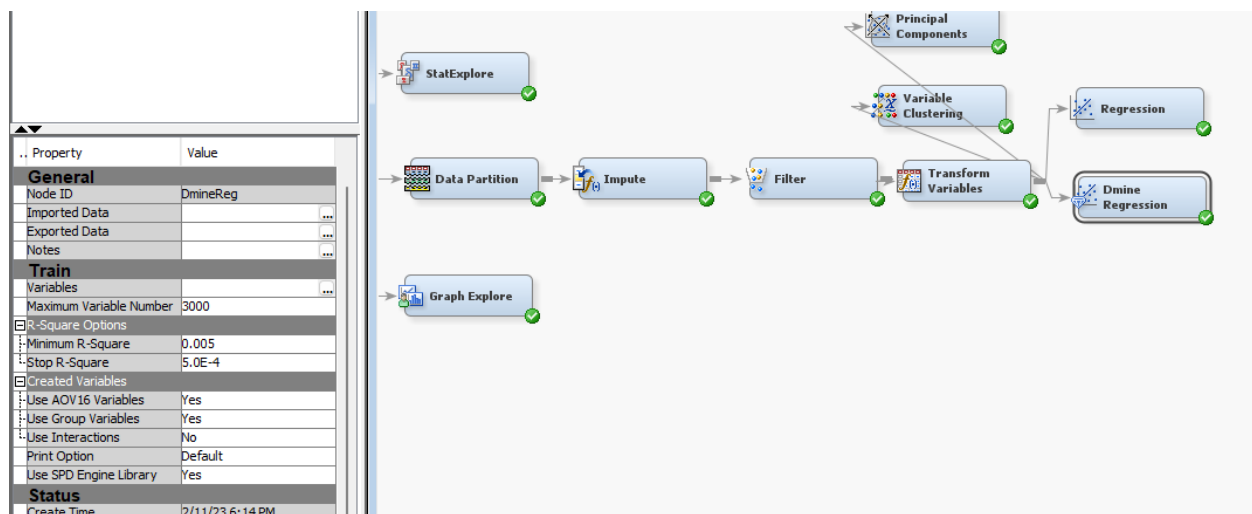
the interval input variable has missing values, it is replaced with the mean value. Any missing values in the target variable can affect the result of the regression model.

I will use the claim fraud data set, drag and drop DMINE Regression in the Model tab on the diagram workplace, and note you can connect Data Partition because the DMINE does not have data filtering and cleaning properties and set properties;

-Maximum Variable Number helps to set the upper limit for the number of input variables for the regression model.

-Minimum R-Square sets the lower limit for the R-squared value of an input variable for the regression model. When setting up all the properties, then Run the DMINE node.

Figure 6: DMINE regression node with properties.



The result of a DMINE regression analyst's Fit statistic window shows the average square at 0.057091 if comparing the stepwise logistic regression average square error of 0.05813 on the Validation partition, which is higher than the DMINE regression average squared error—difference Train and Validation dataset average square error at 0.004631. The DMINE regression

difference average square error is small, which means the model does not appear to be overfitting.

Figure 8: DMINE fit statistic

File Edit View Window

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Fraudule...	Fraudule...	ASE	Average ...	0.05246	0.057091	.
Fraudule...	Fraudule...	DIV	Divisor fo...	5996	4006	.
Fraudule...	Fraudule...	MAX	Maximu...	0.988636	0.996629	.
Fraudule...	Fraudule...	NOBS	Sum of F...	2998	2003	.
Fraudule...	Fraudule...	RASE	Root Ave...	0.229042	0.238936	.
Fraudule...	Fraudule...	SSE	Sum of S...	314.5512	228.7051	.
Fraudule...	Fraudule...	DISF	Frequenc...	2998	2003	.
Fraudule...	Fraudule...	MISC	Misclassi...	0.061374	0.064403	.
Fraudule...	Fraudule...	WRON...	Number ...	184	129	.

The output of the DMINE regression node shows a list of the essential variables for claim fraud: vehicle class, gender, claim cause, claim amount, and months of science last claim. Vehicle class, gender, and claim cause are highly significant, and the claim amount and months of the science of the last claim are moderately significant.

Figure 9: DMINE output result

228
229
230 The DMINE Procedure
231
232 Effects Chosen for Target: Fraudulent_Claim
233
234

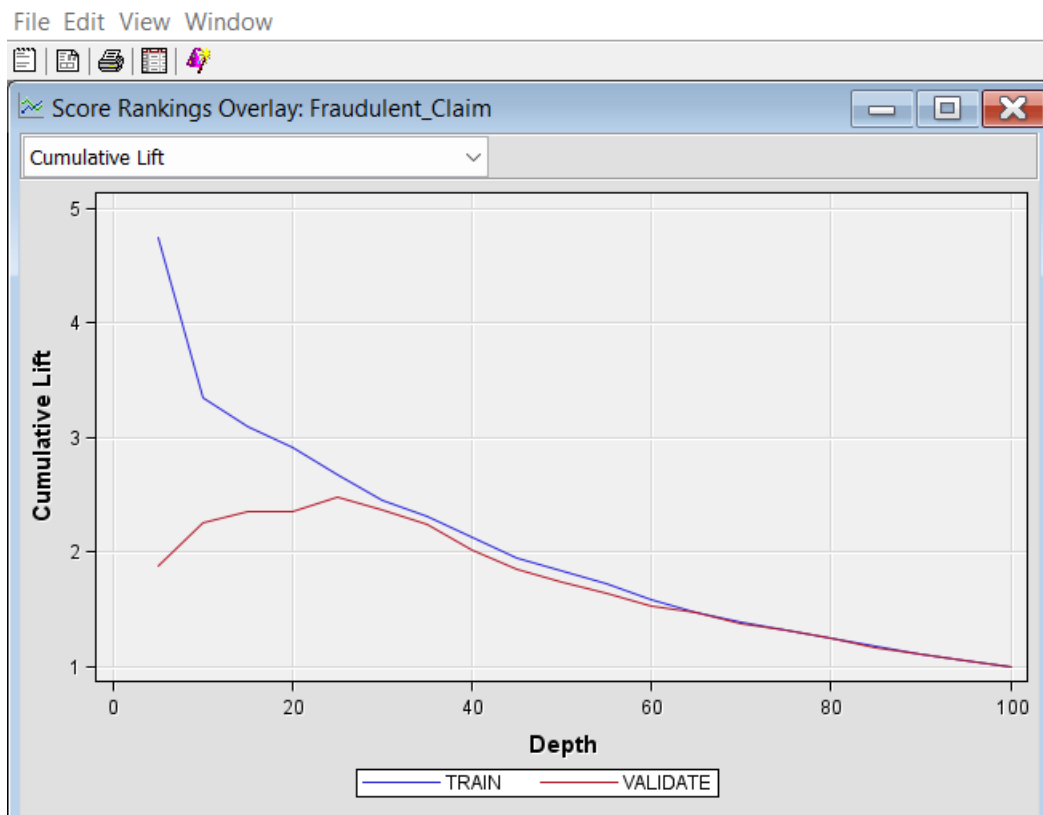
Effect	DF	R-Square	F Value	p-Value	Sum of Squares	Error Mean Square
Group: Vehicle_Class	2	0.022049	33.762833	<.0001	3.788672	0.056107
Class: Gender	1	0.021137	66.141946	<.0001	3.632038	0.054913
Group: Claim_Cause	2	0.013761	21.829506	<.0001	2.364537	0.054159
AOV16: Months_Since_Last_Claim	15	0.006150	1.302786	0.1911	1.056763	0.054077
AOV16: Claim_Amount	13	0.004375	1.069580	0.3811	0.751689	0.054061

242
243
244
245

2:03 PM
2/10/2023

Cumulative lift helps to compare to random guessing or no lift equal to 1. The higher the cumulative lift, the better the result is in the model's predictability. The x-axis depth is the cumulative percentage of data evaluated, and the y-axis shows how far the model is from the random guessing. The accumulative helps us know how to model performance before running the model. The better result is a higher lift. Let's see the result of claim fraud; the depth is 25%, and the cumulative lift is 2.4, indicating the model is two times better than random guessing. The level of the 40% of data left by the cumulative lift 2.0 means that it should continue to improve this model until a better model is developed.

Figure 10: DMINE cumulative lift

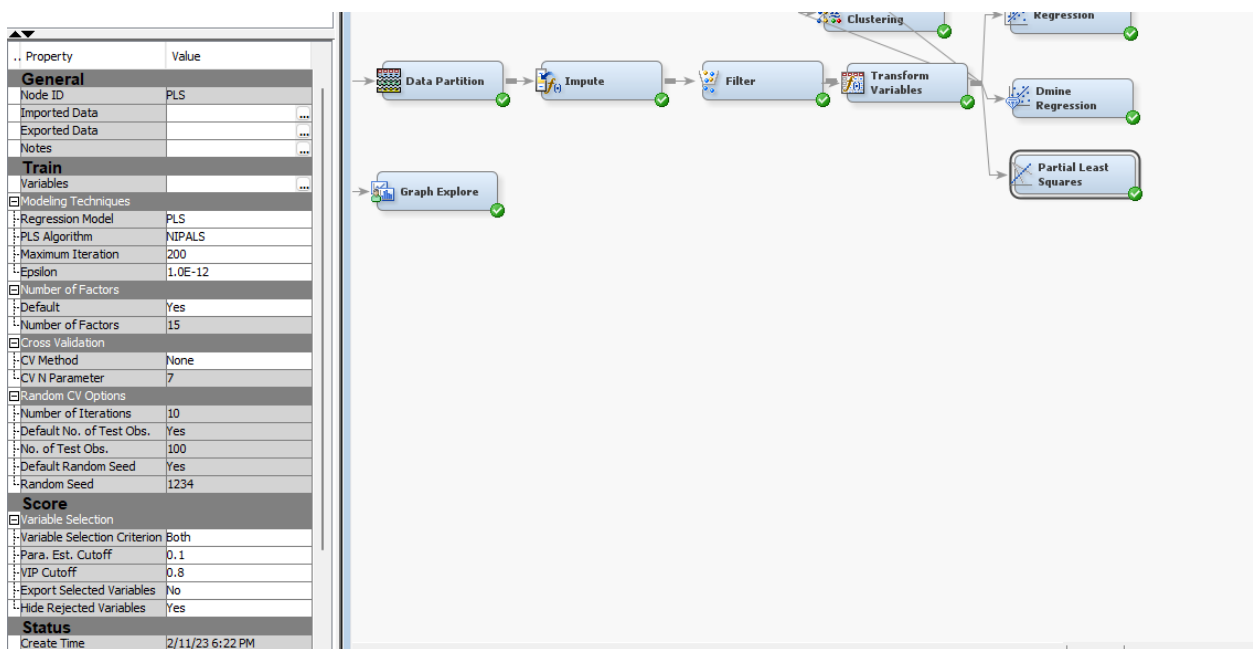


Partial Least Square Regression in SAS Enterprise Miner

The PLS regression node produces an alternative method for multiple linear regression to analyze if the dataset has many input variables and a single continuous or binary target. Also, it gives the best result if the dataset has multicollinearity between input variables. PLS eliminates observations with missing values.

I will apply the PLS regression for the claim fraud dataset. I left the default variable on the original properties.

Figure 11: Partial least squares properties



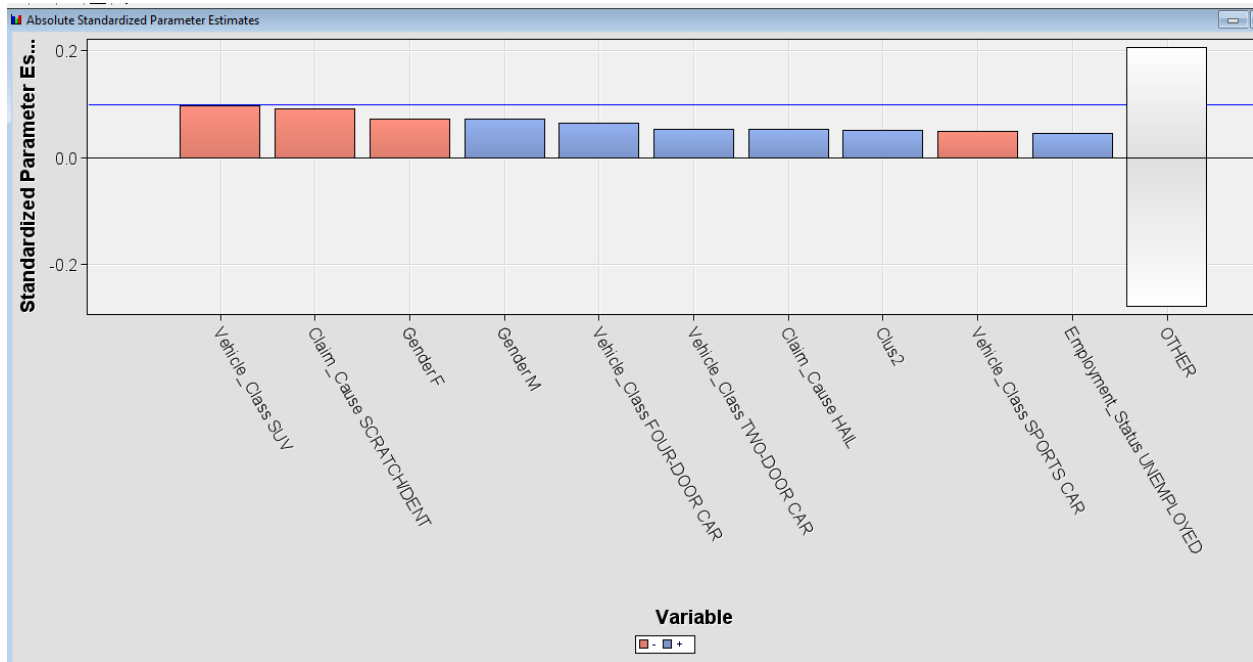
The PLS regression Fit Statistic window results show an average square error of 0.054662, which is more diminutive than the DMINE regression, with the average square at 0.057091. And the difference between the target and validation's average square error is still tiny, so the model does not appear to be overfitting.

Figure 12: Partial least squares fit statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Fraudulent_Claim	Fraudulent_Claim	_ASE_	Average Squared Error	0.053018	0.054662	
Fraudulent_Claim	Fraudulent_Claim	_DIV_	Divisor for ASE	5994	4002	
Fraudulent_Claim	Fraudulent_Claim	_MAX_	Maximum Absolute Error	0.963315	0.988455	
Fraudulent_Claim	Fraudulent_Claim	_NOBS_	Sum of Frequencies	2997	2001	
Fraudulent_Claim	Fraudulent_Claim	_RASE_	Root Average Squared Error	0.230256	0.233798	
Fraudulent_Claim	Fraudulent_Claim	_SSE_	Sum of Squared Errors	317.7875	218.7555	
Fraudulent_Claim	Fraudulent_Claim	_DISF_	Frequency of Classified Cases	2997	2001	
Fraudulent_Claim	Fraudulent_Claim	_MISC_	Misclassification Rate	0.061061	0.061969	
Fraudulent_Claim	Fraudulent_Claim	_WRONG_	Number of Wrong Classificatio...	183	124	

The partial least squares node absolute standardized estimate window shows two color histogram graphs; the red bar specifies a negative value, and the blue bars are positive. The other variable has the highest value, but the color is not meaningful. Vehicle_Class SUV has the highest negative value, and Gender M has the highest positive value.

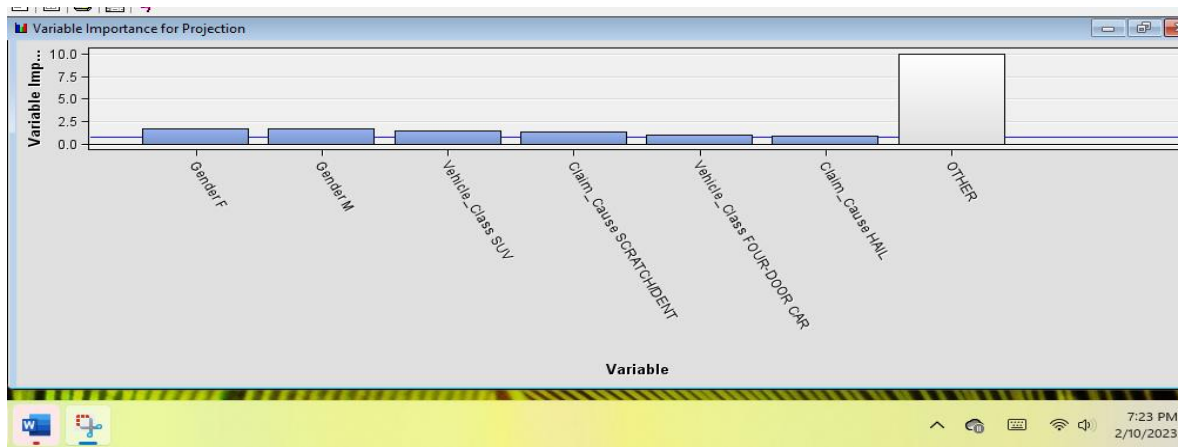
Figure 13: Partial least squares absolute standardized estimate graph



The result of the PLS regression node Variable Importance for the project window helps determine the most crucial variable. The Other column shows the highest point, and then Gender

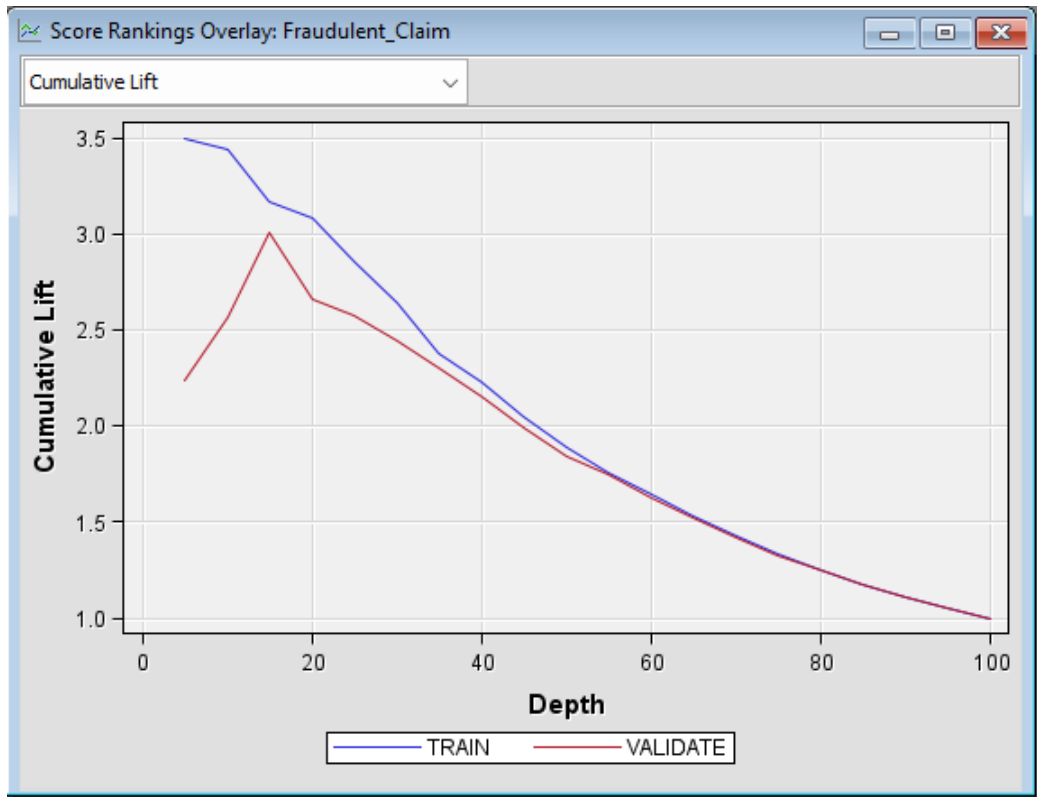
F, Gender M, Vehicle_Class SUV, and other variables are the most critical predictive for the target variable.

Figure 14: Partial least squares variable importance graph



The result of the Score, Rankings Overall window, helps to know what percentage of cumulative data lift, so the depth of 40% is consistent with the cumulative lift above 2.1 that the model needs to be developed. The result DMINE cumulative result is higher than the PLS lift.

Figure 15: Partial least squares cumulative lift graph.

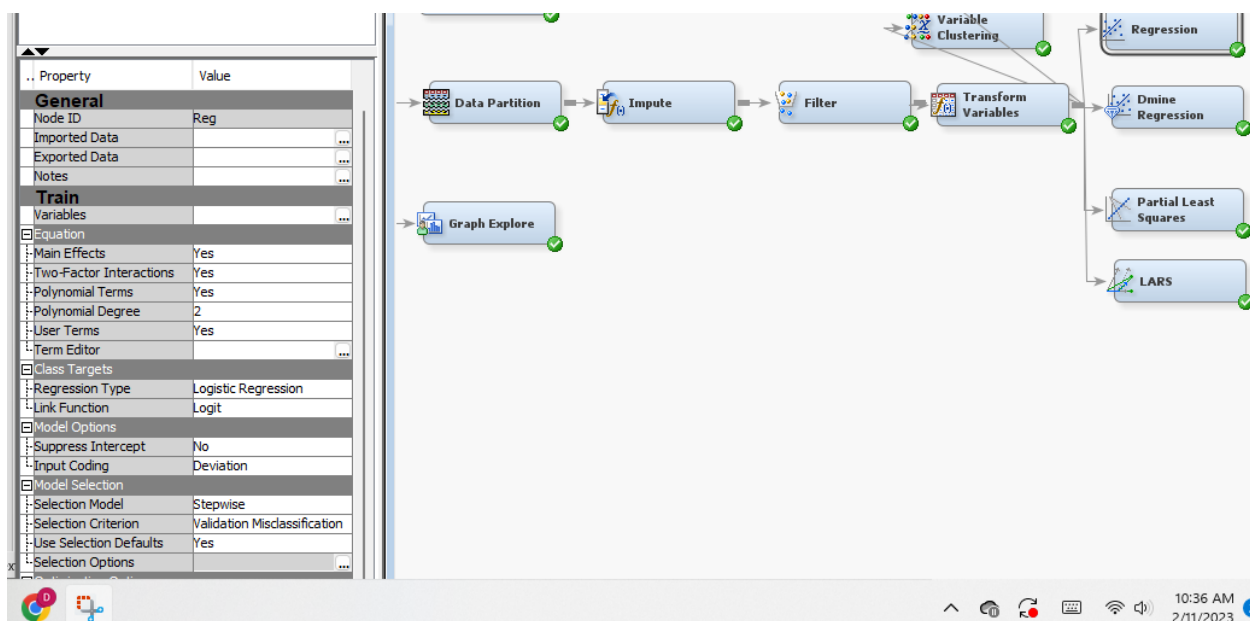


Least Angel Regression in SAS Enterprise Miner

The LARS regression node is similar to the stepwise regression step, which is most helpful if the dataset has many input variables. It does not add a single variable but a single model parameter at a time. Also, the LARS regression node has the least absolute shrinkage, and selection operator (LASSO) regression adds and deletes parameters by the sum of the final regression coefficients for a more accurate model.

I will apply the LARS regression node to the claim fraud data set. Drag and drop the LARS node in the Model tab on the diagram workplace and connect the Transform Variables node.

Figure 16: LARS node and properties in SAS Enterprise Miner.



The result of Fit Statistic windows shows an average square error of 0.055475, slightly higher than the PLS regression average square error of 0.054662.

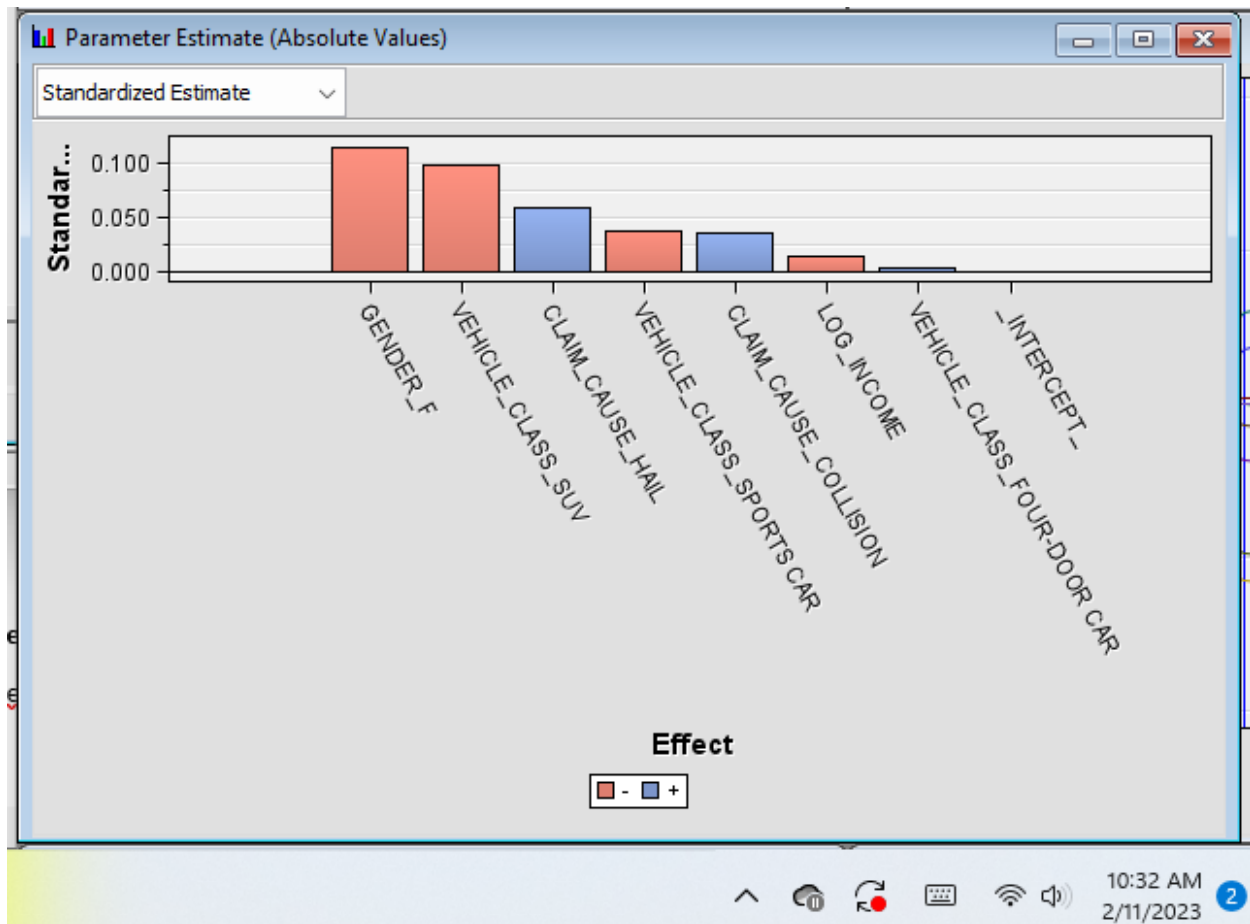
Thus, the partial least square model provided a slightly better result than other models for the claim fraud case, and it is the model I would prefer to use in the following analysis.

Figure 17: Least angle regression Fit statistics window.

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Fraudulent_Claim	Fraudulent_Claim	_ASE_	Average Squared Er...	0.053777	0.055475	
Fraudulent_Claim	Fraudulent_Claim	_DIV_	Divisor for ASE	5994	4002	
Fraudulent_Claim	Fraudulent_Claim	_MAX_	Maximum Absolute ...	0.981017	0.978537	
Fraudulent_Claim	Fraudulent_Claim	_NOBS_	Sum of Frequencies	2997	2001	
Fraudulent_Claim	Fraudulent_Claim	_RASE_	Root Average Squar...	0.231898	0.235532	
Fraudulent_Claim	Fraudulent_Claim	_SSE_	Sum of Squared Err...	322.3367	222.0125	
Fraudulent_Claim	Fraudulent_Claim	_DISF_	Frequency of Classi...	2997	2001	
Fraudulent_Claim	Fraudulent_Claim	_MISC_	Misclassification R...	0.061061	0.061969	
Fraudulent_Claim	Fraudulent_Claim	_WRONG_	Number of Wrong C...	183	124	

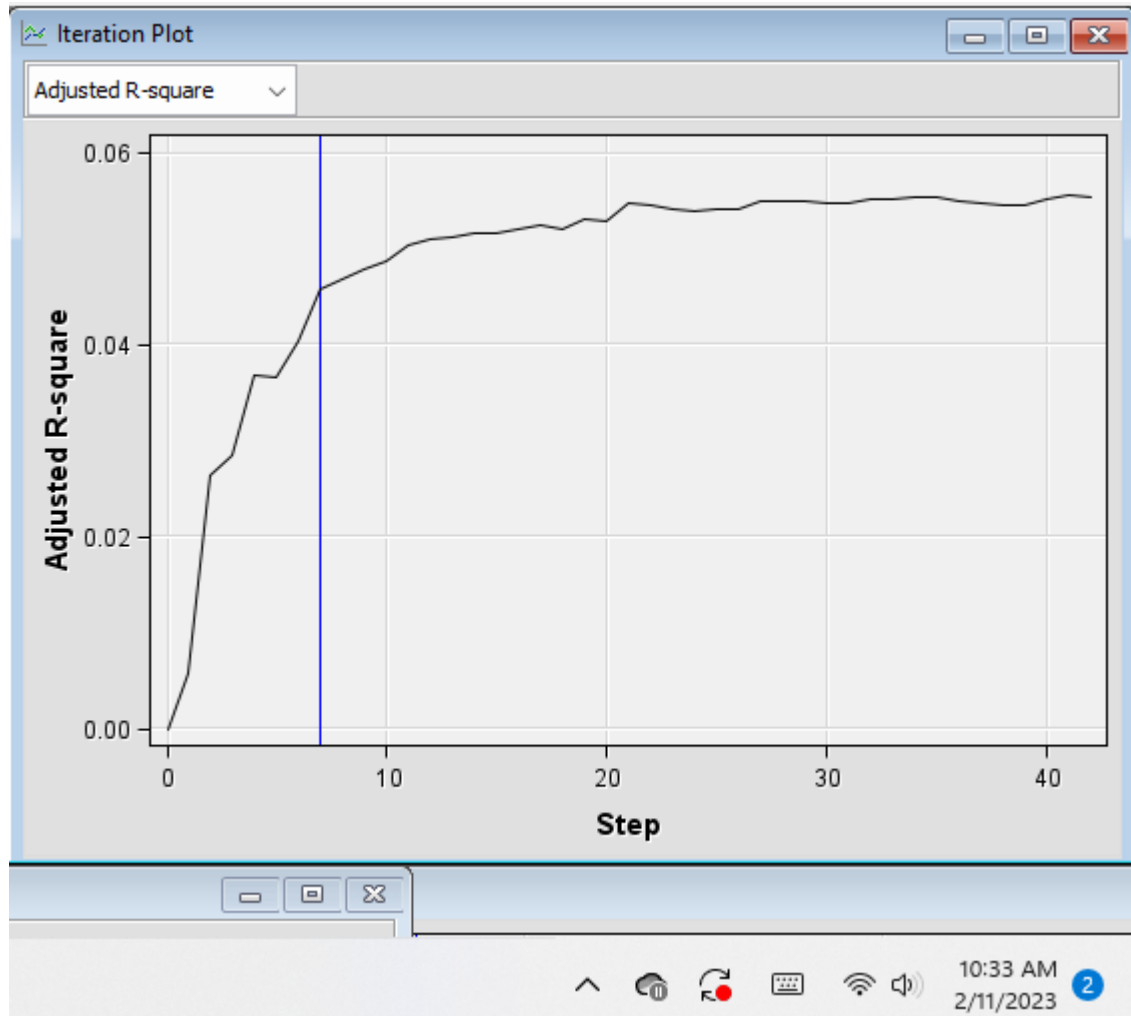
The parameter estimate graph shows four variables with a negative value as the red bar and three bars as positive as the color blue.

Figure 18: Least angle regression Parameter Estimate window.



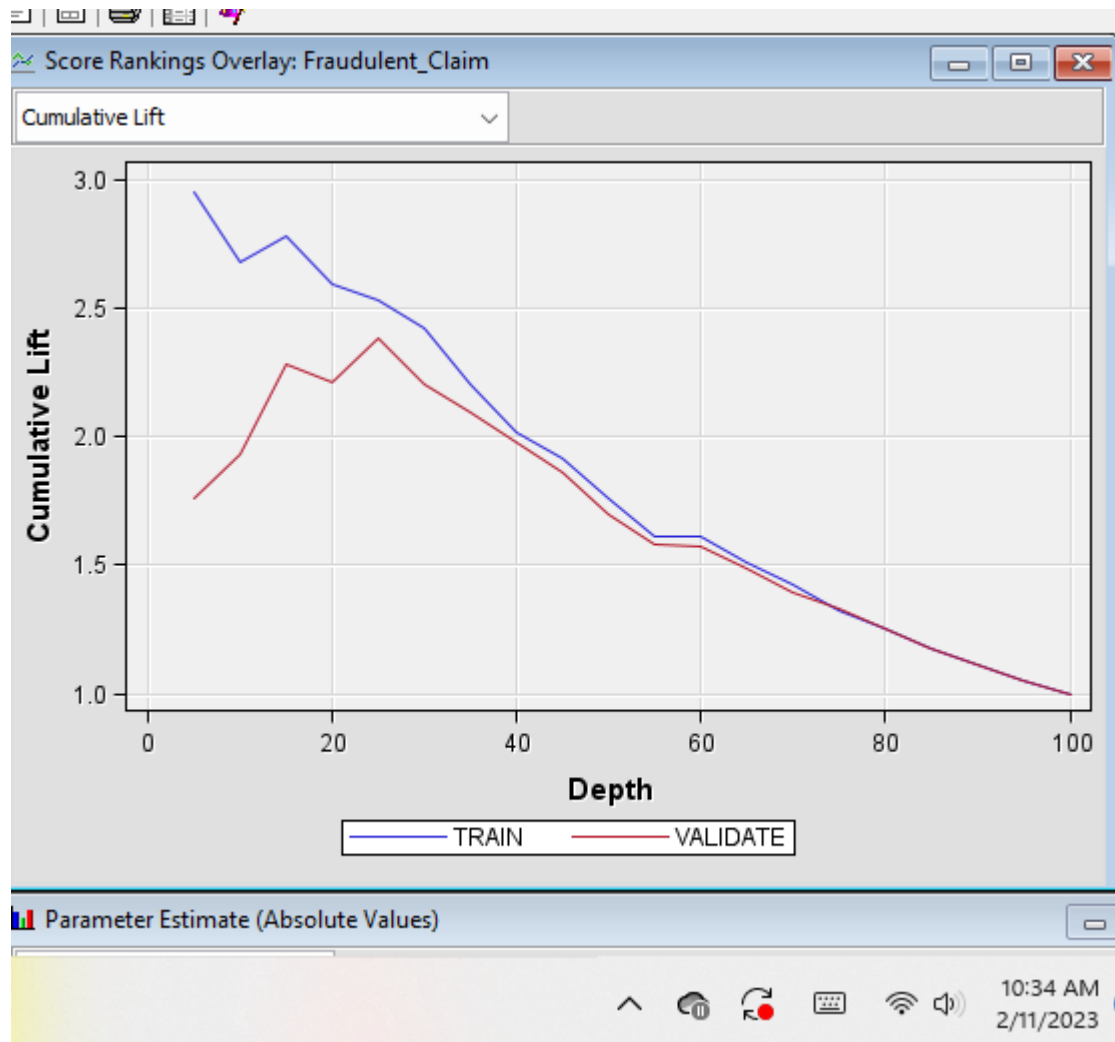
The Iteration Plot window was adjusted to the R-square value to the iterations used to develop the final regression model. The result of the claim fraud dataset by LARS regression adjusted R-square of 0.045 at step 7 is low, which means the model predicts weakly.

Figure 19: Least angle regression adjusted R-squared graph.



The last graph shows that the cumulative fit of the LARS node has a 1.9 lift over the first 40% depth of the claim fraud dataset, which is also weaker than the cumulative lift of the PLS model.

Figure 20: Least angle regression cumulative lift graph.



Conclusion

"Partial least squares (PLS) is a flexible regression technique with principal component analysis features and extends multiple linear regression. PLS is best used when the data set contains fewer observations than input variables and high collinearity exists. PLS can also be used as an exploratory analysis tool to identify input variables to include in the model and outliers. Like multiple linear regression, PLS' overarching goal is to create a linear predictive model." (McCharty,2022)

Thus, I worked on categorical target variable Y/N to know which input variables are predicted: Gender F/M, Vehicle_Class SUV, Claim_Cause SCRATCH/DENT, Vehicle_Class FOUR-DOOR CAR, Claim_CAUSE HAIL. The more accurate model for this case is Partial least square (PLS) with a lower average square error of 0.054662.

Reference

Richard V. McCarthy, Mary M. McCarthy; Wendy Ceccucci, 2022. *Applying Predictive Analytics Finding Value in Data*. Second edition.