

Applied Analytic Modeling

Lab 2

Predictive Modeling Using Regression-SAS Miner

Submitted to
Prof. David Parent

Submitted by

<p>Grace Adaji 301373339</p>
--

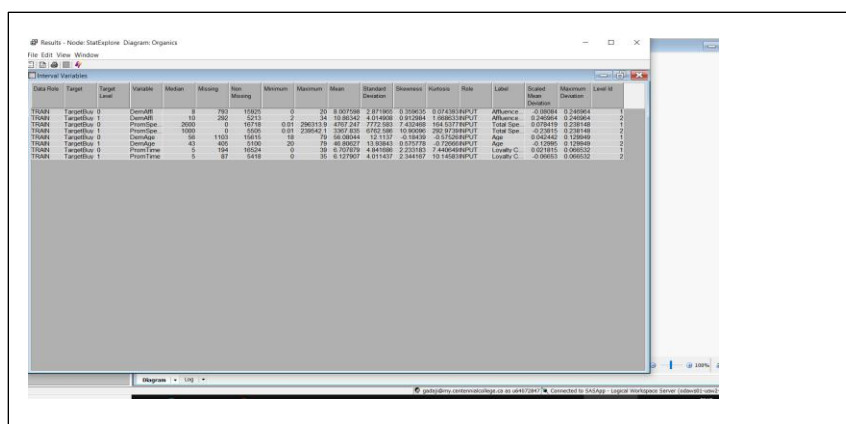
REGRESSION EXERCISE

1. Predictive Modeling Using Regression

- a. Return to the Chapter 3 Organics diagram in the **My Project**. Use the StatExplore tool on the **ORGANICS** data source.

1) First **StatExplore** node is connected to the **ORGANICS** node.

2) StatExplorer node results is generated

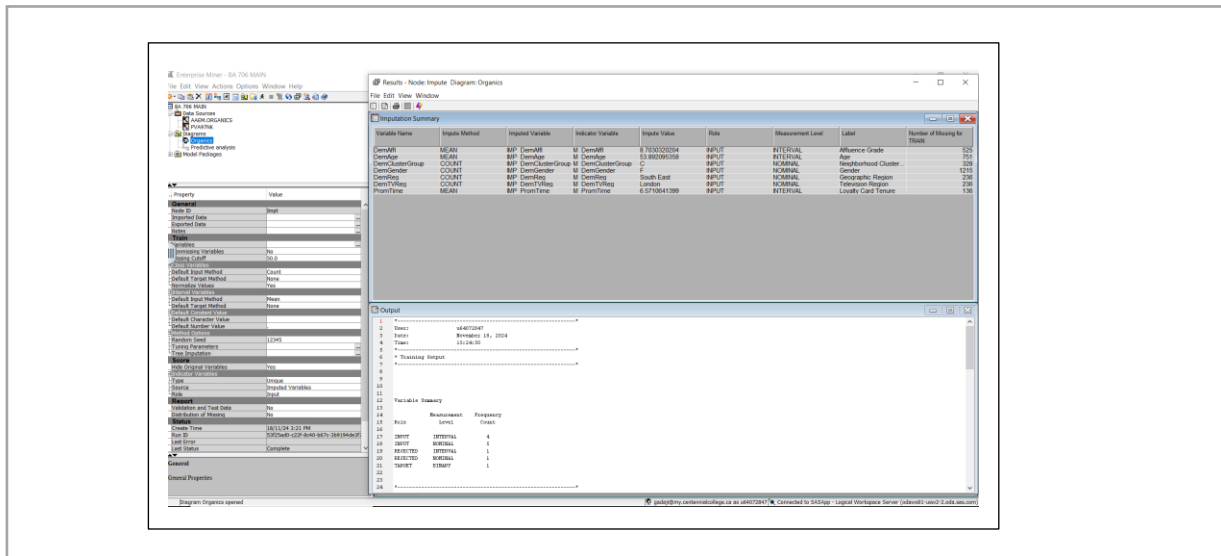


Data Role	Target	Target Label	Variable	Median	Missing	Non-Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Role	Label	Scaled Mean	Scaled Deviation	Level ID
Train	Target	Demand	Demand	10	10	10	0	10	8.027108	3.071001	0.101610	0.1741014	NT	Affluence	0.100000	0.100000	1
Train	Target	Price	Price	280	280	280	0	34	10.00000	4.014000	0.012000	1.000000	NT	Affluence	0.100000	0.100000	2
Train	Target	PromTime	PromTime	2000	0	1070	0.01	2000	1.070000	1.070000	0.010000	0.010000	NT	Total Spent	0.010000	0.010000	1
Train	Target	Price	Price	1000	0	1000	0.01	2000	1.000000	1.000000	0.010000	0.010000	NT	Total Spent	0.010000	0.010000	2
Train	Target	Price	Price	1000	0	1000	0.01	2000	1.000000	1.000000	0.010000	0.010000	NT	Age	0.010000	0.010000	1
Train	Target	Price	Price	1000	0	1000	0.01	2000	1.000000	1.000000	0.010000	0.010000	NT	Age	0.010000	0.010000	2
Train	Target	Price	Price	1000	0	1000	0.01	2000	1.000000	1.000000	0.010000	0.010000	NT	Age	0.010000	0.010000	1
Train	Target	Price	Price	1000	0	1000	0.01	2000	1.000000	1.000000	0.010000	0.010000	NT	Age	0.010000	0.010000	2
Train	Target	Price	Price	1000	0	1000	0.01	2000	1.000000	1.000000	0.010000	0.010000	NT	Age	0.010000	0.010000	1
Train	Target	Price	Price	1000	0	1000	0.01	2000	1.000000	1.000000	0.010000	0.010000	NT	Age	0.010000	0.010000	2

- b. In-order to prepare for regression, missing values are imputed? Why do you think we should impute?

To prepare for regression , missing values were imputed for the following reasons;

- 1.Imputation prevents loss of valuable training data, enhancing model reliability.
2. By imputing, high cost of collecting new data is avoided by filling in gaps effectively.
3. Missing data may hold insights, and imputing helps retain meaningful patterns in the dataset.
4. Regression models require complete data to function effectively since calculation cannot be performed on missing data
5. Imputation fixes biases from missing data, making the model more accurate.



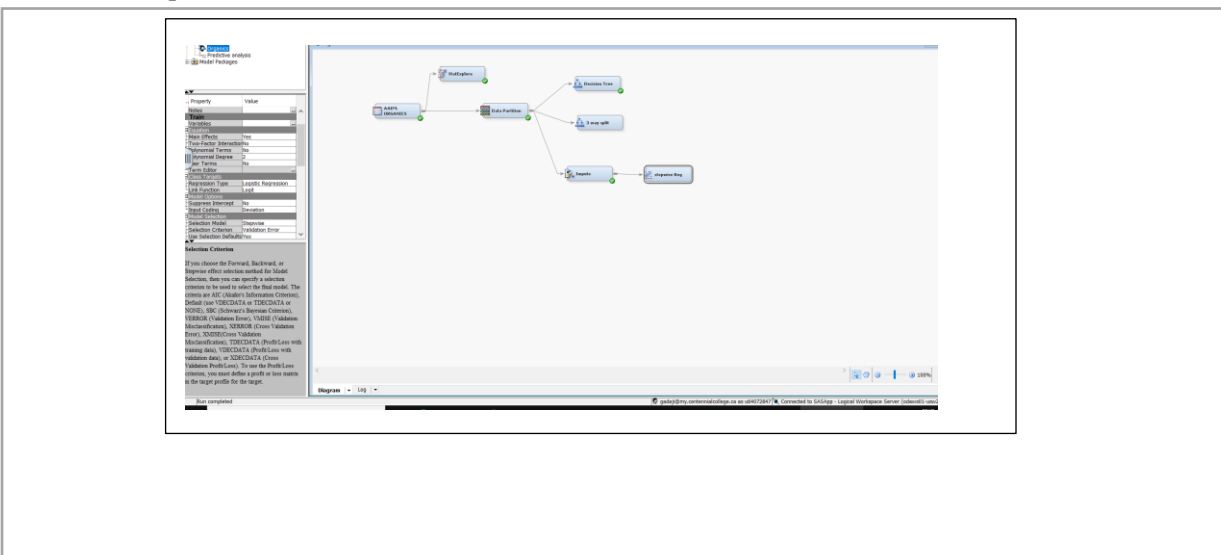
c. Add an **Impute** node from the **Modify** tab into the diagram and connect it to the **Data Partition** node. Create imputation indicators for all imputed inputs.

d. What changed after imputing?

Type your answer here: After imputing, the missing values were replaced using the mean of the existing values for each variable. Also, other missing values were replaced using the count for nominal variables. This ensured that all gaps were addressed, resulting in a complete dataset. The dataset now contains no missing entries, which is crucial for accurate analysis and modeling. Imputation has also helped maintain the dataset's size, avoiding the need to remove rows or columns with missing values, which could have led to a loss of valuable information.

e. Add a **Regression** node to the diagram and connect it to the **Impute** node.

f. Choose stepwise as the selection model and the validation error as the selection criterion.



The screenshot displays the Orange3 data mining software interface. The main workspace shows a workflow with three widgets: 'File', 'Node-Regression', and 'Scatter Plot'. The 'Node-Regression' widget is active, showing a table of regression coefficients for various features. The 'Scatter Plot' widget displays a scatter plot of 'Depth' vs 'Covered Area' with a fitted regression line. The bottom panel shows the 'Fit Statistics' table, which includes metrics like R-squared, Adjusted R-squared, and F-statistic.

Node-Regression Output:

Feature	Estimate	Standard Error	t-Statistic	p-Value	Standardized Estimate	Adjusted R-squared
(Intercept)	-4.4222	0.4226	-10.48	<.0001	0.439	
Depth	1.1284	0.0042	269.35	<.0001	0.805	0.131186
Covered Area	-0.0048	0.0012	-47.17	<.0001	-0.385	0.432380
Depth * Covered Area	0.0004	0.0002	206.43	<.0001	0.001	0.1110
Depth * Depth	0.0002	0.0004	1.04	0.3017	1.005	
Covered Area * Covered Area	-0.0004	0.0006	-16.05	<.0001	-0.004	0.4462
Depth * Depth * Covered Area	-0.0001	0.0007	-0.16	0.881	0.002	
Covered Area * Depth * Depth	0.0001	0.0007	0.16	0.881	0.002	

Fit Statistics:

Target Label	Fit Statistics	Statistics Label	Value	Validation
TargetLabel	Organics Purchases	AIC	6601.257	
TargetLabel	Organics Purchases	Akaike's Information	0.131186	0.131186
TargetLabel	Organics Purchases	Average Squared Error	0.432380	0.432380
TargetLabel	Organics Purchases	Average Error Fx	0.432380	0.432380
TargetLabel	Organics Purchases	Dependent Variable	11110	
TargetLabel	Organics Purchases	Model Degrees of Freedom	0	
TargetLabel	Organics Purchases	Total Degrees of Freedom	11110	
TargetLabel	Organics Purchases	Adjusted R-squared	0.131186	
TargetLabel	Organics Purchases	Error Function	8675.257	8685.81
TargetLabel	Organics Purchases	Final Prediction Error	0.131186	0.131186
TargetLabel	Organics Purchases	Maximum Absolute Error	0.091117	0.091117
TargetLabel	Organics Purchases	Mean Squared Error	0.131186	0.131186
TargetLabel	Organics Purchases	Number of Errors	11110	
TargetLabel	Organics Purchases	Sum of Squared Errors	11110	11111
TargetLabel	Organics Purchases	Root Mean Square Error	0.372222	0.370460
TargetLabel	Organics Purchases	Root Mean Square Error	0.372222	0.370460
TargetLabel	Organics Purchases	Root Mean Square Error	0.372222	0.370460
TargetLabel	Organics Purchases	Sum of Squared Errors	8675.257	8685.81
TargetLabel	Organics Purchases	Sum of Squared Errors	22224	22222
TargetLabel	Organics Purchases	Sum of Squared Errors	0.131186	0.131186

Type your answers here: The variables included in the final model are IMP_DemAffl, IMP_DemGender, IMP_DemAge, M_DemAge, M_DemAffl and M_DemGender.

However, **IMP_DemAffl** and **IMP_DemAge** are the top two predictors (most important) based on their extremely high Chi-square values, signifying strong contributions to the model.

- i) Go to line 632 in the Output window.
- j) The odds ratios indicate the effect that each input has on the logit score. Find the odds ratios in the output and provide a screenshot:

The screenshot displays the SPSS Output Viewer window, showing the results of a multiple regression analysis. The output is organized into several sections:

- Model Summary:** Shows the R Square (0.444), Adjusted R Square (0.401), and Standard Error of the Estimate (1.000).
- ANOVA:** A table showing the Sum of Squares, Degrees of Freedom, Mean Square, F-value, and Sig. for the Regression, Residual, and Total. The Regression is significant (Sig. = 0.000).
- Coefficients:** A table showing the Unstandardized Coefficients, Standardized Coefficients (Beta), t-statistics, and Sig. for each predictor. The predictors are (Constant), X1, X2, X3, and X4. X1, X2, and X3 are significant (Sig. = 0.000), while X4 is not (Sig. = 0.100).
- R Squared Statistics:** A table showing the R Square, Adjusted R Square, and Standard Error of the Estimate for the Regression, Residual, and Total.

The output is displayed in a text-based format within the SPSS Output Viewer window.

k) Interpret the odds ratio estimate:

Type your answer here:

1. **MP_DemAffl (1.283)**: Each one-unit increase in Demographic affluence increases the probability of buying organic products by **28.3%**.
2. **IMP_DemAge (0.947)**: Each one-unit increase in demographic age decrease the probability of buying organic products by **5.3%**.
3. **IMP_DemGender (6.967(Females (F) vs Unknown (U))**: The probability of female with known gender buying organic products are 596.7% higher than for those with unknown gender.
4. **IMP_DemGender Male (2.899 (M) vs Unknown (U))**: The probability of males with known gender buying organic products are 189.9 % higher compared to those with unknown gender.
5. **M_DemAffl (0 vs. 1) point estimate (0.708)**: The probability of people without missing demographic affluence data buying organic products are **29.2% lower** than for people with missing data.
6. **M_DemAge (0 vs. 1), Point Estimate (0.796)**:
The probability of buying organic products is **20.4% lower** for those without missing (complete data) demographic age data than those with missing age data.
7. **M_DemGender (0 vs. 1), Point Estimate (4.769)**:
The probability of people with no missing demographic gender data buying organic products are **376.9% higher** than people with missing demographic gender data.

PART 2

a. In preparation for regression, are any transformations of the data warranted? Why or why not?

Type your answer here: Yes, transformation is warranted if the data is skewed because transformation normalize skewness, which affects model accuracy. However, transformation may not be necessary if the data is normalized and there is no skewness. Additionally, **log transformations** can make the interpretation of results more challenging for clients, as the relationship between variables is transformed into a logarithmic scale, which might not be intuitive.

i. Open the results of the Stat Explore node. Provide a screenshot that includes the variable names and skewness statistics:

resources - node statistics (v) diagram urgencies

File Edit View Windows

Interval Variables

Data Role	Target	Target Level	Variable	Median	Missing	Non-Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Rule	Label	Skewed	Maximum	Level ID
TRAIN	TargetBuy	0	Promoted	2500	0	8108	0.01	110072.4	4768.008	7295.565	1.3462	36.376889PLT	Total Skewed	0.00204	0.146786	1	
TRAIN	TargetBuy	1	Promoted	1000	0	2753	0.01	55000	3305.542	5848.153	3.512415	18.532778PLT	Total Skewed	-0.24879	0.248786	2	
TRAIN	TargetBuy	0	DemAffl	0	0	8108	0	19	8.015167	5.822362	0.129417	0.11105789PLT	Impacted Aff	-0.01702	0.250176	1	
TRAIN	TargetBuy	1	DemAffl	10	0	2753	2	31	10.15414	5.910438	0.860228	1.70380389PLT	Impacted Aff	0.238678	0.238678	2	
TRAIN	TargetBuy	0	DemAge	25	0	8108	18	79	56.05488	17.73538	-0.17028	-0.458778PLT	Impacted Age	0.045488	0.122825	1	
TRAIN	TargetBuy	1	DemAge	44	0	2753	20	79	47.38089	13.40528	0.488884	-0.038889PLT	Impacted Age	-0.12283	0.122825	2	
TRAIN	TargetBuy	0	PromTime	5	0	8108	0	39	6.780027	4.788881	2.284431	7.888888PLT	Impacted L	0.018889	0.088812	1	
TRAIN	TargetBuy	1	PromTime	5	0	2753	0	35	6.178035	4.192229	2.437225	10.438018PLT	Impacted L	-0.05861	0.058612	2	

computed

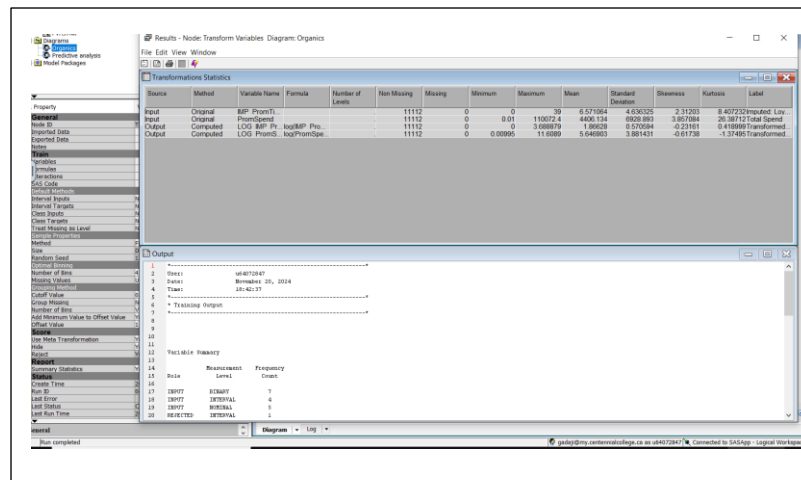
Type here to search

pathfinder.com/college-us-to-40072450 - Connected to Firefox - Legal Workplace Gender (selected view) 2 mb, secure

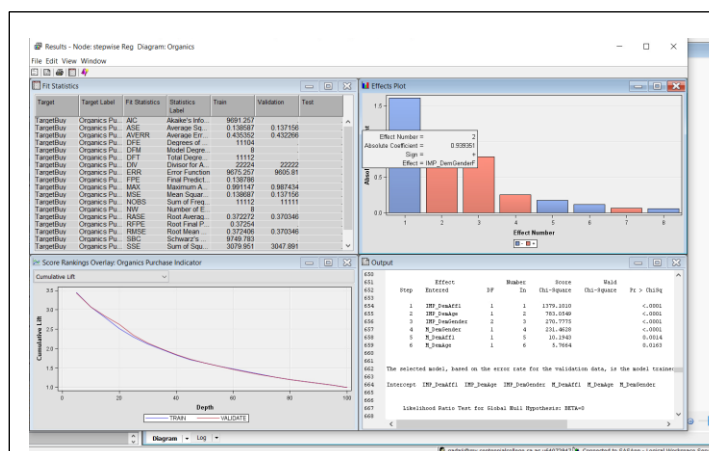
Enterprise Enterprise Enterprise Enterprise Results No. 0

- Type your answer here :PromSpend and IMP_PromTime

- Type your answer here: yes, the log transformation reduced the skewness of PromSpend from 3.8570 to -0.61738, and IMP_PromTime from 2.31203 to -0.23161



- Type your answer here: No. The selected variable : Dem Age, Dem Gender and DemAff did not change. Also, the validation ASE 0.137156 did not change. Meaning they were not important.



g. Go to line 664 of the Output window.

K. In your words, describe what you did in this assignment and why you had to do each of these steps? Plus, how would you describe the independent variables that have an impact on the dependent variable to a client?

- **Data Exploration**

After obtaining the dataset, I conducted an exploratory analysis to understand its structure. This step involved identifying missing values and checking for skewness. Both missing data and skewed distributions can negatively impact regression models, so it's essential to address them. The Staff Explore node was used to assess skewness, and the StatExplore node helped analyze patterns, distributions, and relationships in the data.

- **Data Preparation (Imputation)**

To create a complete dataset, I imputed the missing values. Imputation is necessary because regression models require a full dataset to function properly. For numerical variables, missing values were replaced with the mean, and for categorical variables, they were replaced with the mode. This step reduces bias and prevents the loss of valuable information.

- **Regression Model Selection**

I applied a stepwise selection method to the regression model to identify the most important independent variables. This technique helps simplify the model by retaining only the variables that significantly impact the dependent variable.

- **Result Interpretation**

Odds ratios were calculated to quantify how each independent variable affects the likelihood of the dependent outcome (e.g. Buying organic product probability). For example, demographic affluence and gender significantly influence the likelihood of buying organic products.

- **Skewness and Transformation**

Variables like PromSpend and IMP_PromTime had high skewness, so I log-transformed them to normalize their distributions. Skewed data can distort model results, so reducing skewness improves the model's accuracy.

- **Validation and Model Assessment**

I checked the Average Squared Error (ASE) for validation and confirmed that the transformations did not significantly alter the selected variables or affect model accuracy. The results confirm the robustness of the initial feature selection

Explaining Independent Variables to a Client

Demographic Affluence: A unit increase in affluence increases the probability of donation by 28.3%. More affluent individuals are more likely to buy organic products., hence marketing efforts needs to be focused on high-income neighborhoods and upscale communities to increase the purchase of organic products. Larger households and individuals with lower affluence or income are less likely to buy organic products. Consider offering promotional pricing or emphasizing value in marketing to these segments.

Demographic Age: A higher age slightly reduces the probability of buying of buying organic products, by 5.3%, suggesting that younger individuals may be more inclined to buy organic products. Utilizing social media and digital marketing to reach younger demographics will help drive sales of organic product.

Gender: Females have significantly higher probability of buying organic products compared to unknown genders. Developing marketing campaigns and Offering product lines that cater to women's specific needs and preferences will drive sales of organic product. . Men also show a moderate increase in likelihood, so they should not be ignored entirely.

Recommendation for client

Who to market to: Affluent, younger, and female consumers are the most likely to purchase organic products, so marketing efforts should prioritize these groups.

How to prioritize efforts: Campaigns that emphasize quality, health benefits, and environmental impact are likely to appeal to younger and affluent customers.

Decision-making: The client should allocate resources toward strategies that engage these demographics, such as personalized offers or targeted advertising.