Applied Analytic Modeling

Lab 2
Predictive Modeling Using Regression-SAS Miner

Submitted to

Prof. David Parent

Submitted by

Grace Adaji

301373339

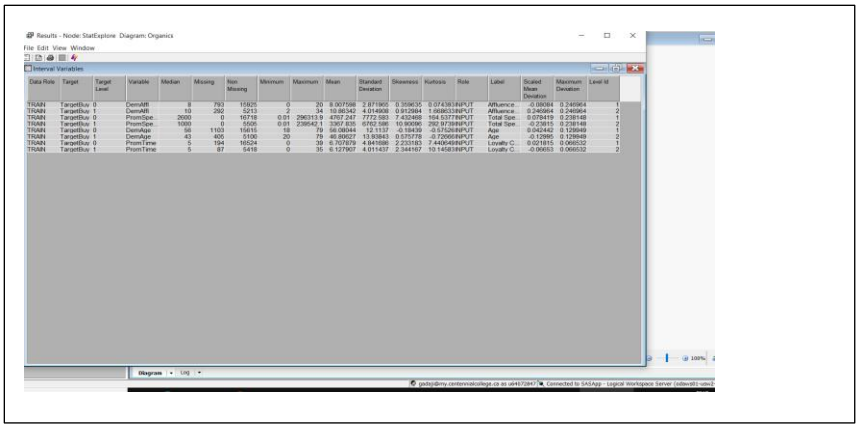
REGRESSION EXERCISE

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



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;

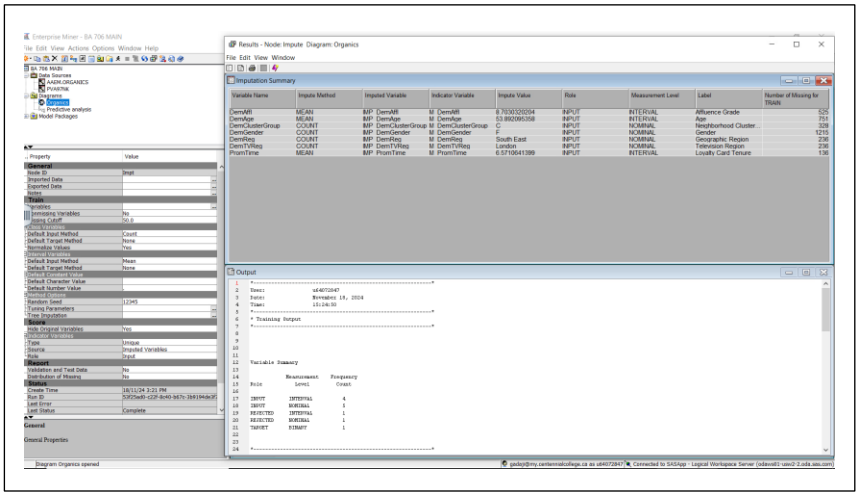
1.Imputation prevents loss of valuable training data, enhancing model reliability.

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.



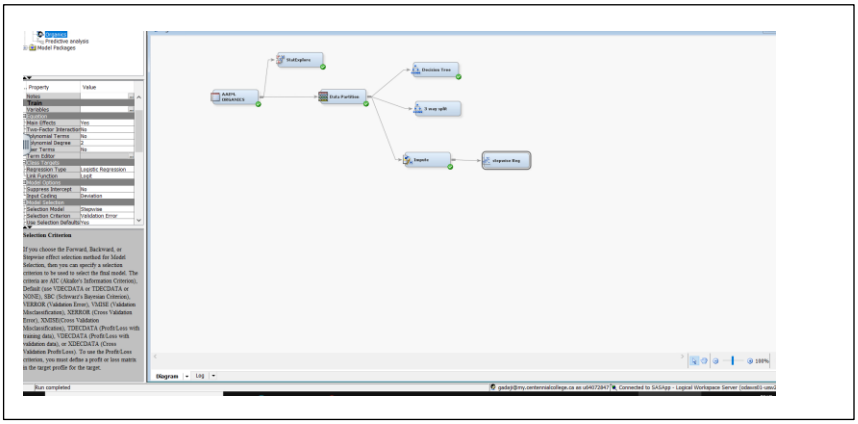
c. Add an **Impute** node from the **Modify** tab into the diagram and connect it to the **Data Partition** node. Create imputation indicators for all imputed inputs.

d. What changed after imputing?

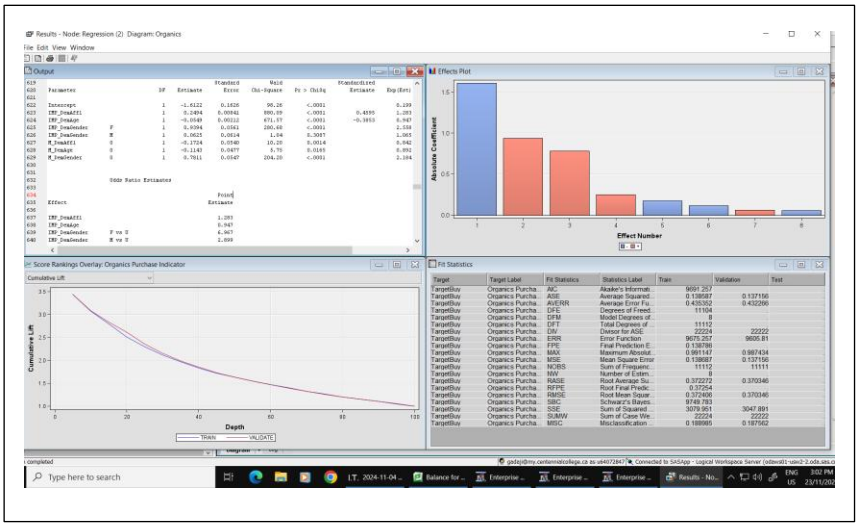
Type your answer here: After imputing, the missing values were replaced using the mean of the existing values for each variable. Also, other missing values were replaced using the count for nominal variables. This ensured that all gaps were addressed, resulting in a complete dataset. The dataset now 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.



g. Run the Regression node and view the results. Maximize the Results Window.



h. Which variables are included in the final model? Which variables are important in this model? What is the validation ASE?

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.

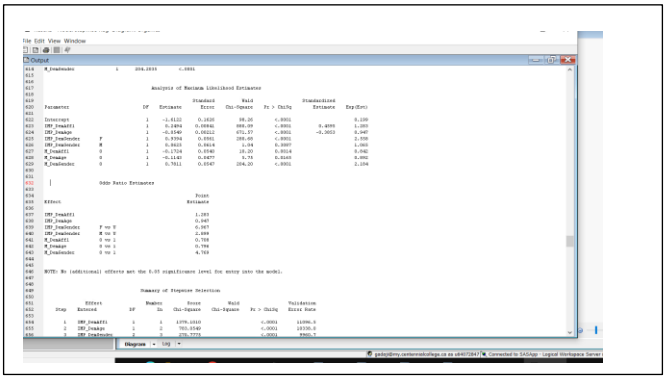
The most important variables are IMP_DemAffl (Chi-square = 880.09), IMP_DemAge (Chi-square = 671.57), IMP_DemGender (Chi-square = 280.68) and M_DemGender (Chi-square = 204.20).

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.

The validation ASE (Average Squared Error) for the model is **0.137156**,

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:



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.

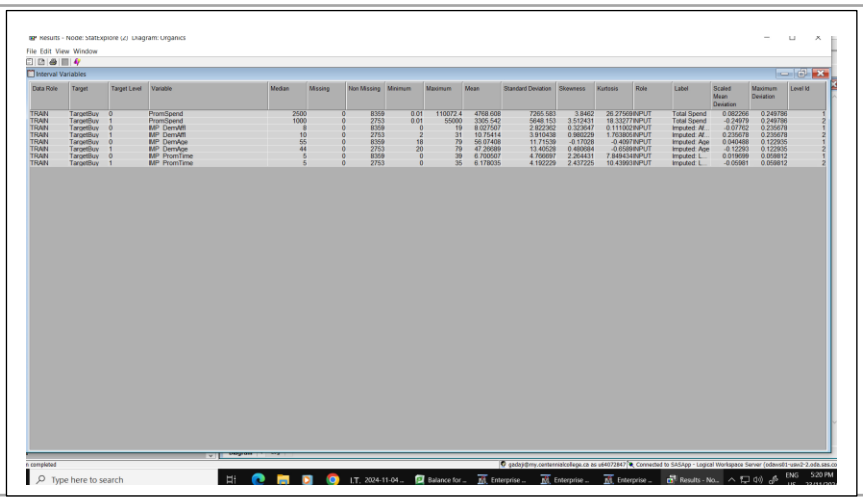
7. **M_DemGender (0 vs. 1), Point Estimate (4.769)**:
The probability of people with no missing demographic gender data buying organic products are **376.9% higher** than people with missing demographic gender data.

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, 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:



b. Which variables are good candidates for transformation using a skewness cutoff of 1.

Type your answer here :PromSpend and IMP_PromTime

c. Add a **Transform Variables** node between the **Impute** node and your **Regression**.

d. Apply a log transformation to the variables you identified in part c.

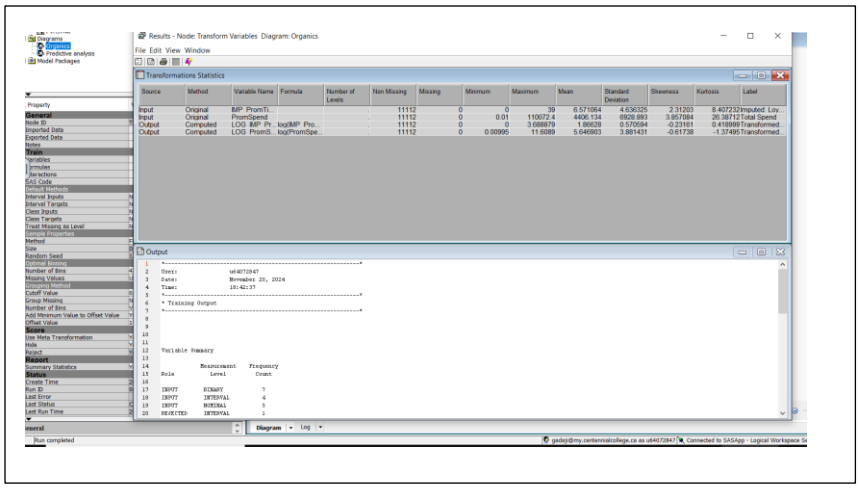
   i. Open the Variables window of the Transform Variables node.

   ii. Select **Method** ⇨ **Log** for the relevant inputs. Select **OK** to close the Variables window.

e. Run the **Transform Variables** node. Explore the exported training data. Did the transformations result in less skewed distributions?

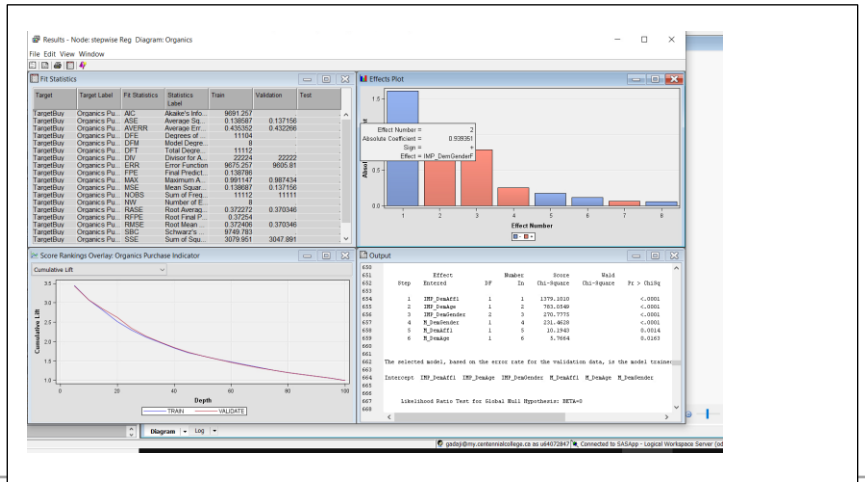
   i. Did the transformations reduce the skewness?

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



f. Rerun the **Regression** node. Do the selected variables change? How about the validation ASE?

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**.**