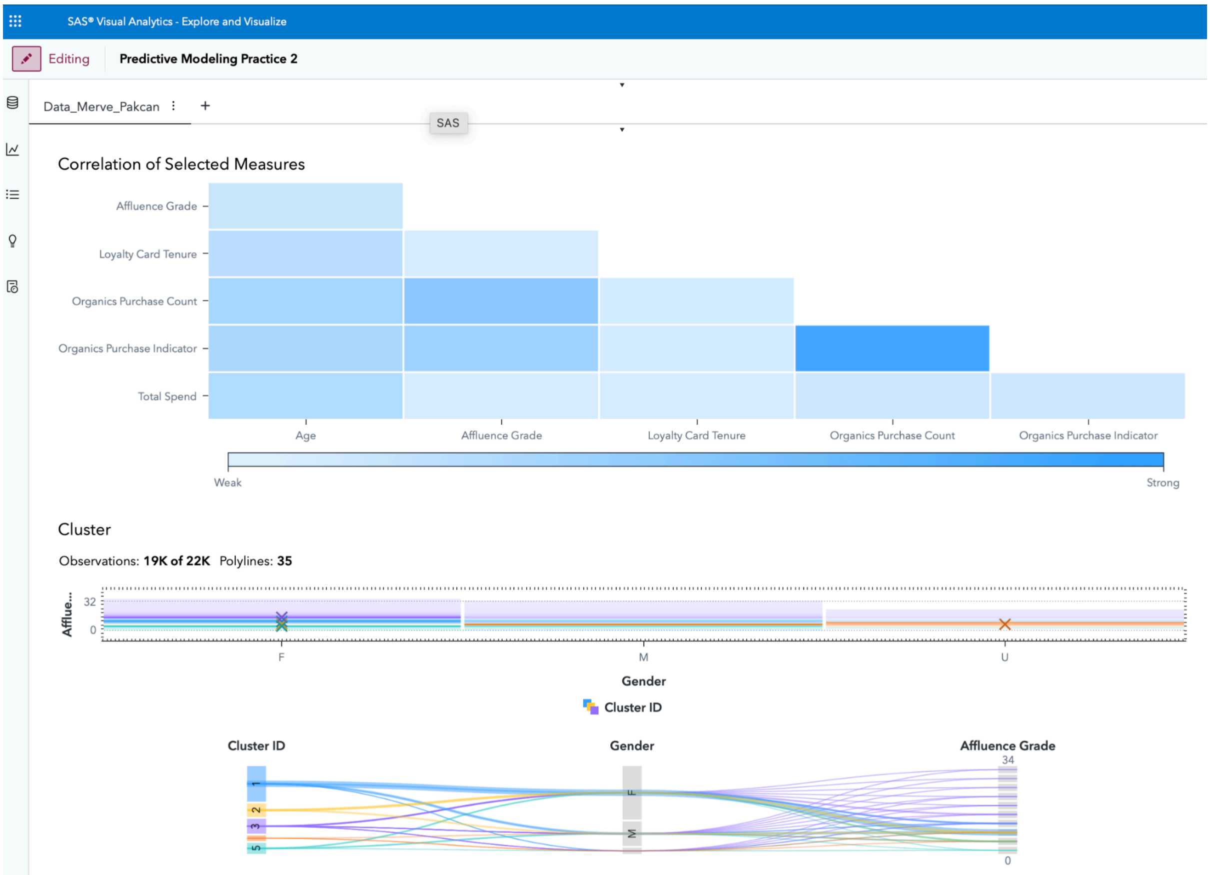
National University of Science and Technology Politehnica Bucharest
Advanced Analytics for Business
Predictive Modeling

Name: Merve Pakcan Tufenk

Practice 2. Handling missing values

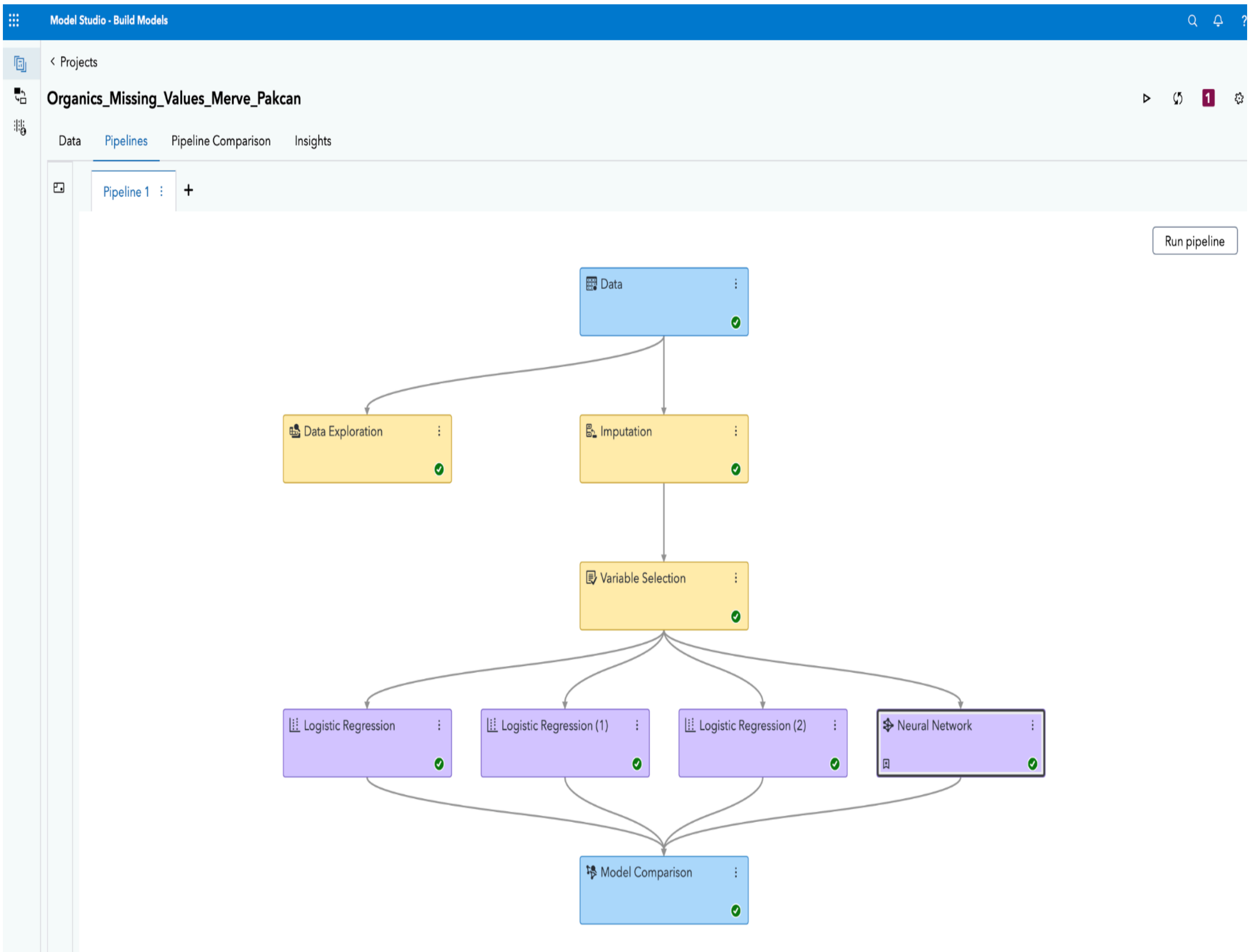
1.Missing value patterns and variable relationships were explored using Visual Analytics. A correlation heatmap revealed moderate relationships between variables such as Affluence Grade and Organics Purchase Count, while a clustering visualization based on Gender and Affluence Grade highlighted the structure of the dataset and missing categories.



2. As part of the data preparation process, I created a pipeline in SAS Model Builder and added a **Data Exploration** node to analyze the structure and quality of the Organics dataset.

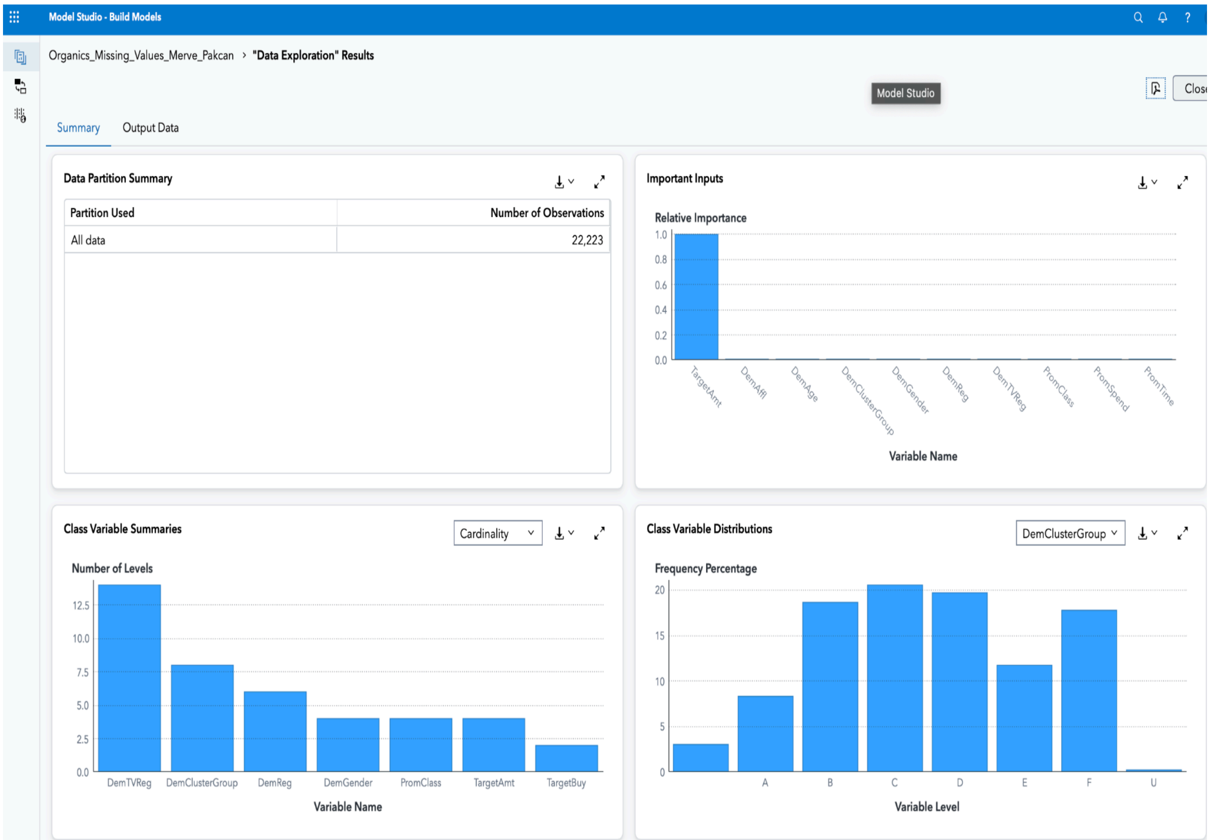
National University of Science and Technology Politehnica Bucharest
Advanced Analytics for Business

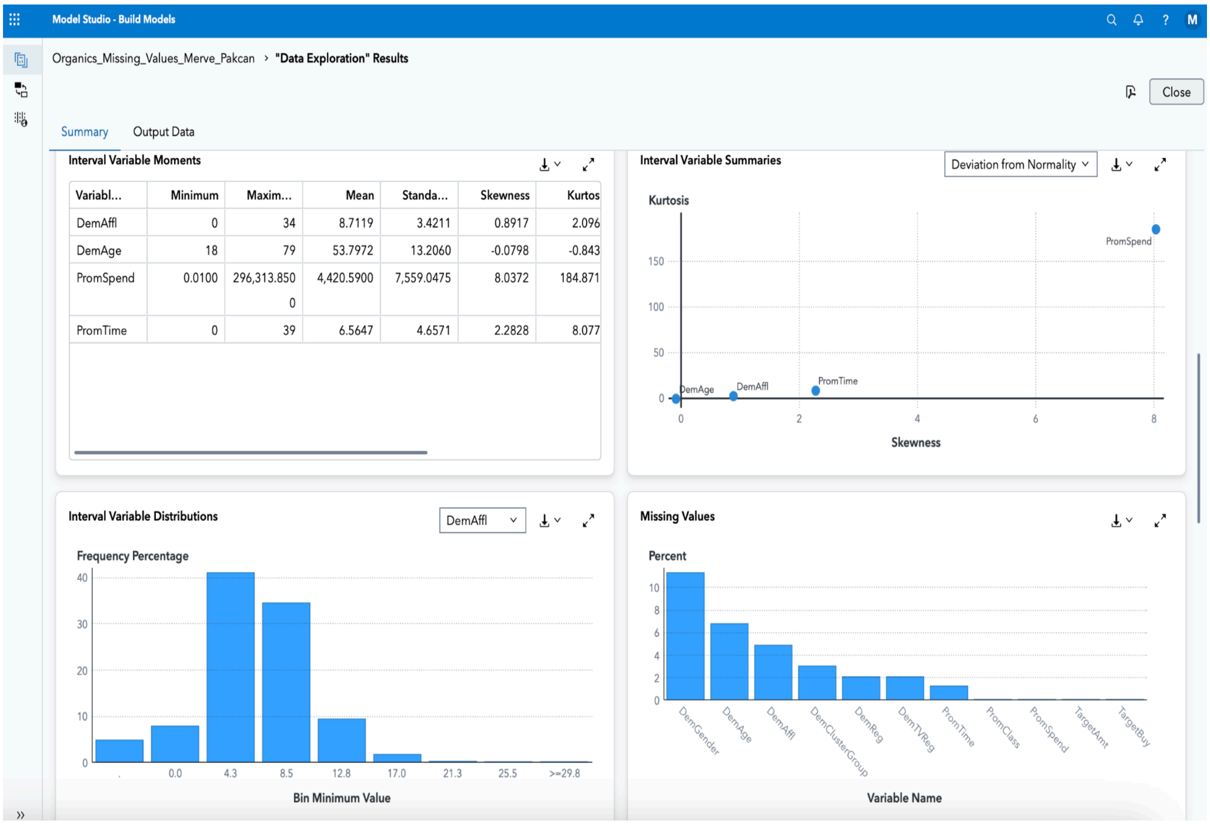
Predictive Modeling



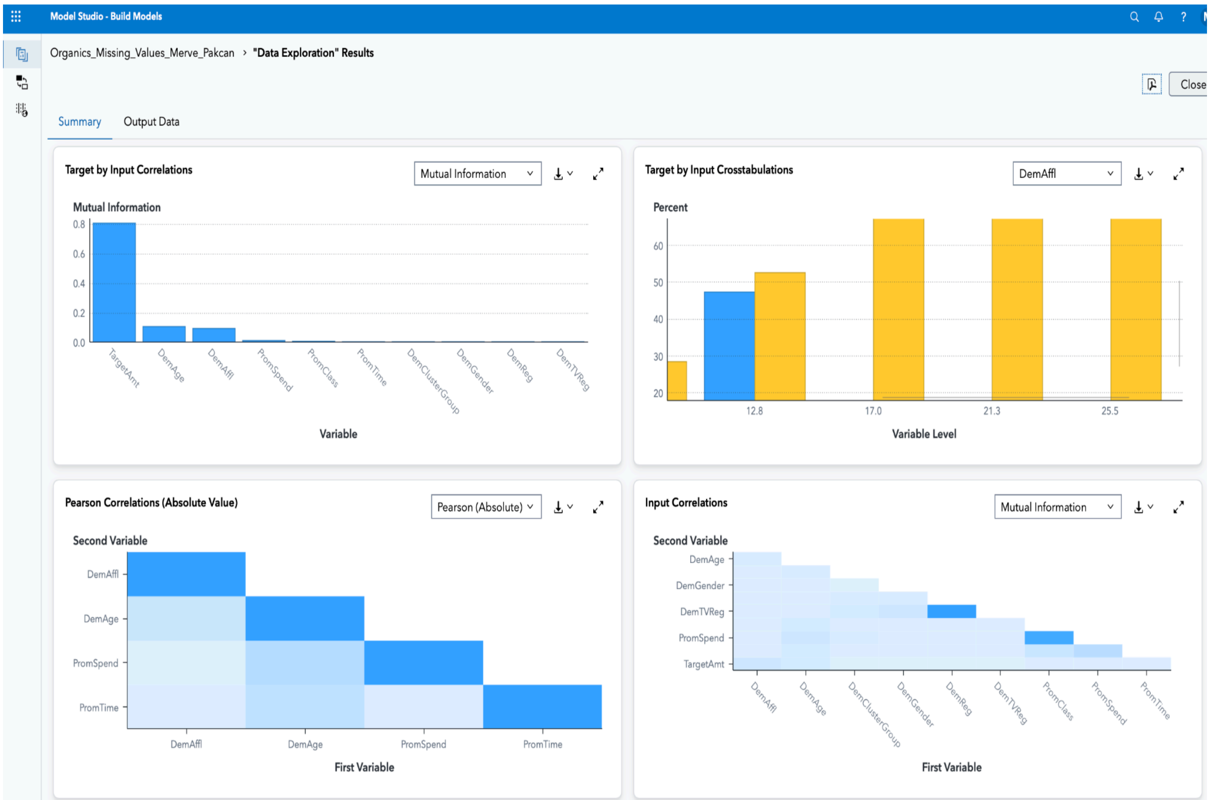
The results has different charts as below and highlighted several key points:
The dataset contains **22,223** observations. **DemGender** has the highest missing value rate **(11.3%)**, followed by DemAge and DemAffl. TargetAmt stands out as the most influential input based on both relative importance and mutual information. PromSpend shows extreme skewness and kurtosis, meaning it is far from normal distribution and may contain outliers.DemTVReg has the highest number of class levels, indicating a need for encoding. Overall, input variables are weakly correlated, with only a moderate relationship observed between PromSpend and PromTime.

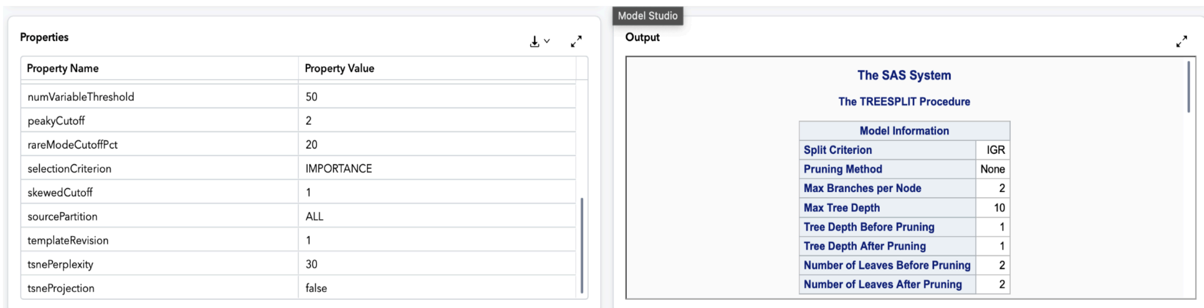
National University of Science and Technology Politehnica Bucharest
Advanced Analytics for Business
Predictive Modeling





National University of Science and Technology Politehnica Bucharest
Advanced Analytics for Business
Predictive Modeling





Property Name	Property Value
numVariableThreshold	50
peakyCutoff	2
rareModeCutoffPct	20
selectionCriterion	IMPORTANCE
skewedCutoff	1
sourcePartition	ALL
templateRevision	1
tsnePerplexity	30
tsneProjection	false

Model Information	
Split Criterion	IGR
Pruning Method	None
Max Branches per Node	2
Max Tree Depth	10
Tree Depth Before Pruning	1
Tree Depth After Pruning	1
Number of Leaves Before Pruning	2
Number of Leaves After Pruning	2

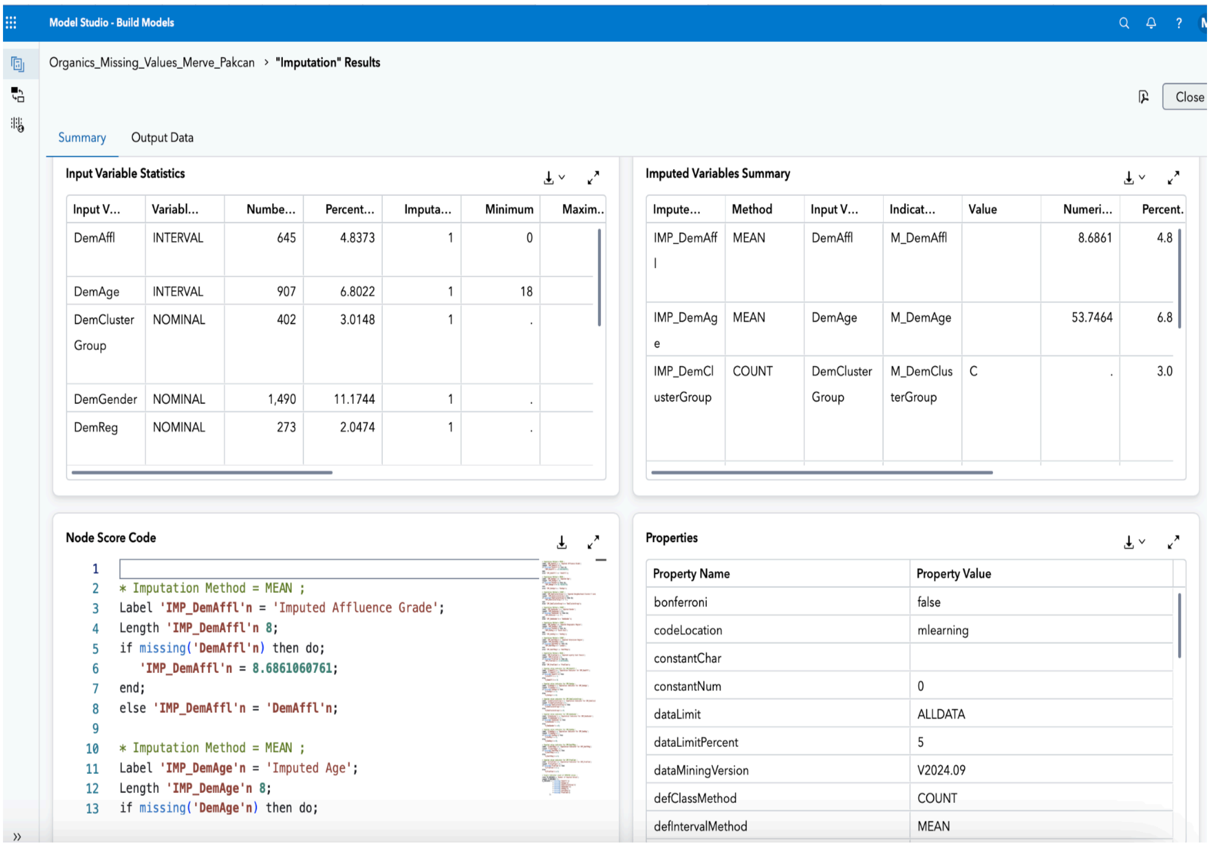
Missing ratio of each variables:
- DemGender: 11.30%
- DemAge: 6.79%
- DemAffl: 4.88%
- DemClusterGroup: 3.03%
- DemReg: 2.09%
- DemTVReg: 2.09%
- PromTime: 1.26%

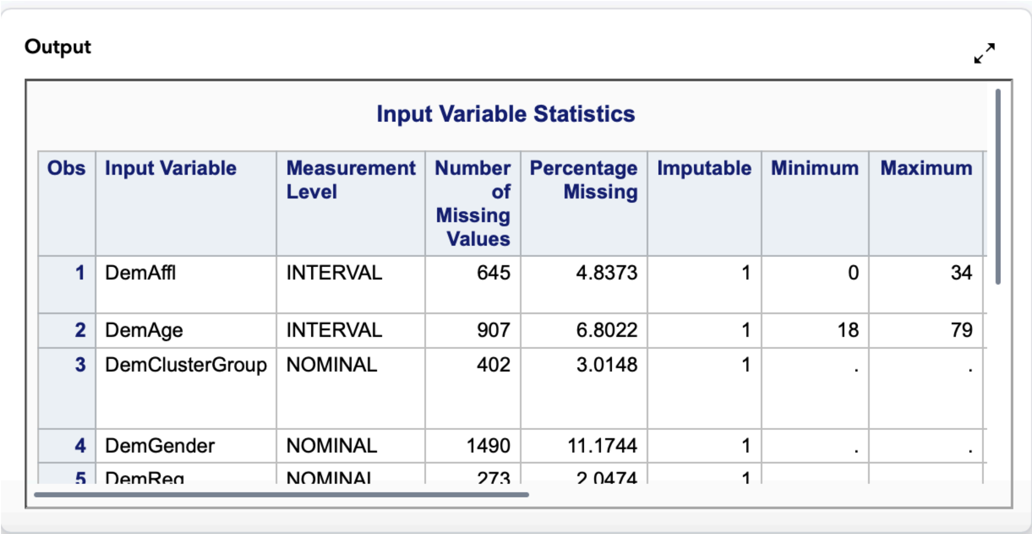
The correlation analysis between input variables was conducted using both Pearson correlation and mutual information heatmaps. Results show that most variables are weakly correlated. A moderate relationship is observed only between DemAge and PromSpend. This means that the input variables are not strongly related to each other, so multicollinearity is not expected to cause an issue in the model.
A complete case analysis was performed to understand how many records have no missing values. Based on the missing value summary, only a subset of the dataset is fully complete, which highlights the importance of handling missing data before modeling.

National University of Science and Technology Politehnica Bucharest
Advanced Analytics for Business

Predictive Modeling

3. An **imputation** block was added to the pipeline to handle missing values. The imputation was performed using common methods: **mean** for interval variables(DemAffl, DemAge) and **mode (count)** for categorical variables(DemClusterGroup).Although the default methods were mean and count, other common imputation techniques include median, regression, zero imputation, and hot-deck, as shown in the course materials.Additionally, both **single** and **unique indicators** were generated for the imputed categorical variables to help track and interpret missingness during modeling.





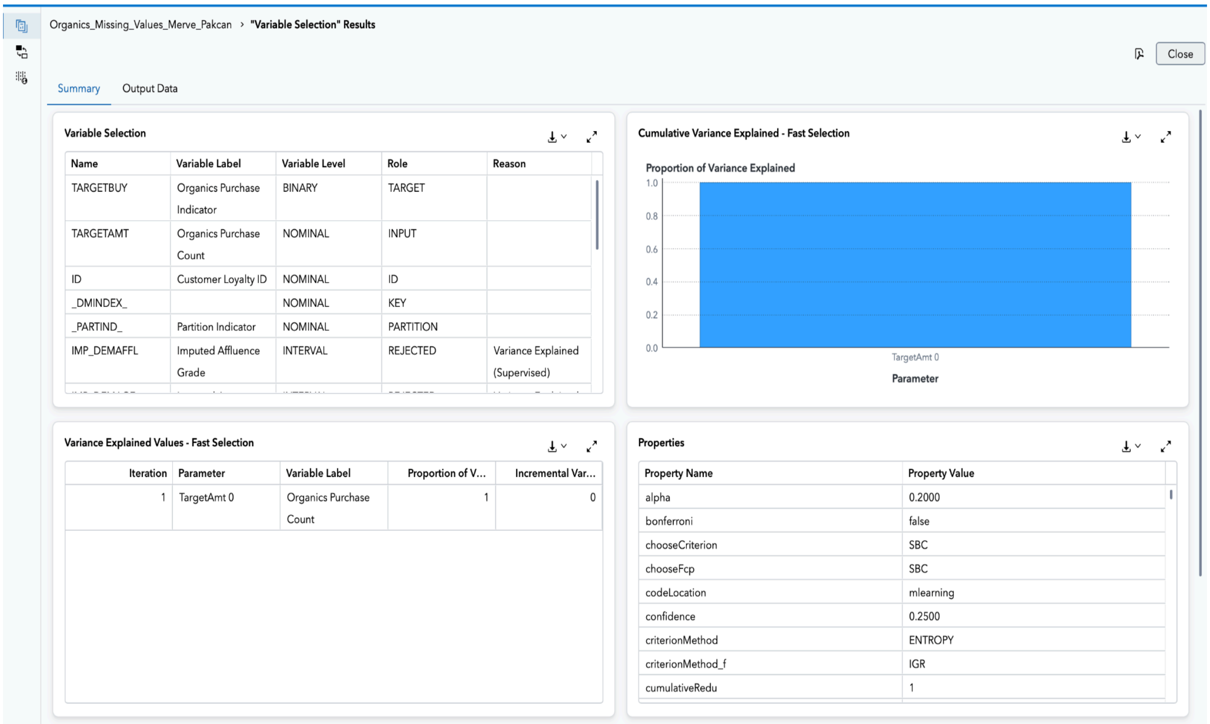
Output

Input Variable Statistics

Obs	Input Variable	Measurement Level	Number of Missing Values	Percentage Missing	Imputable	Minimum	Maximum
1	DemAffl	INTERVAL	645	4.8373	1	0	34
2	DemAge	INTERVAL	907	6.8022	1	18	79
3	DemClusterGroup	NOMINAL	402	3.0148	1	.	.
4	DemGender	NOMINAL	1490	11.1744	1	.	.
5	DemReg	NOMINAL	273	2.0474	1		

After enabling both **Single Indicator** and **Unique Indicators** in the Variable Imputation node, new indicator variables were created for each imputed input (M_DemAffl, M_DemAge, M_DemClusterGroup). These binary indicators help flag which values were originally missing, allowing the model to recognize and adjust for imputation. No change occurred in the number or percentage of missing values, but additional variables were added to the dataset, increasing its dimensionality and potentially improving model interpretability for categorical imputation.

A variable selection block was added to the pipeline to reduce dimensionality and focus on impactful predictors. The Fast Supervised Selection method identified *TargetAmt* as the only variable that explains the full variance of the target variable, with a Proportion of Variance Explained = 1.000000. Other variables were rejected based on this supervised evaluation. This result highlights *TargetAmt*'s strong predictive power for the target and simplifies the model by focusing on the most relevant variable.



Fast Variable Selection

The VARREDUCE Procedure

Number of Observations Read	13334
Number of Observations Used	13334

Selection Summary

Iteration	Parameter	Proportion of Variance Explained	SSE	MSE	AIC	AICC	BIC
1	TargetAmt 0	1.000000	0	0	-Infty	-Infty	-Infty

Selected Effects

Number	Selected Variable	Variable Type
1	TargetAmt	CLASS

Although TargetBuy is automatically recognized as the main target variable in the dataset, the Variable Selection block used TargetAmt as the response variable during its process. This step aimed to explore the variables that explain the variance in TargetAmt, which may provide additional insights.

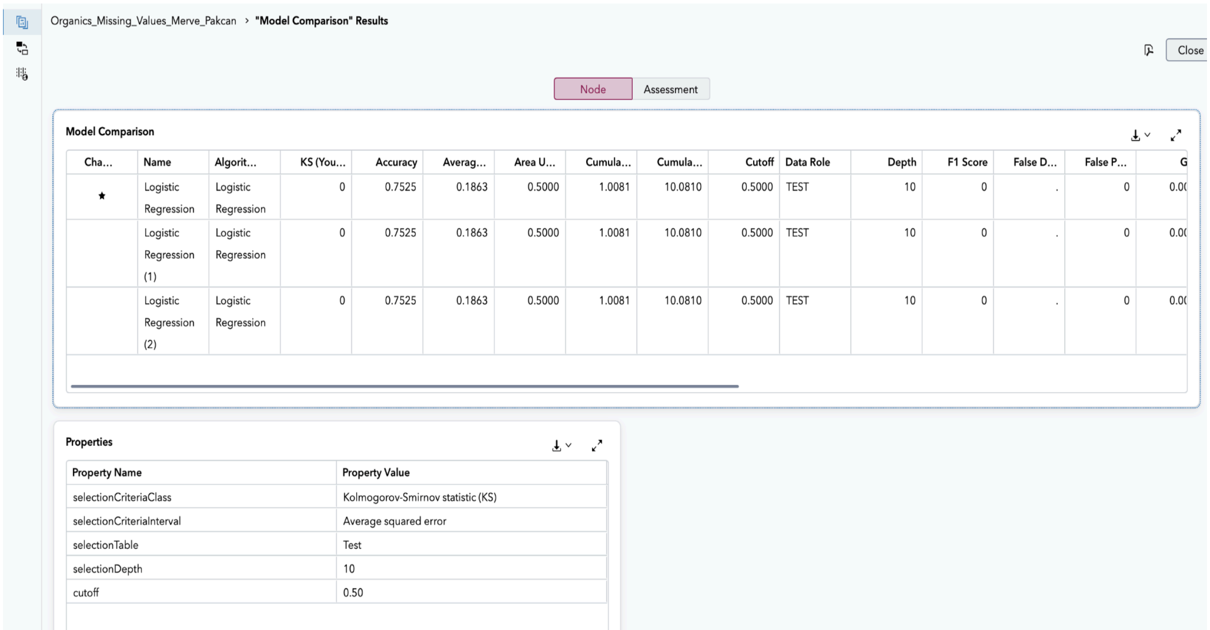
National University of Science and Technology Politehnica Bucharest
Advanced Analytics for Business

Predictive Modeling

4.Three Logistic Regression models were built using different binary link functions: **Logit**, **Probit**, and **Complementary Log-log**, to evaluate model performance under varying assumptions. All models were trained on the same imputed and selected dataset and assessed using a Model Comparison node.
Despite using different link functions, **all models yielded identical results**, with:

- **Accuracy:** 75.25%
- **AUC:** 0.5000
- **Misclassification Rate:** 24.75%
- **F1 Score:** 0
- **Cumulative Lift:** 1.0081

These identical performance metrics suggest that the choice of link function had **no practical impact** in this case. The lack of variation may reflect **weak or uninformative predictors** or a **highly imbalanced binary target (TargetBuy)**.

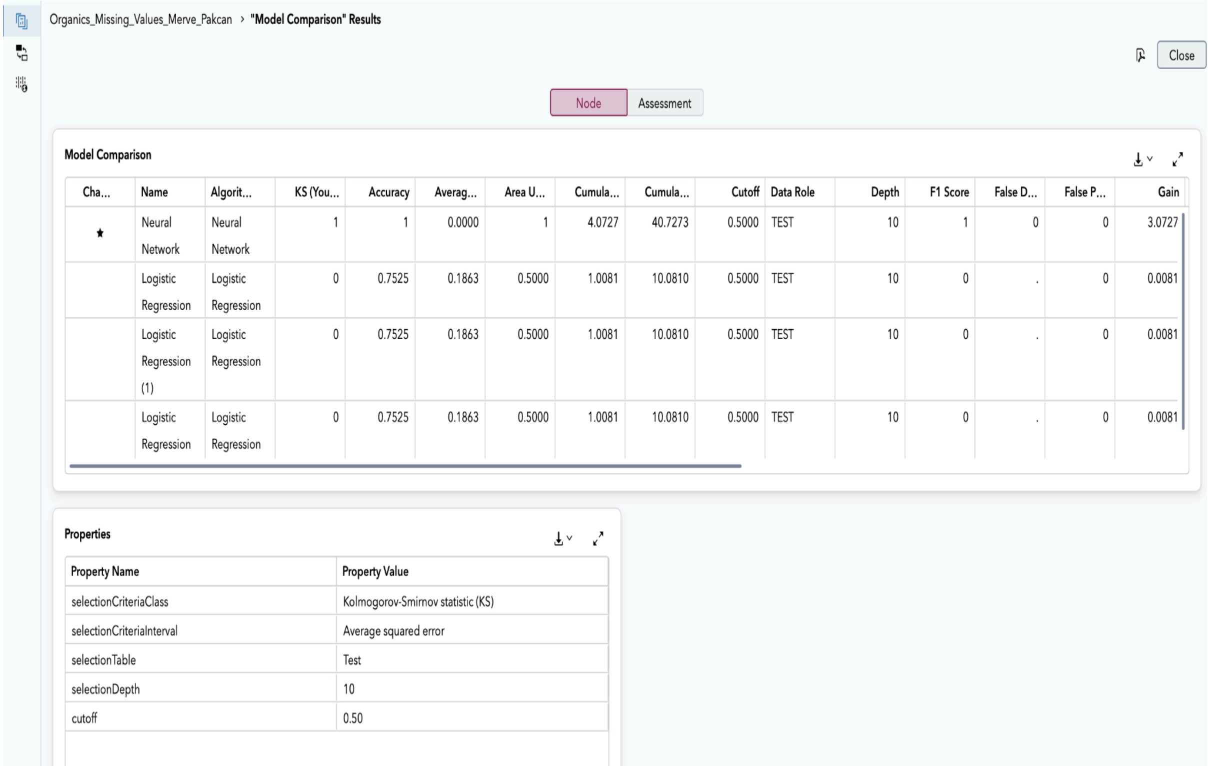


Cha...	Name	Algorit...	KS (You...	Accuracy	Averag...	Area U...	Cumula...	Cumula...	Cutoff	Data Role	Depth	F1 Score	False D...	False P...	G
★	Logistic Regression	Logistic Regression	0	0.7525	0.1863	0.5000	1.0081	10.0810	0.5000	TEST	10	0	.	0	0.0(
	Logistic Regression (1)	Logistic Regression	0	0.7525	0.1863	0.5000	1.0081	10.0810	0.5000	TEST	10	0	.	0	0.0(
	Logistic Regression (2)	Logistic Regression	0	0.7525	0.1863	0.5000	1.0081	10.0810	0.5000	TEST	10	0	.	0	0.0(

Property Name	Property Value
selectionCriteriaClass	Kolmogorov-Smirnov statistic (KS)
selectionCriteriaInterval	Average squared error
selectionTable	Test
selectionDepth	10
cutoff	0.50

5.All models, including three Logistic Regression models with different link functions and a Neural Network model, yielded identical results across all evaluation metrics. This strongly indicates that either the predictors were not informative for the target variable (TargetBuy) or that the classification problem is highly imbalanced. Despite algorithmic differences, no model demonstrated superior predictive capability in this case.

National University of Science and Technology Politehnica Bucharest
Advanced Analytics for Business
Predictive Modeling



Organics_Missing_Values_Merve_Pakcan › "Model Comparison" Results

Close

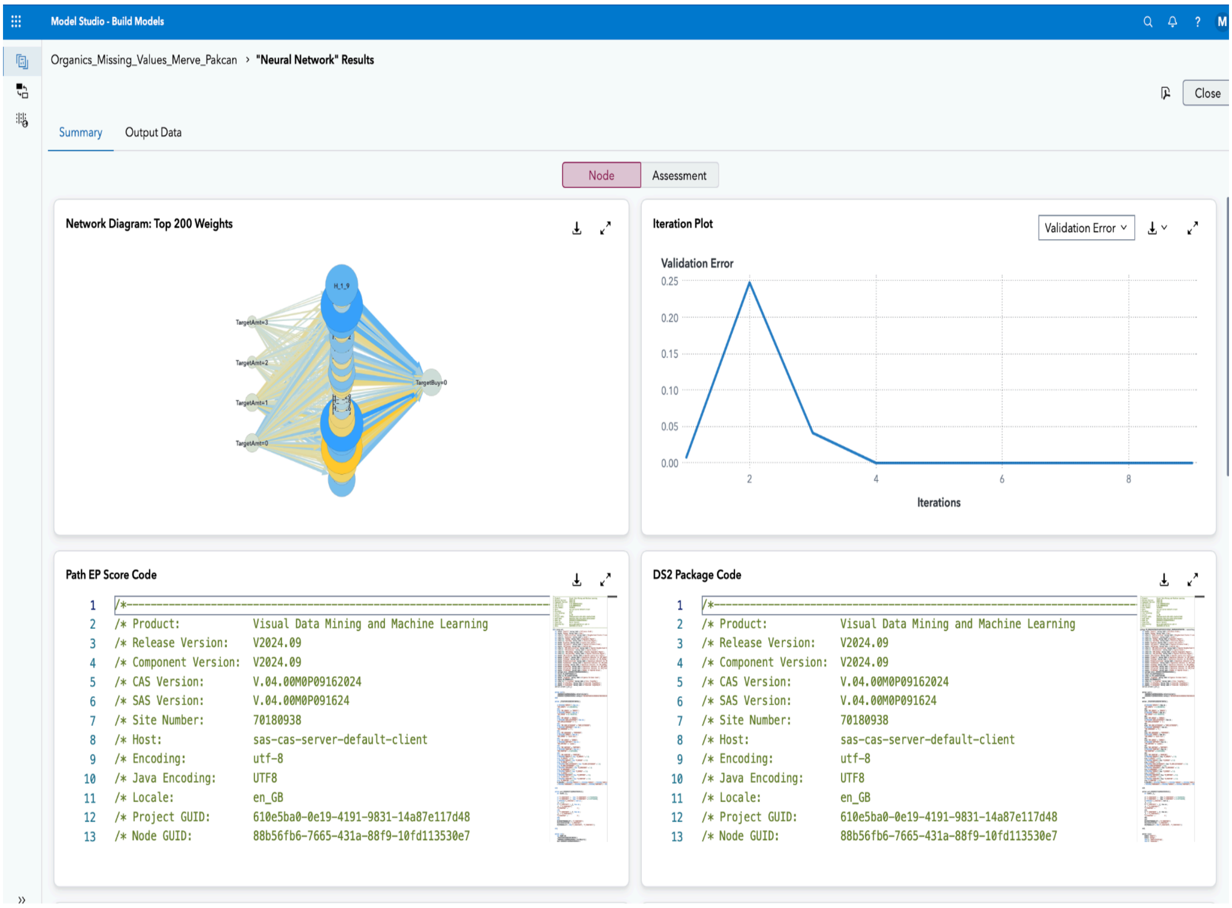
Node | Assessment

Model Comparison

Cha...	Name	Algorit...	KS (You...	Accuracy	Averag...	Area U...	Cumula...	Cumula...	Cutoff	Data Role	Depth	F1 Score	False D...	False P...	Gain
★	Neural Network	Neural Network	1	1	0.0000	1	4.0727	40.7273	0.5000	TEST	10	1	0	0	3.0727
	Logistic Regression	Logistic Regression	0	0.7525	0.1863	0.5000	1.0081	10.0810	0.5000	TEST	10	0	.	0	0.0081
	Logistic Regression (1)	Logistic Regression	0	0.7525	0.1863	0.5000	1.0081	10.0810	0.5000	TEST	10	0	.	0	0.0081
	Logistic Regression	Logistic Regression	0	0.7525	0.1863	0.5000	1.0081	10.0810	0.5000	TEST	10	0	.	0	0.0081

Properties

Property Name	Property Value
selectionCriteriaClass	Kolmogorov-Smirnov statistic (KS)
selectionCriteriaInterval	Average squared error
selectionTable	Test
selectionDepth	10
cutoff	0.50



Path EP Score Code

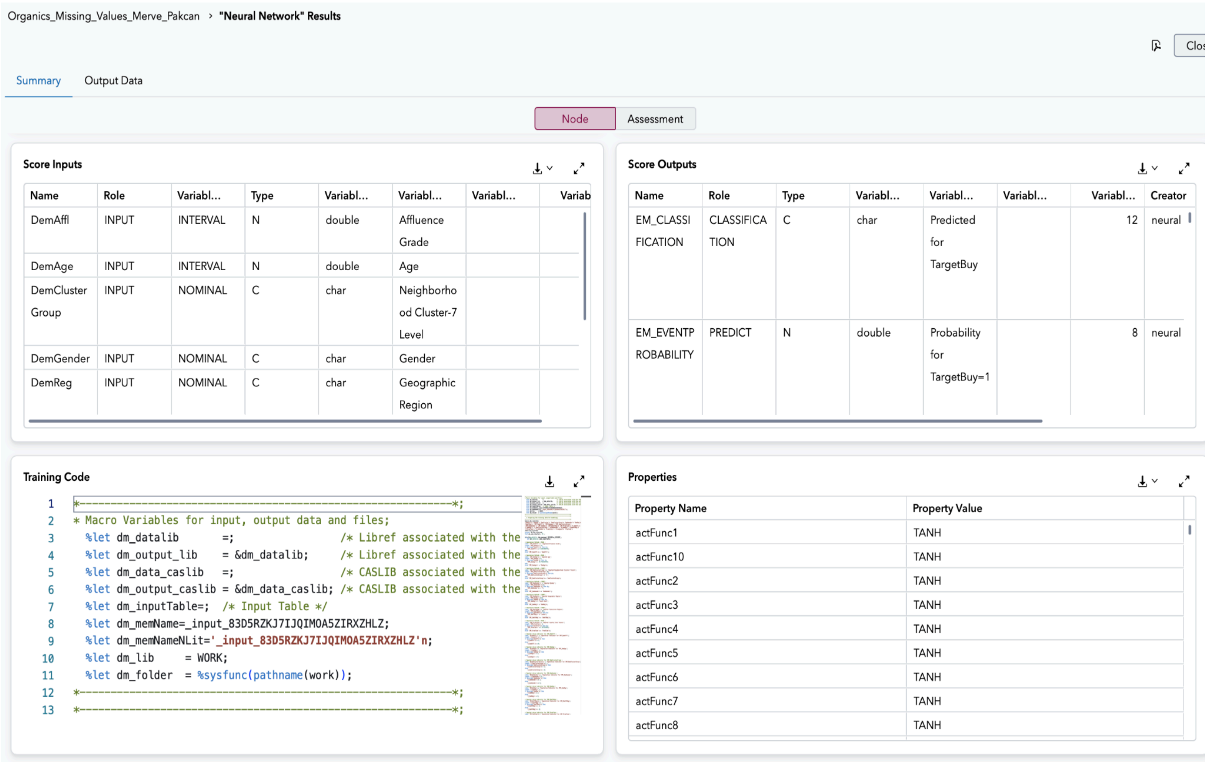
```
/*-----------------------------------------------------------------------
/* Product:            Visual Data Mining and Machine Learning
/* Release Version:    V2024.09
/* Component Version:  V2024.09
/* CAS Version:        V.04.00M0P09162024
/* SAS Version:        V.04.00M0P091624
/* Site Number:        70180938
/* Host:               sas-cas-server-default-client
/* Encoding:           utf-8
/* Java Encoding:      UTF8
/* Locale:             en_GB
/* Project GUID:       610e5ba0-0e19-4191-9831-14a87e117d48
/* Node GUID:          88b56fb6-7665-431a-88f9-10fd113530e7
```

DS2 Package Code


```
/*-----------------------------------------------------------------------
/* Product:            Visual Data Mining and Machine Learning
/* Release Version:    V2024.09
/* Component Version:  V2024.09
/* CAS Version:        V.04.00M0P09162024
/* SAS Version:        V.04.00M0P091624
/* Site Number:        70180938
/* Host:               sas-cas-server-default-client
/* Encoding:           utf-8
/* Java Encoding:      UTF8
/* Locale:             en_GB
/* Project GUID:       610e5ba0-0e19-4191-9831-14a87e117d48
/* Node GUID:          88b56fb6-7665-431a-88f9-10fd113530e7
```

National University of Science and Technology Politehnica Bucharest
Advanced Analytics for Business

Predictive Modeling



Output

The SAS System

The NNET Procedure

Model Information	
Model	Neural Net
Number of Observations Used	13334
Number of Observations Read	13334
Target/Response Variable	TargetBuy
Number of Nodes	56
Number of Input Nodes	4
Number of Output Nodes	2
Number of Hidden Nodes	50