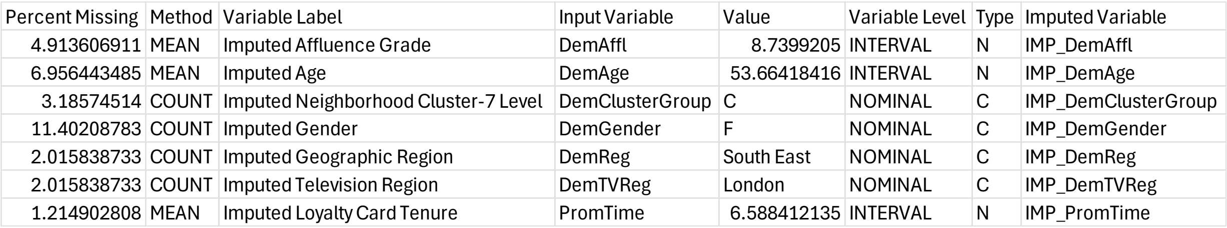
Neural Networks and Model Comparison

Data Handling & Manipulation

We set the dataset to partition as – 50% for training, and 50% for validation, using stratified split – this allows us to get 50% of the target variable “**TargetBuy**” (Organics Purchase Indicator) for both training and validation purposes.
Missing values in data are imputed with the missing percentage cutoff set to **50** which implies that variables whose percentage of missing values exceeds this cutoff are ignored.
The imputation method for class input variables is set to – **“Count”** and for interval variables its set to – **“Mean”**.



Percent Missing	Method	Variable Label	Input Variable	Value	Variable Level	Type	Imputed Variable
4.913606911	MEAN	Imputed Affluence Grade	DemAffl	8.7399205	INTERVAL	N	IMP_DemAffl
6.956443485	MEAN	Imputed Age	DemAge	53.66418416	INTERVAL	N	IMP_DemAge
3.18574514	COUNT	Imputed Neighborhood Cluster-7 Level	DemClusterGroup	C	NOMINAL	C	IMP_DemClusterGroup
11.40208783	COUNT	Imputed Gender	DemGender	F	NOMINAL	C	IMP_DemGender
2.015838733	COUNT	Imputed Geographic Region	DemReg	South East	NOMINAL	C	IMP_DemReg
2.015838733	COUNT	Imputed Television Region	DemTVReg	London	NOMINAL	C	IMP_DemTVReg
1.214902808	MEAN	Imputed Loyalty Card Tenure	PromTime	6.588412135	INTERVAL	N	IMP_PromTime

Fig. (1) Imputation Variable Summary

Neural Network Implementation

NN3 Layer

The network is set up with **3 hidden layers** with **16 neurons per hidden layer**.

The activation function is taken as “**Identify**”. Activation functions introduce non-linearity to the model, allowing it to learn and approximate complex relationships in the data. The “**Identify”** function, also known as the identify function, is one of the most straightforward activation function, where the output is equal to the input.

The function is given by - $\boxed{\boldsymbol{f(x) = x}}$

The error function (used to measure the difference between the actual and the predicted values) is set to **“Normal”** which indicates that the error function could be either mean squared error **(MSE)** or mean absolute error **(MAE)**.

The optimization method is taken to be **“Automatic”** which implies that the algorithm for optimization will be chosen based on the problem and data. It could be any of the commonly used ones such as – Adam, (SGD).

The **epochs** (maximum number of iterations) are set to 15 – this means that the training process will stop after 15 iterations.

Hyperparametric **Autotuning** is set to off – this disables the model’s ability to automatically adjust hyperparameters during training. This is done so that we can check for the model’s performance with the features we input during model initialization.

NN6 Layer

This network is set up with **6 hidden layers** with **16 neurons per hidden layer.** Increasing the number of hidden layers allows the model to learn more complex representations, but they also require more data and computational power to train.

Every other variable in this Neural Network model is the same as the one in NN3 Layer.

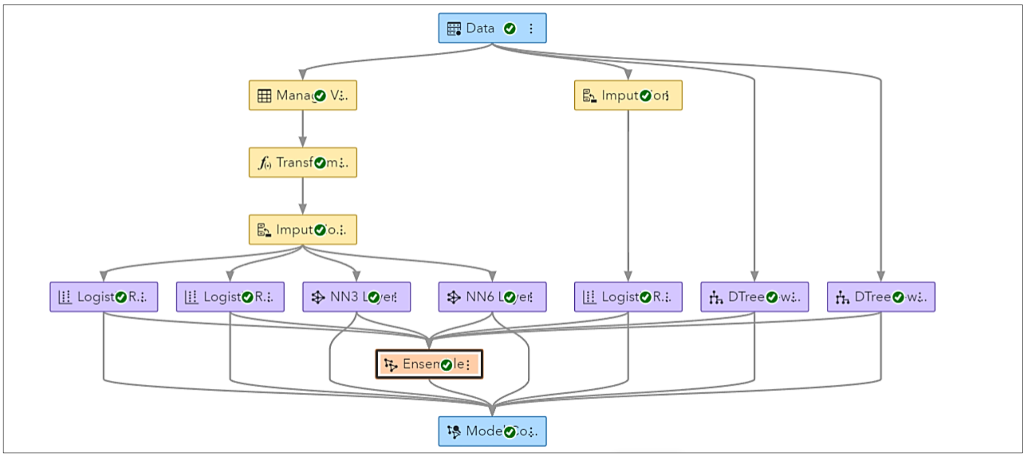


Fig. (2) SAS Model Studio - Pipeline

Ensemble Model

Ensemble Models combine predictions from multiple models to create a single consensus prediction, and its more accurate if the models in the Ensemble model are very different and produce different predictions – in this case the models are:

1. Logistic Regression Models
    a. Logistic Regression
    b. Logistic Regression (1)
    c. Logistic Regression (2)
2. Neural Network Models
    a. NN3 Layer (Neural Network with 3 hidden layers)
    b. NN6 Layer (Neural Network with 6 hidden layers)
3. Decision Tree 3-way split (3-way split means that a single node is split into three child nodes based on threshold criteria at every split)
4. Decision Tree 2-way split (2-way split means that a single node is split into two child nodes based on threshold criteria at every split)

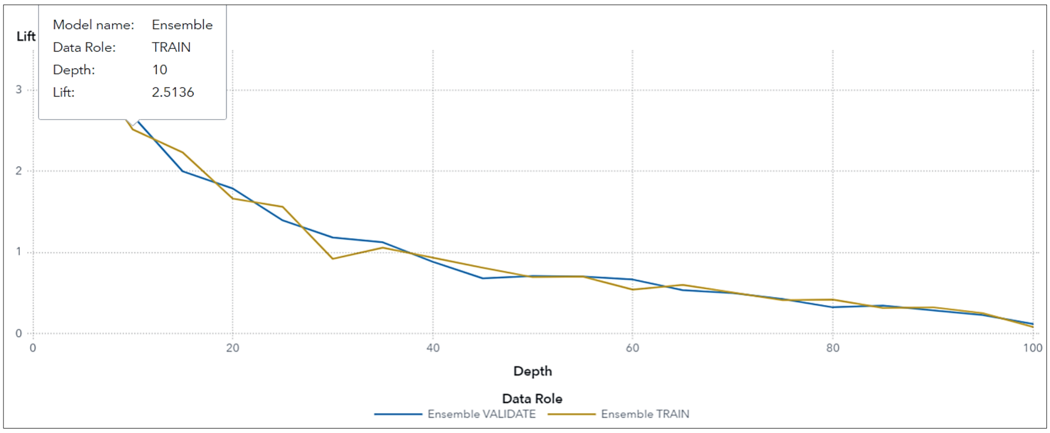


Fig. (3) Lift Value of Ensemble Model

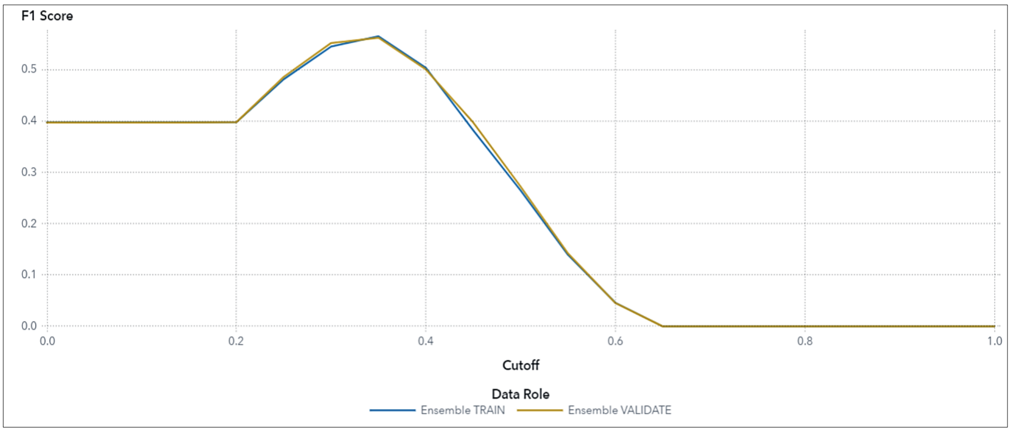
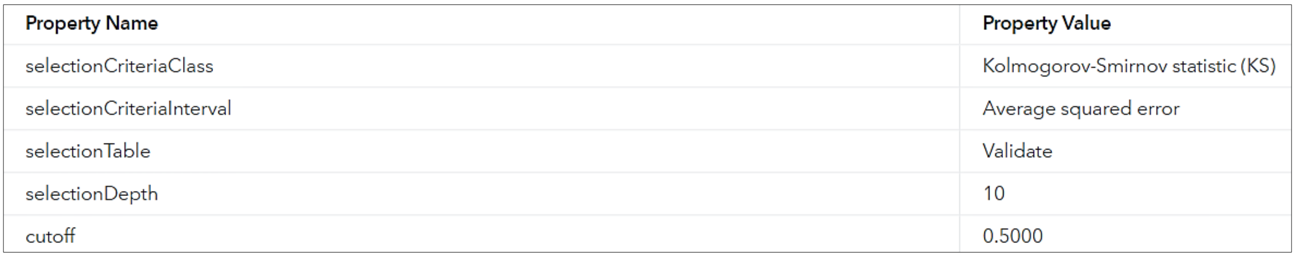


Fig. (4) F1 Score of Ensemble Model

From the Lift Value at **Depth 10** = **2.5136**; we can conclude that the Ensemble Model is 2.5136 times better at predicting the target variable (**TargetBuy**) than a random model.

The **F1 Score** (It represents the proportion of true positive predictions (correctly predicted positive instances) among all instances predicted as positive) in the Validate partition at the Cutoff 0.5 is **0.272**.

Model Comparison



Property Name	Property Value
selectionCriteriaClass	Kolmogorov-Smirnov statistic (KS)
selectionCriteriaInterval	Average squared error
selectionTable	Validate
selectionDepth	10
cutoff	0.5000

Fig. (5) Model Comparison Properties

Name	*Accuracy*	*Misclassification Rate*	*KS (Youden)*	*F1 Score*
Logistic Regression (1)	*0.804158042*	*0.195841958*	*0.433315591*	*0.474649928*
Logistic Regression	*0.804158042*	*0.195841958*	*0.433315591*	*0.474649928*
Logistic Regression (2)	*0.804158042*	*0.195841958*	*0.433315591*	*0.474649928*
Ensemble	*0.785167852*	*0.214832148*	*0.426973164*	*0.272034157*
NN6 Layer	*0.487714877*	*0.512285123*	*0*	*0.263076126*
NN3 Layer	*0.752317523*	*0.247682477*	*0*	*0*
DTree 2-way split	*0.752317523*	*0.247682477*	*0*	*0*
DTree 3-way split	*0.752317523*	*0.247682477*	*0*	*0*

Fig. (6) Comparison factors for different models

Accuracy is the count of correct prediction decisions divided by the total decisions. The goal of any model is to maximize the accuracy.
Misclassification Rate is defined as the count of incorrect prediction decisions divided by total decisions. The goal is to minimize it.
Kolmogorov-Smirnov (KS) Statistic describes the ability of a model to separate the primary and secondary outcomes. The goal of any model is to maximize this (closer to 1)

Based on the goal of the comparison factors the Logistic Regression (1) model outperforms all the other algorithmic models such as Neural Network, Decision Tree, and the Ensemble Model.

Model Champion – “**Logistic Regression (1)**”
This model was chosen based on the KS (Youden) for the Validate partition **(0.43). 80.42%** of the Validate dataset was correctly classified using the trained model of Logistic Regression (1) model.

The selection method in the logistic regression model is **“stepwise”**. In this model, training begins in the forward model, choosing variables based on their statistical significance. Then it also removes variables, then tests to see if the removed variable was statistically significant or not.

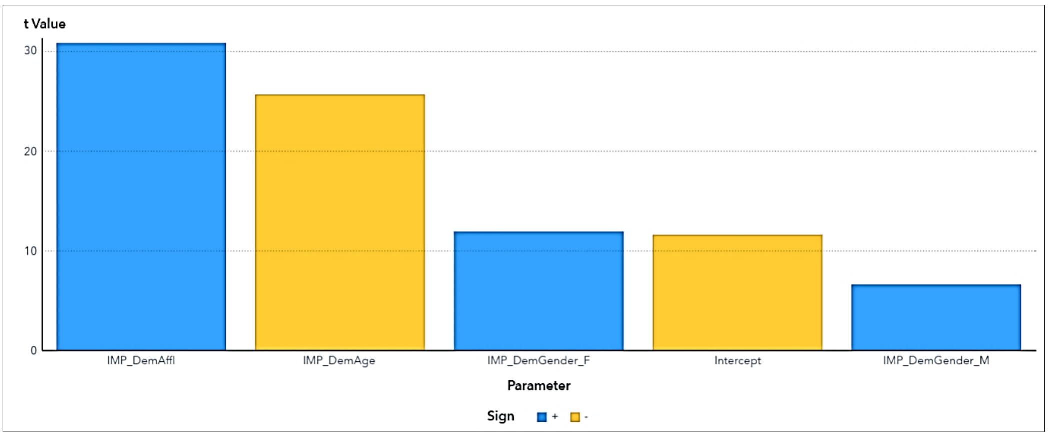


Fig. (7) t-value of the Parameters

The **five most important factors** from the above t-value comparison bar chart are Imputed Affluence Grade, Imputed Age, Imputed Gender, Imputed Neighborhood Cluster-7 Level, and Imputed Geographic Region.

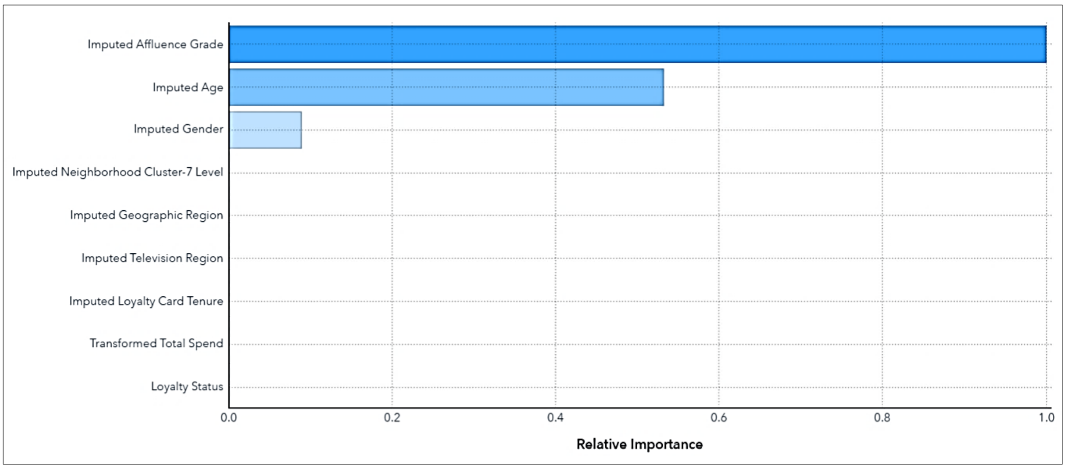


Fig. (8) Relative Importance of features

The most important input for this model in terms of relative importance is **Imputed Affluence Grade**. The input Imputed Gender has a relative importance of 0.089, which implies that it is 0.089 times as important as Imported Affluence Grade.

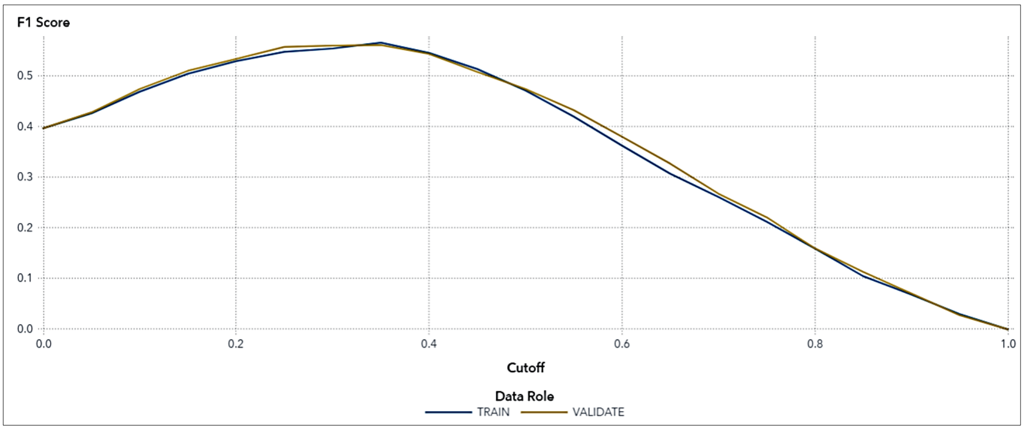


Fig. (9) F1 Score of Logistic Regression (1) Model

F1-Score in the **Validate** partition at the cutoff of 0.5 is **0.475**.
A higher F1 Score indicates that both precision and recall are high, which implies that the model is performing well in terms of both minimizing false positives and false negatives.
F1 Score is the harmonic mean of precision and recall, calculated by $\mathbf{2*(\frac{precision * recall}{precision+recall})}$.

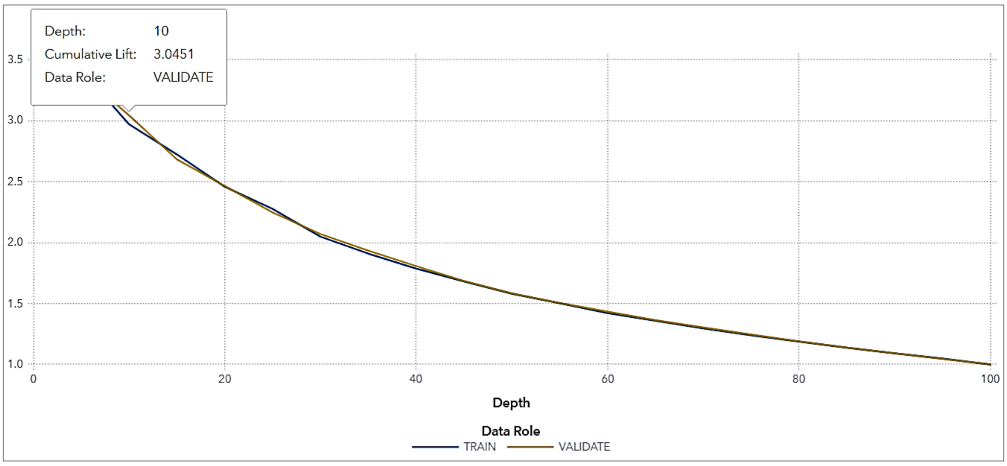


Fig. (10) Cumulative Lift for Logistic Regression (1) Model (with Lift value at Depth = 10)

The Lift value of **3.0451** at **Depth = 10** (Top 10th percentile) implies that this model is 3.0451 times better at identifying the target variable (**TargetBuy**) than a random model.

Conclusion

Based on the winning model – **Logistic Regression (1)** to identify and distinguish customers who are likely to transition to purchasing organic products – the factors which would be instrumental are –
1. Imputed Affluence Grade
2. Imputed Age
3. Imputed Gender
4. Imputed Neighborhood Cluster 7-level
5. Imputed Geographic Region

The model accuracy is **80.42%** - which implies that the logistic regression model can correctly predict those many cases out of all the possible cases.