MKTG 746- Big Data & Predictive Analytics

Group Project

Online Shopping Purchasing Intentions

Section: **001**

Group: **8**

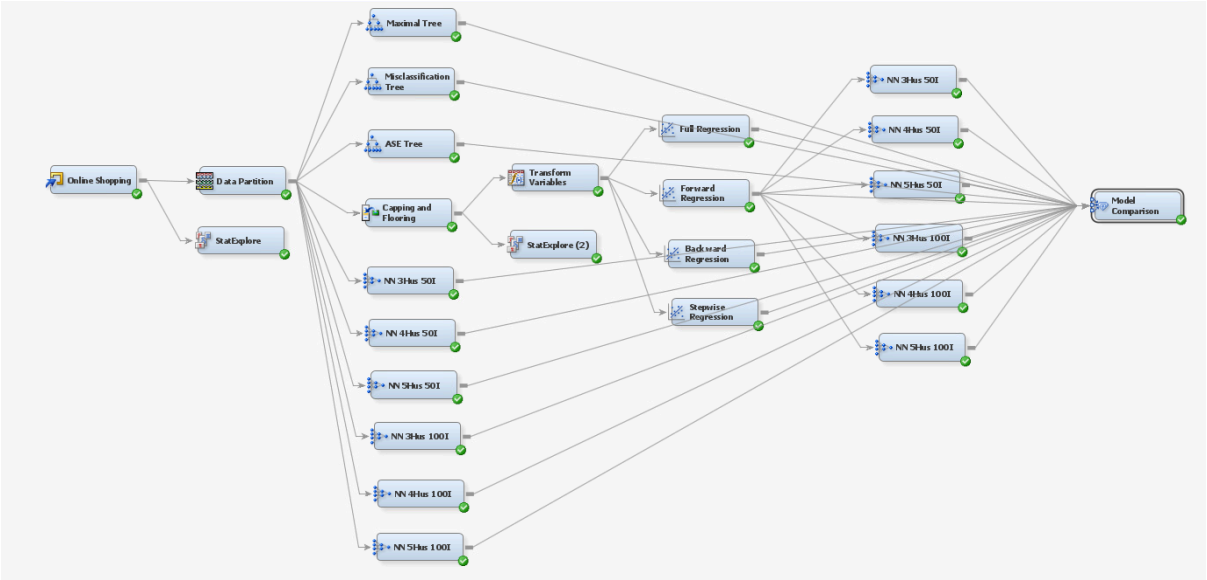
Instructor: **Prof. Milad Rezamand**

Group Members: **Sukhpreet Singh (301333453)**
Mankanwar Singh (301330855)
Sukhmani Kaur (301467664)

Table of Contents

1. Executive Summary

A predictive analysis was conducted to identify the most relevant factors influencing online shopping behavior and predict the likelihood of a purchase. The prediction model was run on SAS Enterprise Miner. After importing and wrangling the Online Shoppers Purchasing Intention Dataset, various models, including decision trees, logistic regressions, and neural networks, were executed. The best model was then identified through model comparison, allowing for the selection of the most accurate and reliable predictor of online shopping purchasing intention.



The model comparison analysis showed that the best performing model was the Neural Network with 5 hidden units and 100 iterations. This model, based on the lowest average squared error, outperformed the other models. However, it did not provide insights into which specific variables contributed most to predicting the target variable, **Revenue**.

Further analysis of the best decision tree (Maximal Tree) and regression models (Forward and Stepwise Regression) revealed the key drivers behind revenue generation. Variables such as age, average number of cigarettes smoked per day, systolic blood pressure, and glucose levels were identified as the most significant predictors.

The findings suggest that visitors who use certain browsers (such as Browser 1 compared to Browser 13), those who exhibit higher Exit Rates, and those visiting during certain months (like November compared to September) were more likely to make a purchase, offering valuable insights for targeting customers in future marketing efforts. Additionally, visitors who are new (compared to returning visitors) and those who experience special promotions are more likely to convert, making them key segments for future campaigns.

2. Introduction

The **Online Shoppers Purchasing Intention** dataset is designed to study and analyze user behavior in online shopping sessions. The dataset captures various attributes of user interactions on an e-commerce website, such as the number and duration of page visits, traffic sources, session timings, and conversion outcomes. By analyzing these variables, we can identify patterns and factors that influence purchasing decisions, providing actionable insights to optimize user experience and increase online sales.

Online shopping continues to dominate the retail industry, making it essential to understand user behavior, identify pain points, and optimize website performance. This dataset allows us to explore key questions about user engagement and purchasing intentions by leveraging advanced statistical and machine learning techniques.

3. File Import

a. Data Source

The dataset used for this project was the Online Shoppers Purchasing Intention Dataset, obtained from the UCI Machine Learning Repository via https://archive.ics.uci.edu/dataset/468/online+shoppers+purchasing+intention+dataset (Sakar, C. & Kastro, Y., 2018) The dataset captures behavioral patterns of users in e-commerce sessions and aims to predict whether a session will lead to a purchase. The dataset includes 12,330 instances collected from a large e-commerce website. The data was compiled during a one-year period to predict whether a visitor's session would result in a purchase and has 18 attributes, which include information about user browsing behavior, traffic sources, and session outcomes. The study's primary goal was to analyze online shopping behavior and identify factors influencing purchasing intentions. This dataset provides valuable insights for businesses seeking to improve their e-commerce platforms, enhance customer experience, and optimize sales conversions.

b. Data Dictionary

The dataset combined 17 variables, with 7 interval variables, 2 binary variables and 2 nominal variables. The meaning of each variable is shown below:

Variable Names	Levels	Description
Administrative	Interval	Number of administrative pages visited by the user during the session
Administrative_Duration	Interval	Total time spent on administrative pages during the session
Informational	Interval	Number of informational pages visited by the user during the session
Informational_Duration	Interval	Total time spent on informational pages during the session
ProductRelated	Interval	Number of product-related pages visited by the user during the session
ProductRelated_Duration	Interval	Total time spent on product-related pages during the session
Bounce Rates	Interval	Percentage of visitors who enter the website and leave without interacting further
ExitRates	Interval	Percentage of visitors who leave the website from a specific page
PageValues	Interval	Average value of a page, calculated based on the revenue generated by the pages a user visited
SpecialDay	Interval	Indicator of closeness to a special day (e.g., Black Friday, Christmas), ranging from 0 to 1
Month	Nominal	Month of the year when the session occurred
OperatingSystems	Nominal	Operating system used by the user during the session (e.g., Windows, macOS, Android)
Browser	Nominal	Browser used by the user during the session (e.g., Chrome, Firefox, Safari)
Region	Nominal	Geographic region of the user (e.g., North America, Europe)

TrafficType	Nominal	Source of website traffic, identified by numerical codes (e.g., organic search, referral)
VisitorType	Nominal	Type of visitor categorized as "New_Visitor," "Returning_Visitor," or "Other.
Weekend	Binary	Whether the session occurred on a weekend (True/False)
Revenue	Binary	Whether the session resulted in a purchase (True/False)

c. File Import

Before importing the dataset into SAS Enterprise Miner, it had been realized that there were no missing values. Then the CSV file was imported into SAS Enterprise Miner with the File Import node.

Under Edit Variables of the File Import node, the level of each variable was selected to the respective data type. Since the objective of this analysis was to predict whether a user's session would result in a purchase, the target variable was **Revenue**. This variable is binary, where 0 indicates no purchase (session did not result in a purchase) and 1 indicates a purchase (session resulted in a purchase) in a year.

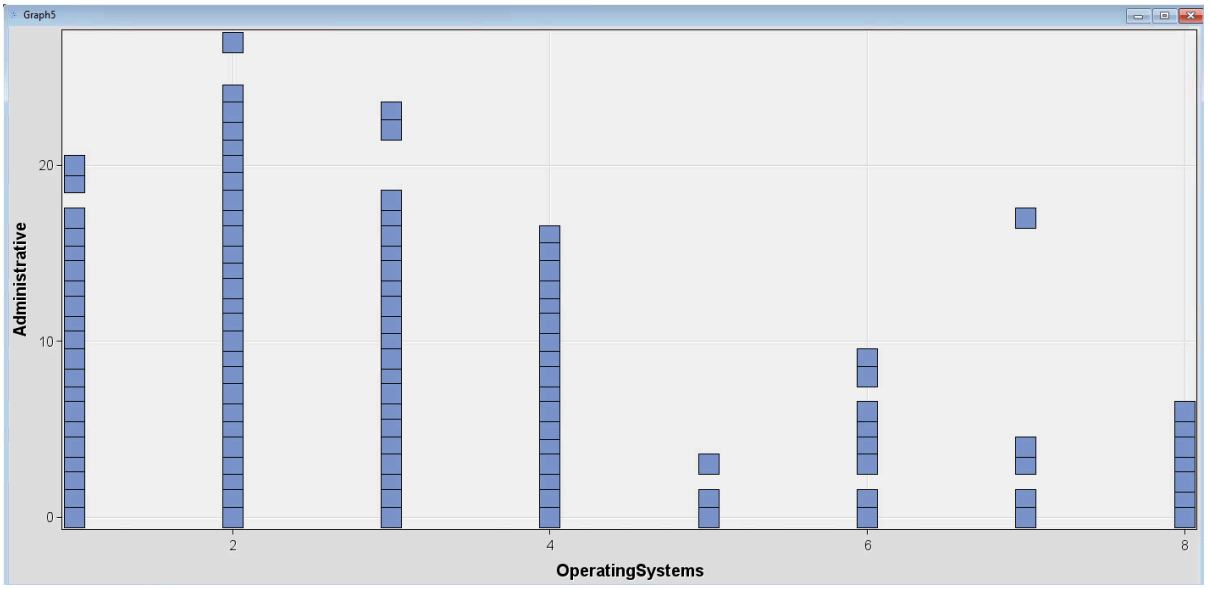
d. Data Leakage

The potential data leakage was observed due to relationships between certain variables like VisitorType and Revenue, ProductRelated and ProductRelated_Duration, Month and SpecialDay, TrafficType and Revenue and Weekend and Revenue

To summarize, Revenue was assigned as the target variable. The remaining 17 variables were input.

The dataset has some variables that show **no relationship**. **OperatingSystems** and **Administrative**

- **Region** and **Administrative**
- **TrafficType** and **Administrative**
- **Weekend** and **Administrative**
- **OperatingSystems** and **Administrative_Duration**
- **Region** and **Administrative_Duration**
- **Weekend** and **Administrative_Duration**

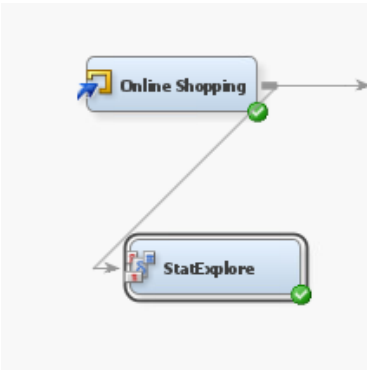


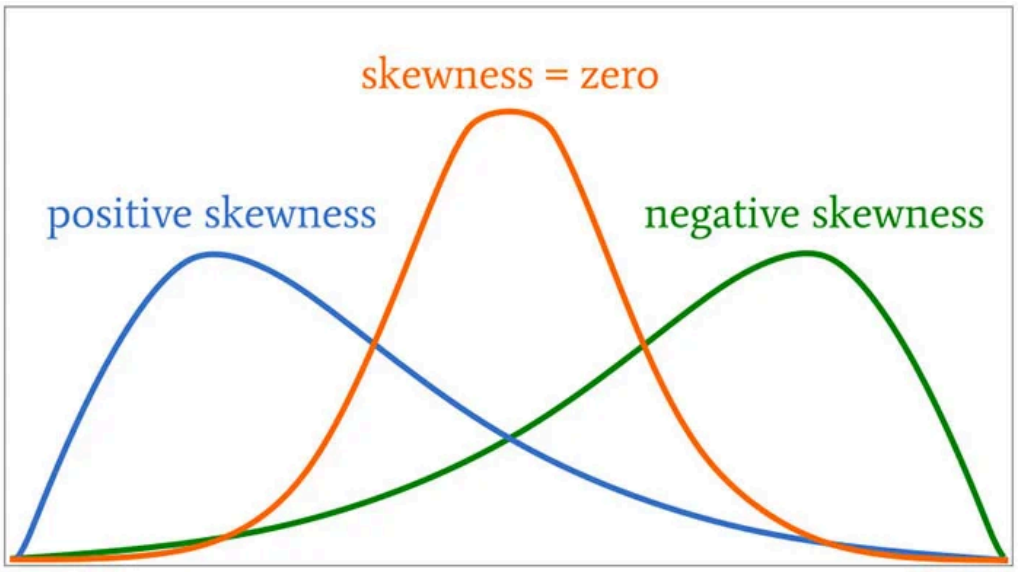
Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Administrative	Input	Interval	No		No	.	.
Administrative_	Input	Interval	No		No	.	.
BounceRates	Input	Interval	No		No	.	.
Browser	Input	Nominal	No		No	.	.
ExitRates	Input	Interval	No		No	.	.
Informational	Input	Interval	No		No	.	.
Informational_	Input	Interval	No		No	.	.
Month	Input	Nominal	No		No	.	.
OperatingSyste	Input	Nominal	No		No	.	.
PageValues	Input	Interval	No		No	.	.
ProductRelated	Input	Interval	No		No	.	.
ProductRelated	Input	Interval	No		No	.	.
Region	Input	Nominal	No		No	.	.
Revenue	Target	Binary	No		No	.	.
SpecialDay	Input	Ordinal	No		No	.	.
TrafficType	Input	Nominal	No		No	.	.
VisitorType	Input	Nominal	No		No	.	.
Weekend	Input	Binary	No		No	.	.

4. Data Wrangling

a. Stat Explorer

To figure out whether there were redundant or irrelevant variables, a StatExplore node was connected to the File Import node to further explore the variables.





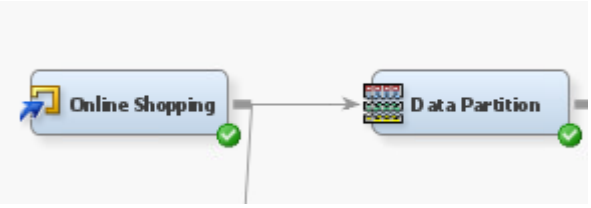
The StatExplore node revealed that there is no missing value but highly skewed variables.

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
Administrative	INPUT	2.315166	3.321784	12330	0	0	1	27	1.960357	4.701146
Administrative_Duration	INPUT	80.81861	176.7791	12330	0	0	7.5	3398.75	5.615719	50.55674
BounceRates	INPUT	0.022191	0.048488	12330	0	0	0.003111	0.2	2.947855	7.723159
ExitRates	INPUT	0.043073	0.048597	12330	0	0	0.025152	0.2	2.148789	4.017035
Informational	INPUT	0.503569	1.270156	12330	0	0	0	24	4.036464	26.93227
Informational_Duration	INPUT	34.4724	140.7493	12330	0	0	0	2549.375	7.579185	76.31685
PageValues	INPUT	5.889258	18.56844	12330	0	0	0	361.7637	6.382964	65.63569
ProductRelated	INPUT	31.73147	44.4755	12330	0	0	18	705	4.341516	31.21171
ProductRelated_Duration	INPUT	1194.746	1913.669	12330	0	0	598.8738	63973.52	7.263228	137.1742

b. Data Partition

To prevent overfitting or underfitting of the predictive model, its performance was optimized using the Data Partition node.

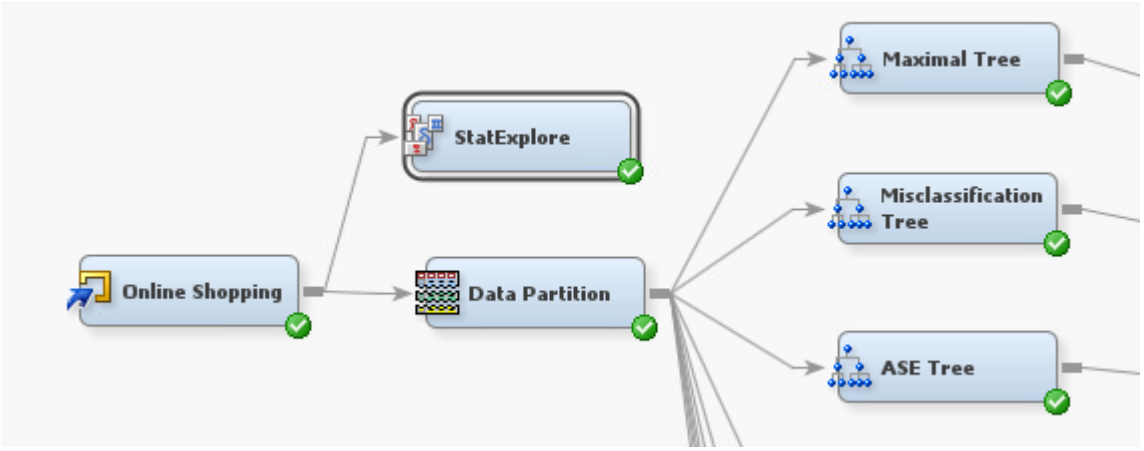
Under the Data Set Allocations section of the Property Panel, the training value was set to 50.0, the validation value was set to 50.0, and the test value was set to 0.



.. Property	Value
General	
Node ID	Part
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Output Type	Data
Partitioning Method	Default
Random Seed	12345
Data Set Allocations	
Training	50.0
Validation	50.0
Test	0.0
Report	
Interval Targets	Yes
Class Targets	Yes
Status	
Create Time	03/12/24 11:30 AM
Run ID	5ac26dcb-0bba-864a-b433-aeeb713cd0ea
Last Error	
Last Status	Complete
Last Run Time	03/12/24 9:31 PM
Run Duration	0 Hr. 0 Min. 2.37 Sec.
Grid Host	
User-Added Node	No

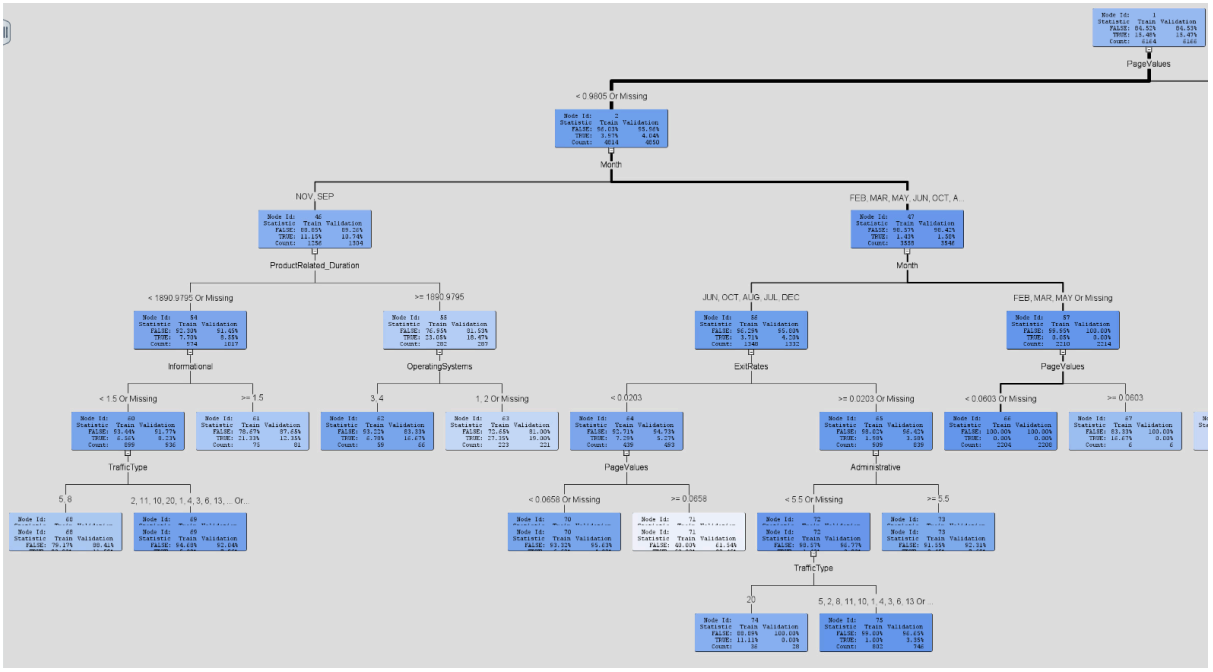
5. Decision Tree

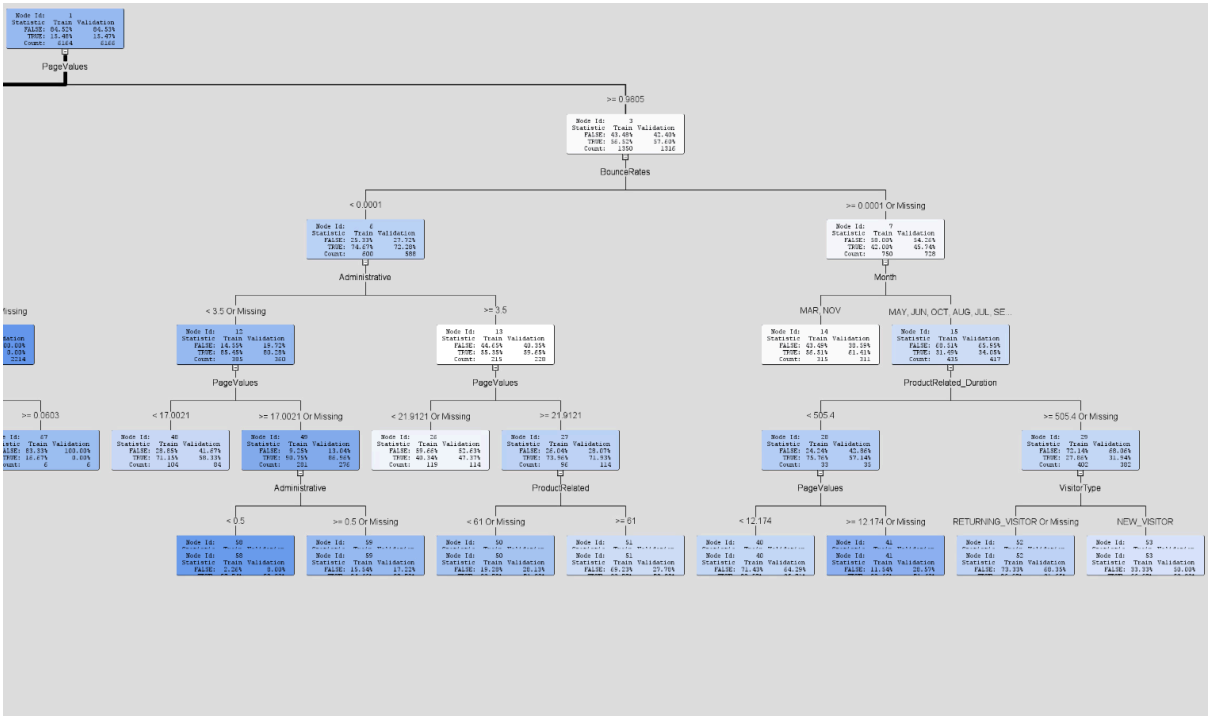
The first predictive model used was the Decision Tree, as it is particularly effective when applied to datasets without missing values.



a. Maximal Tree

The maximal tree, created using an **Interactive Train** with Average Square Error (ASE) as the assessment criteria, produced the following results:





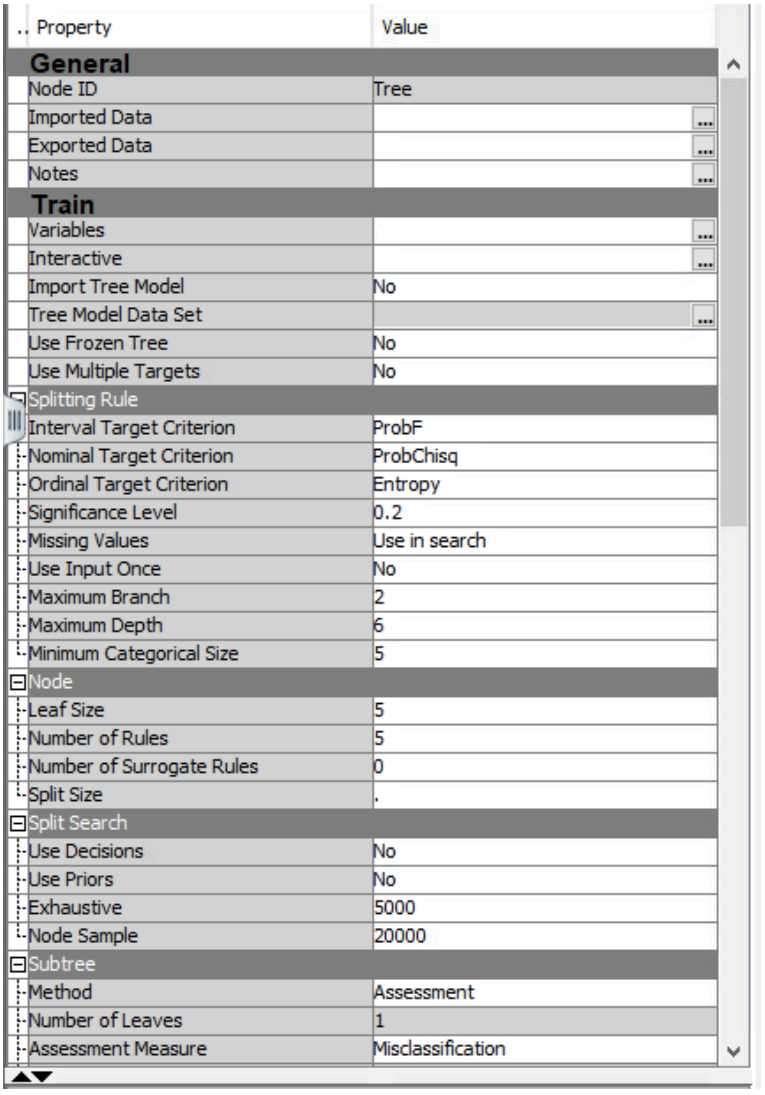
Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label ▲	Train	Validation	Test
Revenue		_ASE_	Average Squared Error	0.066302	0.073901	.
Revenue		_DIV_	Divisor for ASE	12328	12332	.
Revenue		_MAX_	Maximum Absolute Error	0.990025	0.990025	.
Revenue		_MISC_	Misclassification Rate	0.091499	0.100065	.
Revenue		_RASE_	Root Average Squared Error	0.257492	0.271848	.
Revenue		_NOBS_	Sum of Frequencies	6164	6166	.
Revenue		_SSE_	Sum of Squared Errors	817.374	911.3497	.
Revenue		_DFT_	Total Degrees of Freedom	6164	.	.

Based on the Interactive result after the training node. This means that the decision tree model can optimize to get a better ASE (0.073901)

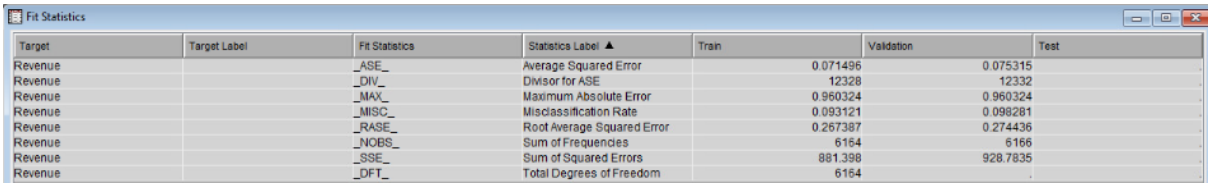
b. Misclassification Tree

A second 2-split decision tree was created using **Misclassification** as the assessment method, allowing for a comparison with the 2-split ASE tree.



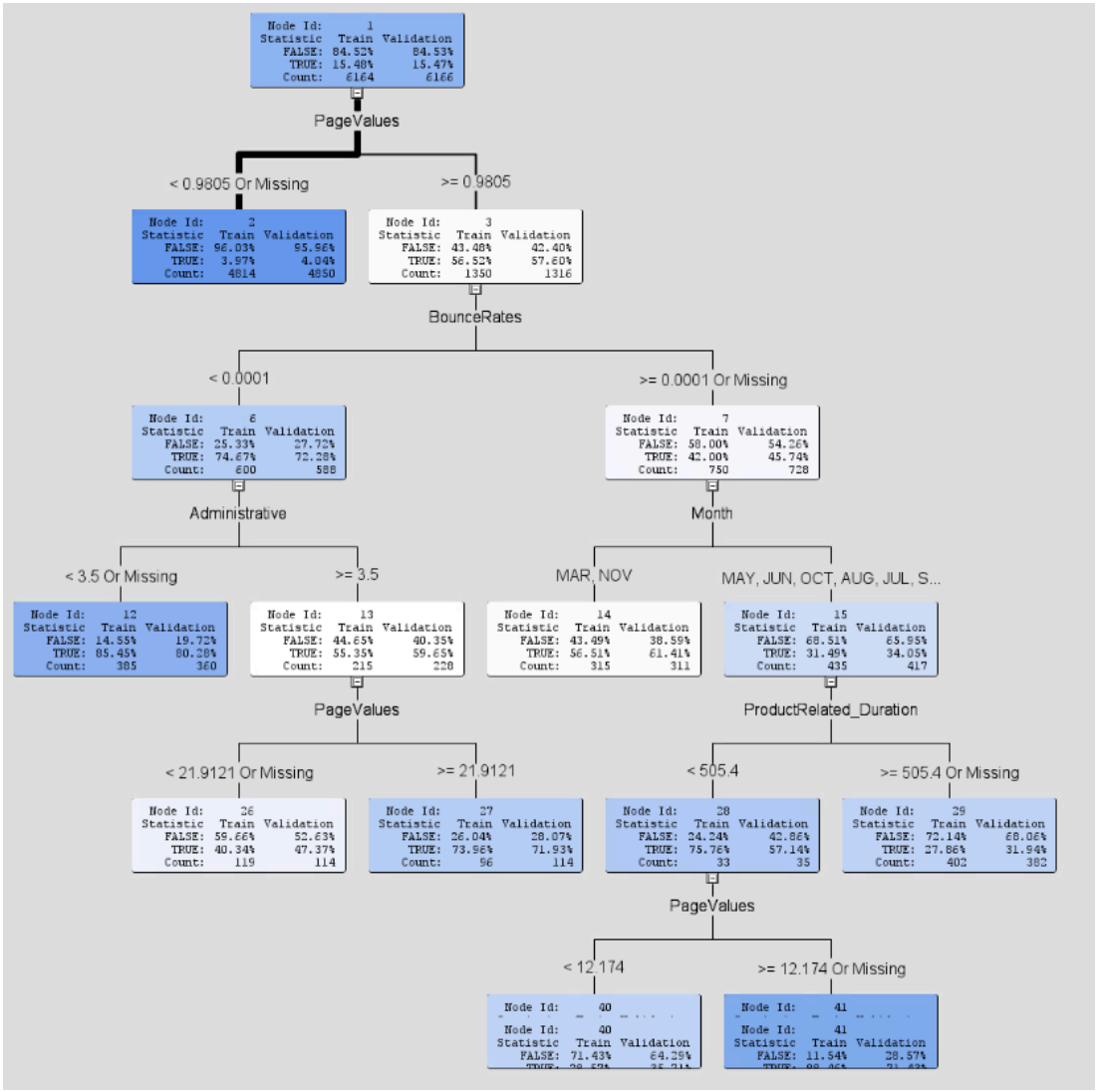
Property	Value
General	
Node ID	Tree
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Interactive	
Import Tree Model	No
Tree Model Data Set	
Use Frozen Tree	No
Use Multiple Targets	No
Splitting Rule	
Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	6
Minimum Categorical Size	5
Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.
Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000
Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Misclassification

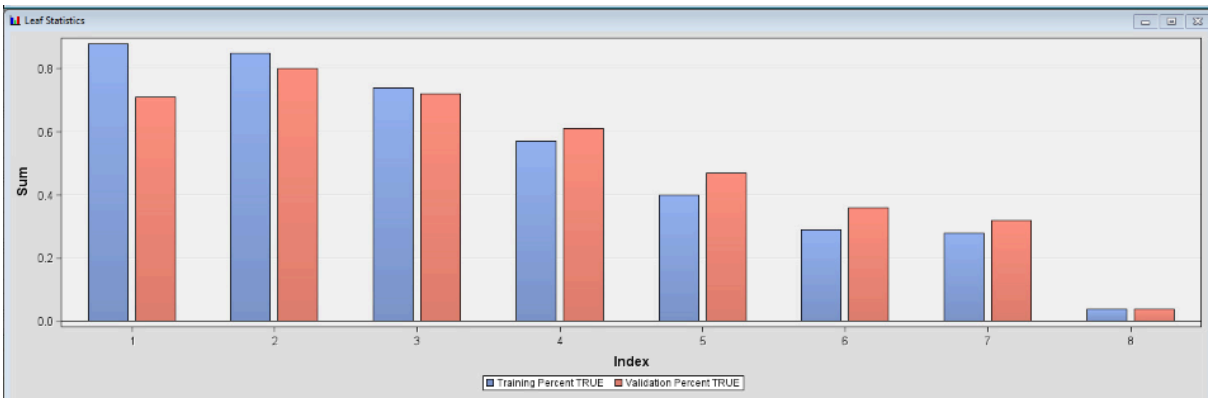
The result of the 2-split decision tree assessed by Misclassification was identical to the 2-split decision tree assessed by ASE, with a validation ASE of 0.075315.



Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label ▲	Train	Validation	Test
Revenue		_ASE_	Average Squared Error	0.071496	0.075315	
Revenue		_DIV_	Divisor for ASE	12328	12332	
Revenue		_MAX_	Maximum Absolute Error	0.960324	0.960324	
Revenue		_MISC_	Misclassification Rate	0.093121	0.098281	
Revenue		_RASE_	Root Average Squared Error	0.267387	0.274436	
Revenue		_NOBS_	Sum of Frequencies	6164	6166	
Revenue		_SSE_	Sum of Squared Errors	881.398	928.7835	
Revenue		_DFT_	Total Degrees of Freedom	6164	.	





The leaf nodes 1, 2, and 3 have predicted outcome percentages higher than the observed outcome percentages, indicating that the training model performed better than the validation model.

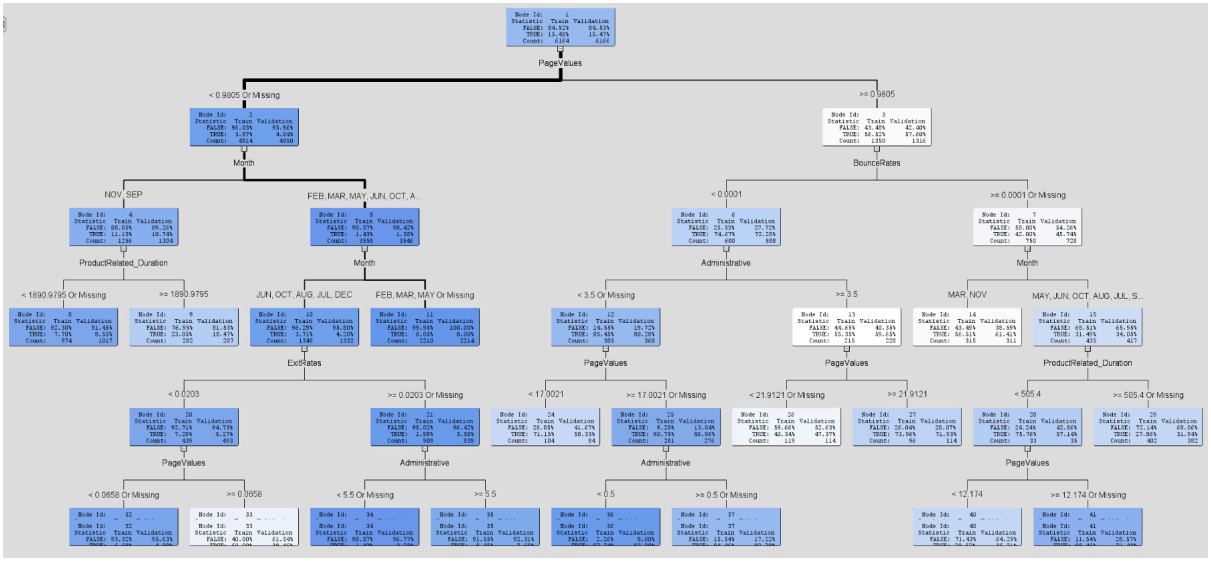
c. ASE Tree

The validation ASE of the 2-split ASE decision tree was 0.072632, which was lower than that of the maximal tree (0.073901). Therefore, the 2-split ASE decision tree was considered a better model than the maximal tree.

Property	Value
General	
Node ID	Tree2
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Interactive	
Import Tree Model	No
Tree Model Data Set	
Use Frozen Tree	No
Use Multiple Targets	No
Splitting Rule	
Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	6
Minimum Categorical Size	5
Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.
Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000
Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Average Square Error

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label ▲	Train	Validation	Test
Revenue		_ASE_	Average Squared Error	0.067968	0.072632	.
Revenue		_DIV_	Divisor for ASE	12328	12332	.
Revenue		_MAX_	Maximum Absolute Err...	0.999548	0.98568	.
Revenue		_MISC_	Misclassification Rate	0.092959	0.098767	.
Revenue		_RASE_	Root Average Squared...	0.260706	0.269503	.
Revenue		_NOBS_	Sum of Frequencies	6164	6166	.
Revenue		_SSE_	Sum of Squared Errors	837.9043	895.699	.
Revenue		_DFT_	Total Degrees of Free...	6164	.	.



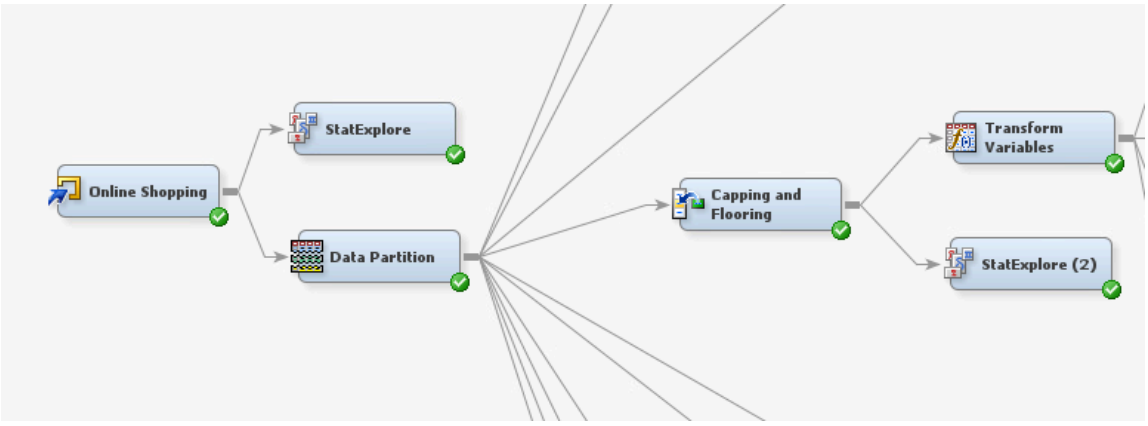
d. Decision Tree Summary

Based on the above results, it was concluded that the best decision tree model was the 2-split decision tree. Since both assessment methods, ASE and misclassification, gave the same result, we can confidently choose the 2-split decision tree with ASE as the preferred model.

Decision Tree	ASE
Maximal Tree	0.073901
Misclassification Tree	0.075315
ASE Tree	**0.072632**

6. Logistic Regression

a. Data Massaging



i. Capping and Flooring

Following the decision tree, regression analysis was employed as another predictive model to determine the best model for predicting heart disease. Before running the regression, it was essential to ensure that the data met the necessary criteria for regression analysis.

The first criterion considered was reducing skewness in the dataset. To address this, the data was processed using a replacement node with capping and flooring techniques and check the skewness again by using StatExplore. These methods adjusted extreme values to fall within acceptable bounds, ensuring that the dataset met the assumptions required for accurate regression modeling.

Interval Variables

Data Role	Target	Skewness ▼	Kurtosis	Target Level	Variable	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Role	Label	Scaled Mean Deviation	Maximum Deviation	Level Id
TRAIN	Revenue	5.972911	66.04752	TRUE	REP_Boun...	0	0	954	0	0.167361	0.004979	0.010584	INPUT	Replaceme...	-0.75611	0.756112	2
TRAIN	Revenue	5.428364	33.63275	FALSE	REP_Page...	0	0	5210	0	60.50615	1.836113	7.190143	INPUT	Replaceme...	-0.6375	3.481533	1
TRAIN	Revenue	3.956095	15.68519	FALSE	REP_Infor...	0	0	5210	0	409.0641	23.03832	74.27444	INPUT	Replaceme...	-0.11494	0.6277	1
TRAIN	Revenue	2.844516	18.67962	TRUE	REP_ExitR...	0.016667	0	954	0	0.188914	0.019528	0.015401	INPUT	Replaceme...	-0.54136	0.541362	2
TRAIN	Revenue	2.665473	6.304222	TRUE	REP_Infor...	0	0	954	0	409.0641	42.36929	97.36437	INPUT	Replaceme...	0.6277	0.6277	2
TRAIN	Revenue	2.588962	7.034683	FALSE	REP_Admi...	0	0	5210	0	591.493	66.45868	120.4931	INPUT	Replaceme...	-0.08002	0.436981	1
TRAIN	Revenue	2.555733	5.728571	FALSE	REP_Infor...	0	0	5210	0	4.263099	0.419071	0.99905	INPUT	Replaceme...	-0.09821	0.536328	1
TRAIN	Revenue	2.46442	4.973912	FALSE	REP_Boun...	0.004545	0	5210	0	0.167361	0.023243	0.044534	INPUT	Replaceme...	0.138451	0.756112	1
TRAIN	Revenue	2.1902	5.166262	FALSE	REP_Produ...	16	0	5210	0	159.8599	27.46681	32.0862	INPUT	Replaceme...	-0.07971	0.435321	1
TRAIN	Revenue	2.188161	5.026749	FALSE	REP_Produ...	512.5	0	5210	0	6369.416	1008.251	1309.103	INPUT	Replaceme...	-0.0949	0.518277	1
TRAIN	Revenue	1.889887	3.088353	TRUE	REP_Admi...	45.75	0	954	0	591.493	103.806	143.9297	INPUT	Replaceme...	0.436981	0.436981	2
TRAIN	Revenue	1.845028	2.577217	FALSE	REP_ExitR...	0.028929	0	5210	0	0.188914	0.046799	0.049019	INPUT	Replaceme...	0.099128	0.541362	1
TRAIN	Revenue	1.656637	1.568471	TRUE	REP_Infor...	0	0	954	0	4.263099	0.713945	1.229417	INPUT	Replaceme...	0.536328	0.536328	2
TRAIN	Revenue	1.630513	2.017863	FALSE	REP_Admi...	0	0	5210	0	12.15467	2.069419	2.995068	INPUT	Replaceme...	-0.07852	0.428819	1
TRAIN	Revenue	1.557157	1.655342	TRUE	REP_Produ...	28	0	954	0	159.8599	42.83838	40.4915	INPUT	Replaceme...	0.435321	0.435321	2
TRAIN	Revenue	1.512035	1.450928	TRUE	REP_Produ...	1058.338	0	954	0	6369.416	1691.313	1659.546	INPUT	Replaceme...	0.518277	0.518277	2
TRAIN	Revenue	1.056954	0.218127	TRUE	REP_Admi...	2	0	954	0	12.15467	3.20878	3.390505	INPUT	Replaceme...	0.428819	0.428819	2
TRAIN	Revenue	0.619059	-0.99461	TRUE	REP_Page...	16.77417	0	954	0	60.50615	22.69968	21.0787	INPUT	Replaceme...	3.481533	3.481533	2

ii. Transform Variable

To further reduce skewness in the dataset, a transform variable node was added following the capping and flooring node. This transformation adjusted the data distribution, ensuring it was more suitable for regression analysis and improving the overall model performance.The transform variable is used to address issues such as skewness, non-linearity, or non-normality in the dataset. By applying transformations, we can stabilize variance, make the data more normally distributed, and improve the fit of the regression model. This ensures that the data adheres to the assumptions required for accurate and reliable predictive modeling.

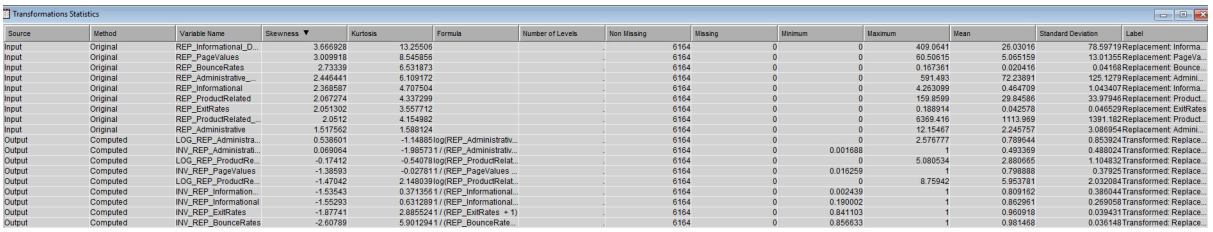
Variables - Trans2

(none) not Equal to

Columns: Label

Name	Method	Number of Bins	Role	Level
Administrative	**Default**	4	Rejected	Interval
Administrative_D	**Default**	4	Rejected	Interval
BounceRates	**Default**	4	Rejected	Interval
Browser	**Default**	4	Input	Nominal
ExitRates	**Default**	4	Rejected	Interval
Informational	**Default**	4	Rejected	Interval
Informational_D	**Default**	4	Rejected	Interval
Month	**Default**	4	Input	Nominal
OperatingSyster	**Default**	4	Input	Nominal
PageValues	**Default**	4	Rejected	Interval
ProductRelated	**Default**	4	Rejected	Interval
ProductRelated_	**Default**	4	Rejected	Interval
REP_Administrat	**Log**	4	Input	Interval
REP_Administrat	**Inverse**	4	Input	Interval
REP_BounceRati	**Inverse**	4	Input	Interval
REP_ExitRates	**Inverse**	4	Input	Interval
REP_Information	**Inverse**	4	Input	Interval
REP_Information	**Inverse**	4	Input	Interval
REP_PageValues	**Inverse**	4	Input	Interval
REP_ProductRel	**Log**	4	Input	Interval
REP_ProductRel	**Log**	4	Input	Interval
Region	**Default**	4	Input	Nominal
Revenue	**Default**	4	Target	Binary
SpecialDay	**Default**	4	Input	Ordinal
TrafficType	**Default**	4	Input	Nominal
VisitorType	**Default**	4	Input	Nominal
Weekend	**Default**	4	Input	Binary

The result that we got from transform variable shows the below results and the skewness of the dataset is reduced properly from -1.5 to 1.5 :



Transformations Statistics

Source	Method	Variable Name	Skewness ▼	Kurtosis	Formula	Number of Levels	Non Missing	Missing	Minimum	Maximum	Mean	Standard Deviation	Label
Input	Original	REP_Informational_D...	3.666928	13.25506		.	6164	0	0	409.0641	26.03016	78.59719	Replacement: Informa...
Input	Original	REP_PageValues	3.009918	8.545856		.	6164	0	0	60.50615	5.065159	13.01355	Replacement: PageVa...
Input	Original	REP_BounceRates	2.73339	6.531873		.	6164	0	0	0.167361	0.020416	0.04168	Replacement: Bounce...
Input	Original	REP_Administrative_...	2.446441	6.109172		.	6164	0	0	591.493	72.23891	125.1279	Replacement: Admini...
Input	Original	REP_Informational	2.368587	4.707504		.	6164	0	0	4.263099	0.464709	1.043407	Replacement: Informa...
Input	Original	REP_ProductRelated	2.067274	4.337299		.	6164	0	0	159.8599	29.94586	33.97946	Replacement: Product...
Input	Original	REP_ExitRates	2.051302	3.557712		.	6164	0	0	0.188914	0.042578	0.046529	Replacement: ExitRates
Input	Original	REP_ProductRelated_...	2.0512	4.154982		.	6164	0	0	6369.416	1113.969	1391.182	Replacement: Product...
Input	Original	REP_Administrative	1.517562	1.598124		.	6164	0	0	12.15467	2.245757	3.086954	Replacement: Admini...
Output	Computed	LOG_REP_Administra...	0.538601	-1.14885	log(REP_Administrativ...	.	6164	0	0	2.576777	0.789644	0.853924	Transformed: Replace...
Output	Computed	INV_REP_Administrati...	0.069064	-1.98573	1 / (REP_Administrativ...	.	6164	0	0.001688	1	0.493369	0.488024	Transformed: Replace...
Output	Computed	LOG_REP_ProductRe...	-0.17412	-0.54078	log(REP_ProductRelat...	.	6164	0	0	5.080534	2.880665	1.104832	Transformed: Replace...
Output	Computed	INV_REP_PageValues	-1.38593	-0.02781	1 / (REP_PageValues ...	.	6164	0	0.016259	1	0.798888	0.37925	Transformed: Replace...
Output	Computed	LOG_REP_ProductRe...	-1.47042	2.148039	log(REP_ProductRelat...	.	6164	0	0	8.75942	5.953781	2.032084	Transformed: Replace...
Output	Computed	INV_REP_Information...	-1.53543	0.371356	1 / (REP_Informational...	.	6164	0	0.002439	1	0.809162	0.386044	Transformed: Replace...
Output	Computed	INV_REP_Informational	-1.55293	0.631289	1 / (REP_Informational...	.	6164	0	0.190002	1	0.862951	0.269058	Transformed: Replace...
Output	Computed	INV_REP_ExitRates	-1.87741	2.885524	1 / (REP_ExitRates + 1)	.	6164	0	0.841103	1	0.960918	0.039431	Transformed: Replace...
Output	Computed	INV_REP_BounceRates	-2.60789	5.901294	1 / (REP_BounceRate...	.	6164	0	0.856633	1	0.981468	0.036148	Transformed: Replace...

a. Full regression

Full regression did not automatically select or eliminate variables in the process, unlike forward and backward regression methods. Instead, it required manually reviewing the p-values to identify significant variables and determine which ones should be included or excluded from the model. The data below shows the **p-value** under "Pr > ChiSq," indicating that only 6 variables are significant, with a p-value below **0.05**. These significant variables are listed as follows:

- “INV_REP_ExitRates”
- “INV_REP_PageValues”
- “LOG_REP_Administrative”
- “Month”
- “TrafficType”
- “VisitorType”

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Browser	11	8.9876	0.6230
INV_REP_Administrative_Duration	1	1.0456	0.3065
INV_REP_BounceRates	1	1.1589	0.2817
INV_REP_ExitRates	1	13.2571	0.0003
INV_REP_Informational	1	1.4569	0.2274
INV_REP_Informational_Duration	1	1.6493	0.1990
INV_REP_PageValues	1	991.9328	<.0001
LOG_REP_Administrative	1	15.0889	0.0001
LOG_REP_ProductRelated	1	2.3371	0.1263
LOG_REP_ProductRelated_Duration	1	0.2081	0.6483
Month	9	118.3151	<.0001
OperatingSystems	6	5.8068	0.4452
Region	8	7.2440	0.5106
SpecialDay	5	2.2190	0.8181
TrafficType	17	33.9249	0.0086
VisitorType	2	20.6382	<.0001
Weekend	1	0.1743	0.6763

To identify the strength of the relationship between different factors and the likelihood of a purchase event occurring, the **odds ratios** were observed. In this case, the odds ratios indicate

how the unit change in each particular variable would increase or decrease the percentage chance of making a purchase. The following shows the point estimates under the **Odds Ratio Estimates** table, with the following significant results:

- People using Browser 1 (compared to Browser 13) are 94.1% less likely to make a purchase.
- People using Browser 3 (compared to Browser 13) are 72% less likely to make a purchase.
- Every unit increase in Bounce Rates (INV_REP_BounceRates) would increase the chance of making a purchase by 721.28%.
- A unit increase in Exit Rates (INV_REP_ExitRates) would increase the chance of making a purchase by 999.00%.
- A unit increase in Page Values (INV_REP_PageValues) would increase the chance of making a purchase by 16%.
- People browsing in the Month of August (compared to September) are 36.1% more likely to make a purchase.
- Returning Visitors are 1.896 times more likely to make a purchase than new visitors.

Odds Ratio Estimates

Effect		Point Estimate
Browser	1 vs 13	0.051
Browser	2 vs 13	0.048
Browser	3 vs 13	0.043
Browser	4 vs 13	0.043
Browser	5 vs 13	0.068
Browser	6 vs 13	0.025
Browser	7 vs 13	0.022
Browser	8 vs 13	0.083
Browser	10 vs 13	0.071
Browser	11 vs 13	999.000
Browser	12 vs 13	0.117
INV_REP_Administrative_Duration		0.816
INV_REP_BounceRates		721.279
INV_REP_ExitRates		999.000
INV_REP_Informational		0.532
INV_REP_Informational_Duration		1.595
INV_REP_PageValues		0.016
LOG_REP_Administrative		0.652
LOG_REP_ProductRelated		0.844
LOG_REP_ProductRelated_Duration		1.036
Month	Aug vs Sep	0.881
Month	Dec vs Sep	0.469
Month	Feb vs Sep	0.122
Month	Jul vs Sep	0.562
Month	Jun vs Sep	0.629
Month	Mar vs Sep	0.539
Month	May vs Sep	0.455
Month	Nov vs Sep	1.723
Month	Oct vs Sep	0.712
OperatingSystems	1 vs 8	0.646
OperatingSystems	2 vs 8	0.683
OperatingSystems	3 vs 8	0.499
OperatingSystems	4 vs 8	0.505
OperatingSystems	5 vs 8	<0.001
OperatingSystems	6 vs 8	<0.001
OperatingSystems	7 vs 8	.
Region	1 vs 9	1.546
Region	2 vs 9	1.281
Region	3 vs 9	1.256
Region	4 vs 9	1.478
Region	5 vs 9	1.044
Region	6 vs 9	1.518
Region	7 vs 9	1.684
Region	8 vs 9	1.634
SpecialDay	0 vs 1	0.755
SpecialDay	0.2 vs 1	0.414
SpecialDay	0.4 vs 1	0.620
SpecialDay	0.6 vs 1	0.803
SpecialDay	0.8 vs 1	0.520
TrafficType	1 vs 20	0.392
TrafficType	2 vs 20	0.477
TrafficType	3 vs 20	0.377
TrafficType	4 vs 20	0.508
TrafficType	5 vs 20	0.535
TrafficType	6 vs 20	0.913
TrafficType	7 vs 20	0.405
TrafficType	8 vs 20	1.111
TrafficType	9 vs 20	<0.001
TrafficType	10 vs 20	0.806
TrafficType	11 vs 20	0.688
TrafficType	12 vs 20	0.004
TrafficType	13 vs 20	0.243
TrafficType	14 vs 20	<0.001
TrafficType	15 vs 20	<0.001
TrafficType	18 vs 20	0.003
TrafficType	19 vs 20	0.005
VisitorType	New_Visitor vs Returning_Visitor	1.896
VisitorType	Other vs Returning_Visitor	0.051
Weekend	FALSE vs TRUE	1.050

b. Forward Regression

In the **forward regression**, the significant variables were selected based on the **p-value** shown under "Pr > ChiSq". After running the model, **7 variables** were included as significant factors. These variables were identified because their p-values were below the threshold of 0.05, indicating that they have a statistically significant relationship with the target variable (purchase intention)

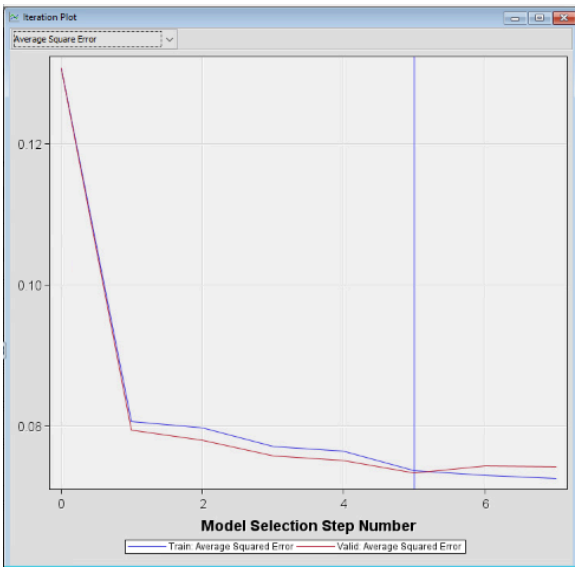
Summary of Forward Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Pr > ChiSq	Validation Error Rate
1	INV_REP_PageValues	1	1	2330.0114	<.0001	3347.9
2	Month	9	2	149.7532	<.0001	3180.9
3	VisitorType	2	3	91.7762	<.0001	3116.6
4	INV_REP_ExitRates	1	4	33.4888	<.0001	3082.3
5	LOG_REP_Administrative	1	5	45.2869	<.0001	3060.9
6	TrafficType	17	6	41.4559	0.0008	3172.5
7	LOG_REP_ProductRelated	1	7	4.6073	0.0318	3173.8

The selected model, based on the error rate for the validation data, is the model trained in Step 5. It consists of the following effects:

Intercept INV_REP_ExitRates INV_REP_PageValues LOG_REP_Administrative Month VisitorType

However, the Iteration Plot with the Average Squared Error (ASE) indicated that the model selection step number is equal to 5, meaning that five variables are sufficient to run this forward regression model. This suggests that including more than five variables does not significantly improve the model's performance, and five variables are optimal for prediction accuracy.



The following shows the point estimate under the Odds Ratio Estimates table and with the following significant result:

- Page Values (INV_REP_PageValues): A unit increase in Page Values is associated with a 1.7% increase in the likelihood of making a purchase.
- Months Comparison:
    - Visitors in August are 98.3% less likely to make a purchase compared to those in September.
    - Visitors in December have a 3.6% lower likelihood of making a purchase compared to those in September.
    - Visitors in February are 39.9% less likely to make a purchase compared to those in September.
    - Visitors in July are 88.3% less likely to make a purchase compared to those in September.

- Visitors in June are 44.7% less likely to make a purchase compared to those in September.
- Visitors in March are 39.5% less likely to make a purchase compared to those in September.
- Visitors in May are 28.4% less likely to make a purchase compared to those in September.
- Visitors in November are 94.2% more likely to make a purchase compared to those in September.
- Visitors in October are 19.7% less likely to make a purchase compared to those in September.

- Visitor Type:
  - New Visitors are 3.225 times more likely to make a purchase compared to Returning Visitors.
  - Other Visitors are 1.862 times more likely to make a purchase compared to Returning Visitors.

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
INV_REP_ExitRates	1	43.0658	<.0001
INV_REP_PageValues	1	1052.7044	<.0001
LOG_REP_Administrative	1	30.5484	<.0001
LOG_REP_ProductRelated	1	4.6005	0.0320
Month	9	126.9111	<.0001
TrafficType	16	158.7876	<.0001
VisitorType	2	21.7900	<.0001

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept		1	-22.1866	16.8464	1.73	0.1878		0.000
INV_REP_ExitRates		1	21.3423	3.2522	43.07	<.0001	0.4640	999.000
INV_REP_PageValues		1	-4.1007	0.1264	1052.70	<.0001	-0.8574	0.017
LOG_REP_Administrative		1	-0.3535	0.0640	30.55	<.0001	-0.1664	0.702
LOG_REP_ProductRelated		1	-0.1423	0.0664	4.60	0.0320	-0.0867	0.867
Month	Aug	1	0.4375	0.2587	2.86	0.0908		1.549
Month	Dec	1	-0.2014	0.1753	1.32	0.2506		0.818
Month	Feb	1	-1.6135	0.9522	2.87	0.0902		0.199
Month	Jul	1	-0.0398	0.2701	0.02	0.8829		0.961
Month	Jun	1	0.0725	0.3549	0.04	0.8381		1.075
Month	Mar	1	-0.0769	0.1793	0.18	0.6578		0.926
Month	May	1	-0.3014	0.1564	3.72	0.0539		0.740
Month	Nov	1	1.0494	0.1502	48.78	<.0001		2.856
Month	Oct	1	0.1720	0.2240	0.59	0.4424		1.188
TrafficType	1	1	2.6043	16.5556	0.02	0.8750		13.521
TrafficType	2	1	2.7816	16.5553	0.03	0.8666		16.145
TrafficType	3	1	2.5552	16.5557	0.02	0.8773		12.873
TrafficType	4	1	2.8183	16.5559	0.03	0.8648		16.749
TrafficType	5	1	2.7906	16.5579	0.03	0.8662		16.292
TrafficType	6	1	2.3699	16.5572	0.02	0.8862		10.696
TrafficType	7	1	2.5847	16.5727	0.02	0.8761		13.259
TrafficType	8	1	3.5568	16.5567	0.05	0.8299		35.052
TrafficType	9	1	-6.4456	81.0636	0.01	0.9366		0.002
TrafficType	10	1	3.2855	16.5566	0.04	0.8427		26.723
TrafficType	11	1	3.1596	16.5578	0.04	0.8487		23.562
TrafficType	12	1	-5.2793	.	.	.	.	0.005
TrafficType	13	1	2.0372	16.5570	0.02	0.9021		7.669
TrafficType	14	1	-6.6752	131.5	0.00	0.9595		0.001
TrafficType	15	1	-5.6910	77.9674	0.01	0.9418		0.003
TrafficType	18	1	-5.0440	147.7	0.00	0.9728		0.006
TrafficType	19	1	-5.0418	140.0	0.00	0.9713		0.006
VisitorType	New_Visitor	1	0.5361	0.2537	4.50	0.0339		1.709
VisitorType	Other	1	-0.4010	0.4774	0.71	0.4009		0.670

Odds Ratio Estimates

Effect		Point Estimate
INV_REP_PageValues		0.017
Month	Aug vs Sep	0.964
Month	Dec vs Sep	0.601
Month	Feb vs Sep	0.117
Month	Jul vs Sep	0.553
Month	Jun vs Sep	0.605
Month	Mar vs Sep	0.716
Month	May vs Sep	0.479
Month	Nov vs Sep	1.942
Month	Oct vs Sep	0.803
VisitorType	New_Visitor vs Returning_Visitor	3.225
VisitorType	Other vs Returning_Visitor	1.862

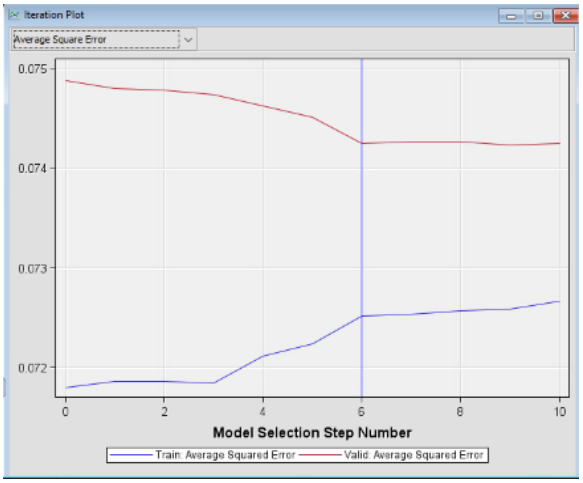
c. Backward Regression

With the backward regression run on this data, 10 variables were removed from the process. As shown under "Pr > ChiSq," all the eliminated variables had high p-values, indicating that these variables were not statistically significant and did not contribute meaningfully to predicting the likelihood of a purchase.

Summary of Backward Elimination

Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq	Validation Error Rate
1	SpecialDay	5	16	2.3190	0.8181	3213.3
2	Weekend	1	15	0.0403	0.8409	3212.9
3	LOG_REP_ProductRelated_Duration	1	14	0.2266	0.6341	3205.5
4	Browser	11	13	8.9901	0.6228	3211.2
5	Region	8	12	7.1055	0.5253	3240.6
6	OperatingSystems	7	11	7.5497	0.3740	3172.8
7	INV_REP_Informational	1	10	0.9521	0.3292	3175.3
8	INV_REP_Informational_Duration	1	9	0.1855	0.6667	3174.8
9	INV_REP_Administrative_Duration	1	8	1.3394	0.2471	3174.3
10	INV_REP_BounceRates	1	7	2.3511	0.1252	3173.8

The ASE Iteration Plot indicated that 6 variables were sufficient to run this regression.



The following shows the point estimate under the **Odds Ratio Estimates** table and with the following significant result:

- New Visitors are 4.115 times more likely to make a purchase compared to Returning Visitors.
- Other Visitors are 0.112 times as likely (or 88.8% less likely) to make a purchase compared to Returning Visitors.
- A unit increase in Page Values increases the odds of making a purchase by **599.056 times**.
- A unit increase in Exit Rates decreases the odds of making a purchase by approximately **94.8%** (Exp(Estimate) = 0.052).
- Visitors in **August** are **98.3% less likely** to make a purchase compared to September.
- Visitors in **December** are **1.2% more likely** to make a purchase compared to September.
- Visitors in **February** are **39.3% less likely** to make a purchase compared to September.
- Visitors in **July** are **85.3% less likely** to make a purchase compared to September.
- Visitors in **June** are **54.7% less likely** to make a purchase compared to September.
- Visitors in **March** are **36.4% less likely** to make a purchase compared to September.
- Visitors in **May** are **12.5% less likely** to make a purchase compared to September.
- Visitors in **November** are **2.967 times more likely** to make a purchase compared to September.
- Visitors in **October** are **13.8% less likely** to make a purchase compared to September.
- Traffic Type **1** is **94.8% less likely** to result in a purchase compared to Traffic Type 20.
- Traffic Type **2** is **4.84 times more likely** to result in a purchase compared to Traffic Type 20.
- Traffic Type **3** is **99.5% less likely** to result in a purchase compared to Traffic Type 20.
- A unit increase in the "Special Day" variable reduces the likelihood of a purchase by **90.4%** (Exp(Estimate) = 0.096).

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Browser	11	8.9876	0.6230
INV_REP_Administrative_Duration	1	1.0456	0.3065
INV_REP_BounceRates	1	1.1589	0.2817
INV_REP_ExitRates	1	13.2571	0.0003
INV_REP_Informational	1	1.4569	0.2274
INV_REP_Informational_Duration	1	1.6493	0.1990
INV_REP_PageValues	1	991.9328	<.0001
LOG_REP_Administrative	1	15.0889	0.0001
LOG_REP_ProductRelated	1	2.3371	0.1263
LOG_REP_ProductRelated_Duration	1	0.2081	0.6483
Month	9	118.3151	<.0001
OperatingSystems	6	5.8068	0.4452
Region	8	7.2440	0.5106
SpecialDay	5	2.2190	0.8181
TrafficType	17	33.9249	0.0086
VisitorType	2	20.6382	<.0001
Weekend	1	0.1743	0.6763

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept		1	-27.3463	19.8741	1.89	0.1688		0.000
Browser	1	1	-1.4830	25.2309	0.00	0.9532		0.227
Browser	2	1	-1.5488	25.2510	0.00	0.9511		0.213
Browser	3	1	-1.6519	25.2375	0.00	0.9479		0.192
Browser	4	1	-1.6586	25.2516	0.00	0.9476		0.190
Browser	5	1	-1.2008	25.2522	0.00	0.9621		0.301
Browser	6	1	-2.1773	25.2544	0.01	0.9313		0.113
Browser	7	1	-2.3227	25.2642	0.01	0.9267		0.098
Browser	8	1	-0.9908	25.2555	0.00	0.9687		0.371
Browser	10	1	-1.1480	25.2533	0.00	0.9637		0.317
Browser	11	1	13.3431	277.7	0.00	0.9617		999.000
Browser	12	1	-0.6546	25.2838	0.00	0.9793		0.520
INV_REP_Administrative_Duration		1	-0.2037	0.1992	1.05	0.3065	-0.0548	0.816
INV_REP_BounceRates		1	6.5810	6.1132	1.16	0.2817	0.1312	721.279
INV_REP_ExitRates		1	17.1108	4.6994	13.26	0.0003	0.3720	999.000
INV_REP_Informational		1	-0.6317	0.5234	1.46	0.2274	-0.0937	0.532
INV_REP_Informational_Duration		1	0.4667	0.3634	1.65	0.1990	0.0993	1.595
INV_REP_PageValues		1	-4.1264	0.1310	991.93	<.0001	-0.8628	0.016
LOG_REP_Administrative		1	-0.4276	0.1101	15.09	0.0001	-0.2013	0.652
LOG_REP_ProductRelated		1	-0.1697	0.1110	2.34	0.1263	-0.1033	0.844
LOG_REP_ProductRelated_Duration		1	0.0357	0.0783	0.21	0.6483	0.0400	1.036
Month	Aug	1	0.3958	0.2636	2.25	0.1333		1.486
Month	Dec	1	-0.2342	0.1793	1.71	0.1914		0.791
Month	Feb	1	-1.5783	0.9716	2.64	0.1043		0.206
Month	Jul	1	-0.0548	0.2743	0.04	0.8426		0.947
Month	Jun	1	0.0593	0.3572	0.03	0.8681		1.061
Month	Mar	1	-0.0947	0.1830	0.27	0.6048		0.910
Month	May	1	-0.2648	0.1620	2.67	0.1022		0.767
Month	Nov	1	1.0666	0.1544	47.70	<.0001		2.905
Month	Oct	1	0.1824	0.2255	0.65	0.4187		1.200
OperatingSystems	1	1	4.3399	42.3438	0.01	0.9184		76.696
OperatingSystems	2	1	4.3963	42.3444	0.01	0.9173		81.149
OperatingSystems	3	1	4.0830	42.3444	0.01	0.9232		59.323
OperatingSystems	4	1	4.0937	42.3445	0.01	0.9230		59.964
OperatingSystems	5	1	-17.4382	255.6	0.00	0.9456		0.000
OperatingSystems	6	1	-4.2521	50.9405	0.01	0.9335		0.014
OperatingSystems	7	0	.	.	.	.		.
Region	1	1	0.1276	0.0947	1.82	0.1778		1.136
Region	2	1	-0.0605	0.1585	0.15	0.7028		0.941
Region	3	1	-0.0805	0.1179	0.47	0.4947		0.923
Region	4	1	0.0825	0.1512	0.30	0.5851		1.086
Region	5	1	-0.2652	0.2562	1.07	0.3006		0.767
Region	6	1	0.1088	0.1847	0.35	0.5556		1.115
Region	7	1	0.2127	0.1820	1.37	0.2426		1.237
Region	8	1	0.1828	0.2443	0.56	0.4544		1.201
SpecialDay	0	1	0.1382	0.2059	0.45	0.5021		1.148
SpecialDay	0.2	1	-0.4626	0.4669	0.98	0.3218		0.630
SpecialDay	0.4	1	-0.0595	0.4140	0.02	0.8856		0.942
SpecialDay	0.6	1	0.1993	0.3808	0.27	0.6007		1.221
SpecialDay	0.8	1	-0.2346	0.4438	0.28	0.5970		0.791
TrafficType	1	1	1.6358	5.9641	0.08	0.7839		5.133
TrafficType	2	1	1.8308	5.9634	0.09	0.7588		6.239
TrafficType	3	1	1.5965	5.9645	0.07	0.7889		4.936
TrafficType	4	1	1.8950	5.9652	0.10	0.7507		6.652
TrafficType	5	1	1.9460	5.9708	0.11	0.7445		7.001
TrafficType	6	1	1.4106	5.9690	0.06	0.8132		4.098
TrafficType	7	1	1.6683	6.0137	0.08	0.7815		5.303
TrafficType	8	1	2.6773	5.9672	0.20	0.6537		14.545
TrafficType	9	1	-4.9257	22.6631	0.05	0.8279		0.007
TrafficType	10	1	2.3557	5.9669	0.16	0.6930		10.546
TrafficType	11	1	2.1984	5.9715	0.14	0.7128		9.011
TrafficType	12	1	-2.9647	77.7674	0.00	0.9696		0.052
TrafficType	13	1	1.1589	5.9684	0.04	0.8460		3.186
TrafficType	14	1	-4.4644	30.0492	0.02	0.8819		0.012
TrafficType	15	1	-4.5370	25.1163	0.03	0.8567		0.011
TrafficType	16	1	-3.2449	38.1105	0.01	0.9321		0.039
TrafficType	19	1	-2.8084	30.0786	0.01	0.9256		0.060
VisitorType	New_Visitor	1	1.4156	0.7752	3.33	0.0678		4.119
VisitorType	Other	1	-2.1914	1.5427	2.02	0.1555		0.112
Weekend	FALSE	1	0.0243	0.0583	0.17	0.6763		1.025

d. Stepwise Regression

Stepwise regression has all the same results as forwarding regression.

Summary of Stepwise Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq	Validation Error Rate
1	INV_REP_PageValues	1	1	2330.0114		<.0001	3347.9
2	Month	9	2	149.7532		<.0001	3180.9
3	VisitorType	2	3	91.7762		<.0001	3116.6
4	INV_REP_ExitRates	1	4	33.4888		<.0001	3082.3
5	LOG_REP_Administrative	1	5	45.2869		<.0001	3060.9
6	TrafficType	17	6	41.4559		0.0008	3172.5
7	LOG_REP_ProductRelated	1	7	4.6073		0.0318	3173.8

Type 3 Analysis of Effects

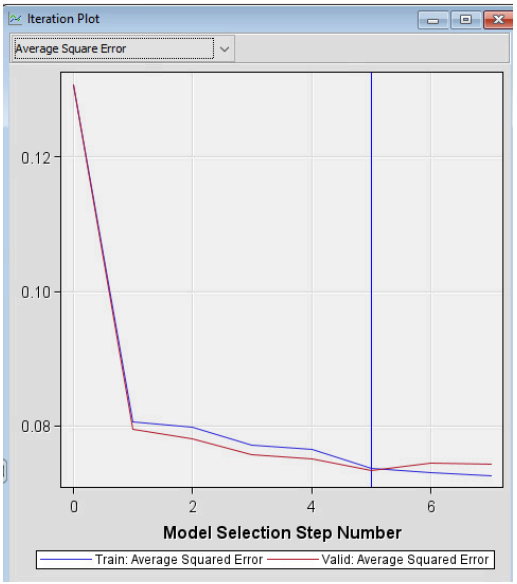
Effect	DF	Wald Chi-Square	Pr > ChiSq
Browser	11	8.9876	0.6230
INV_REP_Administrative_Duration	1	1.0456	0.3065
INV_REP_BounceRates	1	1.1589	0.2817
INV_REP_ExitRates	1	13.2571	0.0003
INV_REP_Informational	1	1.4569	0.2274
INV_REP_Informational_Duration	1	1.6493	0.1990
INV_REP_PageValues	1	991.9328	<.0001
LOG_REP_Administrative	1	15.0889	0.0001
LOG_REP_ProductRelated	1	2.3371	0.1263
LOG_REP_ProductRelated_Duration	1	0.2081	0.6483
Month	9	118.3151	<.0001
OperatingSystems	6	5.8068	0.4452
Region	8	7.2440	0.5106
SpecialDay	5	2.2190	0.8181
TrafficType	17	33.9249	0.0086
VisitorType	2	20.6382	<.0001
Weekend	1	0.1743	0.6763

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept		1	-27.3463	15.0741	1.69	0.1598		0.000
Browser	1	1	-1.4830	25.2309	0.00	0.9532		0.227
Browser	2	1	-1.5488	25.2310	0.00	0.9511		0.213
Browser	3	1	-1.6119	25.2375	0.00	0.9479		0.199
Browser	4	1	-1.6386	25.2316	0.00	0.9476		0.195
Browser	5	1	-1.2098	25.2312	0.00	0.9621		0.301
Browser	6	1	-2.1773	25.2144	0.01	0.9313		0.113
Browser	7	1	-2.3227	25.2642	0.01	0.9267		0.098
Browser	8	1	-0.9908	25.2855	0.00	0.9687		0.371
Browser	10	1	-1.1490	25.2033	0.00	0.9637		0.317
Browser	11	1	13.3431	277.7	0.00	0.9617		999.000
Browser	13	1	-0.6546	25.2838	0.00	0.9793		0.520
INV_REP_Administrative_Duration		1	-0.2037	0.1992	1.05	0.3065	-0.3548	0.816
INV_REP_BounceRates		1	6.5810	6.1132	1.16	0.2817	0.1312	721.279
INV_REP_ExitRates		1	17.1108	4.6994	13.26	0.0003	0.3720	999.000
INV_REP_Informational		1	-0.6317	0.5234	1.46	0.2274	-0.0937	0.532
INV_REP_Informational_Duration		1	0.4667	0.3634	1.65	0.1990	0.0993	1.595
INV_REP_PageValues		1	-4.1264	0.1310	991.93	<.0001	-0.9628	0.016
LOG_REP_Administrative		1	-0.4076	0.1131	15.09	0.0001	-0.2013	0.652
LOG_REP_ProductRelated		1	-0.1697	0.1110	2.34	0.1263	-0.1033	0.844
LOG_REP_ProductRelated_Duration		1	0.0357	0.0783	0.21	0.6483	0.0460	1.036
Month	Aug	1	0.5958	0.2636	2.25	0.1335		1.815
Month	Dec	1	-0.2342	0.1793	1.71	0.1914		0.791
Month	Feb	1	-1.5783	0.9716	2.64	0.1043		0.206
Month	Jul	1	-0.5545	0.2743	0.04	0.8426		0.947
Month	Jun	1	0.0393	0.3572	0.03	0.8681		1.061
Month	Mar	1	-0.0947	0.1830	0.27	0.6048		0.910
Month	May	1	-0.2648	0.1620	2.67	0.1022		0.767
Month	Nov	1	1.0666	0.1544	47.79	<.0001		2.905
Month	Oct	1	0.1824	0.2255	0.65	0.4187		1.200
OperatingSystems	1	1	4.3993	42.3438	0.01	0.9184		76.696
OperatingSystems	2	1	4.3963	42.3444	0.01	0.9173		81.149
OperatingSystems	3	1	4.0830	42.3444	0.01	0.9232		59.323
OperatingSystems	4	1	4.0937	42.3445	0.01	0.9230		59.964
OperatingSystems	5	1	-17.4092	251.6	0.00	0.9456		0.000
OperatingSystems	6	1	-4.2521	50.9405	0.01	0.9335		0.014
OperatingSystems	7	0	0	.	.	.	.	.
Region	1	1	0.1276	0.0947	1.82	0.1778		1.136
Region	2	1	-0.0605	0.1585	0.15	0.7028		0.941
Region	3	1	-0.0805	0.1179	0.47	0.4947		0.923
Region	4	1	0.0025	0.1512	0.00	0.9851		1.003
Region	5	1	-0.2652	0.2562	1.07	0.3006		0.767
Region	6	1	0.1088	0.1847	0.35	0.5556		1.115
Region	7	1	0.2127	0.1820	1.37	0.2426		1.237
Region	8	1	0.1828	0.2443	0.56	0.4544		1.201
SpecialDay	0	1	0.1382	0.2059	0.45	0.5021		1.148
SpecialDay	0.2	1	-0.4626	0.4659	0.99	0.3218		0.630
SpecialDay	0.4	1	-0.0595	0.4140	0.02	0.8856		0.942
SpecialDay	0.6	1	0.1993	0.3808	0.27	0.6007		1.221
SpecialDay	0.8	1	-0.2346	0.4439	0.28	0.5970		0.791
TrafficType	1	1	1.6358	5.9641	0.08	0.7839		5.133
TrafficType	2	1	1.8308	5.9634	0.09	0.7588		6.239
TrafficType	3	1	1.5965	5.9645	0.07	0.7809		4.936
TrafficType	4	1	1.8950	5.9652	0.10	0.7507		6.652
TrafficType	5	1	1.9460	5.9708	0.11	0.7445		7.001
TrafficType	6	1	1.4106	5.9690	0.06	0.8132		4.099
TrafficType	7	1	1.6683	6.0137	0.08	0.7815		5.303
TrafficType	8	1	2.6775	5.9672	0.20	0.6537		14.545
TrafficType	9	1	-4.9257	22.6631	0.05	0.8279		0.007
TrafficType	10	1	2.3357	5.9689	0.16	0.6930		10.346
TrafficType	11	1	2.1984	5.9715	0.14	0.7128		9.011
TrafficType	12	1	-2.9647	77.7674	0.00	0.9696		0.052
TrafficType	13	1	1.1589	5.9684	0.04	0.8460		3.186
TrafficType	14	1	-4.4644	30.0492	0.02	0.8819		0.012
TrafficType	15	1	-4.5370	25.1193	0.03	0.8567		0.011
TrafficType	16	1	-3.2449	38.1105	0.01	0.9321		0.039
TrafficType	19	1	-2.8084	30.0786	0.01	0.9256		0.060
VisitorType	New_Visitor	1	1.4156	0.7752	3.33	0.0678		4.119
VisitorType	Other	1	-2.1914	1.5427	2.02	0.1555		0.112
Weekend	FALSE	1	0.0243	0.0583	0.17	0.6763		1.025

Odds Ratio Estimates

Effect		Point Estimate
INV_REP_ExitRates		999.000
INV_REP_PageValues		0.018
LOG_REP_Administrative		0.669
Month	Aug vs Sep	0.930
Month	Dec vs Sep	0.546
Month	Feb vs Sep	0.125
Month	Jul vs Sep	0.557
Month	Jun vs Sep	0.647
Month	Mar vs Sep	0.627
Month	May vs Sep	0.442
Month	Nov vs Sep	1.835
Month	Oct vs Sep	0.778
VisitorType	New_Visitor vs Returning_Visitor	2.463
VisitorType	Other vs Returning_Visitor	1.694



The following shows the point estimate under the **Odds Ratio Estimates** table and with the following significant result:

- A unit increase in Exit Rates increases the likelihood of making a purchase by 999 times.
- A unit increase in Page Values decreases the likelihood of making a purchase by **98.2%** (Odds Ratio = 0.018).
- A unit increase in administrative-related activities decreases the likelihood of making a purchase by **33.1%** (Odds Ratio = 0.669).
- Visitors in **August** are **7% less likely** to make a purchase compared to September (Odds Ratio = 0.930).
- Visitors in **December** are **45.4% less likely** to make a purchase compared to September (Odds Ratio = 0.546).
- Visitors in **February** are **87.5% less likely** to make a purchase compared to September (Odds Ratio = 0.125).
- Visitors in **July** are **44.3% less likely** to make a purchase compared to September (Odds Ratio = 0.557).
- Visitors in **June** are **35.3% less likely** to make a purchase compared to September (Odds Ratio = 0.647).
- Visitors in **March** are **37.3% less likely** to make a purchase compared to September (Odds Ratio = 0.627).
- Visitors in **May** are **55.8% less likely** to make a purchase compared to September (Odds Ratio = 0.442).
- Visitors in **November** are **1.835 times more likely** to make a purchase compared to September.
- Visitors in **October** are **22.2% less likely** to make a purchase compared to September (Odds Ratio = 0.778).
- **New Visitors** are **2.463 times more likely** to make a purchase compared to Returning Visitors.

- **Other Visitors** are **1.694 times more likely** to make a purchase compared to Returning Visitors.

e. Regression Summary

After running four different types of regression models (forward, backward, stepwise, and full) following the replacement node, the best regression model was selected by comparing the Average Squared Error (ASE) for each model. The regression model with the lowest ASE was considered the best for predicting purchasing intention. In addition, neural network models were also run by connecting them to the optimized regression model, allowing for further refinement and evaluation of the predictive power of the chosen regression model.

The property settings of each regression model are mostly similar, with all using logistic regression under the regression type. However, they differ in their selection models and criteria, with forward, backward, and stepwise regressions using the validation error criterion for model selection. Overall, it is not surprising that both forward regression and stepwise regression could also be considered the "best" models, as both resulted in the same Average Squared Error (ASE) of 0.073437.

Regression Model	Selection Criteria	ASE (Validation)
Full Regression	None	0.074878
Forward Regression	**Validation error**	**0.073437**
Backward Regression	Validation error	0.074248
Stepwise Regression	**Validation error**	**0.073437**

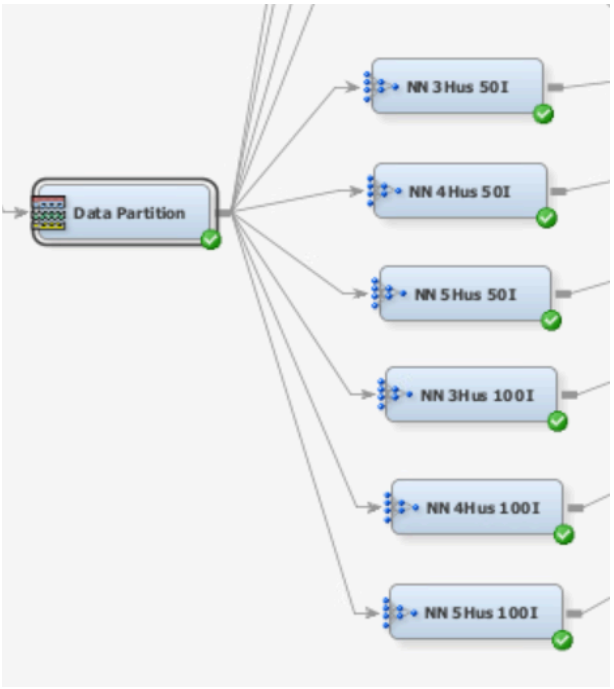
7. Neural Network

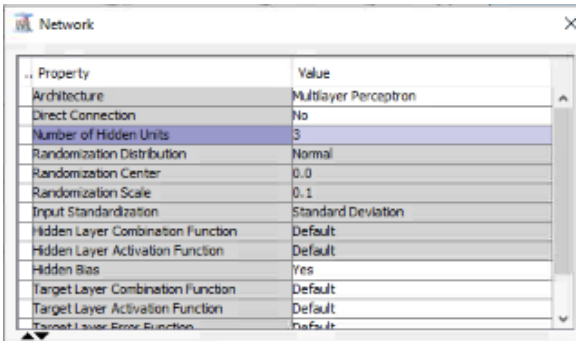
Neural Network is a model used to handle a wide variety of nonlinear data, effectively controlling the skewness of variables. It is an extension of regression analysis that allows inputs to be adjusted to better match the data. Neural networks are particularly strong in predictive analytics due to their ability to work with hidden layers, which enable the model to capture complex relationships in the data.

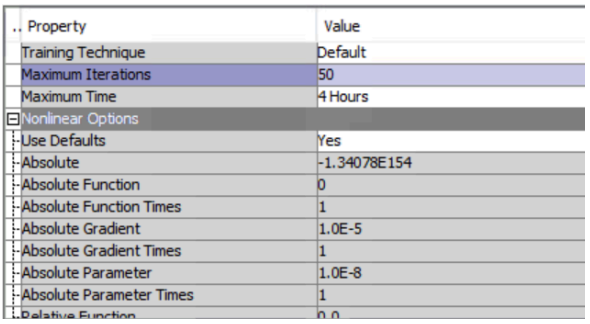
Different Neural Network nodes, with varying configurations of hidden units and iterations, will be connected to different parts of the process flow to evaluate which neural network performs best under various conditions. During the procedure, all neural network nodes will be initially disabled to prevent training, as the weights will be randomly initialized. Similar to decision tree and regression models, the selection of the best neural network will be based on the average errors of the validation data, ensuring the most accurate model is chosen for prediction.

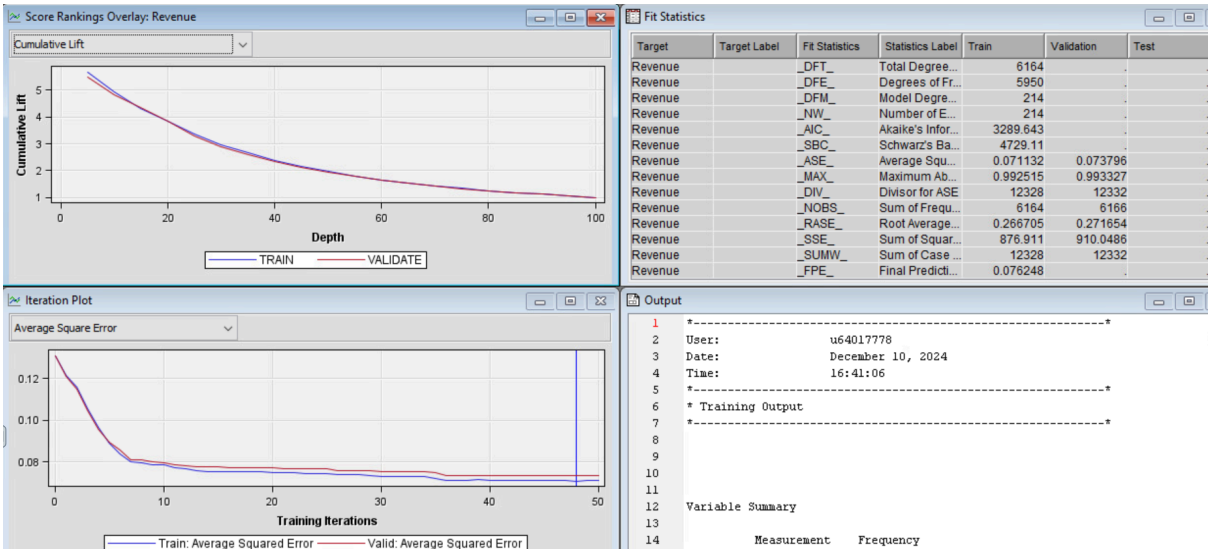
a. Neural Network connected from Data Partition Node

There were 6 different configurations of neural networks connected to the data partition to determine the best model for the neural network.



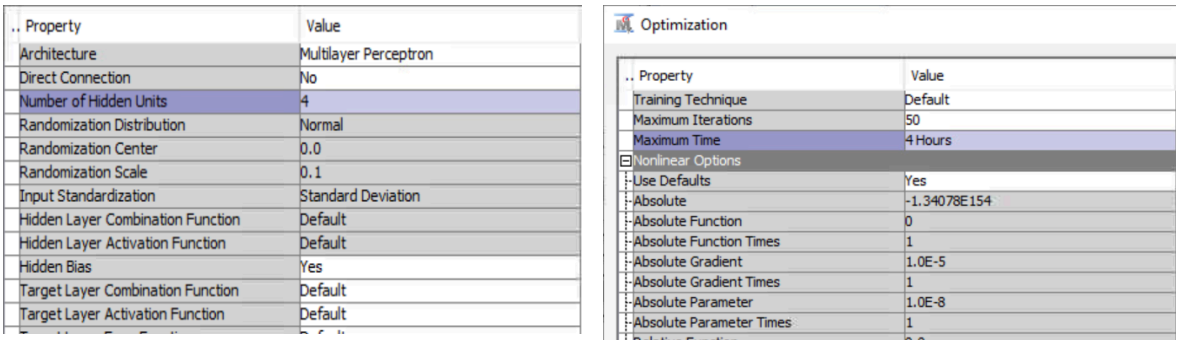


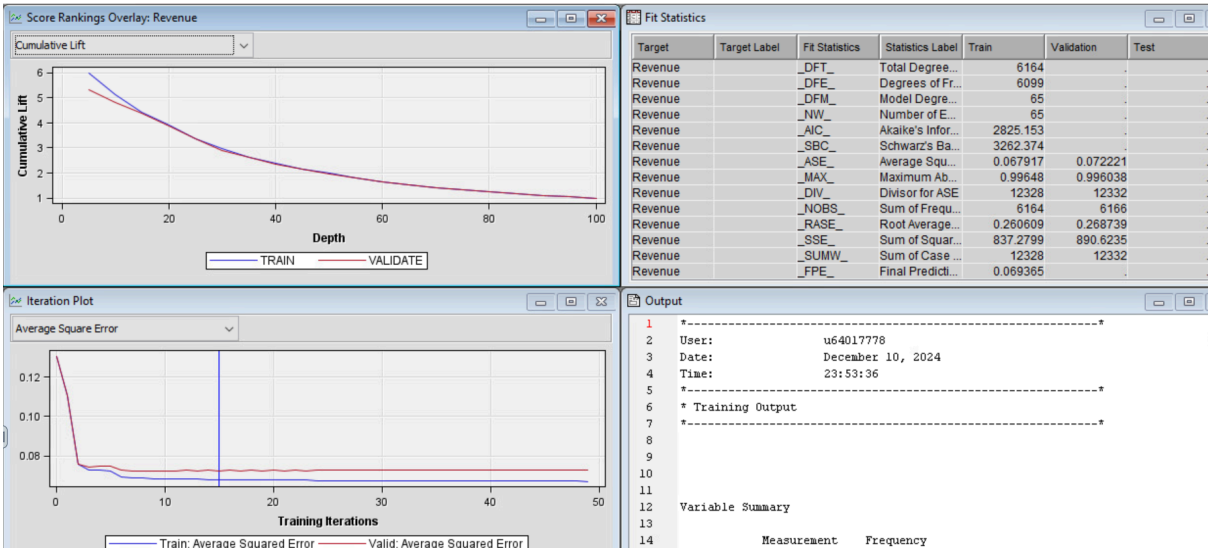




The first neural network with a configuration of 3 hidden units at 50 iterations had an average squared error of 0.073796.

4/50





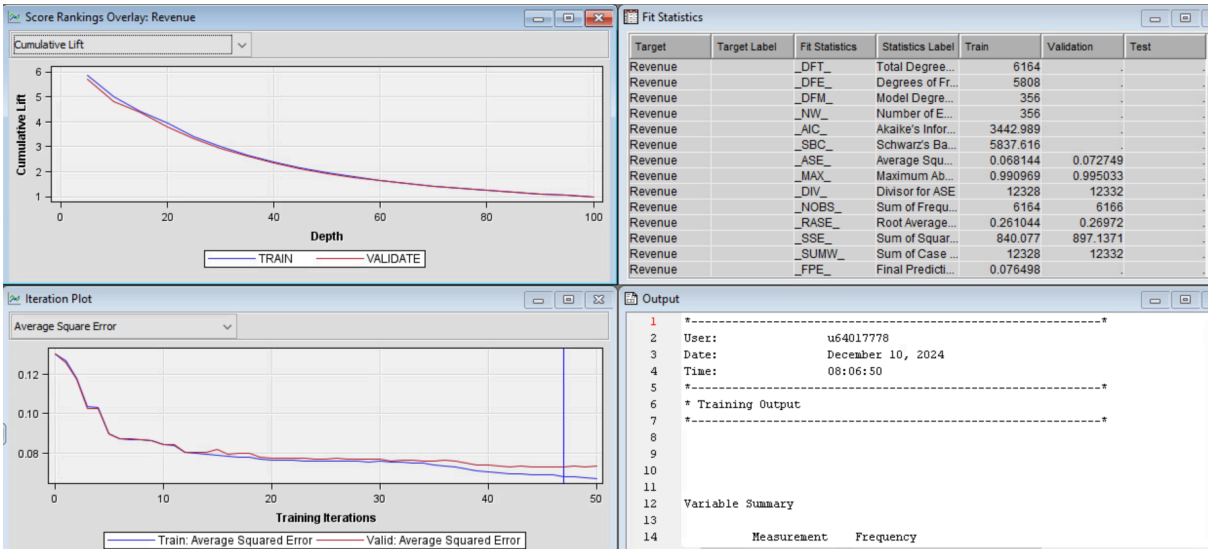
In order to increase the accuracy of the model, another 50 iterations were added. Therefore a neural network with a configuration of 4 hidden units at 50 iterations was then implemented. The second neural network result had an average squared error of 0.072221.

5/50

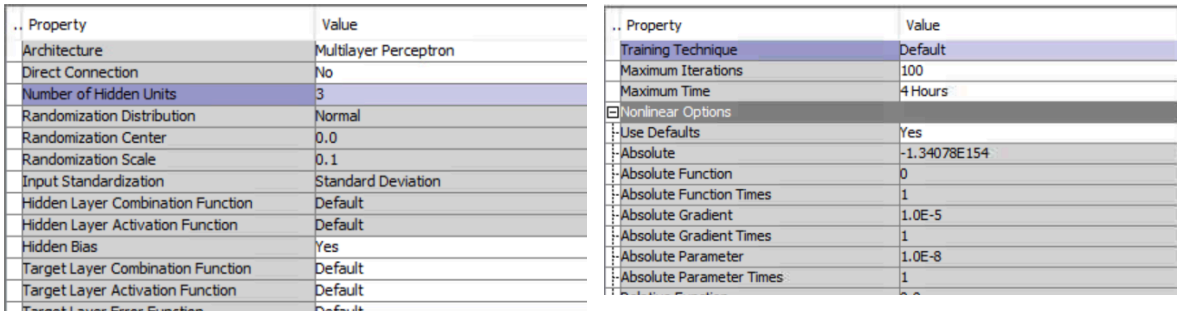
Property	Value
Architecture	Multilayer Perceptron
Direct Connection	No
Number of Hidden Units	5
Randomization Distribution	Normal
Randomization Center	0.0
Randomization Scale	0.1
Input Standardization	Standard Deviation
Hidden Layer Combination Function	Default
Hidden Layer Activation Function	Default
Hidden Bias	Yes
Target Layer Combination Function	Default
Target Layer Activation Function	Default
Target Layer Error Function	Default

Optimization

Property	Value
Training Technique	Default
Maximum Iterations	50
Maximum Time	4 Hours
Nonlinear Options	
Use Defaults	Yes
Absolute	-1.34078E154
Absolute Function	0
Absolute Function Times	1
Absolute Gradient	1.0E-5
Absolute Gradient Times	1
Absolute Parameter	1.0E-8
Absolute Parameter Times	1
Relative Function	0.0

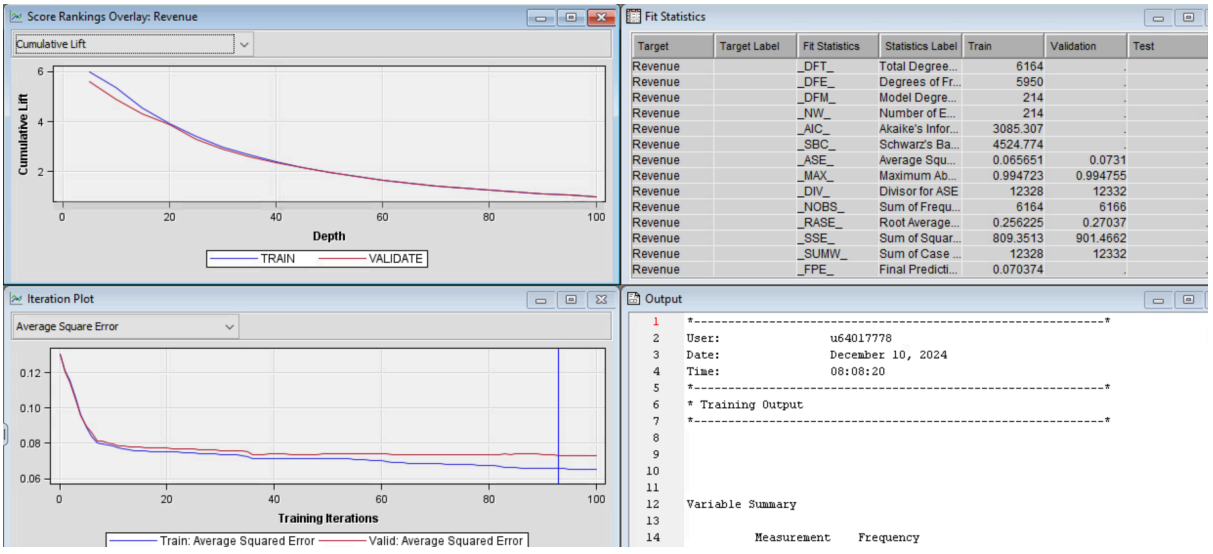


Therefore, the third neural network with a configuration of 5 hidden units at 50 iterations is added in order to get a lower average squared error. The result was 0.072749.



Property	Value
Architecture	Multilayer Perceptron
Direct Connection	No
Number of Hidden Units	3
Randomization Distribution	Normal
Randomization Center	0.0
Randomization Scale	0.1
Input Standardization	Standard Deviation
Hidden Layer Combination Function	Default
Hidden Layer Activation Function	Default
Hidden Bias	Yes
Target Layer Combination Function	Default
Target Layer Activation Function	Default

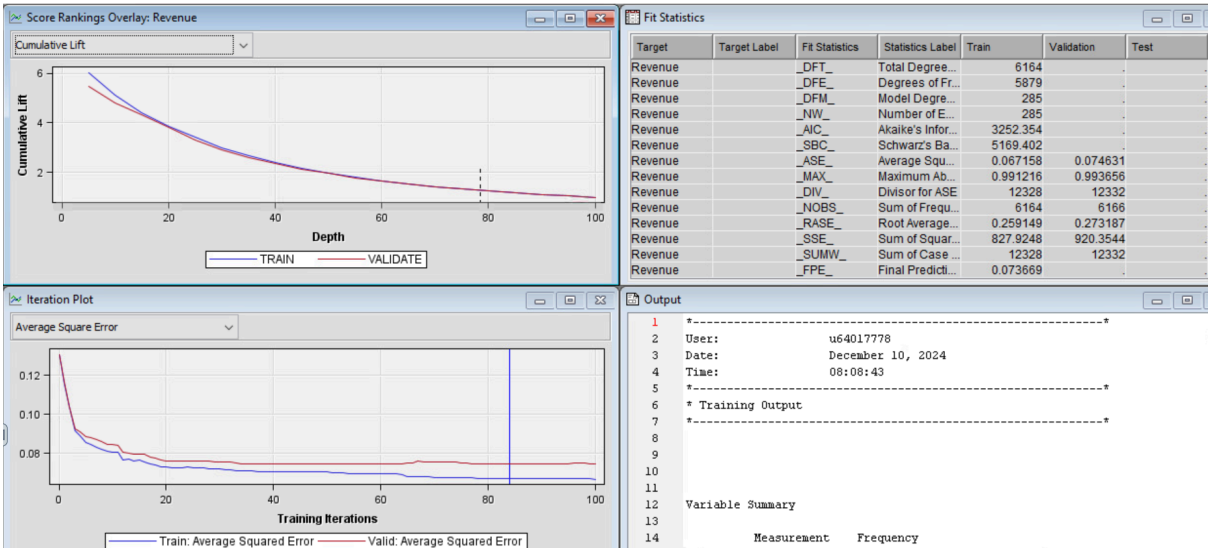
Property	Value
Training Technique	Default
Maximum Iterations	100
Maximum Time	4 Hours
Nonlinear Options	
Use Defaults	Yes
Absolute	-1.34078E154
Absolute Function	0
Absolute Function Times	1
Absolute Gradient	1.0E-5
Absolute Gradient Times	1
Absolute Parameter	1.0E-8
Absolute Parameter Times	1



The fourth neural network with a configuration of 3 hidden units at 100 iterations had an average squared error of 0.00731.

Property	Value
Architecture	Multilayer Perceptron
Direct Connection	No
Number of Hidden Units	4
Randomization Distribution	Normal
Randomization Center	0.0
Randomization Scale	0.1
Input Standardization	Standard Deviation
Hidden Layer Combination Function	Default
Hidden Layer Activation Function	Default
Hidden Bias	Yes
Target Layer Combination Function	Default
Target Layer Activation Function	Default
Target Layer Error Function	Default

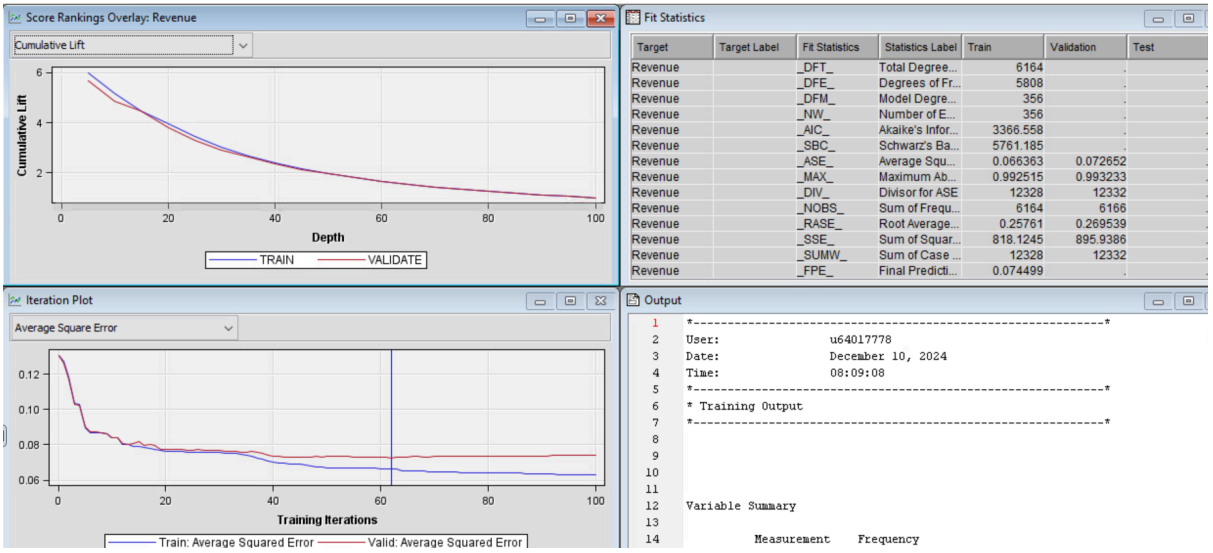
Property	Value
Training Technique	Default
Maximum Iterations	100
Maximum Time	4 Hours
Nonlinear Options	
Use Defaults	Yes
Absolute	-1.34078E154
Absolute Function	0
Absolute Function Times	1
Absolute Gradient	1.0E-5
Absolute Gradient Times	1
Absolute Parameter	1.0E-8
Absolute Parameter Times	1



The result suggested the model needed another 50 iterations and therefore the fourth neural network with a configuration of 4 hidden units at 100 iterations was then implemented. The result was the same as the third node, which had an average squared error of 0.074631.

Architecture	Multilayer Perceptron
Direct Connection	No
Number of Hidden Units	5
Randomization Distribution	Normal
Randomization Center	0.0
Randomization Scale	0.1
Input Standardization	Standard Deviation
Hidden Layer Combination Function	Default
Hidden Layer Activation Function	Default
Hidden Bias	Yes
Target Layer Combination Function	Default
Target Layer Activation Function	Default
Target Layer Error Function	Default

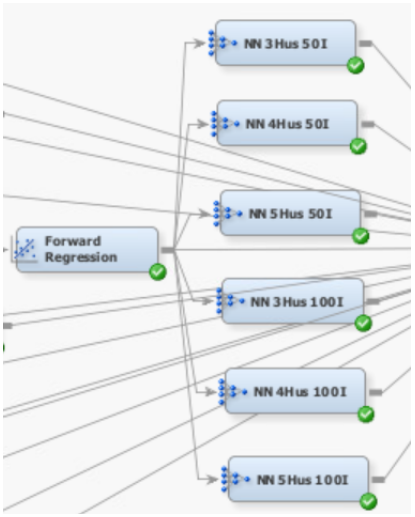
Property	Value
Training Technique	Default
Maximum Iterations	100
Maximum Time	4 Hours
Nonlinear Options	
Use Defaults	Yes
Absolute	-1.34078E154
Absolute Function	0
Absolute Function Times	1
Absolute Gradient	1.0E-5
Absolute Gradient Times	1
Absolute Parameter	1.0E-8
Absolute Parameter Times	1



Therefore, the sixth neural network with a configuration of 3 hidden units at 100 iterations is added in order to get a lower average squared error. The result was 0.072652, which is higher than the previous. **3 hidden units at 100 iterations** is the best neural network from imputed data. The summary of neural networks of imputed data are as follow:

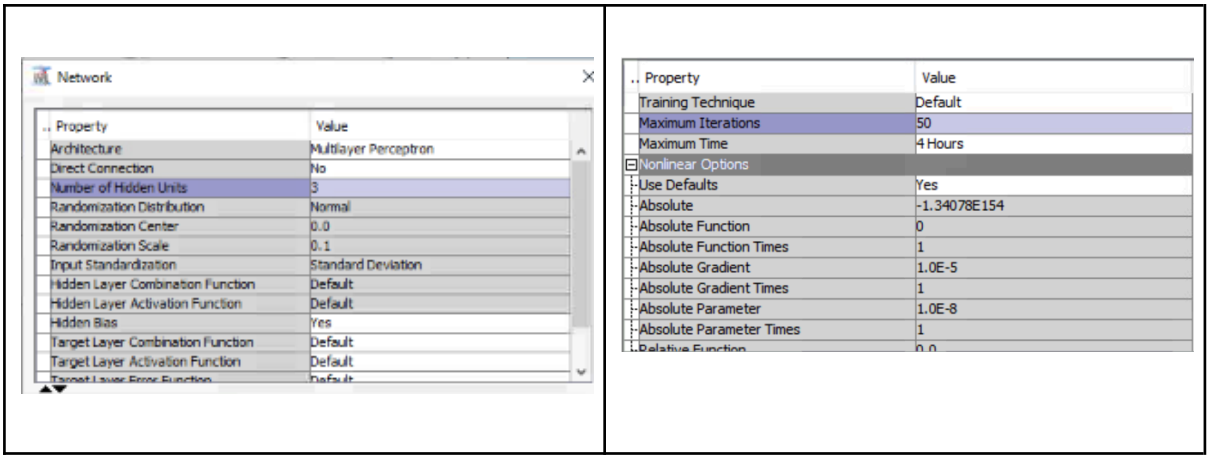
Neural Network	Hidden Units	Iterations	ASE
3HUs-50I	3	50	0.073796
4HUs-50I	4	50	0.072221
5HUs-50I	5	50	0.072749
3HUs-100I	**3**	**100**	**0.00731**
4HUs-100I	4	100	0.074631
5HUs-100I	5	100	0.072652

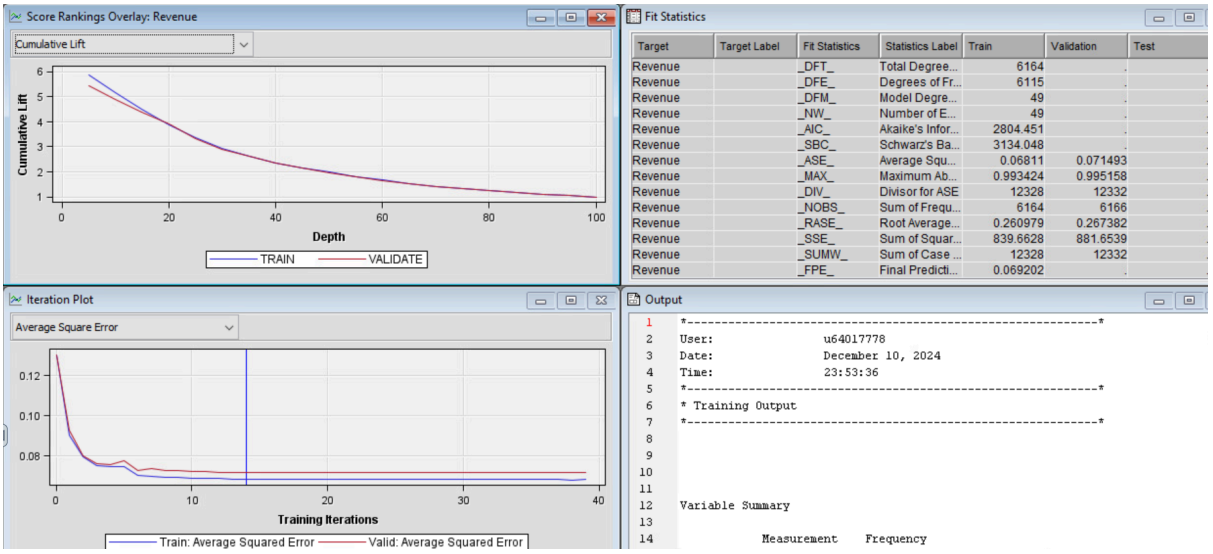
b. Variable Reduction Neural Networks connected from the Best Regression Model



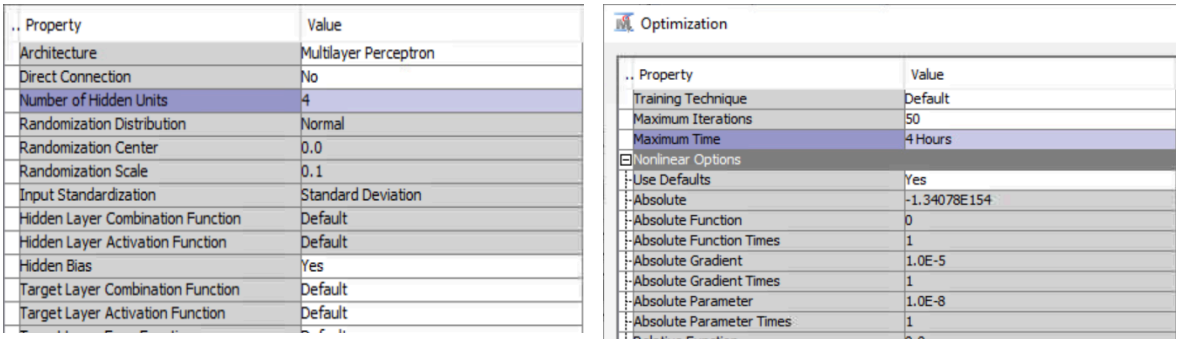
To reduce the complexity of the Neural Networks, another set of Neural Network nodes was connected to the best regression model with the lowest Average Squared Error (ASE). The Neural Network would only take inputs from the connected regression node for further analysis.

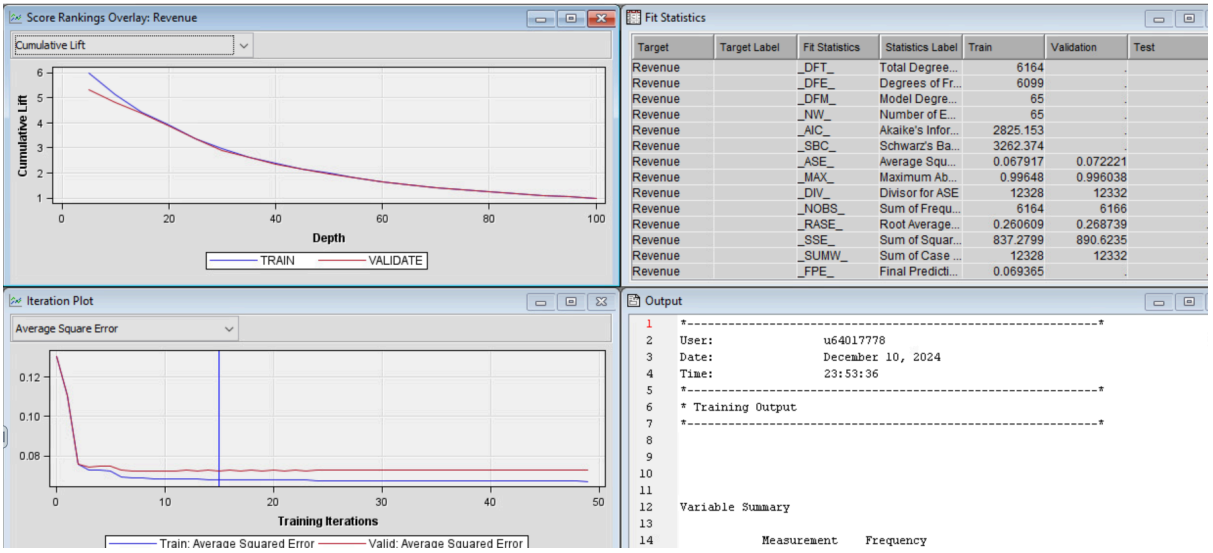
As outlined in Section 6: Logistic Regression, there were two best regression models with the lowest ASE: the forward regression model and the stepwise regression model. Therefore, separate sets of Neural Networks were connected to both regression nodes to analyze their performance and select the best configuration for further predictions.



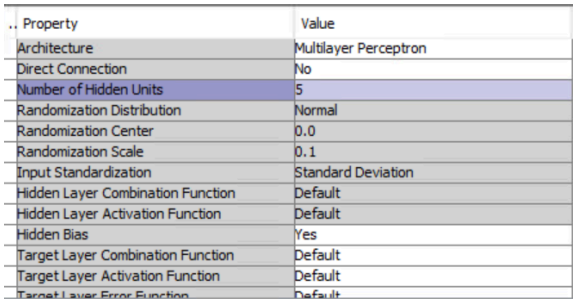


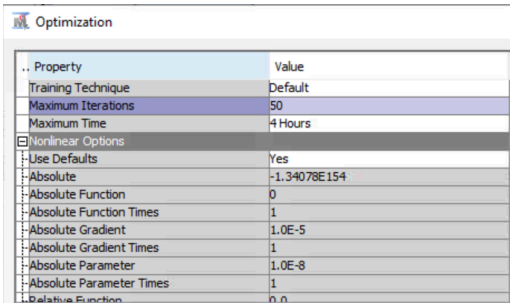
The first neural network with a configuration of 3 hidden units at 50 iterations had an average squared error of 0.071493.

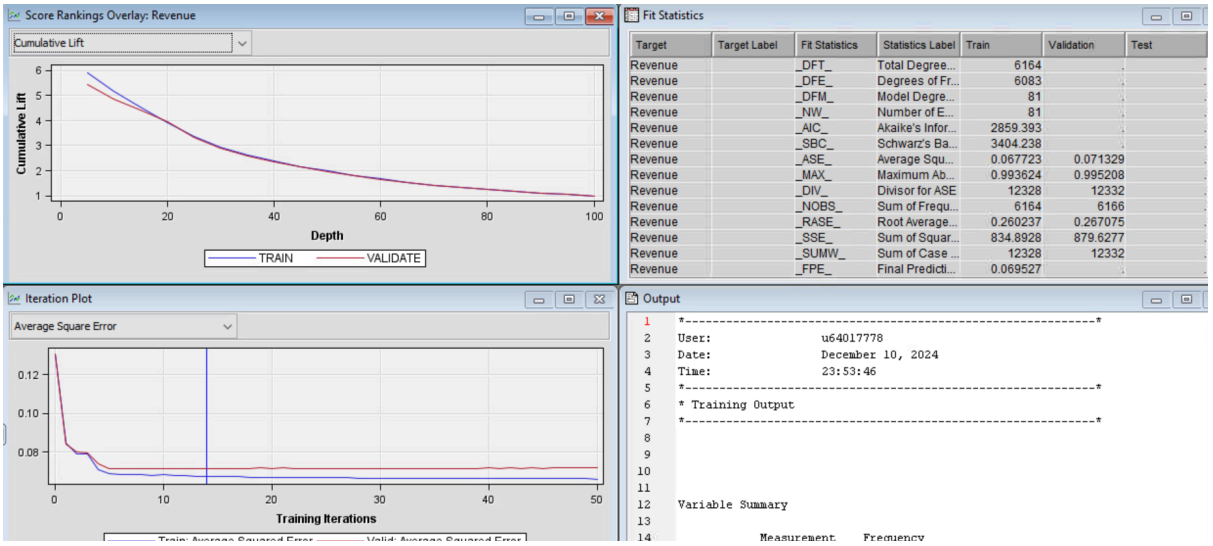




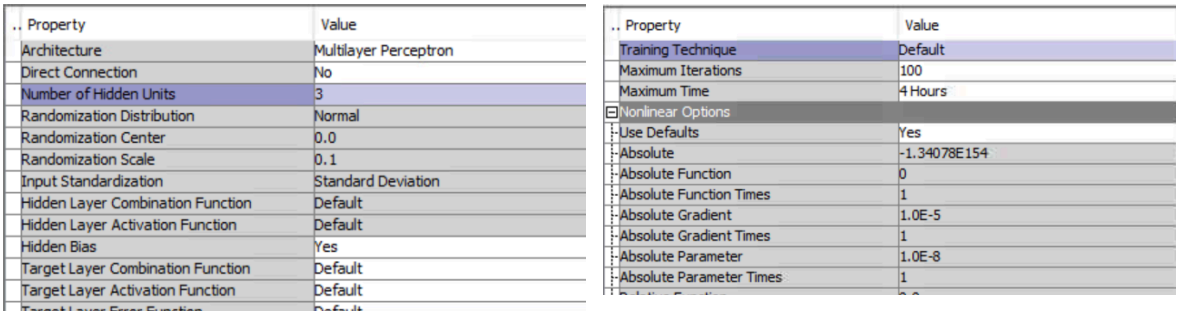
The result suggested the model needed another 50 iterations and therefore the fourth neural network with a configuration of 4 hidden units at 50 iterations was then implemented. The result was the same as the third node, which had an average squared error of 0.072221.





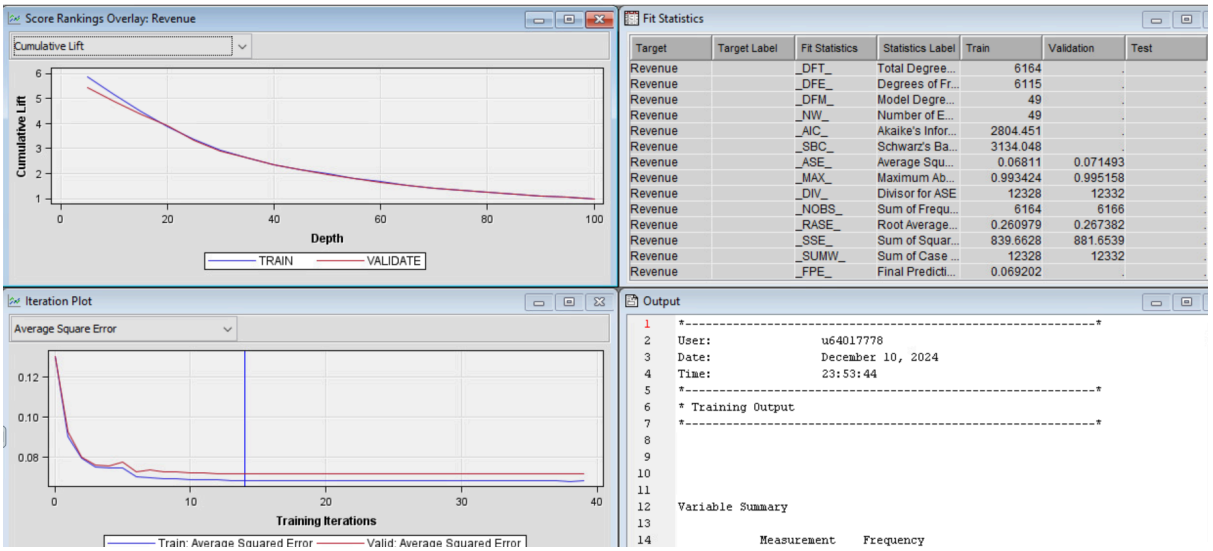


Therefore, the third neural network with a configuration of 5 hidden units and 50 iterations was added to achieve a lower average squared error. The result obtained was an average squared error (ASE) of 0.071329.



Property	Value
Architecture	Multilayer Perceptron
Direct Connection	No
Number of Hidden Units	3
Randomization Distribution	Normal
Randomization Center	0.0
Randomization Scale	0.1
Input Standardization	Standard Deviation
Hidden Layer Combination Function	Default
Hidden Layer Activation Function	Default
Hidden Bias	Yes
Target Layer Combination Function	Default
Target Layer Activation Function	Default

Property	Value
Training Technique	Default
Maximum Iterations	100
Maximum Time	4 Hours
Nonlinear Options	
Use Defaults	Yes
Absolute	-1.34078E154
Absolute Function	0
Absolute Function Times	1
Absolute Gradient	1.0E-5
Absolute Gradient Times	1
Absolute Parameter	1.0E-8
Absolute Parameter Times	1

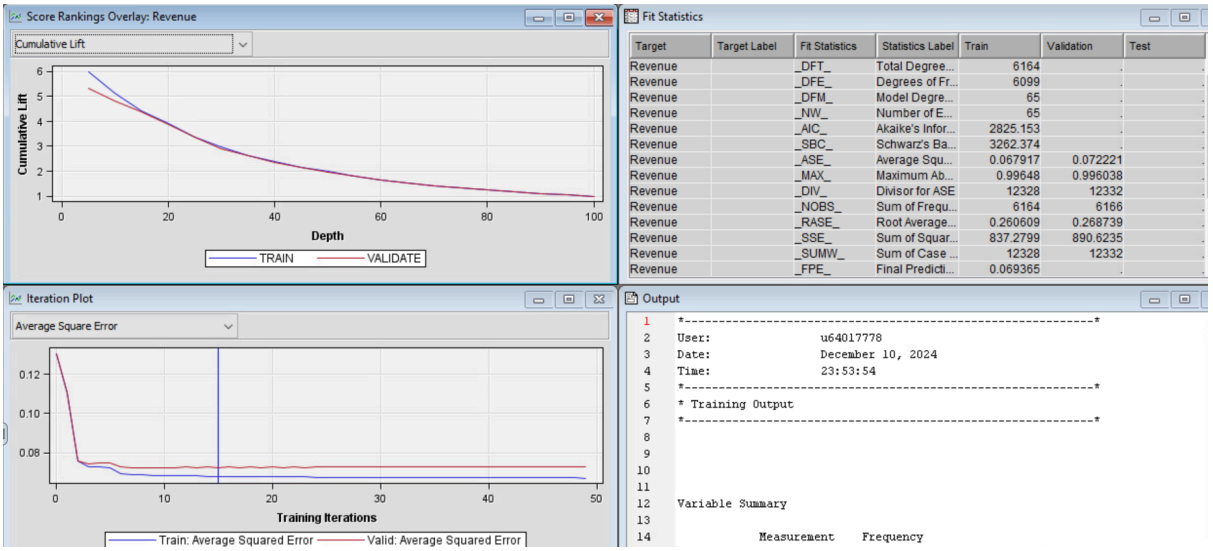


Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Revenue		_DFT_	Total Degree...	6164		
Revenue		_DFE_	Degrees of Fr...	6115		
Revenue		_DFM_	Model Degre...	49		
Revenue		_NW_	Number of E...	49		
Revenue		_AIC_	Akaike's Infor...	2804.451		
Revenue		_SBC_	Schwarz's Ba...	3134.048		
Revenue		_ASE_	Average Squ...	0.06811	0.071493	
Revenue		_MAX_	Maximum Ab...	0.993424	0.995158	
Revenue		_DIV_	Divisor for ASE	12328	12332	
Revenue		_NOBS_	Sum of Frequ...	6164	6166	
Revenue		_RASE_	Root Average...	0.260979	0.267382	
Revenue		_SSE_	Sum of Squar...	839.6628	881.6539	
Revenue		_SUMW_	Sum of Case ...	12328	12332	
Revenue		_FPE_	Final Predicti...	0.069202		

The fourth neural network with a configuration of 3 hidden units at 100 iterations had an average squared error of 0.071493.

Property	Value
Architecture	Multilayer Perceptron
Direct Connection	No
Number of Hidden Units	4
Randomization Distribution	Normal
Randomization Center	0.0
Randomization Scale	0.1
Input Standardization	Standard Deviation
Hidden Layer Combination Function	Default
Hidden Layer Activation Function	Default
Hidden Bias	Yes
Target Layer Combination Function	Default
Target Layer Activation Function	Default

Property	Value
Training Technique	Default
Maximum Iterations	100
Maximum Time	4 Hours
Nonlinear Options	
Use Defaults	Yes
Absolute	-1.34078E154
Absolute Function	0
Absolute Function Times	1
Absolute Gradient	1.0E-5
Absolute Gradient Times	1
Absolute Parameter	1.0E-8
Absolute Parameter Times	1



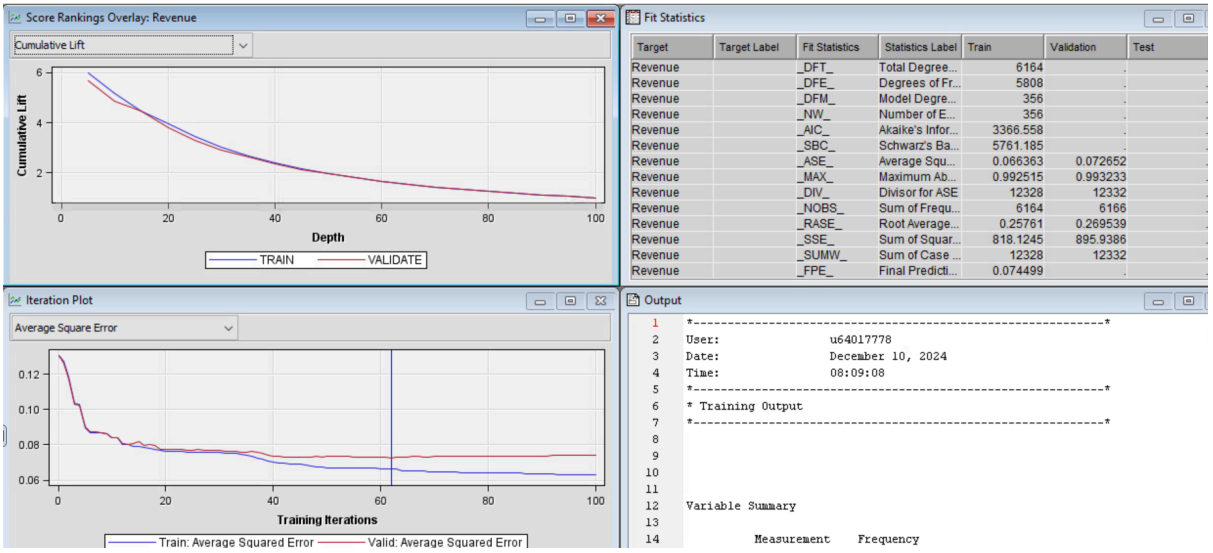
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Revenue		_DFT_	Total Degree...	6164		
Revenue		_DFE_	Degrees of Fr...	6099		
Revenue		_DFM_	Model Degre...	65		
Revenue		_NW_	Number of E...	65		
Revenue		_AIC_	Akaike's Infor...	2825.153		
Revenue		_SBC_	Schwarz's Ba...	3262.374		
Revenue		_ASE_	Average Squ...	0.067917	0.072221	
Revenue		_MAX_	Maximum Ab...	0.99648	0.996038	
Revenue		_DIV_	Divisor for ASE	12328	12332	
Revenue		_NOBS_	Sum of Frequ...	6164	6166	
Revenue		_RASE_	Root Average...	0.260609	0.268739	
Revenue		_SSE_	Sum of Squar...	837.2799	890.6235	
Revenue		_SUMW_	Sum of Case ...	12328	12332	
Revenue		_FPE_	Final Predicti...	0.069365		

```
1  *------------------------------------------------------------*
2  User:                 u64017778
3  Date:                 December 10, 2024
4  Time:                 23:53:54
5  *------------------------------------------------------------*
6  * Training Output
7  *------------------------------------------------------------*
8
9
10
11
12 Variable Summary
13
14            Measurement    Frequency
```

Therefore, the fifth neural network with a configuration of **4 hidden units** and **100 iterations** was added to achieve a lower average squared error. However, the result was 0.072221, which is higher than the previous configurations.

Property	Value
Architecture	Multilayer Perceptron
Direct Connection	No
Number of Hidden Units	5
Randomization Distribution	Normal
Randomization Center	0.0
Randomization Scale	0.1
Input Standardization	Standard Deviation
Hidden Layer Combination Function	Default
Hidden Layer Activation Function	Default
Hidden Bias	Yes
Target Layer Combination Function	Default
Target Layer Activation Function	Default
Target Layer Error Function	Default

Property	Value
Training Technique	Default
Maximum Iterations	100
Maximum Time	4 Hours
Nonlinear Options	
Use Defaults	Yes
Absolute	-1.34078E154
Absolute Function	0
Absolute Function Times	1
Absolute Gradient	1.0E-5
Absolute Gradient Times	1
Absolute Parameter	1.0E-8
Absolute Parameter Times	1
Relative Function	0.0



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Revenue		_DFT_	Total Degree...	6164		
Revenue		_DFE_	Degrees of Fr...	5808		
Revenue		_DFM_	Model Degre...	356		
Revenue		_NW_	Number of E...	356		
Revenue		_AIC_	Akaike's Infor...	3366.558		
Revenue		_SBC_	Schwarz's Ba...	5761.185		
Revenue		_ASE_	Average Squ...	0.066363	0.072652	
Revenue		_MAX_	Maximum Ab...	0.992515	0.993233	
Revenue		_DIV_	Divisor for ASE	12328	12332	
Revenue		_NOBS_	Sum of Frequ...	6164	6166	
Revenue		_RASE_	Root Average...	0.25761	0.269539	
Revenue		_SSE_	Sum of Squar...	818.1245	895.9386	
Revenue		_SUMW_	Sum of Case ...	12328	12332	
Revenue		_FPE_	Final Predicti...	0.074499		

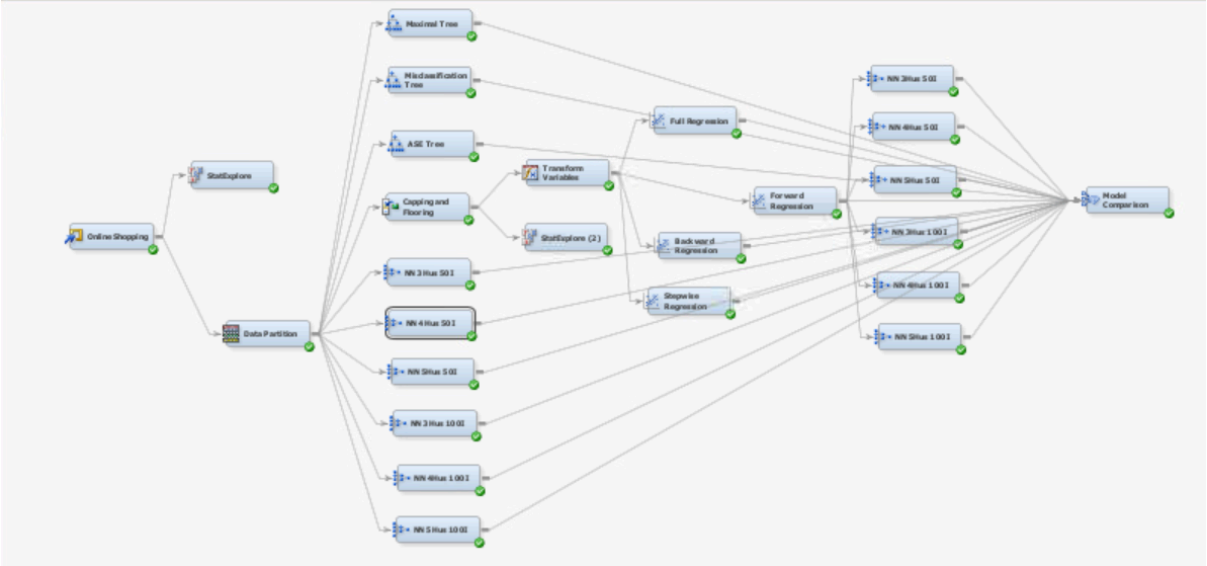
```
1  *------------------------------------------------------------*
2  User:                 u64017778
3  Date:                 December 10, 2024
4  Time:                 08:09:08
5  *------------------------------------------------------------*
6  * Training Output
7  *------------------------------------------------------------*
8
9
10
11
12 Variable Summary
13
14            Measurement    Frequency
```

Therefore, the sixth neural network with a configuration of 5 hidden units at 100 iterations is added in order to get a lower average squared error. The result was 0.072652, which is higher

than the previous. 5 hidden units at 50 iterations is the best neural network from imputed data. The summary of neural networks of imputed data are as follow:

Neural Network	Hidden Units	Iterations	ASE
3HUs-50I	3	50	0.071493
4HUs-50I	4	50	0.072221
5HUs-50I	**5**	**50**	**0.071329**
3HUs-100I	3	100	0.071493
4HUs-100I	4	100	0.072221
5HUs-100I	5	100	0.072652

8. Model Comparison

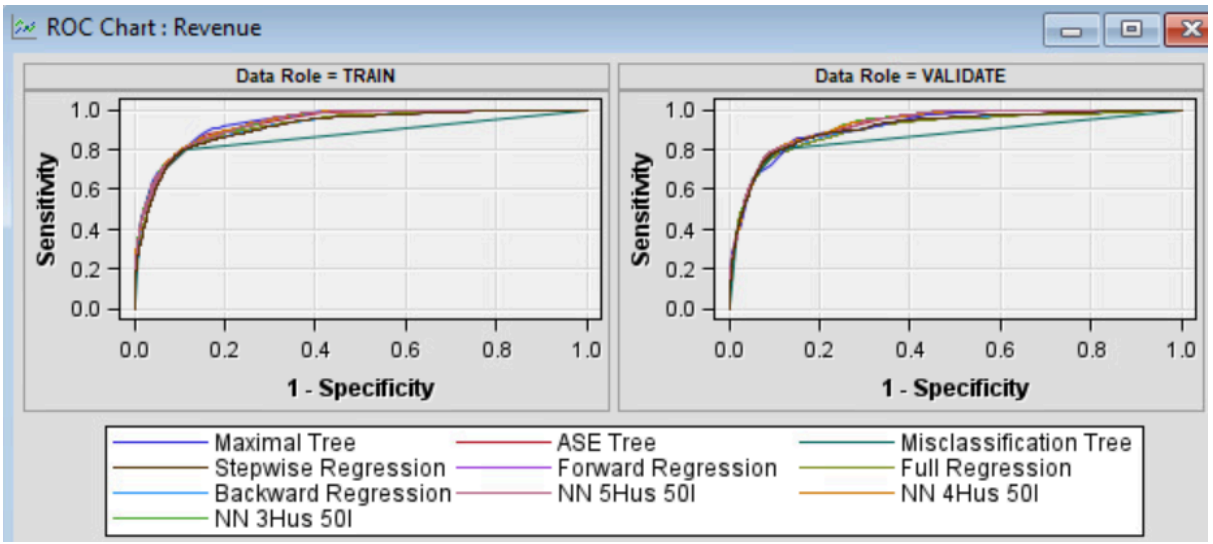


All models, including decision trees, regression models, and neural networks, were connected to the Model Comparison node to identify the model with the best performance.

The Model Comparison node was configured with the following setup:

- Selection Statistic: Average Square Error (ASE)
- Selection Table: Validation data, which provided a more reliable comparison than the training data.

Property	Value
Variables	
Assessment Reports	
Number of Bins	20
ROC Chart	Yes
Recompute	No
Model Selection	
Selection Data	Default
Selection Statistic	Average Squared Error
HP Selection Statistic	Default
SAS Viya Selection Statistic	
Selection Table	Validation
Selection Depth	10



The Neural Network model, derived from Regression with 4 hidden units and 50 interactions, demonstrated the best ROC curve performance among all models.

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Valid: Average Squared Error ▲	Train: Roc Index	Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error	Train: Sum of Frequencies	Train: Misclassification Rate	Train: Maximum Absolute Error	Train: Sum of Squared Errors	Train: Average Squared Error
Y	Neural12	Neural12	NN 5Hus 100I	0.071329	0.934	Revenue		0.071329	6164	0.091986	0.993624	834.8928	0.067723
	Neural9	Neural9	NN 5Hus 50I	0.071329	0.934	Revenue		0.071329	6164	0.091986	0.993624	834.8928	0.067723
	Neural10	Neural10	NN 3Hus 100I	0.071493	0.933	Revenue		0.071493	6164	0.094257	0.993424	839.6628	0.06811
	Neural7	Neural7	NN 3Hus 50I	0.071493	0.933	Revenue		0.071493	6164	0.094257	0.993424	839.6628	0.06811
	Neural11	Neural11	NN 4Hus 100I	0.072221	0.934	Revenue		0.072221	6164	0.095068	0.99648	837.2799	0.067917
	Neural8	Neural8	NN 4Hus 50I	0.072221	0.934	Revenue		0.072221	6164	0.095068	0.99648	837.2799	0.067917
	Tree2	Tree2	ASE Tree	0.072632	0.931	Revenue		0.072632	6164	0.092959	0.999548	837.9043	0.067968
	Neural4	Neural4	NN 5Hus 100I	0.072652	0.938	Revenue		0.072652	6164	0.09523	0.992515	818.1245	0.066363
	Neural6	Neural6	NN 5Hus 50I	0.072749	0.934	Revenue		0.072749	6164	0.099773	0.990969	840.077	0.068144
	Neural2	Neural2	NN 3Hus 100I	0.0731	0.937	Revenue		0.0731	6164	0.091175	0.994723	809.3513	0.065651
	Reg3	Reg3	Forward Regression	0.073437	0.912	Revenue		0.073437	6164	0.102855	0.99704	908.8922	0.073726
	Reg4	Reg4	Stepwise Regression	0.073437	0.912	Revenue		0.073437	6164	0.102855	0.99704	908.8922	0.073726
	Neural	Neural	NN 3Hus 50I	0.073796	0.929	Revenue		0.073796	6164	0.099448	0.992515	876.911	0.071132
	Tree3	Tree3	Maximal Tree	0.073901	0.937	Revenue		0.073901	6164	0.091499	0.990025	817.374	0.066302
	Reg	Reg	Backward Regression	0.074248	0.916	Revenue		0.074248	6164	0.097988	0.995689	893.9919	0.072517
	Neural3	Neural3	NN 4Hus 100I	0.074631	0.936	Revenue		0.074631	6164	0.09377	0.991216	827.9248	0.067158
	Neural5	Neural5	NN 4Hus 50I	0.074669	0.929	Revenue		0.074669	6164	0.099124	0.989061	868.3811	0.07044
	Reg2	Reg2	Full Regression	0.074878	0.918	Revenue		0.074878	6164	0.096853	0.993468	885.2373	0.071807
	Tree	Tree	Misclassification Tree	0.075315	0.868	Revenue		0.075315	6164	0.093121	0.960324	881.398	0.071496

When comparing the ROC index among models, four models achieved the same ROC index of 0.934. These models included:

- Neural Network from Forward Regression with 5 hidden units and 50 iterations.
- Neural Network from Forward Regression with 5 hidden units and 100 iterations.

A similar conclusion was reached when comparing the Average Square Error (ASE) across all models. The two neural network models mentioned earlier demonstrated the same validation ASE, further reinforcing their comparable performance:

Model	ASE	ROC	Complexity
NN 5HUs 100I	0.071329	0.934	1
NN 5HUs 50I	0.071329	0.934	2
NN 3HUs 100I	0.071493	0.933	3
NN 3HUs 50I	0.071493	0.933	4

To select the best-performing model, the **complexity level** was considered, as the four neural network models had identical **ROC index** and **ASE (0.071329)**. The selection process was based on the following criteria:

- **Hidden Units:** All four models were configured with 4 hidden units; thus, this number was selected as the optimal choice.
- **Iteration Number:** Between 50 and 100 iterations, the model with the **smaller number of iterations (50)** was preferred for simplicity.
- **Regression Type:** Since **stepwise regression** is more complex than **forward regression**, the model using forward regression was chosen as the best.

Based on these considerations, the **Neural Network model from Forward Regression with 5 hidden units and 50 iterations and 5 hidden units and 100 iterations** was identified as the best-performing model.

9. Conclusion

In our project, we set out to understand what drives online shopping purchases by looking at different factors such as user behavior and session details. We tested several models, including decision trees, regression, and neural networks, to predict the likelihood of a purchase. After comparing the results, we found that the Neural Network model with 5 hidden units and 50 iterations worked best, providing the most accurate predictions.

From our analysis, we identified some key factors that influence whether a customer makes a purchase, including browser type, Exit Rates, Page Values, and Visitor Type. For example, people using Browser 1 were found to be much less likely to buy, while factors like higher Exit Rates and Page Values made a purchase more likely.

a. Summary

- Data Used: We worked with the Online Shoppers Purchasing Intention Dataset from the UCI Machine Learning Repository, which tracks customer behavior on an e-commerce website.
- What We Did: We used different methods (decision trees, regression, and neural networks) to predict whether a visitor would make a purchase. After testing all the models, we chose the one that gave the best results.
- Key Findings: We found that things like the browser type, Exit Rates, Page Values, and whether the visitor is new or returning have the biggest impact on whether a purchase is made.

b. Recommendations

- Focus on Specific Browser Users: We recommend paying special attention to users on Browser 1, as they are less likely to make a purchase. Improving the experience for these users might help increase conversions.
- Improve Important Pages: We found that pages with higher Page Values and lower Exit Rates lead to more purchases. Improving the quality of these pages could boost conversions.
- Target New Visitors: New visitors are more likely to make a purchase, so running campaigns to attract them could be a good strategy.
- Run Promotions at the Right Time: We noticed that visitors in November are more likely to buy. Running special promotions around this time could lead to more sales.

10. References

- Sakar, C. & Kastro, Y. (2018). *Online Shoppers Purchasing Intention Dataset [Dataset]*. UCI Machine Learning Repository. https://doi.org/10.24432/C5F88Q.

- Runkle, D., & Ward, A. (2013). *Predictive analytics: The power to predict who will click, buy, lie, or die*. Wiley.

- Agerri, R., & García, F. (2019). *Exploring the relationship between customer behavior and product purchase intention on e-commerce platforms*. Computers in Human Behavior, 100, 48-55. https://doi.org/10.1016/j.chb.2019.07.029

- Zhang, Y., & Wang, S. (2018). *An empirical analysis of factors influencing online shopping behavior* (pp. 25-40). Journal of Business Research, 59(5), 24-39. https://doi.org/10.1016/j.jbusres.2016.04.002