BA 706 Project

DATASET: RBC CHURN DATASET

Grace Adaji

Chukwudera Stephen Chukelu

Alicia Waters

Chidiebere Odurukwe

Table of Contents

Dataset: RBC Churn Dataset

1. Executive Summary

Customer churn presents a significant challenge for RBC, impacting revenue and growth. Bailyn (2024) highlights that customer acquisition costs (CAC) in banking vary significantly, with retail banks averaging $561 per customer, emphasizing the importance of cost-effective retention strategies to sustain profitability. These high acquisition costs underscore the importance of customer retention as a cost-effective strategy to sustain profitability. This report focuses on predicting customer churn for RBC using a dataset of customer demographics, account details, and behavioral attributes. The goal was to identify patterns and factors contributing to churn and to develop predictive models that accurately forecast whether a customer would stay with or leave the bank.
The project involved several key steps:

- **Data Exploration**:

The dataset was analyzed to understand the distribution and characteristics of key variables such as age, balance, credit score, and tenure. Initial data exploration highlighted skewness in two variables: age and number of products.

- **Data Preparation**:

Transformations (i.e., Cap and Floor and log transformations) were applied to reduce skewness in variables. The log transformation reduced skewness for Age. New binary variables like **HasBalance** were engineered to enhance model performance. **The variable HasProducts** was added to fix the skewness of the variable **Number of Products**. The dataset was partitioned into 50% training and 50% validation sets.

- **Model Development**:

Three types of predictive models were built and compared the average squared error (ASE) to determine the most accurate model:

  - **Decision Tree**: A simple and interpretable model with an ASE of 0.124137.
  - **Logistic Regression**: Provided insights into key predictors with an ASE of 0.129104.
  - **Neural Network (Cap and floor)**: Achieved the best performance with an ASE of 0.01755 at 60 iterations and a ROC of 0.115123.

- **Model Comparison**:

The Neural Network (Cap & Floor) model achieved the best performance with the highest **ROC of 0.833** and second lowest **ASE of 0.115865.**

- **Key Findings**:

- **Factors Influencing Churn**: Gender, fewer products, inactive membership, and certain geographic regions (e.g., Germany) were significant predictors of churn.
- **Churn Rate**: The overall churn rate in the dataset was **[20.37%]** as seen in the dataset.

- **Recommendations**:
  - Implement targeted retention strategies for inactive members and those with low credit scores.
  - Develop region-specific outreach programs to address higher churn rates in specific areas.
  - Enhance customer engagement through personalized services and loyalty programs.

This analysis provides actionable insights for RBC to reduce customer churn and improve retention strategies. The neural network model serves as a robust tool for predicting at-risk customers, helping RBC take proactive measures to retain them.

Introduction

Customer churn poses a significant challenge for the banking industry. Studies show that financial institutions lose approximately $1 trillion annually due to customer attrition. Furthermore, acquiring new customers can cost up to 10 times more than retaining existing ones. This highlights the critical need for effective churn prediction and proactive retention strategies.

The goal of this project was to analyze RBC's customer data to identify patterns and predictors of churn. By developing robust predictive models, we aim to provide RBC with actionable insights to reduce churn rates and foster long-term customer relationships.

The dataset includes demographic details, account information, and behavioral metrics. The target variable, `Exited`, indicates whether a customer has left the bank (1) or remained (0). Our analysis focuses on understanding the drivers of churn and recommending strategies to retain high-risk customers.

2. Objective

The objective of this project is to develop predictive models to estimate which customers will churn for RBC. The goal is to identify key factors that influence whether a customer will stay or leave the bank. By analyzing customer demographics, account details, and behavioral attributes, the project aims to:

1. **Build and compare predictive models** (Decision Tree, Logistic Regression, and Neural Network) to forecast churn accurately.
2. **Identify significant predictors** of customer churn.
3. **Provide actionable insights** and recommendations to help RBC reduce churn rates and improve customer retention strategies.

Dataset Source

- **Dataset:** RBC Churn Dataset
- **Source:** https://www.kaggle.com/datasets/saadsalim997/rbcchurndataset

3. Data Exploration

Variables Overview

After obtaining the dataset, we conducted an exploratory analysis to understand its structure. The dataset comprises **10,000 customer records** with the following key variables:

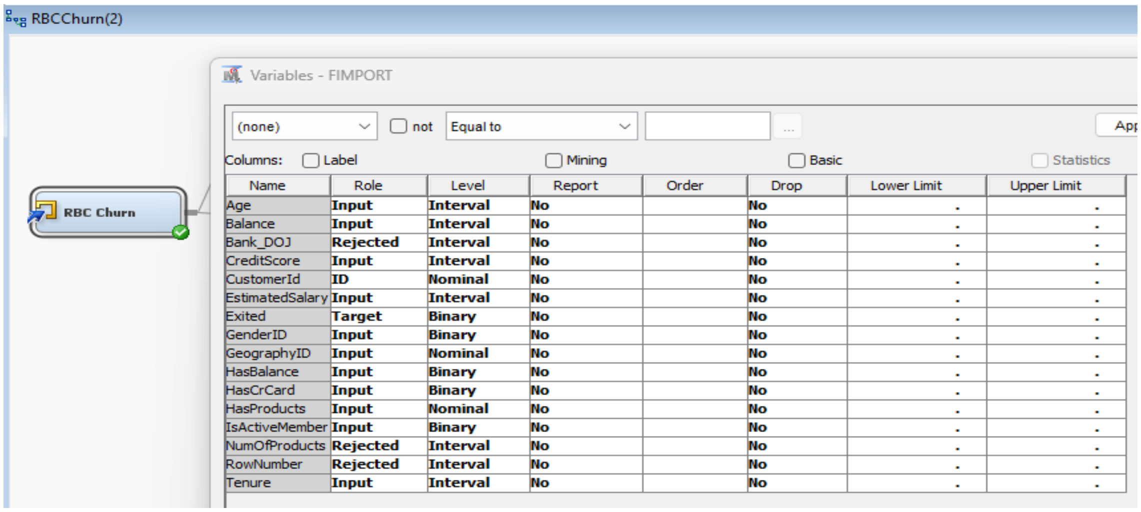
Variable	Description	Type
Age	Age of customer	Interval
Balance	Customer bank balance	Interval
Credit Score	Customer's credit score	Interval
CustomerId	Unique ID number of customer	ID
Estimated Salary	Approx. salary for each customer	Interval
GenderID	Female (2), Male (1)	Binary
GeographyID	France (1), Spain (2), Germany (3)	Nominal
HasBalance	Binary indicator for bank balance [Yes (1) / No (0)]	Binary
HasCrCard	Binary Indicator for credit card ownership [Yes (1) / No (0)]	Binary
IsActiveMember	Active (1), Inactive (0)	Binary
Tenure	Number of years that the customer has been with the bank	Interval
Has Products	1 product (0), multiple products (1)	Nominal

Rejected Variables

- **Bank DOJ:** Rejected due to difficulty in modeling datetime fields.
- **RowNumber:** Rejected due to lack of relevance for prediction.
- **NumofProducts:** Rejected and used HasProducts which has more significance.

Dataset Summary

- **Total Records:** 160,000 records
- **Missing Values:** 0 missing values
- **Variable Types:**
    - **Interval Variables:** Age, Balance, Credit Score, Estimated Salary, Tenure.
    - **Binary Variables:** Exited, GenderID, HasBalance, HasCrCard, IsActiveMember.
    - **Nominal Variables:** CustomerID, GeographyID, HasProducts



Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Age	Input	Interval	No		No	.	.
Balance	Input	Interval	No		No	.	.
Bank_DOJ	Rejected	Interval	No		No	.	.
CreditScore	Input	Interval	No		No	.	.
CustomerId	ID	Nominal	No		No	.	.
EstimatedSalary	Input	Interval	No		No	.	.
Exited	Target	Binary	No		No	.	.
GenderID	Input	Binary	No		No	.	.
GeographyID	Input	Nominal	No		No	.	.
HasBalance	Input	Binary	No		No	.	.
HasCrCard	Input	Binary	No		No	.	.
HasProducts	Input	Nominal	No		No	.	.
IsActiveMember	Input	Binary	No		No	.	.
NumOfProducts	Rejected	Interval	No		No	.	.
RowNumber	Rejected	Interval	No		No	.	.
Tenure	Input	Interval	No		No	.	.

Figure 1

4. Data Preparation

Steps Taken

1. Rejected Variables:
    - **Bank_DOJ, RowNumber,** and **NumofProducts** were excluded as they were deemed irrelevant.
2. Handling Missing Values:
    - No missing values were included in the dataset
3. Data Partition:
    - Split the data into **50% Train** and **50% Validation** sets.
4. Transformations:
    - Impute Variables: There were no missing or null values for variables in this dataset, therefore imputation was not required.
    - Skewed Variables:
        - **Cap and Floor** was applied to reduce skewness in the **Age** variable. Skewness was reduced but remained. (Figure 3)
        - **Log Transformation** was applied to further reduce skewness in the **Age** variable. After transformation, the **Age** variable was no longer skewed. (Figure 4)
5. Feature Engineering (creating new feature):
    - We created a binary feature **HasProducts** as an indicator of customers with one product (0) and more than one product (1)
    - We also created another binary feature **HasBalance** as an indicator of customers who have no balance (0) or a positive balance in their account (1)

Results - Node: StatExplore Diagram: RBCChurn(2)

Interval Variables

Data Role	Target	Target Level	Variable	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Role	Label	Scaled Mean Deviation	Maximum Deviation	Level Id
TRAIN	Exited	0	Balance	92072.68	0	7963	0	221532.8	72745.3	62848.04	-0.04689	-1.55769	INPUT	Balance	-0.04891	0.191181	1
TRAIN	Exited	1	Balance	109339.2	0	2037	0	250898.1	91108.54	58360.79	-0.51273	-0.9278	INPUT	Balance	0.191181	0.191181	2
TRAIN	Exited	0	Age	36	0	7963	18	92	37.40839	10.12536	1.377699	2.850252	INPUT	Age	-0.03888	0.152002	1
TRAIN	Exited	1	Age	45	0	2037	18	84	44.838	9.761562	0.077978	-0.12254	INPUT	Age	0.152002	0.152002	2
TRAIN	Exited	0	EstimatedSalary	99645.04	0	7963	90.07	199992.5	99738.39	57405.59	0.010996	-1.17284	INPUT	EstimatedSalary	-0.00352	0.013742	1
TRAIN	Exited	1	EstimatedSalary	102402.9	0	2037	11.58	199808.1	101465.7	57912.42	-0.0331	-1.21228	INPUT	EstimatedSalary	0.013742	0.013742	2
TRAIN	Exited	0	CreditScore	653	0	7963	405	850	651.8532	95.65384	-0.04702	-0.48473	INPUT	CreditScore	0.002036	0.007959	1
TRAIN	Exited	1	CreditScore	646	0	2037	350	850	645.3515	100.3215	-0.14108	-0.27408	INPUT	CreditScore	-0.00796	0.007959	2
TRAIN	Exited	0	Tenure	5	0	7963	3	7	4.86387	1.219166	0.198491	-0.94513	INPUT	Tenure	-8.83E-5	.0003452	1
TRAIN	Exited	1	Tenure	5	0	2037	3	7	4.865979	1.200117	0.160111	-0.8952	INPUT	Tenure	.0003452	.0003452	2

Figure 2

Results - Node: StatExplore (C&F) Diagram: RBCChurn(2)

Interval Variables

Data Role	Target	Target Level	Variable	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Role	Label	Scaled Mean Deviation	Maximum Deviation	Level Id
TRAIN	Exited	0	REP_Balance	92466.21	0	3981	0	214347	72949.37	62917.25	-0.05155	-1.56107	INPUT	Replacement: Ba...	-0.05178	0.202491	1
TRAIN	Exited	1	REP_Balance	110067.7	0	1018	0	250898.1	92511.13	58676.82	-0.5279	-0.86719	INPUT	Replacement: Ba...	0.202491	0.202491	2
TRAIN	Exited	0	REP_Age	36	0	3981	18	70.28176	37.39726	9.896105	1.185271	1.843091	INPUT	Replacement: Age	-0.0375	0.146653	1
TRAIN	Exited	1	REP_Age	45	0	1018	20	70.28176	44.55246	9.705595	0.032388	-0.25602	INPUT	Replacement: Age	0.146653	0.146653	2
TRAIN	Exited	0	REP_CreditScore	651	0	3981	405	850	650.5293	95.20168	-0.04137	-0.43209	INPUT	Replacement: Cr...	0.00167	0.006531	1
TRAIN	Exited	1	REP_CreditScore	645	0	1018	360.5515	850	645.2034	100.3697	-0.05286	-0.38744	INPUT	Replacement: Cr...	-0.00653	0.006531	2
TRAIN	Exited	0	REP_EstimatedS...	100324	0	3981	106.67	199992.5	100381.6	57546.61	-0.00152	-1.19076	INPUT	Replacement: Es...	0.001571	0.006143	1
TRAIN	Exited	1	REP_EstimatedS...	100496.8	0	1018	91.75	199725.4	99608.51	57887.6	-0.00359	-1.18399	INPUT	Replacement: Es...	-0.00614	0.006143	2
TRAIN	Exited	0	REP_Tenure	5	0	3981	3	7	4.872896	1.222043	0.19485	-0.9422	INPUT	Replacement: Te...	.0001482	.0005794	1
TRAIN	Exited	1	REP_Tenure	5	0	1018	3	7	4.869352	1.182311	0.146926	-0.87988	INPUT	Replacement: Te...	-.000579	.0005794	2

Figure 3

Results - Node: StatExplore (Trans) Diagram: RBCChurn(2)

Interval Variables

Data Role	Target	Target Level	Variable	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Role	Label	Scaled Mean Deviation	Maximum Deviation	Level Id
TRAIN	Exited	0	REP_Balance	92466.21	0	3981	0	214347	72949.37	62917.25	-0.05155	-1.56107	INPUT	Replacement: Ba...	-0.05178	0.202491	1
TRAIN	Exited	1	REP_Balance	110067.7	0	1018	0	250898.1	92511.13	58676.82	-0.5279	-0.86719	INPUT	Replacement: Ba...	0.202491	0.202491	2
TRAIN	Exited	0	LOG_REP_Age	3.610918	0	3981	2.944439	4.266641	3.617714	0.242572	0.336399	0.454906	INPUT	Transformed: Re...	-0.00987	0.038584	1
TRAIN	Exited	1	LOG_REP_Age	3.828641	0	1018	3.044522	4.266641	3.794742	0.224395	-0.55985	0.165358	INPUT	Transformed: Re...	0.038584	0.038584	2
TRAIN	Exited	0	REP_CreditScore	651	0	3981	405	850	650.5293	95.20168	-0.04137	-0.43209	INPUT	Replacement: Cr...	0.00167	0.006531	1
TRAIN	Exited	1	REP_CreditScore	645	0	1018	360.5515	850	645.2034	100.3697	-0.05286	-0.38744	INPUT	Replacement: Cr...	-0.00653	0.006531	2
TRAIN	Exited	0	REP_EstimatedS...	100324	0	3981	106.67	199992.5	100381.6	57546.61	-0.00152	-1.19076	INPUT	Replacement: Es...	0.001571	0.006143	1
TRAIN	Exited	1	REP_EstimatedS...	100496.8	0	1018	91.75	199725.4	99608.51	57887.6	-0.00359	-1.18399	INPUT	Replacement: Es...	-0.00614	0.006143	2
TRAIN	Exited	0	REP_Tenure	5	0	3981	3	7	4.872896	1.222043	0.19485	-0.9422	INPUT	Replacement: Te...	.0001482	.0005794	1
TRAIN	Exited	1	REP_Tenure	5	0	1018	3	7	4.869352	1.182311	0.146926	-0.87988	INPUT	Replacement: Te...	-.000579	.0005794	2

Figure 4

5. Model Development

5.1 Decision Trees

We developed three types of decision tree models to predict customer churn: **Assessment Tree**, **Maximal Tree**, and **Average Squared Error (ASE) Tree**. Each tree was configured with specific settings and evaluated based on their **ASE** and other fit statistics.

Assessment Tree

In the properties panel, we kept the subtree method to “Assessment” and the assessment measure to “Decision”. After running the model, the ASE was 0.130702. There were 9 leaves and we had **Age** as the first split and competing splits were **IsActiveMember**, **GeographyID** and **HasProduct**.

Configuration:

- **Subtree Method**: Assessment
- **Assessment Measure**: Decision
- **First Split**: Age
- **Competing Splits**: IsActiveMember”, “GeographyID” and “HasProduct”

Fit Statistics:

Statistic	Train	Validation
Misclassification Rate	0.157832	0.163367
Average Squared Error (ASE)	0.124893	0.130702
Root ASE	0.353402	0.361527

Tree Details:

- Number of Leaves: 9
- First Split: Age
- Key Variables:
    - IsActiveMember
    - HasProduct
    - GeographyID

Insights:

This tree balances simplicity and performance, providing a clear understanding of how churn is influenced by age, product count, and balance status.

Screenshots:

.. 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	Decision
Assessment Fraction	0.25
Cross Validation	
Perform Cross Validation	No
Number of Subsets	10
Number of Repeats	1
Seed	12345

Figure 5

Results - Node: Decision Tree Diagram: RBCChurn(2)

File Edit View Window

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Exited		_NOBS_	Sum of Frequencies	4999	5001	.
Exited		_MISC_	Misclassification Rate	0.157832	0.163367	.
Exited		_MAX_	Maximum Absolute Error	0.888856	1	.
Exited		_SSE_	Sum of Squared Errors	1248.681	1307.277	.
Exited		_ASE_	Average Squared Error	0.124893	0.130702	.
Exited		_RASE_	Root Average Squared Error	0.353402	0.361527	.
Exited		_DIV_	Divisor for ASE	9998	10002	.
Exited		_DFT_	Total Degrees of Freedom	4999	.	.

Figure 6

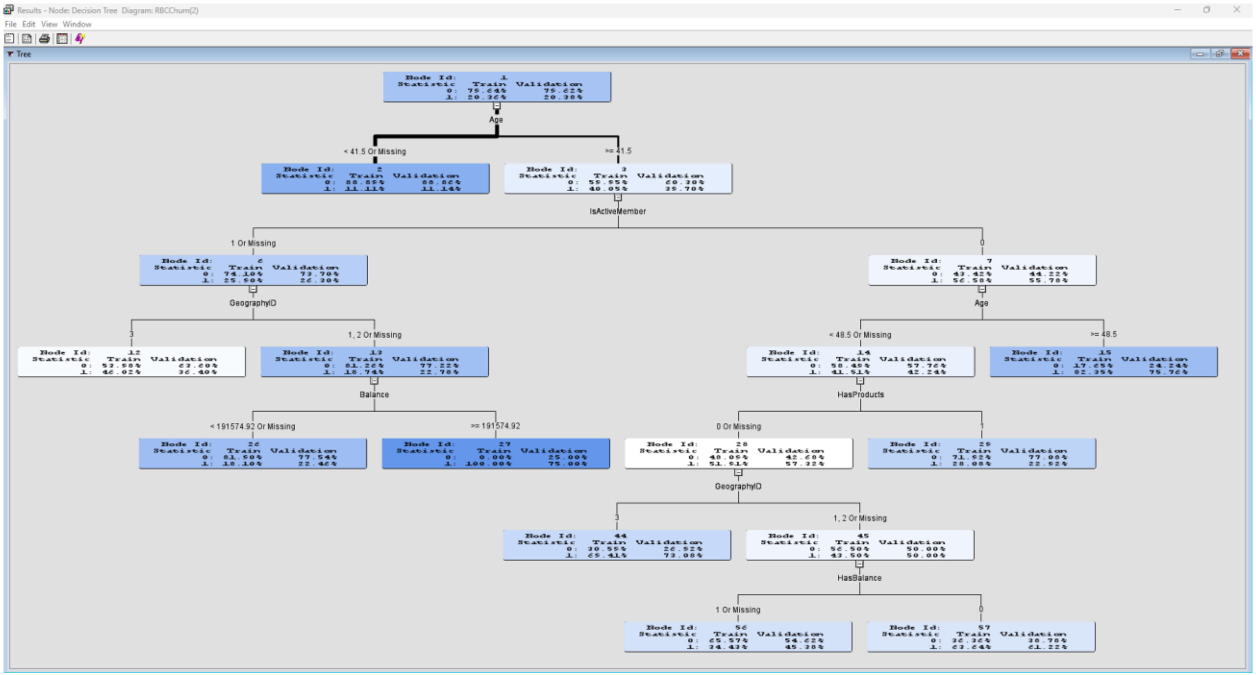


Figure 7

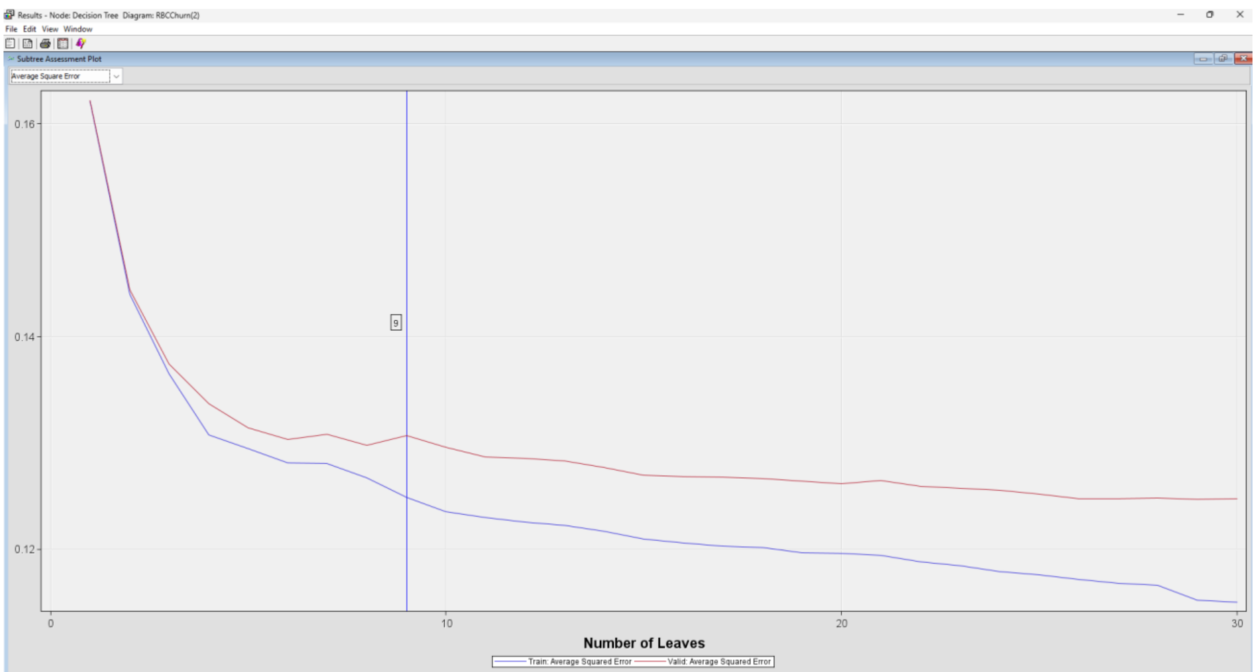


Figure 8

Maximal Tree

We built the maximal tree to determine all the possible splits. We changed the subtree method to "Largest" and the assessment measure to "Decision". After running the model, the ASE was 0.124772. There were 30 leaves and we had **Age** as the first split, with competing splits being **IsActiveMember**, **GeographyID** and **Has products**.

Configuration:

- **Subtree Method**: Largest
- **Assessment Measure**: Decision
- **First Split**: Age
- **Competing Splits**: IsActiveMember, GeographyID, HasProducts, Balance, HasBalance and GenderID.

Fit Statistics:

Statistic	Train	Validation
Misclassification Rate	0.15103	0.163967
Average Squared Error (ASE)	0.115048	0.124772
Root ASE	0.339187	0.35323

Tree Details:

- **Number of Leaves**: 30
- **First Split**: Age
- **Key Variables**:
  - IsActiveMember
  - GeographyID
  - HasProducts

Insights:

The maximal tree captures all possible splits, providing detailed insights but potentially overfitting the data. It highlights complex interactions among variables.

Screenshots:

Property	Value
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	Largest
Number of Leaves	1
Assessment Measure	Decision
Assessment Fraction	0.25
Cross Validation	
Perform Cross Validation	No
Number of Subsets	10
Number of Repeats	1
Seed	12345
Observation Based Importance	
Observation Based Importance	No
Number Single Var Importance	5
P-Value Adjustment	
Bonferroni Adjustment	Yes
Time of Bonferroni Adjustment	Before
Inputs	No
Number of Inputs	1
Depth Adjustment	Yes
Output Variables	
Leaf Variable	Yes

Figure 9

Results - Node: Maximal Tree Diagram: RBCChurn(2)

File Edit View Window

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Exited		_NOBS_	Sum of Frequencies	4999	5001	.
Exited		_MISC_	Misclassification Rate	0.15103	0.163967	.
Exited		_MAX_	Maximum Absolute Error	0.971963	1	.
Exited		_SSE_	Sum of Squared Errors	1150.251	1247.966	.
Exited		_ASE_	Average Squared Error	0.115048	0.124772	.
Exited		_RASE_	Root Average Squared Error	0.339187	0.35323	.
Exited		_DIV_	Divisor for ASE	9998	10002	.
Exited		_DFT_	Total Degrees of Freedom	4999	.	.

Figure 10

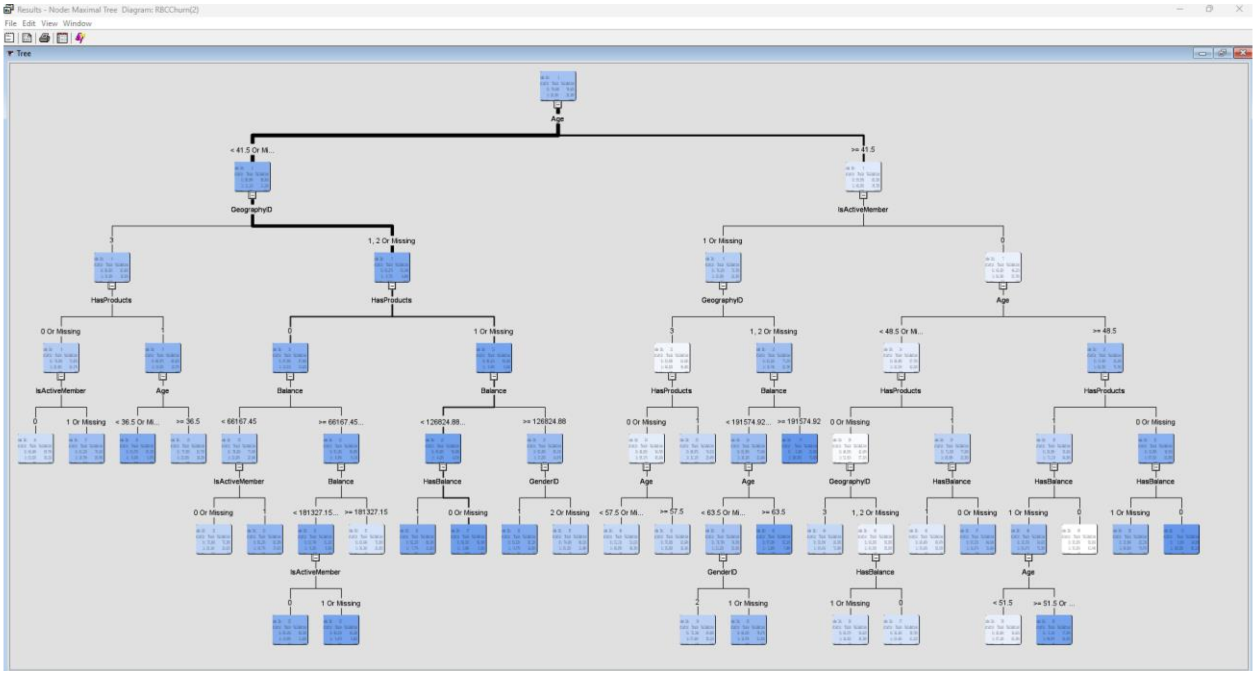


Figure 11

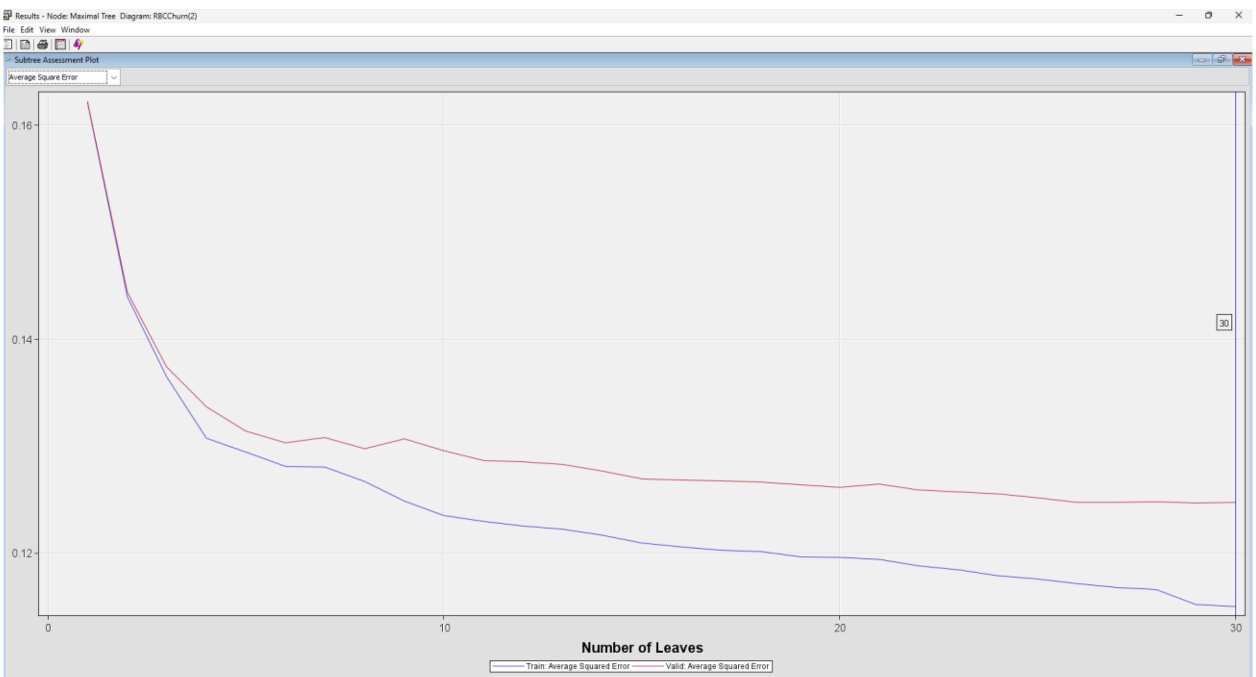


Figure 12

ASE Tree

We built the ASE tree to optimize the best split. We changed the subtree method to "Assessment" and the assessment measure to "Average Square Error". After running the model, the ASE was 0.124137. There were 26 leaves with "Age" as the first split and competing splits were **IsActiveMember, GeographyID, HasProducts, Balance, HasBalance** and **GenderID**.

Configuration:

- **Subtree Method**: Assessment
- **Assessment Measure**: Average Square Error
- **First Split**: Age
- **Competing Splits**: IsActiveMember, GeographyID, HasProducts

Fit Statistics:

Statistic	**Train**	**Validation**
Misclassification Rate	0.153431	0.165767
Average Squared Error (ASE)	0.116497	0.124137
Root ASE	0.341316	0.352331

Tree Details:

- **Number of Leaves**: 26
- **First Split**: Age
- **Key Variables**:
    - IsActiveMember
    - GeographyID
    - HasProducts

Insights:

The ASE tree was optimized to minimize the Average Squared Error. It offers a balance between detail and performance, and avoids excessive overfitting.

Screenshots:

Property	Value
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
Assessment Fraction	0.25
Cross Validation	
Perform Cross Validation	No
Number of Subsets	10
Number of Repeats	1
Seed	12345
Observation Based Importance	
Observation Based Importance	No
Number Single Var Importance	5
P-Value Adjustment	
Bonferroni Adjustment	Yes
Time of Bonferroni Adjustment	Before
Inputs	No
Number of Inputs	1
Depth Adjustment	Yes
Output Variables	
Leaf Variable	Yes

Figure 13

Results - Node: ASE Tree Diagram: RBCChurn(2)

File Edit View Window

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Exited		_NOBS_	Sum of Frequencies	4999	5001	.
Exited		_MISC_	Misclassification Rate	0.153431	0.165767	.
Exited		_MAX_	Maximum Absolute Error	0.971963	1	.
Exited		_SSE_	Sum of Squared Errors	1164.735	1241.619	.
Exited		_ASE_	Average Squared Error	0.116497	0.124137	.
Exited		_RASE_	Root Average Squared Error	0.341316	0.352331	.
Exited		_DIV_	Divisor for ASE	9998	10002	.
Exited		_DFT_	Total Degrees of Freedom	4999	.	.

Figure 14

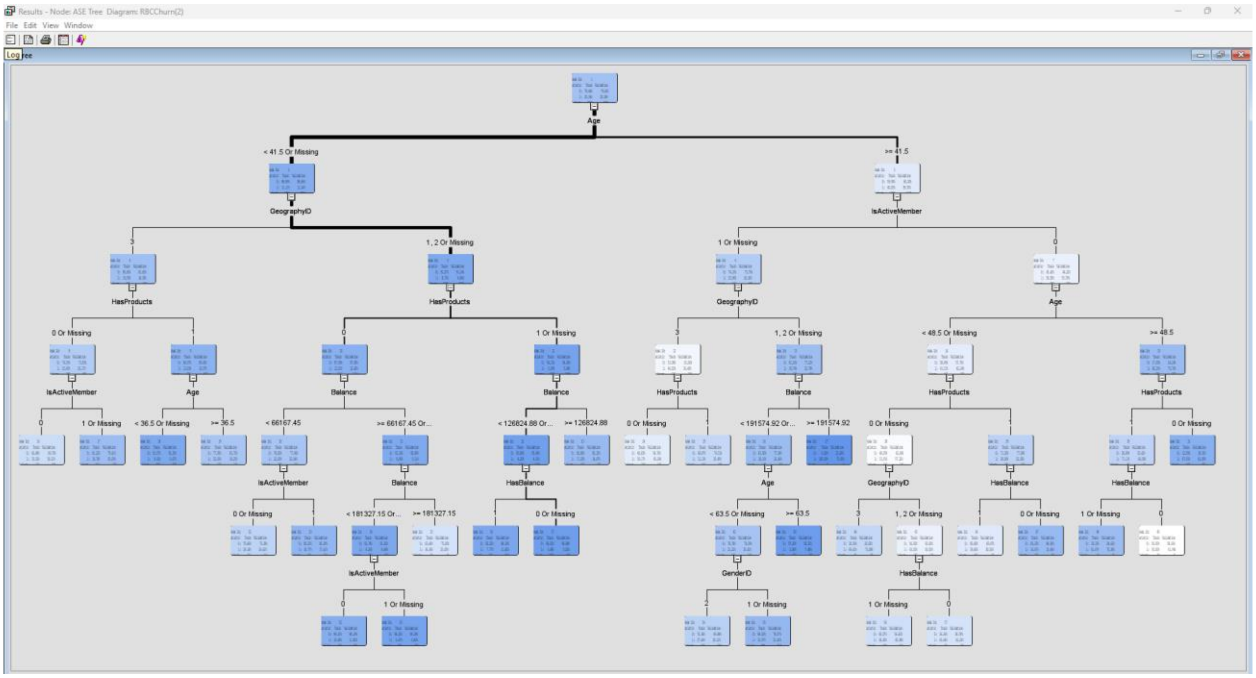


Figure 15

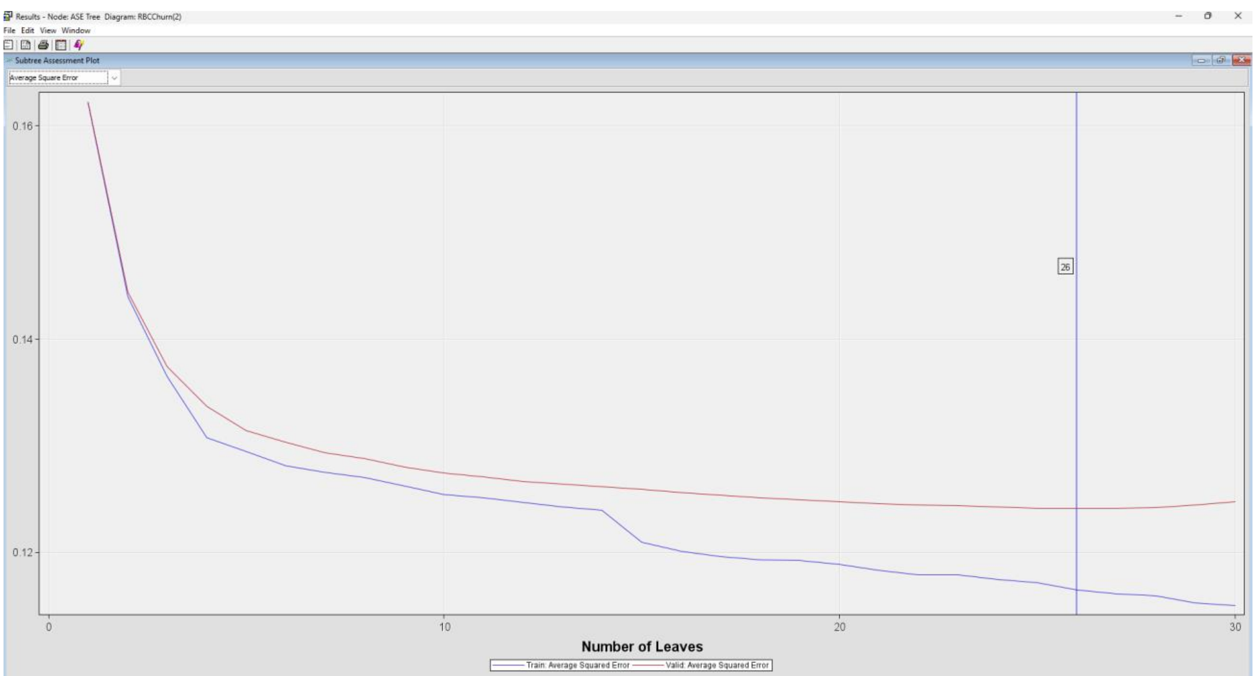


Figure 16

Decision Tree Comparison

Tree Type	ASE	Misclassification Rate	Number of Leaves	Key Splits
Assessment Tree	0.130702	0.163367	9	Age, IsActiveMember, GeographyID, HasProducts
Maximal Tree	0.124772	0.163967	30	Age, IsActiveMember, GeographyID
ASE Tree	0.124137	0.165767	26	Age, IsActiveMember, GeographyID

Conclusion

- The ASE Tree provided the lowest ASE (0.124137), currently making it the most accurate decision tree model.
- The Maximal Tree offered the most detailed splits but risked overfitting.
- The Assessment Tree was simpler and interpretable but had a higher ASE.

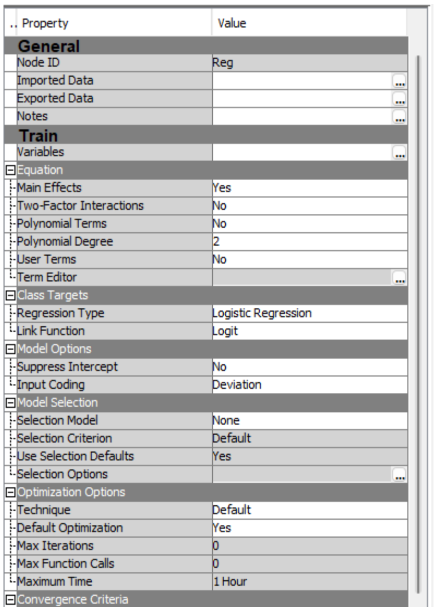
These decision tree models highlight key factors influencing customer churn, such as **Age**, **IsActiveMember**, and **GeographyID**.

5.2 Regressions

After building a decision tree, we developed regression models (i.e., full regression, backward elimination, forward regression, and stepwise regression). These models helped us refine our analysis, validate variable significance, and optimize the model for improved accuracy and interpretability

Full Regression

We brought in the full regression and ran the model. The ASE was 0.129104. We had significant variables with p-values lower than 0.0001 to be **GenderID**, **GeographyID, HasProducts, IsActiveMember,** and **Log_REP_Age.**



Property	Value
General	
Node ID	Reg
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	No
Term Editor	...
Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	None
Selection Criterion	Default
Use Selection Defaults	Yes
Selection Options	...
Optimization Options	
Technique	Default
Default Optimization	Yes
Max Iterations	0
Max Function Calls	0
Maximum Time	1 Hour
Convergence Criteria	

Figure 17

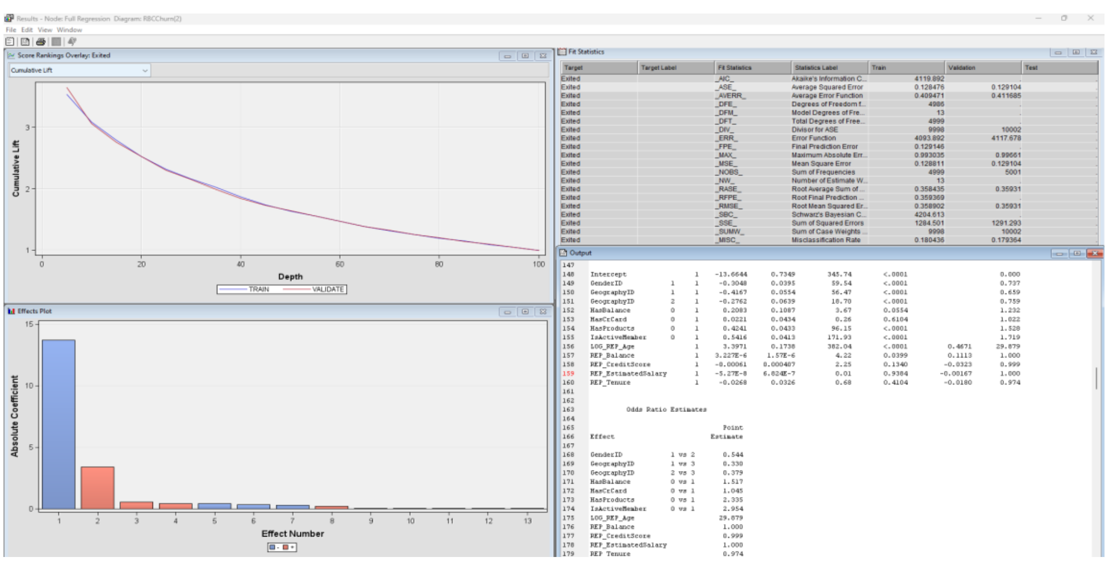


Figure 18

Backward Regression

We modeled backward regression to remove the least significant variable to optimize model complexity. We changed the selection model to “Backward” and the Selection Criterion to “Validation Error” and ran the model. The ASEof the backward regression model was 0.12913.

We had significant variables with p-values lower than 0.0001 to be “GenderID”, “GeographyID”, “Has Products”, “IsActiveMember” and “Log_REP_Age”.

Viewing the points estimate, we determined how these significant variable influenced customer churn:

- GenderID” (1 vs 2) at 0.547, which indicates males are 45.3% less likely to churn compared to females.
- “GeographyID” (1 vs 3) at 0.334, which indicates that customers in France are 66.6% less likely to churn compared to customers in Germany.
- “GeographyID” (2 vs 3) at 0.384, which indicates that customers in Spain are 61.6% less likely to churn when compared to customers in Germany.
- “HasProducts” (0 vs 1) at 2.316, which indicates that people with one product are 131.6% more likely to churn than people with more than one product.
- “IsActiveMember” (0 vs 1) at 2.953, which indicates that customers who are inactive are nearly three times more likely to churn compared to customers that are active,
- “Log_REP_Age” (29.669) indicates that for every 1-unit increase in natural log of age (equivalent to a 2.718 -fold increase in age), customers are approximately 29.7 times more likely to churn. This suggests that older customers are more likely to churn than the younger ones.

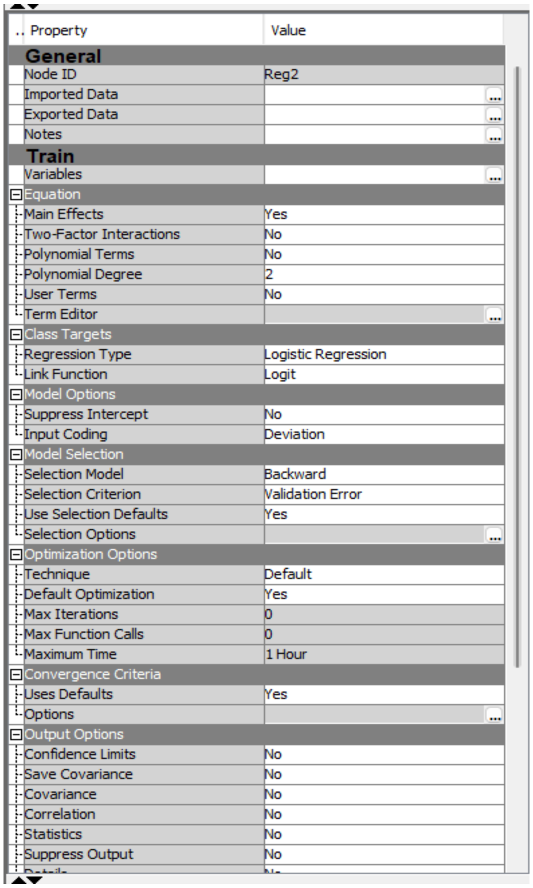


Figure 19

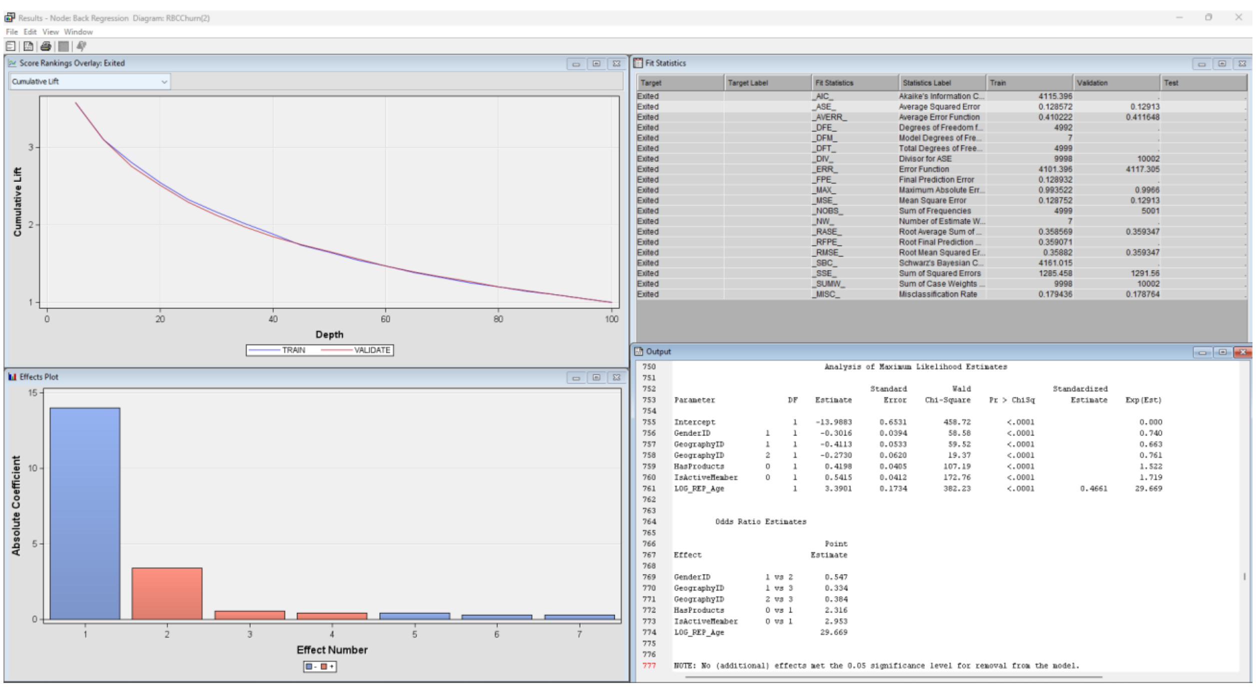
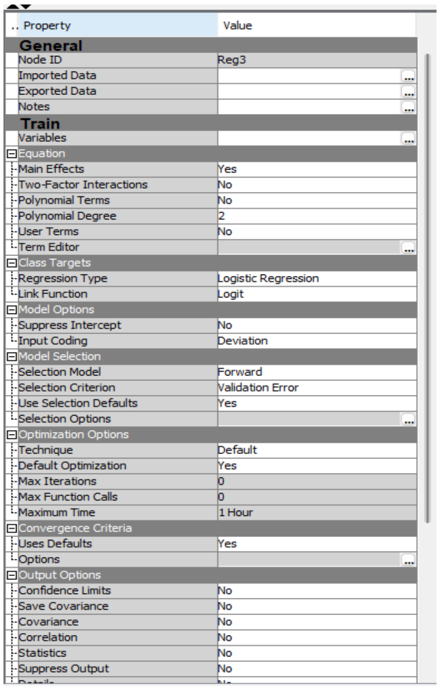


Figure 20

Forward Regression

We utilized forward regression to refine our model by sequentially adding variables based on their statistical significance to optimize model complexity. We changed the selection model to "Forward" and the Selection Criterion to "Validation Error" and ran the model. The ASE for forward regression was 0.12913.

Points estimates and p-values were the same as the backward regression model. For further analysis, please refer to the Backward regression.



Property	Value
General	
Node ID	Reg3
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	No
Term Editor	...
Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	Forward
Selection Criterion	Validation Error
Use Selection Defaults	Yes
Selection Options	...
Optimization Options	
Technique	Default
Default Optimization	Yes
Max Iterations	0
Max Function Calls	0
Maximum Time	1 Hour
Convergence Criteria	
Uses Defaults	Yes
Options	...
Output Options	
Confidence Limits	No
Save Covariance	No
Covariance	No
Correlation	No
Statistics	No
Suppress Output	No

Figure 21

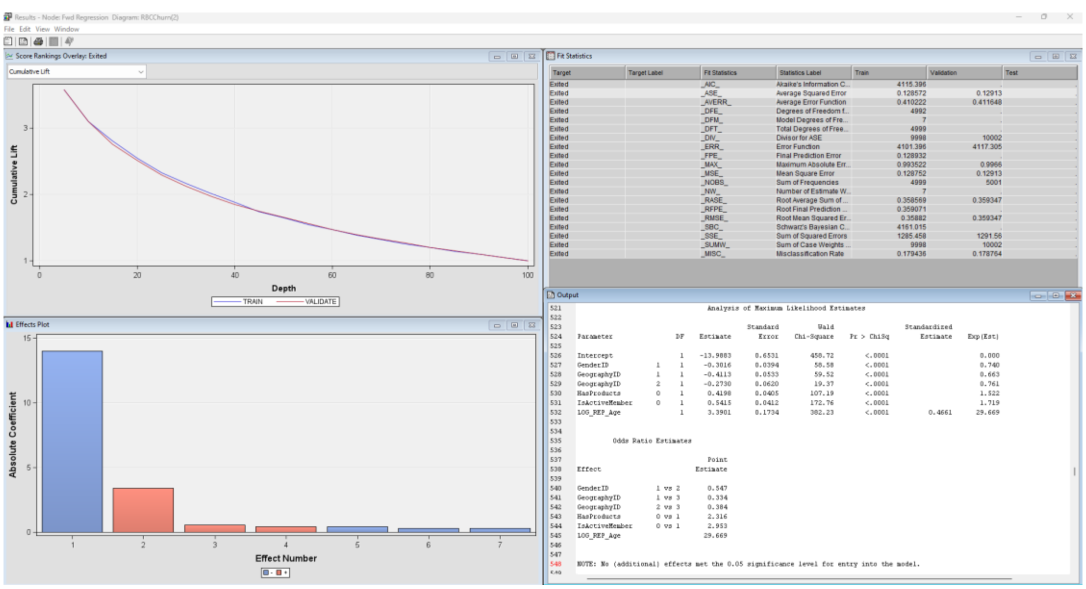


Figure 22

Stepwise

We used stepwise regression to iteratively add and remove variables based on their statistical significance and contribution to the model. This approach allowed us to identify the optimal set of predictors, balancing model complexity, and predictive accuracy.

For the stepwise model, we changed the selection model to "Stepwise" and the Selection Criterion to "Validation Error" and ran the model. The ASE of the stepwise model was 0.12913.

Points estimates and p-values were the same as the backward regression model. For further analysis, please refer to the Backward regression.

Property	Value
General	
Node ID	Reg4
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	No
Term Editor	...
Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	Stepwise
Selection Criterion	Validation Error
Use Selection Defaults	Yes
Selection Options	...
Optimization Options	
Technique	Default
Default Optimization	Yes
Max Iterations	0
Max Function Calls	0
Maximum Time	1 Hour
Convergence Criteria	
Uses Defaults	Yes
Options	...
Output Options	
Confidence Limits	No
Save Covariance	No
Covariance	No
Correlation	No
Statistics	No
Suppress Output	No

Figure 23

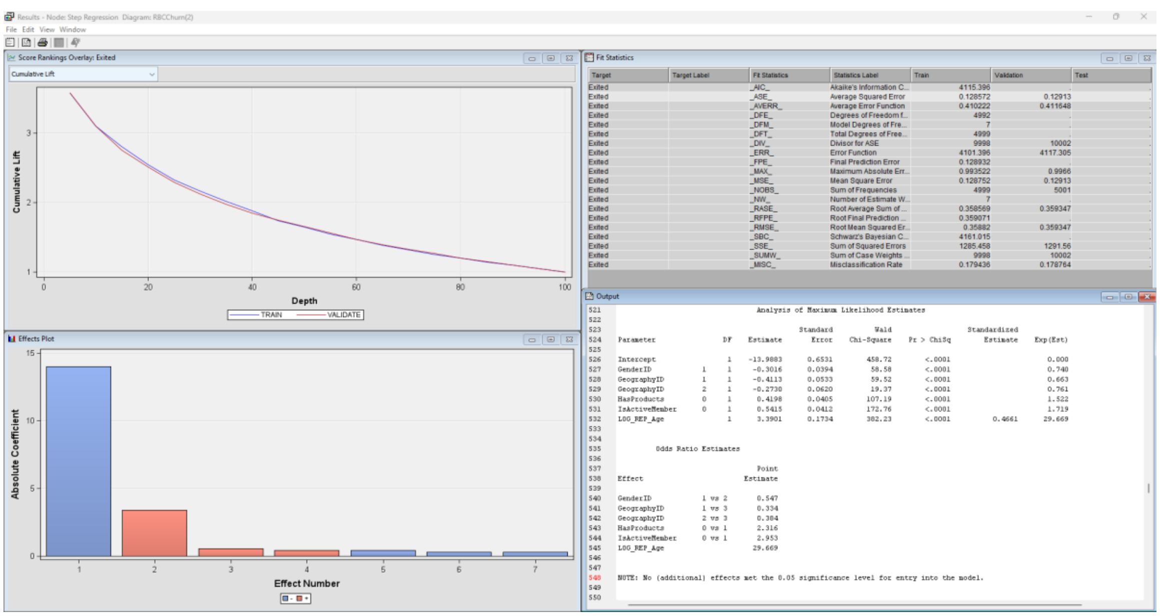


Figure 24

5.3 Neural Networks

We modeled using neural networks to capture complex non-linear relationships between variables and enhance the predictive accuracy of our analysis.

Neural Network: Cap and Floor

We connected the neural network model to the cap and floor node. We changed the Model selection criterion to “Average Error”, and Maximum iterations to “100”. We also disabled the Preliminary training then ran the model.

The model converged at 63 iterations and the ASE was 0.115865.

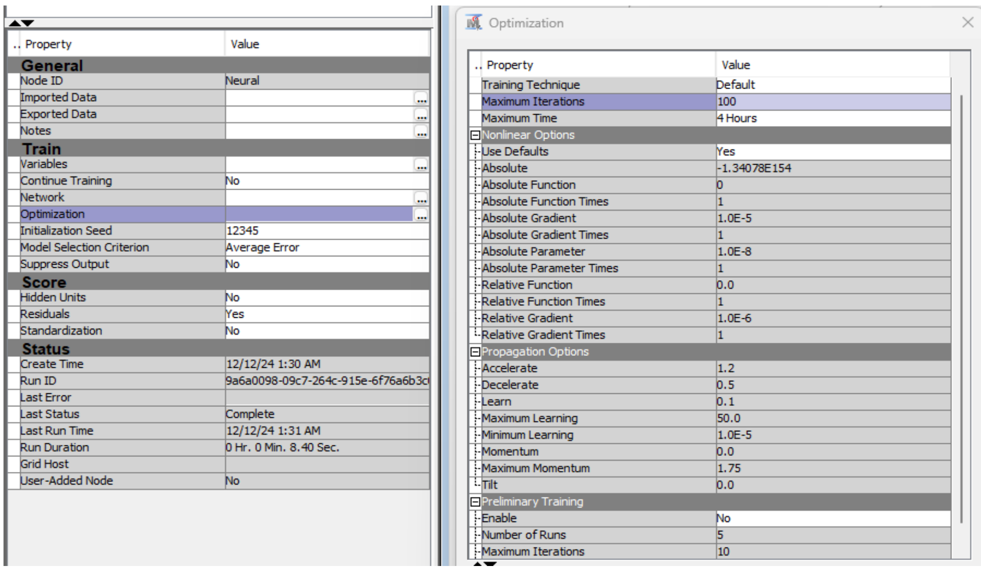


Figure 25

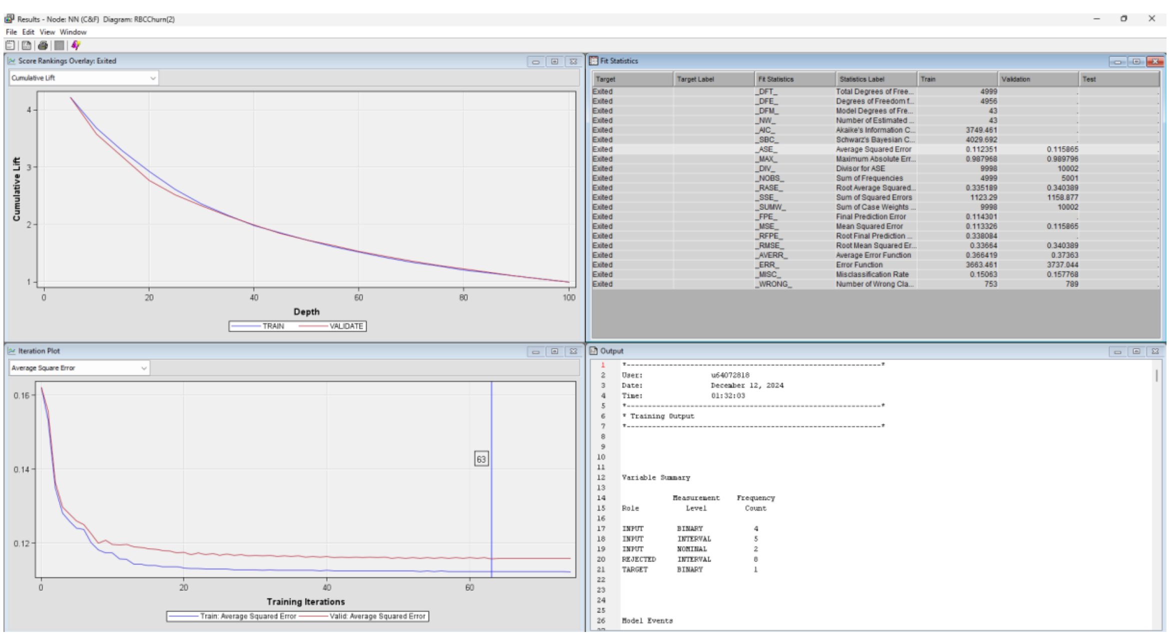


Figure 26

Neural Network: Transform

We connected the neural network model to the transform node and ran the model. We had 30 iterations, and the ASE was 0.121844.

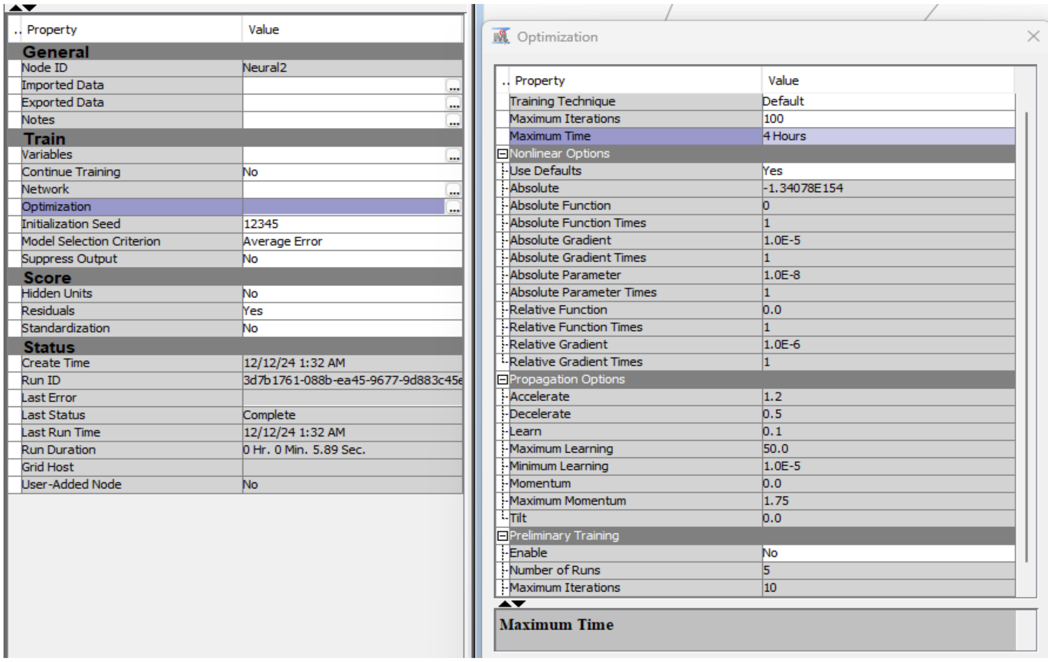


Figure 27

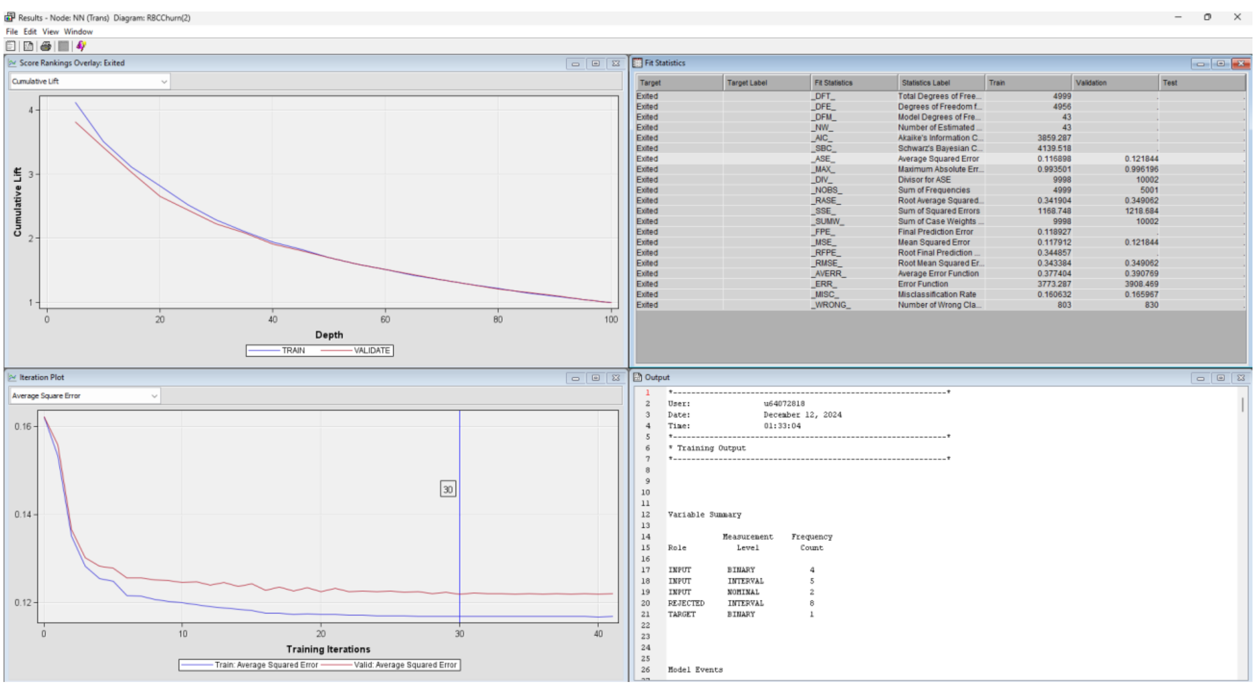


Figure 28

Since the Neural Network: Cap & Floor has the lowest ASE, we decided to run additional neural networks with different hidden units. By default, the neutral network runs with 3 hidden units.

NN Cap and Floor 2H

We changed the number of hidden units to 2 to see if the model will perform better. The model had 48 iterations with an ASE of 0.119794. This was higher than the model with 3 hidden units.

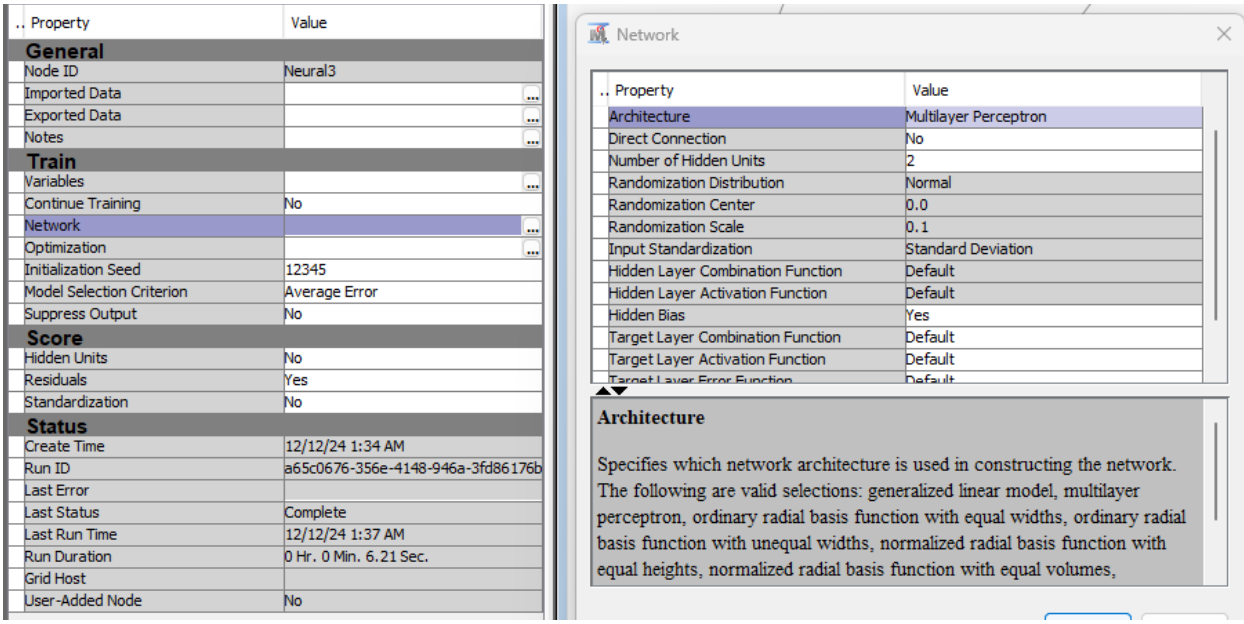


Figure 29

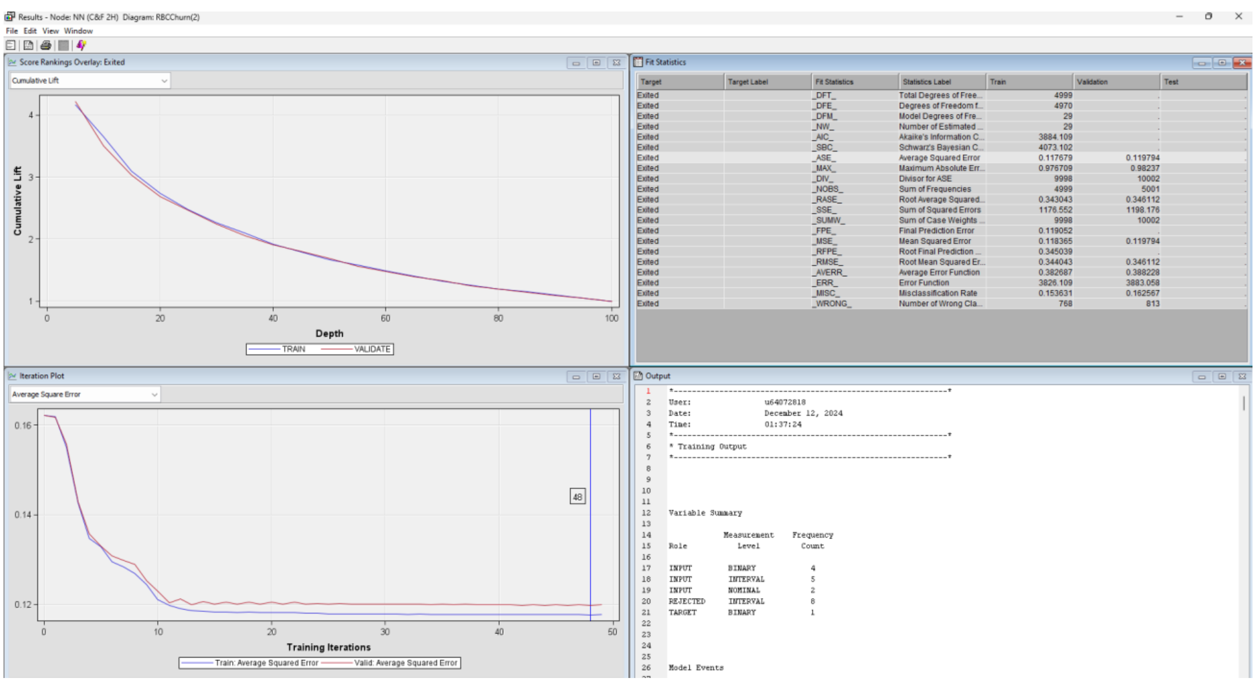


Figure 30

NN Cap and Floor 4H

We changed the number of hidden units to 4 to see if the model will perform better. This model had 66 iterations with an ASE of 0.115123. This was lower than the models with 2 and 3 hidden units.

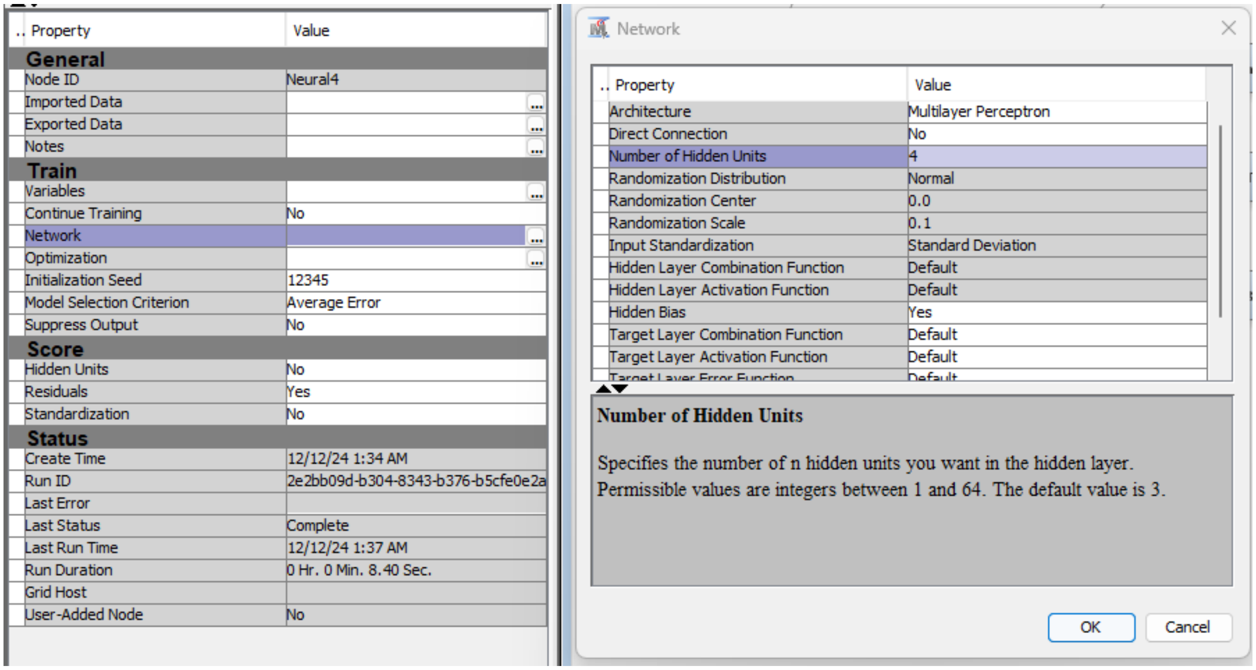


Figure 31

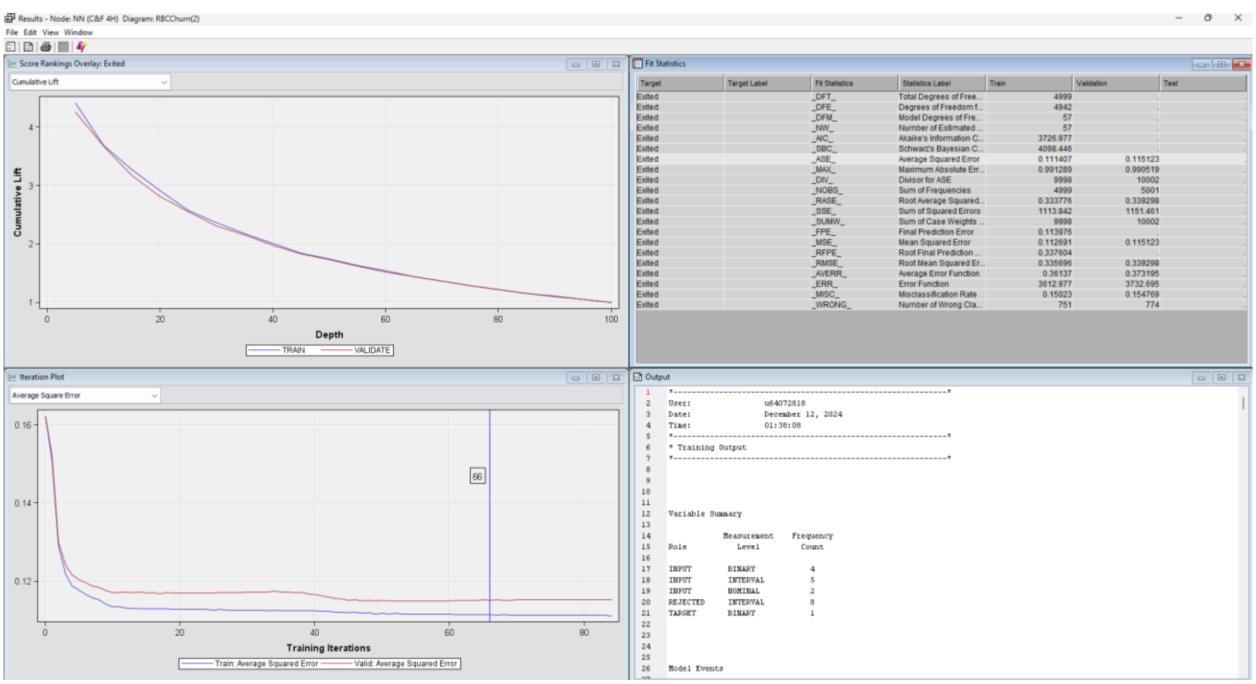


Figure 32

NN Cap and Floor 5H

We changed the number of hidden units to 5 to see if the model will perform better. This model had 20 iterations with an ASE of 0.116445. This was higher than the Cap & Floor with 4 hidden units.

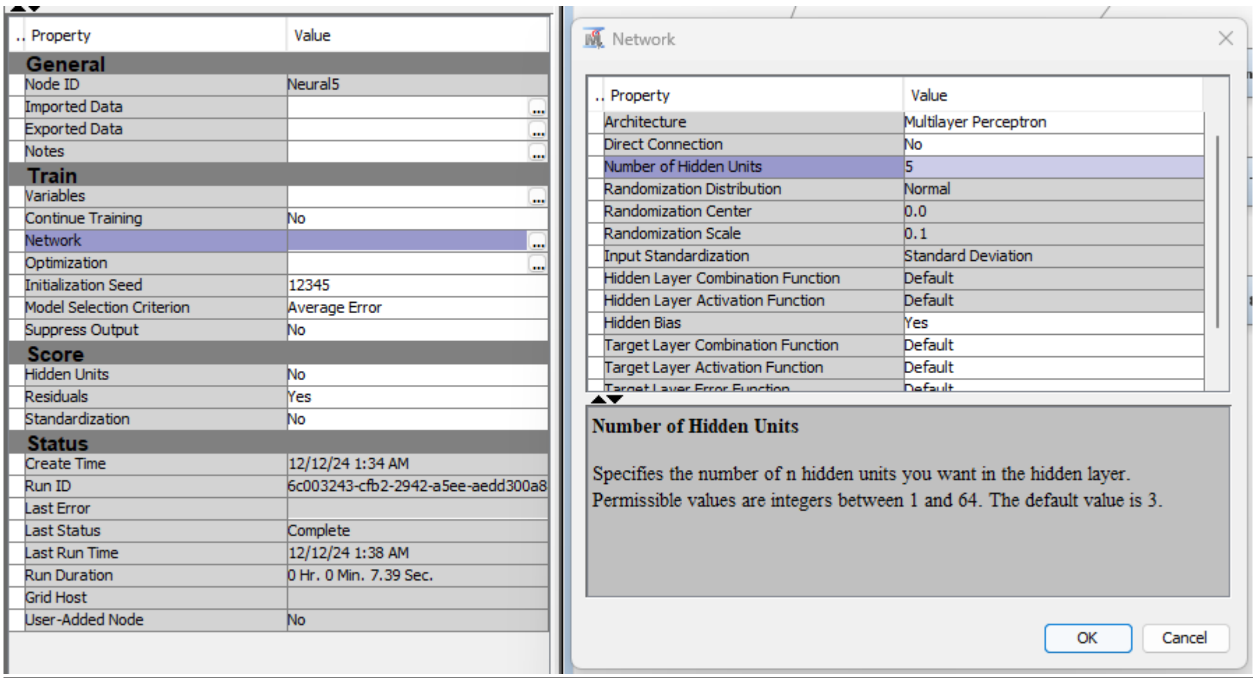


Figure 33

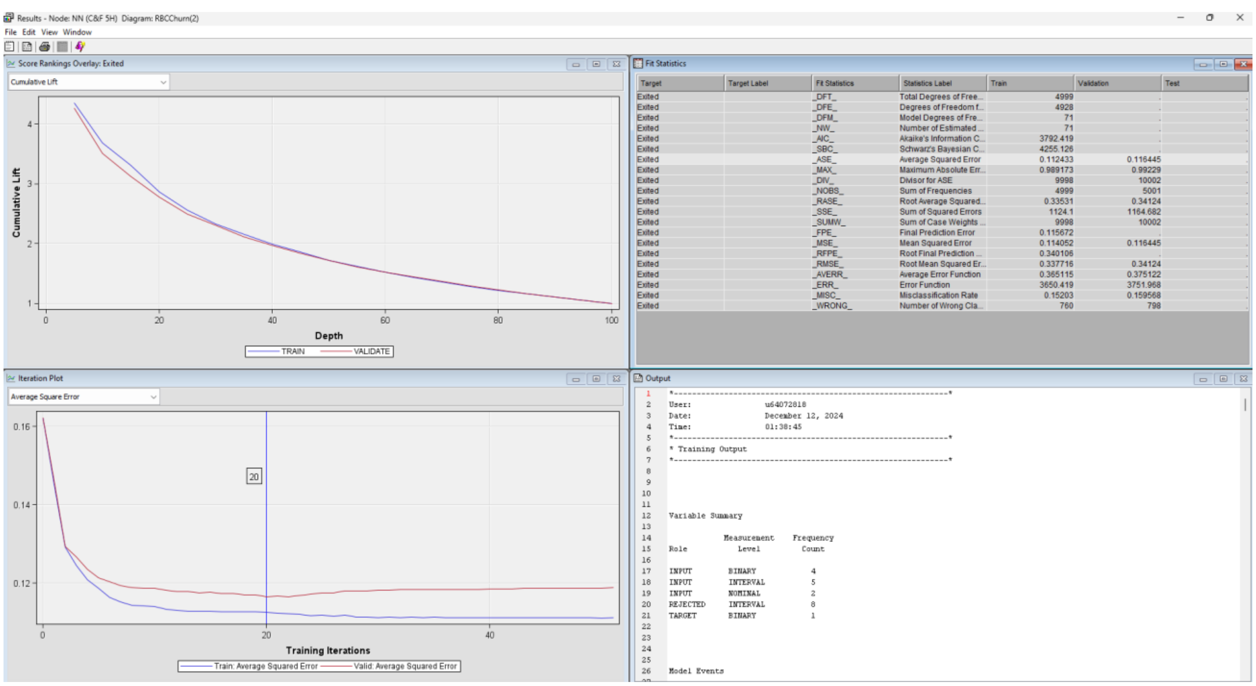


Figure 34

NN Cap and Floor 6H

We changed the number of hidden units to 6 to see if the model will perform better. This model had 11 iterations with an ASE of 0.116965. This was higher than the models with 3, 4 and 5 hidden units but lower than the model with 2 hidden units.

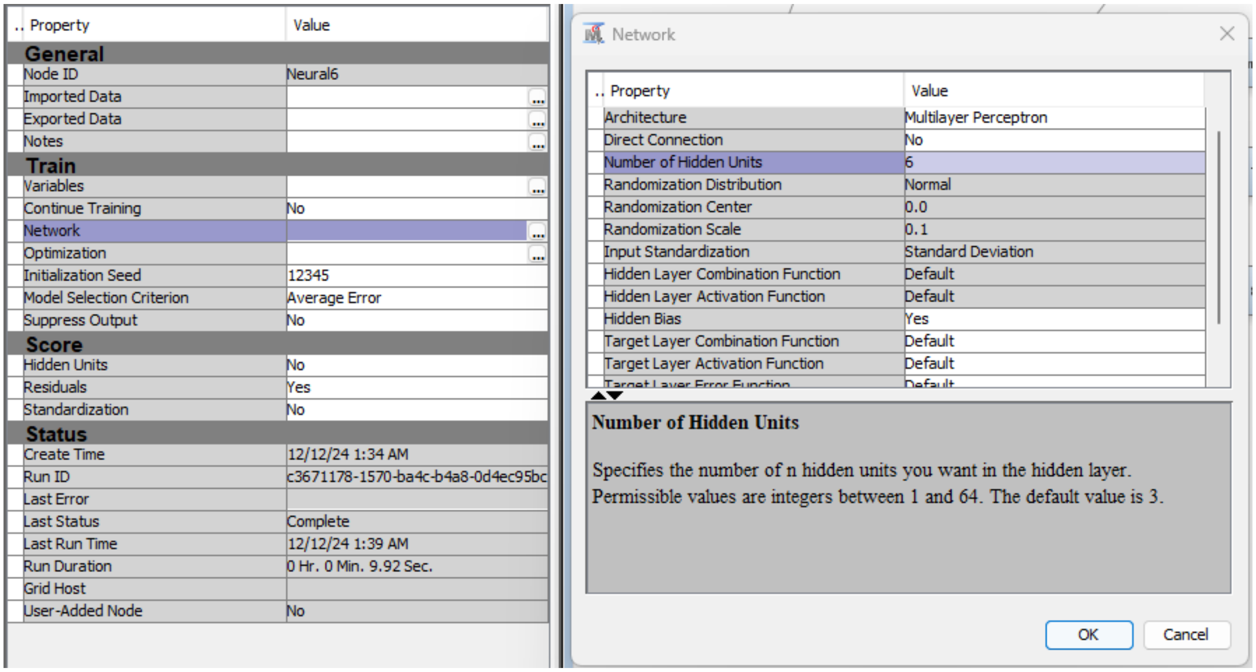


Figure 35

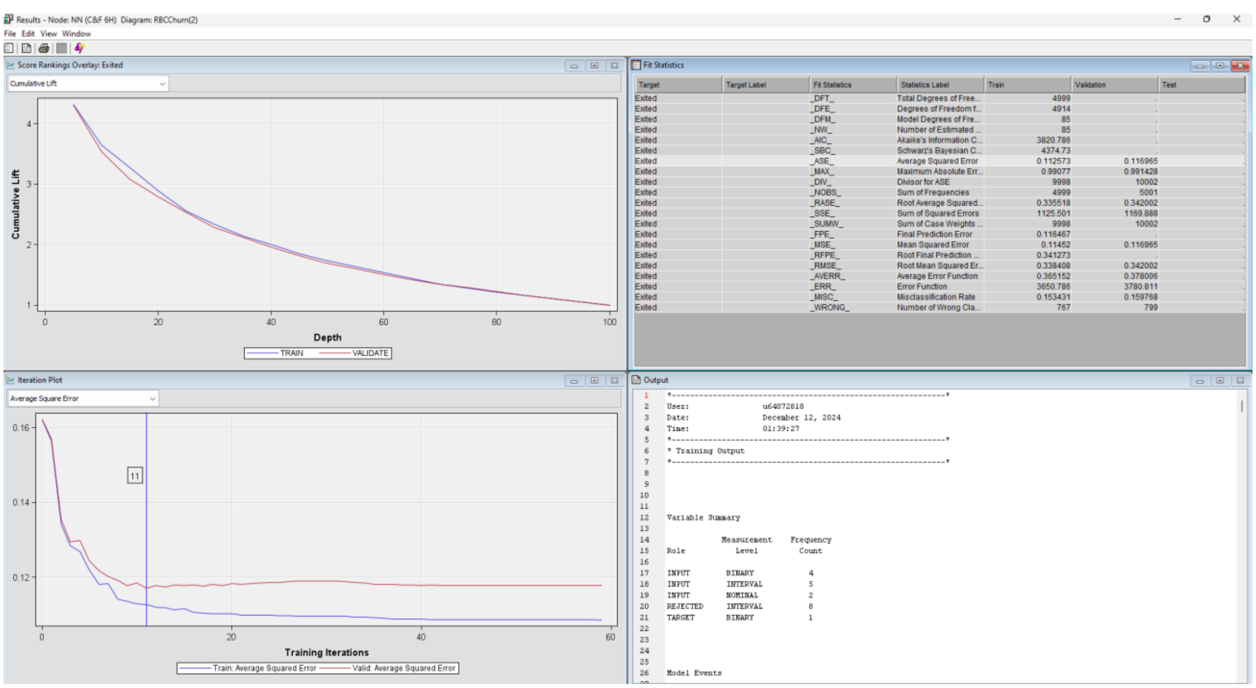


Figure 36

NN Cap and Floor 7H

We changed the number of hidden units to 7 to see if the model will perform better. This model had 29 iterations with an ASE of 0.118729. This was the second highest ASE so far after the model with 2 hidden units.

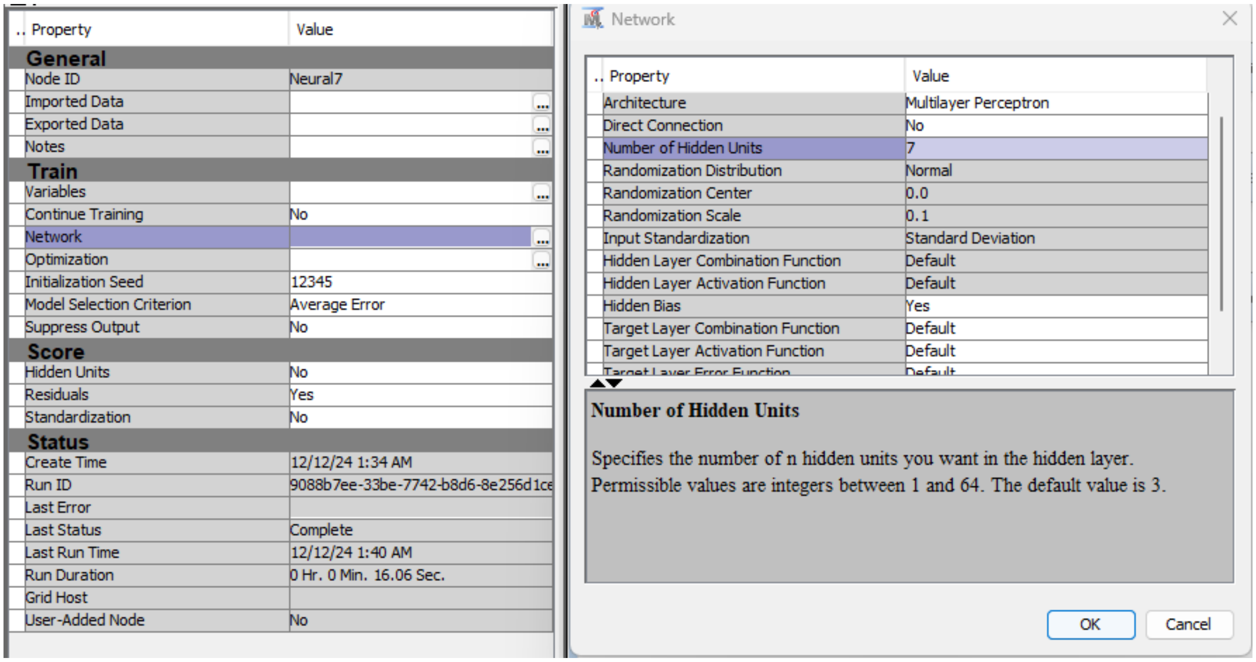


Figure 37

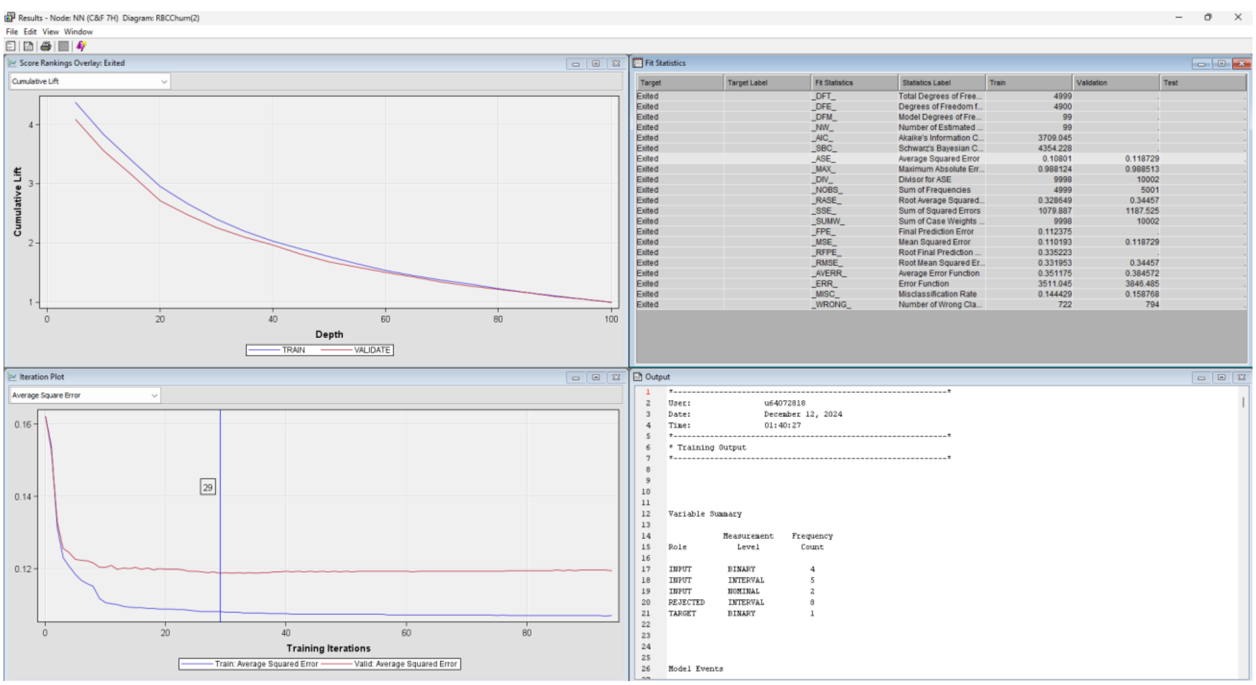


Figure 38

NN Cap and Floor 8H

We changed the number of hidden units to 8 to see if the model will perform better. This model had 21 iterations with an ASE of 0.118568. This was the third highest ASE of the hidden unit Neural Network models.

We decided to stop at 8 hidden units because the ASE kept increasing, therefore the accuracy of the models would likely worsen with each additional hidden unit.

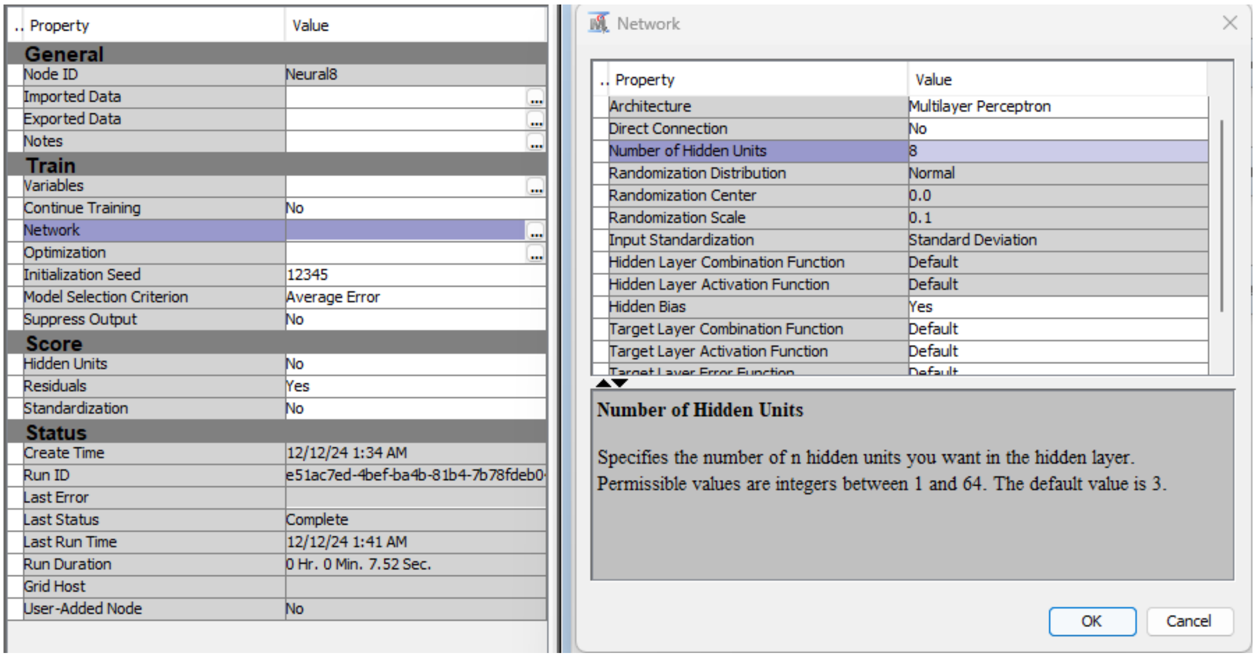


Figure 39

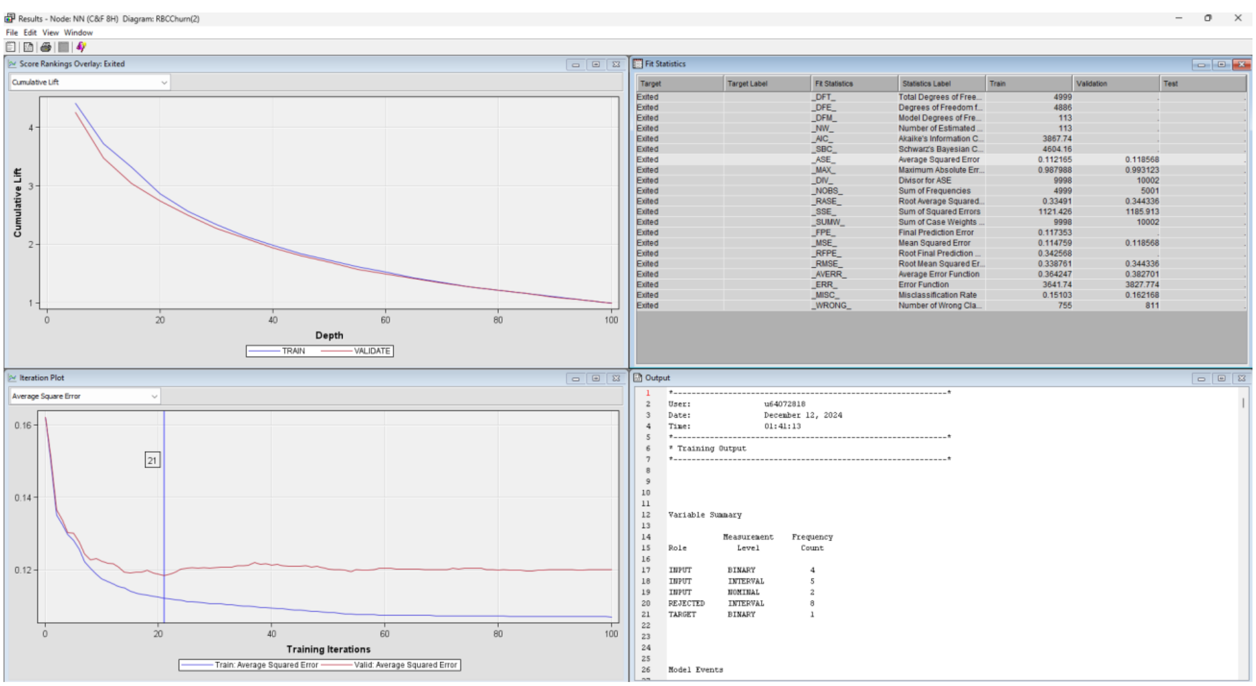


Figure 40

5.4 Model Assessment

We brought in our Model Assessment node and connected it to all our models. We changed the Selection Statistic to “ROC” and the Selection Table “Validation”.

After running the Model Assessment node, the results indicated our best model to be Neural Network: Cap & Floor (3H) which had a ROC value of 0.833.

In the cumulative lift chart, the NN Cap & Floor (6H) model has the best response rate at a depth of 5, with results being 4.301616.

Property	Value
General	
Node ID	MdlComp
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
⊟Assessment Reports	
Number of Bins	20
ROC Chart	Yes
Recompute	No
⊟Model Selection	
Selection Data	Default
Selection Statistic	ROC
HP Selection Statistic	Default
SAS Viya Selection Statistic	...
Selection Table	Validation
Selection Depth	10
Score	
Selection Editor	...
Report	
⊟Selected Model	
Target	Exited
Model Node	Neural
Model Description	NN (C&F)
Selection Criteria	Valid: Roc Index
Status	
Create Time	12/12/24 1:41 AM
Run ID	c9cab0ed-6cc2-ca46-ae7a-586aaec51
Last Error	
Last Status	Complete
Last Run Time	12/12/24 1:43 AM
Run Duration	0 Hr. 0 Min. 15.74 Sec.
Grid Host	
User-Added Node	No

Figure 41

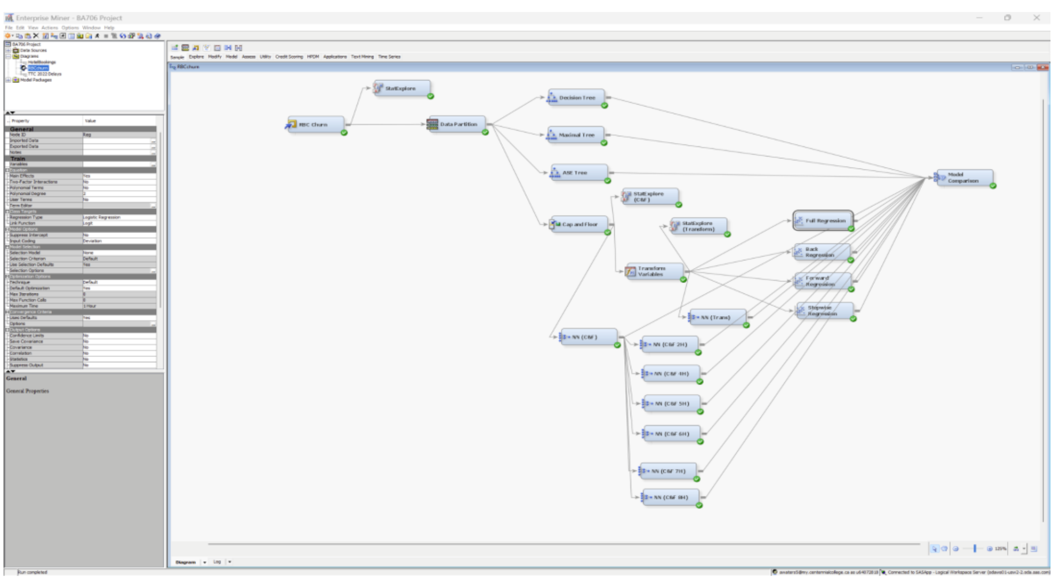


Figure 42

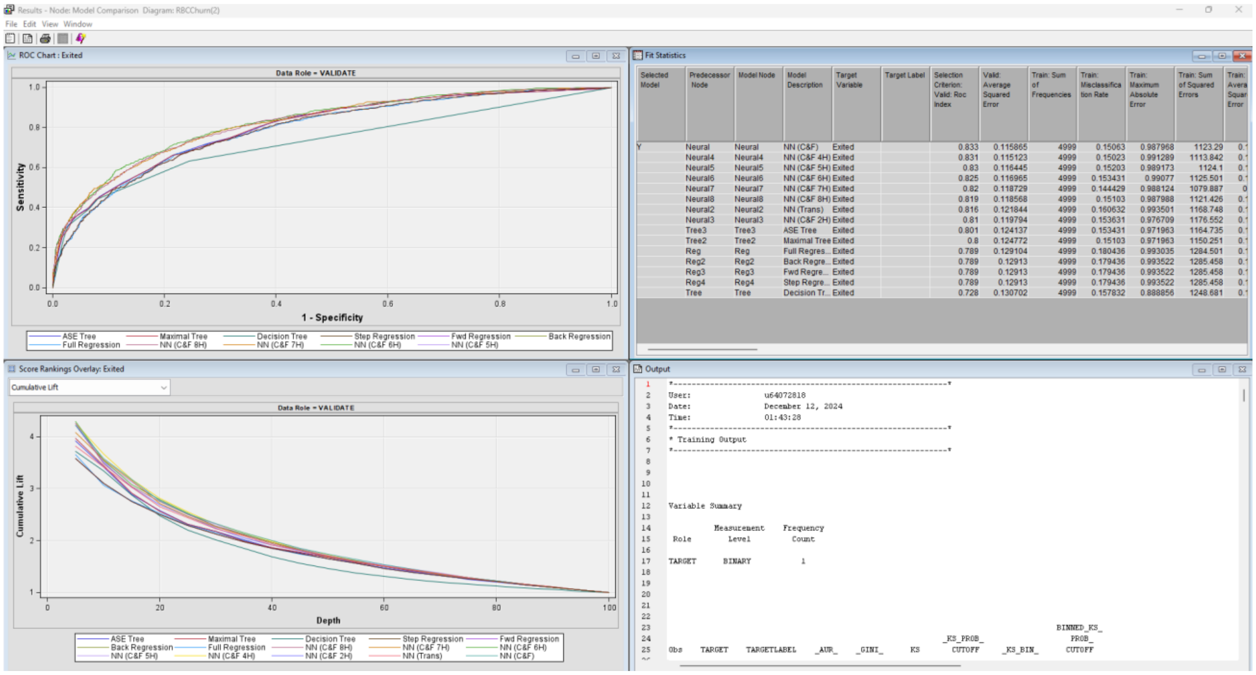


Figure 43

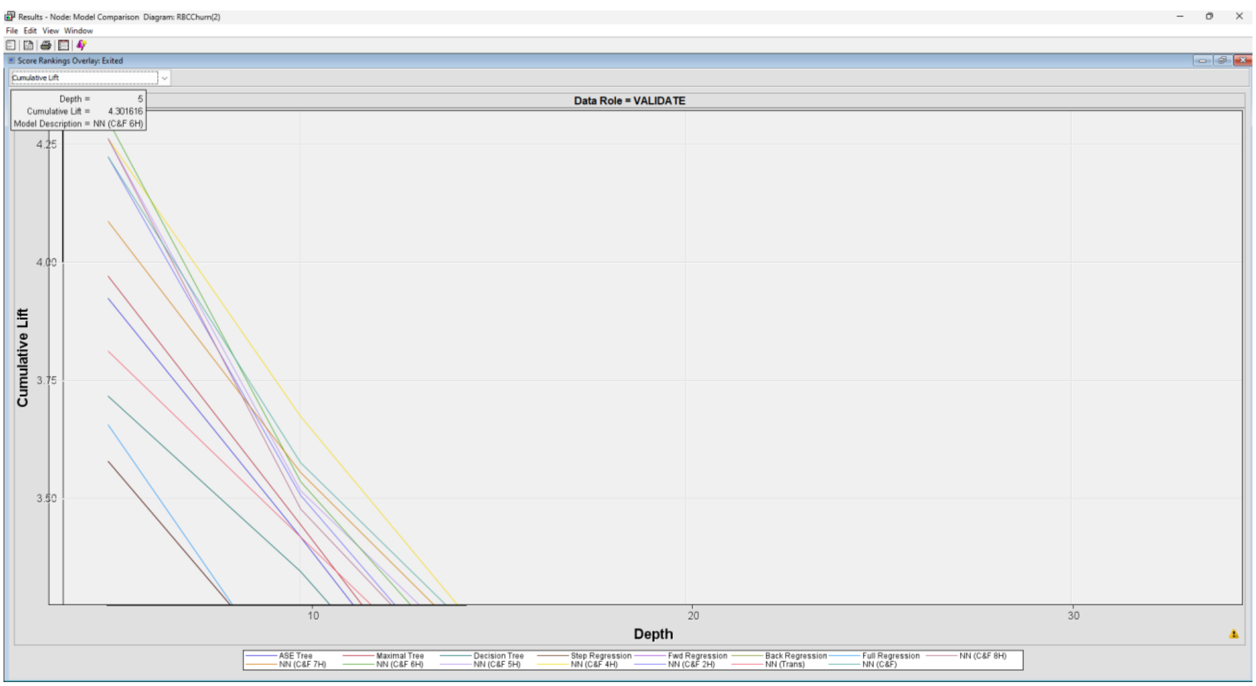


Figure 44

6. Recommendations & Insights

Based on the results of our models, we have identified several key insights and actionable recommendations to reduce customer churn and improve customer retention for RBC.

6.1. Key Insights

Demographic Factors

Gender:

- **Insight**: Males are **45.3% less likely** to churn compared to females.
- **Recommendation**: Develop targeted retention campaigns specifically aimed at female customers, addressing their needs and preferences.

Age:

- **Insight**: Older customers are more likely to churn compared to younger customers.
- **Recommendation**: Implement loyalty programs and personalized offers for older customers to increase engagement and satisfaction.

Geographic Factors

Geography:

- **Insight**: Customers in **France** and **Spain** are significantly less likely to churn compared to those in **Germany**.
- **Recommendations**:
  - Investigate potential reasons for higher churn rates in Germany (e.g., product offerings, customer service quality).
  - Develop region-specific strategies and incentives for German customers to improve retention.

Customer Behavior

Number of Products:

- **Insight**: Customers with only **one product** are significantly more likely to churn.
- **Recommendations**:
  - Encourage customers to adopt multiple products through bundling offers or cross-selling strategies.
  - Provide incentives or discounts for customers who add additional products.

Active Membership Status

- **Insight**: Inactive customers are about **2 times more likely** to churn compared to active members.
- **Recommendations**:
  - Launch engagement campaigns to reactivate dormant accounts.
  - Offer personalized incentives to encourage activity (e.g., cashback, special promotions).

6.2. Model-Specific Insights

From Decision Trees

- Age is consistently the first and most important split in the decision trees.
- **Active Membership Status, GeographyID** and **HasProducts** are key factors influencing churn.

Recommendation: Focus retention efforts on younger customers with fewer products and inactive accounts.

From Regression Models

- **Significant Predictors:**
  - **GenderID, GeographyID, HasProducts, IsActiveMember, and Log_REP_Age.**

Recommendation:

- Target customers based on these predictors for personalized retention strategies.
- Use regression insights to understand customer segments and tailor communication accordingly.

From Neural Networks

- **The Neural Network (Cap & Floor)** model achieved the best performance with the highest **ROC of 0.833** and second lowest **ASE of 0.115865.**
- The model identified complex non-linear relationships between features.

Recommendations:

- Leverage the neural network insights to build advanced churn prediction tools.
- Use these tools for real-time churn risk assessments and proactive customer interventions.

6.3. Actionable Recommendations

6.3.1. Offer Diverse Product Sets

Older customers are at a higher risk of churning, often due to a lack of products that cater to their evolving financial needs. As customers age, their priorities shift towards retirement planning, investment growth, and wealth preservation, which may not be addressed by standard product offerings.

Recommendation

RBC should introduce personalized product bundles tailored specifically to older customers. These bundles should address their financial goals at different life stages, such as retirement, estate planning, and wealth management. By offering comprehensive solutions that evolve with their needs, RBC can foster long-term relationships and reduce churn.

Strategies to Implement

1. **Segment Analysis**:
    - Conduct a thorough analysis to identify different segments within the older customer demographic (e.g., pre-retirement, early retirement, and late retirement).
    - Use data analytics to understand their financial behaviors, product preferences, and pain points.
2. **Personalized Bundling**:
    - Develop product bundles tailored to each segment, such as:
        - **"Retirement Ready Bundle"**: Includes retirement savings accounts, annuities, and estate planning services.
        - **"Investment Growth Plan"**: Combines investment portfolios, wealth advisory services, and tax-efficient investment strategies.
        - **"Legacy Protection Package"**: Focuses on wills, trusts, and life insurance products.
3. **Educational Campaigns**:
    - Launch educational initiatives like webinars, workshops, and personalized consultations to help older customers understand the benefits of these products.
    - Develop resources (blogs, e-books, and videos) on topics such as retirement planning, wealth preservation, and legacy planning.
4. **Exclusive Incentives**:
    - Offer preferential rates on mortgages, reduced fees on investment services, and cashback for adopting bundled products.
    - Create loyalty programs that reward older customers for engaging with multiple RBC products.
5. **Feedback Mechanisms**:
    - Regularly collect feedback through surveys and focus groups to refine and enhance these bundles based on customer needs.

6.3.2. Gender-Specific Strategies

Female customers are **42.5% more likely** to churn compared to male customers. This indicates a need for more targeted engagement strategies that address their unique financial needs, preferences, and experiences.

Recommendation

RBC should design and implement marketing campaigns and financial services tailored to female customers. These initiatives should focus on building trust, offering personalized solutions, and empowering women to achieve their financial goals.

Strategies to Implement

1. **Financial Empowerment Workshops**:
    - Host workshops and webinars specifically for women, covering topics such as:
        - Investment planning and portfolio management.
        - Budgeting, savings strategies, and debt management.
        - Retirement planning and financial independence.
2. **Personalized Communication**:
    - Use data-driven insights to create personalized communication strategies.
    - Send tailored messages highlighting products and services that resonate with female customers, such as investment opportunities, savings plans, and family-oriented financial products.
3. **Success Stories and Case Studies**:
    - Share testimonials and case studies featuring successful female clients.
    - Highlight how RBC's services have helped women achieve their financial goals, reinforcing RBC's commitment to supporting female customers.
4. **Dedicated Relationship Managers**:
    - Assign relationship managers who specialize in understanding and addressing the needs of female clients.
    - Train staff to be sensitive to the unique financial challenges women may face, such as career breaks, wage gaps, and caregiving responsibilities.
5. **Women-Focused Financial Products**:
    - Develop products designed with women in mind, such as flexible investment plans, joint savings accounts for families, and financial planning services for working mothers.

6.3.3. Customer Engagement Initiatives

Inactive customers are twice as likely to churn compared to active customers. This highlights the importance of proactively re-engaging dormant customers to rekindle their interest in RBC's services.

Recommendation

RBC should develop targeted reactivation campaigns to encourage inactive customers to become active users again. Personalized incentives, timely communication, and value-driven offers can reignite customer engagement.

Strategies to Implement

1. **Cashback and Rewards Programs**:
    - Introduce cashback offers or loyalty rewards for completing specific transactions, such as making deposits, applying for a credit card, or using online banking services.
    - Create limited-time reward campaigns to create a sense of urgency and excitement.
2. **Exclusive Time-Sensitive Offers**:
    - Provide promotions like discounted loan rates, waived fees, or bonus interest on savings accounts for a limited period.
    - Tailor offers to each customer's historical behavior to maximize relevance.
3. **Personalized Outreach**:
    - Use personalized emails, SMS, or phone calls to reach out to inactive customers.
    - Craft messages that acknowledge their inactivity and present solutions or incentives to re-engage, such as "We've missed you! Here's an exclusive offer to welcome you back."
4. **Feedback Collection**:
    - Reach out to inactive customers to understand the reasons behind their disengagement.
    - Use surveys and one-on-one calls to gather insights and identify areas for service improvement.
5. **Re-Engagement Campaigns**:
    - Launch campaigns with themes like "Welcome Back" or "Reconnect with RBC" to make customers feel valued and appreciated.
    - Offer personalized product recommendations based on their previous engagement history.

6.3.4. Pilot Programs and A/B Testing

Implementing new strategies at scale without testing can lead to inefficient use of resources and suboptimal outcomes. Testing ensures that only the most effective initiatives are rolled out widely.

Recommendation

RBC should conduct pilot programs and A/B testing to evaluate the effectiveness of new strategies before full-scale implementation. This method minimizes risk, optimizes resource allocation, and improves the likelihood of success.

Strategies to Implement

1. **Controlled Experiments**:
    - Select a representative sample of customers to test new initiatives such as product bundles, marketing campaigns, or engagement tactics.
    - Create control and test groups to compare outcomes objectively.
2. **Key Metrics Tracking**:
    - Define and track critical metrics such as churn rate, customer engagement, product adoption rates, and return on investment (ROI).
    - Analyze performance data to determine which strategies yield the best results.
3. **Iterative Refinements**:
    - Based on the results of A/B tests, refine and adjust strategies to improve effectiveness.
    - Implement a continuous improvement loop to ensure ongoing optimization.
4. **Documentation and Reporting**:
    - Maintain detailed documentation of the testing process, results, and insights gained.
    - Share reports with stakeholders to facilitate informed decision-making and ensure transparency.
5. **Scaling Successful Strategies**:
    - Once a strategy proves successful in pilot testing, roll it out to the broader customer base with confidence.
    - Develop implementation guidelines to ensure consistency and effectiveness across all regions.

6.3.5. Product Bundling

Customers who hold only a single product are significantly more likely to churn. Encouraging customers to adopt multiple products enhances their engagement, satisfaction, and loyalty, making them less likely to leave.

Recommendation

RBC should implement comprehensive product bundling strategies to incentivize customers to adopt multiple financial products. Bundling services such as **savings accounts, credit cards, mortgages, investment services, and insurance products** creates more value for customers and strengthens their relationship with RBC.

Strategies to Implement

1. **Service Bundling**:
- Offer bundled packages that combine complementary services, such as:
  - **"Everyday Banking Bundle"**: Savings account + credit card + online banking.
  - **"Family Financial Bundle"**: Joint savings account + children's education savings plan + family insurance.
  - **"Home Ownership Bundle"**: Mortgage + home insurance + home equity line of credit.

2. **Cross-Selling Campaigns**:
- Launch targeted cross-selling campaigns that recommend additional products based on customers' current holdings.
- Use personalized communication to highlight how adding products can benefit customers (e.g., "You already have a savings account! Add a credit card and enjoy cashback rewards on purchases.").

3. **Incentives and Discounts**:
- Offer discounts or perks for customers who adopt multiple products, such as:
  - **Fee Waivers**: Waive monthly fees for customers who bundle at least three services.
  - **Cashback Rewards**: Provide cashback on purchases made with bundled products.
  - **Interest Rate Benefits**: Offer lower mortgage rates or higher savings rates for bundled customers.

4. **Promotional Campaigns**:
- Develop seasonal or limited-time promotions to encourage customers to adopt bundles (e.g., "Bundle now and receive a $100 bonus!").
- Highlight these promotions through emails, SMS, RBC's website, and social media.

5. **Customer Education**:
- Educate customers on the benefits of product bundling through workshops, webinars, and one-on-one consultations.
- Share success stories and testimonials from customers who have benefited from bundled services.

6. **Tracking and Personalization**:
- Use customer data to identify which bundles are most relevant to specific segments.
- Implement AI-driven personalization to recommend the right bundle at the right time.

6.3.6. Data-Driven Monitoring

Predictive models, especially neural networks, provide powerful tools for identifying at-risk customers. Continuous monitoring of churn risk allows RBC to intervene in real time and proactively prevent customer attrition.

Recommendation

Leverage the capabilities of the **Neural Network model** to establish a robust data-driven monitoring system. This system should provide ongoing insights into churn risk, offer early warnings, and facilitate timely interventions to retain at-risk customers.

Strategies to Implement

1. **Continuous Risk Monitoring**:
    - Integrate the neural network model into RBC's customer relationship management (CRM) system to monitor churn risk continuously.
    - Automate daily or weekly risk assessments to keep track of changes in customer behavior and churn likelihood.
2. **Early Warning System**:
    - Develop an early warning system that flags customers who exhibit high churn risk based on predictive scores.
    - Set thresholds for churn probability (e.g., customers with a churn risk above **70%**) to trigger alerts for immediate follow-up by relationship managers.
3. **Real-Time Intervention Strategies**:
    - Implement real-time intervention strategies such as personalized offers, engagement calls, and targeted emails to retain high-risk customers.
    - For example, if a customer shows signs of disengagement (e.g., reduced account activity), automatically send a personalized offer like "Enjoy a $50 reward for your next transaction."
4. **Dynamic Customer Segmentation**:
    - Continuously update customer segments based on their churn risk scores.
    - Create dynamic profiles that help RBC understand which segments are most vulnerable and tailor interventions accordingly.
5. **Dashboard and Reporting Tools**:
    - Develop interactive dashboards to visualize churn risk data, trends, and intervention outcomes.
    - Provide regular reports to stakeholders highlighting key insights, intervention success rates, and areas for improvement.
6. **Feedback Loop**:
    - Establish a feedback mechanism to evaluate the effectiveness of interventions.
    - Use insights from successful and unsuccessful interventions to continuously refine the monitoring and response processes.
7. **Integration with Customer Support**:
    - Equip customer support teams with churn risk data to personalize interactions.
    - Train support agents to recognize churn risk indicators and respond with empathy and appropriate solutions.

6.4. Strategic Implementation Plan

Action	Timeline	Responsibility	Key Metrics
Launch gender-specific campaigns	3 months	Marketing Team	Churn rate by gender
Develop age-based loyalty offers	3-6 months	Customer Experience Team	Retention rate among older customers
Implement region-specific plans	6 months	Regional Managers	Churn rate in Germany
Product bundling offers	3 months	Product Management Team	Number of multi-product customers
Reactivation campaigns	3 months	Customer Engagement Team	Activity rate among dormant accounts
Neural network deployment	6-12 months	Data Science Team	Accuracy and ROC of churn predictions

7. Conclusion

To effectively reduce churn, RBC must adopt a **comprehensive, data-driven approach** that focuses on personalized solutions, targeted engagement, and strategic testing. By leveraging insights from predictive models, RBC can:

1. **Engage Older Customers**: Offer diverse product sets that cater to the unique needs of older demographics, such as bundled mortgages, investment services, and retirement planning tools.
2. **Support Female Customers**: Develop gender-specific strategies that foster trust, financial empowerment, and personalized service for female clients.
3. **Reactivate Inactive Customers**: Launch targeted campaigns to incentivize dormant customers to re-engage with RBC services.
4. **Test and Refine Strategies**: Implement pilot programs and A/B testing to ensure that initiatives are effective and resource efficient.

By addressing the specific needs of different customer segments and continuously refining strategies based on data, RBC can improve customer satisfaction, enhance loyalty, and drive long-term business growth. This proactive approach not only reduces churn but also strengthens RBC's position as a customer-centric financial institution.

References

Bailyn, E. (2024, August 1). Average customer acquisition cost (CAC) in banking. First Page Sage. https://firstpagesage.com/seo-blog/average-customer-acquisition-cost-cac-in-banking/