

Predictive Modeling Homework

Churn Prediction

Merve Pakcan

Introduction

Customer churn prediction is a crucial challenge in many industries, particularly in telecommunications, finance, and e-commerce, where retaining existing customers is more cost-effective than acquiring new ones. This project focuses on building a predictive model to determine whether a customer will discontinue using a company's services based on their past interactions and usage patterns.

Using the **Customer Churn Prediction** dataset from **Kaggle**, the goal is to uncover key drivers behind customer attrition and develop a robust classification model using **SAS**. The dataset comprises **7043 customer records with 21 features**, offering insights into customer behavior, demographics, and service engagement.

By identifying customers at high risk of churn, businesses can implement targeted retention strategies, enhance customer satisfaction, and ultimately reduce churn rates. This project illustrates a practical and real-world application of machine learning in customer analytics, leveraging data-driven decisions to improve business outcomes.

Description of Variables

The dataset includes the following 21 variables:

Variable	Description	Type
customerID	Unique identifier for each customer	Character
gender	Customer's gender (Male, Female)	Character
SeniorCitizen	Indicates if customer is a senior citizen (0 = No, 1 = Yes)	Numeric
Partner	Whether the customer has a partner (Yes/No)	Character
Dependents	Whether the customer has dependents (Yes/No)	Character
tenure	Number of months with the company	Numeric
PhoneService	Whether the customer has phone service	Character

MultipleLines	Whether the customer has multiple phone lines	<i>Character</i>
InternetService	Type of internet service (DSL, Fiber optic, or No)	<i>Character</i>
OnlineSecurity	Subscribed to online security service	<i>Character</i>
OnlineBackup	Subscribed to online backup service	<i>Character</i>
DeviceProtection	Customer has device protection	<i>Character</i>
TechSupport	Customer has technical support	<i>Character</i>
StreamingTV	Customer uses streaming TV services	<i>Character</i>
StreamingMovies	Customer uses streaming movie services	<i>Character</i>
Contract	Type of contract (Month-to-month, One year, Two year)	<i>Character</i>
PaperlessBilling	Uses paperless billing	<i>Character</i>
PaymentMethod	Payment method (credit card, bank transfer)	<i>Character</i>
MonthlyCharges	Monthly charges billed	<i>Numeric</i>
TotalCharges	Total charges accumulated	<i>Numeric</i>
Churn	Target: Churned (Yes/No)	<i>Character</i>

Measure

- Frequency
- MonthlyCharges
- SeniorCitizen
- tenure
- TotalCharges

The variables I selected for initial frequency analysis MonthlyCharges, Contract, PaymentMethod, and Churn are considered among the most important in predicting customer churn. MonthlyCharges directly reflects the financial burden on the customer and may influence satisfaction or cancellation decisions. The Contract type is highly indicative of customer commitment; for instance, month-to-month contracts typically imply higher churn risk. PaymentMethod can also reveal patterns in churn behavior, such as customers using electronic checks potentially being more likely to churn. Finally, Churn is the target variable, and analyzing its distribution is essential to understand class balance and overall problem structure. These variables together provide strong early insights into customer behavior and retention risk.

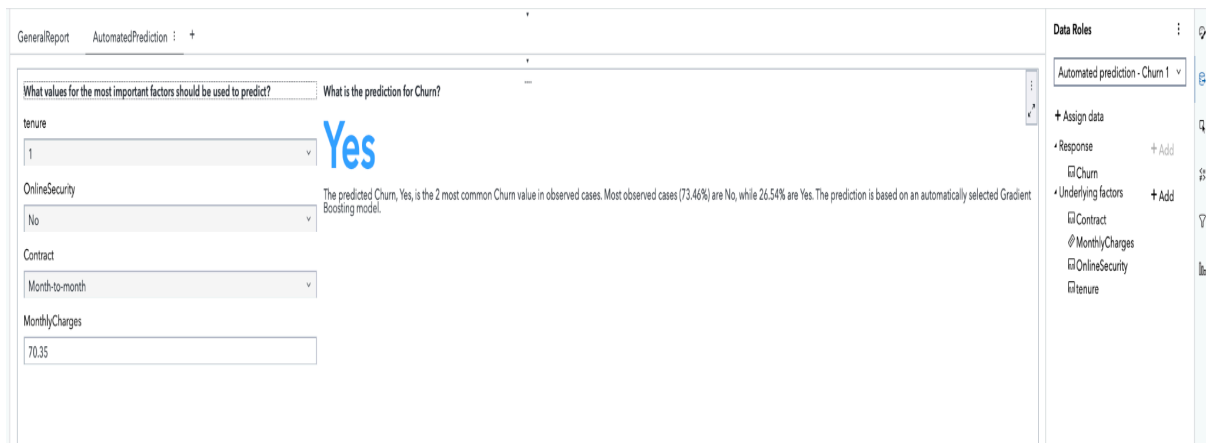
General Report





In this project, I grouped the numeric variable `MonthlyCharges` into three meaningful categories Low, Medium, and High by first keeping it as a measure and then using the New data item *Custom Category* feature in SAS Visual Analytics. This method is more appropriate for continuous variables with a wide range of values, as it allows for flexible and precise range definitions. In contrast, the professor demonstrated a similar grouping process for the `Age` variable, which was already categorical or had a limited set of integer values. While manually assigning groups to such a variable is practical, doing the same for `MonthlyCharges` which contains many decimal values would be inefficient and error-prone. Therefore, maintaining `MonthlyCharges` as numeric and creating range-based categories ensures both analytical accuracy and modeling compatibility.

Automated Prediction



GeneralReport AutomatedPrediction : +

What values for the most important factors should be used to predict? What is the prediction for Churn?

tenure 1 **Yes**

OnlineSecurity No The predicted Churn, Yes, is the 2 most common Churn value in observed cases. Most observed cases (73.46%) are No, while 26.54% are Yes. The prediction is based on an automatically selected Gradient Boosting model.

Contract Month-to-month

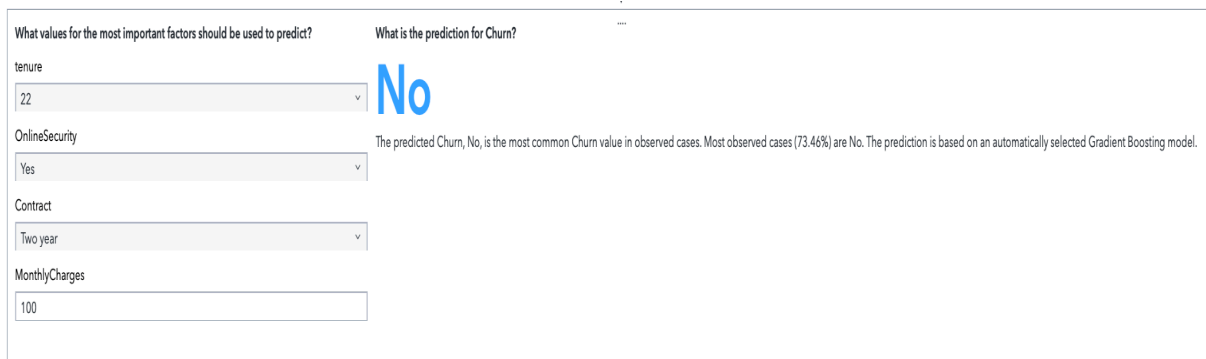
MonthlyCharges 70.35

Data Roles

Automated prediction - Churn 1

- + Assign data
- + Response + Add
 - Churn
- + Underlying factors + Add
 - Contract
 - MonthlyCharges
 - OnlineSecurity
 - tenure

In this prediction scenario, the model forecasted that the customer would churn ("Yes"). The decision was based on key input factors: the customer had a very short tenure (1 month), did not subscribe to online security services, was on a month-to-month contract, and had relatively high monthly charges. These characteristics align with typical churn-prone profiles, validating the model's logic and interpretability.



What values for the most important factors should be used to predict? What is the prediction for Churn?

tenure 22 **No**

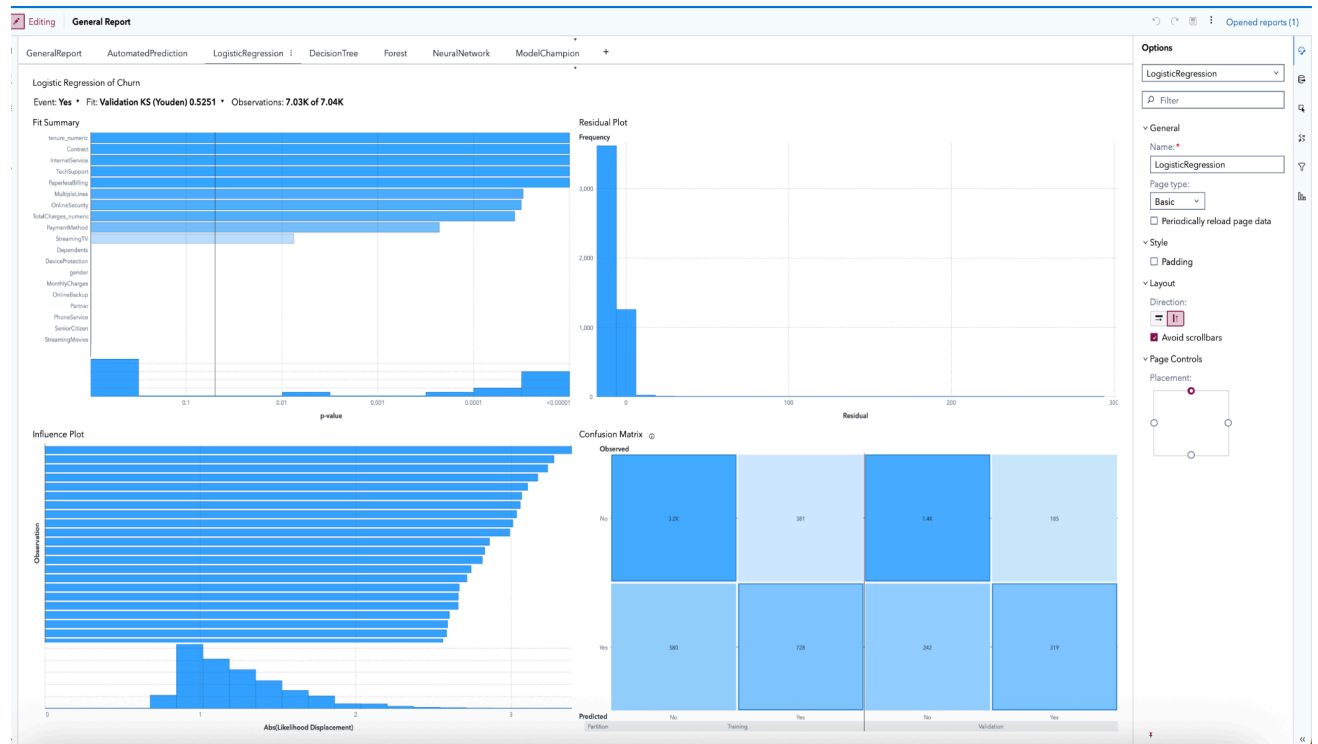
OnlineSecurity Yes The predicted Churn, No, is the most common Churn value in observed cases. Most observed cases (73.46%) are No. The prediction is based on an automatically selected Gradient Boosting model.

Contract Two year

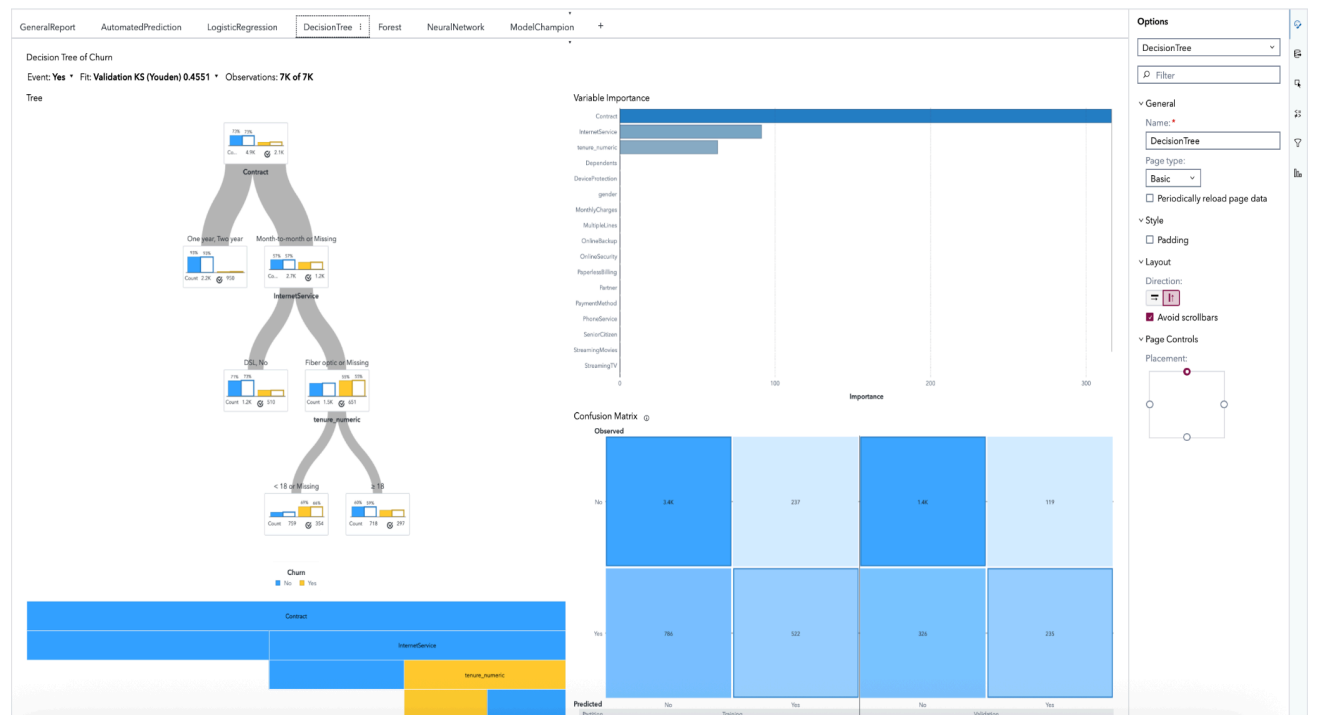
MonthlyCharges 100

In this scenario, the model predicted that the customer would not churn ("No"). This decision was influenced by several strong retention indicators: the customer has a two-year contract, subscribes to online security services, and has been with the company for 22 months. Although the monthly charge is relatively high (100), the presence of long tenure and engagement with additional services likely outweighs the price factor, leading to a low churn risk.

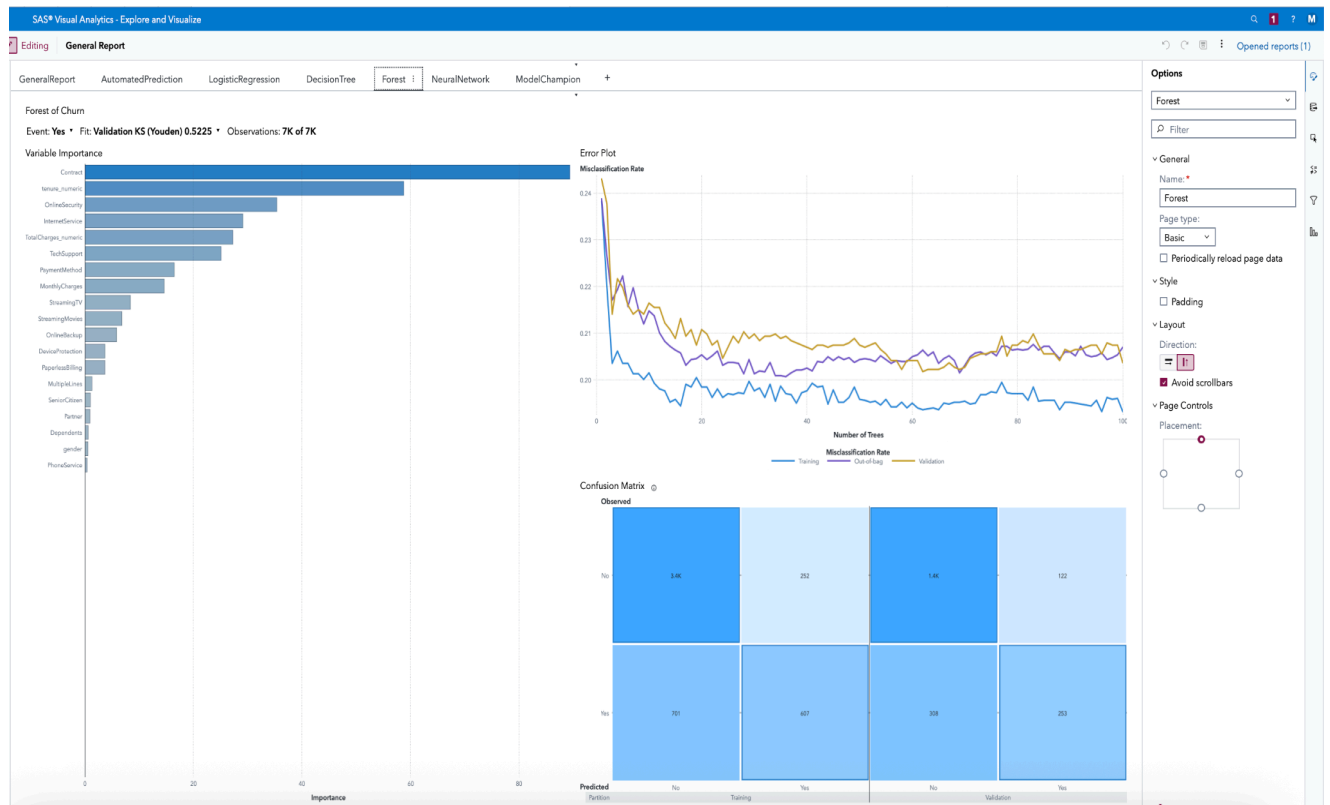
Logistic Regression



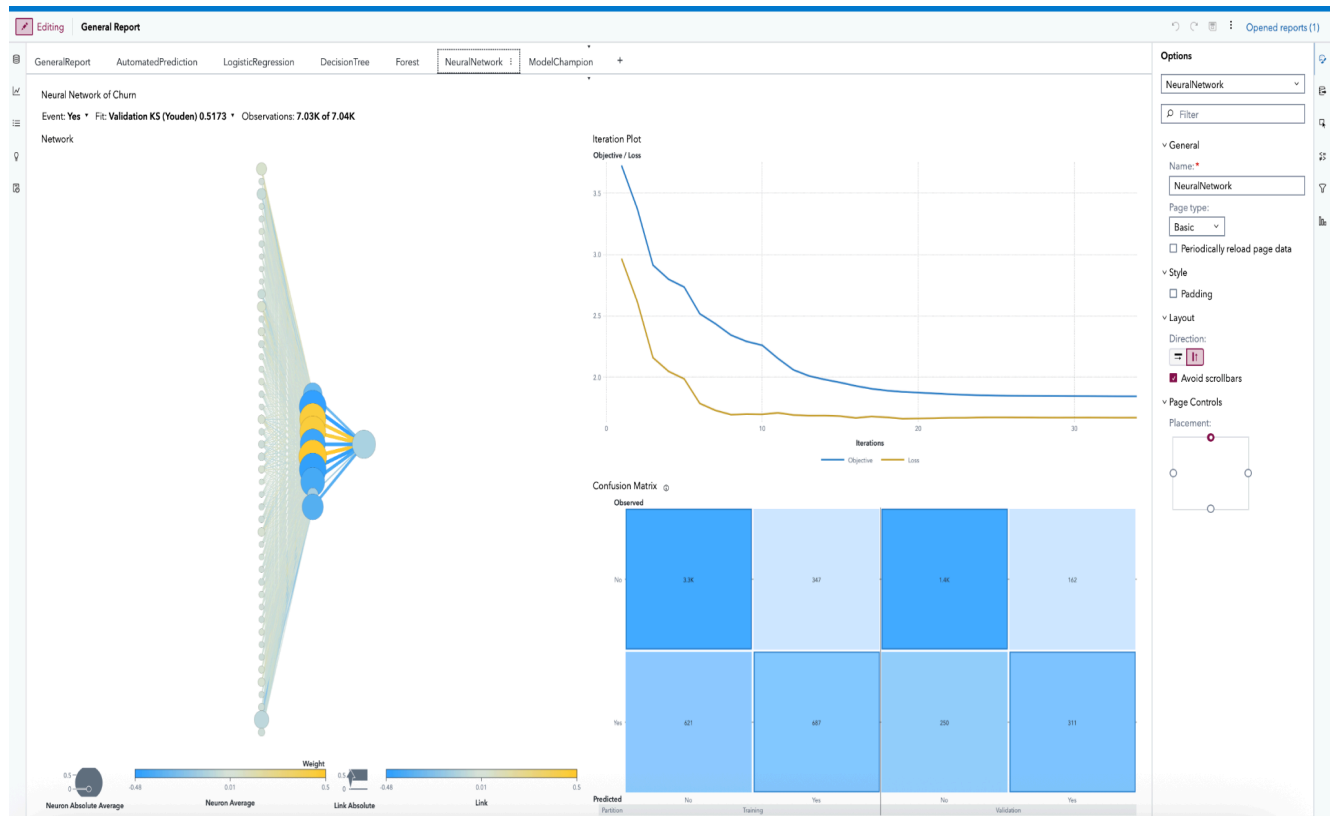
Decision Tree



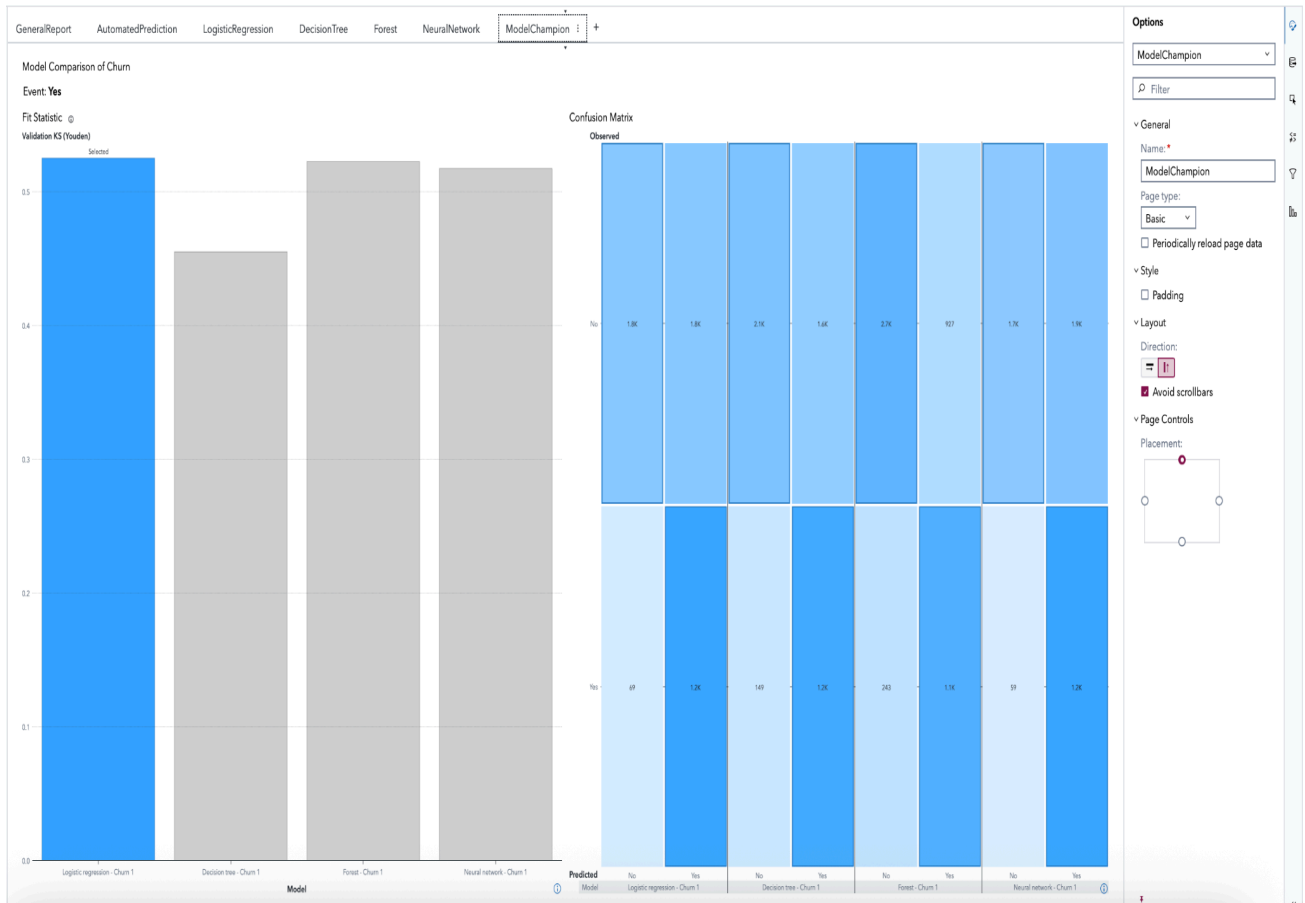
Forest



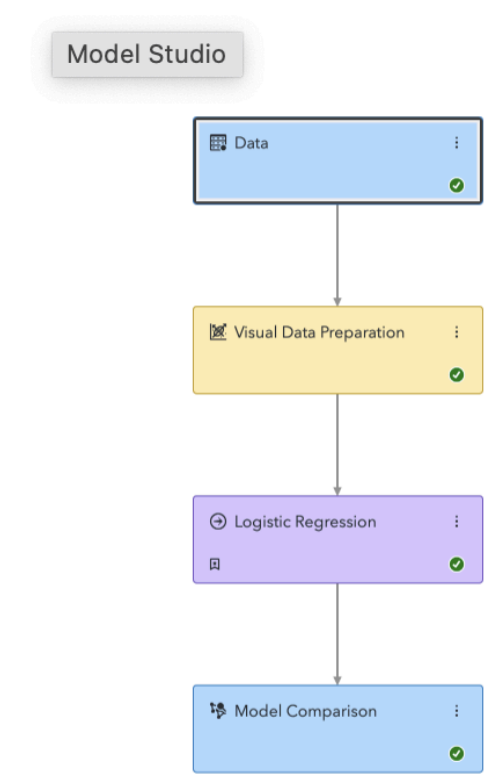
Neural Network



Model Comparison



Pipeline For Logistic Regression



Logistic Regression Results

Model Studio - Build Models

General Report > "Logistic Regression" Results

Summary Output Data

Node

Assessment

Path EP Score Code

```
1 /* Model Studio
2 /* Product: Visual Data Mining and Machine Learning
3 /* Release Version: V2024.09
4 /* Component Version: V2024.09
5 /* CAS Version: V.04.00M0P09162024
6 /* SAS Version: V.04.00M0P091624
7 /* Site Number: 70180930
8 /* Host: sas-cas-server-default-client
9 /* Encoding: utf-8
10 /* Java Encoding: UTF8
11 /* Locale: en_GB
12 /* Project GUID: 59189c41-1416-47a2-94fa-406fad89ed94
13 /* Node GUID: 5912cd93-6dcd-43c1-9ed7-8e6cc73ed6
14 /* Node ID: 590A26CNRG2XMSN8Z1D1ZMU
```

DS2 Package Code

```
1 /*
2 /* Product: Visual Data Mining and Machine Learning
3 /* Release Version: V2024.09
4 /* Component Version: V2024.09
5 /* CAS Version: V.04.00M0P09162024
6 /* SAS Version: V.04.00M0P091624
7 /* Site Number: 70180930
8 /* Host: sas-cas-server-default-client
9 /* Encoding: utf-8
10 /* Java Encoding: UTF8
11 /* Locale: en_GB
12 /* Project GUID: 59189c41-1416-47a2-94fa-406fad89ed94
13 /* Node GUID: 5912cd93-6dcd-43c1-9ed7-8e6cc73ed6
14 /* Node ID: 590A26CNRG2XMSN8Z1D1ZMU
```

Score Inputs

Name	Role	Variable Level	Type	Variable Type	Variable Label	Variable For...	Variable Len...
Contract	INPUT	NOMINAL	C	vchar			14
InternetService	INPUT	NOMINAL	C	vchar			11
MultipleLines	INPUT	NOMINAL	C	vchar			16
OnlineSecurity	INPUT	NOMINAL	C	vchar			19
PaperlessBilling	INPUT	BINARY	C	vchar			3
PaymentMethod	INPUT	NOMINAL	C	vchar			25
StreamingTV	INPUT	NOMINAL	C	vchar			19
TechSupport	INPUT	NOMINAL	C	vchar			19

Score Outputs

Name	Role	Type	Variable ...	Variable ...	Variable ...	Variable ...	Creator	Function	Creator ...
EM_CLASSIFICATION	CLASSIFICATION	C	char	Predicted for Churn		3	logisticreg	CLASSIFICATION	5912cd93-6dcd-43c1-9ed7-8e6cc73ed6
EM_EVENT PROBABILITY	PREDICT	N	double	Probability for Churn=Yes		8	logisticreg	PREDICT	5912cd93-6dcd-43c1-9ed7-8e6cc73ed6

Properties

Property Name	Property Value
binaryProbCutoff	0.5000
chooseCriterion	SBC
classCoding	GLM
classOrder	FMTASC
codeLocation	mlearning
dataMiningVersion	V2024.09
derivedFromCmModel	true
exactBrillift	true

Output

The SAS System

The CONTENTS Procedure

Data Set Name	DMCASLIB_INPUT_590A26CNRG2XMSN8Z1D1ZMU	Observations	7043
Member Type	DATA	Variables	22
Engine	CAS	Indexes	0
Created	17/05/2025 01:05:56	Observation Length	304
Last Modified	17/05/2025 01:05:56	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			

Model Comparison Results

Model Studio - Build Models

General Report > "Model Comparison" Results

Node

Assessment

Model Comparison

Ch...	Name	Algor...	KS (Y...	Accu...	Aver...	Area ...	Cum...	Cum...	Cutoff	Data ...	Depth	F1 Sc...	False ...	False ...	Gain	Gini ...	ROC ...	Lift	Misc...	Multi...	Misc...	Misc...	Num...	Root ...	Capt...
*	Logistic Regression	Logistic Regression	0.5254	0.7977	0.1378	0.8402	2.8342	28.3422	0.5000	VALIDATE	10	0.5991	0.3671	0.1194	1.8342	0.6804	0.4493	2.6025	0.2023	0.4237	0.2023	0.2023	2,111	0.3712	13.0125

Properties

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