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.

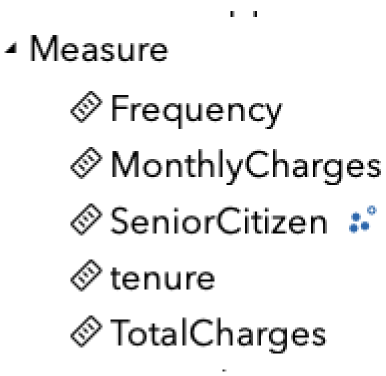
By identifying customers at high risk of churn, businesses can implement targeted retention strategies, enhance customer satisfaction, and ultimately reduce churn rates. This project illustrates a practical and real-world application of machine learning in customer analytics, leveraging data-driven decisions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

Description of Variables

The dataset includes the following 21 variables:

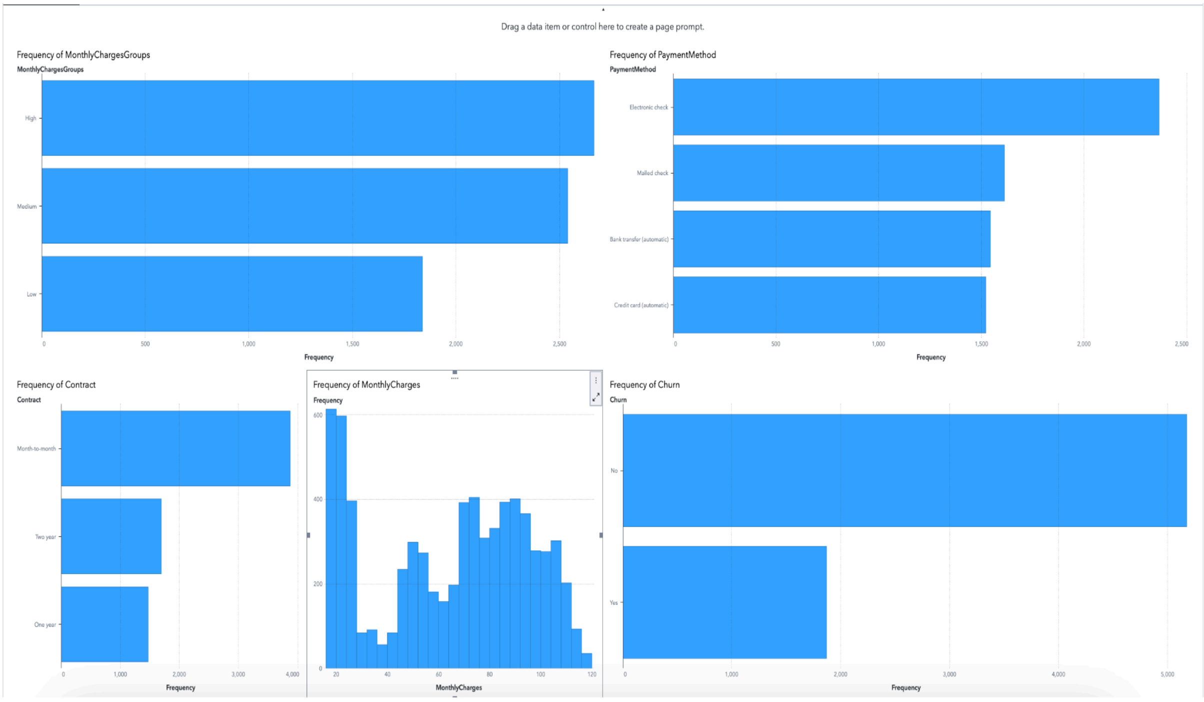
Variable	**Description**	**Type**
customerID	Unique identifier for each customer	*Character*
gender	Customer’s gender (Male, Female)	*Character*
SeniorCitizen	Indicates if customer is a senior citizen (0 = No, 1 = Yes)	*Numeric*
Partner	Whether the customer has a partner (Yes/No)	*Character*
Dependents	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	*Character*
InternetService	Type of internet service (DSL, Fiber optic, or No)	*Character*
OnlineSecurity	Subscribed to online security service	*Character*
OnlineBackup	Subscribed to online backup service	*Character*
DeviceProtection	Customer has device protection	*Character*
TechSupport	Customer has technical support	*Character*
StreamingTV	Customer uses streaming TV services	*Character*
StreamingMovies	Customer uses streaming movie services	*Character*
Contract	Type of contract (Month-to-month, One year, Two year)	*Character*
PaperlessBilling	Uses paperless billing	*Character*
PaymentMethod	Payment method (credit card, bank transfer)	*Character*
MonthlyCharges	Monthly charges billed	*Numeric*
TotalCharges	Total charges accumulated	*Numeric*
Churn	Target: Churned (Yes/No)	*Character*



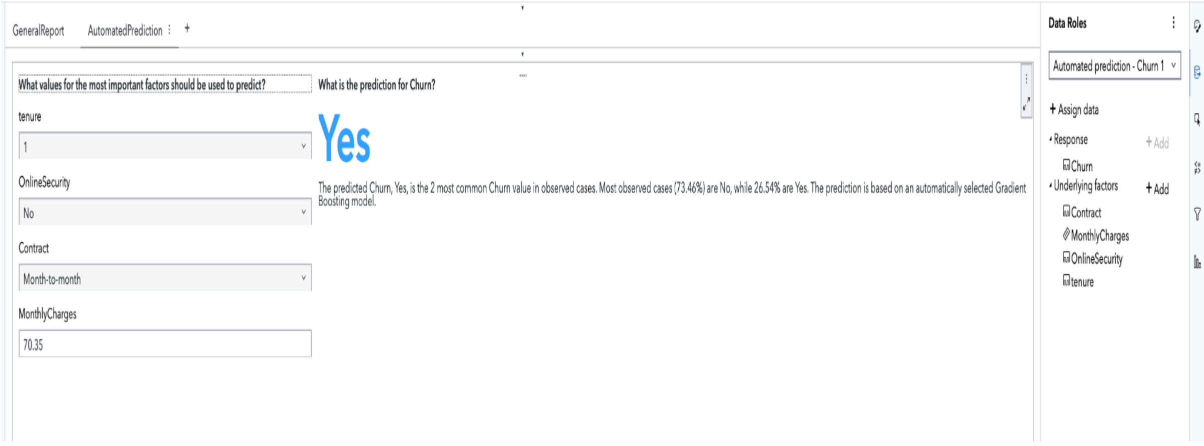
The variables I selected for initial frequency analysis MonthlyCharges, Contract, PaymentMethod, and Churn are considered among the most important in predicting customer churn. MonthlyCharges directly reflects the financial burden on the customer and may influence satisfaction or cancellation decisions. 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

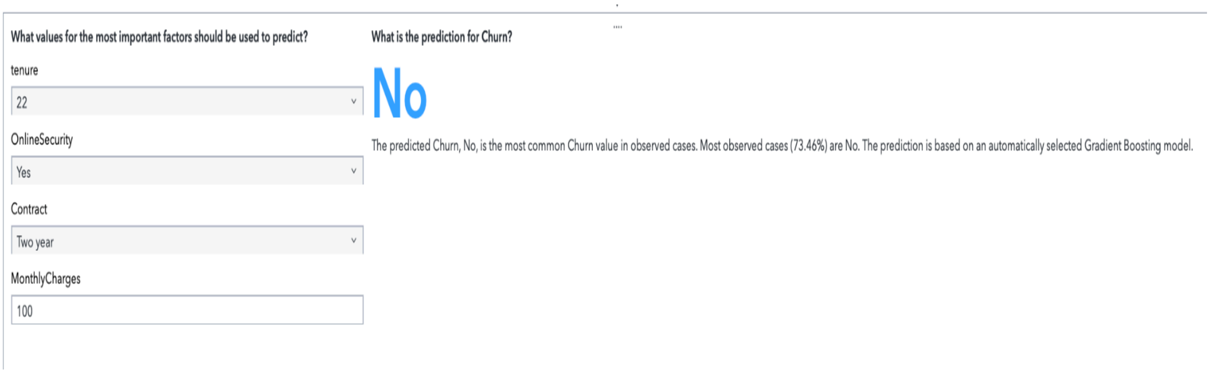


In this project, I grouped the numeric variable MonthlyCharges into three meaningful categories Low, Medium, and High by first keeping it as a measure and then using the New data item *Custom Category* feature in SAS Visual Analytics. This method is more appropriate for continuous variables with a wide 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

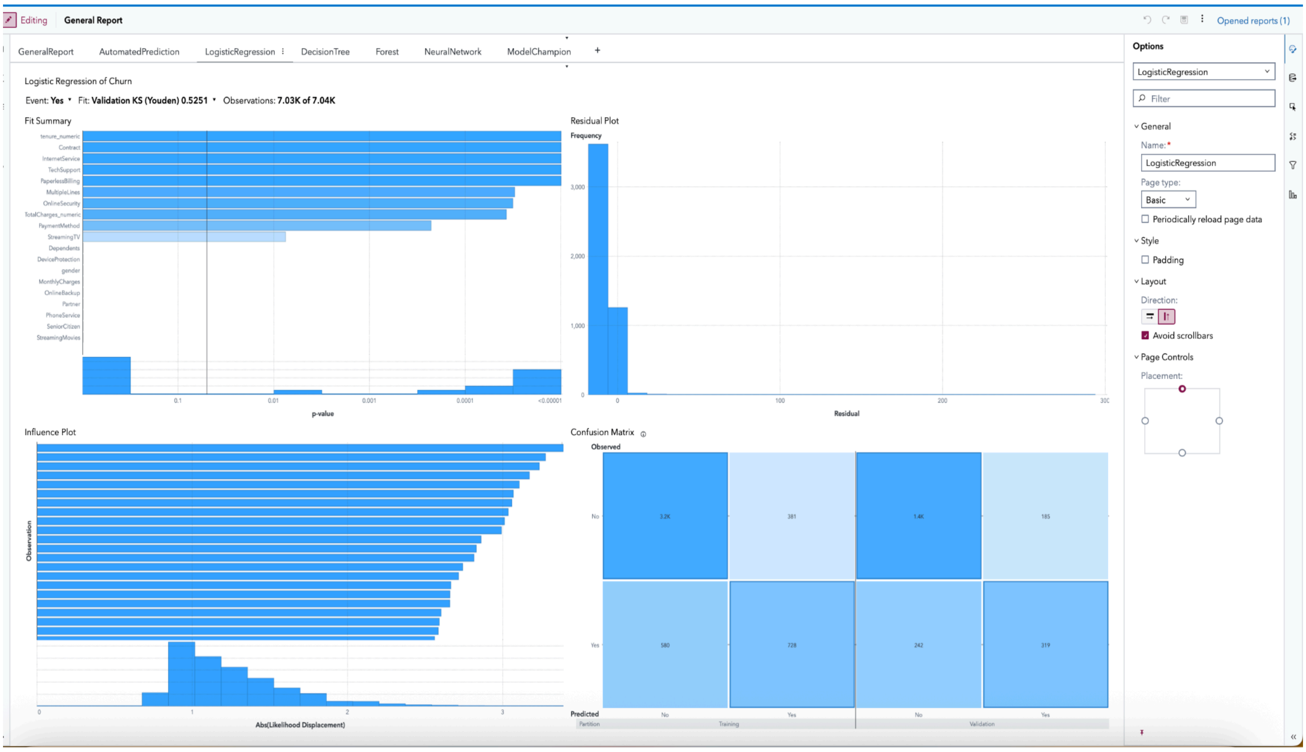


In this prediction scenario, the model forecasted that the customer would churn ("Yes"). The decision was based on key input factors: the customer had a very short tenure (1 month), did not subscribe to online security services, was on a month-to-month contract, and had relatively high monthly charges.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

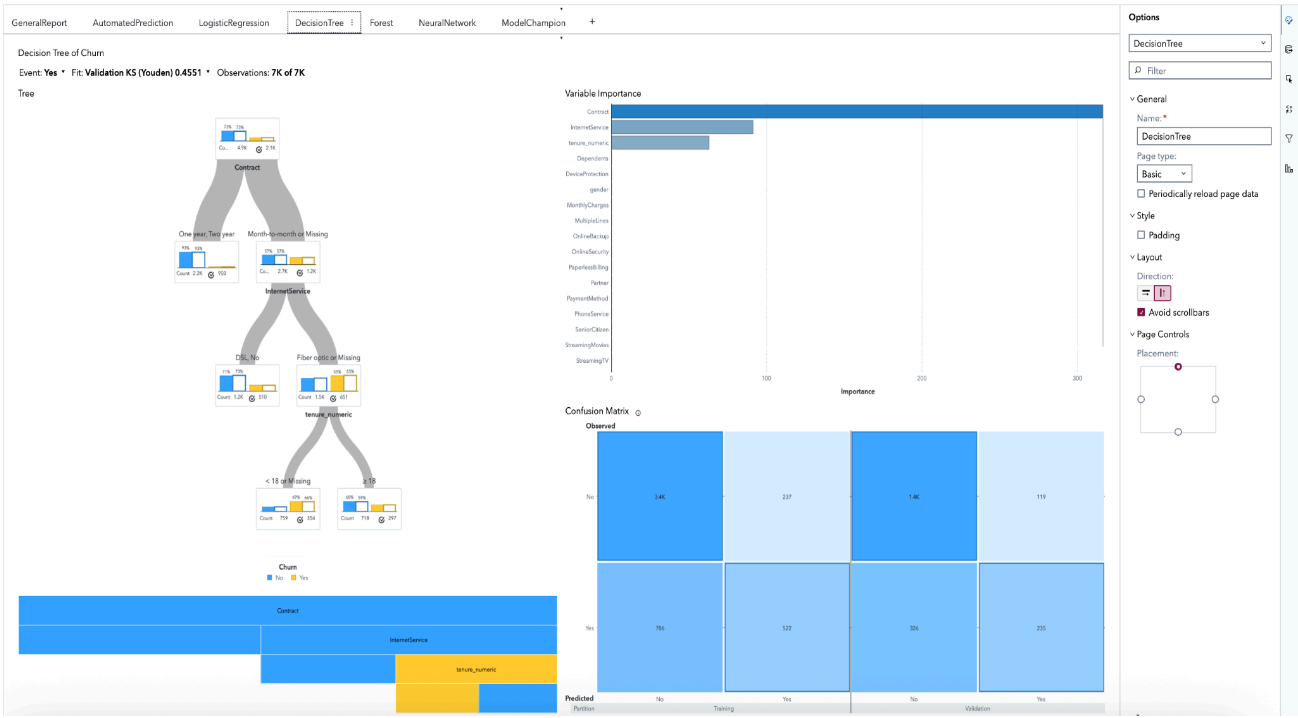


In this scenario, the model predicted that the customer would not churn ("No"). This decision was influenced by several strong retention indicators: the customer has a two-year contract, subscribes to online security services, and has been with the company for 22 months. Although the monthly charge is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

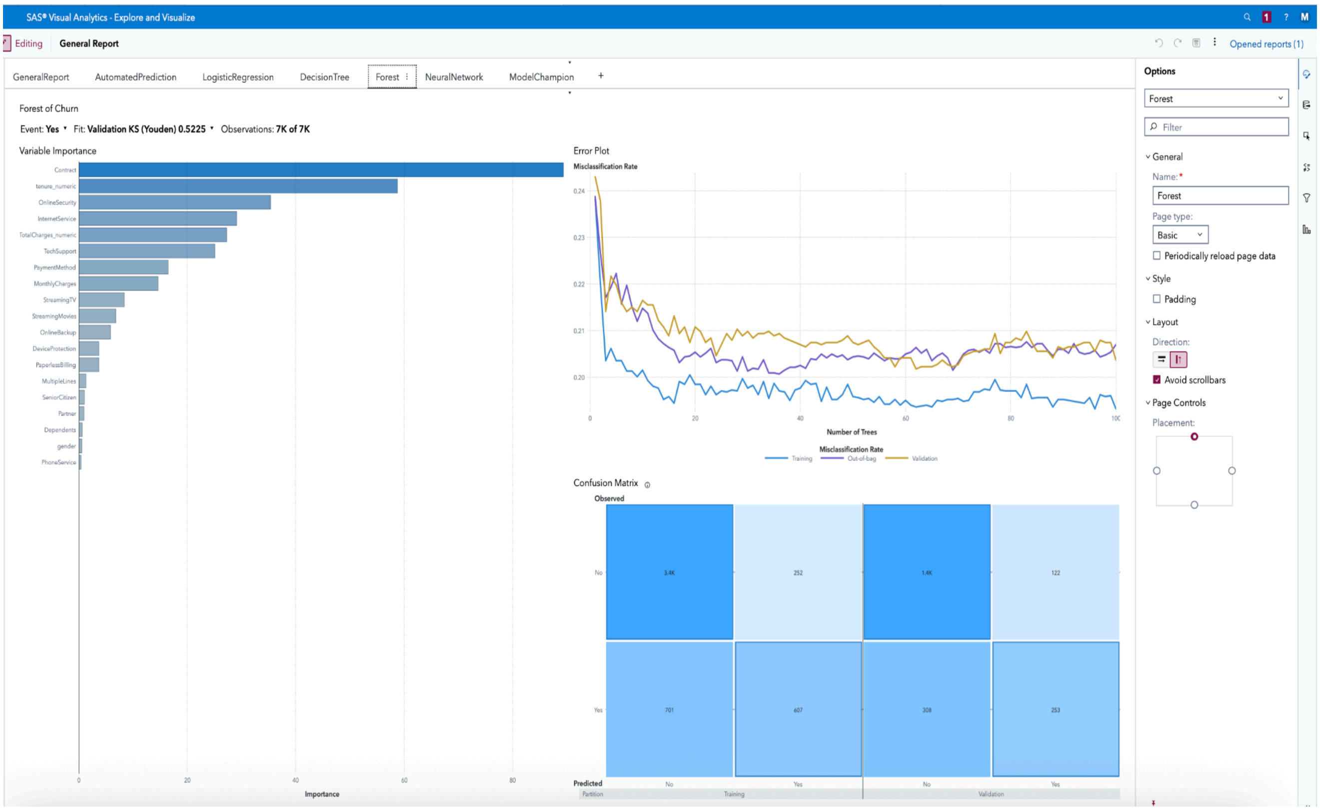
Logistic Regression



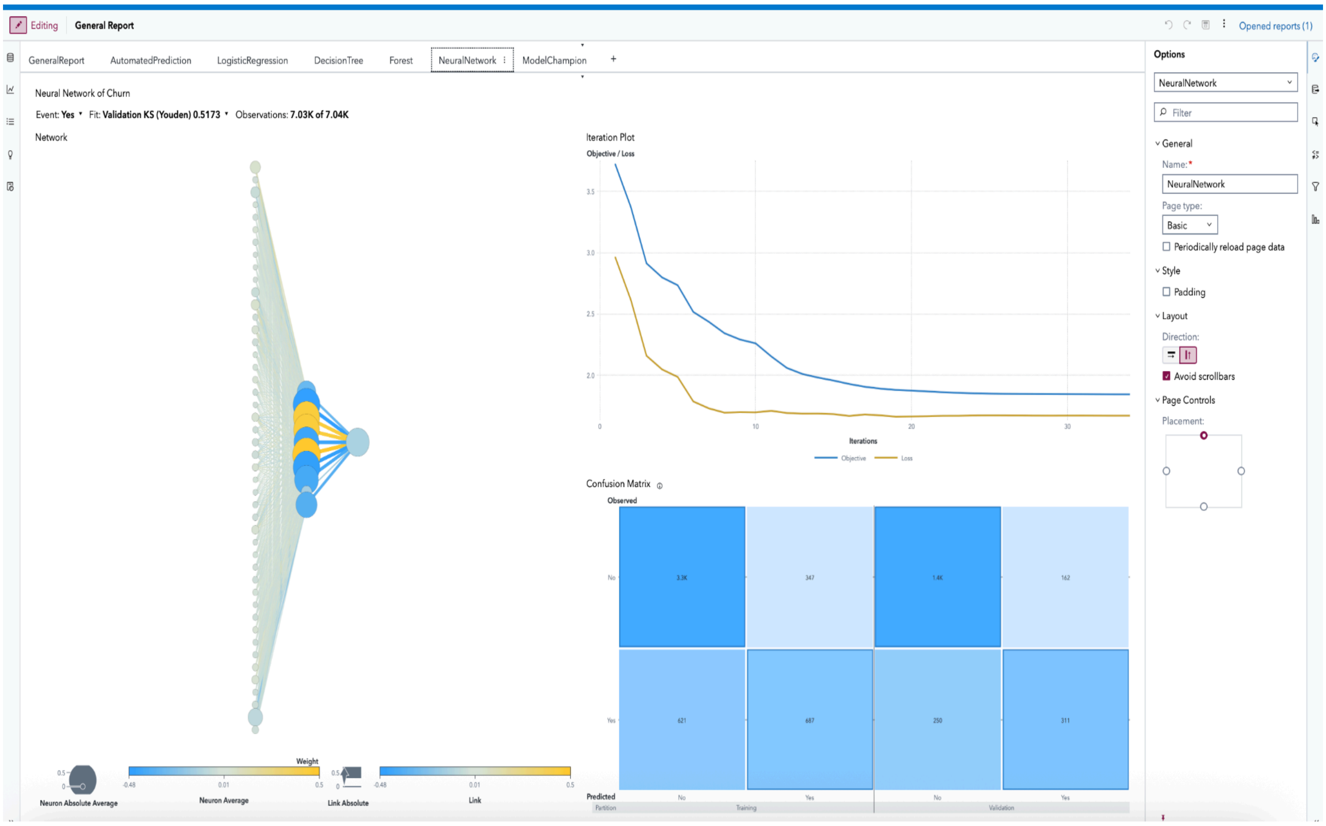
Decision Tree



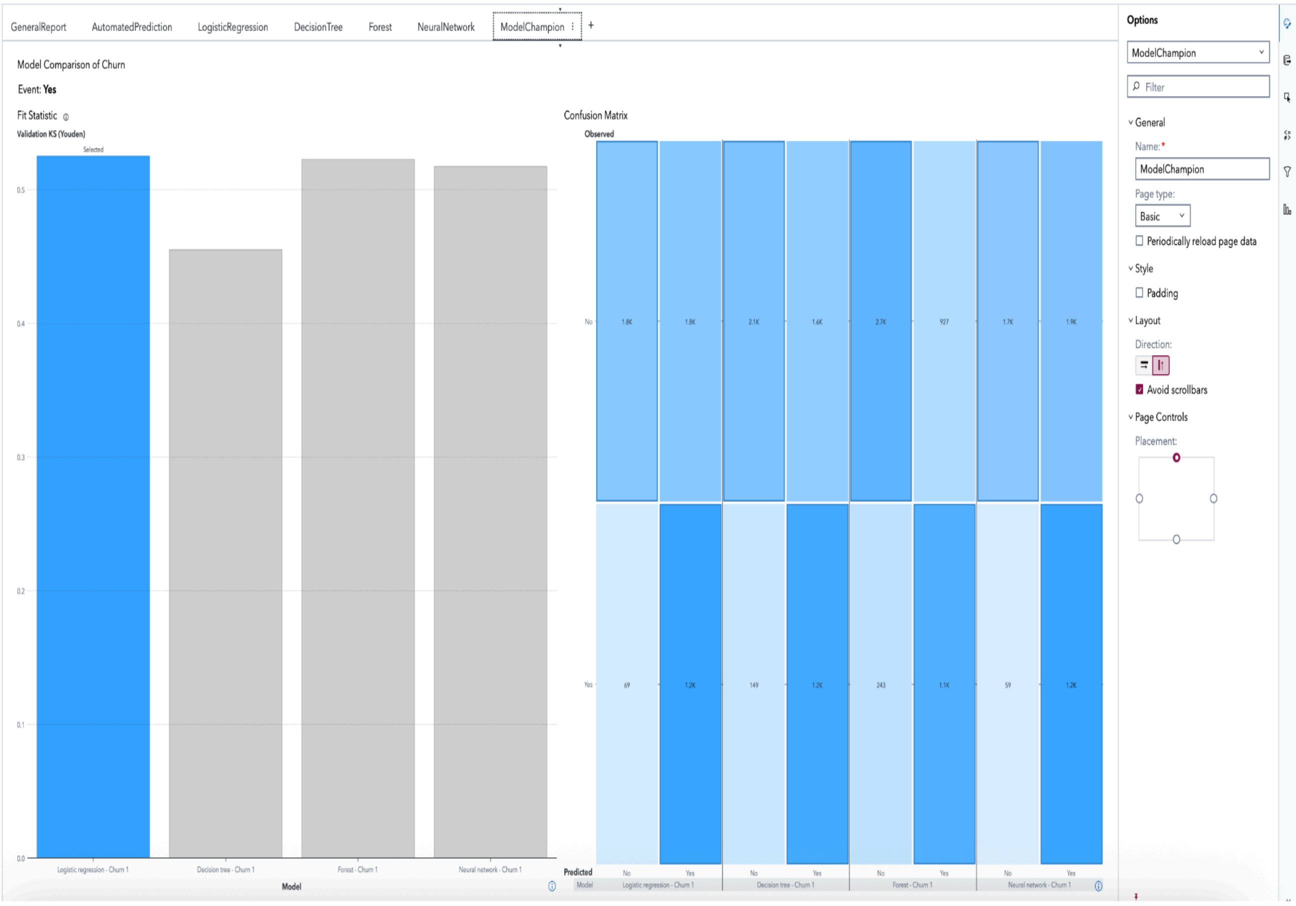
Forest



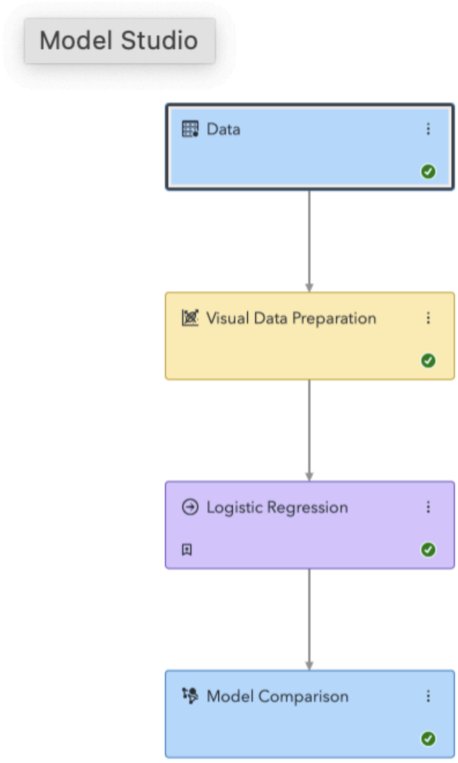
Neural Network



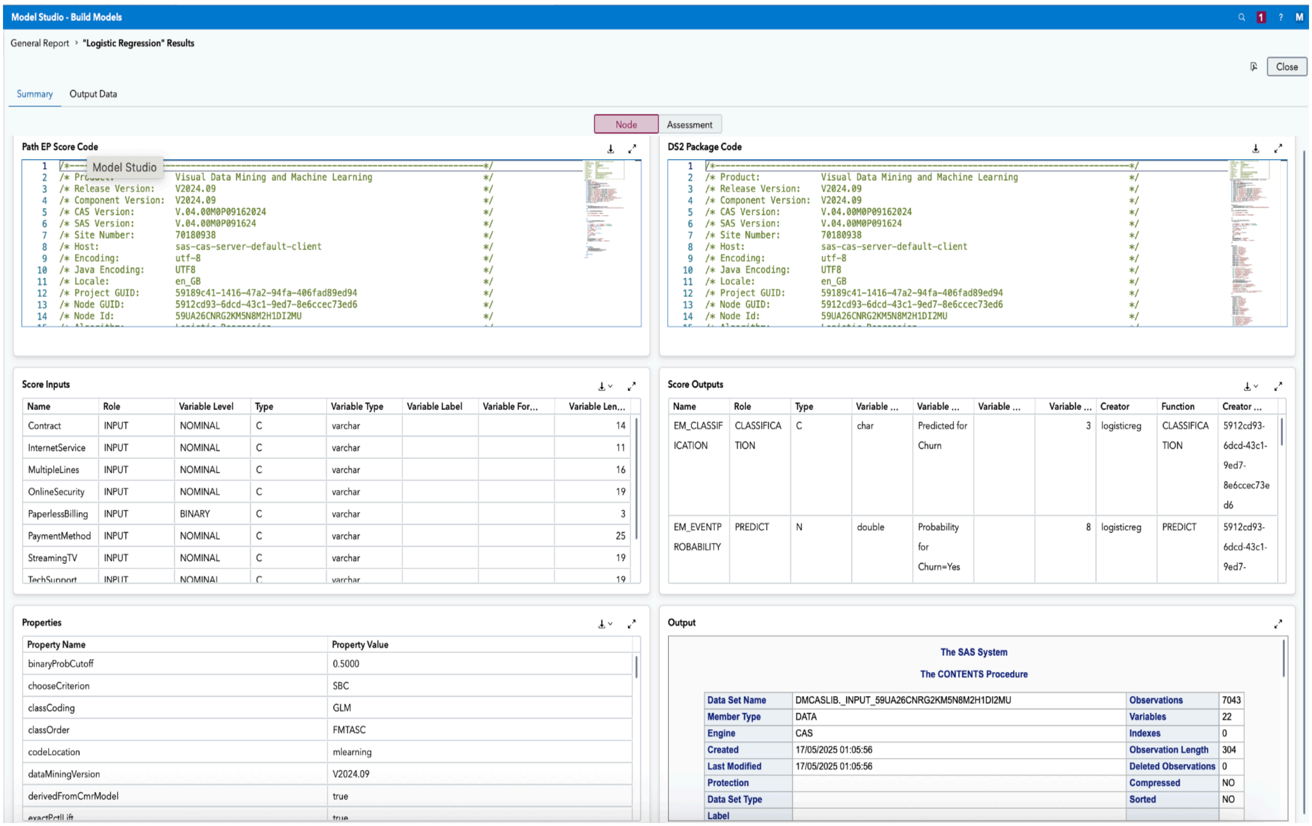
Model Comparison



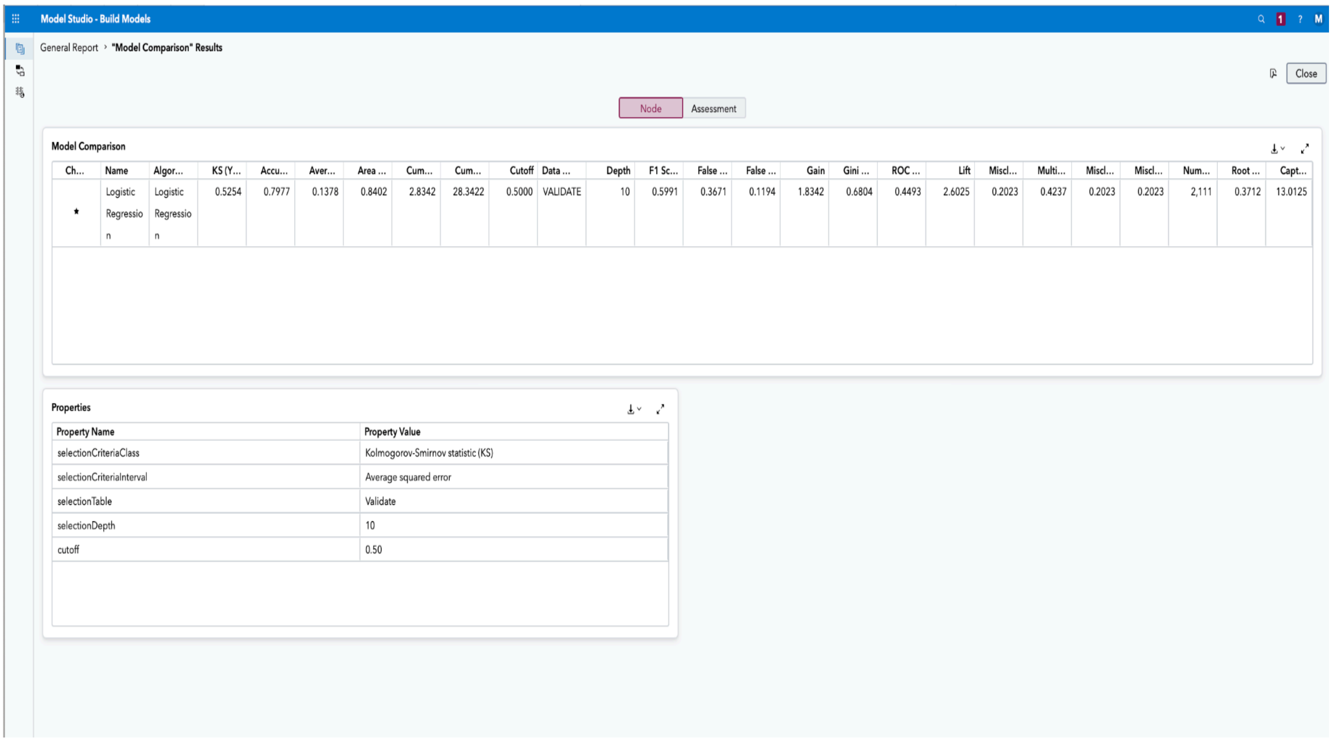
Pipeline For Logistic Regression



Logistic Regression Results



Model Comparison Results



Ch...	Name	Algor...	KS (Y...	Accu...	Aver...	Area ...	Cum...	Cum...	Cutoff	Data ...	Depth	F1 Sc...	False ...	False ...	Gain	Gini ...	ROC ...	Lift	Miscl...	Multi...	Miscl...	Miscl...	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