SAS Viya Project

Customer Churn Insights

by: Emmanuel Kwesi Ocrah

Executive Summary

This project evaluates predictive models for customer churn using the Telco Customer Churn dataset from Kaggle. Customer churn, defined as the percentage of customers who terminate their subscriptions over a specific period, is a critical challenge for subscription-based businesses, particularly in highly competitive industries such as telecommunications. Accurate churn prediction enables organizations to proactively retain at-risk customers, reducing revenue losses and increasing profitability.

The analysis leveraged SAS Viya to test four machine learning models: Logistic Regression, Decision Trees (including a Gini variant), and Neural Networks. After extensive evaluation, Logistic Regression emerged as the champion model. It achieved the highest performance metrics, including an accuracy of 79.81%, a Kolmogorov-Smirnov (KS) statistic of 0.5374, and an Area Under the Receiver Operating Characteristic Curve (AUC-ROC) of 0.85. These metrics confirm the model's ability to distinguish between churned and retained customers effectively, making it an invaluable tool for churn prediction.

Decision Trees, including the Gini variant, demonstrated reasonable predictive performance with accuracy values close to 79%, but their KS statistics and AUC-ROC scores were slightly lower than those of Logistic Regression. While Decision Trees offer interpretability and ease of implementation, their limitations in capturing complex patterns likely contributed to their lower performance. Neural Networks, despite their capacity to model intricate relationships, significantly underperformed. With an accuracy of only 73.46% and a KS statistic of 0.1192, the Neural Network model struggled to provide actionable insights, possibly due to overfitting or insufficient feature engineering.

The Logistic Regression model, selected from the "Churn Analysis" pipeline, was evaluated based on the KS statistic for the Validate partition. This partition, which represented unseen data, revealed that almost 80% of observations were correctly classified by the model. Such robust performance underscores the importance of rigorous data preprocessing and model selection techniques. The analysis identified five key predictors of churn: Tenure, Contract Type, Internet Service, Monthly Charges, and Online Security. These variables provide actionable insights, enabling businesses to design targeted retention strategies focused on high-risk customer segments.

The findings highlight the critical role of predictive analytics in addressing customer churn. The superior performance of Logistic Regression demonstrates that simple, interpretable models, when supported by strong data preprocessing, can outperform more complex algorithms. These results provide a clear roadmap for telecommunications companies seeking to enhance customer retention through data-driven decision-making.

Introduction

Customer churn, the phenomenon where customers discontinue using a service, represents a critical issue for subscription-based industries like telecommunications. Churn prediction aims to identify customers likely to discontinue their service subscriptions. Customer retention is more cost-effective than acquiring new customers, making churn prediction a vital aspect of business strategy. Telecom companies face increasing competition, necessitating precise identification of at-risk customers. Accurate churn prediction models enable targeted interventions, reducing churn rates and improving profitability. This makes churn prediction essential for business sustainability. This project employs data mining techniques to analyze the Telco Customer Churn dataset, aiming to identify the most effective predictive model. By leveraging data science, organizations can proactively engage at-risk customers to improve retention rates.

Methodology

The Telco Customer Churn dataset consists of customer demographic, account, and service details, as well as whether they churned. Key variables include tenure, monthly charges, contract type, payment method, and churn status.

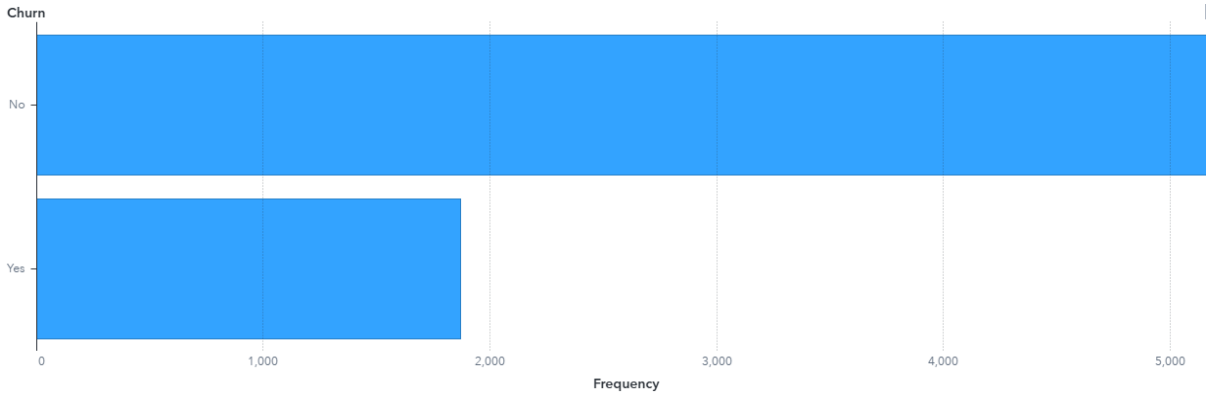
Data preprocessing included handling missing values, encoding categorical variables, and scaling numerical variables. The dataset was divided into training (60%) and validation (40%) subsets to evaluate model performance.

Four models were analyzed:

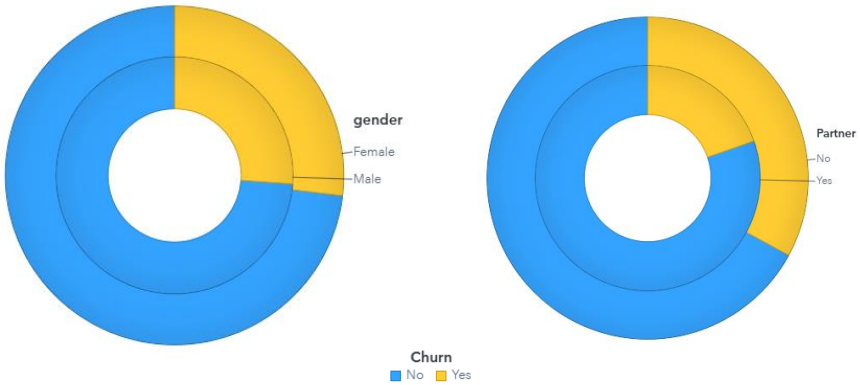
- Logistic Regression: Chosen for its interpretability and strong baseline performance.
- Decision Trees: Selected for ease of implementation and ability to capture non-linear relationships.
- Decision Tree (Gini Variant): Incorporates Gini impurity for improved splits.
- Neural Networks: Included to explore potential advantages in capturing complex patterns.

The Kolmogorov-Smirnov (KS) statistic and Area Under the ROC Curve (AUC-ROC) were key metrics for model evaluation. The Telco Customer Churn dataset includes 7,043 customer records with 21 variables, such as demographic, account, and service-related information. The target variable is churn (Yes/No).

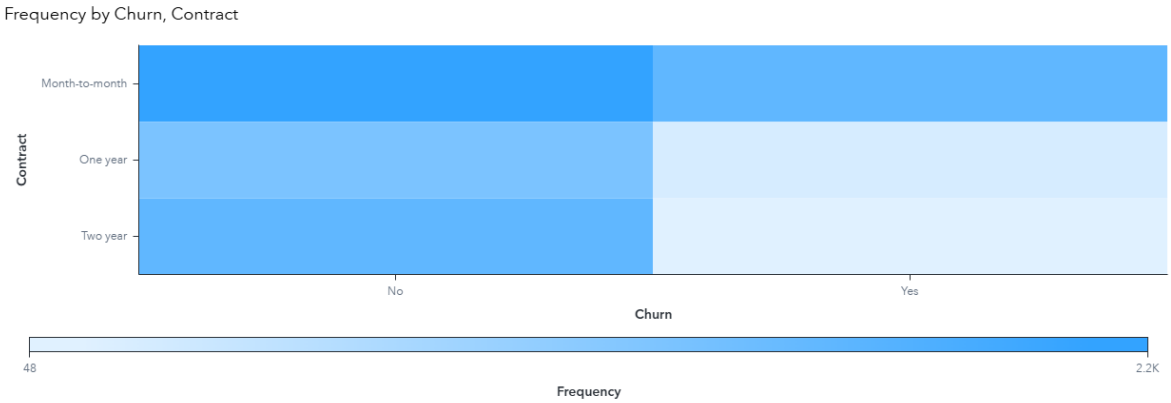
Results



The bar chart above shows that the dataset has a little over 1800 customers churning and a little over 5000 doing otherwise.



The pie charts above show the proportion of the population who churn by gender and whether the subscriber has a partner or not.



The heat map above illustrates the churn frequency based on the subscriber's contract type.

The table below summarizes the performance metrics for all four models that were used in the project.

Model	Accuracy	KS (Youden)	AUC	F1 Score	Gain
Logistic Regression	79.81%	0.5374	0.85	0.5782	1.8743
Decision Tree	78.99%	0.5024	0.808	0.5368	1.7931
Decision Tree (Gini)	78.81%	0.5155	0.8108	0.5733	1.4940
Neural Network	73.46%	0.1192	0.535	0	0.8316

Logistic Regression achieved the best overall performance with the highest KS (0.5374), AUC (0.85), and cumulative lift (2.8743). Decision Trees showed moderate success, with the Gini variant slightly outperforming the standard version with respect to KS, AUC and F1 Score. Neural Networks failed to deliver meaningful predictions, reflecting either inadequate training or poor suitability for the dataset.

Discussion and Conclusion

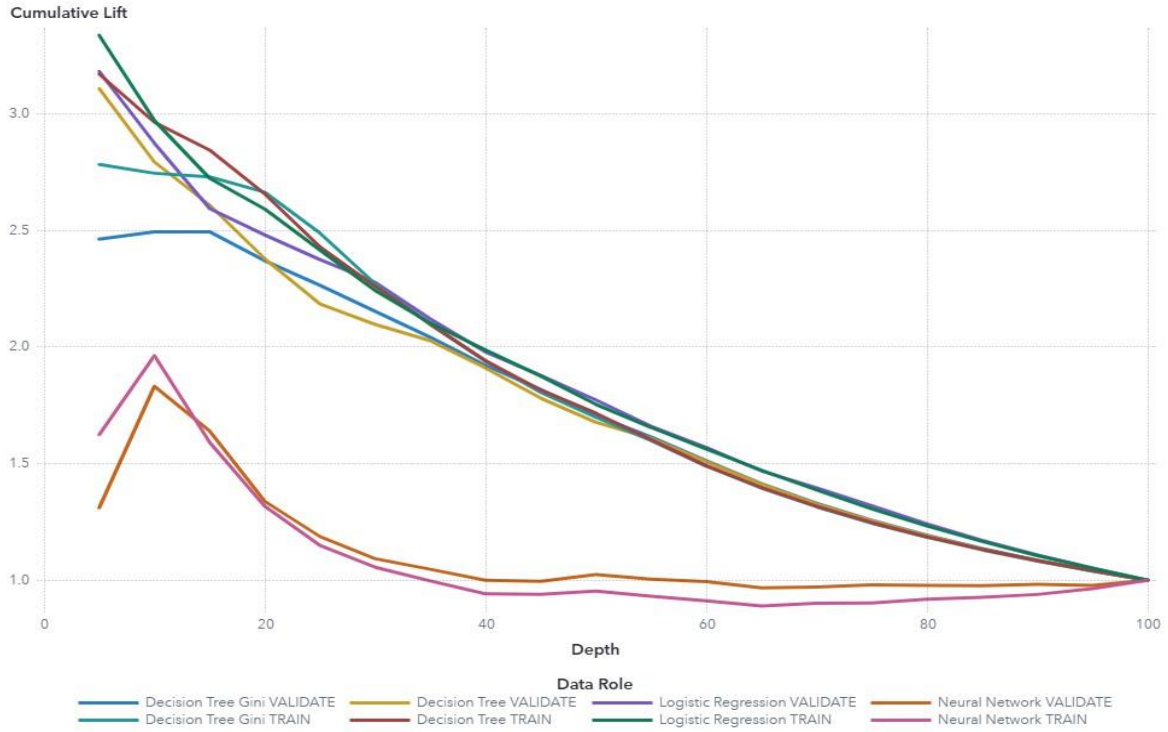
This analysis underscores the strength of Logistic Regression in churn prediction for the Telco Customer Churn dataset. Its superior performance highlights its utility for business applications requiring interpretable and actionable insights. While Decision Trees offer competitive performance, their low accuracy and KS values make them less ideal for this dataset. Neural Networks require further optimization to address their lackluster results.

The recommendations are to implement Logistic Regression for churn prediction due to its proven performance and ease of deployment. Secondly, explore ensemble methods such as Random Forests or Gradient Boosted Machines for potential improvement in performance. Analysis could also be extended to incorporate real-time data streams for dynamic prediction and intervention.

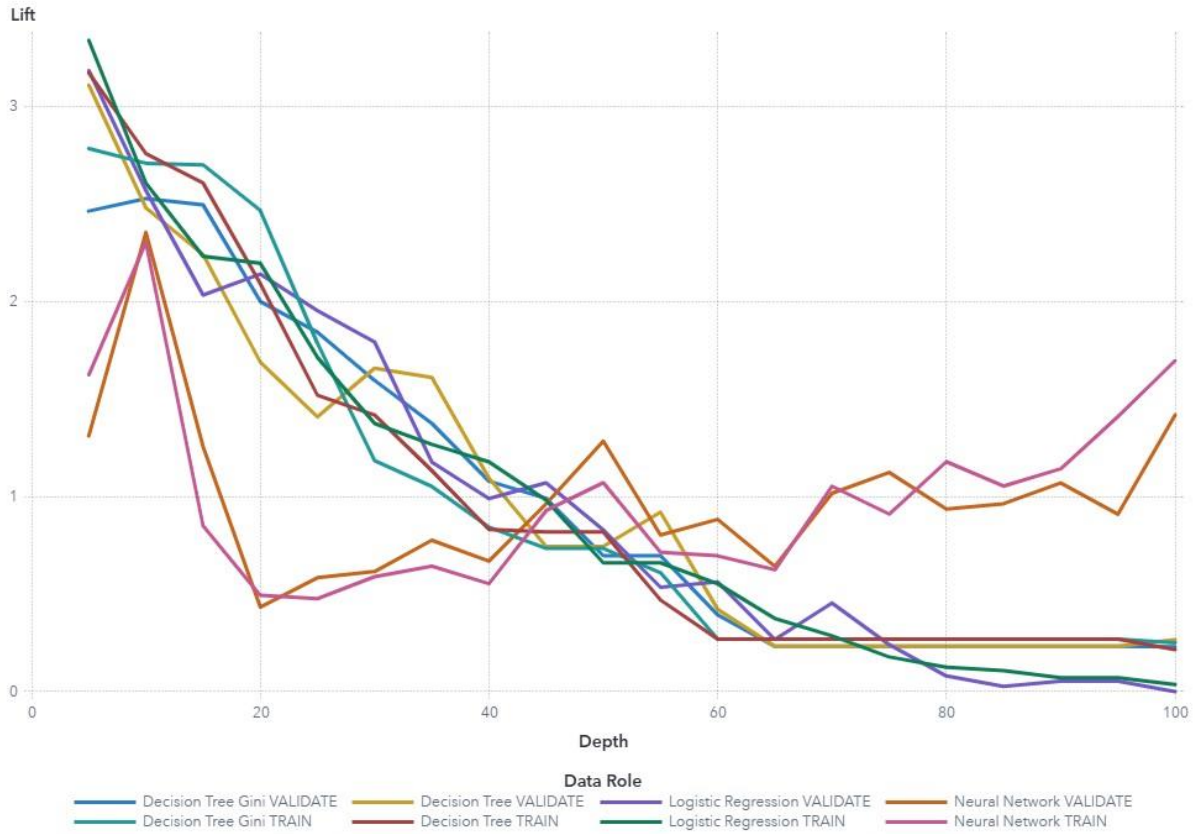
Future Research should incorporate customer sentiments from service interaction logs and investigate the effects of feature engineering, particularly for Neural Networks. Also, cost-sensitive learning approaches to account for unequal churn retention costs should be evaluated.

Appendix

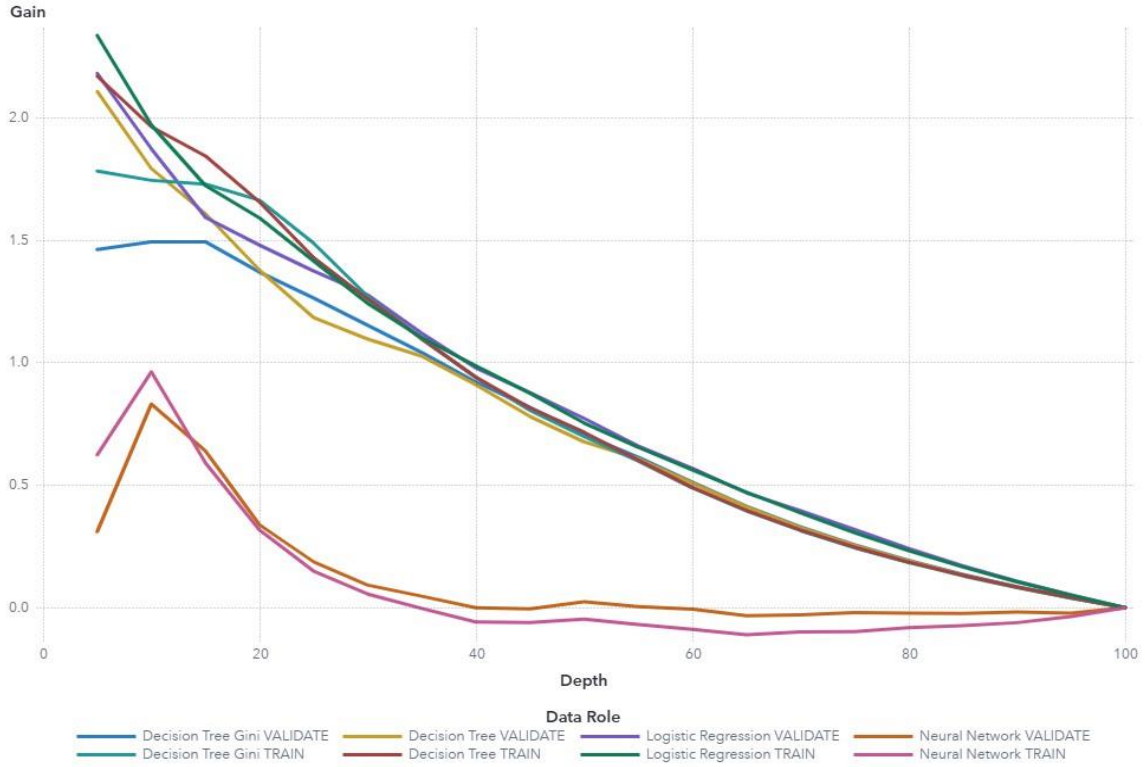
1. Cumulative Lift



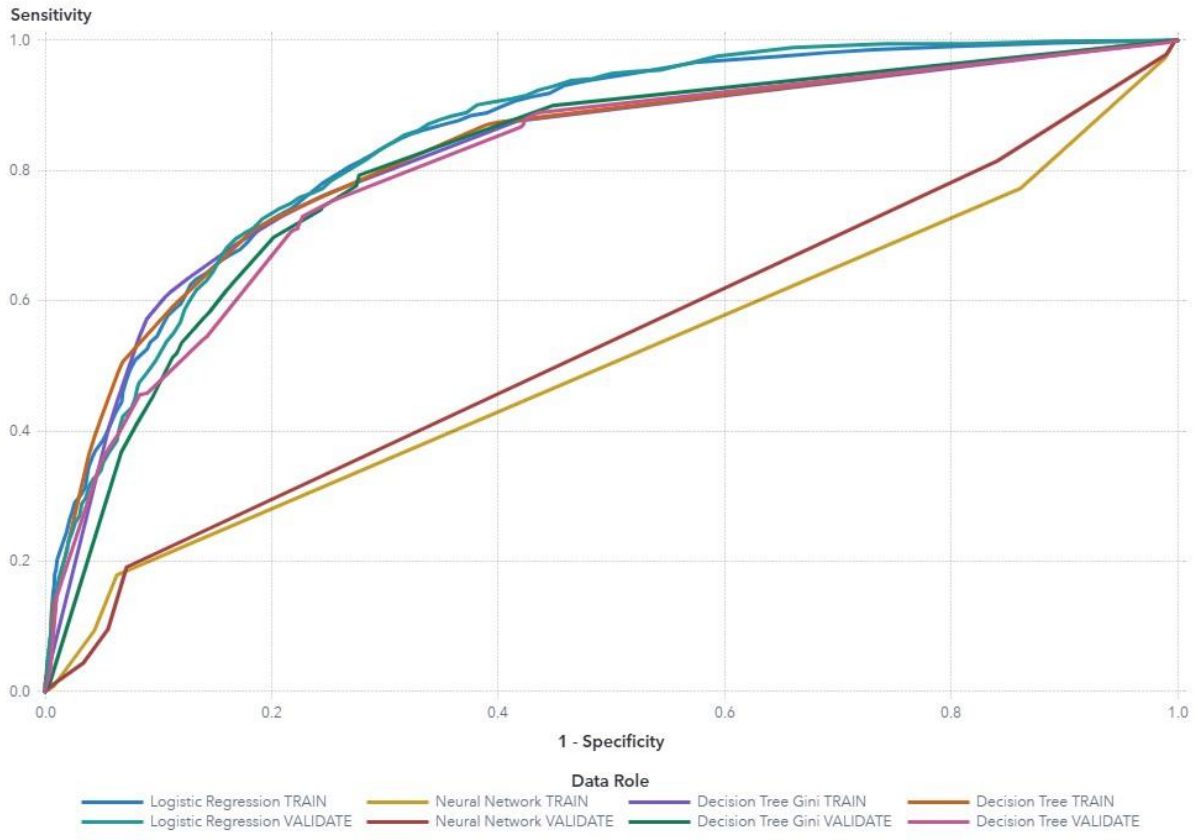
2. Lift



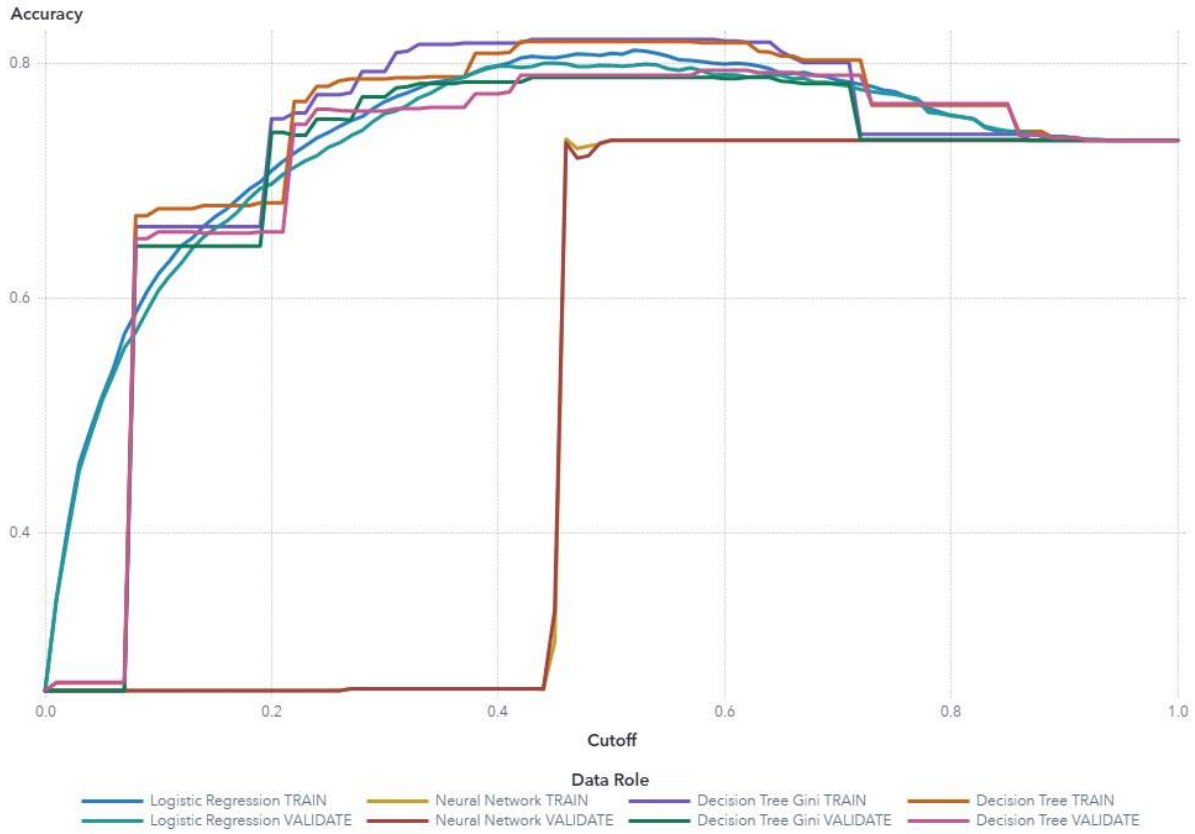
3. Gain


Gain
2.0
1.5
1.0
0.5
0.0
0
20
40
60
80
100
Depth
Data Role
Decision Tree Gini VALIDATE
Decision Tree VALIDATE
Logistic Regression VALIDATE
Neural Network VALIDATE
Decision Tree Gini TRAIN
Decision Tree TRAIN
Logistic Regression TRAIN
Neural Network TRAIN

4. ROC


Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
1 - Specificity
Data Role
Logistic Regression TRAIN
Neural Network TRAIN
Decision Tree Gini TRAIN
Decision Tree TRAIN
Logistic Regression VALIDATE
Neural Network VALIDATE
Decision Tree Gini VALIDATE
Decision Tree VALIDATE

5. Accuracy



6. Model Comparison

Champion	Name	KS (Youden)	Accuracy	ASE	AUC	Cumulative Lift	Misclassification Rate
TRUE	Logistic Regression	0.5374	0.7981	0.1343	0.85	2.8743	0.2019
FALSE	Decision Tree	0.5024	0.7899	0.1453	0.808	2.7931	0.2101
FALSE	Decision Tree Gini	0.5155	0.7881	0.1475	0.8108	2.494	0.2119
FALSE	Neural Network	0.1192	0.7346	0.2302	0.535	1.8316	0.2654