Machine Learning for Customer Churn Analytics

SCH MGMT 655 - Group 2 Project

Sakshi Agarwal • Noynicaa Santani • Haley Hoang

Project Overview

The Benefit

Identify high-risk customers early and support targeted retention strategies

7,043
Customers

21
Variables

including their demographic profile, account information, services used, and churn status

The Challenge

Customer churn is a critical issue for subscription-based businesses. The cost of acquiring a new customer often exceeds the cost of retaining an existing one.

Our Objective

Predict customer churn and understand key drivers affecting customer retention

Understanding the Dataset

IBM Telco Customer Churn

7,043 telecommunications customers with detailed profiles including demographics, services, account information, and churn status.

21 variables spanning customer characteristics, service subscriptions, contract details, and billing methods.

Key Data Categories

- Demographics: gender, age, family status
- Services: phone, internet, security, streaming
- Account: tenure, contract type, charges
- Target: Churn (Yes/No)

Three Critical Business Questions

01

High-Revenue Customer ROI

What is the dollar value of reducing churn in the top 20% of high-revenue customers, and which interventions maximize return on investment?

02

Customer Lifetime Value

Which customer segments by contract type, tenure, and billing method contribute most to total lifetime revenue?

03

Service Bundle Impact

Which service combinations most strongly affect churn risk and customer satisfaction?

Data Preparation & Cleaning

Missing Data Handling

11 missing values found in TotalCharges variable (0.2% of records). Removed to ensure data quality.

Final dataset: 7,043 → 7,032 complete records

Variable Classification

3 Numerical: tenure, MonthlyCharges, TotalCharges

18 Categorical: demographics, services, contract details

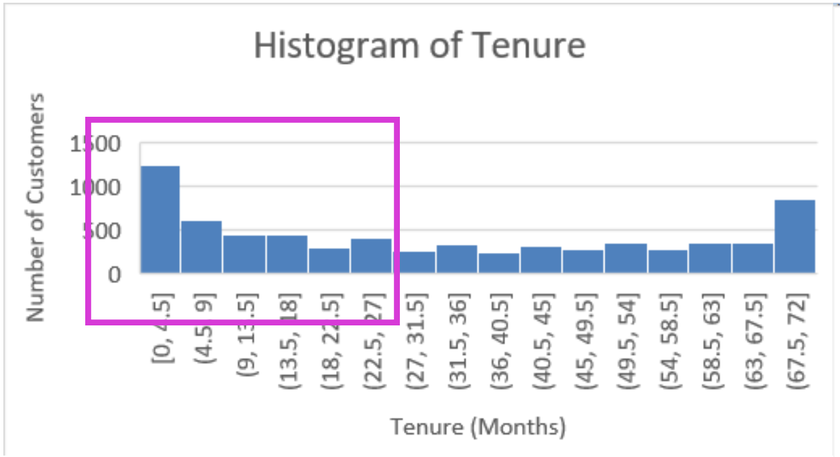
Churn (Yes/No) is target variable, while others are predictors

Correlation Insights & Multicollinearity Warnings

	MonthlyCharges	*TotalCharges*	*tenure*
MonthlyCharges	1		
TotalCharges	0.651173832	1	
tenure	0.247899856	0.826178398	1

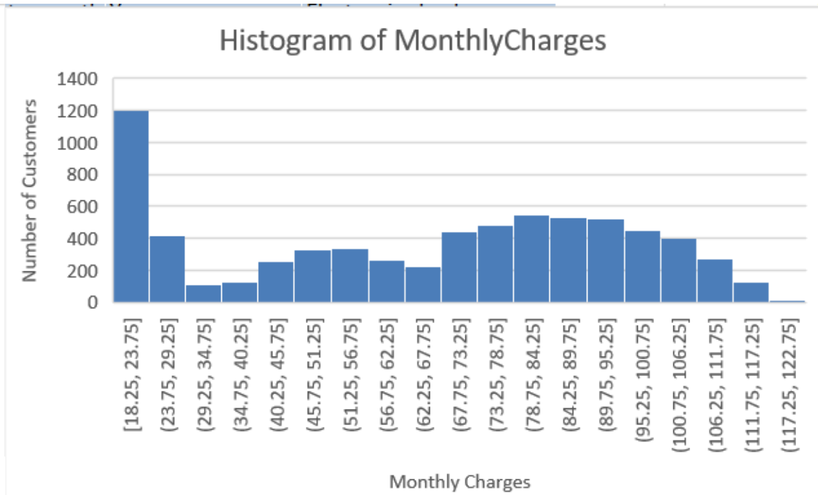
Key Findings	Business Implication
TotalCharges ≈ Tenure × MonthlyCharges	High correlation (expected) – reflects cumulative spend
Tenure & MonthlyCharges are weakly correlated	Plan price does **not** strongly drive loyalty
No severe multicollinearity among predictors	All variables add unique predictive power → safe to keep in model

Data Visualization of Numerical Variables



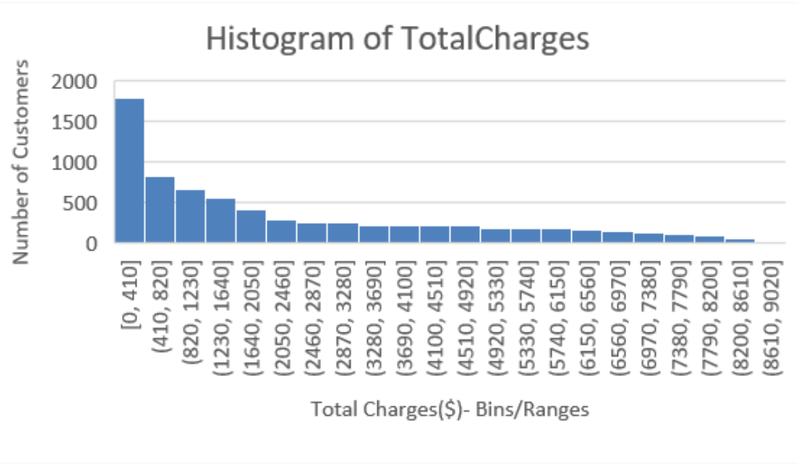
Highly right-skewed tenure & total spend

Most customers are new or low-value; only a small loyal tail (60+ months) drives the majority of lifetime revenue



Monthly/visit spend is concentrated in mid-tier

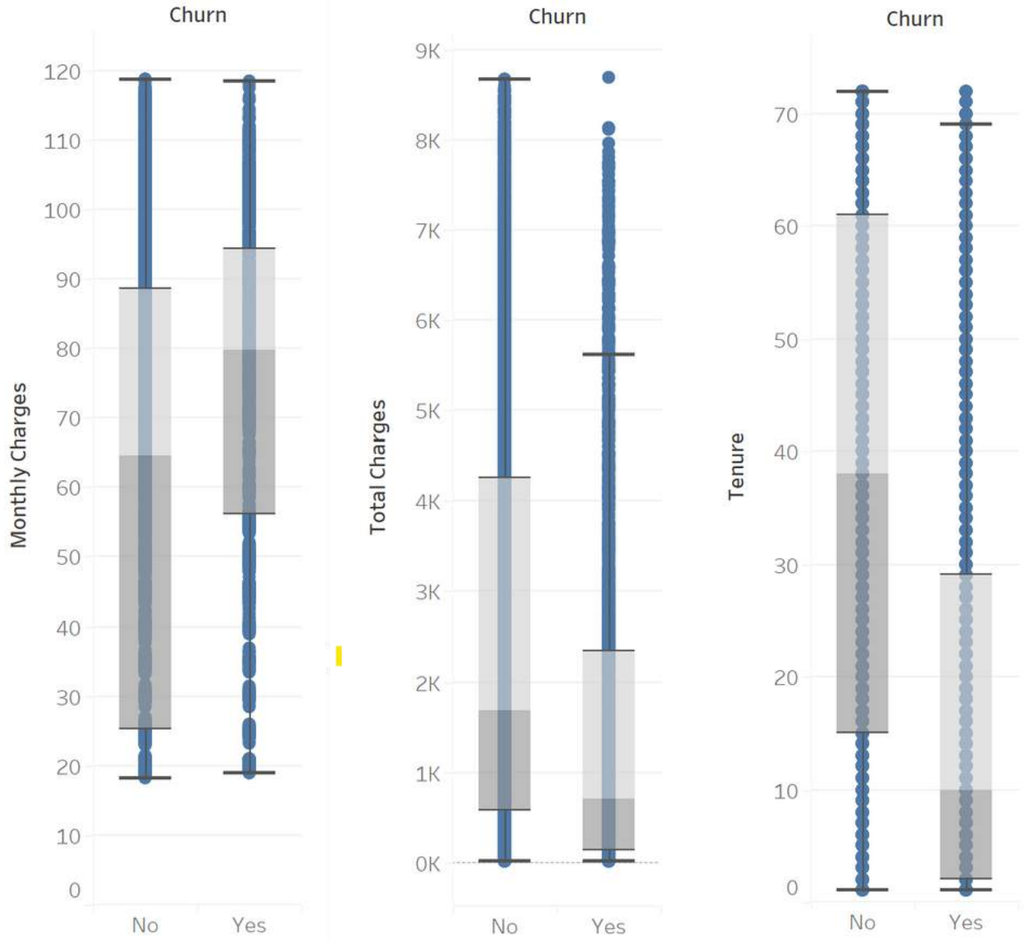
Bulk of customers spend $20–$90 per trip → clear “sweet spot” for basket-building promotions; very few extreme low or premium spenders



Recent shoppers dominate the base

Huge concentration in 0–6 months tenure → grocery customers shop very frequently, but loyalty is fragile if not nurtured early

Data Visualization of Categorical Variables



Monthly Charges

- Churners pay higher monthly charges on average.
- Higher-priced plans → higher churn risk.

Total Charges

- Churners have much lower total charges.
- Indicates shorter customer lifetime before leaving.

Tenure

- Churners show very low tenure.
- Long-tenure customers are far less likely to churn.

Classification Models Used

1

Logistic Regression

for interpretability

2

Neural Network

for maximum predictive power

3

Decision Tree

for intuitive rules

Logistic Regression

Logistic Regression Model

Feature Selection

Stepwise selection identified 11 significant predictors from 21 variables, ensuring model efficiency without overfitting.

Predictor	Estimate	Confidence Interval: Low	Confidence Interval: Upp	Odds	Standard Erro	Chi2-Statisti	P-Value
Intercept	-2.036387	-2.633747339	-1.439026934	0.130499	0.304781214	44.64206758	2.37E-11
tenure	-0.062595	-0.077737535	-0.047453031	0.939324	0.007725781	65.64451652	5.4E-16
TotalCharges	0.0003806	0.000215907	0.000545236	1.000381	8.4014E-05	20.51959206	5.9E-06
SeniorCitizen_1	0.2573681	0.049899029	0.464837181	1.293521	0.105853514	5.911516244	0.015042
PhoneService_Yes	-0.602506	-0.935108751	-0.26990288	0.547438	0.169698493	12.60568683	0.000385
InternetService_DSL	0.6371501	0.297417907	0.976882337	1.891084	0.173335948	13.51158386	0.000237
InternetService_Fiber optic	1.4690829	1.127176165	1.810989593	4.345248	0.174445407	70.92078244	3.72E-17
OnlineSecurity_Yes	-0.424464	-0.635721624	-0.213206146	0.65412	0.107786541	15.50787943	8.22E-05
TechSupport_Yes	-0.429733	-0.647387562	-0.212078919	0.650683	0.111050164	14.9747469	0.000109
StreamingMovies_Yes	0.2751269	0.07508086	0.475172863	1.316698	0.102066162	7.266117535	0.007027
Contract_Month-to-month	1.3292382	0.901486178	1.75699014	3.778164	0.218244817	37.09520049	1.13E-09
Contract_One year	0.6560336	0.229874723	1.082192518	1.927133	0.217432004	9.103434149	0.002551
PaperlessBilling_Yes	0.3969613	0.208348882	0.585573622	1.487298	0.096232569	17.01578821	3.71E-05
PaymentMethod_Electronic check	0.4138958	0.238532378	0.589259244	1.5127	0.089472783	21.39932957	3.73E-06

Equation

$$P(\text{Churn} = 1) = \frac{1}{1 + e^{-(-2.76 + 0.62*Dt\_Customer\_Years - 0.89*Teenhome - 0.03*Recency + 0.001*MntWines + 0.001*MntMeatProducts + 0.004*GoldProducts + 0.08*NumWebPurchases - 1.8*NumStorePurchases + 0.16*NumWebPurchasesMonth + 0.66*Education\_Phd + 0.88*Marital\_Status\_Divorced + 0.96*Marital\_Status\_Single)}}$$

Key Churn Drivers: Odds Ratios

77%

Month-to-Month Contract

3.77× higher churn odds vs. two-year contracts

77%

Fiber Optic Internet

4.34× higher churn odds vs. no internet service

34%

Electronic Check Payment

1.51× higher churn odds vs. other payment methods

35%

OnlineSecurity Service

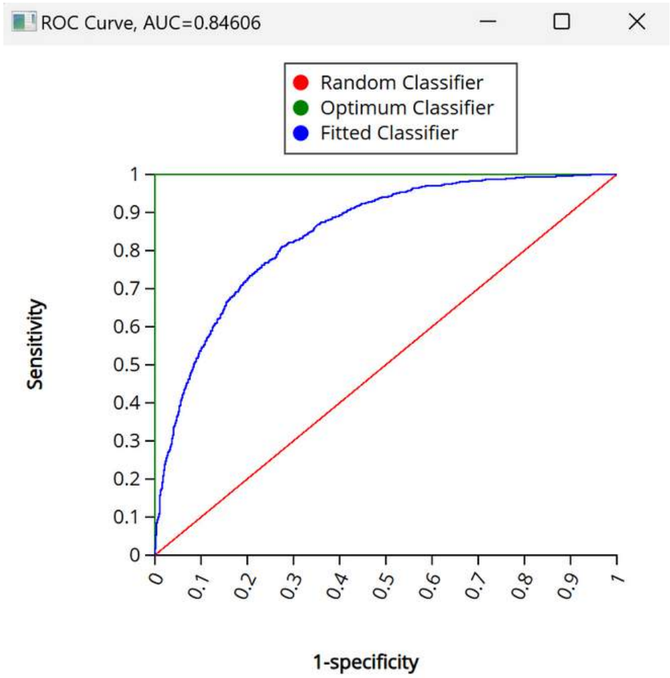
0.65× odds ratio - protective factor reducing churn

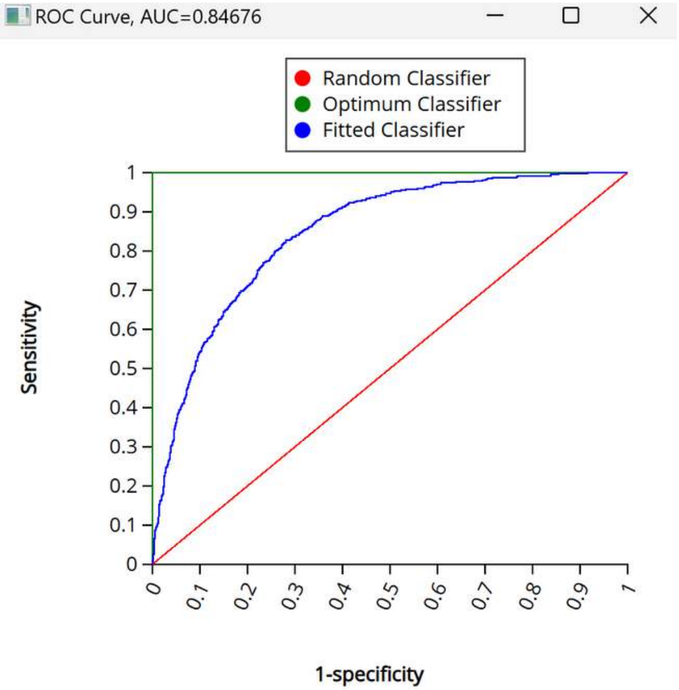
Model Performance

0.846 **0.8467**

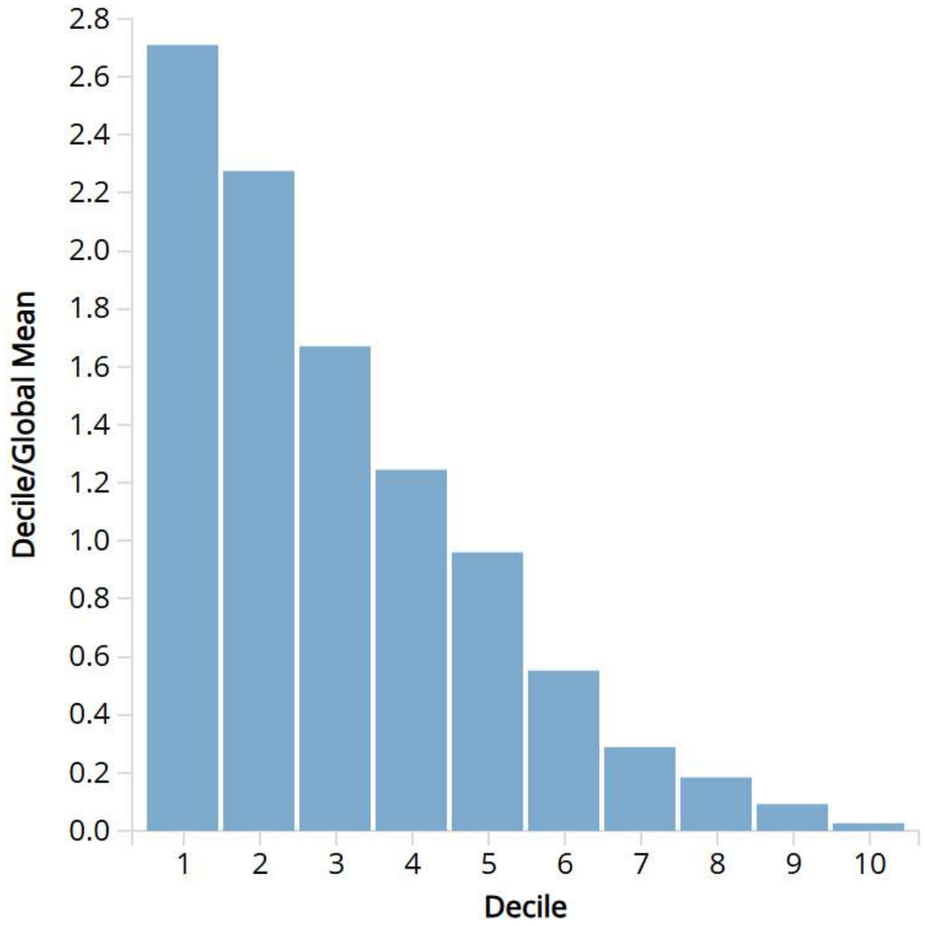
Training AUC **Validation AUC**

This model works very well since AUC > 0.5.
Training and validation AUC are nearly identical meaning there is no overfitting issue and it can work well with new dataset





Determine Appropriate Cutoff



To maximize campaign efficiency, we recommend targeting the top 40% of customers ranked by predicted churn probability (Deciles 1–4).

New cutoff = **0.290166343**

Confusion Matrix		
Actual\Predicted	No	Yes
No	1832	220
Yes	335	426

Error Report			
Class	# Cases	# Errors	% Error
No	2052	220	10.72124756
Yes	761	335	44.02102497
Overall	2813	555	19.72982581

Metrics	
Metric	Value
Accuracy (#correct)	2258
Accuracy (%correct)	80.27017
Specificity	0.892788
Sensitivity (Recall)	0.55979
Precision	0.659443
F1 score	0.605544
Success Class	Yes
Success Probability	0.5

Confusion Matrix		
Actual\Predicted	No	Yes
No	1528	524
Yes	160	601

Error Report			
Class	# Cases	# Errors	% Error
No	2052	524	25.53606238
Yes	761	160	21.02496715
Overall	2813	684	24.31567721

Metrics	
Metric	Value
Accuracy (#correct)	2129
Accuracy (%correct)	75.68432
Specificity	0.744639
Sensitivity (Recall)	0.78975
Precision	0.534222
F1 score	0.637328
Success Class	Yes
Success Probability	0.290166

Model Performance Comparison

Default Cutoff (0.50)

- Accuracy: 80.27%
- Recall: 0.5598
- Precision: 0.6594
- F1-Score: 0.6055

Optimized Cutoff (0.29)

- Accuracy: 75.68%
- Recall: 0.7898 ⬆
- Precision: 0.5342
- F1-Score: 0.6373 ⬆

Lowering the cutoff dramatically improved recall from 56% to 79%, capturing far more at-risk customers. The F1-score improvement confirms better balance for churn prevention priorities.

Neural Network

Model Architecture

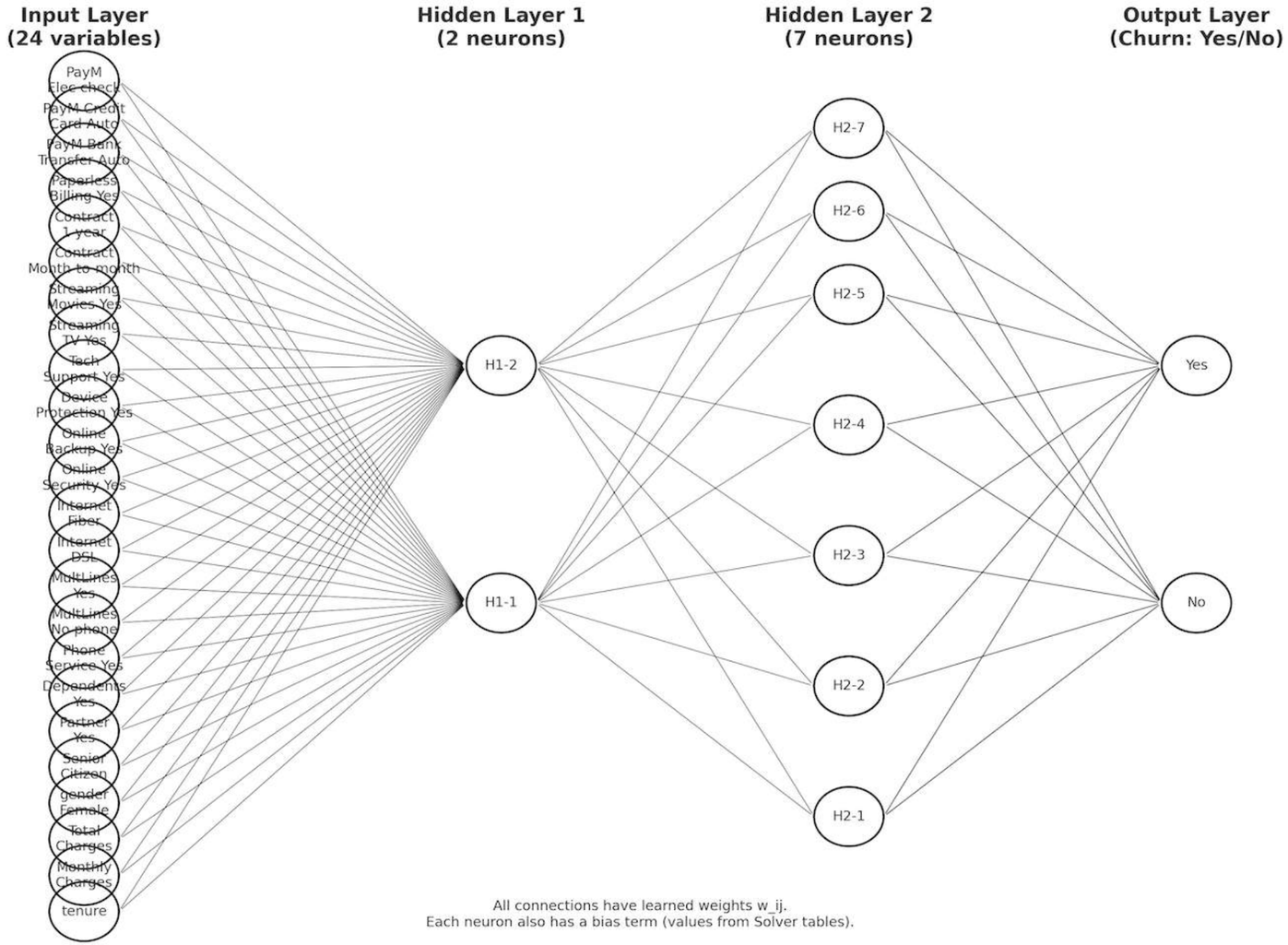
Best Model:

Tested 5 top networks with highest F1-score; Net36 achieved highest discriminatory power.

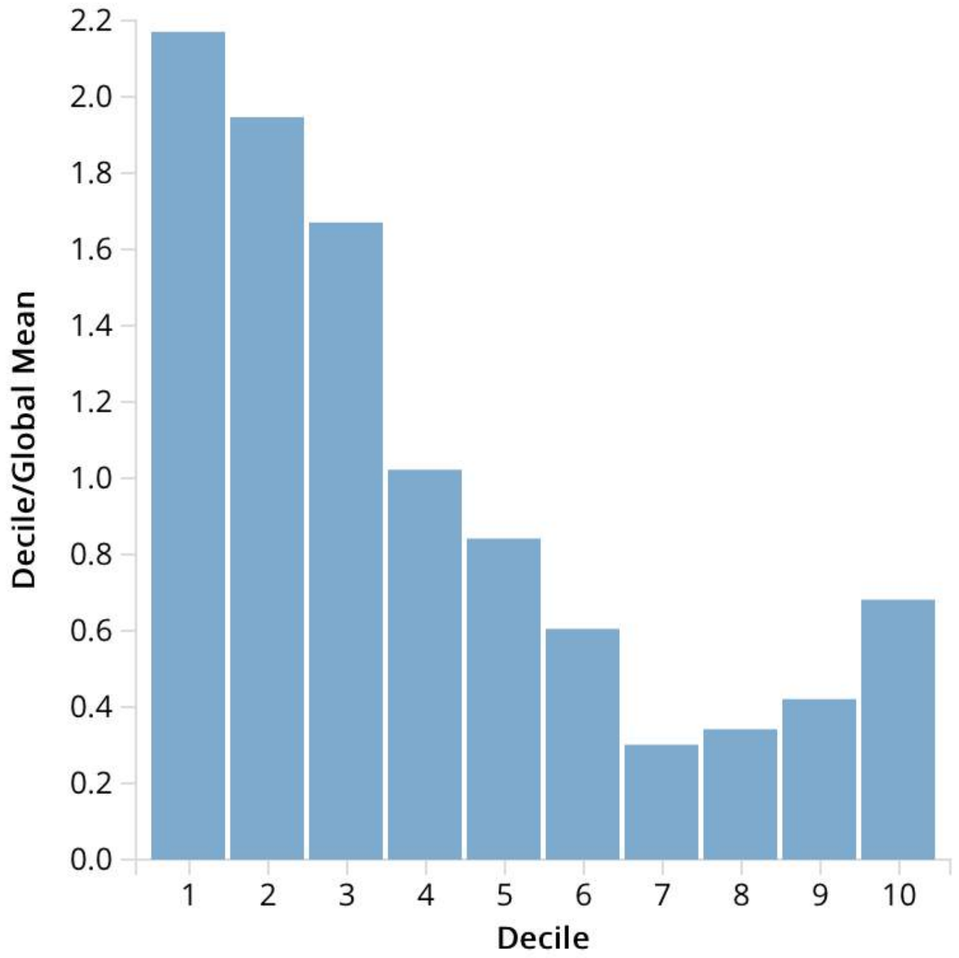
2 hidden layers:

- Hidden layer 1: 2 nodes
- Hidden layer 2: 7 nodes

Learning Rate: 0.1



Determine Appropriate Cutoff



To maximize campaign efficiency, we recommend targeting the top 40% of customers ranked by predicted churn probability (Deciles 1–4).

New cutoff = **0.29318**

Validation: Classification Summary cutoff = 0.29

Confusion Matrix		
Actual\Predicted	No	Yes
No	1445	607
Yes	243	518

Error Report			
Class	# Cases	# Errors	% Error
No	2052	607	29.58089669
Yes	761	243	31.93166886
Overall	2813	850	30.21685034

Metrics	
Metric	Value
Accuracy (#correct)	1963
Accuracy (%correct)	69.78315
Specificity	0.704191
Sensitivity (Recall)	0.680683
Precision	0.460444
F1 score	0.549311
Success Class	Yes
Success Probability	0.293181

Validation: Classification Summary cutoff = 0.3

Confusion Matrix		
Actual\Predicted	No	Yes
No	1806	246
Yes	455	306

Error Report			
Class	# Cases	# Errors	% Error
No	2052	607	29.5809
Yes	761	243	31.93167
Overall	2813	850	30.21685

Metrics	
Metric	Value
Accuracy (#correct)	2112
Accuracy (%correct)	75.07999
Specificity	0.880117
Sensitivity (Recall)	0.402102
Precision	0.554348
F1 score	0.466108
Success Class	Yes
Success Probability	0.3

Model Performance Comparison

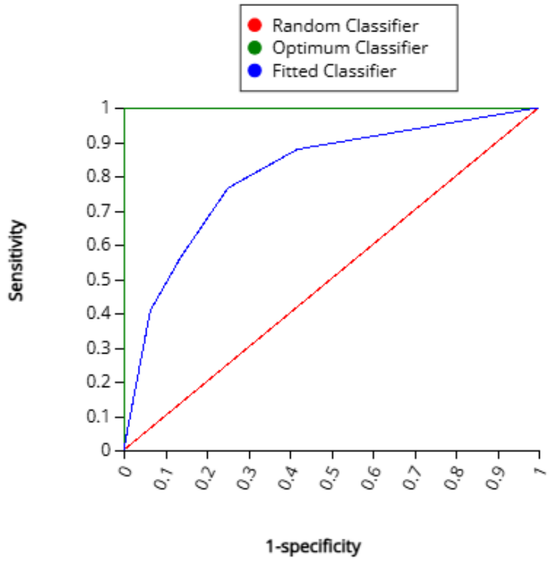
The cutoff = 0.29 is superior for churn prediction because it provides a stronger Recall + F1-score, which are the priority metrics.

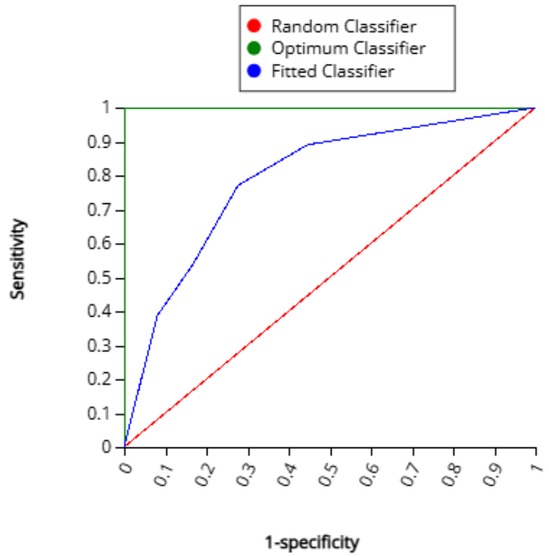
Decision Tree

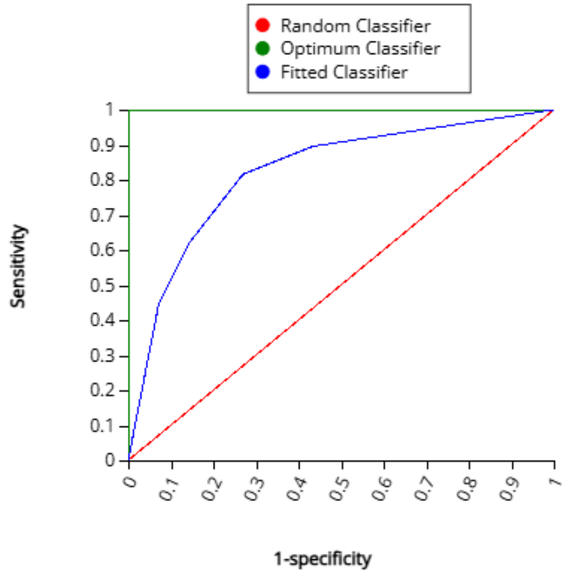
Decision Tree: Optimal Model Selection

We tested minimum leaf sizes of 300, 400, 500, 600, and 700 to identify the optimal model complexity. The tree with a limit of **300 achieved the highest validation ROC AUC**, while remaining interpretable. This model offered the best balance between predictive performance and structural simplicity.

Training ROC AUC	Validation ROC AUC	Test ROC AUC
AUC = 0.80776	**AUC = 0.79147**	**AUC = 0.822**

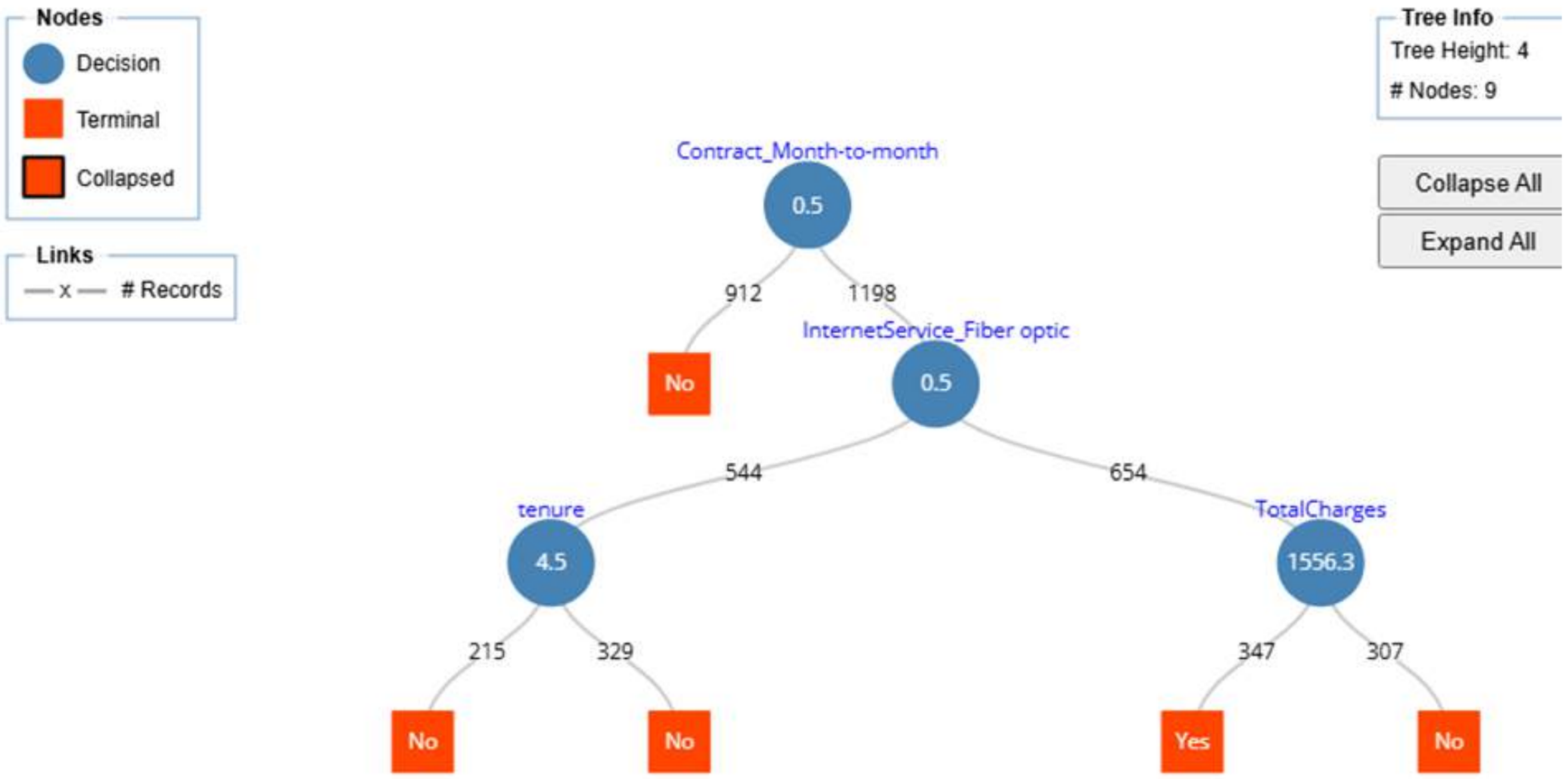




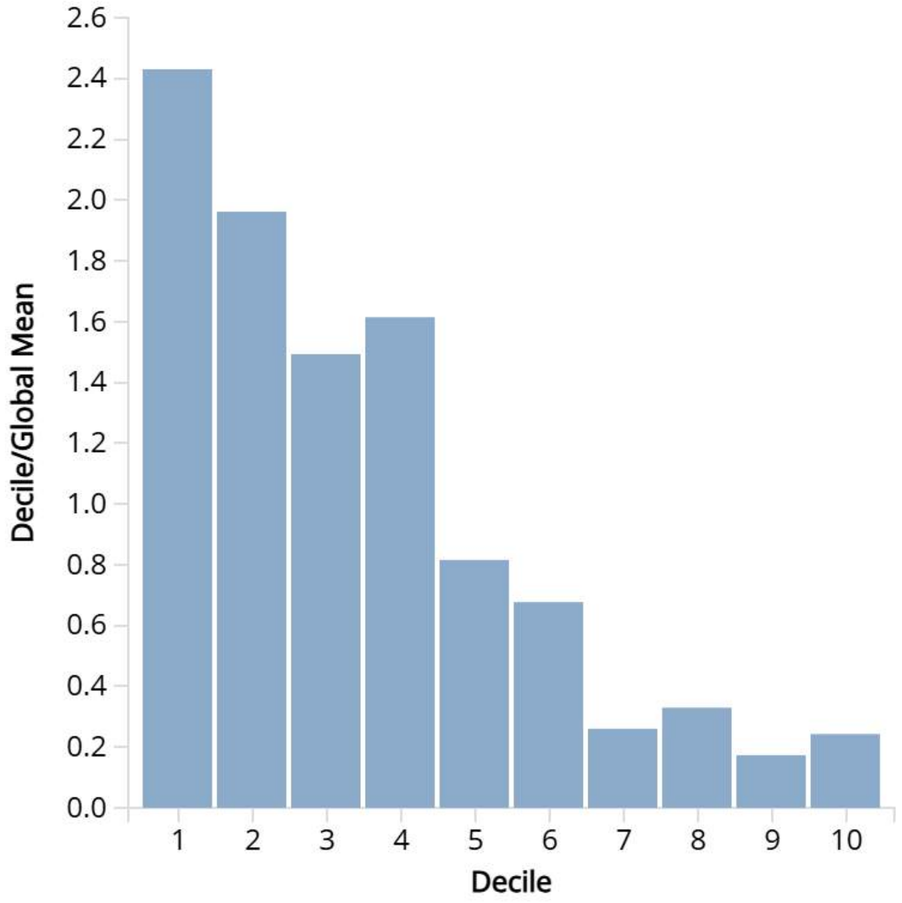


Best Pruned Tree:

- Limit 300
- Cut-off = 0.5



Determine Appropriate Cutoff

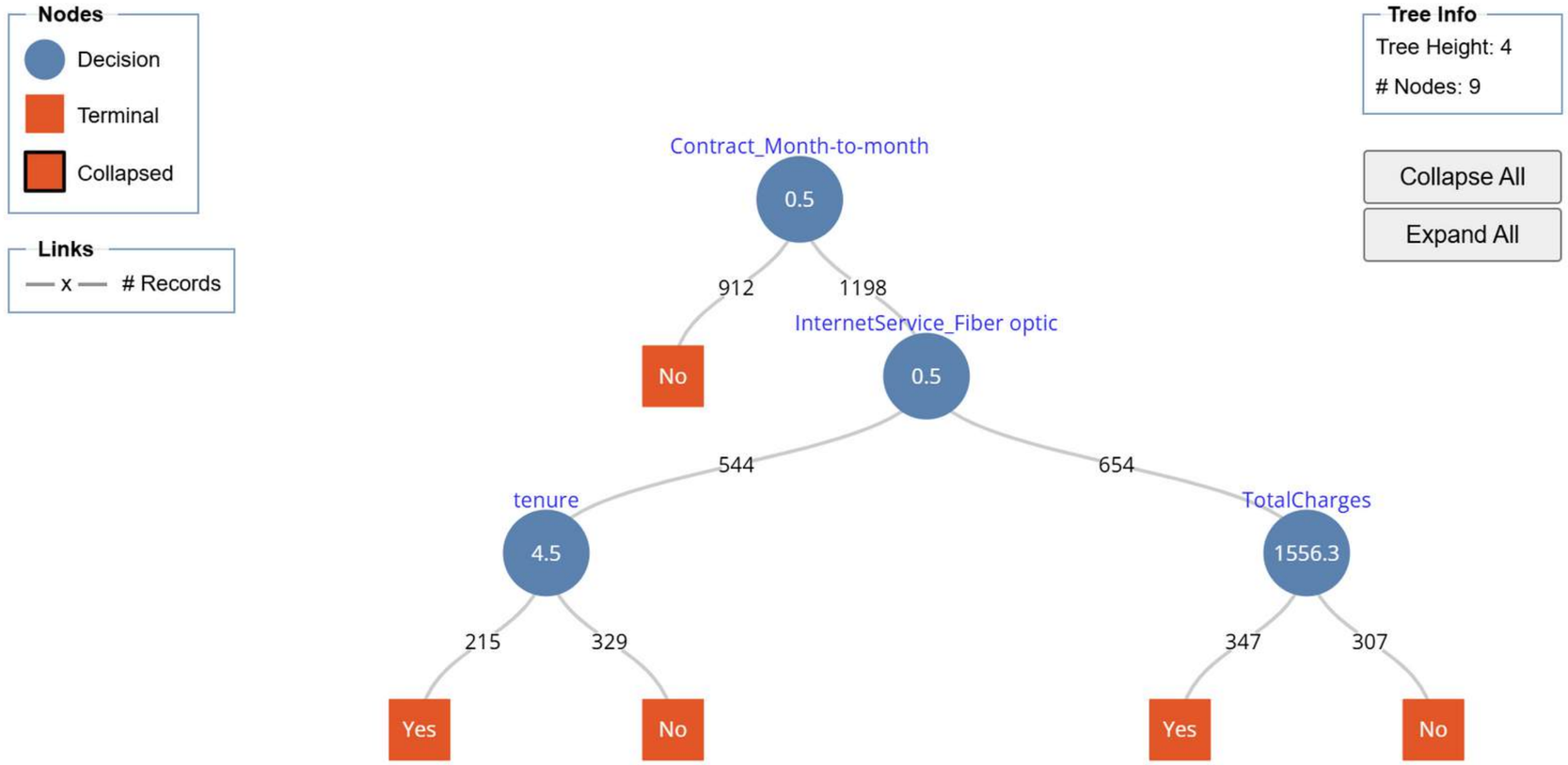


To maximize retention strategy efficiency, we analyzed decile groups for churn probability. We recommend targeting customers in the top 40% risk group (Deciles 1–4).

New cutoff = **0.389796**

Best Pruned Tree:

- Limit 300
- New cut-off = 0.38976



Validation: Classification Summary

Confusion Matrix		
Actual\Predicted	No	Yes
No	1410	124
Yes	353	223

Error Report			
Class	# Cases	# Errors	% Error
No	1534	124	8.083441982
Yes	576	353	61.28472222
Overall	2110	477	22.60663507

Metrics	
Metric	Value
Accuracy (#correct)	1633
Accuracy (%correct)	77.393365
Specificity	0.9191656
Sensitivity (Recall)	0.3871528
Precision	0.6426513
F1 score	0.4832069
Success Class	Yes
Success Probability	0.5

Cutoff = 0.5

Validation: Classification Summary

Confusion Matrix		
Actual\Predicted	No	Yes
No	1280	254
Yes	268	308

Error Report			
Class	# Cases	# Errors	% Error
No	1534	254	16.55801825
Yes	576	268	46.52777778
Overall	2110	522	24.73933649

Metrics	
Metric	Value
Accuracy (#correct)	1588
Accuracy (%correct)	75.26066
Specificity	0.83442
Sensitivity (Recall)	0.534722
Precision	0.548043
F1 score	0.541301
Success Class	Yes
Success Probability	0.389796

Cutoff = 0.38976

Model Performance Comparison

Lowering the cutoff increased recall, allowing the model to identify more churners, while keeping a balanced F1-score. This improves retention-focused decision-making.

Decision Rules from Best Tree

01

Long-term contracts

If the Contract is NOT Month-to-Month

Predict: No Churn

02

Short tenure, no fiber

IF Month-to-Month AND No Fiber AND

Tenure < 4.5

Predict: Yes churn

03

Longer tenure, no fiber

IF Month-to-Month AND No Fiber AND

Tenure ≥ 4.5

Predict: No churn

04

High-risk group

IF Month-to-Month AND Fiber AND TotalCharges ≤ $1,556 →

Predict: **Yes, churn**

05

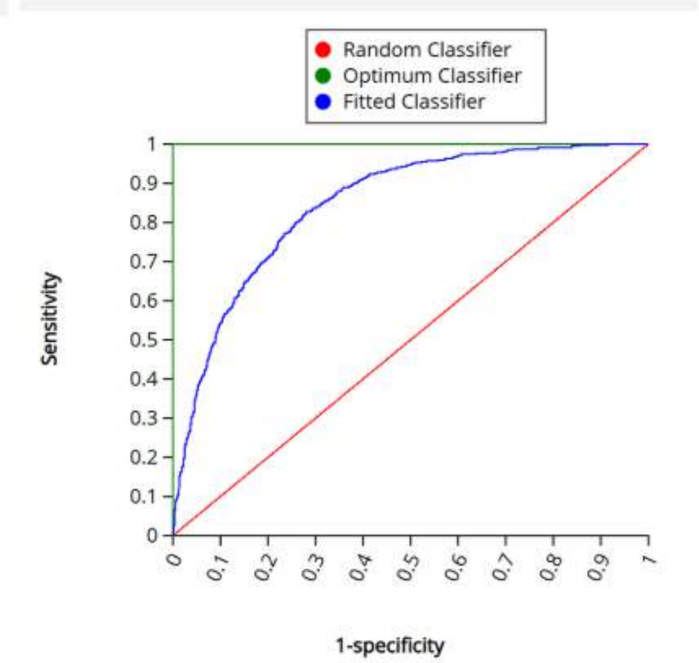
Loyal fiber customers

IF Month-to-Month AND Fiber AND TotalCharges > $1,556 →

Predict: No churn

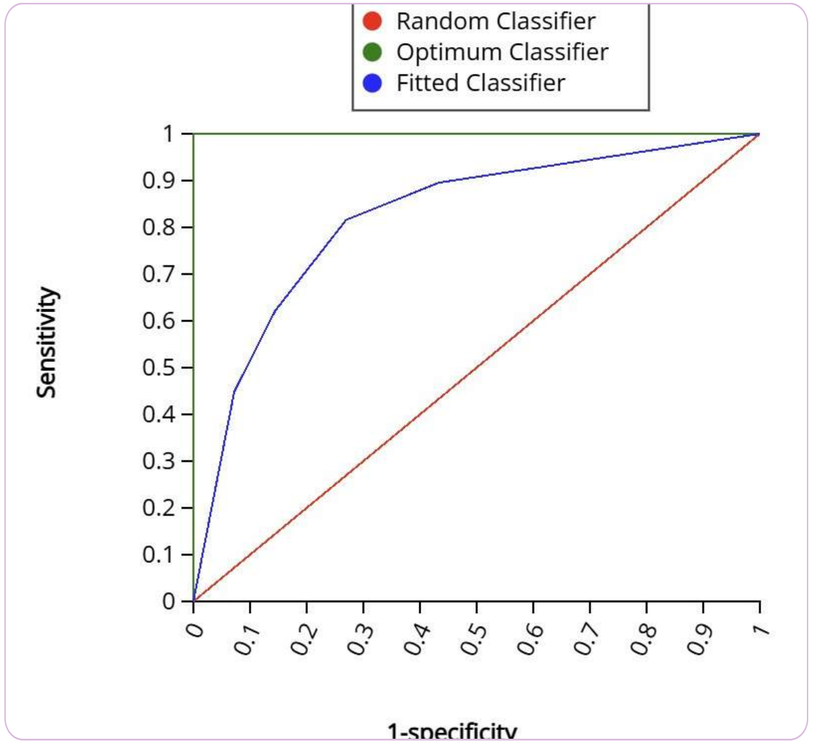
Model Comparison

Logistic Regression



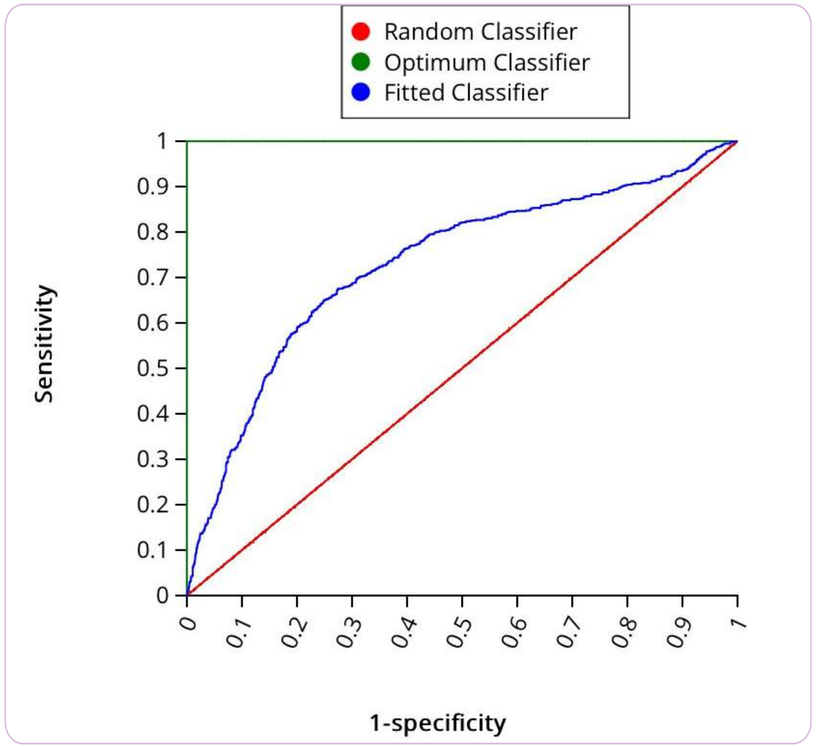
AUC= 0.8467

Decision Tree



AUC= 0.822

Neural Network



AUC= 0.73

Hypothetical Example

Logistic Regression

P(Churn=1) = 0.83 > Cutoff = 0.29

=> This customer would be likely to churn.

Decision Tree

1. Contract = Month-to-Month → go right
2. Internet Service = Fiber Optic → go right
3. Total Charges = 250 ≤ 1556.3 → go left → YES

=> This customer would be likely to churn.

Neural Network

Output_1 $= \frac{1}{1+e^{-1.46}} \approx 0.81 > 0.29$

=> This customer would be likely to churn.

Variable	Value
Tenure	5 months
Total Charges	$250
Senior Citizen	No
Phone Service	Yes
Internet Service	Fiber Optic
Online Security	No
Tech Support	No
Streaming Movies	Yes
Contract	Month-to-Month
PaperlessBilling	Yes
Payment Method	Electronic Check

Business Question Impact

Highest churn risk occurs in Month-to-Month, Fiber users with low tenure/low TotalCharges. Targeting the top 40% highest-risk customers in the high-revenue segment yields the greatest ROI, since preventing their churn directly protects future revenue (probability × monthly revenue × expected months).

Long-tenure, high-TotalCharges customers on 1–2 year contracts have the highest CLV and lowest churn risk. In contrast, Month-to-Month, low-tenure customers churn early and have low CLV. Moving customers to longer contracts and supporting them through their first few months increases lifetime value.

OnlineSecurity and TechSupport bundles lower churn, while Fiber-only or streaming-only customers are much higher risk. Bundling Fiber + support services reduces churn, and these high-risk unbundled customers should be prioritized for retention or upsell offers.

Key Insights

Contract Type Matters Most

All models agree Month-to-month customers are far more likely to churn than customers on 1–2 year contracts.

Fiber Optic Risk

Across Logistic, Tree, and Neural Net, new Fiber users with low total charges show the highest churn probabilities and should be prioritized for retention.

Security Services Reduce Churn

Online security, tech support, and longer contracts consistently appear as protective factors, creating stickier, higher-lifetime-value customers.

Strategic Recommendations

1

Convert Month-to-Month Customers to Longer Contracts

Month-to-month users have the highest churn risk → offer discounts, loyalty benefits, or limited-time upgrades to 1-year or 2-year plans.

2

Improve Experience for Fiber Optic Users

Fiber optic customers show significantly higher churn → address service quality issues, provide proactive support, and offer performance guarantees.

3

Promote OnlineSecurity & TechSupport Bundles

These features strongly reduce churn → provide free trials or bundle them with mid-tier plans to improve retention.

4

Focus Retention Efforts on Early-Tenure Customers

Most churn happens within the first few months → implement onboarding programs, follow-up calls, and early troubleshooting assistance.

Thank You