**PROJECT REPORT**

**Project Name**: Customer Churn Analysis for Telecom

**Course**: SCH-MGMT 655 - Machine Learning for Analytics

**Team Members**: Haley Hoang, Sakshi Agarwal, Noynicaa Santani

Table of Contents

[1. Project description 5](#_Toc216642205)

[1.1. Business context and goals of the project 5](#_Toc216642206)

[1.1.1. Context of the project 5](#_Toc216642207)

[1.1.2. Goal of the project 5](#_Toc216642208)

[1.2. Business questions 5](#_Toc216642209)

[1.2.1. What is the dollar value of reducing churn in the top 20% of high-revenue customers, and which interventions (contracts, bundles, offers) maximize ROI? 5](#_Toc216642210)

[1.2.2. Which customer segments (by contract type, tenure, and billing method) contribute most to total lifetime revenue, and how could retention offers improve customer lifetime value (CLV)? 6](#_Toc216642211)

[1.2.3. Which service combinations most strongly affect churn risk? 6](#_Toc216642212)

[2. Data Preprocessing 6](#_Toc216642213)

[2.1. Random sampling 6](#_Toc216642214)

[2.2. Handling missing data 6](#_Toc216642215)

[2.3. Summary characterizes 7](#_Toc216642216)

[2.4. Correlation table 8](#_Toc216642217)

[2.5. Histogram, scatterplot, boxplot (dependent variable vs. important independent variables) 8](#_Toc216642218)

[2.5.1. Histogram Interpretation 8](#_Toc216642219)

[2.5.2. Boxplot Interpretation 10](#_Toc216642220)

[3. Model 14](#_Toc216642221)

[3.1. Model 1 – Logistic Regression 14](#_Toc216642222)

[3.1.1. Model identification and rationale 14](#_Toc216642223)

[3.1.2. Variables selection 14](#_Toc216642224)

[3.1.3. Variable selection techniques applied 15](#_Toc216642225)

[3.1.4. Model output 16](#_Toc216642226)

[3.1.5. Summary report for training, validation, and test data 18](#_Toc216642227)

[3.1.6. Model performance 19](#_Toc216642228)

[3.1.7. Determine optimal cutoff value 20](#_Toc216642229)

[3.2. Model 2 – Neural Network 21](#_Toc216642230)

[3.2.1. Model identification 21](#_Toc216642231)

[3.2.2. Variable selection 21](#_Toc216642232)

[3.2.3. Variable selection techniques applied 22](#_Toc216642233)

[3.2.4. Model output 23](#_Toc216642234)

[3.2.5. Determine optimal cutoff value 24](#_Toc216642235)

[3.2.6. Model performance 25](#_Toc216642236)

[3.3. Model 3 – Decision Tree 25](#_Toc216642237)

[3.3.1. Model identification 25](#_Toc216642238)

[3.3.2. Variables selection 26](#_Toc216642239)

[3.3.3. Variable selection techniques applied 26](#_Toc216642240)

[3.3.4. Model output 27](#_Toc216642241)

[3.3.5. Model Selection – Full Tree, Pruned Tree 29](#_Toc216642242)

[3.3.6. Determine optimal cutoff value 32](#_Toc216642243)

[3.3.7. Decision Rules from Final Tree (Minimum Leaf Size = 300) 33](#_Toc216642244)

[4. Compare models 35](#_Toc216642245)

[5. Hypothetical example 37](#_Toc216642246)

[6. Recommendations 38](#_Toc216642247)

[6.1. Explain how the results address your business questions 38](#_Toc216642248)

[6.2. Business Strategy 39](#_Toc216642249)

[7. Project Summary 42](#_Toc216642250)

[7.1. Lessons learned from the project 42](#_Toc216642251)

[7.2. Dataset limitations and suggestions for future data extensions 42](#_Toc216642252)

[8. Task Allocation 43](#_Toc216642253)

# Project description

# Business context and goals of the project

# *Context of the project*

This project focuses on understanding and predicting customer churn in the telecom industry using the IBM Telco Customer Churn dataset, which contains detailed demographic, service usage, billing, and contract information for over 7,000 customers.

Churn is a critical challenge for subscription-based businesses because retaining customers is far more cost-effective than acquiring new ones. By analyzing customer behavior patterns and identifying early warning signs of churn, the company can take proactive steps to retain high-value customers, improve satisfaction, and enhance long-term profitability.

The dataset provides a comprehensive view of customer characteristics and behaviors, enabling us to analyze why customers leave and which attributes distinguish high-risk individuals from loyal ones.

# *Goal of the project*

The goal of this project is to build predictive models that estimate each customer's probability of churning and to uncover the key factors that drive churning behavior. These predictions help translate raw customer data into actionable insights that can guide retention strategies, identify vulnerable segments, and support business decisions related to contracts, service bundles, and customer engagement.

Ultimately, the goal is not only to predict churn but to understand it - so the business can design targeted interventions, reduce attrition, and improve customer lifetime value.

# Business questions

# *What is the dollar value of reducing churn in the top 20% of high-revenue customers, and which interventions (contracts, bundles, offers) maximize ROI?*

High-revenue customers generate a disproportionate share of overall income, making their retention crucial to business performance. Reducing churn in this segment not only preserves significant revenue but also highlights the financial upside of targeted retention efforts. The objective is to quantify the revenue at risk if these customers leave and to determine which interventions - such as contract shifts, service bundles, or loyalty offers - provide the greatest return on investment.

# *Which customer segments (by contract type, tenure, and billing method) contribute most to total lifetime revenue, and how could retention offers improve customer lifetime value (CLV)?*

This helps the company find out which types of customers bring in the most money over time based on their contract type, how long they’ve stayed (tenure), and their billing method. By predicting which of these customers are likely to leave, the analysis can show where to focus on retention offers or discounts to keep them longer. This way, the company can increase customer lifetime value (CLV), reduce churn, and maintain steady revenue growth.

# *Which service combinations most strongly affect churn risk?*

IBM Telco Customer Churn dataset includes detailed information about which add-on services (such as Online Security, Tech Support, Streaming TV, and Streaming Movies) each customer subscribes to. These features directly represent the service bundle choices customers make and are likely major drivers of customer satisfaction or frustration. Unlike demographic or payment features, service combinations are actionable for levers the company can control, it can change packages, pricing, or promotional offers.

Understanding how different service bundles affect churn risk allows the business to design smarter retention strategies, not just predict churn.

# Data Preprocessing

# Random sampling

The cleaned dataset contains 7,032 records, which is a manageable size for analysis and modeling. Therefore, random sampling is not required. Instead, the full dataset is used, and training/validation partitions will be created in later steps to ensure unbiased model evaluation.

# Handling missing data

After importing the dataset into Excel, the total number of records was 7,043. To identify missing values, the Missing Data Handling tool in Excel (Data Science → Transform → Missing Data Handling) was used.

Steps performed:

* Imported the dataset into Excel
* Navigated to Data Science → Transform → Missing Data Handling
* Selected all variables to ensure a complete scan for missing values
* Choose Delete Records to remove any rows containing missing data
* Verified that the transformation was applied correctly

The variable Total Charges contained 11 missing values. Since this represents less than 0.2% of all observations, removing these records does not materially impact the analysis. After deletion, the dataset was reduced from 7,043 to 7,032 for complete records, ensuring that all remaining observations are valid and ready for modeling.

**Why Deleting Missing Values Was the Correct Choice**

We deleted the 11 incomplete rows instead of imputing values because:

* The missing percentage was extremely small (<0.2%) → Imputation adds unnecessary complexity without meaningful benefit.
* The variable with missing values (Total Charges) is highly correlated with Tenure, so imputing it could introduce bias or distort the relationship.
* Deletion maintains data purity for modeling → Imputing financial variables (like Total Charges) risks artificially inflating or deflating churn risk.
* No pattern existed in the missing data → It appeared random, not systematic, so removing those rows does not damage representativeness.

This makes deletion the most appropriate, simple, and statistically safe approach.

# Summary characterizes

We summarized the three numerical variables in the dataset: Tenure, Monthly Charges, and Total Charges.

**Tenure:**

* The mean is higher than the median, indicating that many customers have short tenures while fewer stay long-term.
* The standard deviation is moderate, showing normal variation in customer loyalty.
* A few long-tenure values appear as outliers, but these represent genuine long-standing customers.

**Monthly Charges:**

* The mean and median are close, with only mild skewness.
* The standard deviation is moderate, reflecting differences in plan types and service bundles.
* High Monthly Charges values are valid and correspond to premium service users.

**Total Charges:**

* The mean is much higher than the median, due to many low-spending short-tenure customers and fewer high-spending long-tenure customers.
* The standard deviation is high, reflecting large variability in customer lifetime spending.
* Very high Total Charges values are valid; however, records with Total Charges = 0 but tenure > 0 were identified as data errors and removed.

# Correlation table

A screenshot of a table

AI-generated content may be incorrect.

**Interpretation:**

* Total Charges and tenure have a strong positive correlation (r ≈ 0.83). This makes sense because the longer a customer stays with the company, the more they pay overall.
* Monthly Charges and Total Charges have a moderate positive correlation (r ≈ 0.65), indicating customers with higher monthly bills tend to accumulate higher total payments.
* Monthly charges and tenure have a weak positive correlation (r = 0.248), suggesting that the amount billed per month does not strongly depend on how long a customer has stayed (likely due to different plans).

**Overall:**

Total Charges are strongly driven by tenure, while Monthly Charges vary based on service plan choices rather than customer loyalty duration.

# Histogram, scatterplot, boxplot (dependent variable vs. important independent variables)

# Histogram Interpretation

**Tenure by Churn**

A graph of a number of numbers and a bar

AI-generated content may be incorrect.

The tenure distribution is heavily right-skewed, meaning most customers leave within the first few months of service, while relatively few remain loyal long-term. From a business perspective, this suggests that the early stage of the customer lifecycle is the most critical period for churn prevention - if customers stay past the initial months, they are much more likely to remain with the company.

**Total Charges by Churn**

A graph of a graph of a number of charge

AI-generated content may be incorrect.

The Total Charges distribution is strongly right skewed, meaning most customers have paid relatively low total amounts due to short tenures, while only long-standing customers accumulate high total charges over time. This highlights that customer retention directly drives revenue — the longer customers stay, the more value they generate for the business.

**Monthly Charges by Churn**

A screenshot of a graph

AI-generated content may be incorrect.

The Monthly Charges distribution appears bimodal, showing two distinct customer groups, one paying lower subscription fees and another paying significantly higher rates. This indicates the presence of multiple pricing tiers or service bundles. From a customer behavior standpoint, spending varies more by plan type than by loyalty or usage.

# Boxplot Interpretation

**Monthly Charges by Churn**

A screenshot of a graph

AI-generated content may be incorrect.

The boxplot compares the monthly charges of customers who churned (Yes) and those who did not churn (No).

Customers who churned generally have higher median monthly charges compared to those who stayed.

The interquartile range (IQR) for both groups overlaps, but the upper quartile for churned customers extends slightly higher, indicating that customers with expensive monthly plans are more likely to leave.

Both groups have similar minimum and maximum charge ranges (around $20–$120), showing that churn occurs across all plan levels, though it is more frequent among high-paying customers.

There are a few outliers visible in both groups, which could represent customers with unusually high service combinations or add-ons.

**Total Charges by Churn**

A graph of a graph showing the amount of charge

AI-generated content may be incorrect.

The median Total Charges is much higher for customers who did not churn compared to those who left.

This indicates that long-term customers (who stayed) have accumulated significantly higher total payments, reflecting stronger loyalty and longer tenure.

Customers who churned tend to have lower Total Charges, meaning they left early in their subscription period before generating large revenue.

The spread (IQR) for non-churned customers is wider, showing a broader range of total spending levels among loyal customers.

A few outliers above the upper whisker (above roughly $8,000–$9,000) represent customers with very high lifetime payments, likely from long-term, high-value users.

These high-value outliers all belong mostly to the non-churn group, reinforcing that retaining customers longer directly increases revenue.

**Tenure by Churn**

A graph of a graph showing the different types of numbers

AI-generated content may be incorrect.

The median tenure is much higher for customers who did not churn, indicating they have been with the company longer.

Customers who churn (Churn = Yes) tend to have shorter tenures, showing that many leave within their first few months or years of service.

The spread (IQR) for non-churned customers is broader, reflecting that loyal customers vary widely in how long they have stayed - from mid-term to long-term contracts.

Outliers above the upper whisker represent a few customers with extremely long tenures (above ~70 months); these long-term users are highly retained and likely satisfied with service quality.

The overall pattern confirms that the longer a customer stays, the lower their likelihood of churning, emphasizing the importance of early-stage retention strategies to prevent customers from leaving within their first year.

# Model

# Model 1 – Logistic Regression

# Model identification and rationale

We have used Logistic Regression, Decision Tree, and Neural Network as our three distinct models.

Logistic Regression was selected because it directly supports all three of our business questions by providing customer-level churn probabilities that can be translated into financial and strategic insights. For high-revenue customers, logistic regression allows us to quantify expected revenue at risk by multiplying churn probability with monthly revenue and expected remaining months and then compare pre- and post-intervention probabilities to estimate dollar savings and ROI. It also helps evaluate customer segments—such as contract types, tenure groups, and billing methods—by assigning each segment an average churn probability, which allows us to infer their relative Customer Lifetime Value (CLV), since lower churn probability indicates higher CLV. This makes it possible to identify which segments benefit most from retention offers or contract/billing adjustments. Additionally, logistic regression reveals how individual services and service combinations influence churn through coefficients and odds ratios, helping us distinguish protective bundles (like Internet + TechSupport + Security) from high-risk ones.

Overall, logistic regression provides both predictive accuracy and deep behavioral insight, making it the most suitable model for linking churn risk to revenue, customer value, and service-mix decisions.

# Variables selection

In the logistic regression model, we included all 21 available predictor variables from the Telco Customer Churn dataset (excluding only the customerID). The decision to use all variables was intentional because each feature contributes meaningful information that supports both accurate churn prediction and the three strategic business questions addressed in this project.

First, all demographic variables (gender, SeniorCitizen, Partner, Dependents) were retained because they help capture differences in customer stability and behavior across population groups, improving the model’s ability to identify high-risk customer profiles.

Next, all account and billing variables (tenure, Contract, Payment Method, Paperless Billing) were included since they directly relate to customer retention patterns and are essential for segment-level analysis of lifetime value.

Similarly, revenue variables such as Monthly Charges and Total Charges were used because they allow us to link churn probabilities with financial outcomes, which is crucial for estimating the dollar value of reducing churn among high-revenue customers.

Finally, all service subscription variables (Internet Service, Online Security, Online Backup, Device Protection, Tech Support, Streaming TV, Streaming Movies, Phone Service, Multiple Lines) were kept because they define the customer’s service bundle and are required to identify which combinations of services most strongly increase or reduce churn risk.

By including all 21 variables, the model captures the full behavioral, financial, and usage profile of each customer, ensuring that the churn predictions appropriately support the analysis of revenue risk, customer segment lifetime value, and the retention impact of different service bundles.

# Variable selection techniques applied

In this project, we used stepwise selection combined with best‐subset evaluation to identify the most statistically significant and business-relevant predictors of customer churn.

Stepwise selection was applied first to iteratively add and remove variables based on their contribution to model fit, statistical significance, and improvement in prediction accuracy. This technique allowed the model to automatically filter out weaker or redundant variables, ensuring that each predictor included an added meaningful value.

After stepwise screening, we used the best-subset comparison to confirm that the final chosen to set of variables produced the strongest model performance among all possible combinations.

The chosen subset includes the following variables:

* tenure
* TotalCharges
* SeniorCitizen
* PhoneService
* InternetService (DSL, Fiber)
* OnlineSecurity
* TechSupport
* StreamingMovies
* Contract (Month-to-month, One-year)
* PaperlessBilling
* PaymentMethod (Electronic check)

# Model output

A screenshot of a computer

AI-generated content may be incorrect.

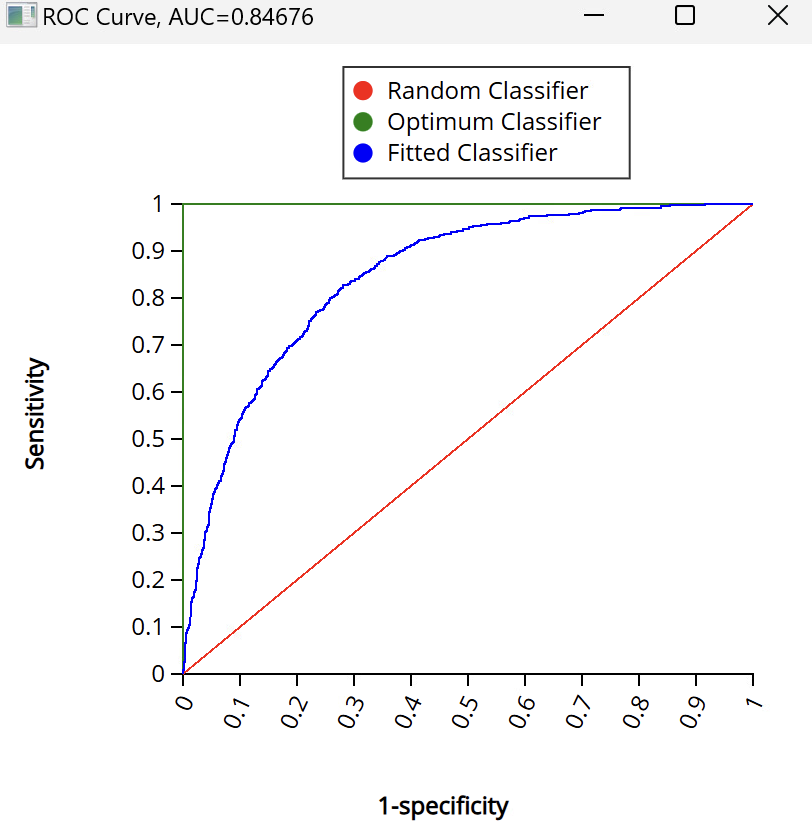
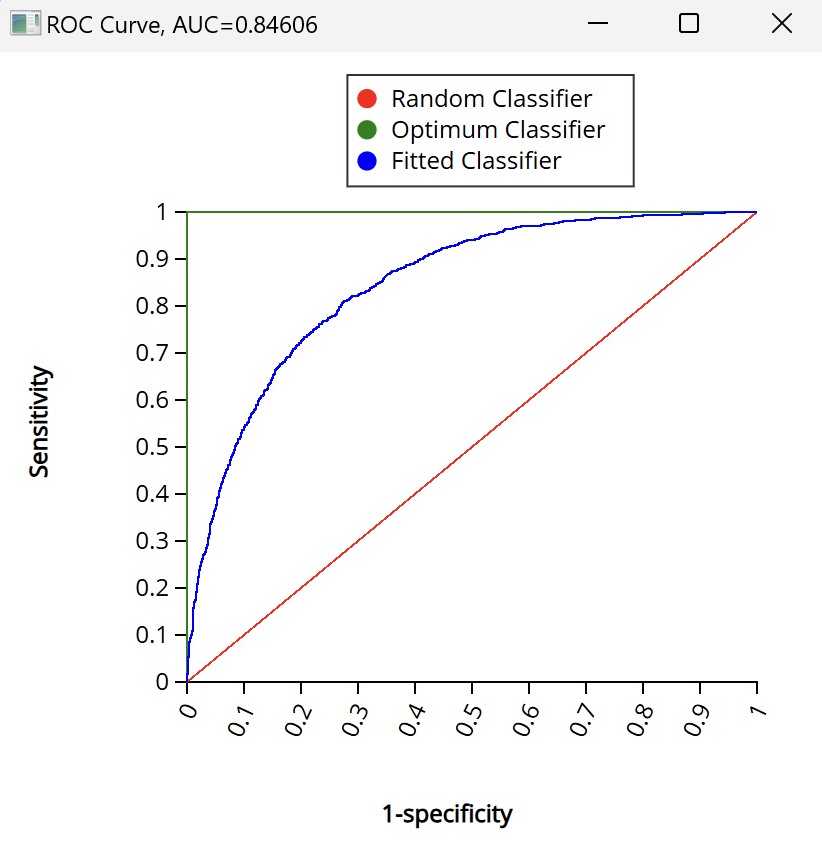
**Equation**

* Logit(Churn=1) = -2.03 - 0.06tenure + 0.0003TotalCharges + 0.25SeniorCitizen\_1 - 0.6PhoneService\_Yes + 0.63InternetService\_DSL + 1.46InternetService\_Fiber optic - 0.42OnlineSecurity\_Yes - 0.42TechSupport\_Yes + 0.27StreamingMovies\_Yes + 1.32Contract\_Month-to-month + 0.65Contract\_One year + 0.39PaperlessBilling\_Yes + 0.41PaymentMethod\_Electronic check
* Odds(Churn=1) = e^(-2.03 - 0.06tenure + 0.0003TotalCharges + 0.25SeniorCitizen\_1 - 0.6PhoneService\_Yes + 0.63InternetService\_DSL + 1.46InternetService\_Fiber optic - 0.42OnlineSecurity\_Yes - 0.42TechSupport\_Yes + 0.27StreamingMovies\_Yes + 1.32Contract\_Month-to-month + 0.65Contract\_One year + 0.39PaperlessBilling\_Yes + 0.41PaymentMethod\_Electronic check)
* P(Churn=1) = 1 / 1+e^-(-2.03 - 0.06tenure + 0.0003TotalCharges + 0.25SeniorCitizen\_1 - 0.6PhoneService\_Yes + 0.63InternetService\_DSL + 1.46InternetService\_Fiber optic - 0.42OnlineSecurity\_Yes - 0.42TechSupport\_Yes + 0.27StreamingMovies\_Yes + 1.32Contract\_Month-to-month + 0.65Contract\_One year + 0.39PaperlessBilling\_Yes + 0.41PaymentMethod\_Electronic check)

**Coefficient Interpretation**

* For each additional month the customer stays with the company; odds of customer leaving the company would be multiplied by 0.93, holding other variables constant.
* For each additional total amount charged during the customer’s tenure, odds of customer churn would be multiplied by 1, holding other variables constant.
* The odds of a customer who is a senior citizen leaving the company would be 1.29 times the odds of a customer who is not a senior citizen leaving the company, holding other variables constant.
* The odds of a customer having an active phone line leaving the company would be 0.54 times the odds of a customer not having an active phone line leaving the company, holding other variables constant.
* The odds of a customer subscribing to DSL internet service leaving the company would be 1.89 times the odds of a customer not subscribing to an internet service leaving the company, holding other variables constant.
* The odds of a customer subscribing to Fiber optic internet service leaving the company would be 4.34 times the odds of a customer not subscribing to an internet service leaving the company, holding other variables constant.
* The odds of a customer having an online security add-on service leaving the company would be 0.65 times the odds of a customer not having an online security add-on service or not having internet service leaving the company, holding other variables constant.
* The odds of a customer having technical support service leaving the company would be 0.65 times the odds of a customer not having technical support service or not having internet service leaving the company, holding other variables constant.
* The odds of a customer subscribing to a streaming movies service leaving the company would be 1.31 times the odds of a customer not subscribing to a streaming movies service or not having internet service leaving the company, holding other variables constant.
* The odds of a customer having a month-to-month contract leaving the company would be 3.77 times the odds of a customer having a two-year contract leaving the company, holding other variables constant.
* The odds of a customer having a one-year contract leaving the company would be 1.92 times the odds of a customer having a two-year contract leaving the company, holding other variables constant.
* The odds of a customer using paperless billing leaving the company would be 1.48 times the odds of a customer not using paperless billing leaving the company, holding other variables constant.
* The odds of a customer using electronic check payment leaving the company would be 1.51 times the odds of a customer using Mailed check/Bank transfer (automatic)/Credit card (automatic) payment leaving the company, holding other variables constant.

# Summary report for training, validation, and test data

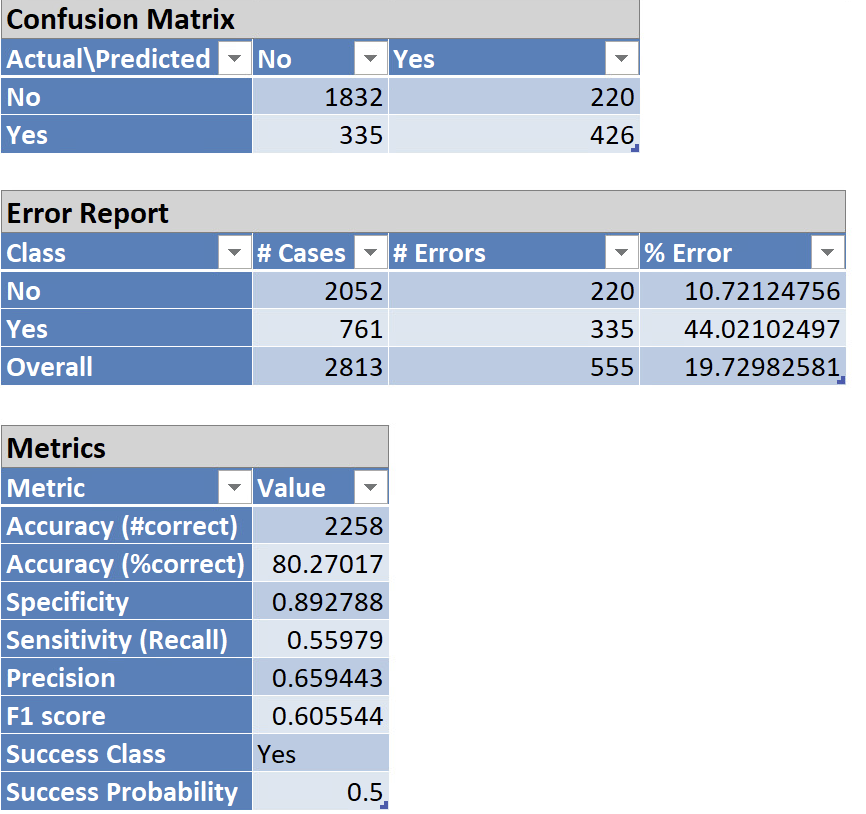


* Training AUC = 0.84606
* Validation AUC = 0.84676

This model works extremely well since AUC > 0.8. Training and validation AUC are nearly identical meaning there is no overfitting issue, and it can work well with new dataset.

Since logistic regression does not require a separate test dataset unless specified, training and validation AUC are sufficient.

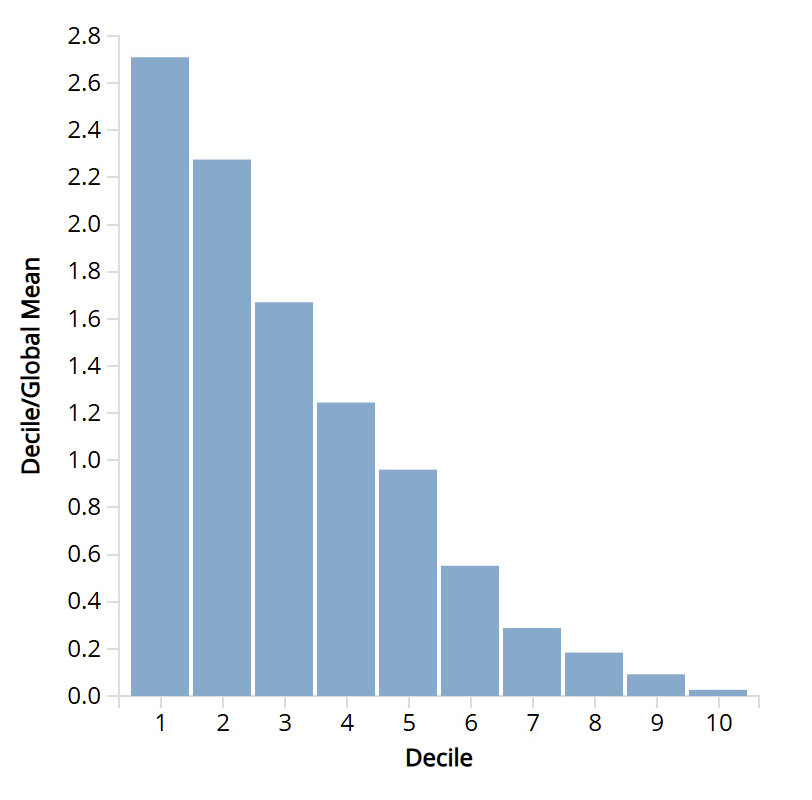
# Model performance



* **Precision = 0.659** => Out of all customers the model predicted would churn, 65.9% churned. This means the marketing budget is not wasted heavily on false alarms.
* **Sensitivity (Recall) = 0.559** => Out of all customers who truly churned, the model correctly identified 55.9% of them. This is crucial because the business goal is often to catch as many churners as possible.

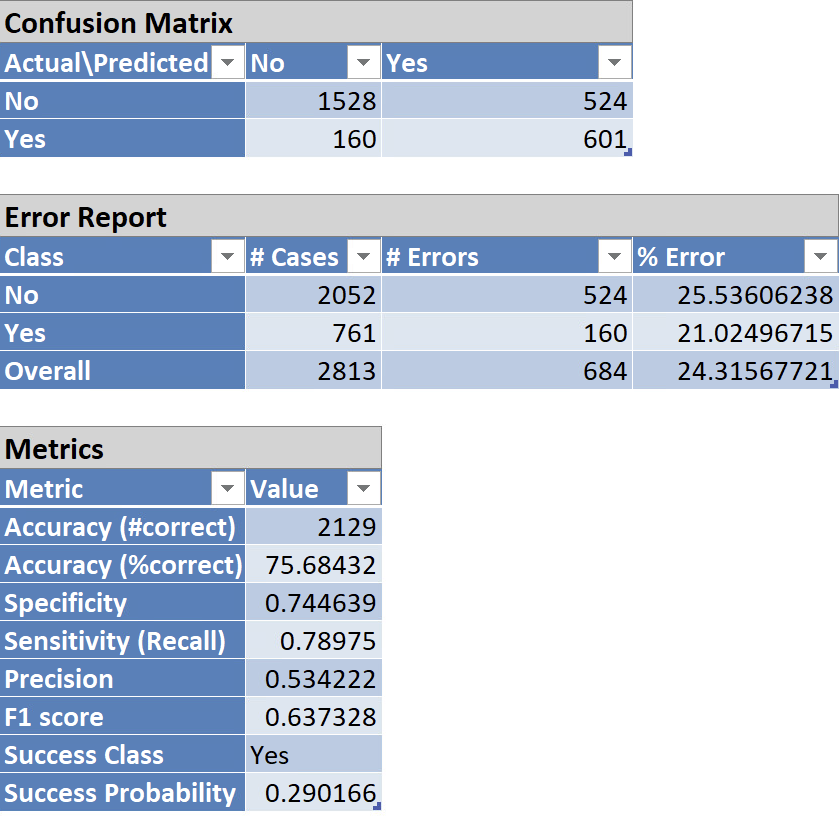
Together, these show balanced performance, not overly biased toward either class.

# Determine optimal cutoff value



Based on the decile chart, we choose to use decile 4 as the cutoff. In total, we have 2813 records in the validation data. After ordering the predicted probability of belonging to class 1, we choose the 1125th largest probability, which is in row 1161 = 2813 x 0.4 + 36 (the starting row).

We use 0.290166343 as the cutoff value for the new regression.



After lowering the cutoff from 0.5 to 0.29, the model becomes much more sensitive to churners (Recall increased from 0.56 to 0.79). Although accuracy and precision decrease, the F1 score improves, indicating better balance between capturing churners and controlling false positives. Since churn prevention prioritizes identifying as many at-risk customers as possible, the cutoff of 0.29 is more appropriate.

# Model 2 – Neural Network

# Model identification

The second classification model employed in this study is neural network. Neural networks are well suited for churn prediction problems due to their ability to capture nonlinear relationships and complex interactions among customer demographics, service subscriptions, contract attributes, and billing behaviors.

While logistic regression provides strong interpretability, it assumes linear relationships between predictors and the log-odds of churn. In contrast, customer churn behavior is often driven by nonlinear combinations of factors (e.g., interactions between contract type, internet service, and tenure). Therefore, a neural network was selected to maximize predictive performance and to assess whether nonlinear modeling could improve churn detection compared to traditional methods.

The neural network serves as a complementary model, prioritizing classification accuracy and recall, particularly for identifying high-risk churn customers, which aligns with the business objective of proactive retention.

# Variable selection

Predictor set includes 11 variables out of total 24 variables. The input variables for best network were chosen to concentrate model capacity on features that are both empirically strong churn correlates and operationally actionable levers for retention. The final predictor set includes: tenure, TotalCharges, SeniorCitizen, PhoneService, InternetService (DSL/Fiber optic), OnlineSecurity, TechSupport, StreamingMovies, Contract (Month-to-month/One-year), PaperlessBilling, and PaymentMethod (Electronic check).

The selected predictors capture three complementary dimensions of churn risk:

* Customer tenure and cumulative value (tenure, TotalCharges): represent lifecycle stage and realized spending, which are consistently associated with churn propensity in subscription settings.
* Contract and billing frictions (Contract, PaperlessBilling, PaymentMethod): encode commitment level and payment behaviors that often differentiate stable vs. high-risk segments and are directly targetable via retention offers (e.g., contract conversion, autopay incentives).
* Service experience and support “stickiness” (InternetService, OnlineSecurity, TechSupport, StreamingMovies, PhoneService): represent service type and add-ons that influence perceived value and switching costs; these are actionable through bundling, service assurance, and onboarding interventions.

# Variable selection techniques applied

The final selected model (Net36) was estimated using a reduced input set of 11 predictors rather than the full one-hot expanded design. Model selection followed a two-stage screening approach:

* Identify the top candidate networks by highest validation F1-score under an adjusted operational cutoff
* Select the candidate with the highest ROC/AUC among those top-F1 networks; this procedure yielded Net36 with AUC = 0.73.

Categorical predictors were dummy-coded relative to a baseline profile. The baseline levels used in the neural network specification were:

* Gender = Male
* Senior Citizen = 0
* Partner = No
* Dependents = No
* Phone Service = No
* Multiple Lines = No
* Internet Service = No
* Online Security = No
* Online Backup = No
* Device Protection = No
* Tech Support = No
* Streaming TV = No
* Streaming Movies = No
* Contract = Two year
* Paperless Billing = No
* Payment Method = Mailed check

# Model output

A screenshot of a screen

AI-generated content may be incorrect.

Among the five candidate neural networks ranked by validation F1-score, Net36 achieved the highest discriminatory performance, with a ROC AUC of 0.73, and was therefore selected as the final neural network model.

The chosen architecture consists of 24 nodes in the input layer; two hidden layers with 2 neurons in the first hidden layer and 7 neurons in the second hidden layer, representing a parsimonious structure that balances model complexity and predictive capability. The corresponding neuron weight matrix is reported to document the learned parameter structure.

A computer screen with a blue and white line

AI-generated content may be incorrect.

# Determine optimal cutoff value

A graph of a bar graph

AI-generated content may be incorrect.A yellow and blue table with numbers

AI-generated content may be incorrect.

Following model selection, an optimal classification cutoff was determined to align model outputs with churn-prevention objectives. Using the decile chart, customers were ranked by predicted probability of churn, and Decile 4 (top 40% highest-risk customers) was selected as the operational threshold. The validation dataset contains 2,813 observations; therefore, the cutoff corresponds to the 1,125th highest predicted probability (approximately 40% of the ranked observations), located at row 1,161 after accounting for the dataset’s starting index. This procedure yielded a cutoff value of 0.29318.

Notably, the optimized cutoff of 0.29318 is very close to the trial threshold of 0.30, providing evidence of model stability with respect to cutoff selection. The neural network was subsequently re-estimated using the refined cutoff, and performance metrics were compared against the results obtained at the 0.30 threshold.

# Model performance

A screenshot of a computer

AI-generated content may be incorrect.

Lowering the cutoff from 0.30 to 0.29 led to a substantial improvement in sensitivity, which increased from 0.402 to 0.680. This indicates a markedly enhanced ability to correctly identify customers who churn, a critical consideration in churn prediction where false negatives are particularly costly. Precision declined modestly from 0.55 to 0.46, reflecting the expected trade-off associated with identifying a larger proportion of true churners at the expense of additional false positives. Importantly, the F1-score, which balances precision and recall, increased from 0.466 to 0.549, confirming that the lower cutoff yields superior overall classification balance.

Although accuracy decreased from 75.1% to 69.8%, this decline is not indicative of inferior model quality. In imbalanced classification settings such as customer churn, accuracy is a misleading metric, as high accuracy can be achieved by predominantly predicting the majority class. Consequently, the cutoff of 0.29 is preferred, as it better aligns model performance with the primary business objective of maximizing churn detection.

# Model 3 – Decision Tree

# Model identification

The third classification model implemented in this project is a Classification Tree (CT) model for binary churn prediction for three reasons:

First, it provides a high degree of interpretability, as model outputs can be expressed as transparent if–then rules that are easily understood by non-technical stakeholders. This interpretability is particularly valuable in churn management, where frontline teams (e.g., customer service and retention units) must be able to act on model insights without requiring statistical expertise.

Second, CT is well suited to capturing nonlinear relationships and interaction effects among predictors, such as the combined influence of contract type, internet service, tenure, and total charges on churn risk. Unlike logistic regression, which assumes additive effects on the log-odds scale, the classification tree naturally models threshold effects and hierarchical decision structures that reflect real-world customer behavior.

Third, the classification tree enables a clear trade-off between predictive performance and model simplicity through pruning and minimum leaf-size constraints. By testing multiple tree complexities and selecting the pruned tree with the strongest validation and test ROC AUC, the model achieves good generalization performance while avoiding overfitting. This balance ensures that the resulting decision rules remain both reliable and operationally deployable.

Overall, the CT model complements the logistic regression and neural network approaches by prioritizing actionability and transparency. While it does not achieve the highest overall AUC, its ability to translate churn risk into intuitive decision rules makes it particularly well suited for supporting retention strategies and real-time decision-making in a business context.

# Variables selection

We used the same predictor set as in the other classification models to enable a fair comparison: demographic variables (gender, Senior Citizen, Partner, Dependents), account variables (tenure, Contract, Payment Method, Paperless Billing, Monthly Charges, Total Charges), and service variables (Phone Service, Multiple Lines, Internet Service, Online Security, Online Backup, Device Protection, Tech Support, Streaming TV, Streaming Movies). All these factors can reasonably influence churn, and including them allows the tree to discover the most important split variables and interactions, especially around contract type, fiber internet, tenure, and total charges, which repeatedly appear as key drivers.

# Variable selection techniques applied

Decision trees do not require a separate variable selection process. The model automatically selects variables during training by choosing splits that maximize the reduction in Gini impurity. Predictors that do not contribute meaningfully simply never appear in the tree.

Instead of applying an external feature selection method, we controlled model complexity through pre-pruning parameters (such as minimum leaf size), which helps prevent overfitting while allowing the tree to determine the most important variables.

# Model output

**Step 1: Initial Tree (Default Parameters)**

First, we fit a decision tree using the default settings in Excel’s Data Science → Analyze → Decision Tree tool. The cutoff was set to 0.50 (customers with predicted churn probability ≥ 0.5 are classified as churners), and the number of terminal nodes was set to Automatic, allowing the algorithm to grow a full tree.

This produced a large initial full tree (Tree height = 7, 99 nodes) that captured many detailed patterns in the data. However, the tree was too complex for practical interpretation and likely overfit some noise in the training sample.

A screenshot of a computer screen

AI-generated content may be incorrect.

Excel then generated an initial best-pruned tree, which simplified the structure while keeping the most important splits.

A diagram of a network

AI-generated content may be incorrect.

In this pruned version, the top splits were:

Contract\_Month-to-month (root node)

InternetService\_Fiber optic

TotalCharges ≈ 1,556.3

The ROC AUC scores for this initial tree were:

A graph with red and blue lines

AI-generated content may be incorrect. A graph of a function

AI-generated content may be incorrect. A graph with a line and a red line

AI-generated content may be incorrect.

* Training AUC: 0.79605
* Validation AUC: 0.7876
* Test AUC: 0.80294

These values show that even with default settings, the tree has reasonably good predictive power and similar performance across training, validation, and test sets, indicating only mild overfitting. In the next steps, we further tune the model by controlling the minimum leaf size.

**Step 2: Testing Different Minimum Leaf Sizes**

After generating the initial tree with default settings, the next step was to control model complexity by adjusting the Minimum Leaf Size parameter.  
Excel allows the tree to grow very deep by default, which can lead to overfitting—the model learns noise instead of meaningful structure.

To prevent this, we tested different minimum leaf sizes:

* 300
* 400
* 500
* 600
* 700

Each value changes how many observations must be present in a leaf node before a split is allowed.

Because decision trees naturally overfit, tuning the minimum leaf size helps:

* Reduce model complexity
* Improve generalization to unseen data
* Stabilize splits (ensuring they reflect real patterns)
* Improve validation AUC
* Make the tree more interpretable

ROC AUC Details for each Leaf node limit tested:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Limit/ Env | 300 | 400 | 500 | 600 | 700 |
| Training | 0.80776 | 0.81463 | 0.81267 | 0.77913 | 0.77913 |
| Validation | 0.79147 | 0.79616 | 0.7979 | 0.77556 | 0.77556 |
| Testing | 0.822 | 0.81677 | 0.81444 | 0.78181 | 0.78181 |

# Model Selection – Full Tree, Pruned Tree

Following the initial model generation, we evaluated the decision tree produced by Excel under two configurations: the fully grown tree using default settings and the automatically pruned version generated by Excel’s internal validation process. This step documents how each model behaved, why pruning was necessary, and why an additional tuning stage was required before selecting the final model.

The full tree, created using the default terminal node settings and a cutoff of 0.50, resulted in a highly complex structure with a tree height of 7 and a total of 99 nodes. While this tree captured many detailed splits, the depth and number of nodes indicated that the model was highly susceptible to overfitting. The presence of several branches built from very small subsets of the data made the full tree unstable and difficult to interpret, limiting its usefulness for practical decision-making. Although the ROC AUC scores for this model were respectable (Training AUC = 0.79605, Validation AUC = 0.7876, Test AUC = 0.80294), the complexity of the full tree made it unsuitable as a final model.

Excel subsequently generated a best-pruned tree, which reduced the tree height to 4 and the total number of nodes to 7. Pruning resulted in the elimination of unnecessary splits that did not significantly improve prediction accuracy. This pruned structure focused on the most influential variables: Contract type, InternetService (Fiber optic versus non-Fiber), Tenure, and TotalCharges. These variables consistently appeared near the upper levels of the tree, indicating their strong predictive power in explaining customer churn. The pruned tree retained AUC values similar to the full tree but offered substantially greater interpretability and stability, both of which are essential for managerial use.

Although the best-pruned tree was significantly improved, Excel’s automatic pruning does not account for model stability across different minimum leaf sizes. Because churn models benefit from reduced variance and stronger generalization, further tuning was required. Therefore, in the next step, the model was refined by systematically adjusting the minimum leaf size (300, 400, 500, 600, and 700) to identify the configuration that provided the best balance between predictive performance and interpretability.

* Limit: 300
* Cut off: 0.5

A screenshot of a graph

AI-generated content may be incorrect.A screen shot of a graph

AI-generated content may be incorrect.A screen shot of a graph

AI-generated content may be incorrect.

Full tree:

A diagram of a company

AI-generated content may be incorrect.

Best Tree:

A diagram of a diagram

AI-generated content may be incorrect.

# Determine optimal cutoff value

A graph of a bar graph

AI-generated content may be incorrect.A table with numbers and symbols

AI-generated content may be incorrect.

The initial version of the decision tree applied the default classification cutoff of 0.50, meaning customers with a predicted churn probability of 50% or higher were classified as churners. While this threshold is commonly used in binary classification, it is not always optimal in churn prediction contexts, where the business objective prioritizes identifying as many at-risk customers as possible. Relying on a 0.50 cutoff can result in a lower sensitivity, leading the model to miss a substantial proportion of true churners. Such false negatives are costly in practice because they prevent timely intervention.

To determine a more appropriate threshold, we examined the decile chart produced by Excel, which groups customers into ten segments based on predicted churn probability. The highest-risk customers appear in Deciles 1 through 4, representing approximately the top 40% of the distribution. The probability boundary separating Decile 4 from Decile 5 was 0.38976, indicating that customers above this threshold exhibit meaningfully elevated churn risk. This value therefore served as a natural cutoff point for classification.

A screenshot of a computer

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.

The cutoff was subsequently lowered from 0.50 to 0.38976. This adjustment increased the model’s ability to capture true churners, improving sensitivity while maintaining an acceptable level of precision. The revised cutoff thus better aligns the model with the business objective of minimizing customer attrition by ensuring that a larger proportion of at-risk customers are flagged for proactive retention strategies.

**Best Tree: Cutoff: 0.38976**

A diagram of a network

AI-generated content may be incorrect.

# Decision Rules from Final Tree (Minimum Leaf Size = 300)

The final decision tree, constructed using a minimum leaf size of 300 and the adjusted cutoff, produced a concise and interpretable set of decision rules. These rules represent the key pathways that the model uses to classify customers as likely to churn or not churn based on their contract type, service characteristics, tenure, and accumulated charges.

A consistent pattern across the tree is the dominant role of contract type. Customers who do not subscribe to month-to-month contracts were consistently placed into non-churn terminal nodes, indicating that longer-term contractual commitments strongly reduce churn likelihood. Among customers with month-to-month contracts, the model further distinguishes risk levels based on Internet service type. Individuals without Fiber optic service and with very short tenure—specifically less than 4.5 months—were classified as high churn risk, reflecting instability among newly enrolled customers. In contrast, customers in the same service category with tenure exceeding 4.5 months were directed to non-churn outcomes, suggesting that retention improves significantly with service duration.

For customers with month-to-month contracts who use Fiber optic service, the model relies on TotalCharges to determine churn probability. Those with relatively low cumulative spending (≤ $1,556) were classified as likely to churn, a pattern consistent with early-stage or low-value customers who may be more price-sensitive or less committed to the service. Conversely, Fiber customers with total charges above this threshold were predicted not to churn, indicating that higher-value customers exhibit greater stability even within flexible contract arrangements.

These rules collectively demonstrate that churn risk is influenced by a combination of contract flexibility, service type, customer tenure, and total historical spending. The resulting structure provides clear guidance for identifying vulnerable customer segments and informs targeted retention strategies.

# Compare models

A graph of a tree

AI-generated content may be incorrect.

All three models—Logistic Regression, Classification Tree, and Neural Network—were first built and evaluated using their default classification cutoff of 0.50 and then re-estimated with business-oriented cutoffs chosen from the decile charts. The final comparison below is based on these tuned versions:

* Logistic Regression: cutoff = 0.29
* Decision Tree: cutoff = 0.38976
* Neural Network: cutoff = 0.29

**Logistic Regression**

Logistic regression achieves the highest discrimination among the three models, with a validation AUC of 0.8468, and nearly identical training AUC, indicating excellent generalization and no overfitting. After lowering the cutoff from 0.50 to 0.29 (top 40% by predicted probability), recall increases from roughly 0.56 to 0.79, and the F1-score improves from 0.606 to 0.637, while accuracy declines only modestly from 80.3% to 75.7%. Thus the tuned model simultaneously preserves strong overall accuracy and becomes much more effective at identifying churners. In addition, logistic regression provides clear coefficient estimates and odds ratios that explain how contract type, payment method, internet service, and support features drive churn risk.

**Classification Tree**

The classification tree was tuned by varying the minimum leaf size (300–700) and then applying a decile-based cutoff. The final tree uses a leaf size of 300 and a cutoff of 0.38976, corresponding to the top 40% of customers ranked by churn probability. This model attains a training AUC of 0.8078, validation AUC of 0.7915, and test AUC of 0.8220, showing strong and balanced performance across all datasets. Lowering the cutoff increases recall and improves the F1-score, allowing the tree to capture more churners while maintaining reasonable precision. Its main advantage is interpretability: the final tree reduces a small set of intuitive IF–THEN rules involving contract type, fiber usage, tenure, and total charges.

**Neural Network**

The neural network (Net36) required even more aggressive cutoff tuning. With the optimized cutoff of 0.29, it achieves an AUC of about 0.73, accuracy of 69.8%, precision of 0.46, recall of 0.681, and F1-score of 0.549. Compared to its default threshold, the tuned network substantially improves recall and F1-score, meaning it identifies a larger share of true churners, but this comes at the cost of lower accuracy and precision. Moreover, the neural network behaves as a “black box” and does not provide clear interpretable rules, which limits its usefulness for managerial explanation.

**Overall Comparison and Final Choice**

When all models are evaluated after cutoff adjustment, three patterns emerge:

* Predictive power: Logistic Regression retains the highest AUC (~0.85), followed by the Decision Tree (~0.82), and then the Neural Network (~0.73).
* Recall–precision balance: With tuned cutoffs, logistic regression and the neural network both achieve high recall, but logistic regression maintains better overall balance (higher F1-score and accuracy), while the neural network sacrifices too much accuracy.
* Interpretability and actionability: Logistic Regression and the Decision Tree both provide transparent structures—coefficients for the former and IF–THEN rules for the latter—whereas the neural network is difficult to interpret and harder to operationalize.

Given these results, Logistic Regression is selected as the primary model for decision-making because it combines the strongest AUC, high recall under the optimized cutoff, and rich interpretability for business insight. The Decision Tree is recommended as a complementary model for operational use, where simple rules are needed for frontline teams. The Neural Network can be used as a secondary tool in scenarios where maximizing recall is more important than accuracy or interpretability, such as broad early-warning campaigns.

# Hypothetical example

To demonstrate how the three model makes predictions, below is a hypothetical example with a potential customer profile including:

|  |  |
| --- | --- |
| Variable | Value |
| **Tenure** | 5 months |
| **Senior Citizen** | No |
| **Phone Service** | Yes |
| **Internet Service** | Fiber Optic |
| **Online Security** | No |
| **Tech Support** | No |
| **Streaming Movies** | Yes |
| **Contract** | Month-to-Month |
| **Paperless Billing** | Yes |
| **Payment Method** | Electronic Check |

To illustrate how the predictive models can be applied in practice, we consider a hypothetical customer profile characterized by short tenure, a month-to-month contract, fiber optic internet service, and limited add-on services. Specifically, this customer has a tenure of 5 months, total charges of $250, subscribes to fiber optic internet without online security or technical support, and uses electronic check as the payment method

* Using the logistic regression model, the predicted probability of churn for this customer is 0.83, which exceeds the optimized cutoff value of 0.29. Therefore, the model classifies this customer as likely to churn.
* The decision tree model arrives at the same conclusion through a sequence of interpretable decision rules. Starting from the root node, the customer follows the branch corresponding to a month-to-month contract, then the fiber optic internet branch, and finally the low total charges (≤ $1,556) branch, which leads to a terminal node predicting churn.
* Similarly, the neural network model produces a high output probability for the churn class, again exceeding the cutoff threshold. Although the neural network lacks the explicit interpretability of the decision tree, its prediction is directionally consistent with both the logistic regression and tree-based models.

Overall, this hypothetical example demonstrates strong agreement across all three modeling approaches. Customers with short tenure, month-to-month contracts, fiber optic service, and limited service bundling consistently exhibit high churn risk, reinforcing the robustness of the modeling results.

# Recommendations

# Explain how the results address your business questions

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

The modeling results indicate that the highest churn risk is concentrated among month-to-month customers using fiber optic internet with low tenure and low total charges. These customers are at an early stage in their customer lifecycle, meaning their future revenue potential is substantial if churn can be prevented. By targeting the top 40% of customers ranked by predicted churn probability, the company can focus retention resources on those customers where the expected return on investment is greatest, calculated as the product of churn probability, monthly revenue, and expected remaining tenure.

***Which customer segments (contract type, tenure, billing method) contribute most to total revenue, and how could retention offers improve CLV?***

The analysis also highlights clear differences in customer lifetime value (CLV) across segments. Customers with long tenure, high total charges, and one- or two-year contracts exhibit the lowest churn risk and the highest CLV. In contrast, month-to-month customers with low tenure tend to churn early, resulting in significantly lower lifetime value. These findings suggest that encouraging customers to transition from month-to-month contracts to longer-term agreements—particularly during the early months of service—can substantially improve retention and long-term profitability.

***Which service combinations most strongly affect churn risk?***

Service bundling emerges as a key protective factor against churn. Across all models, customers who subscribe to online security and technical support services show consistently lower churn probabilities. Conversely, customers who use fiber optic internet without complementary support services face significantly higher churn risk. This pattern suggests that bundling fiber optic plans with security and support add-ons can improve customer satisfaction and reduce churn, while also creating opportunities for upselling higher-value service packages.

# Business Strategy

Based on the empirical results from the logistic regression, neural network, and classification tree models, several consistent and actionable insights emerge regarding customer churn behavior.

**Key Insights for the Organization**

* Customer churn is highly predictable using existing customer, contract, billing, and service data.
* Early tenure is the most critical churn window—customers are far more likely to leave in their first few months.
* Contract structure is the strongest driver of churn. Month-to-month customers are significantly more likely to churn than customers on longer-term contracts.
* Service bundles matter. Customers with support-related services (Online Security and Tech Support) are much more likely to stay, while customers with only entertainment add-ons show higher churn risk.
* Payment behavior signals risk. Customers using electronic check payments churn more frequently than those using automatic or traditional payment methods.
* High-value customers tend to be more stable, meaning retention strategies for them should focus on loyalty reinforcement rather than aggressive discounting.

**Recommended Actions**

***Prioritize Contract Conversion for Month-to-Month Customers***

All three models consistently indicate that customers on month-to-month contracts exhibit substantially higher churn risk compared to those on one-year or two-year contracts. Logistic regression shows significantly elevated odds of churn for month-to-month customers, while both the neural network and decision tree isolate this contract type as a primary split in identifying high-risk customers.

Strategic implication:

The company should prioritize converting month-to-month customers to longer-term contracts through targeted incentives such as discounted pricing, loyalty rewards, or bundled service upgrades. These interventions are likely to generate high returns on investment, particularly when directed toward customers with high predicted churn probabilities and moderate-to-high monthly charges.

***Improve Retention Efforts for Fiber-Optic Internet Customers***

Customers subscribing to fiber-optic internet service, especially those with short tenure and lower total charges, consistently demonstrate elevated churn risk. This pattern appears across all models and suggests that fiber service may be associated with unmet expectations, service quality concerns, or pricing sensitivity during the early stages of adoption.

Strategic implication:

Rather than treating fiber-optic customers as inherently high-value, the firm should recognize them as a high-risk, high-impact segment. Targeted onboarding programs, proactive service quality monitoring, and early-stage customer support for new fiber users may substantially reduce churn. Addressing service experience issues early can prevent revenue loss from customers who otherwise have strong long-term value potential.

***Focus Retention Resources on Early-Tenure Customers***

Tenure emerges as one of the strongest protective factors against churn. Customers with low tenure are far more likely to churn, while long-tenure customers demonstrate significantly lower churn probabilities and higher accumulated revenue.

Strategic implication:

Retention efforts should be concentrated during the early customer lifecycle, particularly within the first few months of service. Structured onboarding initiatives—such as welcome communications, usage guidance, early troubleshooting support, and follow-up outreach—can meaningfully improve customer engagement and reduce early attrition, thereby increasing customer lifetime value.

***Promote “Sticky” Service Bundles That Reduce Churn***

Support-related services, particularly OnlineSecurity and TechSupport, consistently appear as protective factors across models, reducing churn probability when bundled with core internet services. In contrast, entertainment-focused add-ons (e.g., StreamingMovies) without support services do not provide the same retention benefits.

Strategic implication:

The company should actively promote bundles that combine internet service with support and security features. Offering free trials or discounted introductory periods for these services—especially to high-risk customers—can increase perceived value and switching costs, thereby strengthening customer retention. Bundle design should emphasize reliability, support, and peace of mind rather than entertainment alone.

***Address High-Risk Payment and Billing Behaviors***

Payment method and billing preferences also play a meaningful role in churn risk. Customers using electronic check payments and paperless billing exhibit higher churn probabilities, potentially reflecting weaker customer commitment or payment friction.

Strategic implication:

The firm should incentivize customers to transition toward more stable and automated payment methods, such as credit cards or bank transfers. Incentives may include small discounts, loyalty points, or bundled benefits. Reducing billing-related friction can improve retention while also lowering operational risk.

***Operationalize Decision Tree Rules for Frontline Teams***

While logistic regression provides the strongest overall predictive performance and neural networks improve churn detection sensitivity, the classification tree offers the clearest operational value through simple, interpretable rules.

Strategic implication:

Decision tree rules can be embedded into customer service workflows, dashboards, or call-center scripts to enable real-time churn risk identification. Frontline teams can use these rules to quickly classify customers and trigger appropriate retention actions without requiring advanced analytical expertise.

# Project Summary

# Lessons learned from the project

Several key lessons emerged from the modeling process and results.

First, model selection should be guided by business objectives, not accuracy alone. While logistic regression achieved the highest ROC AUC, the neural network proved valuable when recall was prioritized, and the classification tree excelled in interpretability. This highlights that no single model is universally optimal; instead, model choice depends on whether the primary goal is explanation, detection, or operational deployment.

Second, cutoff optimization is critical in imbalanced classification problems such as churn prediction. Using the default cutoff of 0.50 led to poor detection of churners, particularly for the neural network. Decile-based cutoff selection substantially improved recall and F1-score, demonstrating that probability thresholds must be aligned with business costs and objectives rather than statistical convention.

Third, the project underscored the importance of early customer lifecycle management. Tenure consistently emerged as a dominant protective factor across all models, indicating that the highest return on retention efforts occurs early in the customer relationship. This insight reinforces the strategic value of onboarding and early engagement programs.

Fourth, the analysis showed that actionable variables matter more than purely descriptive ones. Contract type, billing method, and service bundles are controllable levers for the business and were among the strongest churn predictors. This confirms that machine learning models are most valuable when they focus on features that can directly inform managerial action.

Finally, the project demonstrated the benefit of a multi-model approach. Using logistic regression for strategic insight, decision trees for operational rules, and neural networks for enhanced churn detection provides a more comprehensive solution than relying on a single modeling technique.

# Dataset limitations and suggestions for future data extensions

While the dataset provided a strong foundation for churn analysis, it also has several limitations that restrict the depth of insights and accuracy of predictions.

**Dataset Limitations**

* The data represents a single snapshot in time, rather than customer behavior over multiple periods, limiting the ability to capture churn trends dynamically.
* There is no direct measure of customer satisfaction, such as complaint history, service quality issues, or customer support interactions.
* The dataset lacks pricing and promotion history, making it difficult to evaluate how past discounts or offers influenced retention.
* External factors such as competitor pricing, regional market conditions, or service outages are not captured.
* Revenue impact is inferred indirectly; explicit profitability or cost-to-serve data is missing.

**Suggestions for Future Data Extensions**

To improve future churn modeling and business impact, the organization should consider collecting:

* Time-series customer data (monthly snapshots) to model churn as a dynamic process.
* Customer support and complaint logs, including call frequency, issue types, and resolution times.
* Promotion and offer history, enabling direct ROI measurement of retention strategies.
* Usage intensity metrics, such as data consumption or service utilization.
* Customer feedback and satisfaction scores (e.g., surveys or Net Promoter Score).
* Profitability metrics, including margins and service costs, to move from churn prediction to profit-optimized retention strategies.

Expanding the dataset in these directions would allow the company to move beyond predicting churn toward prescriptive analytics, where the model recommends the best action for each customer.

# Task Allocation

|  |  |
| --- | --- |
| **Member** | **Tasks** |
| ***Haley Hoang*** | * Handling missing data * Random sampling * Identify numerical variables vs. categorical variables * Based on your questions, which variable is the dependent variable? Which variables could be used as independent variables? * Run the Logistic Regression & Neural Network models following:  1. Identify the models used and provide a rationale for each selection. 2. Specify the variables included and justify their choice. Describe any variable selection techniques applied. 3. Present the model output, including equations (if applicable) and coefficient interpretations 4. Provide a summary report for training, validation, and test data (if applicable), along with lift charts. Assess the model’s performance. 5. For classification models, determine an appropriate cutoff value based on your results. Run the model with alternative cutoff values and compare performance. 6. Explain how the results address your business questions. 7. Offer a hypothetical example of a new data record and demonstrate prediction or classification.  * Final report structure, format & edit * Final slide design & format |
| ***Noynicaa Santani*** | * Correlation table of numerical variables, comment on the correlations * Histogram of numerical variables, comment on the distributions * Scatterplot of the Dependent variable vs the independent variables, comment on the relationships * Define the business questions (note any changes to the questions from the first two milestones). * Identify the model used and provide a rationale for your selection. * Specify the variables included and justify their choice. * Describe any variable selection techniques applied. * Business Recommendation * Business Questions Analysis * Data Limitation & Suggestion for future * Slide design |
| ***Sakshi Agarwal*** | * Boxplot of the Dependent variable vs the independent variables, comment on the relationships * From the previous plots, detect outliers and find out whether they are errors or extreme values * In general, are there any potential issues with the dataset? Like, if you could get data from the company directly, how would you extend the current dataset? * Ran the Decision Tree model and addressed the following question for each model:  1. Identify the models used and provide a rationale for each selection. 2. Specify the variables included and justify their choice. 3. Describe any variable selection techniques applied. 4. Present the model output, including equations (if applicable) and coefficient interpretations 5. Provide a summary report for training, validation, and test data (if applicable), along with lift charts. Assess the model’s performance. 6. For classification models, determine an appropriate cutoff value based on your results. Run the model with alternative cutoff values and compare performance. 7. Explain how the results address your business questions. 8. Offer a hypothetical example of a new data record and demonstrate prediction or classification.  * Model Comparison * Report format * Slide design |