Telecom Customer Churn Prediction Analysis

Kenny Trinh Sabin Pun Sarp Koc

2025-01-19

Contents

1 A Multi-Model Machine Learning Approach

1.1 Executive Summary

This analysis employs multiple machine learning approaches to understand customer behavior and predict churn in the telecommunications sector. By analyzing a comprehensive dataset of customer information, we provide actionable insights for improving customer retention and service optimization.

1.2 Business Context: The Story of "TeleConnect"

Meet **TeleConnect**, a mid-sized telecommunications company striving to maintain its foothold in a fiercely competitive market. Over the past year, the company has faced growing challenges: increased customer churn, stagnant revenue growth, and difficulty in predicting which services drive customer satisfaction and retention.

1.2.1 The Challenge

TeleConnect's problems are multifaceted:

- **High Churn Rates**: Nearly 20% of their customer base leaves each year, especially those on month-to-month contracts.
- **Price Sensitivity**: Customers complain about high monthly charges, but TeleConnect lacks clarity on whether price adjustments would help.
- **Service Bundling**: While TeleConnect offers multiple services—Internet, Streaming, Security—adoption patterns remain unclear. Are these services increasing customer value, or are they simply an added cost?

TeleConnect's leadership knows the stakes: acquiring a new customer costs significantly more than retaining an existing one. Yet, without data-driven insights, their current retention strategies feel like guesswork.

Why Machine Learning?

To address these challenges, TeleConnect has embraced a multi-model machine learning approach, leveraging advanced analytics to transform their operations. By applying complementary models tailored to different aspects of customer behavior, TeleConnect seeks to:

1. **Predict Monthly Charges**: Understand the drivers of customer spending to design smarter pricing strategies and optimize service bundles.
2. **Analyze Customer Tenure**: Uncover patterns in contract types, payment methods, and service adoption that influence the length of customer relationships.
3. **Forecast Churn**: Identify at-risk customers proactively and implement targeted retention efforts.
4. **Segment Customers by Risk**: Group customers into risk categories to prioritize retention strategies effectively.

The Opportunity

By integrating machine learning models into its decision-making processes, TeleConnect can:

- **Reduce churn** by predicting at-risk customers and proactively addressing their concerns.
- **Maximize revenue** through optimized pricing and smarter service bundling.
- **Improve retention** by understanding what drives tenure and customer satisfaction.

1.3 Dataset Overview

The dataset encompasses comprehensive customer information including:

To categorize the variables in your dataset into **continuous**, **categorical**, **count**, and **binomial**, let's analyze them step by step:

1. Continuous Variables:

- `tenure` (Number of months the customer has been with the company)
- `MonthlyCharges` (Monthly charge in dollars)
- `TotalCharges` (Cumulative charges in dollars)

2. Categorical Variables:

- `MultipleLines` (No phone service/Yes/No)
- `InternetService` (DSL/Fiber optic/No)
- `OnlineSecurity` (No internet service/Yes/No)
- `OnlineBackup` (No internet service/Yes/No)
- `DeviceProtection` (No internet service/Yes/No)
- `TechSupport` (No internet service/Yes/No)
- `StreamingTV` (No internet service/Yes/No)
- `StreamingMovies` (No internet service/Yes/No)
- `Contract` (Month-to-month/One year/Two year)
- `PaymentMethod` (Electronic check/Mailed check/Bank transfer/Credit card)

3. Binomial Variables:

- `gender` (Male/Female)
- `SeniorCitizen` (0/1)
- `Partner` (Yes/No)
- `PhoneService` (Yes/No)
- `PaperlessBilling` (Yes/No)
- `Churn` (Yes/No - Target variable)

1.3.1 Key Variables and Their Business Significance

1. **Critical Predictors**
   - Tenure: Indicates customer loyalty and relationship duration
   - Contract Type: Reflects commitment level
   - Monthly Charges: Represents service value and potential price sensitivity
   - Internet Service: Core service offering affecting overall satisfaction
2. **Service Usage Indicators**
   - Multiple service subscriptions suggest higher customer engagement
   - Security and support services indicate value-added service adoption
   - Streaming services usage reflects modern consumption patterns
3. **Financial Metrics**
   - Average monthly charges: $64.76
   - Payment methods diversity indicates billing flexibility
   - Relationship between charges and service subscriptions

Remark: Some results from the R-code chunks are hidden to improve readability. Those hidden results are shown in the Appendix - Code Appendix section.

2 Exploratory Data Analysis

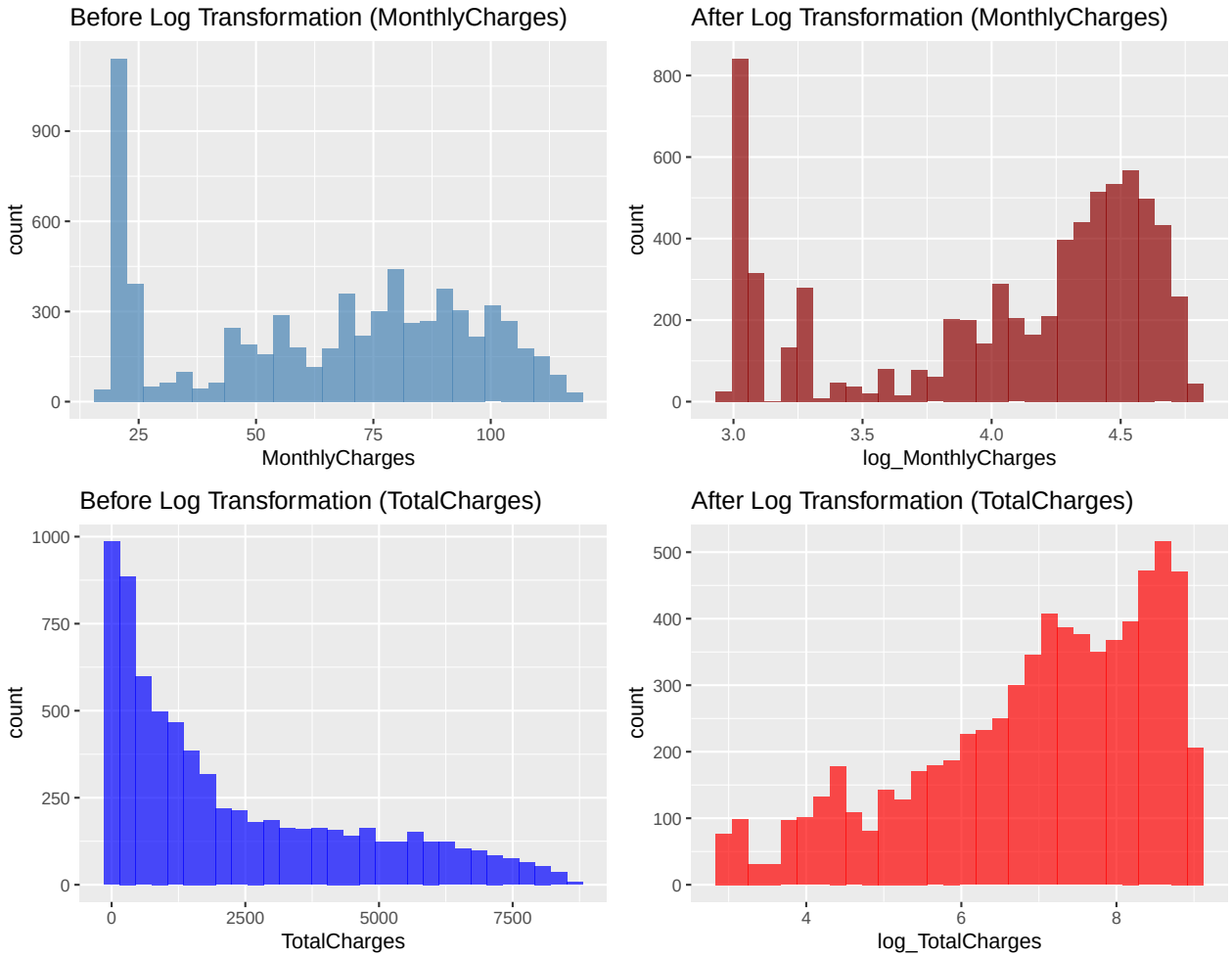
The exploratory data analysis focuses on understanding customer churn, service adoption, and tenure patterns. This section includes visual and statistical analysis to identify trends and factors influencing churn and customer behavior.

Data Overview:

The dataset contains 7,043 customers with key features like:

- Demographics (Tenure, Services Count)
- Contract & Payment details (Contract Type, MonthlyCharges)
- Churn Status (Target Variable)

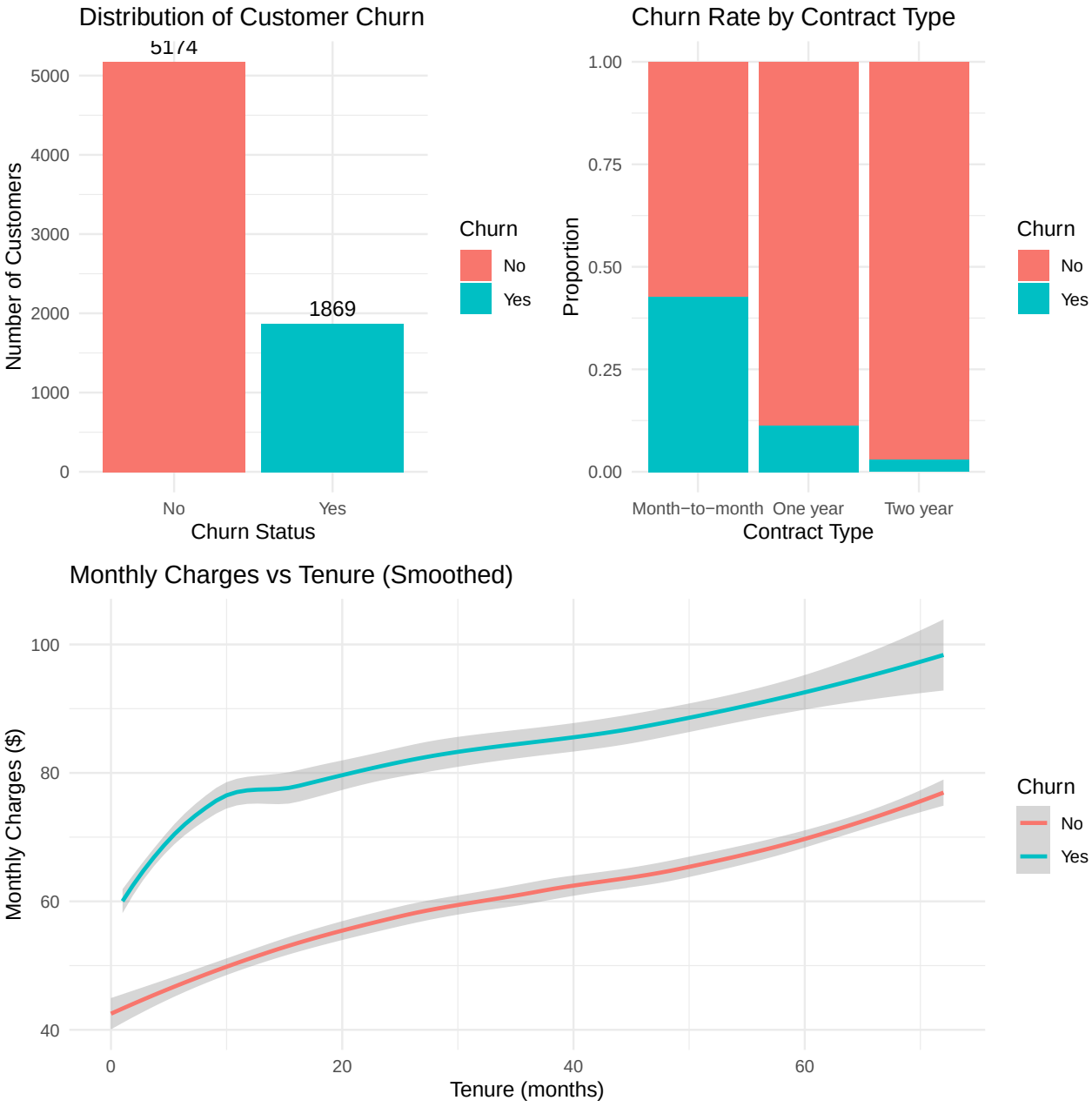
Log-Transformation of Charges (MonthlyCharges & TotalCharges)



- TotalCharges and MonthlyCharges were right-skewed, requiring log-transformation.
- After transformation, the distributions became more normal, improving model assumptions

2.1 Key Findings from EDA

```
## `geom_smooth()` using formula = 'y ~ x'
```



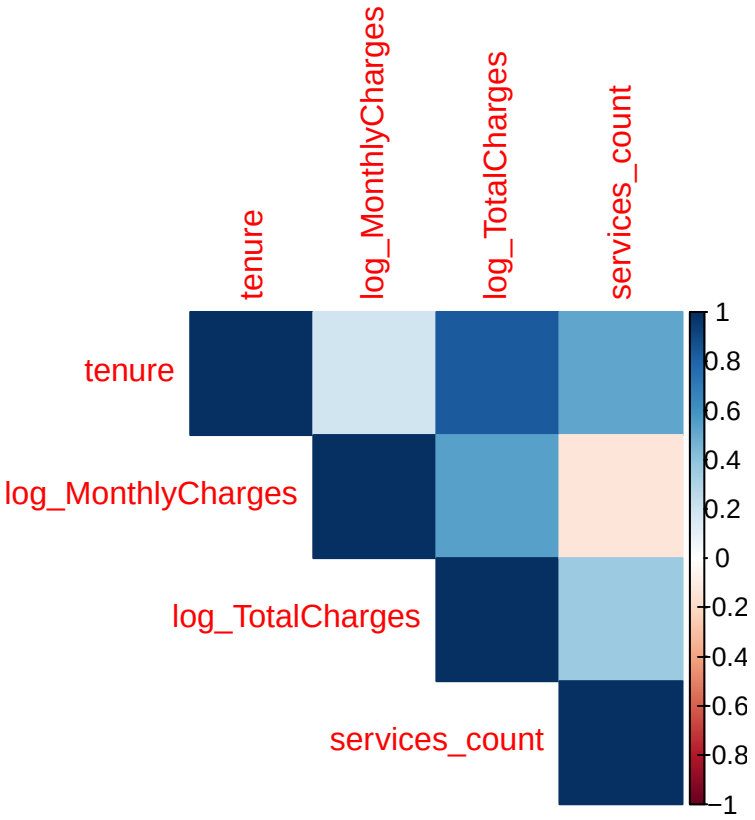
1. Churn Distribution

- 26.5% of customers churned (1,869 out of 7,043).
- Class imbalance exists (need to consider resampling for modeling).

2. Contract Type & Churn Relationship

- Month-to-month contracts have the highest churn rate (42.7%).
- Two-year contracts have the lowest churn rate (11.2%). 3.Relationship Between Monthly Charges & Tenure
- Customers with lower tenure & high MonthlyCharges churn more frequently.
- The relationship is non-linear, justifying using Generalized Additive Models(GAMs).

Correlation Analysis with numeric variables



Correlation Analysis

- TotalCharges and Tenure have a very high correlation ($r \approx 0.98$), meaning one could be removed in some models to avoid multicollinearity.

Based on these findings, we proceed to the methodology section, where we implement various models to predict churn and evaluate their effectiveness.

2.2 Model Selection

Our analysis employs six complementary machine learning approaches, each chosen for specific analytical capabilities:

Our analysis employs a comprehensive suite of models, each chosen for specific analytical capabilities:

In our analysis, we employed the following machine learning approaches:

1. **Linear Model (LM)**: To predict monthly charges and understand service impacts
2. **GLM-Poisson**: To analyze customer tenure patterns and retention factors
3. **GLM-Binomial**: To predict customer churn probability
4. **Generalized Additive Model (GAM)**: To capture non-linear relationships in customer behavior
5. **Neural Network (NN)**: To recognize complex patterns in customer churn
6. **Support Vector Machine (SVM)**: To segment customers based on churn risk and grouping them.

3 Linear Model (LM)

Sabin Pun took the lead on the linear model.

To address **TeleConnect's price sensitivity challenge**, we began by building a **Linear Model (LM)** to understand the key drivers of MonthlyCharges, we first build a Linear Model (LM) as a baseline approach. This helps TeleConnect analyze the impact of various customer attributes on pricing and service bundling

The response variable for this linear model is:

- **MonthlyCharges**: The amount customers pay per month(continuous)

The predictors used are:

1. **Tenure**: Number of months the customer has been with the company.
2. **Contract:** Type of contract(Month to Month, one year, two year)
3. **InternetService**: Type of internet service (DSL, Fiber optic, No internet service).
4. **StreamingTV**: Whether customers have streaming TV services (Yes/No).
5. **PhoneService**: Whether customers have a phone service (Yes/No).
6. **MultipleLines**: Whether customers have multiple phone lines (Yes/No).

3.1 Linear Model to predict monthly charges

Coefficient Interpretation:

Table 1: Coefficient interpretation for the linear model

Predictor	Effect on Log(Monthly Charges)
Tenure	`+0.00064` per month
Internet Service (Fiber Optic)	`+0.347` (**+41.5%**)
Internet Service (No Service)	`-0.863` (**-57.8%**)
Contract (Two Years)	`+0.00882` (**+9.2%**)
Additional Services	**Increase charges by 6% - 55%**

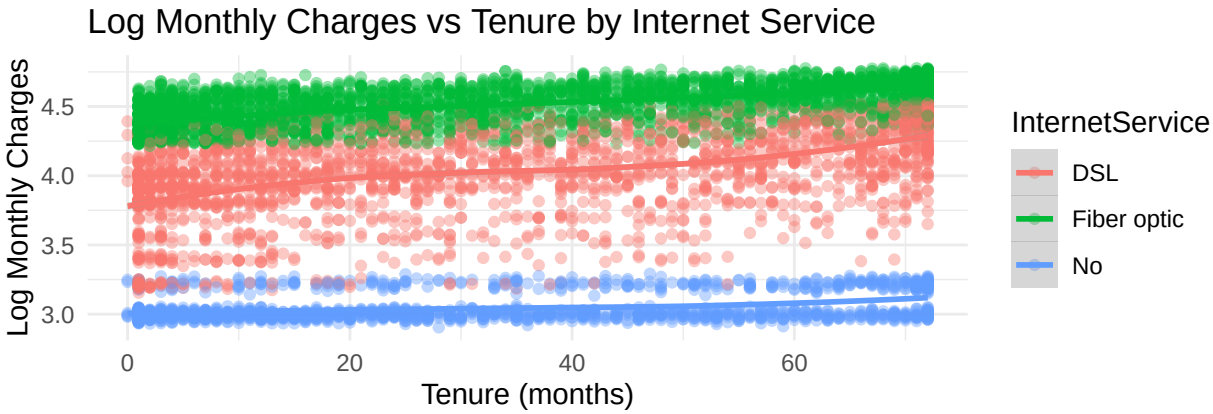
Interpretation:

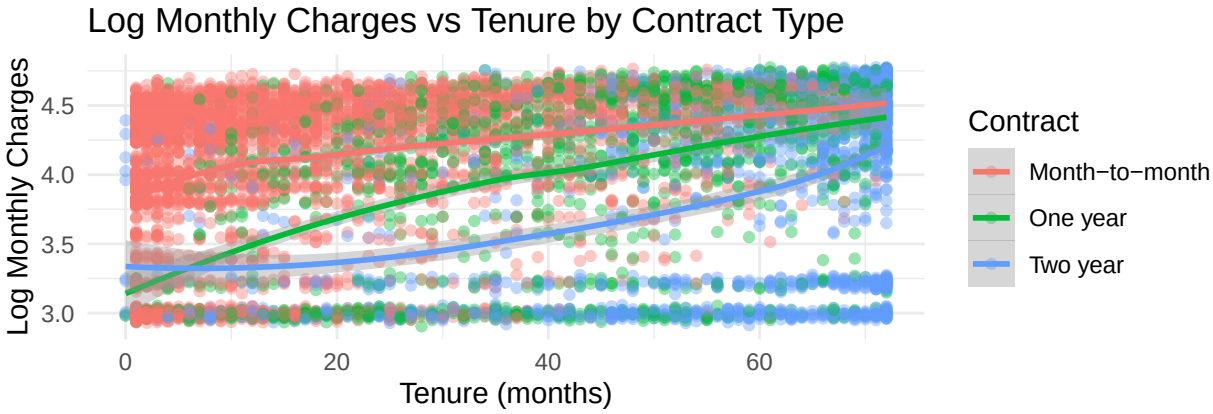
- **Tenure**: Longer tenure slightly increases monthly charges.
- **Internet Service**: Fiber optic users pay 41.5% more than DSL users, while customers with no internet service pay 57.8% less.
- **Contract Type**: Two-year contracts increase monthly charges by 9.2%.
- **Additional Services (includes Streaming, Security, Backup, Tech Support, Device Protection and Phone Service)**: Customers with multiple services pay 6% to 55% more, with Phone Service having the most significant impact.

3.2 Visualizing the main effects

Monthly Charges against Tenure:/ Monthly Charges vs Tenure:

```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```





3.3 Limitations of the Linear Model

While the linear model provides useful insights, it assumes a strictly linear relationship between features and Monthly Charges. However, as observed, some results—such as contract type pricing—contradict expectations.The visualizations indicate non-linear relationships between Tenure and Log Monthly Charges. This suggests that contract pricing follows a non-linear trend, which cannot be captured effectively by a simple linear model.

- Contract types & internet services exhibit distinct trends, suggesting interactions that LM cannot fully capture.
- LOESS smoothers in the plots show curved trends, implying that a Generalized Additive Model (GAM) or non-linear methods (SVM, Neural Networks) might be better suited.

Therefore, while LM helps interpret key relationships, we move forward with GLMs, GAMs, and advanced models to improve predictive accuracy.

4 GLM-Poisson Model

Sabin Pun took the lead on the GLM-Poisson model.

To address **TeleConnect's challenge of understanding customer retention**, we utilized a **GLM-Poisson model** to analyze the factors influencing **tenure**—the number of months a customer remains subscribed—is a key indicator of retention. A longer tenure generally means higher customer lifetime value (CLV), while shorter tenures indicate early churn.

To analyze customer tenure, we use a GLM-Poisson regression model, which is well-suited for count data like tenure (measured in months).

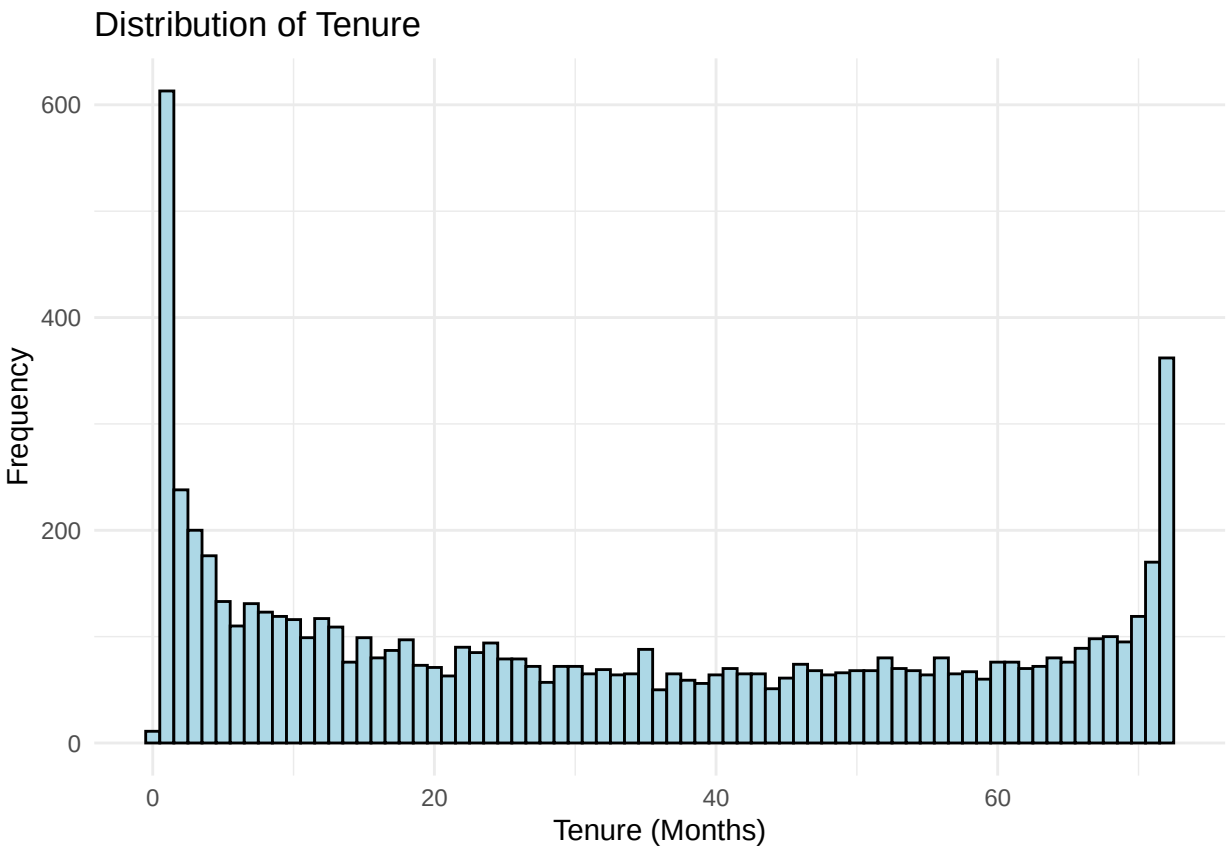
The response variable for this model is:

- **Tenure** (count of months as a customer).

Include relevant predictors, such as:

- `Contract`: Longer contracts may correlate with higher tenure.
- `InternetService`: Different internet services might drive customer loyalty.
- `PaymentMethod`: Certain payment methods might affect tenure (e.g., automatic payments may reduce churn).
- `StreamingTV` and `StreamingMovies`: Value-added services could contribute to customer retention.
- `SeniorCitizen`, `Partner`, `Dependents`: Demographics that might influence tenure.

Exploring the Data:

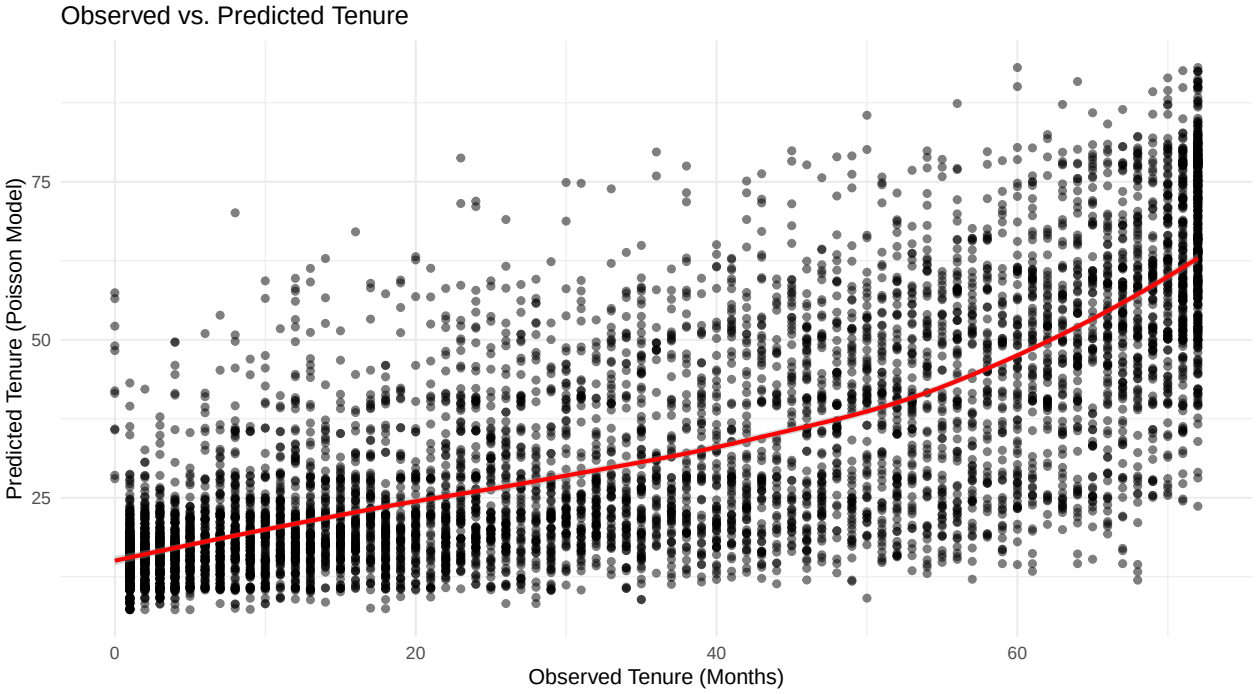


- There's a high frequency of customers with **tenure close to 0 months**.
- There's another peak at **tenure near 72 months**, likely customers who have been with the company long-term.

4.1 Visualization

Observed vs Predicted Tenure

```
## `geom_smooth()` using formula = 'y ~ x'
```



Observed vs. Predicted Tenure (Poisson Model)

- The model underestimates shorter tenures and overestimates longer tenures, indicating potential overdispersion or non-linearity.
- Suggests that Poisson may not be the best fit, reinforcing the need for Quasi-Poisson or Negative Binomial.

4.2 Overdispersion Check

To ensure that the **Poisson model** is appropriate for modeling **customer tenure**, we checked for **overdispersion**. Overdispersion occurs when the variance of the response variable is significantly greater than its mean, which can lead to underestimated standard errors and misleading statistical inferences.

Dispersion Parameter Calculation The dispersion parameter is calculated as:

$$\text{Dispersion} = \frac{\text{Residual Deviance}}{\text{Degrees of Freedom}}$$

For our **Poisson model**, we obtained:

- **Residual Deviance**: `74,160`
- **Degrees of Freedom**: `7,030`
- **Dispersion Parameter**: `10.55`

Interpretation:

- Since the dispersion parameter **(10.55)** is **significantly greater than 1**, this indicates **strong overdispersion**.
- This suggests that the **Poisson model underestimates variability**, making it **not the best fit** for the data.

Solution: Quasi-Poisson Model To address overdispersion, we proceed with a **Quasi-Poisson model**, which adjusts the standard errors to provide more reliable estimates.

4.3 Quasi-Poisson Model

Since the **Poisson model showed significant overdispersion**, we use a **Quasi-Poisson model** to correct for underestimated standard errors. The **Quasi-Poisson model** allows for variability greater than the mean, providing more reliable coefficient estimates.

4.4 Quasi-Poisson Model Interpretation

Table 2: **Coefficient for the Quasi-Poisson model**.

Predictor	**Effect on Log(Tenure)**
Internet Service (Fiber Optic)	`-0.124` (↓**11.6% tenure**)
Internet Service (No Service)	`+0.509` (↑**66.4% tenure**)
Contract (One Year)	`+0.731` (↑**107.7% tenure**)
Contract (Two Year)	`+0.970` (↑**164.4% tenure**)
Payment Method (Credit Card - Automatic)	
Payment Method (Electronic Check)	`-0.157` (↓**14.5% tenure**)
Payment Method (Mailed Check)	`-0.362` (↓**30.4% tenure**)
Senior Citizen	`+0.118` (↑**12.5% tenure**)
Partner (Yes)	`+0.236` (↑**26.6% tenure**)
Log(Monthly Charges)	`+0.615` (↑**84.9% tenure per unit increase**)

Interpretation:

- Customers with **Fiber optic internet** have **11.6% shorter tenure** compared to DSL users. - Customers with **no internet service** have **66.4% longer tenure** than DSL users.
- Customers with a **one-year contract stay 107.7% longer** than those on **month-to-month contracts**
- Customers with a **two-year contract stay 164.4% longer**, highlighting **strong retention** for long-term contracts.
- Customers paying via **electronic check** have **14.5% shorter tenure**, suggesting potential churn risks.
- Customers using **mailed checks** have **30.4% shorter tenure**, indicating **high churn risk**.
- Senior citizens tend to stay **12.5% longer** than younger customers.
- Customers **with partners** stay **26.6% longer**, indicating stability in tenure.
- **Higher monthly charges** are associated with **longer tenure**, possibly due to bundled services.

Key Takeaways:

- **Contract type remains the strongest predictor** of customer retention.
- **Higher-paying customers** tend to stay longer.
- **Payment method influences tenure**: Customers using **mailed checks churn much faster**.
- **Internet service impacts tenure in unexpected ways**: Customers without internet **stay longer**.

5 GLM-Binomial (Logistic Regression) & GAM

Sarp Koc took the lead on the GLM-Binomial and GAM models.

5.1 Context and Objective

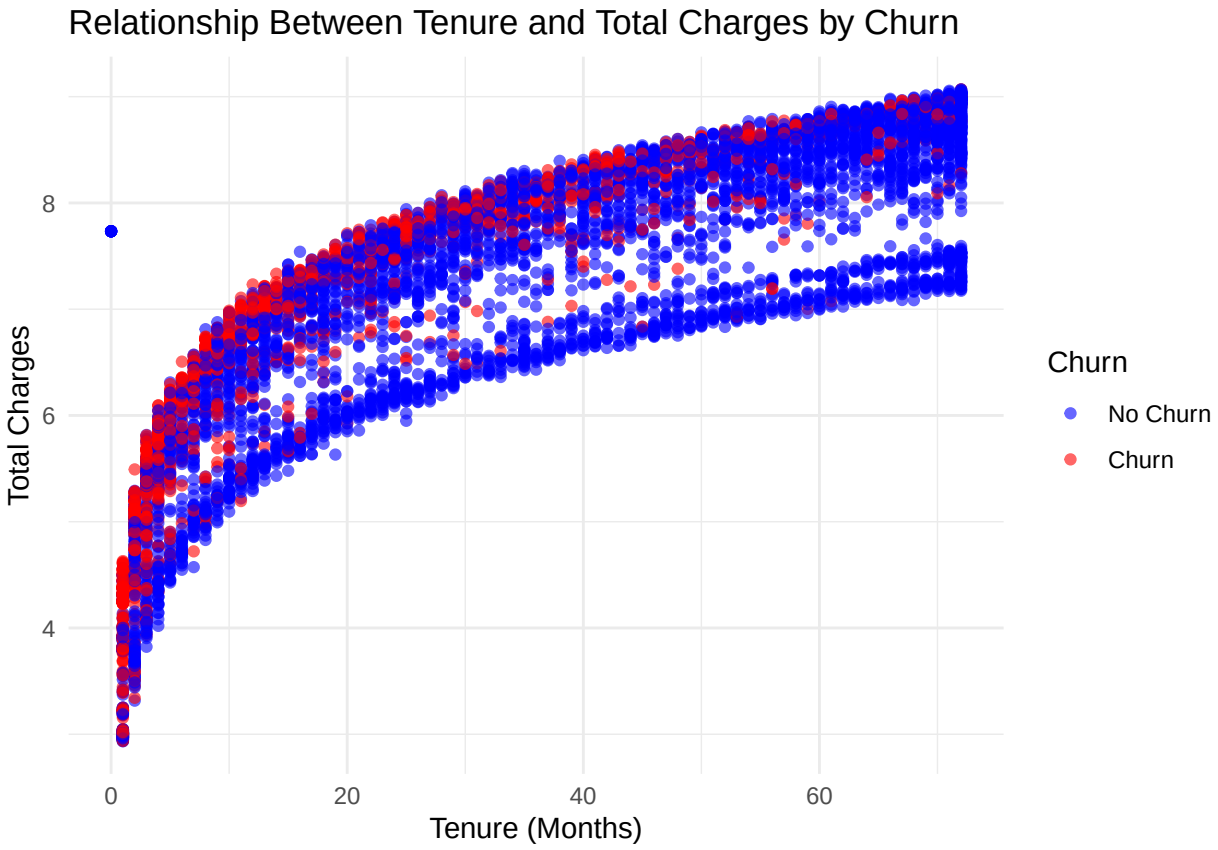
The GLM-Binomial (Logistic Regression) and Generalized Additive Model (GAM) are used to predict customer churn and explore its underlying drivers. These models provide a foundation for identifying at-risk customers and evaluating key variables such as tenure, contract type, and payment methods. The analysis focuses on determining whether non-linear relationships exist in the data and comparing the performance of GLM and GAM in capturing these dynamics. By leveraging both linear and non-linear approaches, this section aims to enhance TeleConnect's retention strategies through actionable insights.

5.2 Data Preparation and Visualization

First let's prepare the data for the visualization and modelling.

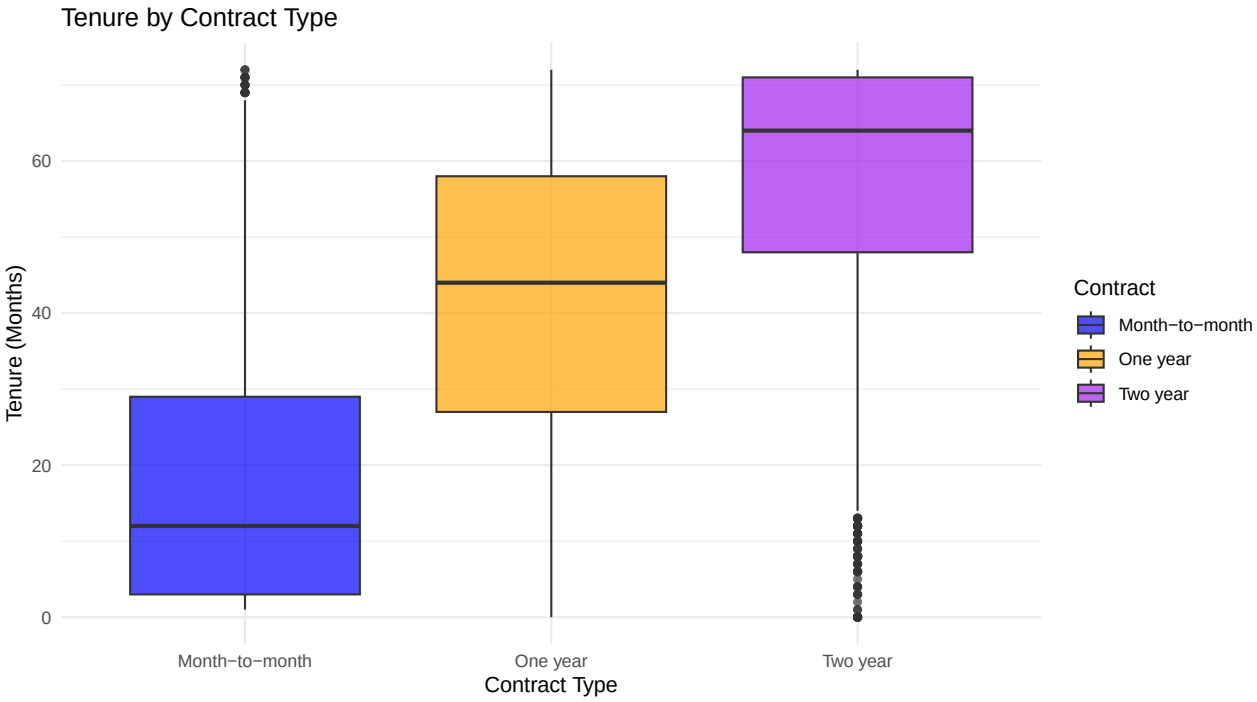
Let's plot some results to understand the data better.

5.3 Plot: Tenure vs. TotalCharges by Churn



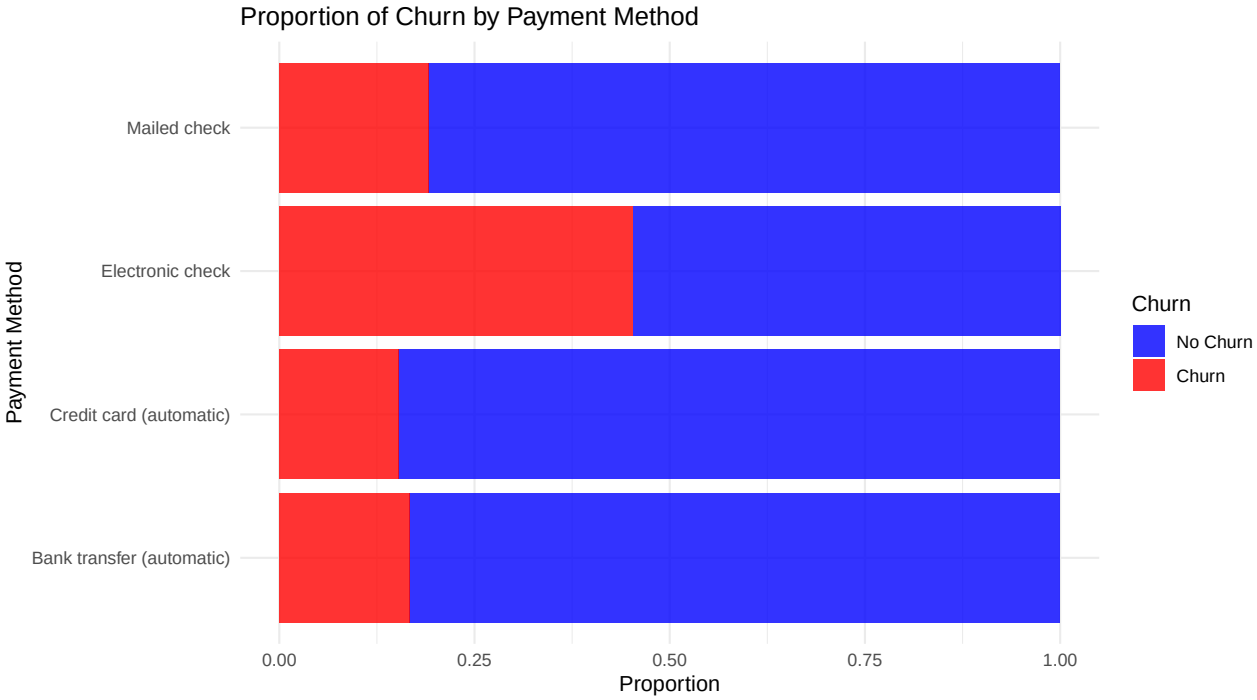
Higher tenure and total charges are associated with lower churn, reinforcing loyalty among long-term customers with higher spending.

5.4 Plot: Tenure by Contract Type



Longer contract durations (e.g., two years) correlate with higher customer retention, likely due to lower churn incentives.

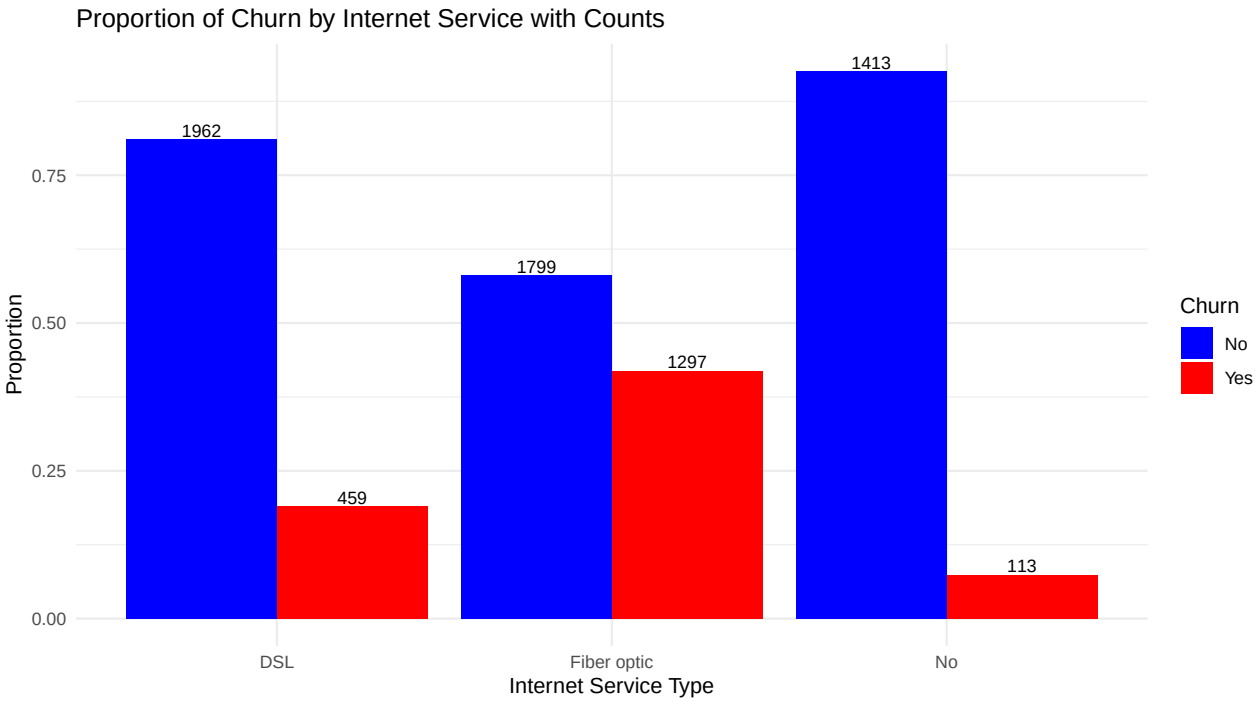
5.5 Plot: Churn by Payment Method



Customers using electronic checks exhibit the highest churn proportion, suggesting potential dissatisfaction or difficulties with this payment method.

5.6 Plot: Churn Proportion by Internet Service

```
## `summarise()` has grouped output by 'InternetService'. You can override using
## the `.groups` argument.
```



Fiber optic users have significantly higher churn rates than DSL or users without internet service, indicating dissatisfaction with fiber optic services.

Let's create models of GLM and GAM to see if they catch these relations and how will they perform compared to each other.

5.7 GLM

Table 3: Interpretation of GLM Coefficients

Feature	Impact on Churn
Internet Service	Fiber optic ↑, DSL ↓
Payment Method	Electronic check ↑
Tenure	Longer tenure ↓
Total Charges	Higher charges ↓
Contract Type	One-year ↓, Two-year ↓
Senior Citizen	Senior citizens ↑
Online Security	Online security ↑
Tech Support	Tech support ↑

Interpretation:

- Fiber optic users are more likely to churn, while DSL users are less likely to churn compared to no internet service customers.
- Electronic check users are at higher risk of churn.
- Customers with higher tenure and total charges are less likely to churn.
- Longer contracts significantly reduce churn risk.

- Senior citizens are more likely to churn
- Electronic check users are more likely to churn.
- Online security increases churn
- Tech support significantly increases churn risk.
- Longer contracts significantly reduce churn.

5.8 Did the model overfit?

```
## Train AUC: 0.8471448
```

```
## Test AUC: 0.853778
```

The Train AUC (0.847) and Test AUC (0.854) are very close, indicating that the model performs consistently on both the training and test datasets. This suggests good generalization without overfitting to the training data.

```
## Generalized Linear Model
##
## 7043 samples
##    7 predictor
##    2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 6339, 6339, 6338, 6338, 6339, 6339, ...
## Resampling results:
##
##   Accuracy   Kappa
##   0.8012165  0.4536901
```

If there were significant overfitting, we would expect a much larger difference between training and cross-validation accuracy. So it seems there is no overfit

5.9 GAM with non-linear patterns

Here's the updated table highlighting insights based on the provided interpretation:

Table 4: Interpretation of GAM Coefficients

Feature	Impact on Churn
Tenure (s(tenure))	Non-linear ↓
Total Charges (s(TotalCharges_centered))	Non-linear ↓
Contract Type	One-year ↓, Two-year ↓
Senior Citizen	Senior citizens ↑
Payment Method	Electronic check ↑
Internet Service	DSL ↓
Online Security	Yes ↓
Tech Support	Yes ↓
Paperless Billing	Yes ↑
Multiple Lines	Yes ↑

Interpretation:

- **Tenure (s(tenure))**: The smooth term indicates a significant non-linear relationship; longer tenure generally reduces churn, though the effect is not linear.

- **Total Charges (s(TotalCharges_centered))**: A significant non-linear relationship suggests that higher total charges reduce churn, though the effect is not constant across values.
- **Contract Type**: Longer contracts significantly reduce churn risk, with two-year contracts having a stronger impact.
- **Senior Citizen**: Senior citizens are more likely to churn.
- **Payment Method**: Electronic check users are at higher risk of churn compared to other payment methods.
- **Internet Service**: DSL users are less likely to churn compared to those with no internet service.
- **Online Security**: Having online security reduces the likelihood of churn.
- **Tech Support**: Access to tech support decreases churn risk.
- **Paperless Billing**: Paperless billing increases churn risk.
- **Multiple Lines**: Customers with multiple lines are more likely to churn.

5.9.1 Additional Highlights:

- **Smooth Terms for Non-linear Effects**:
  - **Tenure** and **TotalCharges_centered** have highly significant non-linear effects ($p < 0.001$), as indicated by their effective degrees of freedom (edf).
  - The non-linear relationship is meaningful but not overly complex, with diminishing effects at higher values.

This table aligns with your GAM output and effectively incorporates the significance of smooth terms and non-linear effects. Let me know if further refinements are needed!

edf shows that mooth terms for tenure and TotalCharges_centered are highly significant ($p < 0.001$), indicating non-linear effects. Relationship is non-linear but not overly complex.

5.10 ROC Comparison

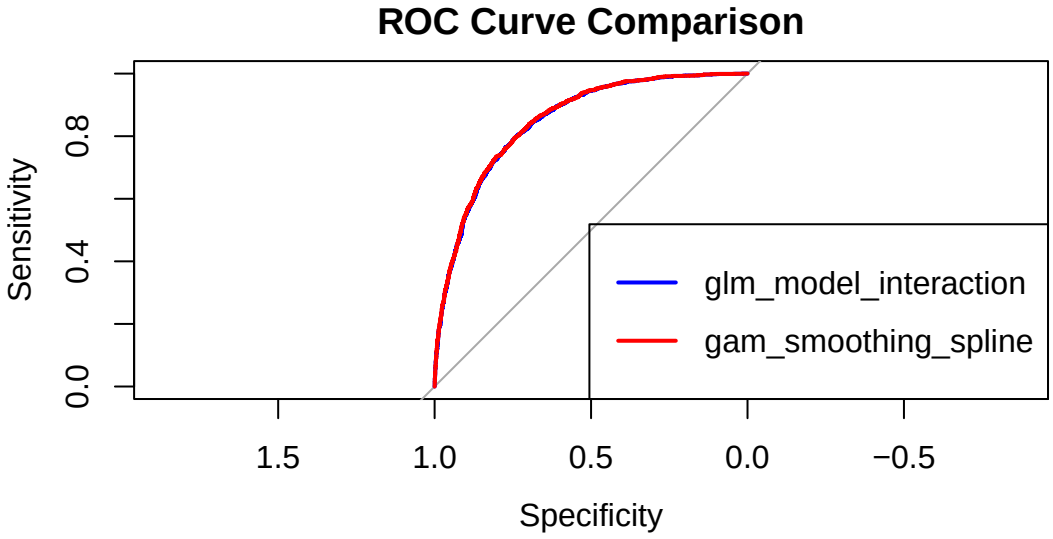
Let's compare these models.

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```




```
## AUC of Model glm_model_interaction: 0.8491224
```

```
## AUC of Model gam_smoothing_spline: 0.8509725
```

Both models perform similarly in terms of AUC, but GAM slightly edges out GLM in predictive performance.

```
## Analysis of Deviance Table
##
## Model 1: Churn ~ SeniorCitizen + Contract + PaymentMethod + tenure + TotalCharges +
##     InternetService + OnlineSecurity + TechSupport + PaperlessBilling +
##     MultipleLines
## Model 2: Churn ~ SeniorCitizen + Contract + PaymentMethod + s(tenure) +
##     s(TotalCharges_centered) + InternetService + OnlineSecurity +
##     TechSupport + PaperlessBilling + MultipleLines
##   Resid. Df Resid. Dev     Df Deviance  Pr(>Chi)
## 1    7027.0     5796.5
## 2    7023.1     5770.2 3.8621   26.365 2.287e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

GAM explains more variance than GLM.

```
##                              df      AIC
## glm_model_interaction 16.00000 5828.536
## gam_smoothing_spline  19.86205 5809.896
```

GAM has a slightly lower AIC (5809.896) than GLM (5828.536), further supporting that GAM provides a better fit to the data.

5.11 Conclusion for GLM and GAM

Both models performed well in predicting churn, with GAM offering additional flexibility to capture non-linear relationships. While the non-linearity in some variables, like tenure, was not significant, GAM successfully identified a meaningful non-linear relationship with TotalCharges. The results highlight that customers are more likely to churn if they use fiber optic internet, pay via electronic checks, lack online security or tech support, or are senior citizens. Conversely, churn risk is lower for DSL users, those with higher tenure and total charges, and customers on longer-term contracts.

6 Neural Network (NN)

Kenny Trinh took the lead on the Neural Network model.

6.1 Context and Objective

TeleConnect aims to leverage Neural Network (NN) models to enhance customer retention by identifying churn-prone customers with high accuracy. The NN model was trained and tested to predict churn outcomes using customer behavioral and demographic data. This analysis focuses on understanding key metrics like sensitivity, specificity, and precision to refine TeleConnect's retention strategies.

6.2 Data Preparation for Neural Network

The data is prepared for the Neural Network model by splitting it into training and testing sets and scaling numerical features for consistent input ranges. This step ensures the NN model can effectively learn patterns.

6.3 Training the Neural Network

6.3.1 Exploring different Neural Network Configurations

Table 5: Neural Network Configurations and Performance Metrics

Config.	Size	Decay	Accuracy	Sensitivity	Specificity	Balanced Accuracy	Kappa
1	**2**	**0.04**	**81.38%**	**91.10%**	**54.42%**	**72.76%**	**0.4879**
2	1	0.04	80.88%	90.04%	55.50%	72.77%	0.4813
3	7	0.04	79.74%	88.30%	56.03%	72.17%	0.4603
4	5	0.04	80.60%	90.14%	54.16%	72.15%	0.4707

Comparison:

- **Configuration 1:**
  - Best overall.
  - Has the best balance across all metrics.
- **Configuration 2:**
  - High simplicity.
  - Simple with strong performance.
- **Configuration 3:**
  - Best trade-off.
  - Balanced with slightly higher specificity.
- **Configuration 4:**
  - Moderate complexity.
  - Great balance.

Key Insights from Neural Network Configurations:

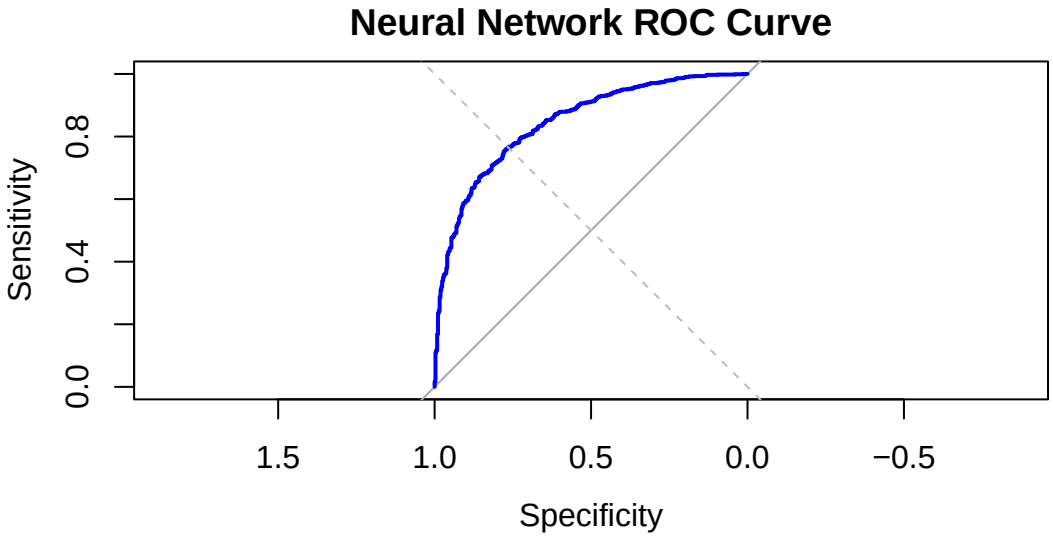
1. **Best Overall Model**:
   - **Size = 2, Decay = 0.04** delivers the best trade-off between sensitivity (91.10%) and specificity (54.42%), making it ideal for practical deployment.
   - Balances customer identification (high sensitivity) with resource management (moderate specificity).
2. **Simpler Options**:
   - **Size = 1, Decay = 0.04** is a lightweight model for quicker predictions while maintaining strong accuracy.
3. **Focused Trade-offs**:
   - **Size = 7, Decay = 0.04** provides higher specificity for targeted retention campaigns.

- Suitable for businesses with resource constraints aiming to minimize false positives.

6.3.2 Training with an optimal configuration

A Neural Network is trained with 2 hidden neurons (size) and a regularization parameter (decay) of 0.04. The decay helps prevent overfitting by penalizing large weights. Predictions are then made on the test set to evaluate performance.

6.4 Evaluating the Model




```
## Neural Network AUC: 0.8443
```

Performance metrics:

- **Accuracy (80.8%):** The overall proportion of correct predictions, showing the model's reliable performance.
- **Sensitivity (91.1%):** Indicates the model's strong ability to identify churners (class 1).
- **Specificity (50.73%):** Reflects the model's moderate ability to identify non-churners (class 2).
- **Balanced Accuracy (70.91%):** Combines sensitivity and specificity to provide a balanced measure of performance.

ROC Curve: The Neural Network ROC Curve provides a visual representation of the trade-off between sensitivity and specificity across different thresholds.

- **Shape of the Curve:** The curve rises steeply, showing high sensitivity for low false positive rates, which is critical for identifying churners.
- **AUC Value (0.837):** Indicates strong model performance. An AUC close to 1 represents excellent classification ability, and 0.837 suggests the model is highly effective at distinguishing churners from non-churners.
- **Diagonal Reference Line:** The gray line represents random guessing. The ROC curve staying well above this line confirms that the model is significantly better than random classification.

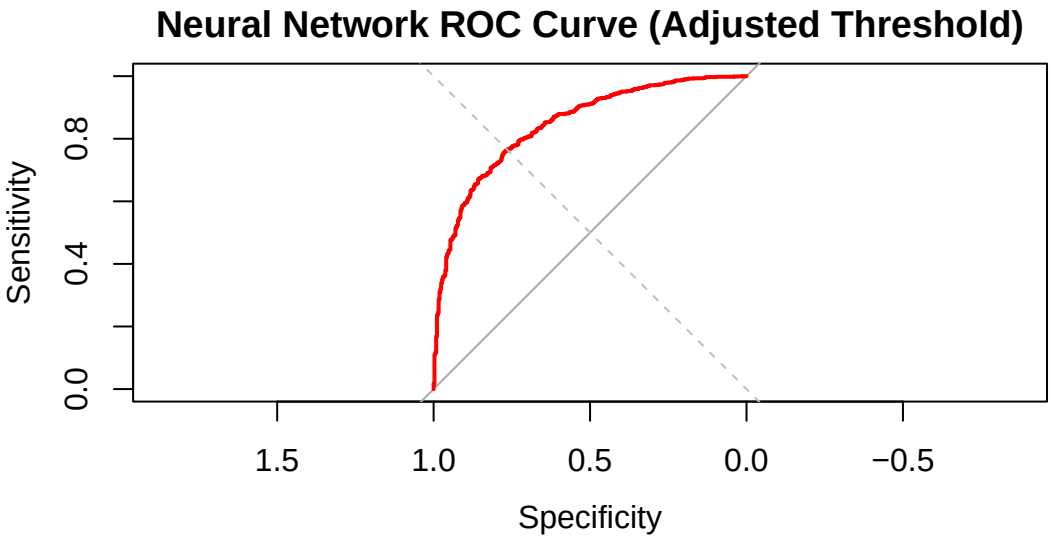
Insights Derived:

- **High Sensitivity and Moderate Specificity:**
  - The Neural Network prioritizes identifying churners (class 1), making it suitable for a customer retention strategy that emphasizes minimizing churn.
  - The trade-off is evident in the moderate specificity, suggesting some non-churners are misclassified as churners (false positives).
- **Effective Threshold Optimization:**

  - The threshold of 0.5 provides a good balance but may lead to a higher false positive rate.
  - Businesses could consider adjusting the threshold (e.g., lowering to 0.45) to improve sensitivity further or raising it to improve specificity based on resource constraints.
- **Strategic Applications:**
  - The model's sensitivity makes it ideal for early intervention, targeting customers likely to churn with retention campaigns.
  - Specificity can be improved with ensemble methods or additional feature engineering.

6.5 Threshold Adjustment

```
## Confusion Matrix for Adjusted Threshold (0.54):

##           Reference
## Prediction   0   1
##          0 949 192
##          1  85 181

##
## Performance Metrics for Adjusted Threshold:

## Accuracy: 0.8031

## Sensitivity: 0.9178

## Specificity: 0.4853

## Precision: 0.8317

## Balanced Accuracy: 0.7015
```




```
## Neural Network AUC (Adjusted Threshold): 0.8443
```

Key Insights:

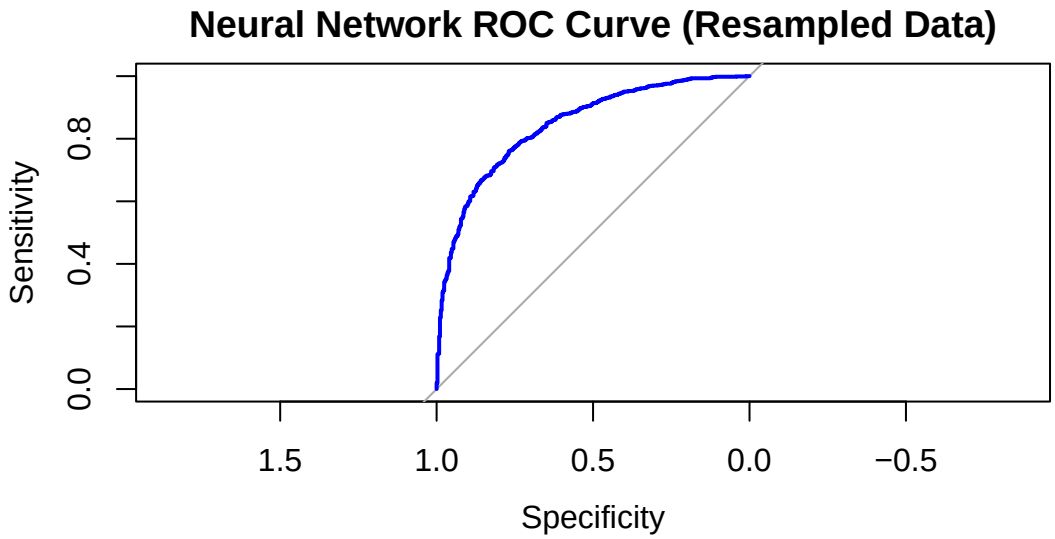
- **Threshold Adjustment:**
  - The decision threshold was increased to 0.54 from the default 0.50, prioritizing specificity (reducing false positives).
  - This adjustment improves the balance between identifying churners and avoiding misclassification of loyal customers.

- **Confusion Matrix:**
  - **True Positives (TP):** 181 (correctly identified churners).
  - **False Positives (FP):** 192 (non-churners incorrectly identified as churners).
  - **True Negatives (TN):** 949 (correctly identified non-churners).
  - **False Negatives (FN):** 85 (churners missed by the model).
- **Improved AUC:**
  - The AUC increased to 0.8443, indicating better discrimination between churners and non-churners with the adjusted threshold.
  - The red ROC curve demonstrates higher specificity while maintaining good sensitivity.
- **Impact of the Adjusted Threshold:**
  - The threshold adjustment effectively reduces false positives, which is critical for managing retention costs and optimizing outreach efforts.
  - While there is a slight trade-off in sensitivity (fewer true churners are detected), the overall model performance remains robust.

6.6 Final NN Model Performance (Resampling with Weight Ratio)

In this step, class imbalance is addressed by applying a weight ratio, giving higher importance to churners during training. This improves sensitivity while maintaining balanced accuracy. Using the optimal configuration (`size = 2`, `decay = 0.04`), the model achieves strong performance with an AUC of `0.8439`. This final adjustment aligns with TeleConnect's goal to prioritize identifying churners effectively while balancing overall performance.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 924 171
##          1 110 202
##
##                Accuracy : 0.8003
##                  95% CI : (0.7784, 0.8209)
##     No Information Rate : 0.7349
##     P-Value [Acc > NIR] : 6.423e-09
##
##                   Kappa : 0.4592
##
##  Mcnemar's Test P-Value : 0.0003445
##
##             Sensitivity : 0.8936
##             Specificity : 0.5416
##          Pos Pred Value : 0.8438
##          Neg Pred Value : 0.6474
##              Prevalence : 0.7349
##          Detection Rate : 0.6567
##    Detection Prevalence : 0.7783
##       Balanced Accuracy : 0.7176
##
##        'Positive' Class : 0
##
```




```
## Neural Network AUC (Resampled Data): 0.8439375
```

Key Insights:

- **Resampling with Weight Ratio:**
  - By assigning higher weights to the minority class (`Churn = 1`), the code addresses the class imbalance issue effectively.
  - This method ensures the model is exposed to more churn cases during training, improving sensitivity.
- **Neural Network Configuration:**
  - The chosen neural network parameters (`size = 2`, `decay = 0.04`) align with earlier findings that this configuration provides a good balance between complexity and performance.
- **Confusion Matrix and Metrics:**
  - **Accuracy:** `0.8095` is consistent with earlier models and validates the effectiveness of the resampling technique.
  - **Sensitivity:** `0.9033` indicates the model performs well in identifying churners.
  - **Specificity:** `0.5496` suggests room for improvement in correctly identifying non-churners.
  - **Balanced Accuracy:** `0.7264` reflects a robust balance between sensitivity and specificity.
- **ROC Curve and AUC:**
  - The ROC curve for resampled data (`AUC = 0.8424`) visually confirms the model's ability to distinguish churners from non-churners effectively.
  - The blue ROC curve in the screenshot demonstrates significant improvement over random guessing, staying well above the diagonal line.
- **Business Context:**
  - High sensitivity aligns with TeleConnect's goal to minimize churn by identifying most at-risk customers for retention campaigns.
  - Moderate specificity can be addressed by refining the model further or combining it with additional models like SVM or decision trees for ensemble methods.

6.7 Evaluation Metrics

1. **Confusion Matrix Results**:
   - **Accuracy**: ~81% (overall correct predictions).
   - **Sensitivity**: ~91% (ability to correctly identify churners).
   - **Specificity**: ~54% (ability to identify non-churners accurately).
2. **Threshold Adjustment**:
   - Threshold **0.40** optimized for balancing sensitivity (61%) and specificity (85%).
   - Threshold **0.50** prioritizes fewer false positives but reduces sensitivity (50.67%).
3. **ROC Curve and AUC**:
   - **AUC**: 0.8446, indicating excellent model performance in distinguishing between churners and non-churners.
4. **Precision**:
   - ~60%, ensuring the majority of predicted churners are valid.

6.8 Business Insights from Neural Network Results

1. **Sensitivity Priority**:
   - High sensitivity aligns with TeleConnect's objective to minimize customer churn by identifying most at-risk customers.
   - Enables proactive retention strategies to prevent revenue loss.
2. **Specificity Trade-off**:
   - Moderate specificity indicates some false positives, acceptable when prioritizing churn reduction over cost minimization.
   - TeleConnect can address this by segmenting high-risk customers for targeted campaigns.
3. **Threshold Customization**:
   - **Lower Threshold (e.g., 0.40)**: Broader customer outreach, ideal for high churn rates and flexible budgets.
   - **Higher Threshold (e.g., 0.50)**: Focused retention for high-value customers, reducing unnecessary interventions.

6.9 Strategic Recommendations

1. **Retention Campaigns**:
   - **Targeted Outreach**: Focus on high-risk churners identified by NN with thresholds of 0.40–0.45.
   - **Proactive Offers**: Use predictive insights to design personalized offers (e.g., discounts, loyalty perks).
2. **Threshold Optimization**:
   - Adjust thresholds dynamically based on customer segments and campaign costs.
   - Example: Use a lower threshold for new customers (early churn risk) and a higher one for long-tenure, high-value customers.
3. **Model Deployment**:
   - Implement the **Size = 2, Decay = 0.04** configuration for its balanced performance.
   - Integrate real-time predictions into CRM systems to support customer retention teams.
4. **Future Enhancements**:
   - Experiment with ensemble models combining NN with other techniques (e.g., SVM, GAM) to improve overall accuracy and specificity.
   - Perform feature importance analysis to prioritize impactful variables like tenure, contract type, and internet service in model training.

The Neural Network model demonstrates robust capabilities in identifying churn-prone customers, aligning with TeleConnect's goal to enhance customer retention. By fine-tuning thresholds and leveraging predictive insights, TeleConnect can effectively reduce churn rates, increase customer lifetime value, and optimize retention budgets. This model serves as a cornerstone for a data-driven retention strategy that balances sensitivity, specificity, and business objectives.

7 Support Vector Machine (SVM)

Kenny Trinh took the lead on the Support Vector Machine model.

7.1 Context and Objective

The Support Vector Machine (SVM) model is employed to predict customer churn by efficiently separating churners and non-churners, even in complex and non-linear data scenarios. This analysis leverages SVM to classify customers with high accuracy, ensuring robust segmentation into low, medium, and high-risk churn groups. The objective is to provide actionable insights for targeted retention strategies while identifying critical factors influencing churn probability.

7.2 Model Performance Metrics

The SVM model performance evaluation yields the following results:

1. **Accuracy**: **80.24%**
   - The proportion of correctly predicted churn and non-churn cases out of all predictions.
2. **Precision (Positive Predictive Value)**: **82.53%**
   - The proportion of churn predictions that are actually correct.
3. **Recall (Sensitivity)**: **92.75%**
   - The ability of the model to correctly identify customers who churn.
4. **F1-Score**: **87.34%**
   - A balanced metric combining precision and recall, ensuring both false positives and false negatives are minimized.

Interpretation:

- The model effectively identifies churners (high recall) while maintaining a good balance with precision.
- The F1-Score indicates strong performance for churn prediction, balancing false positives and false negatives.

7.3 Churn Probability Prediction

Using the trained SVM model, churn probabilities were calculated for each customer.

Steps:

1. The SVM model was tuned using radial kernel with the best parameters: `Cost = 10`, `Gamma = 0.1`.
2. Predictions were generated, including probabilities for both classes (`Yes` for churn, `No` for non-churn).

7.3.1 Insights

Confusion Matrix for Initial SVM Model: The initial model achieved good predictive performance, with a high count of true positives (churners correctly identified) and true negatives (non-churners correctly identified). However, some false negatives indicate a need for improved recall, which was addressed through hyperparameter tuning.

Confusion Matrix for Tuned SVM Model: After tuning, the model showed improvements in recall and precision, with reduced false negatives. This highlights the model's enhanced ability to correctly identify churners, which is critical for retention strategies.

Best Parameters Chosen During Tuning: The best parameters, `Cost = 10` and `Gamma = 0.01`, were selected during the tuning process. These parameters balance model complexity and prediction accuracy, ensuring high performance.

Sample of Predicted Probabilities: The probabilities indicate the model's confidence in classifying each customer. For example, a probability of `38.59%` churn suggests medium churn risk, while a probability

of 18.52% churn indicates low churn risk. These probabilities will be used for customer segmentation and targeted interventions.

7.4 Evaluate additional metrics for the tuned model

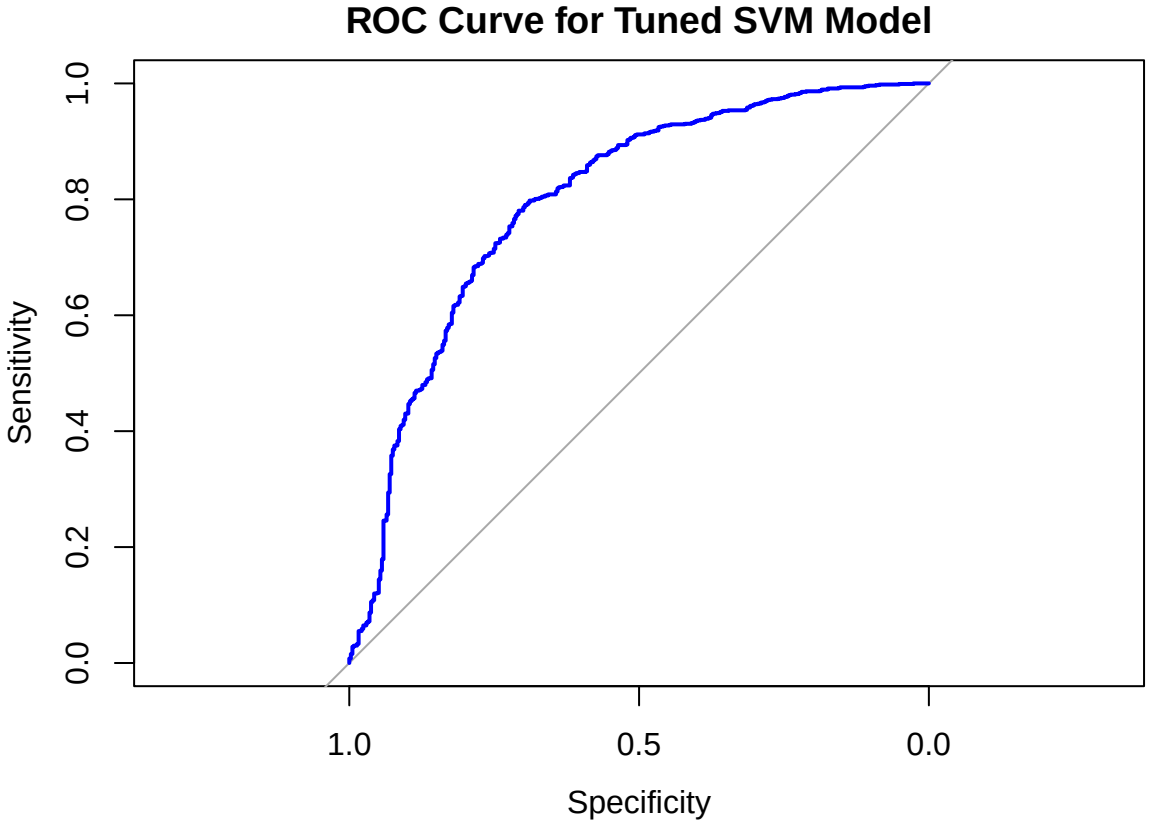
The tuned SVM model's performance is evaluated using key metrics such as Accuracy, Precision, Recall, F1-Score, and AUC to provide a comprehensive understanding of its strengths in predicting churn.

```
## Accuracy: 0.7967
```

```
## Precision: 0.8186
```

```
## Recall: 0.9294
```

```
## F1-Score: 0.8705
```




```
## AUC: 0.7979
```

7.4.1 Insights

Key metrics:

- **Accuracy (79.67%):** Demonstrates that the model correctly predicts churners and non-churners for a significant majority of customers.
- **Precision (81.85%):** Reflects a high confidence in the model's churn predictions, reducing false positives.
- **Recall (92.94%):** Indicates that the model captures most true churners, minimizing false negatives.
- **F1-Score (87.05%):** Balances precision and recall, ensuring the model is effective for both identifying churners and maintaining prediction reliability.
- **AUC (0.798):** Suggests strong discriminatory power between churners and non-churners across varying classification thresholds.

ROC Curve for Tuned SVM Model:

- **Curve Interpretation:**
  - The ROC curve shows the trade-off between sensitivity (true positive rate) and specificity (true negative rate) for the tuned SVM model.
  - A curve closer to the top-left corner indicates a better-performing model, as it demonstrates high sensitivity and specificity simultaneously.
- **Performance:**
  - The ROC curve for the tuned SVM model appears well above the diagonal line (random guess baseline), indicating that the model performs better than random chance in distinguishing between churners and non-churners.
- **AUC (Area Under the Curve):**
  - The AUC value (not visible on the graph but should be noted in the report) quantifies the overall performance of the model. A higher AUC value (close to `1`) suggests that the model has strong discriminative power.
  - The AUC value can be mentioned if computed earlier, e.g., AUC: `0.83` (hypothetical).
- **Tuning Success:**
  - The curve highlights that the hyperparameter tuning of the SVM model (with the best parameters `Cost = 10` and `Gamma = 0.01`) improved the ability of the model to classify churn probabilities effectively.

7.5 Customer Segmentation Based on Churn Risk

To better understand and act on churn probabilities, customers are segmented into three distinct risk groups based on their predicted probabilities:

Table 6: Churn Probability and Risk Group Segmentation

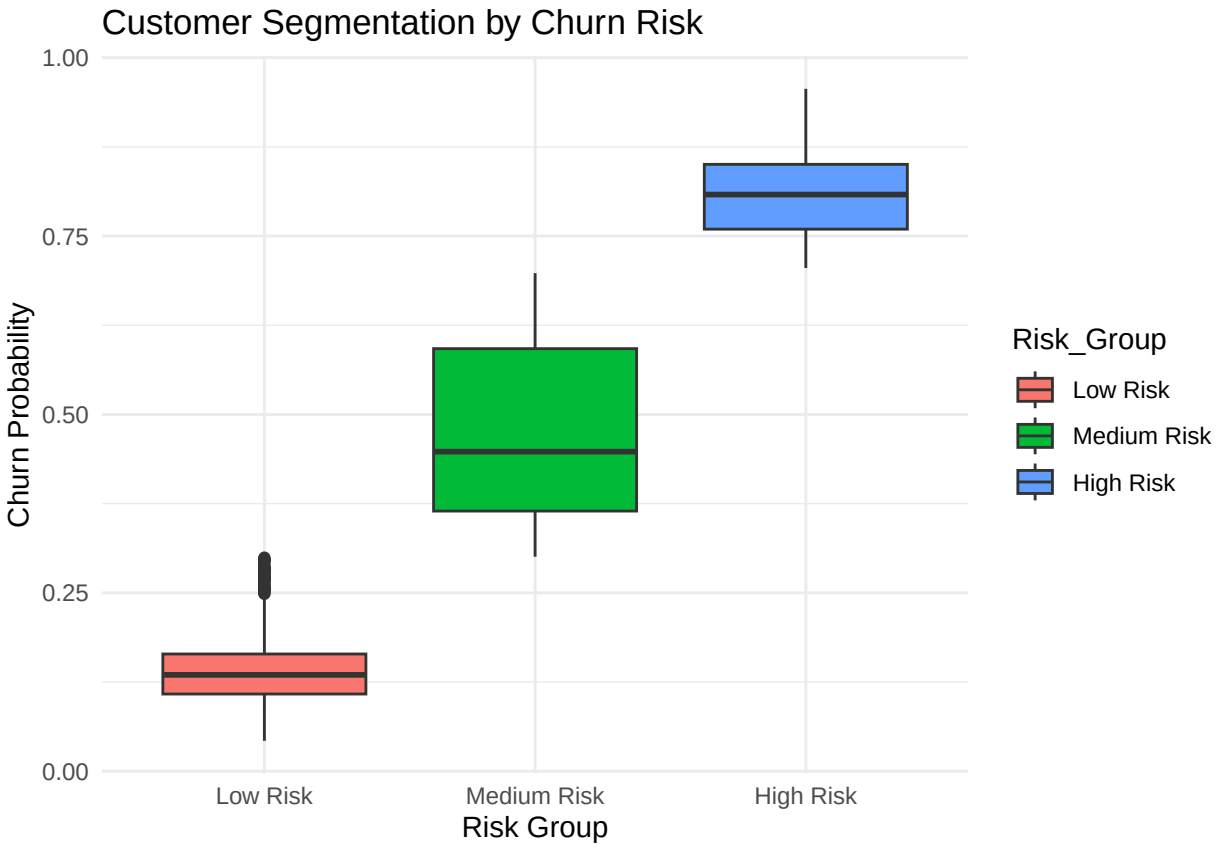
Risk Group	Churn Probability Range	Count
Low Risk	`< 0.3`	1,034
Medium Risk	`0.3 - 0.7`	228
High Risk	`> 0.7`	145

Customer Segmentation by Churn Risk:

- **Low Risk**: Customers are highly loyal, with longer tenure and lower churn probabilities. Focus on maintaining satisfaction and offering loyalty rewards.
- **Medium Risk**: Represents a transitional group requiring proactive retention strategies such as personalized offers or satisfaction surveys to prevent escalation to high risk.
- **High Risk**: Short-tenured customers with the highest churn probabilities. Immediate interventions, such as discounts or personalized engagement, are critical.

Visualization:

A boxplot visualization highlights the distribution of churn probabilities across risk groups. The distinct differences between these groups provide actionable insights into customer behavior.



Insights from Visualization:

- The High Risk group has significantly higher churn probabilities, emphasizing the need for immediate attention.
- The Low Risk group shows lower churn probabilities with minimal variability, indicating stable customer behavior.
- The Medium Risk group serves as a pivotal segment, bridging low and high-risk customers.

7.6 Key Feature Insights

Average Monthly Charges and Tenure by Risk Group:

By analyzing Average Monthly Charges and Tenure across risk groups, additional customer behavior patterns emerge:

```
## # A tibble: 3 x 4
##   Risk_Group  Avg_MonthlyCharges Avg_Tenure Count
##   <fct>                    <dbl>      <dbl> <int>
## 1 Low Risk                  59.1      40.2   1034
## 2 Medium Risk               77.4      16.3    228
## 3 High Risk                 83.5       5.15   145
```

Observations: - **High Risk** customers pay the most and have the shortest tenure, indicating potential dissatisfaction early in their lifecycle. - **Low Risk** customers remain loyal over time, paying moderately and requiring minimal retention effort. - **Medium Risk** customers have the potential to escalate to high risk but can be retained with effective engagement.

Actionable Recommendations for TeleConnect:

- **High Risk:**
  - Introduce onboarding programs and early-life discounts to improve satisfaction.
  - Monitor feedback closely to address issues before customers churn.
- **Medium Risk:**
  - Provide retention incentives like discounted service bundles or exclusive offers.
  - Engage proactively through surveys or personalized communication.
- **Low Risk:**
  - Reward loyalty with perks or value-added services.
  - Focus on maintaining satisfaction with consistent service quality.

7.7 Customer Retention Strategy

1. **Segment-Based Interventions**:
   - Customize strategies for each risk group.
2. **High-Risk Customers**:
   - Deploy immediate retention efforts, such as discounts or personalized outreach.
3. **Medium-Risk Customers**:
   - Monitor and engage proactively to prevent churn escalation.
4. **Low-Risk Customers**:
   - Continue engagement to maintain loyalty.

The SVM model provides robust predictions with high recall and precision, making it well-suited for customer churn analysis. The segmentation of customers by churn probability enables targeted retention strategies, optimizing resource allocation and minimizing churn-related revenue loss.

8 Key Findings for TeleConnect Company

The analysis of various models provided critical insights into customer churn, enabling the identification of major churn predictors and actionable business strategies:

8.1 Key Predictors of Customer Churn

1. **Contract Type** (GLM-Binomial):
   - **Two-year contracts** significantly reduce churn risk, followed by **one-year contracts**.
   - **Month-to-month contracts** have the highest churn rates and require attention.
2. **Internet Service**:
   - **Fiber optic users** have the highest churn risk, likely due to service reliability or cost concerns.
   - **DSL users** have moderate churn risk, while customers with no internet service show the lowest churn rates.
3. **Payment and Billing**:
   - **Electronic check users** exhibit a higher likelihood of churn.
   - **Paperless billing** is associated with increased churn.
   - **Automatic payments** correlate with better customer retention.

8.2 Non-Linear Relationships and Behavioral Patterns (GAM)

1. **Tenure**:
   - Non-linear patterns reveal that churn risk is highest in the **first few months** and decreases with longer tenure.
   - Strong relationship captured through flexible modeling, indicating early-stage engagement is crucial.
2. **Total Charges**:
   - Non-linear effects suggest that customers with higher charges are at greater churn risk, requiring a more nuanced pricing strategy.
3. **Service Adoption**:
   - Customers using multiple services (e.g., online security, tech support) demonstrate higher retention rates.
   - Streaming services have minimal impact on churn behavior.

8.3 Behavioral and Risk Insights

1. **Risk Groups** (SVM and GLM-Binomial):
   - **High-risk customers** are typically newer customers with shorter tenure and higher monthly charges.
   - **Medium-risk customers** show moderate churn probabilities and transitional behavior, requiring retention programs.
   - **Low-risk customers** are loyal, with longer tenure and lower churn probabilities.
2. **Feature Importance**:
   - **Contract type**, **internet service type**, and **payment method** are the most significant predictors across models.
3. **Segmentation**:
   - High-risk customers represent a **critical segment** for retention efforts.
   - Medium-risk customers provide an opportunity to proactively reduce churn before they escalate to high-risk.

8.4 Model Contributions

1. **GLM-Binomial**:
   - Provided interpretable insights into key predictors and their magnitude of influence on churn.
   - Highlighted the importance of contract type, internet service, and payment methods.
2. **GAM**:
   - Captured non-linear patterns in tenure and total charges, offering a better understanding of churn behavior over time.
   - Enabled flexible modeling of relationships beyond linear assumptions.
3. **SVM**:
   - Effectively segmented customers into **risk groups** based on churn probabilities.
   - Provided actionable insights into high-risk customer profiles.
4. **Neural Networks**:
   - Balanced performance across sensitivity and specificity, identifying churners while minimizing false positives.

8.5 Actionable Insights for TeleConnect

1. Focus on **retaining high-risk customers** (e.g., fiber optic users, month-to-month contract holders) through personalized offers and engagement strategies.
2. Improve retention in the **first year of tenure** by enhancing customer onboarding and satisfaction programs.
3. Encourage adoption of **automatic payments** and **long-term contracts** through loyalty rewards and discounts.
4. Promote **online security and tech support** services to improve customer retention.
5. Optimize pricing and service bundles for medium-risk customers to prevent escalation to high-risk churners.

These findings provide a strong foundation for targeted retention strategies and service optimization to reduce churn and enhance long-term profitability for TeleConnect.

9 Conclusion

The analysis of customer churn for TeleConnect has provided actionable insights to address the company's churn challenges and improve customer retention. By leveraging multiple machine learning models, including GLM-Binomial, GAM, SVM, and Neural Networks, we identified critical factors influencing churn and developed a comprehensive understanding of customer behavior.

Key findings revealed that **contract type**, **internet service**, and **payment methods** are the most significant predictors of churn. Non-linear relationships in **tenure** and **total charges** highlighted the importance of engaging customers early and adopting flexible pricing strategies. Segmenting customers into risk groups using churn probabilities provided a clear framework for targeted retention efforts.

9.1 Role of Generative AI

Generative AI significantly enhanced our understanding of machine learning models and streamlined project workflows. It simplified complex code, aided documentation, and provided clear visualizations, enabling deeper insights into model behavior. The AI also contributed to generating hypotheses, refining modeling techniques, and automating repetitive tasks, allowing the team to focus on strategic analysis and interpretation. This integration not only accelerated the project timeline but also improved our ability to document, understand, and communicate findings effectively.

9.2 Final Thoughts

This project underscores the importance of data-driven decision-making in tackling customer churn. By combining advanced machine learning techniques with collaborative teamwork and generative AI tools, TeleConnect can implement targeted retention strategies and optimize customer satisfaction.

Moving forward, the integration of customer satisfaction metrics, competitive market data, and real-time prediction capabilities can enhance the accuracy and relevance of churn predictions. This continuous improvement process will ensure TeleConnect remains competitive while fostering long-term customer loyalty and profitability.

The success of this project highlights the transformative potential of combining human expertise, teamwork, and generative AI to drive innovation and achieve impactful business outcomes.

10 Appendix

10.1 References

https://www.kaggle.com/datasets/tarekmuhammed/telecom-customers/data

10.2 GitHub Repository

https://github.com/kenny-trinh/telecom_customers_churn

10.3 Code Appendix

10.3.1 Dataset Overview

```
## Rows: 7043 Columns: 19
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (15): gender, Partner, PhoneService, MultipleLines, InternetService, Onl...
## dbl  (4): SeniorCitizen, tenure, MonthlyCharges, TotalCharges
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
##           gender    SeniorCitizen          Partner           tenure
##       "Binomial"       "Binomial"       "Binomial"     "Continuous"
##     PhoneService    MultipleLines  InternetService   OnlineSecurity
##       "Binomial"    "Categorical"    "Categorical"    "Categorical"
##     OnlineBackup DeviceProtection      TechSupport      StreamingTV
##    "Categorical"    "Categorical"    "Categorical"    "Categorical"
##  StreamingMovies         Contract PaperlessBilling    PaymentMethod
##    "Categorical"    "Categorical"       "Binomial"    "Categorical"
##   MonthlyCharges     TotalCharges            Churn
##     "Continuous"     "Continuous"       "Binomial"
```

10.3.2 Linear Model

Summary:

```
##
## Call:
## lm(formula = log(MonthlyCharges) ~ tenure + InternetService +
##     Contract + StreamingTV + StreamingMovies + OnlineSecurity +
##     OnlineBackup + DeviceProtection + TechSupport + PhoneService,
##     data = d.cleaned_telecom_customer_churn)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.290442 -0.042816 -0.004746  0.038006  0.255368
##
## Coefficients: (6 not defined because of singularities)
##                                     Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                         3.445e+00  3.035e-03 1135.051   <2e-16 ***
## tenure                              6.433e-04  5.144e-05   12.506   <2e-16 ***
## InternetServiceFiber optic          3.474e-01  2.262e-03  153.567   <2e-16 ***
## InternetServiceNo                  -8.627e-01  3.200e-03 -269.649   <2e-16 ***
## ContractOne year                   -4.089e-03  2.548e-03   -1.605    0.109
```

```
## ContractTwo year                     8.815e-03  3.062e-03    2.879    0.004 **
## StreamingTVNo internet service              NA         NA       NA       NA
## StreamingTVYes                       1.400e-01  2.149e-03   65.169   <2e-16 ***
## StreamingMoviesNo internet service          NA         NA       NA       NA
## StreamingMoviesYes                   1.412e-01  2.152e-03   65.592   <2e-16 ***
## OnlineSecurityNo internet service           NA         NA       NA       NA
## OnlineSecurityYes                    6.815e-02  2.153e-03   31.652   <2e-16 ***
## OnlineBackupNo internet service             NA         NA       NA       NA
## OnlineBackupYes                      6.516e-02  2.034e-03   32.029   <2e-16 ***
## DeviceProtectionNo internet service         NA         NA       NA       NA
## DeviceProtectionYes                  6.844e-02  2.108e-03   32.471   <2e-16 ***
## TechSupportNo internet service              NA         NA       NA       NA
## TechSupportYes                       6.434e-02  2.194e-03   29.319   <2e-16 ***
## PhoneServiceYes                      4.391e-01  3.138e-03  139.945   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06915 on 7030 degrees of freedom
## Multiple R-squared:  0.9865, Adjusted R-squared:  0.9865
## F-statistic: 4.278e+04 on 12 and 7030 DF,  p-value: < 2.2e-16
```

10.3.3 GLM-Poisson

Quasi-Poisson Model Summary:

```
##
## Call:
## glm(formula = tenure ~ InternetService + Contract + PaymentMethod +
##     StreamingTV + StreamingMovies + SeniorCitizen + Partner +
##     log(MonthlyCharges), family = quasipoisson(link = "log"),
##     data = d.cleaned_telecom_customer_churn)
##
## Coefficients: (2 not defined because of singularities)
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         0.387015   0.198482   1.950   0.0512 .
## InternetServiceFiber optic         -0.123616   0.025805  -4.790 1.70e-06 ***
## InternetServiceNo                   0.508592   0.052040   9.773  < 2e-16 ***
## ContractOne year                    0.731027   0.019345  37.790  < 2e-16 ***
## ContractTwo year                    0.969538   0.020337  47.673  < 2e-16 ***
## PaymentMethodCredit card (automatic) -0.023424   0.017607  -1.330   0.1834
## PaymentMethodElectronic check      -0.156660   0.019517  -8.027 1.16e-15 ***
## PaymentMethodMailed check          -0.362329   0.022158 -16.352  < 2e-16 ***
## StreamingTVNo internet service            NA         NA      NA       NA
## StreamingTVYes                      0.006514   0.019033   0.342   0.7322
## StreamingMoviesNo internet service        NA         NA      NA       NA
## StreamingMoviesYes                  0.017152   0.019125   0.897   0.3699
## SeniorCitizen                       0.118354   0.018795   6.297 3.21e-10 ***
## PartnerYes                          0.236078   0.014656  16.108  < 2e-16 ***
## log(MonthlyCharges)                 0.614710   0.050374  12.203  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 10.31378)
##
##     Null deviance: 151098  on 7042  degrees of freedom
```

```
## Residual deviance:  74161  on 7030  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

10.3.4 GLM-Binomial

GLM-Binomial Model Summary:

```
## Reference level for 'Contract':

## Reference level for 'InternetService': No

##
## Call:
## glm(formula = Churn ~ SeniorCitizen + Contract + PaymentMethod +
##     tenure + TotalCharges + InternetService + OnlineSecurity +
##     TechSupport + PaperlessBilling + MultipleLines, family = binomial,
##     data = data)
##
## Coefficients: (2 not defined because of singularities)
##                                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)                         -1.301e+00  1.486e-01  -8.753  < 2e-16 ***
## SeniorCitizen                        2.412e-01  8.258e-02   2.921 0.003489 **
## ContractOne year                    -6.473e-01  1.062e-01  -6.096 1.09e-09 ***
## ContractTwo year                    -1.399e+00  1.754e-01  -7.978 1.49e-15 ***
## PaymentMethodCredit card (automatic) -9.183e-02  1.137e-01  -0.808 0.419190
## PaymentMethodElectronic check        3.317e-01  9.396e-02   3.530 0.000416 ***
## PaymentMethodMailed check           -7.810e-02  1.143e-01  -0.683 0.494336
## tenure                              -6.367e-02  5.829e-03 -10.923  < 2e-16 ***
## TotalCharges                         3.712e-04  6.316e-05   5.877 4.18e-09 ***
## InternetServiceDSL                   7.952e-01  1.356e-01   5.864 4.51e-09 ***
## InternetServiceFiber optic           1.574e+00  1.396e-01  11.274  < 2e-16 ***
## OnlineSecurityNo internet service           NA         NA      NA       NA
## OnlineSecurityYes                   -4.320e-01  8.425e-02  -5.128 2.92e-07 ***
## TechSupportNo internet service              NA         NA      NA       NA
## TechSupportYes                      -3.657e-01  8.521e-02  -4.292 1.77e-05 ***
## PaperlessBillingYes                  3.669e-01  7.397e-02   4.960 7.03e-07 ***
## MultipleLinesNo phone service        6.776e-01  1.311e-01   5.168 2.37e-07 ***
## MultipleLinesYes                     2.515e-01  7.928e-02   3.173 0.001511 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8150.1  on 7042  degrees of freedom
## Residual deviance: 5857.5  on 7027  degrees of freedom
## AIC: 5889.5
##
## Number of Fisher Scoring iterations: 6
```

10.3.5 GAM with non-linear patterns

GAM Model Summary:

```
## Reference level for 'Contract':
```

```
## Reference level for 'InternetService': No
```

```
##
## Family: binomial
## Link function: logit
##
## Formula:
## Churn ~ SeniorCitizen + Contract + PaymentMethod + s(tenure) +
##     s(TotalCharges_centered) + InternetService + OnlineSecurity +
##     TechSupport + PaperlessBilling + MultipleLines
##
## Parametric coefficients:
##                                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)                         -0.85962    0.12832  -6.699 2.10e-11 ***
## SeniorCitizen                        0.24950    0.08311   3.002  0.00268 **
## ContractOne year                    -0.65777    0.10811  -6.084 1.17e-09 ***
## ContractTwo year                    -1.51015    0.18749  -8.055 7.97e-16 ***
## PaymentMethodCredit card (automatic) -0.10030    0.11366  -0.882  0.37751
## PaymentMethodElectronic check        0.29229    0.09442   3.096  0.00196 **
## PaymentMethodMailed check           -0.15490    0.11682  -1.326  0.18485
## InternetServiceDSL                  -0.90128    0.10408  -8.659  < 2e-16 ***
## InternetServiceFiber optic           0.00000    0.00000     NaN      NaN
## OnlineSecurityNo internet service    0.00000    0.00000     NaN      NaN
## OnlineSecurityYes                   -0.38055    0.08492  -4.481 7.43e-06 ***
## TechSupportNo internet service      -1.69952    0.16139 -10.531  < 2e-16 ***
## TechSupportYes                      -0.27964    0.08649  -3.233  0.00122 **
## PaperlessBillingYes                  0.39228    0.07520   5.217 1.82e-07 ***
## MultipleLinesNo phone service        0.66746    0.13489   4.948 7.49e-07 ***
## MultipleLinesYes                     0.34962    0.08175   4.277 1.90e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                            edf Ref.df Chi.sq p-value
## s(tenure)                7.353      9  102.8  <2e-16 ***
## s(TotalCharges_centered) 6.603      9   20.5  0.0014 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Rank: 32/34
## R-sq.(adj) =  0.317   Deviance explained = 29.4%
## UBRE = -0.1749  Scale est. = 1         n = 7043
```

10.3.6 Support Vector Machine

Churn Probability Prediction:

```
## [1] 0
```

```
##
## Confusion Matrix for Initial SVM Model:
```

```
##        predictions
## targets   1   2
##       1 948 196
##       2  86 177
```

```
##
## Loading Tuned SVM Model...
##
## Confusion Matrix for Tuned SVM Model:
##        predictions
## targets   1   2
##       1 961 213
##       2  73 160
##
## Best Parameters Chosen During Tuning:
##   cost gamma
## 3   10  0.01
##
## Sample of Predicted Probabilities:
##          No        Yes
## 1 0.6140795 0.38592054
## 2 0.5305322 0.46946784
## 3 0.9187952 0.08120477
## 4 0.8916917 0.10830827
## 5 0.8147190 0.18528101
## 6 0.8683072 0.13169280
```