Telecom Customer Churn Prediction Analysis

Kenny Trinh Sabin Pun Sarp Koc

2025-01-10

Contents

1 A Multi-Model Machine Learning Approach **1**
1.1 Executive Summary . . . 1
1.2 Business Context: The Story of "TeleConnect" . . . 1
1.3 Dataset Overview . . . 1

2 Exploratory Data Analysis **3**

3 Methodology and Model Selection **5**

4 Linear Model (LM) **5**
4.1 Visualizing the main effects . . . 7
4.2 Enhancing the models with interactions: . . . 8
4.3 Model Summary . . . 9
4.4 Model Comparison . . . 9
4.5 Business Value and Insights for Teleconnet . . . 9

5 GLM-Poisson Model **10**
5.1 Fit Poisson regression . . . 10
5.2 Key Findings . . . 11
5.3 Evaluating the model: . . . 12
5.4 Visualization. . . . 12

6 GLM-Binomial (Logistic Regression) & GAM **17**
6.1 Context and Objective . . . 17
6.2 Data Preparation and Visualization . . . 17
6.3 Plot: Tenure vs. TotalCharges by Churn . . . 17
6.4 Plot: Tenure by Contract Type . . . 18
6.5 Plot: Churn by Payment Method . . . 19
6.6 Plot: Churn Proportion by Internet Service . . . 19
6.7 GLM . . . 20
6.8 Did the model overfit? . . . 21
6.9 GAM with non-linear patterns . . . 22
6.10 ROC Comparison . . . 23
6.11 Conclusion . . . 24

7 Neural Network (NN) **24**
7.1 Context and Objective . . . 24
7.2 Data Preparation for Neural Network . . . 24
7.3 Training the Neural Network . . . 24
7.4 Evaluating the Model . . . 25
7.5 Threshold Adjustment . . . 26

7.6 Final NN Model Performance (Resampling with Weight Ratio) . . . 27
7.7 Neural Network with RSNNS . . . 28
7.8 Neural Network with RSNNS: Trying Alternative Activation Functions . . . 28
7.9 Neural Network with Regularization and Adjusted Architecture . . . 29
7.10 Neural Network Experimentation with Different Architectures . . . 29
7.11 Threshold Analysis and Visualization . . . 30
7.12 Evaluation Metrics . . . 30
7.13 Business Insights from Neural Network Results . . . 31
7.14 Strategic Recommendations . . . 31

8 Support Vector Machine (SVM) 32
8.1 Context and Objective . . . 32
8.2 Model Performance Metrics . . . 32
8.3 Churn Probability Prediction . . . 32
8.4 Evaluate additional metrics for the tuned model . . . 33
8.5 Churn Probability Prediction . . . 34
8.6 Customer Segmentation Based on Churn Risk . . . 34
8.7 Key Feature Insights . . . 36
8.8 Customer Retention Strategy . . . 36
8.9 Key Findings for TeleConnect Company . . . 36

9 Conclusion 38
9.1 Role of Generative AI . . . 38
9.2 Final Thoughts . . . 38

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:

```
## 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"
```

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

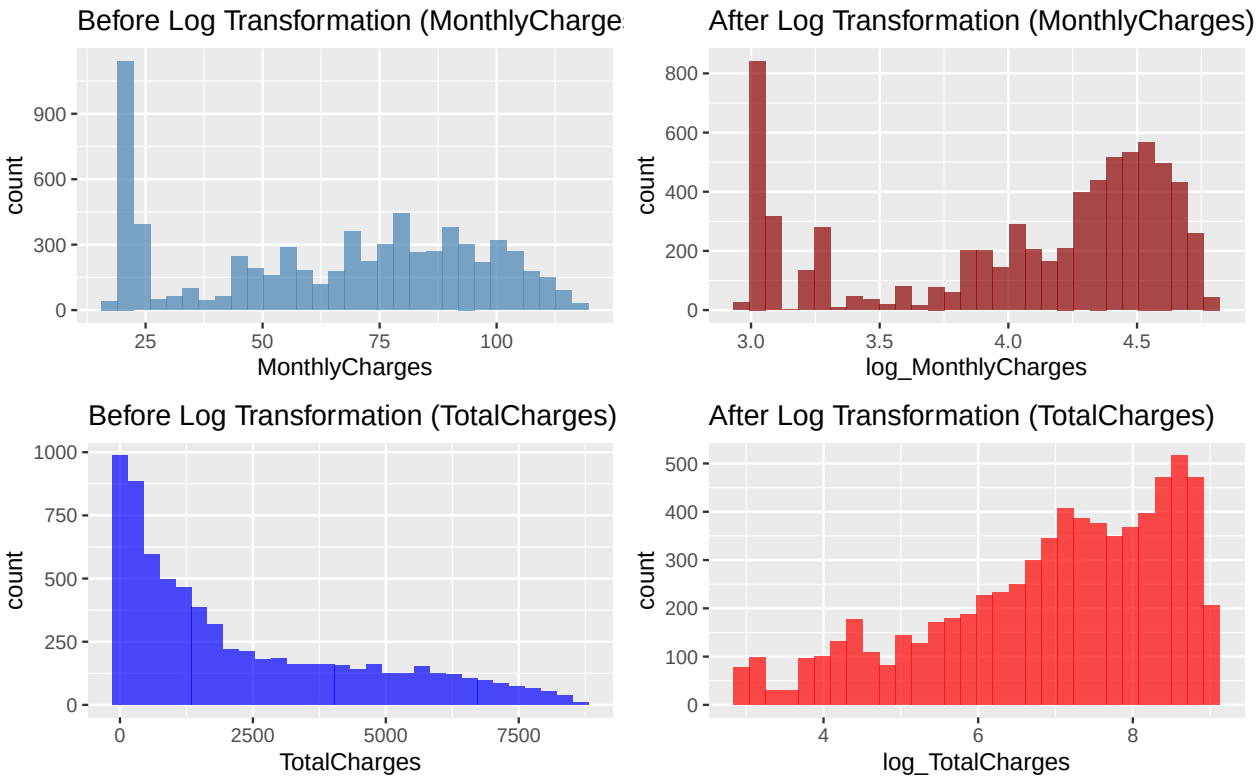
2 Exploratory Data Analysis

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)

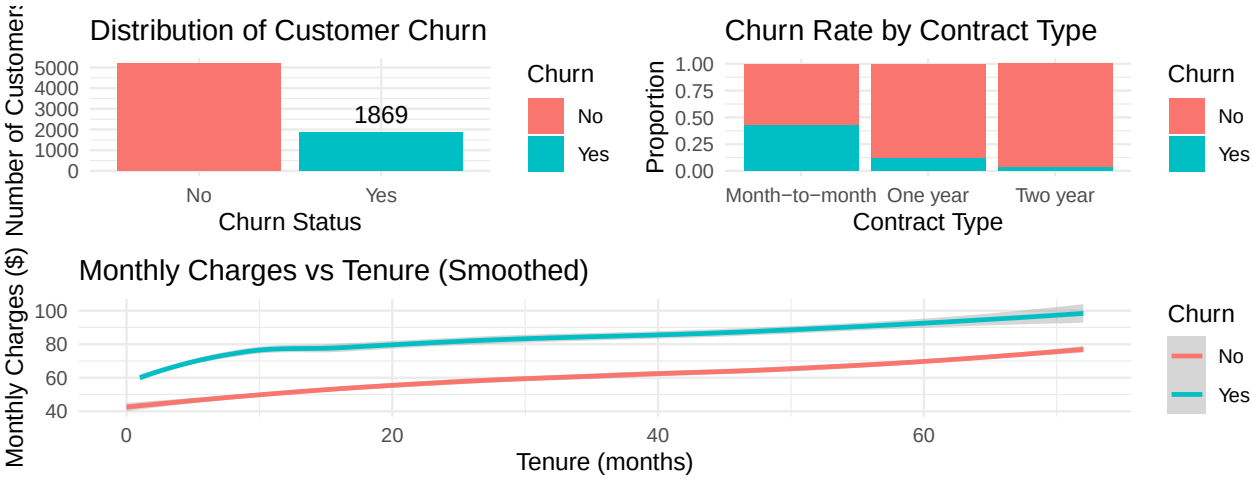
Log-Transformation of Charges (MonthlyCharges & TotalCharges)



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

Key Findings from EDA

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



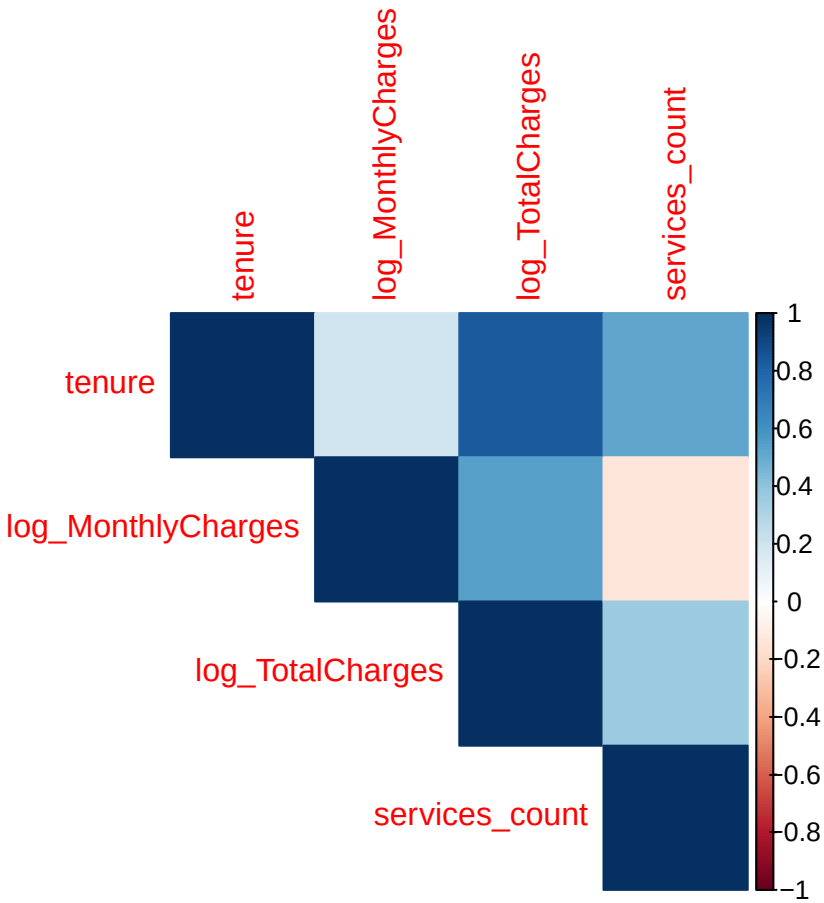
1. Churn Distribution

- 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 ≈ 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.

3 Methodology and 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.

4 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 analyze the factors influencing **MonthlyCharges**, the amount customers pay per month. Understanding what drives these charges provides crucial insights into how TeleConnect can design competitive pricing strategies and service bundles.

The response variable for this linear model is:

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

The predictors used are:

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

Linear Model to predict monthly charges:

```
##
## 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
```

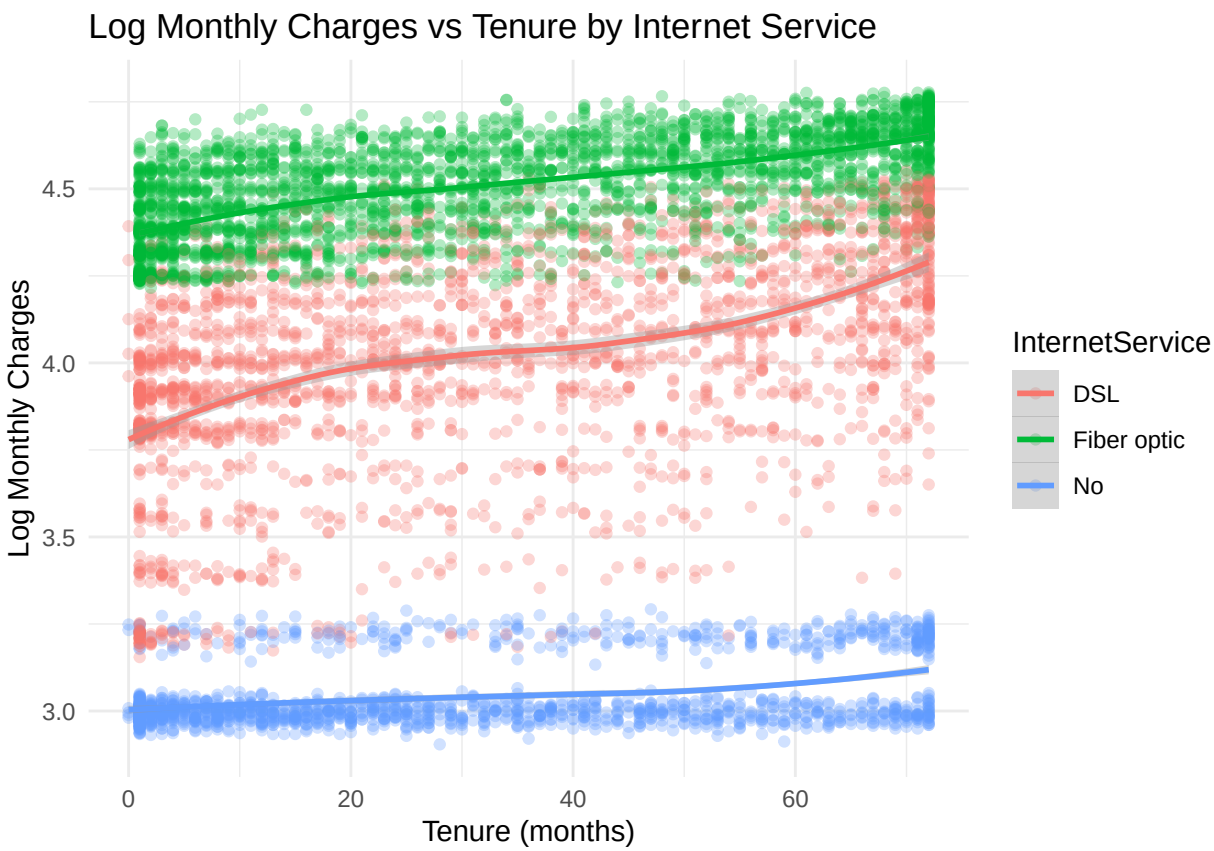
Coefficient Interpretation:

- InternetService has the largest impact:
  - Fiber optic: +0.347 (≈ 41.5% increase in charges)
  - No service: -0.863 (≈ 57.8% decrease in charges)
- Contract Two year: +0.0882(≈ 9.2%) significant effect
- PhoneService: +0.439 (≈ 55.1% increase)
- Streaming services both have similar effects (about 15%)
- Extremely low p-value (< 2.2e-16), indicating the model is significant

4.1 Visualizing the main effects

Monthly Charges against Tenure:

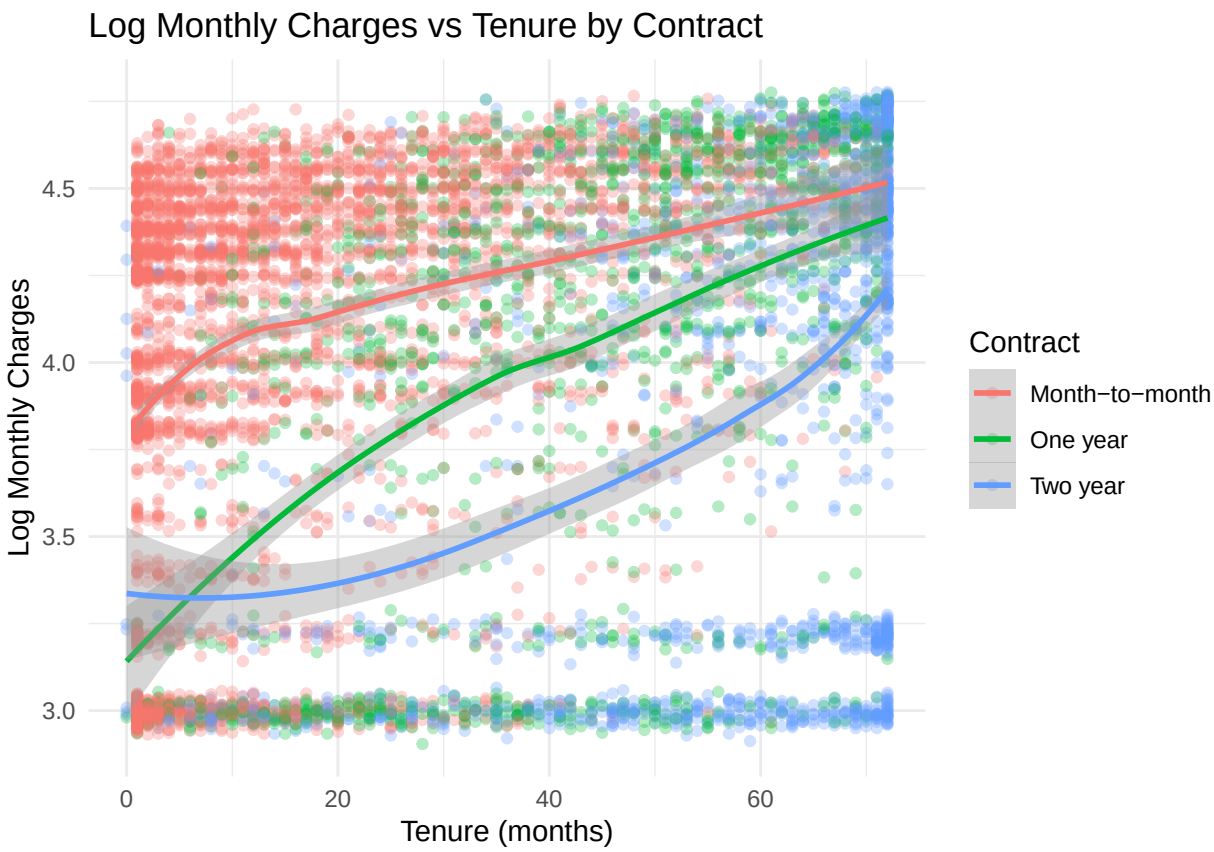
```
## `geom_smooth()` using formula = 'y ~ x'
```



- Fiber optic shows a steeper increase with tenure
- DSL shows a moderate increase
- No internet service stays flat

Monthly Charges vs Tenure:

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



4.2 Enhancing the models with interactions:

```
##
## Call:
## lm(formula = log(MonthlyCharges) ~ tenure * InternetService +
##     tenure * Contract + StreamingTV + StreamingMovies + OnlineSecurity +
##     OnlineBackup + DeviceProtection + TechSupport + PhoneService,
##     data = d.cleaned_telecom_customer_churn)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.289403 -0.038118 -0.004837  0.034807  0.268165
##
## Coefficients: (6 not defined because of singularities)
##                                      Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                         3.402e+00  3.231e-03 1052.897  < 2e-16 ***
## tenure                              1.936e-03  7.964e-05   24.308  < 2e-16 ***
## InternetServiceFiber optic          4.167e-01  3.155e-03  132.097  < 2e-16 ***
## InternetServiceNo                  -8.484e-01  3.756e-03 -225.869  < 2e-16 ***
## ContractOne year                    1.068e-02  4.512e-03    2.366    0.018 *
## ContractTwo year                    3.759e-02  5.599e-03    6.713 2.05e-11 ***
## StreamingTVNo internet service             NA         NA       NA       NA
## StreamingTVYes                      1.421e-01  2.015e-03   70.531  < 2e-16 ***
## StreamingMoviesNo internet service         NA         NA       NA       NA
## StreamingMoviesYes                  1.435e-01  2.019e-03   71.070  < 2e-16 ***
## OnlineSecurityNo internet service          NA         NA       NA       NA
## OnlineSecurityYes                   7.168e-02  2.046e-03   35.034  < 2e-16 ***
```

```
## OnlineBackupNo internet service             NA         NA       NA       NA
## OnlineBackupYes                       7.098e-02  1.935e-03   36.679  < 2e-16 ***
## DeviceProtectionNo internet service         NA         NA       NA       NA
## DeviceProtectionYes                   7.153e-02  1.995e-03   35.862  < 2e-16 ***
## TechSupportNo internet service              NA         NA       NA       NA
## TechSupportYes                        6.835e-02  2.074e-03   32.959  < 2e-16 ***
## PhoneServiceYes                       4.389e-01  2.935e-03  149.527  < 2e-16 ***
## tenure:InternetServiceFiber optic    -2.138e-03  7.215e-05  -29.634  < 2e-16 ***
## tenure:InternetServiceNo             -1.011e-04  9.297e-05   -1.087    0.277
## tenure:ContractOne year              -5.241e-04  1.071e-04   -4.893 1.02e-06 ***
## tenure:ContractTwo year              -9.581e-04  1.071e-04   -8.949  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06458 on 7026 degrees of freedom
## Multiple R-squared:  0.9882, Adjusted R-squared:  0.9882
## F-statistic: 3.685e+04 on 16 and 7026 DF,  p-value: < 2.2e-16
```

4.3 Model Summary

Our model explains 98.82% of the variability in monthly charges, showing strong predictive power. The key factors affecting monthly charges are:

- **Internet Service Type**:
  - Fiber optic: +41.7%
  - No internet: -57.2%
- **Contract Type**:
  - One-year: +1.1%
  - Two-year: +3.8%
- **Additional Services**:
  - Streaming (TV or Movies): +15.3% each
  - Phone Service: +15.1%
  - Security/Support services: +7.5% each

4.4 Model Comparison

Initial Model: R-squared = 0.9865 Enhanced Model: R-squared = 0.9882 Improvement: 0.0017 (only about 0.17%)

This small improvement suggests that:

1. The simpler model (without interactions) might be preferable due to:
   - Easier interpretation
   - Greater parsimony (fewer parameters)
   - Nearly identical predictive power (98.65% vs 98.82%)

4.5 Business Value and Insights for Teleconnet

1. **Internet Service is the Primary Revenue Driver**:
   - **Value**: Internet service type (e.g., Fiber optic) has the largest impact on monthly charges, making it the most critical factor for revenue growth.
2. **Contract Type Has Limited Revenue Impact**:
   - **Value**: Customers pay similar monthly charges regardless of contract type.
3. **Add-On Services Boost Revenue Incrementally**:
   - **Value**: Services like Streaming, Security, and Tech Support add 6-14% to monthly charges, providing consistent opportunities for revenue growth.

5 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 with the company. By identifying the key drivers of tenure, this model helps TeleConnect develop targeted retention strategies.

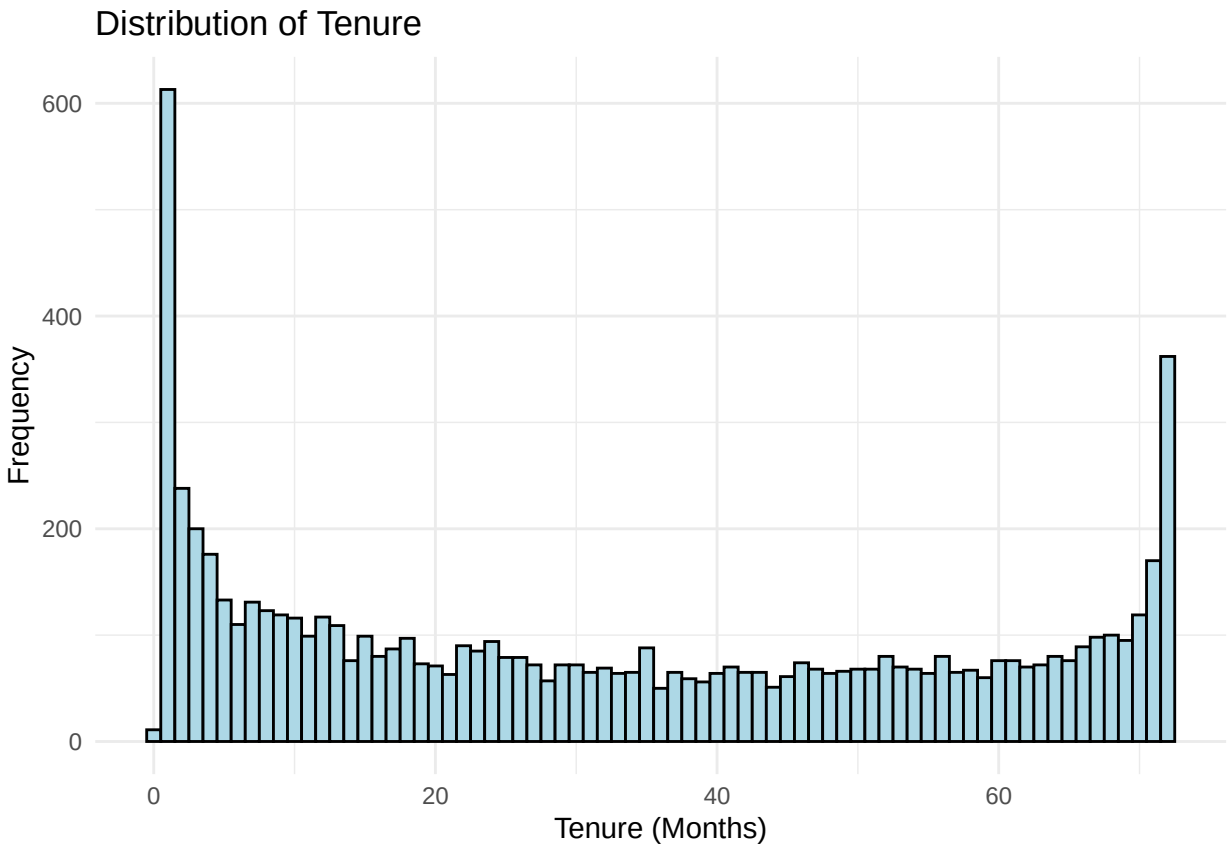
The response variable for Poisson regression must be **count data** (non-negative integers).

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.

5.1 Fit Poisson regression

```
##
```

```
## Call:
## glm(formula = tenure ~ InternetService + Contract + PaymentMethod +
##     StreamingTV + StreamingMovies + SeniorCitizen + Partner +
##     MonthlyCharges + TotalCharges, family = poisson(link = "log"),
##     data = d.cleaned_telecom_customer_churn)
##
## Coefficients: (2 not defined because of singularities)
##                                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)                           3.428e+00  1.532e-02 223.750  < 2e-16 ***
## InternetServiceFiber optic            2.282e-02  9.674e-03   2.359 0.018342 *
## InternetServiceNo                    -4.822e-02  1.074e-02  -4.489 7.16e-06 ***
## ContractOne year                      2.735e-01  6.389e-03  42.807  < 2e-16 ***
## ContractTwo year                      2.760e-01  7.176e-03  38.458  < 2e-16 ***
## PaymentMethodCredit card (automatic) -2.054e-02  5.481e-03  -3.748 0.000178 ***
## PaymentMethodElectronic check        -2.909e-02  6.045e-03  -4.813 1.49e-06 ***
## PaymentMethodMailed check            -2.457e-01  7.017e-03 -35.011  < 2e-16 ***
## StreamingTVNo internet service               NA         NA      NA       NA
## StreamingTVYes                       -3.331e-03  6.321e-03  -0.527 0.598205
## StreamingMoviesNo internet service           NA         NA      NA       NA
## StreamingMoviesYes                   -8.651e-04  6.337e-03  -0.137 0.891417
## SeniorCitizen                         5.749e-02  5.858e-03   9.814  < 2e-16 ***
## PartnerYes                            9.300e-02  4.640e-03  20.043  < 2e-16 ***
## MonthlyCharges                       -1.650e-02  3.062e-04 -53.889  < 2e-16 ***
## TotalCharges                          3.359e-04  1.990e-06 168.794  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 151098  on 7042  degrees of freedom
## Residual deviance:  43105  on 7029  degrees of freedom
## AIC: 76757
##
## Number of Fisher Scoring iterations: 5
```

5.2 Key Findings

1. **Contract Type**:
   - Customers with a **One-year contract** are expected to have a tenure **31.5% longer** than those with a month-to-month contract.
   - Customers with a **Two-year contract** have a **31.7% longer tenure**, making contract type a key predictor of retention.
2. **Internet Service**:
   - **Fiber optic users** have a slightly **higher tenure rate (+2.9%)** compared to DSL users.
   - Customers with **No internet service** show a significant **decrease in tenure (-4.7%)**, suggesting bundled services may improve retention.
3. **Payment Method**:
   - Customers paying via **Mailed check** have significantly **shorter tenures (-27%)** compared to electronic payment users.
   - **Credit card automatic payments** also show slightly shorter tenure (-20%), indicating that payment methods might influence retention.
4. **Streaming Services**:
   - **Streaming services (TV and Movies)** do not appear to significantly impact tenure, suggesting these features may not drive long-term customer retention.

5. **Senior Citizen**:
   - **Senior citizens** have a **7.4% longer tenure** than non-senior citizens, showing their loyalty to services.
6. **Monthly and Total Charges**:
   - **Monthly Charges** negatively impact tenure, with higher monthly charges leading to shorter tenures.
   - **Total Charges** positively influence tenure, as expected for long-term customers.

5.3 Evaluating the model:

1. Goodness of Fit:

- **Residual Deviance**:
  - The residual deviance is **43,105** on **7,029 degrees of freedom**.
  - This ratio (Degrees of Freedom/ Residual Deviance) is approximately **6.13**, which is significantly greater than 1. This indicates **overdispersion** (variance > mean), suggesting the Poisson model may not be a perfect fit.

2. Significant Predictors:

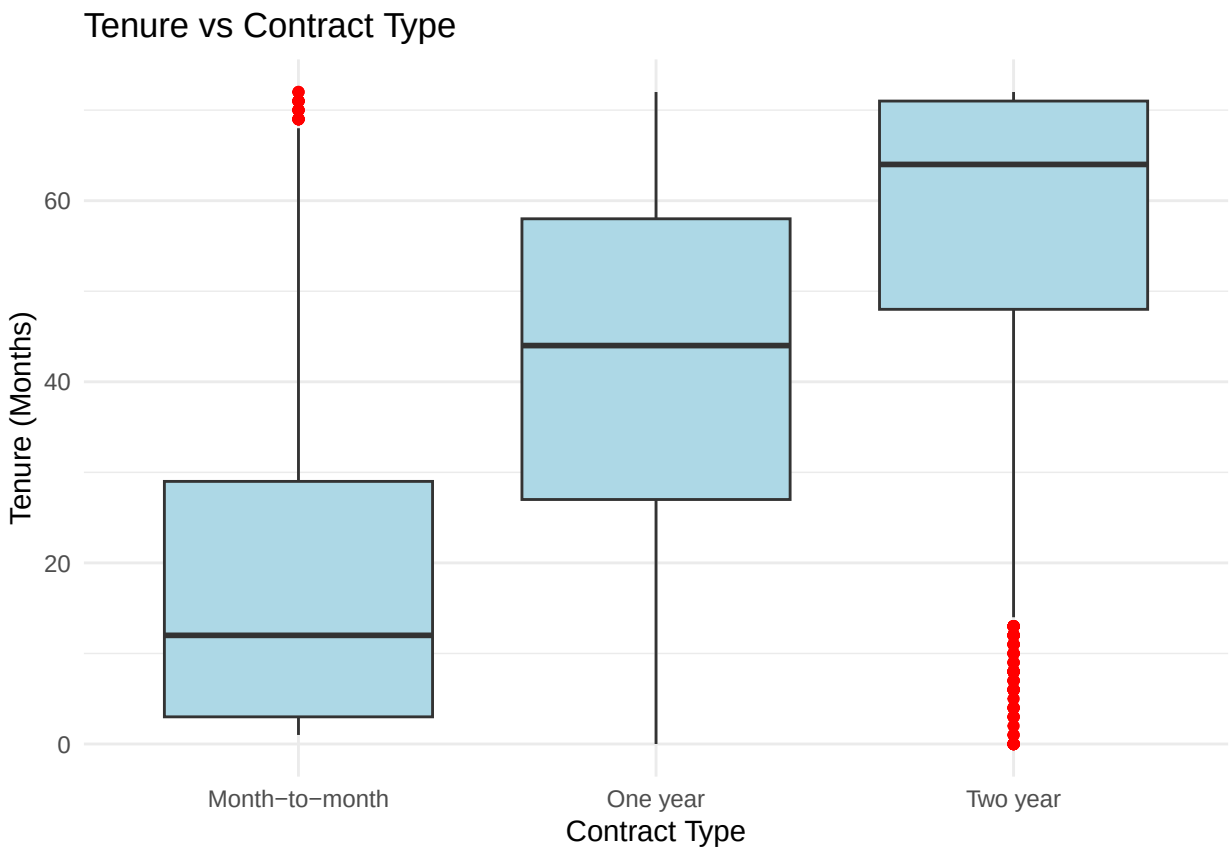
- Predictors with extremely low p-values (p<0.001p < 0.001p<0.001) include:
  - **Contract type** (One-year and Two-year contracts)
  - **Internet Service** (Fiber optic and No Internet Service)
  - **Payment Method** (Mailed check and Credit card)
  - **Senior Citizen**, **Partner**, **Monthly Charges**, and **Total Charges**
- Non-significant predictors:
  - **StreamingTV** and **StreamingMovies** were not significant, indicating they may not have a strong impact on tenure.

3. Overdispersion Check:

- Overdispersion occurs when the variance of the response variable is greater than the mean.
  - **Mean of Tenure**: ~32 months
  - **Variance of Tenure**: Much larger than the mean, indicating overdispersion.
  - This can be addressed by using:
    - **Quasi-Poisson Regression**: Allows for overdispersion by adjusting the standard errors.

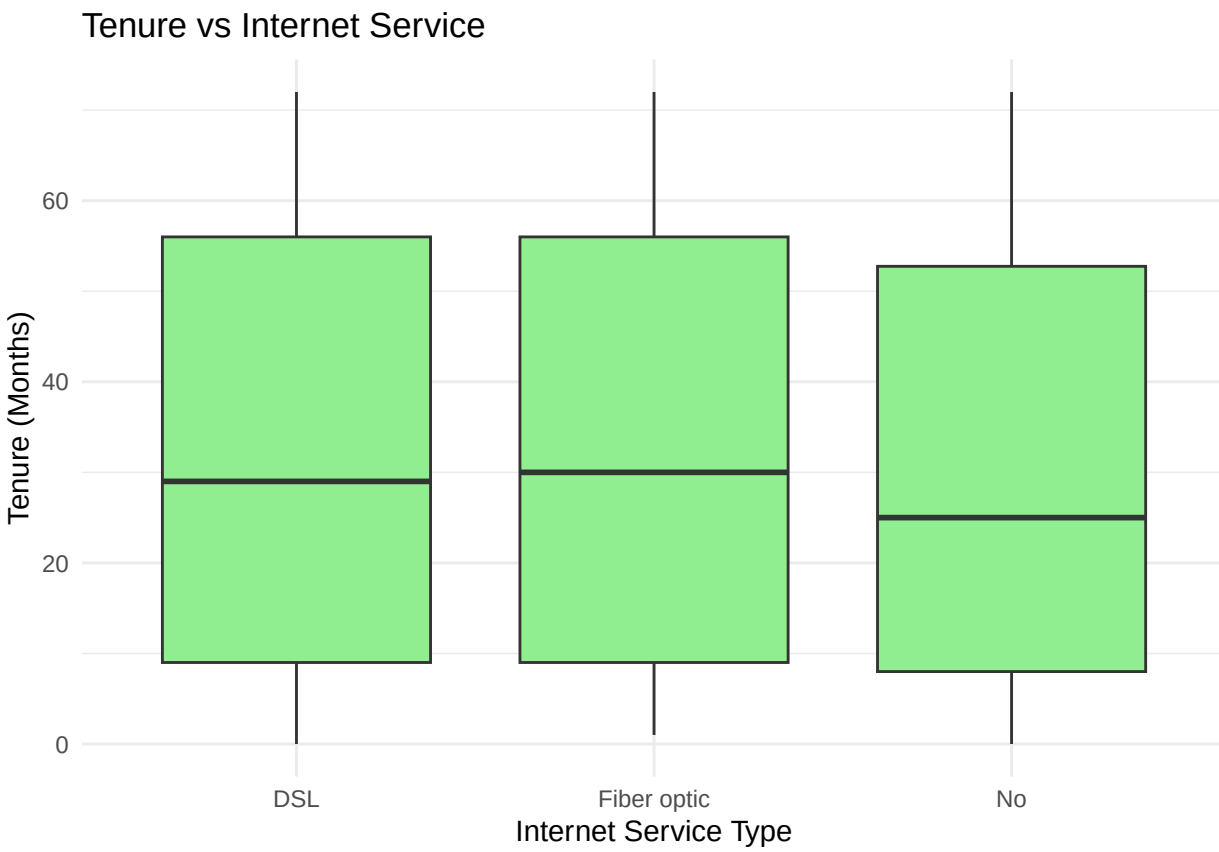
5.4 Visualization.

Boxplot for Contract



- **Month-to-month contracts** show the shortest median tenure, with many outliers indicating customers leaving quickly.
- **One-year** and **Two-year contracts** significantly increase tenure, with Two-year contracts having the longest median tenure.

Boxplot for Internet Service



- Customers with **No internet service** tend to have shorter tenures.
- **DSL** and **Fiber optic** services show similar median tenures, indicating these internet options retain customers for similar durations

Model with Quasi- Poisson Model with interaction Terms:

```
##
## Call:
## glm(formula = tenure ~ InternetService * Contract + PaymentMethod *
##     Contract + StreamingTV + StreamingMovies + SeniorCitizen +
##     Partner + MonthlyCharges + TotalCharges, family = quasipoisson(link = "log"),
##     data = d.cleaned_telecom_customer_churn)
##
## Coefficients: (2 not defined because of singularities)
##                                                      Estimate Std. Error
## (Intercept)                                         3.444e+00  3.376e-02
## InternetServiceFiber optic                          3.313e-01  2.634e-02
## InternetServiceNo                                  -6.899e-01  3.615e-02
## ContractOne year                                    3.248e-01  2.801e-02
## ContractTwo year                                    1.717e-01  2.672e-02
## PaymentMethodCredit card (automatic)               -2.960e-02  2.407e-02
## PaymentMethodElectronic check                      -1.489e-01  2.004e-02
## PaymentMethodMailed check                          -3.915e-01  2.839e-02
## StreamingTVNo internet service                             NA         NA
## StreamingTVYes                                      2.055e-02  1.259e-02
## StreamingMoviesNo internet service                         NA         NA
## StreamingMoviesYes                                  1.372e-02  1.260e-02
## SeniorCitizen                                       2.545e-03  1.181e-02
```

```
## PartnerYes                                          4.159e-02  9.248e-03
## MonthlyCharges                                     -1.884e-02  6.275e-04
## TotalCharges                                        3.929e-04  4.083e-06
## InternetServiceFiber optic:ContractOne year        -5.121e-01  2.573e-02
## InternetServiceNo:ContractOne year                  4.738e-01  4.128e-02
## InternetServiceFiber optic:ContractTwo year        -6.072e-01  2.439e-02
## InternetServiceNo:ContractTwo year                  9.393e-01  3.812e-02
## ContractOne year:PaymentMethodCredit card (automatic) -3.664e-04  3.198e-02
## ContractTwo year:PaymentMethodCredit card (automatic)  1.314e-02  2.845e-02
## ContractOne year:PaymentMethodElectronic check      1.392e-01  2.963e-02
## ContractTwo year:PaymentMethodElectronic check      1.548e-01  3.028e-02
## ContractOne year:PaymentMethodMailed check          2.820e-01  3.844e-02
## ContractTwo year:PaymentMethodMailed check          2.871e-01  3.450e-02
##                                                    t value Pr(>|t|)
## (Intercept)                                        102.005  < 2e-16 ***
## InternetServiceFiber optic                          12.579  < 2e-16 ***
## InternetServiceNo                                  -19.083  < 2e-16 ***
## ContractOne year                                    11.596  < 2e-16 ***
## ContractTwo year                                     6.425 1.40e-10 ***
## PaymentMethodCredit card (automatic)                -1.230    0.219
## PaymentMethodElectronic check                       -7.430 1.22e-13 ***
## PaymentMethodMailed check                          -13.792  < 2e-16 ***
## StreamingTVNo internet service                          NA       NA
## StreamingTVYes                                       1.632    0.103
## StreamingMoviesNo internet service                      NA       NA
## StreamingMoviesYes                                   1.088    0.277
## SeniorCitizen                                        0.216    0.829
## PartnerYes                                           4.497 7.01e-06 ***
## MonthlyCharges                                     -30.018  < 2e-16 ***
## TotalCharges                                        96.209  < 2e-16 ***
## InternetServiceFiber optic:ContractOne year        -19.903  < 2e-16 ***
## InternetServiceNo:ContractOne year                  11.477  < 2e-16 ***
## InternetServiceFiber optic:ContractTwo year        -24.898  < 2e-16 ***
## InternetServiceNo:ContractTwo year                  24.638  < 2e-16 ***
## ContractOne year:PaymentMethodCredit card (automatic)  -0.011    0.991
## ContractTwo year:PaymentMethodCredit card (automatic)   0.462    0.644
## ContractOne year:PaymentMethodElectronic check          4.698 2.68e-06 ***
## ContractTwo year:PaymentMethodElectronic check          5.113 3.26e-07 ***
## ContractOne year:PaymentMethodMailed check              7.337 2.43e-13 ***
## ContractTwo year:PaymentMethodMailed check              8.323  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 3.98992)
##
##     Null deviance: 151098  on 7042  degrees of freedom
## Residual deviance:  31910  on 7019  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

Model Performance

1. Dispersion Parameter:
   - The dispersion parameter for the quasi-Poisson model is **3.99**, confirming overdispersion (variance

> mean).
    - This indicates that the quasi-Poisson model is better suited for this dataset compared to the Poisson model.
2. Residual Deviance:
    - **Residual Deviance**: 31,910 (on 7,019 degrees of freedom), consistent with the Poisson model.
    - This shows that the fit remains the same, but the quasi-Poisson model adjusts the standard errors for overdispersion, leading to more reliable p-values.

Insights

1. Interaction Terms are Significant:
    - Significant interaction effects for **Internet Service x Contract** and **Payment Method x Contract** highlight the importance of combining these factors to predict tenure effectively.
2. Adjusted Standard Errors:
    - Quasi-Poisson correction adjusts the standard errors, resulting in more reliable significance levels compared to the Poisson model.
3. Non-Significant Predictors:
    - **Streaming Services** and **Credit card (automatic)** are not significant, suggesting they may not strongly impact tenure.

5.4.1 Business Insights and Recommendations

1. **Encourage Long-Term Contracts**:
    - Promote One-year and Two-year contracts, particularly for customers with **No Internet Service**, to boost retention.
2. **Optimize Internet Service Offerings**:
    - Address potential dissatisfaction among **Fiber optic** customers by enhancing service quality or offering competitive pricing.
    - Bundling internet services with contracts can significantly improve tenure.
3. **Reevaluate Payment Methods**:
    - Reduce reliance on **Mailed checks**, as they correlate with shorter tenure.
    - Incentivize automatic payments to retain customers longer.
4. **Pricing Sensitivity**:
    - Monitor **Monthly Charges** closely and consider pricing strategies or discounts to improve retention among price-sensitive customers.

Customer **tenure** is strongly influenced by contract type, internet service, and payment methods, with significant interaction effects between these factors. Long-term contracts and bundled internet services emerge as the most effective strategies to enhance customer retention.

6 GLM-Binomial (Logistic Regression) & GAM

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

6.1 Context and Objective

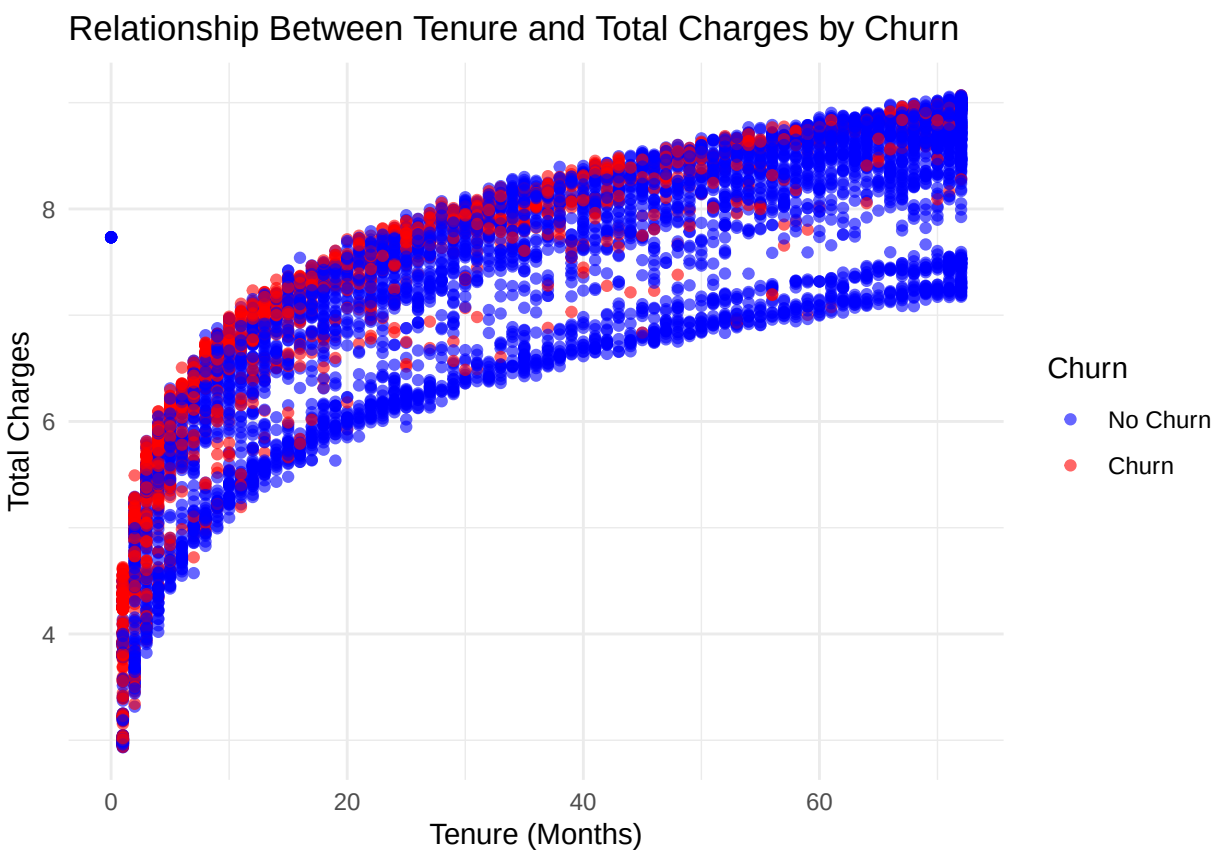
The GLM-Binomial (Logistic Regression) and Generalized Additive Model (GAM) are used to predict customer churn and explore its underlying drivers. These models provide a foundation for identifying at-risk customers and evaluating key variables such as tenure, contract type, and payment methods. The analysis 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.

6.2 Data Preparation and Visualization

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

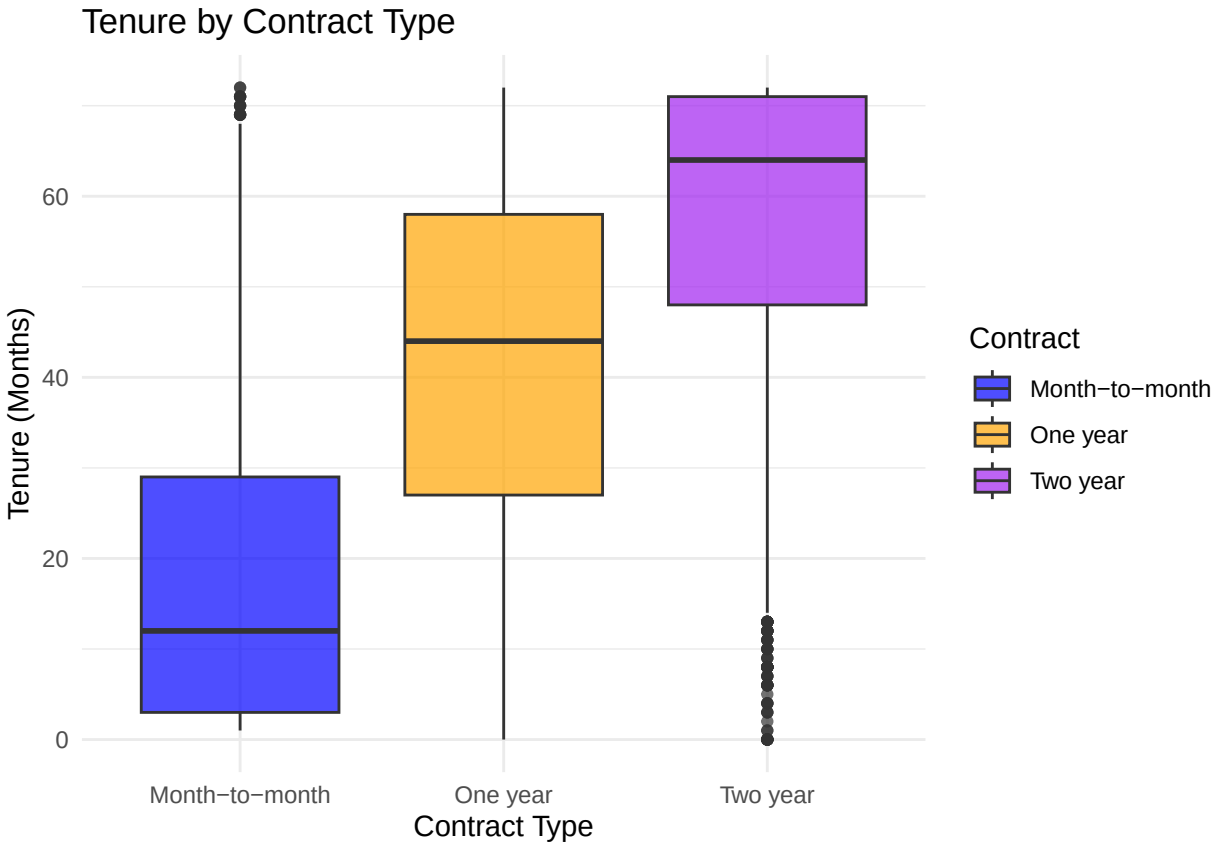
Let's plot some results to understand the data better.

6.3 Plot: Tenure vs. TotalCharges by Churn



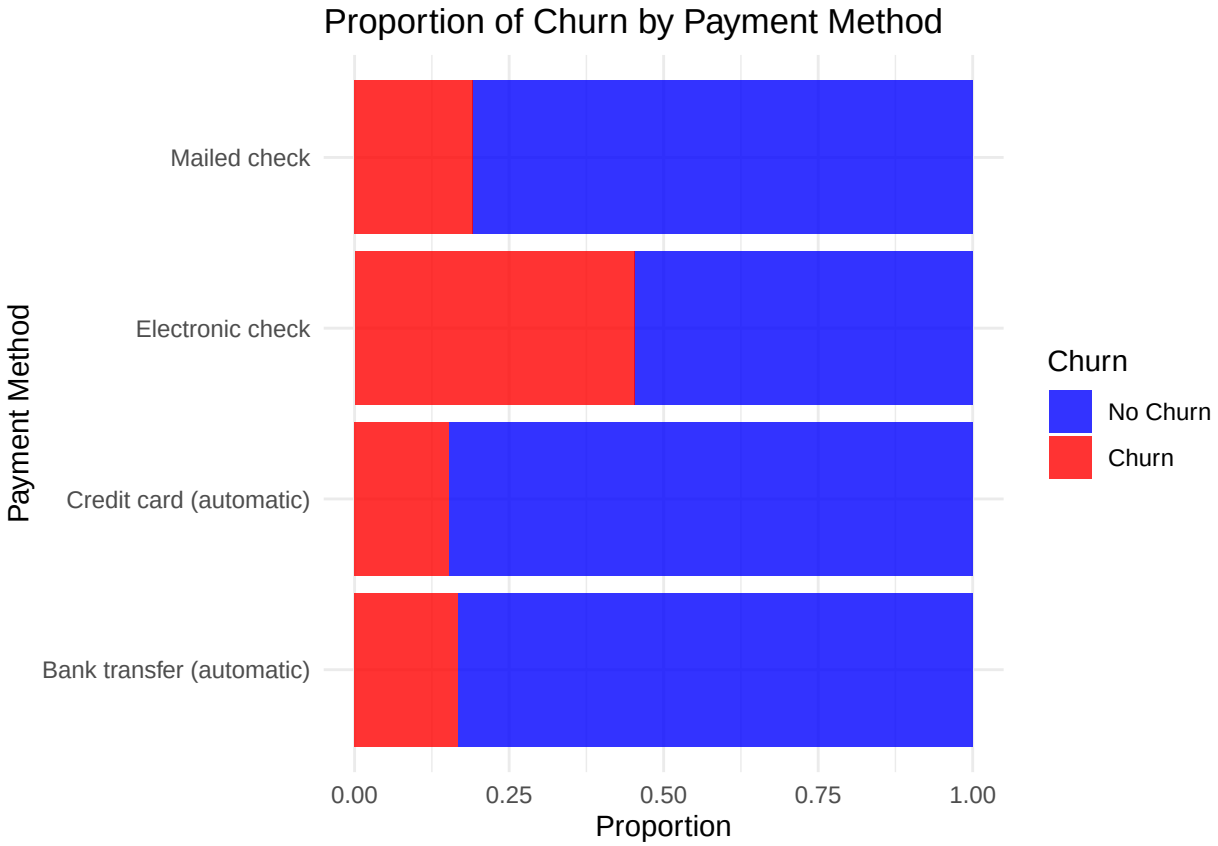
Higher tenure and total charges are associated with lower churn, reinforcing loyalty among long-term customers with higher spending.

6.4 Plot: Tenure by Contract Type



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

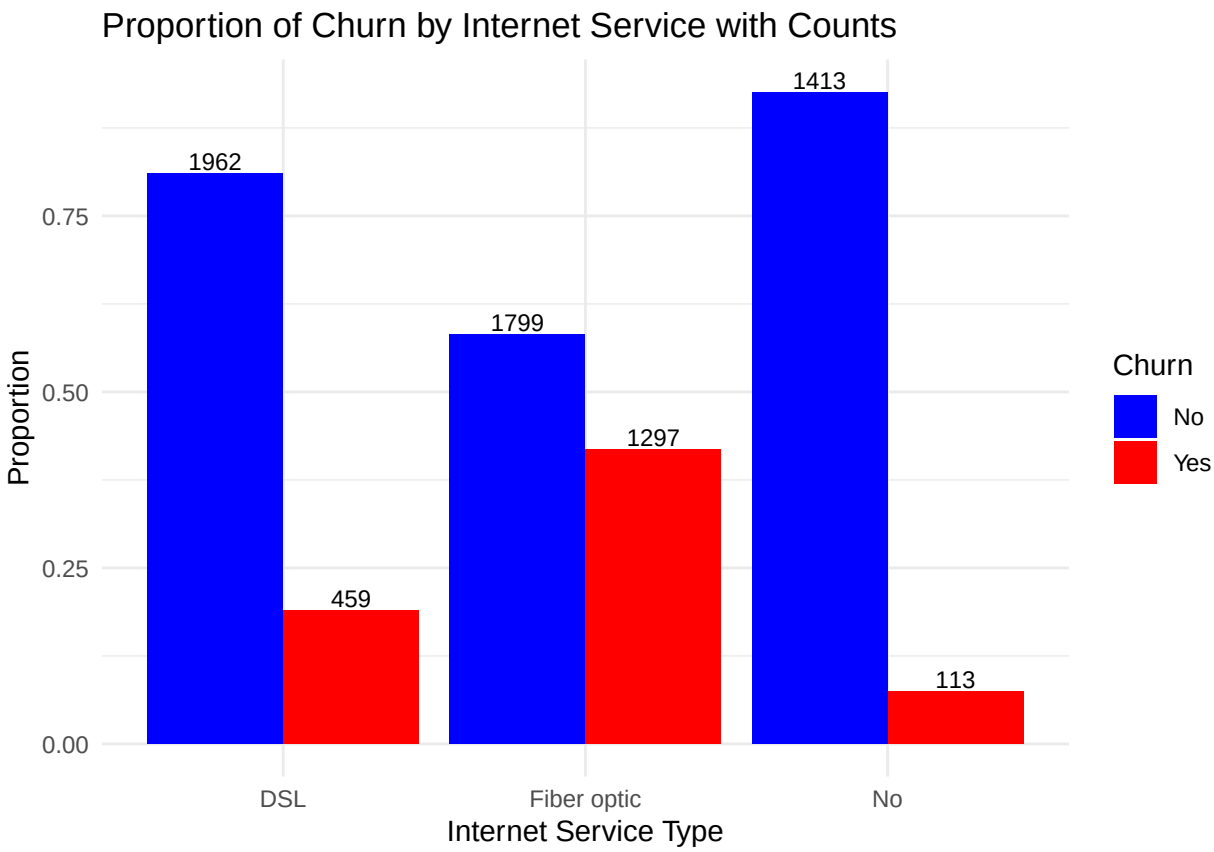
6.5 Plot: Churn by Payment Method



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

6.6 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.

6.7 GLM

```
## Reference level for 'Contract': Month-to-month

## 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)                           0.399390   0.247458   1.614 0.106534
## SeniorCitizen                         0.253345   0.083059   3.050 0.002287 **
## ContractOne year                     -0.599254   0.107003  -5.600 2.14e-08 ***
## ContractTwo year                     -1.506555   0.177593  -8.483  < 2e-16 ***
## PaymentMethodCredit card (automatic) -0.091846   0.113045  -0.812 0.416518
## PaymentMethodElectronic check         0.318315   0.094082   3.383 0.000716 ***
## PaymentMethodMailed check            -0.140741   0.116109  -1.212 0.225455
## tenure                               -0.005313   0.003504  -1.516 0.129403
```

```
## TotalCharges                        -0.486935   0.050138  -9.712  < 2e-16 ***
## InternetServiceDSL                   1.352561   0.145821   9.276  < 2e-16 ***
## InternetServiceFiber optic           2.588159   0.153860  16.822  < 2e-16 ***
## OnlineSecurityNo internet service          NA         NA      NA       NA
## OnlineSecurityYes                   -0.317249   0.084517  -3.754 0.000174 ***
## TechSupportNo internet service             NA         NA      NA       NA
## TechSupportYes                      -0.177401   0.085055  -2.086 0.037004 *
## PaperlessBillingYes                  0.400922   0.074954   5.349 8.85e-08 ***
## MultipleLinesNo phone service        0.396823   0.130512   3.041 0.002362 **
## MultipleLinesYes                     0.435195   0.080293   5.420 5.96e-08 ***
## ---
## 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: 5796.5  on 7027  degrees of freedom
## AIC: 5828.5
##
## Number of Fisher Scoring iterations: 6
```

So let's interpret the results:

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.

6.8 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

6.9 GAM with non-linear patterns

```
## Reference level for 'Contract': Month-to-month
```

```
## 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)                        -0.75971    0.12557  -6.050 1.45e-09 ***
## SeniorCitizen                       0.24749    0.08290   2.985  0.00283 **
## ContractOne year                   -0.60800    0.10750  -5.656 1.55e-08 ***
## ContractTwo year                   -1.43393    0.18586  -7.715 1.21e-14 ***
## PaymentMethodCredit card (automatic) -0.08378    0.11335  -0.739  0.45982
## PaymentMethodElectronic check       0.30880    0.09417   3.279  0.00104 **
## PaymentMethodMailed check          -0.15686    0.11675  -1.344  0.17911
## InternetServiceDSL                 -1.00916    0.10021 -10.070  < 2e-16 ***
## InternetServiceFiber optic          0.00000    0.00000     NaN      NaN
## OnlineSecurityNo internet service  -2.12423    0.16807 -12.639  < 2e-16 ***
## OnlineSecurityYes                  -0.36158    0.08458  -4.275 1.91e-05 ***
## TechSupportNo internet service      0.00000    0.00000     NaN      NaN
## TechSupportYes                     -0.25100    0.08548  -2.936  0.00332 **
## PaperlessBillingYes                 0.39353    0.07504   5.244 1.57e-07 ***
## MultipleLinesNo phone service       0.57035    0.13447   4.241 2.22e-05 ***
## MultipleLinesYes                    0.36540    0.08044   4.542 5.57e-06 ***
## ---
## 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)                3.592      9  48.08  <2e-16 ***
## s(TotalCharges_centered) 2.270      9  26.08  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Rank: 32/34
## R-sq.(adj) =  0.317   Deviance explained = 29.2%
## UBRE = -0.17508  Scale est. = 1         n = 7043
```

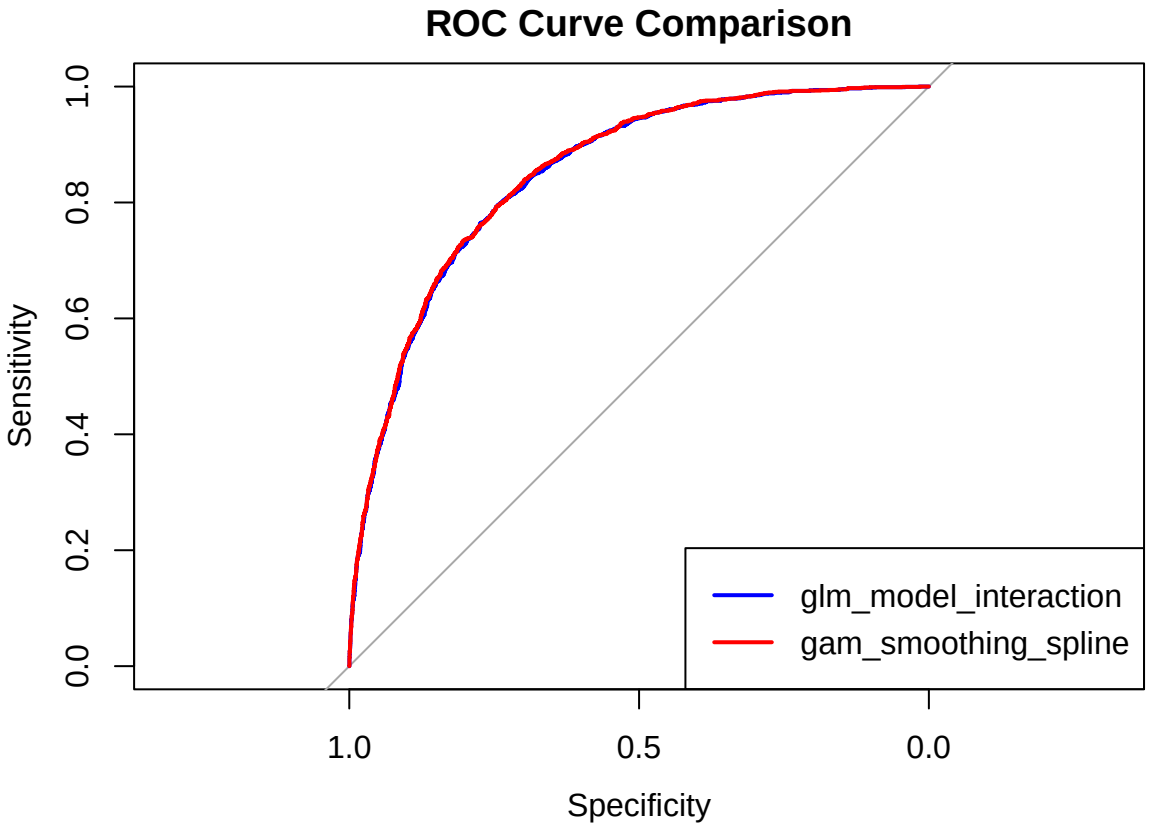
edf shows that mooth terms for tenure and TotalCharges_centered are highly significant ($p < 0.001$),

indicating non-linear effects. Relationship is non-linear but not overly complex.

6.10 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.

6.11 Conclusion

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.

7 Neural Network (NN)

Kenny Trinh took the lead on the Neural Network model.

7.1 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.

7.2 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.

7.3 Training the Neural Network

7.3.1 Exploring different Neural Network Configurations

Table 1: 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.

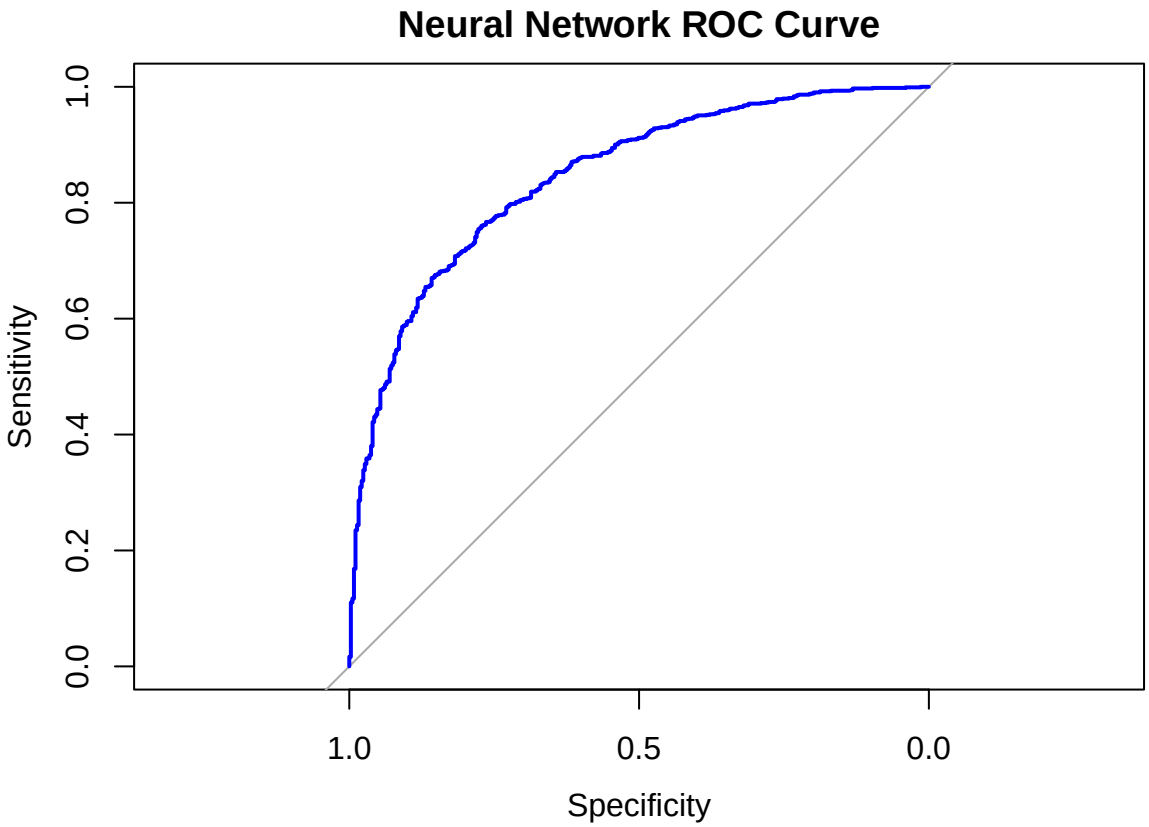
7.3.2 Training with an optimal configuration

```r
# Train the Neural Network
nn_model <- nnet(
  Churn ~ ., data = trainData,
  size = 2, decay = 0.04, maxit = 300, trace = FALSE
)

# Predict on the test set
nn_prob <- predict(nn_model, testData, type = "raw")
```

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.

7.4 Evaluating the Model

```
##        predictions
## targets   1   2
##       1 931 171
##       2 103 202
```

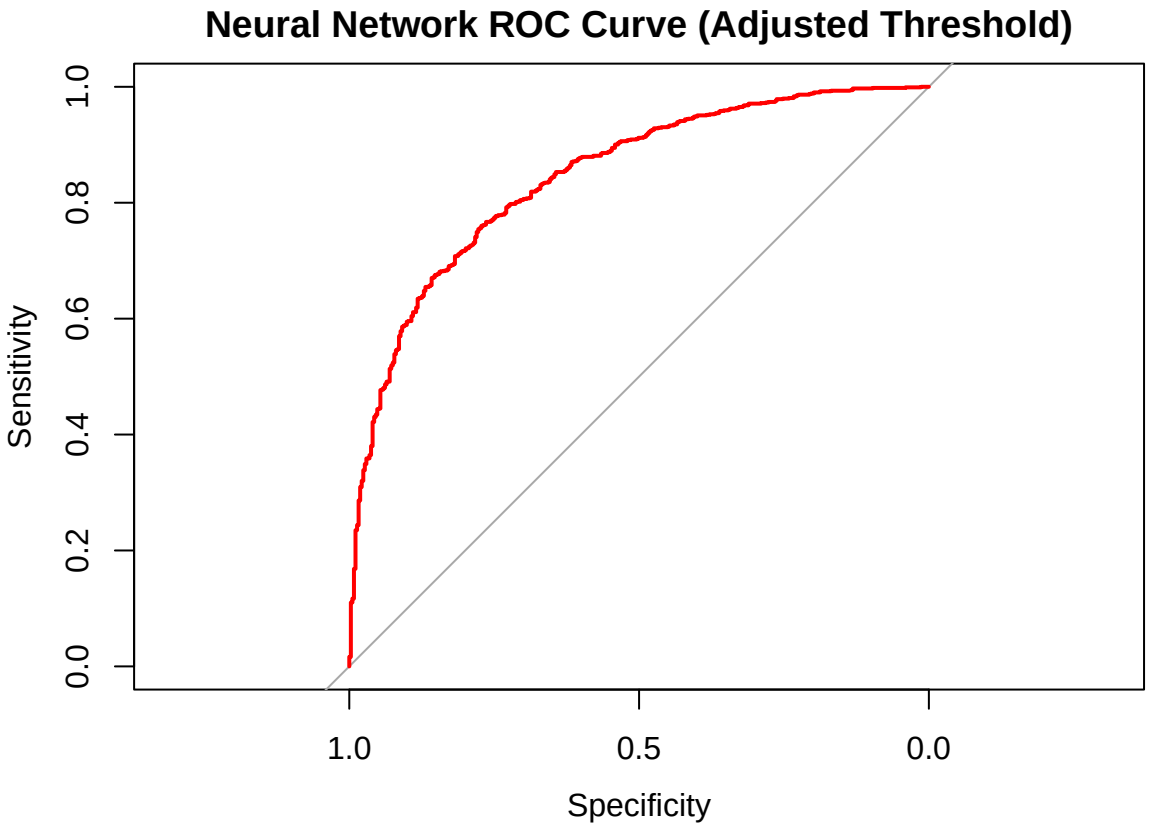



```
## Neural Network AUC: 0.8442849
```

The Neural Network's performance is evaluated using a confusion matrix and ROC curve. The Area Under the Curve (AUC) quantifies how well the model distinguishes between churners and non-churners.

7.5 Threshold Adjustment

```
##        predictions
## targets   1   2
##       1 949 192
##       2  85 181
```



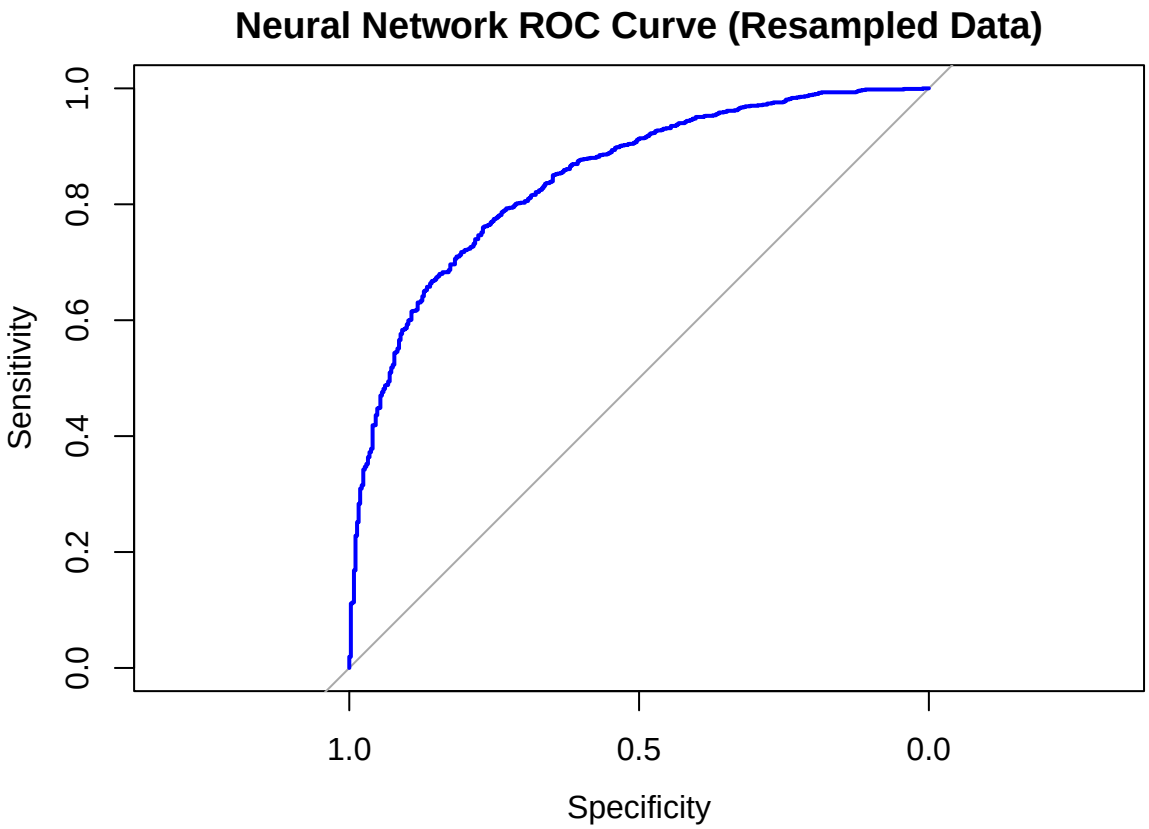

```
## Neural Network AUC (Adjusted Threshold): 0.8442849
```

The decision threshold for classifying churners is adjusted to 0.54, focusing on improving specificity (reducing false positives) while maintaining acceptable sensitivity. The updated confusion matrix is then evaluated.

7.6 Final NN Model Performance (Resampling with Weight Ratio)

```
## 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
```

7.7 Neural Network with RSNNS

```
##        predictions
## targets   1   2
##       1 937 205
##       2  97 168
```

```
## Accuracy: 0.7854
```

```
## Sensitivity (Recall): 0.4504
```

```
## Specificity: 0.9062
```

```
## Precision: 0.634
```

7.8 Neural Network with RSNNS: Trying Alternative Activation Functions

[1] "Confusion Matrix (Act_TanH):" predictions targets 1 2 1 895 163 2 139 210 [1] "Confusion Matrix (Act_ReLU):" predictions targets 1 2 1 936 187 2 98 186 Metrics for Act_TanH: Accuracy: 0.7854 Sensitivity (Recall): 0.563 Specificity: 0.8656 Precision: 0.6017

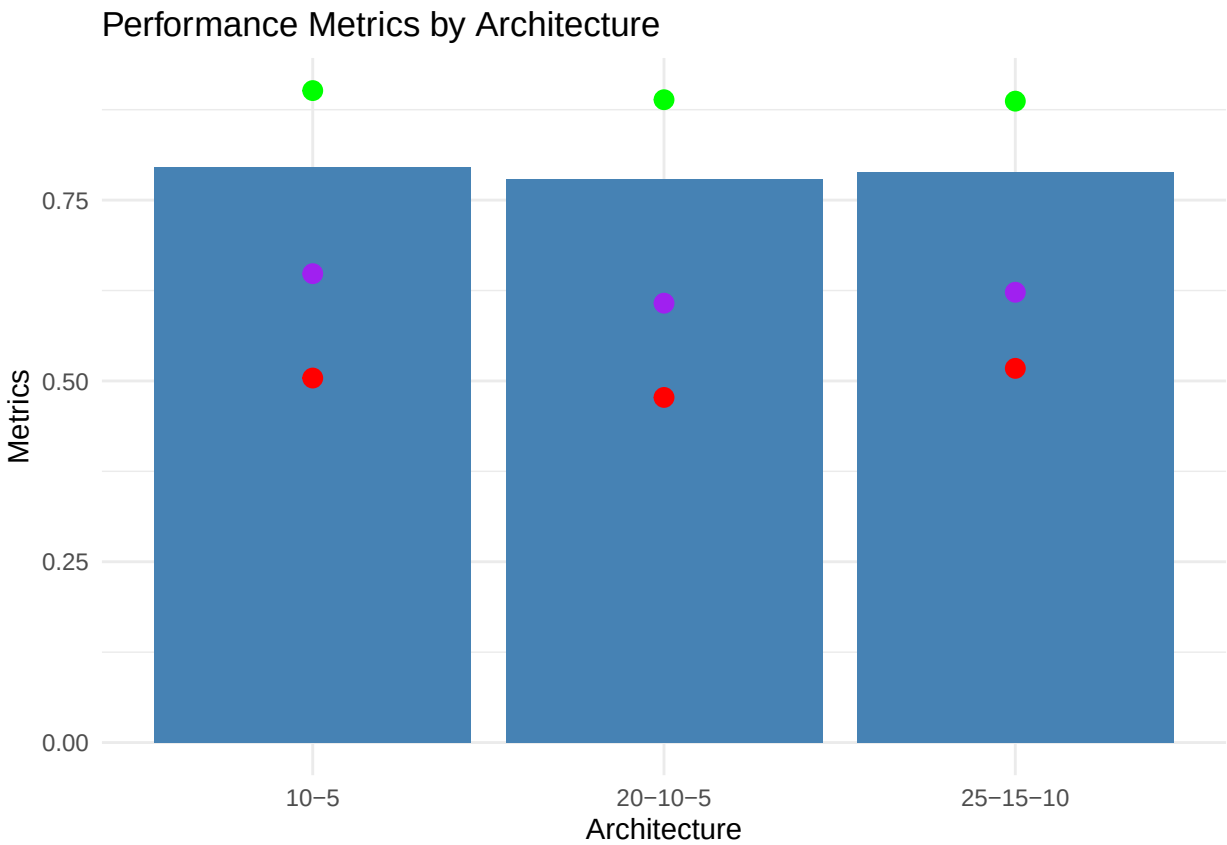
Metrics for Act_ReLU: Accuracy: 0.7974 Sensitivity (Recall): 0.4987 Specificity: 0.9052 Precision: 0.6549

7.9 Neural Network with Regularization and Adjusted Architecture

Metrics for Regularized Neural Network with Adjusted Architecture: Accuracy: 0.7861 Sensitivity (Recall): 0.4987 Specificity: 0.8897 Precision: 0.62

7.10 Neural Network Experimentation with Different Architectures

```
## Testing architecture: 20-10-5

## Testing architecture: 10-5

## Testing architecture: 25-15-10

##   Architecture Accuracy Sensitivity Specificity Precision
## 1      20-10-5   0.7797      0.4772      0.8888    0.6075
## 2         10-5   0.7960      0.5040      0.9014    0.6483
## 3     25-15-10   0.7889      0.5174      0.8868    0.6226
```

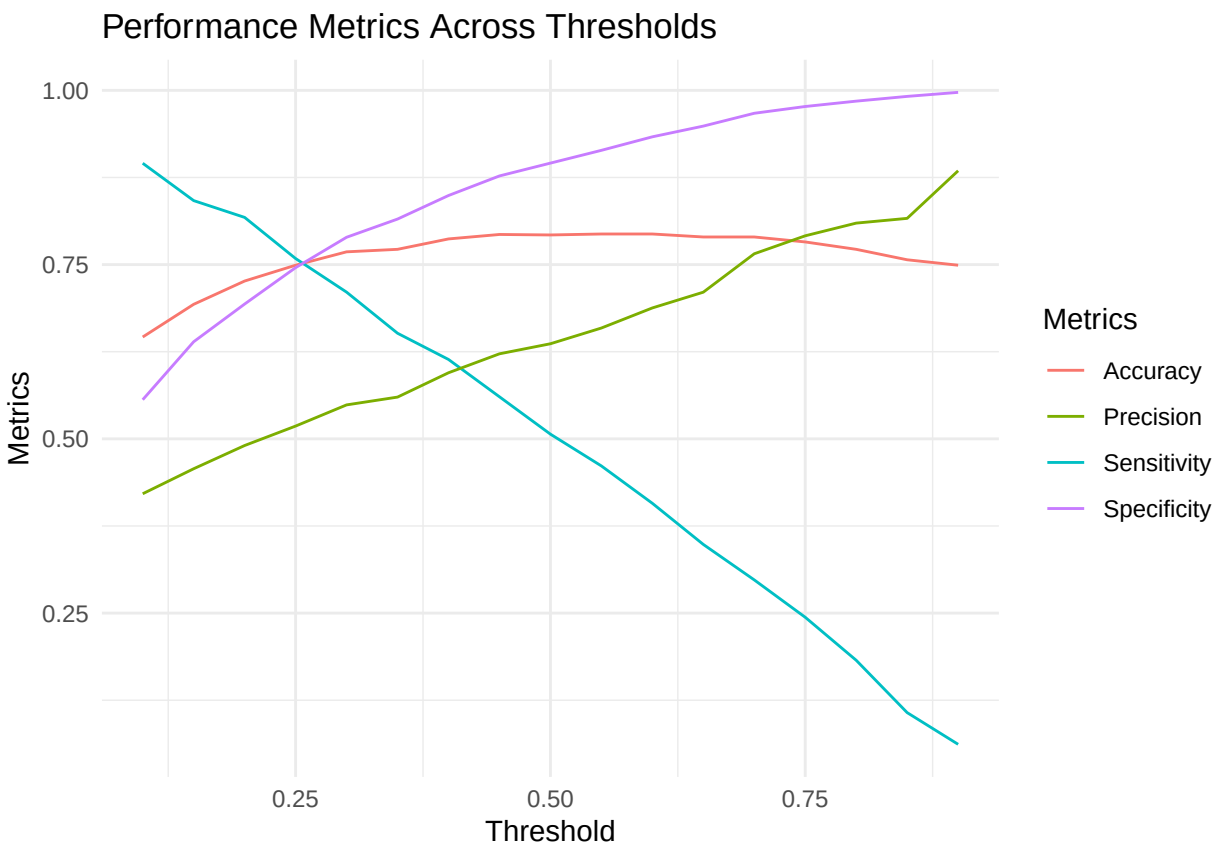



```
## Metrics for Neural Network with Dropout Regularization:

## Accuracy: 0.7925

## Sensitivity (Recall): 0.5067

## Specificity: 0.8956

## Precision: 0.6364

## Metrics with Adjusted Threshold (0.4):

## Accuracy: 0.7868

## Sensitivity (Recall): 0.6139
```

```
## Specificity: 0.8491
## Precision: 0.5948
```

7.11 Threshold Analysis and Visualization

```
##    Threshold Accuracy Sensitivity Specificity Precision
## 1       0.10   0.6461      0.8954      0.5561    0.4212
## 2       0.15   0.6930      0.8418      0.6393    0.4571
## 3       0.20   0.7264      0.8177      0.6934    0.4904
## 4       0.25   0.7491      0.7587      0.7456    0.5183
## 5       0.30   0.7683      0.7105      0.7892    0.5487
## 6       0.35   0.7719      0.6515      0.8153    0.5599
## 7       0.40   0.7868      0.6139      0.8491    0.5948
## 8       0.45   0.7932      0.5603      0.8772    0.6220
## 9       0.50   0.7925      0.5067      0.8956    0.6364
## 10      0.55   0.7939      0.4611      0.9139    0.6590
## 11      0.60   0.7939      0.4075      0.9333    0.6878
## 12      0.65   0.7896      0.3485      0.9487    0.7104
## 13      0.70   0.7896      0.2976      0.9671    0.7655
## 14      0.75   0.7825      0.2440      0.9768    0.7913
## 15      0.80   0.7719      0.1823      0.9845    0.8095
## 16      0.85   0.7569      0.1072      0.9913    0.8163
## 17      0.90   0.7491      0.0617      0.9971    0.8846
```



7.12 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.

7.13 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.

7.14 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.

8 Support Vector Machine (SVM)

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

8.1 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.

8.2 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.

8.3 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).

```
## [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
```

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.

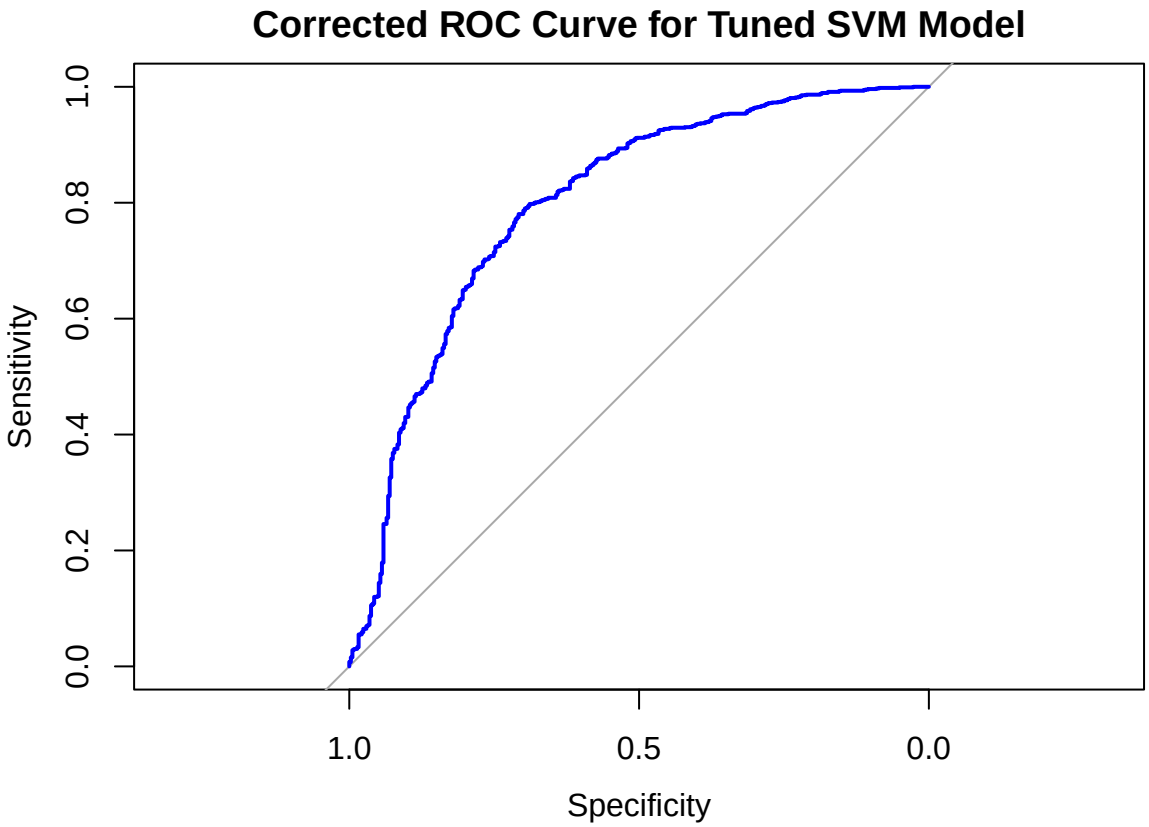
8.4 Evaluate additional metrics for the tuned model

```
## Accuracy: 0.7967306
```

```
## Precision: 0.818569
```

```
## Recall: 0.9294004
```

```
## F1-Score: 0.870471
```




```
## Corrected AUC: 0.7978581
```

8.5 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).

Sample Output:

```
##   cost gamma
## 3   10  0.01
```

```
##          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
```

8.6 Customer Segmentation Based on Churn Risk

Customers were segmented into three risk groups based on churn probabilities:

Table 2: Churn Probability and Risk Group Segmentation

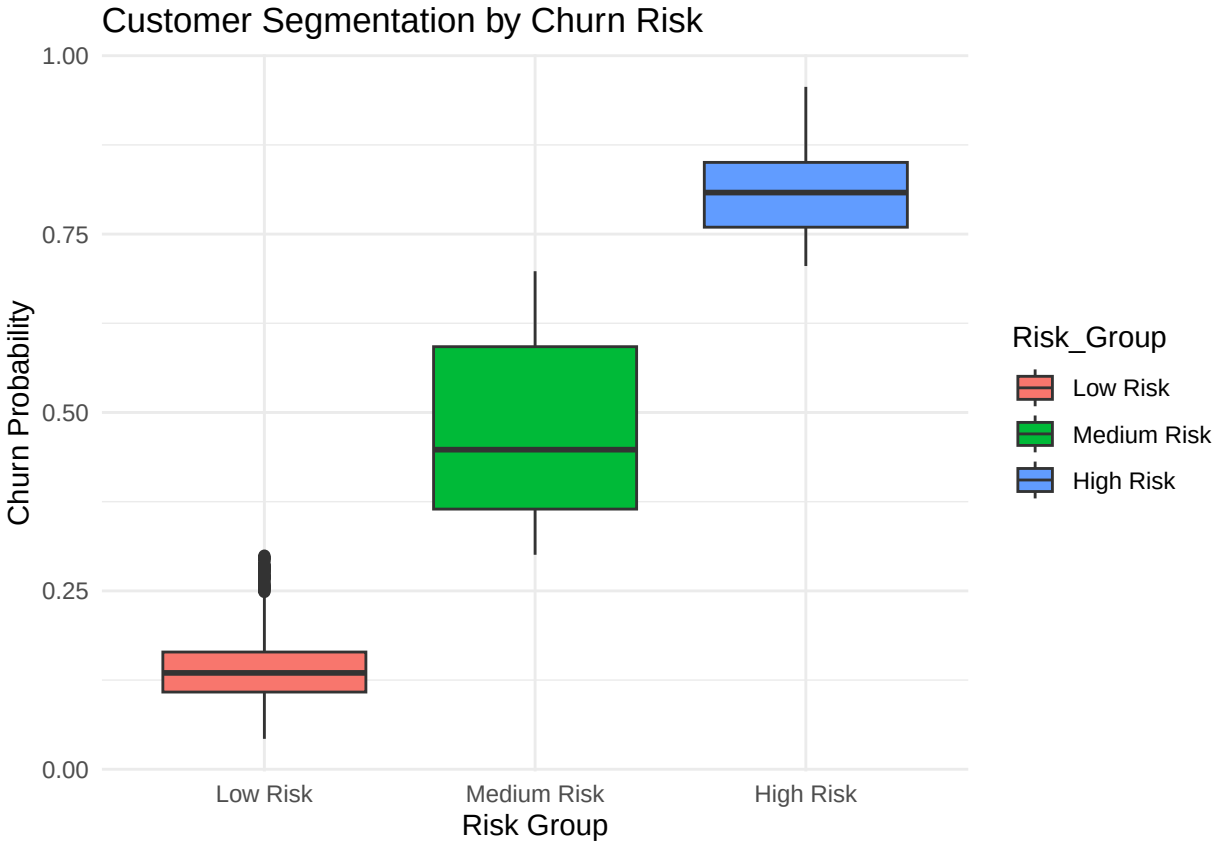
Risk Group	Churn Probability Range	Count
Low Risk	`< 0.3`	1,034
Medium Risk	`0.3 - 0.7`	228
High Risk	`> 0.7`	145

```
## # A tibble: 6 x 2
##   Churn_Prob Risk_Group
##        <dbl> <fct>
## 1     0.386  Medium Risk
## 2     0.469  Medium Risk
## 3     0.0812 Low Risk
## 4     0.108  Low Risk
## 5     0.185  Low Risk
## 6     0.132  Low Risk
```

Insights:

- **Low Risk**: Loyal customers with longer tenure and lower churn probability.
- **Medium Risk**: A transitional group that requires targeted retention efforts to prevent escalation to high risk.
- **High Risk**: Customers with the shortest tenure and highest churn probability, requiring immediate attention.

Visualization:



A **boxplot** was used to visualize the distribution of churn probabilities across risk groups, highlighting

distinct differences.

8.7 Key Feature Insights

Average Monthly Charges and Tenure by Risk Group:

```
## # 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 have the highest monthly charges and shortest tenure.
- **Low Risk** customers pay less on average but stay significantly longer.

Actionable Steps:

1. **High-Risk Customers**:
   - Offer discounts or loyalty incentives to reduce churn.
   - Focus on engagement strategies during the early months of the customer lifecycle.
2. **Medium-Risk Customers**:
   - Deploy retention campaigns, such as satisfaction surveys or targeted offers, to prevent escalation to high risk.
3. **Low-Risk Customers**:
   - Reward loyalty with perks, such as exclusive deals or access to premium services.

8.8 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.

8.9 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:

8.9.1 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.

8.9.2 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.

8.9.3 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.

8.9.4 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.

8.9.5 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.