GLM AND GAM

```r
# Load required libraries
library(ggplot2)

# Load the data
data <- read.csv("telecom_customers_churn_cleaned.csv")
data$Churn <- ifelse(data$Churn == "Yes", 1, 0)
#str(data)
#colSums(is.na(data))
```

GLM

```r
# Ensure variables are factors
data$Contract <- as.factor(data$Contract)
data$PaymentMethod <- as.factor(data$PaymentMethod)
data$Churn <- as.factor(data$Churn)
data$TotalCharges <- log(data$TotalCharges)
data$MonthlyCharges <- log(data$MonthlyCharges)
data$InternetService <- as.factor(data$InternetService)


# Set 'No service' as the reference level for InternetService
data$InternetService <- relevel(data$InternetService, ref = "No")
# Confirm reference level
cat("Reference level for 'Contract':", levels(data$Contract)[1], "\n")
```

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

```r
cat("Reference level for 'InternetService':", levels(data$InternetService)[1], "\n")
```

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

```r
# Fit the GLM with interaction terms
glm_model_interaction <- glm(Churn ~ SeniorCitizen + Contract + PaymentMethod +
                              tenure + TotalCharges +
                              InternetService + OnlineSecurity + TechSupport +
                              PaperlessBilling + MultipleLines,
                              family = binomial, data = data)
# Check the summary
summary(glm_model_interaction)
```

```
##
```

```
## Call:
## glm(formula = Churn ~ SeniorCitizen + Contract + PaymentMethod +
##     tenure + TotalCharges + InternetService + OnlineSecurity +
##     TechSupport + PaperlessBilling + MultipleLines, family = binomial,
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1917  -0.6772  -0.2991   0.5990   3.1554
##
## 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.

Did the model overfit?

```
library(pROC)
```

```
## Warning: package 'pROC' was built under R version 4.1.3
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
# Split data into training and test sets
set.seed(123)
train_indices <- sample(1:nrow(data), 0.7 * nrow(data))
train_data <- data[train_indices, ]
test_data <- data[-train_indices, ]

# Predict on training and test data
train_preds <- predict(glm_model_interaction, train_data, type = "response")
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading
```

```r
test_preds <- predict(glm_model_interaction, test_data, type = "response")
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading
```

```r
# Calculate performance metrics
library(pROC)
train_auc <- roc(train_data$Churn, train_preds)$auc
```

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

## Setting direction: controls < cases
```

```r
test_auc <- roc(test_data$Churn, test_preds)$auc
```

```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```

```r
cat("Train AUC:", train_auc, "\n")
```

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

```r
cat("Test AUC:", test_auc, "\n")
```

```
## 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.

```r
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.1.3

## Loading required package: lattice
```

```r
# Set up cross-validation
control <- trainControl(method = "cv", number = 10) # 10-fold cross-validation

# Train the model
cv_model <- train(Churn ~ SeniorCitizen + Contract + PaymentMethod +
                  tenure + MonthlyCharges + TotalCharges +
                  InternetService + InternetService:MonthlyCharges,
                  data = data, method = "glm", family = "binomial", trControl = control)

# Check cross-validation results
print(cv_model)
```

```
## 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

GAM WITH NON-LINEAR PATTERNS

```r
library(mgcv)
```

```
## Loading required package: nlme
```

```
## This is mgcv 1.8-36. For overview type 'help("mgcv-package")'.
```

```r
data$TotalCharges_centered <- data$TotalCharges- mean(data$TotalCharges)
cat("Reference level for 'Contract':", levels(data$Contract)[1], "\n")
```

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

```r
cat("Reference level for 'InternetService':", levels(data$InternetService)[1], "\n")
```

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

```r
gam_smoothing_spline <- gam(Churn ~ SeniorCitizen + Contract + PaymentMethod +
                             s(tenure) + s(TotalCharges_centered) + tenure +  TotalCharges +
                             InternetService + OnlineSecurity + TechSupport +
                             PaperlessBilling + MultipleLines,
                             family = binomial, data = data)

summary(gam_smoothing_spline)
```

```
##
## Family: binomial
## Link function: logit
##
## Formula:
## Churn ~ SeniorCitizen + Contract + PaymentMethod + s(tenure) +
```

```
##     s(TotalCharges_centered) + tenure + TotalCharges + InternetService +
##     OnlineSecurity + TechSupport + PaperlessBilling + MultipleLines
##
## Parametric coefficients:
##                                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)                           0.00000    0.00000     NaN      NaN
## SeniorCitizen                         0.24804    0.08294   2.991 0.002784 **
## ContractOne year                     -0.60206    0.10785  -5.582 2.37e-08 ***
## ContractTwo year                     -1.42503    0.18646  -7.642 2.13e-14 ***
## PaymentMethodCredit card (automatic) -0.08295    0.11336  -0.732 0.464342
## PaymentMethodElectronic check         0.31008    0.09422   3.291 0.000999 ***
## PaymentMethodMailed check            -0.15815    0.11687  -1.353 0.175990
## tenure                               -0.05358    0.01797  -2.982 0.002863 **
## TotalCharges                          0.14308    0.08046   1.778 0.075363 .
## InternetServiceDSL                   -1.03226    0.10889  -9.480  < 2e-16 ***
## InternetServiceFiber optic            0.00000    0.00000     NaN      NaN
## OnlineSecurityNo internet service     0.00000    0.00000     NaN      NaN
## OnlineSecurityYes                    -0.35708    0.08500  -4.201 2.66e-05 ***
## TechSupportNo internet service       -2.19811    0.20640 -10.650  < 2e-16 ***
## TechSupportYes                       -0.24301    0.08664  -2.805 0.005036 **
## PaperlessBillingYes                   0.39515    0.07512   5.260 1.44e-07 ***
## MultipleLinesNo phone service         0.55051    0.13923   3.954 7.68e-05 ***
## MultipleLinesYes                      0.37316    0.08162   4.572 4.83e-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)                2.684  3.615  7.547   0.114
## s(TotalCharges_centered) 3.108  3.972 42.474  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Rank: 32/36
## R-sq.(adj) =  0.317   Deviance explained = 29.2%
## UBRE = -0.17486  Scale est. = 1         n = 7043
```

Smooth terms for tenure and TotalCharges_centered are highly significant ($p < 0.001$), indicating non-linear effects.

Let's compare these models

ROC COMPARISON

```r
library(pROC)

# Predictions for glm_model
pred1 <- predict(glm_model_interaction, type = "response")
roc1 <- roc(data$Churn, pred1)
```

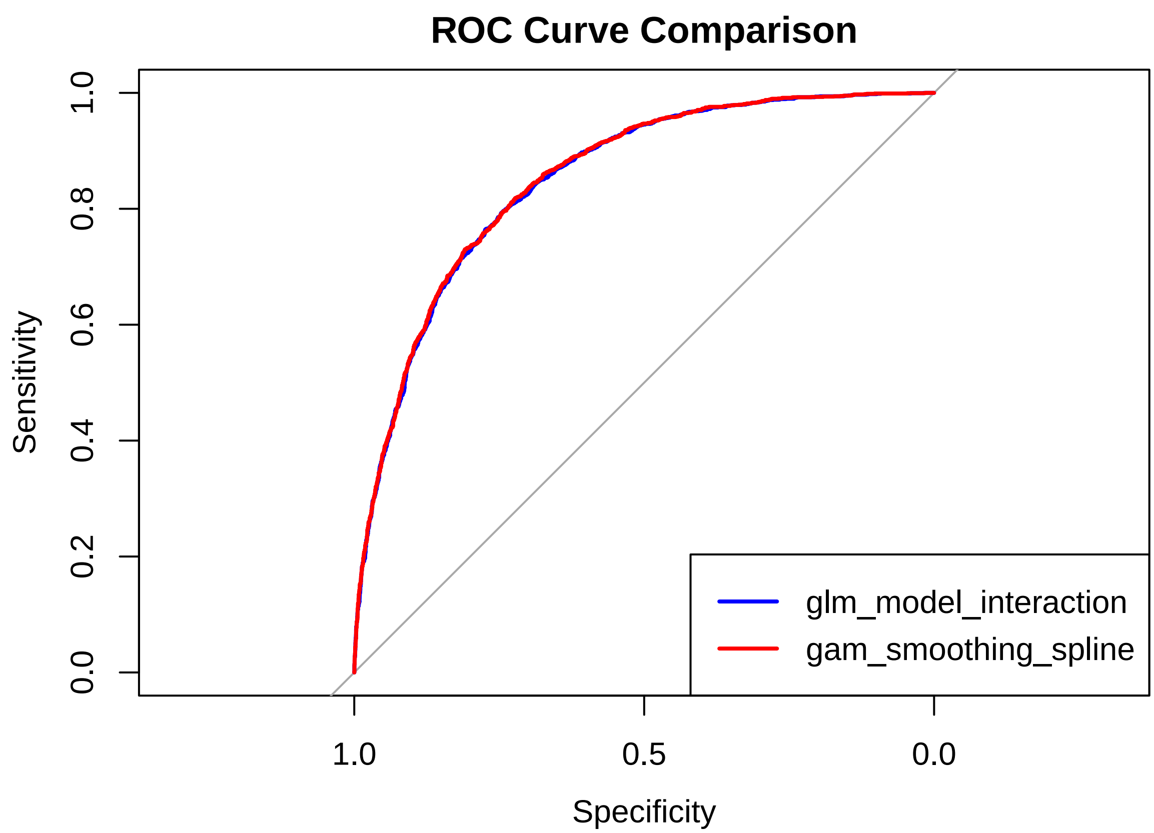
```
## Setting levels: control = 0, case = 1
```

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

```r
pred2 <- predict(gam_smoothing_spline, type = "response")
roc2 <- roc(data$Churn, pred2)
```

```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```

```r
# Plot ROC curves
plot(roc1, col = "blue", lwd = 2, main = "ROC Curve Comparison")
lines(roc2, col = "red", lwd = 2)
legend("bottomright", legend = c("glm_model_interaction", "gam_smoothing_spline"),
       col = c("blue", "red"), lwd = 2)
```




```r
auc1 <- auc(roc1)
auc2 <- auc(roc2)
cat("AUC of Model glm_model_interaction:", auc1, "\n")
```

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

```r
cat("AUC of Model gam_smoothing_spline:", auc2, "\n")
```

```
## AUC of Model gam_smoothing_spline: 0.8509579
```

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

```
anova(glm_model_interaction, gam_smoothing_spline, test = "Chisq")
```

```
## 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) + tenure + TotalCharges + InternetService +
##     OnlineSecurity + TechSupport + PaperlessBilling + MultipleLines
##   Resid. Df Resid. Dev     Df Deviance  Pr(>Chi)
## 1    7027.0     5796.5
## 2    7022.2     5769.9 4.7925   26.659 5.407e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

GAM explains more variance than GLM.

```
AIC(glm_model_interaction, gam_smoothing_spline)
```

```
##                             df      AIC
## glm_model_interaction 16.00000 5828.536
## gam_smoothing_spline  20.79248 5811.463
```

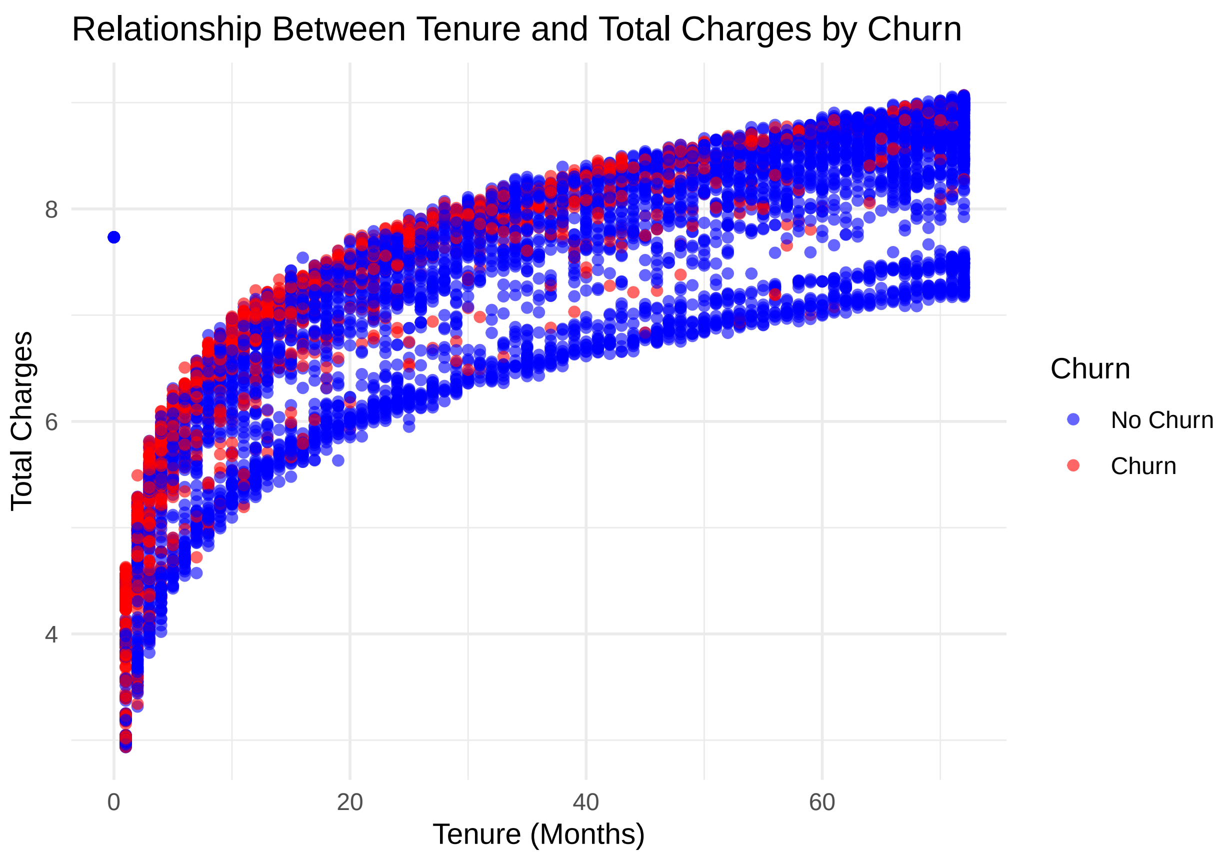
GAM has a slightly lower AIC (5810.983) than GLM (5828.536), further supporting that GAM provides a better fit to the data.

Let's plot some results

Plot: Tenure vs. TotalCharges by Churn

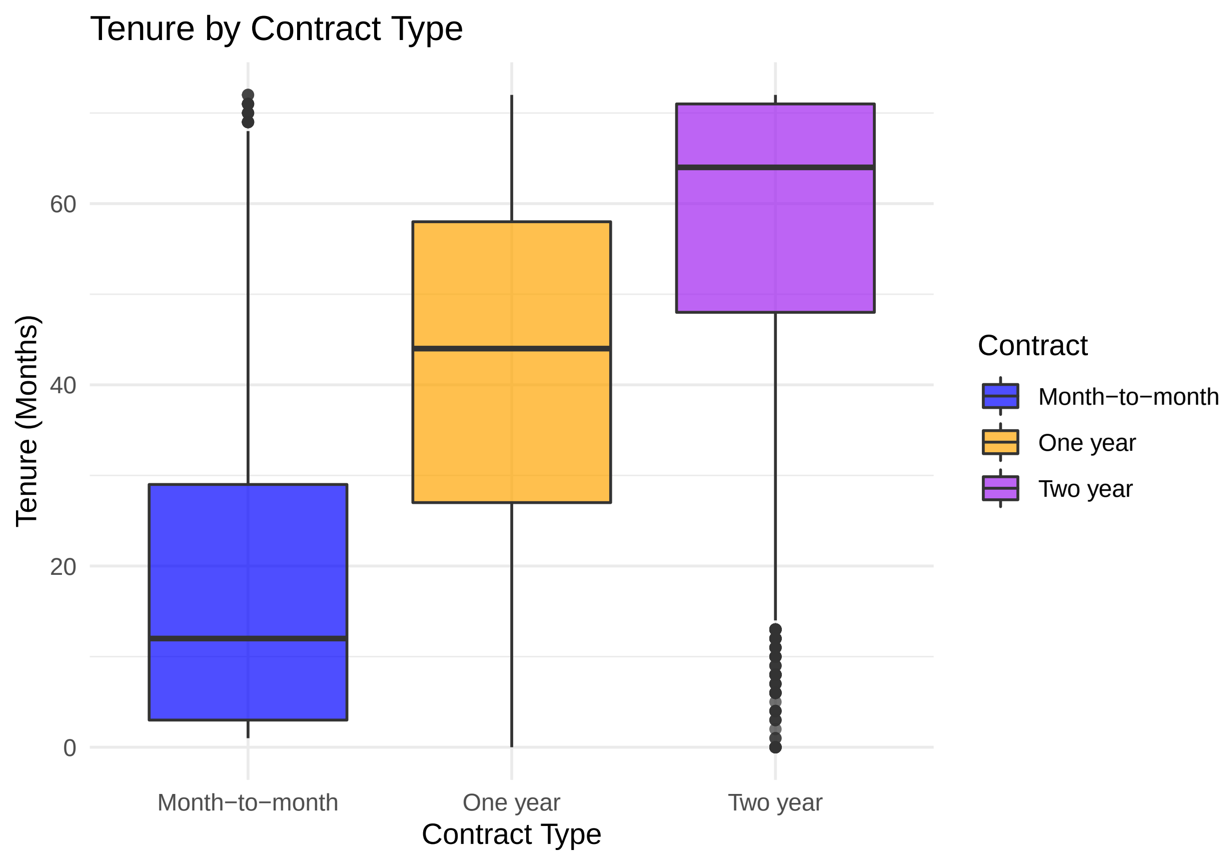
```
# Ensure necessary libraries are installed
if(!require(ggplot2)) install.packages("ggplot2")

# Load the library
library(ggplot2)


ggplot(data, aes(x = tenure, y = TotalCharges, color = Churn)) +
  geom_point(alpha = 0.6) +
  labs(
    title = "Relationship Between Tenure and Total Charges by Churn",
    x = "Tenure (Months)",
```

```
    y = "Total Charges"
  ) +
  scale_color_manual(values = c("blue", "red"), labels = c("No Churn", "Churn")) +
  theme_minimal()
```



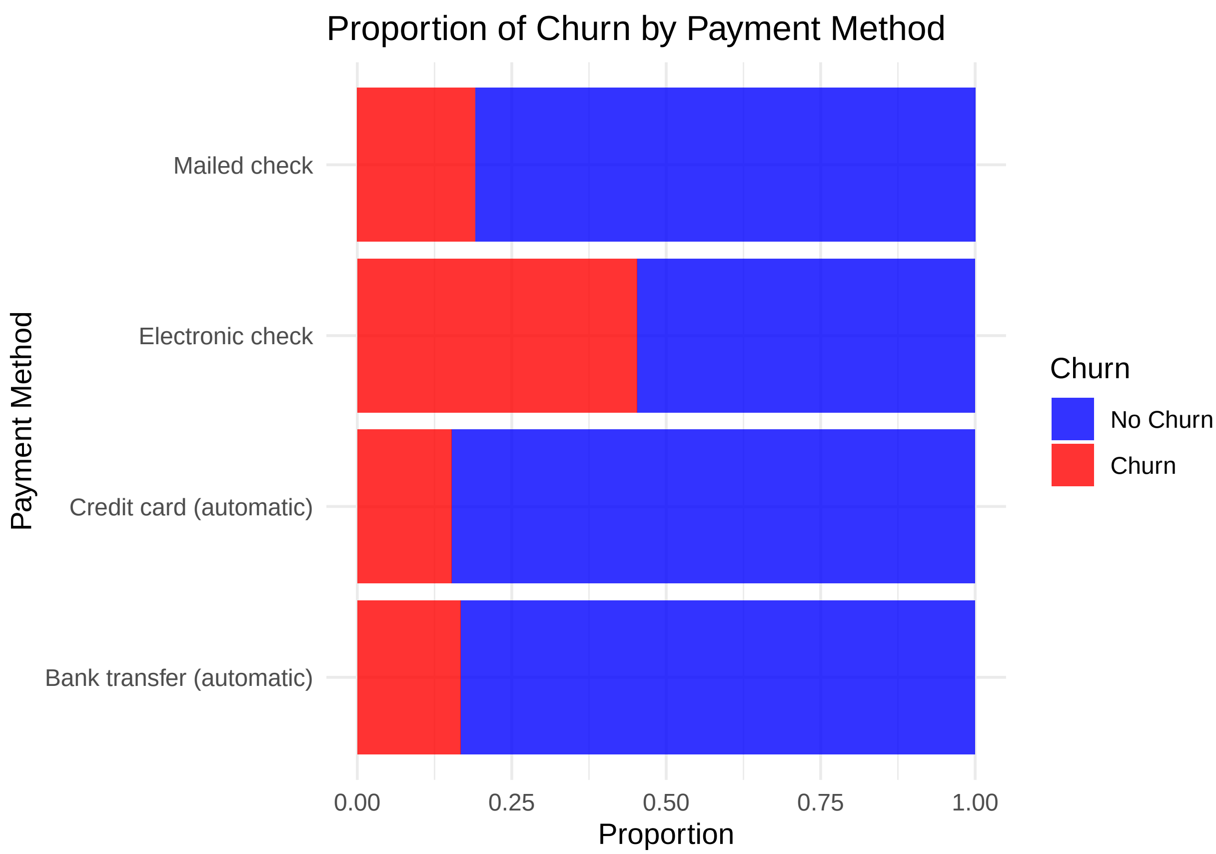
Plot: Tenure by Contract Type

```
ggplot(data, aes(x = Contract, y = tenure, fill = Contract)) +
  geom_boxplot(alpha = 0.7) +
  labs(
    title = "Tenure by Contract Type",
    x = "Contract Type",
    y = "Tenure (Months)"
  ) +
  scale_fill_manual(values = c("blue", "orange", "purple")) +
  theme_minimal()
```



Plot: Churn by Payment Method

```
ggplot(data, aes(x = PaymentMethod, fill = Churn)) +
  geom_bar(position = "fill", alpha = 0.8) +
  labs(
    title = "Proportion of Churn by Payment Method",
    x = "Payment Method",
    y = "Proportion"
  ) +
  scale_fill_manual(values = c("blue", "red"), labels = c("No Churn", "Churn")) +
  theme_minimal() +
  coord_flip()
```



Plot: Churn Proportion by Internet Service

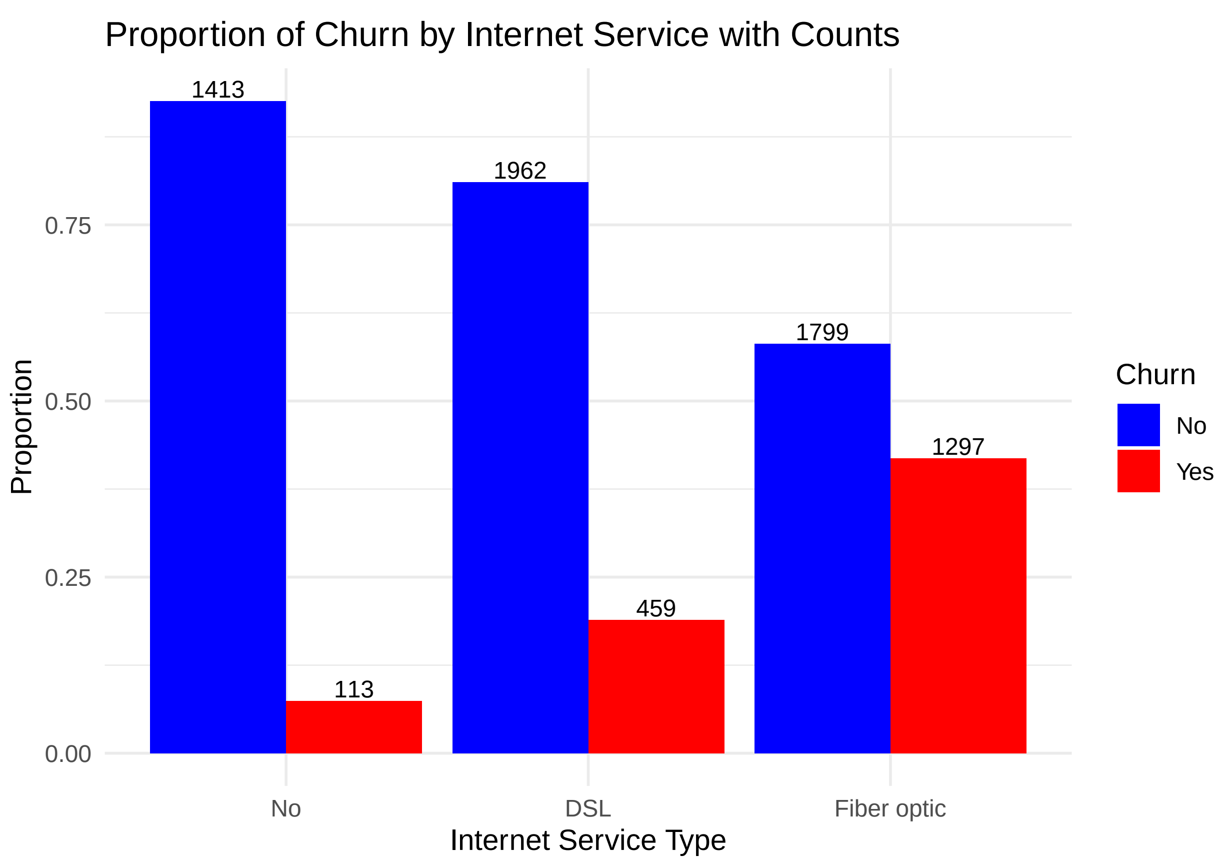
```
# Load necessary libraries
if (!require(ggplot2)) install.packages("ggplot2")
if (!require(dplyr)) install.packages("dplyr")
```

```
## Loading required package: dplyr

## Warning: package 'dplyr' was built under R version 4.1.3

##
## Attaching package: 'dplyr'

## The following object is masked from 'package:nlme':
##
##     collapse

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(ggplot2)
library(dplyr)

# Create a dataset for proportions and counts
proportion_df <- data %>%
  group_by(InternetService, Churn) %>%
  summarise(Count = n()) %>%
  mutate(Total = sum(Count),
         Proportion = Count / Total)
```

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

```r
# Plot the bar chart with counts displayed
ggplot(proportion_df, aes(x = InternetService, y = Proportion, fill = as.factor(Churn))) +
  geom_bar(stat = "identity", position = "dodge") +
  geom_text(aes(label = Count), position = position_dodge(width = 0.9), vjust = -0.2, size = 3) +
  labs(
    title = "Proportion of Churn by Internet Service with Counts",
    x = "Internet Service Type",
    y = "Proportion",
    fill = "Churn"
  ) +
  scale_fill_manual(values = c("blue", "red"), labels = c("No", "Yes")) +
  theme_minimal()
```



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