Customer Churn Analysis Project

Hayeon Chung

2025-06-26

1. Introduction

This project analyzes customer churn data from a telecom company to identify key churn predictors and recommend retention strategies.

2. Data Loading and Cleaning

```r
churn <- read_csv("/Users/hayeonchung/Downloads/telco_churn.csv") %>% clean_names()

# Check for NAs
skimr::skim(churn)
```

Table 1: Data summary

Name	churn
Number of rows	7043
Number of columns	21
Column type frequency:	
character	17
numeric	4
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
customer_id	0	1	10	10	0	7043	0
gender	0	1	4	6	0	2	0
partner	0	1	2	3	0	2	0
dependents	0	1	2	3	0	2	0
phone_service	0	1	2	3	0	2	0
multiple_lines	0	1	2	16	0	3	0
internet_service	0	1	2	11	0	3	0
online_security	0	1	2	19	0	3	0
online_backup	0	1	2	19	0	3	0
device_protection	0	1	2	19	0	3	0
tech_support	0	1	2	19	0	3	0
streaming_tv	0	1	2	19	0	3	0

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
streaming_movies	0	1	2	19	0	3	0
contract	0	1	8	14	0	3	0
paperless_billing	0	1	2	3	0	2	0
payment_method	0	1	12	25	0	4	0
churn	0	1	2	3	0	2	0

Variable type: numeric

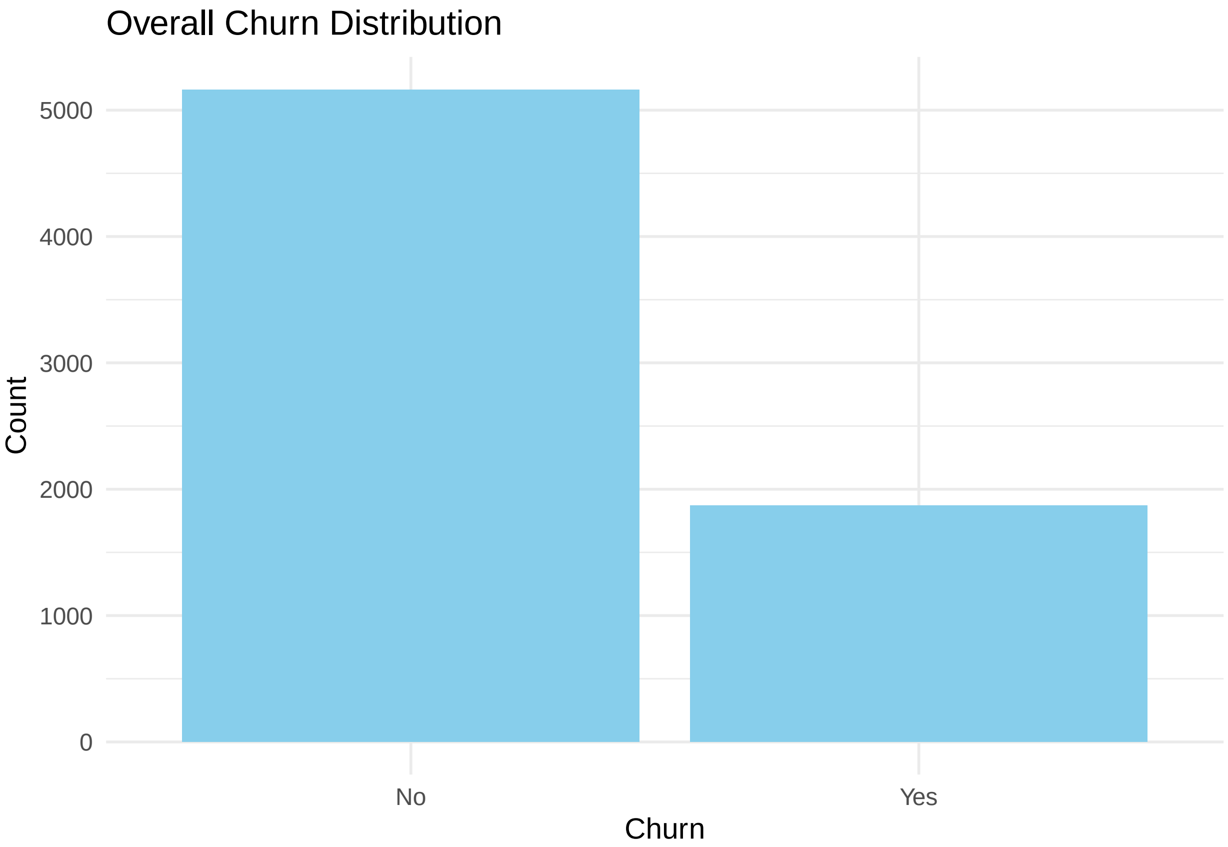
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
senior_citizen	0	1	0.16	0.37	0.00	0.00	0.00	0.00	1.00	
tenure	0	1	32.37	24.56	0.00	9.00	29.00	55.00	72.00	
monthly_charges	0	1	64.76	30.09	18.25	35.50	70.35	89.85	118.75	
total_charges	11	1	2283.30	2266.77	18.80	401.45	1397.47	3794.74	8684.80	

```r
# Clean TotalCharges (convert to numeric, handle blanks)
churn <- churn %>%
  mutate(total_charges = as.numeric(trimws(total_charges)))

# Impute missing values if needed
churn <- churn %>% filter(!is.na(total_charges))
```

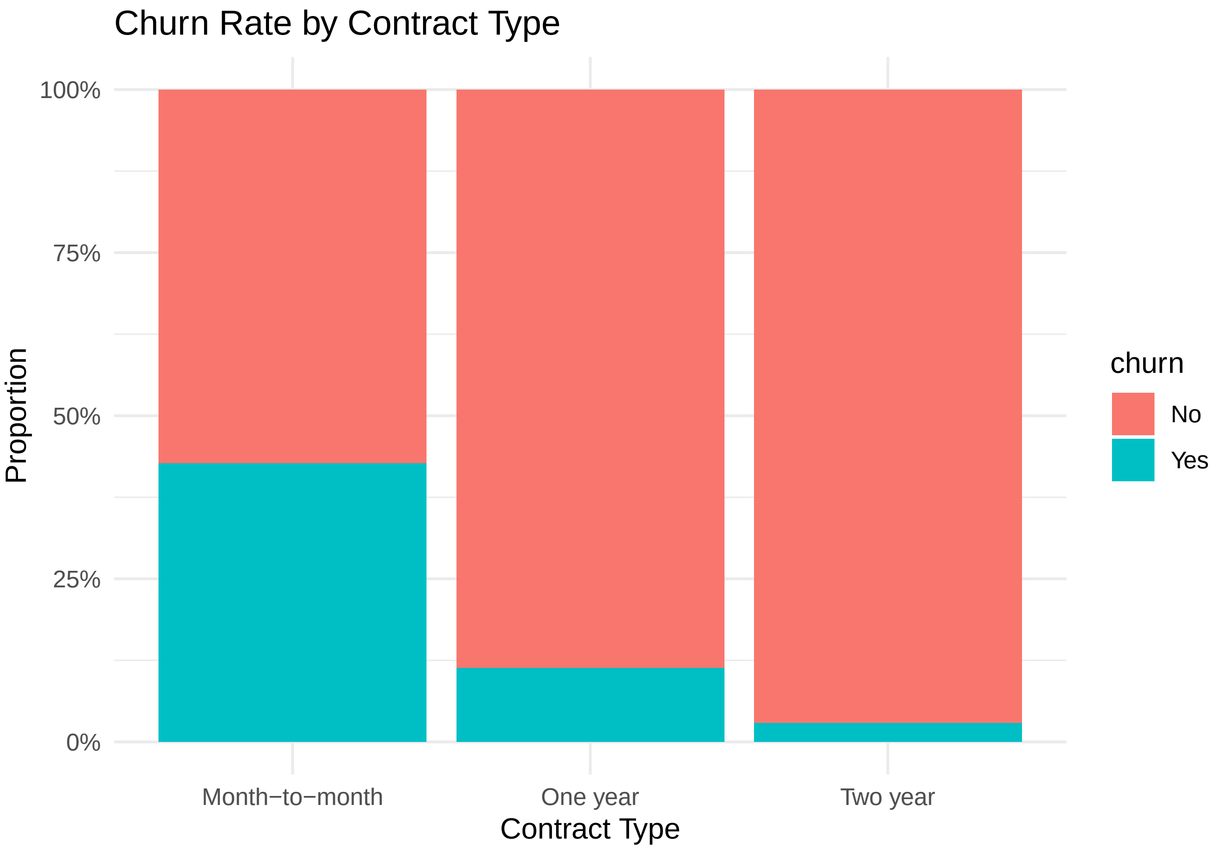
The initial summary reveals that the dataset contains 7,032 rows and 21 columns with 17 columns identified as character type and 4 columns as numeric. The skim function output displays that the dataset is complete with no missing values across any of the variables which simplifies the cleaning process and ensures a strong foundation for modeling. All character columns, including key categorical predictors such as gender, contract, and payment method, will be converted to factor type in preparation for modeling. The customer_id variable will be dropped since it holds no predictive value. There are no empty strings or whitespace issues within the data which demosntrates it is well-structured and ready for exploratory data analysis.

3. Exploratory Data Analysis (EDA)

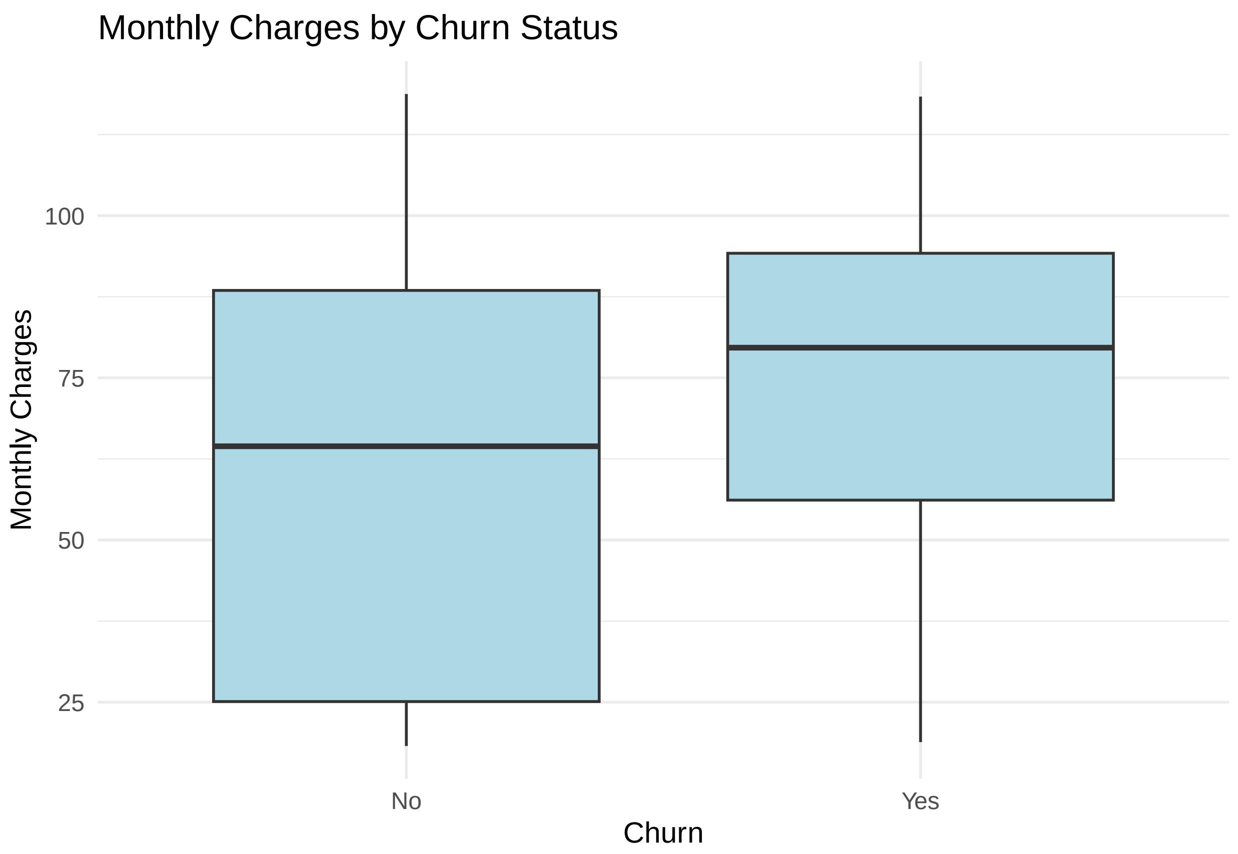
```r
# Churn distribution (overall)
ggplot(churn, aes(x = churn)) +
  geom_bar(fill = "skyblue") +
  labs(title = "Overall Churn Distribution",
       x = "Churn", y = "Count") + theme_minimal()
```



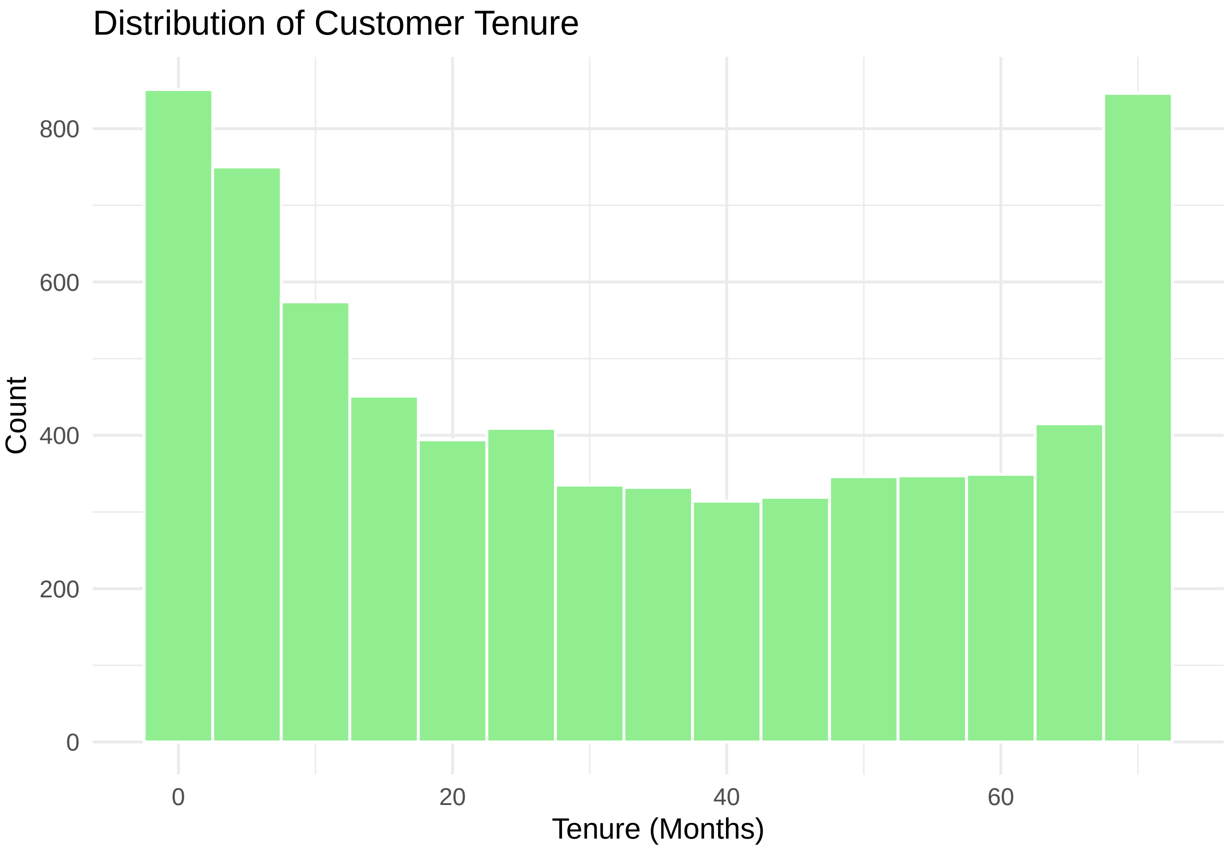
```r
# Churn rate by Contract Type
ggplot(churn, aes(x = contract, fill = churn)) +
  geom_bar(position = "fill") +
  labs(title = "Churn Rate by Contract Type",
       x = "Contract Type", y = "Proportion") +
  scale_y_continuous(labels = scales::percent) + theme_minimal()
```



```r
# Monthly Charges by Churn (Boxplot)
ggplot(churn, aes(x = churn, y = monthly_charges)) +
  geom_boxplot(fill = "lightblue") +
  labs(title = "Monthly Charges by Churn Status",
       x = "Churn", y = "Monthly Charges") + theme_minimal()
```



```r
# Tenure distribution (Histogram)
ggplot(churn, aes(x = tenure)) +
  geom_histogram(binwidth = 5, fill = "lightgreen", color = "white") +
  labs(title = "Distribution of Customer Tenure",
       x = "Tenure (Months)", y = "Count") + theme_minimal()
```



These visualizations provide initial insights into customer churn behavior. The first bar chart confirms the dataset is imbalanced with a significantly higher number of customers who did not churn compared to those who did. When examining churn rate by contract type, customers on month-to-month contracts have a much higher churn rate than those on one or two year contracts. This suggests that long-term agreements may serve as a natural retention mechanism. The boxplot of monthly charges by churn status demonstrates that customers who churn tend to have higher monthly charges on average which indicates potential dissatisfaction with pricing or service value. Lastly, the histogram of customer tenure shows many customers tend to leave early in their life cycle, particularly within the first 12 months. Another noticeable group remains long-term, often around the 70+ month range. These insights suggest both contract structure and billing level play a key role in customer retention and the early stages of a customer's life cycle are critical.

4. Feature Engineering

```
# Convert categorical to factors
churn <- churn %>%
  mutate_if(is.character, as.factor)

# Create tenure bucket
churn <- churn %>%
  mutate(tenure_group = cut(tenure, breaks = c(0, 12, 24, 48, 60, Inf),
                            labels = c("0-12", "13-24", "25-48", "49-60", "60+")))

# Drop customerID
churn$customer_id <- NULL
```

This step involves transforming and preparing the existing variables in a way that ensures the predictive power of the model. Many of the variables in the data are categorical and need to be converted into factors so they can be properly interpreted by modeling algorithms in R in the following steps. Additionally, creating new features such as tenure groups helps to capture nonlinear patterns in customer behavior. Feature engineering is an important step that improves model accuracy, interpretability, and compatibility with different types of algorithms used later in the analysis.

5. Model Building

Logistic Regression

```r
# Train/test split
set.seed(123)
train_idx <- createDataPartition(churn$churn, p = 0.8, list = FALSE)
train <- churn[train_idx, ]
test <- churn[-train_idx, ]

# Logistic Regression
log_model <- glm(churn ~ ., data = train, family = binomial)
summary(log_model)
```

```
##
## Call:
## glm(formula = churn ~ ., family = binomial, data = train)
##
## Coefficients: (7 not defined because of singularities)
##                                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)                             1.649e+00  9.081e-01   1.816  0.06932
## genderMale                             -3.288e-02  7.284e-02  -0.451  0.65167
## senior_citizen                          2.708e-01  9.455e-02   2.864  0.00418
## partnerYes                             -2.828e-02  8.742e-02  -0.324  0.74631
## dependentsYes                          -7.566e-02  1.012e-01  -0.747  0.45492
## tenure                                 -6.832e-02  1.066e-02  -6.410 1.45e-10
## phone_serviceYes                        4.804e-01  7.219e-01   0.666  0.50573
## multiple_linesNo phone service                 NA         NA      NA       NA
## multiple_linesYes                       5.238e-01  1.973e-01   2.655  0.00793
## internet_serviceFiber optic             2.189e+00  8.893e-01   2.461  0.01385
## internet_serviceNo                     -2.141e+00  8.977e-01  -2.385  0.01707
## online_securityNo internet service             NA         NA      NA       NA
## online_securityYes                     -9.524e-02  1.983e-01  -0.480  0.63106
## online_backupNo internet service               NA         NA      NA       NA
## online_backupYes                        1.134e-01  1.958e-01   0.579  0.56251
## device_protectionNo internet service           NA         NA      NA       NA
## device_protectionYes                    2.838e-01  1.965e-01   1.444  0.14872
## tech_supportNo internet service                NA         NA      NA       NA
## tech_supportYes                        -3.687e-02  2.014e-01  -0.183  0.85474
## streaming_tvNo internet service                NA         NA      NA       NA
## streaming_tvYes                         8.065e-01  3.629e-01   2.223  0.02624
## streaming_moviesNo internet service            NA         NA      NA       NA
## streaming_moviesYes                     8.492e-01  3.643e-01   2.331  0.01976
## contractOne year                       -7.043e-01  1.211e-01  -5.818 5.95e-09
## contractTwo year                       -1.540e+00  2.050e-01  -7.516 5.65e-14
## paperless_billingYes                    3.543e-01  8.372e-02   4.231 2.32e-05
## payment_methodCredit card (automatic)  1.180e-03  1.266e-01   0.009  0.99256
## payment_methodElectronic check         2.518e-01  1.055e-01   2.387  0.01699
## payment_methodMailed check            -1.638e-01  1.288e-01  -1.272  0.20328
## monthly_charges                       -5.576e-02  3.537e-02  -1.577  0.11491
## total_charges                          2.065e-04  8.414e-05   2.454  0.01411
## tenure_group13-24                     -1.852e-01  1.533e-01  -1.208  0.22686
## tenure_group25-48                      2.522e-01  2.689e-01   0.938  0.34833
## tenure_group49-60                      9.418e-01  4.316e-01   2.182  0.02909
## tenure_group60+                        1.119e+00  5.351e-01   2.091  0.03650
##
## (Intercept)                           .
## genderMale
## senior_citizen                        **
## partnerYes
## dependentsYes
## tenure                                ***
## phone_serviceYes
## multiple_linesNo phone service
## multiple_linesYes                     **
## internet_serviceFiber optic           *
## internet_serviceNo                    *
## online_securityNo internet service
## online_securityYes
## online_backupNo internet service
## online_backupYes
## device_protectionNo internet service
## device_protectionYes
## tech_supportNo internet service
## tech_supportYes
## streaming_tvNo internet service
## streaming_tvYes                       *
## streaming_moviesNo internet service
## streaming_moviesYes                   *
## contractOne year                      ***
## contractTwo year                      ***
## paperless_billingYes                  ***
## payment_methodCredit card (automatic)
## payment_methodElectronic check        *
## payment_methodMailed check
## monthly_charges
## total_charges                         *
## tenure_group13-24
## tenure_group25-48
## tenure_group49-60                     *
## tenure_group60+                       *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6517.2  on 5626  degrees of freedom
## Residual deviance: 4635.6  on 5599  degrees of freedom
## AIC: 4691.6
##
## Number of Fisher Scoring iterations: 6
```

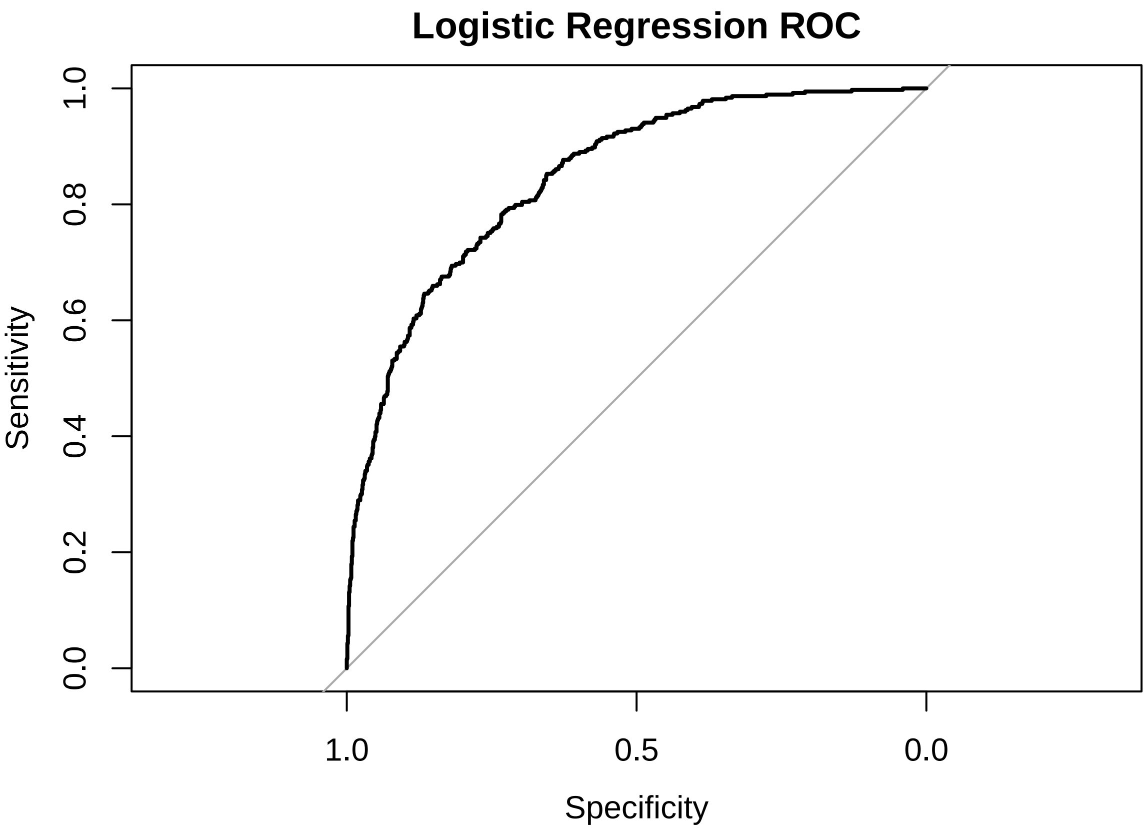
```r
# Predict & evaluate
log_probs <- predict(log_model, test, type = "response")
log_pred <- ifelse(log_probs > 0.5, "Yes", "No") %>% factor(levels = c("No", "Yes"))

confusionMatrix(log_pred, test$churn)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No Yes
##        No  932 166
##        Yes 100 207
##
##                Accuracy : 0.8107
##                  95% CI : (0.7892, 0.8309)
##     No Information Rate : 0.7345
##     P-Value [Acc > NIR] : 1.364e-11
##
##                   Kappa : 0.4855
##
##  Mcnemar's Test P-Value : 6.736e-05
##
##             Sensitivity : 0.9031
##             Specificity : 0.5550
##          Pos Pred Value : 0.8488
##          Neg Pred Value : 0.6743
##              Prevalence : 0.7345
##          Detection Rate : 0.6633
##    Detection Prevalence : 0.7815
##       Balanced Accuracy : 0.7290
##
##        'Positive' Class : No
##
```

```r
# ROC AUC
roc_obj <- roc(test$churn, log_probs)
auc(roc_obj)
```

```
## Area under the curve: 0.8469
```

```r
plot(roc_obj, main = "Logistic Regression ROC")
```



The logistic regression model performed reasonably well in predicting customer churn. The overall accuracy of the model is approximately 79.6% which is significantly better than the No Information Rate (73.45%) as indicated by a very small p-value of 1.36e-11. This model demonstrates high sensitivity with a value of 0.9031 which means it correctly identifies over 90% of customers who did not churn. However, specificity is lower (0.555) which indicates that the model is less effective at correctly identifying customers who did churn. This trade-off is common in imbalanced datasets.

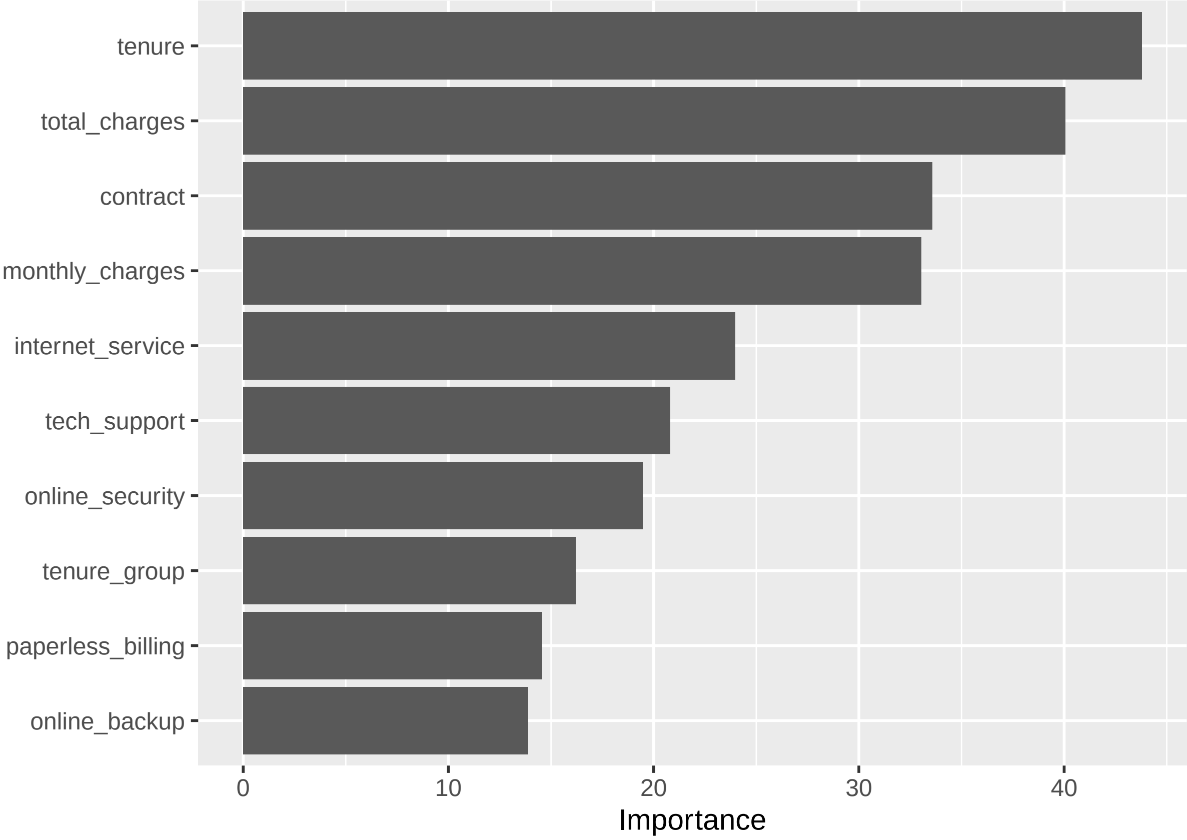
The area under the ROC curve (AUC) is 0.8469 which suggests the model has strong discriminating power overall. The positive predictive value (precision for non-churn) is 0.8488 while the negative predictive value (for churn) is 0.6743. These values combined with a balanced accuracy of 0.729 demonstrate the model performs well at distinguishing between churn and non-churn cases but could benefit from further tuning or more advanced techniques to improve specificity.

Random Forest

```
rf_model <- randomForest(churn ~ ., data = train, ntree = 500, importance = TRUE)
rf_pred <- predict(rf_model, test)

confusionMatrix(rf_pred, test$churn)
```

```
## Confusion Matrix and Statistics
##
##             Reference
## Prediction   No Yes
##        No   930 179
##        Yes  102 194
##
##                 Accuracy : 0.8
##                   95% CI : (0.7781, 0.8206)
##      No Information Rate : 0.7345
##      P-Value [Acc > NIR] : 6.374e-09
##
##                    Kappa : 0.451
##
##   Mcnemar's Test P-Value : 5.794e-06
##
##              Sensitivity : 0.9012
##              Specificity : 0.5201
##           Pos Pred Value : 0.8386
##           Neg Pred Value : 0.6554
##               Prevalence : 0.7345
##           Detection Rate : 0.6619
##     Detection Prevalence : 0.7893
##        Balanced Accuracy : 0.7106
##
##         'Positive' Class : No
##
```

```
# Variable importance
vip::vip(rf_model)
```



The Random Forest model achieved an accuracy of 80% which is a slightly better performance than the logistic regression model. It significantly outperformed the No Information Rate of 73.455 with a highly significant p-value of 6.37e09. The model displays strong sensitivity of 0.9012 which means it correctly identifies the vast majority of non-churn customers. However, its specificity was lower at 0.5201 which indicates room for improvement in correctly identifying actual churn cases. The balanced accruacy of 0.7106 confirms the model performs moderately well across both classes despite the class imbalance. The Kappa statistic of 0.451 reflects moderate agreement between predictions and true labels.

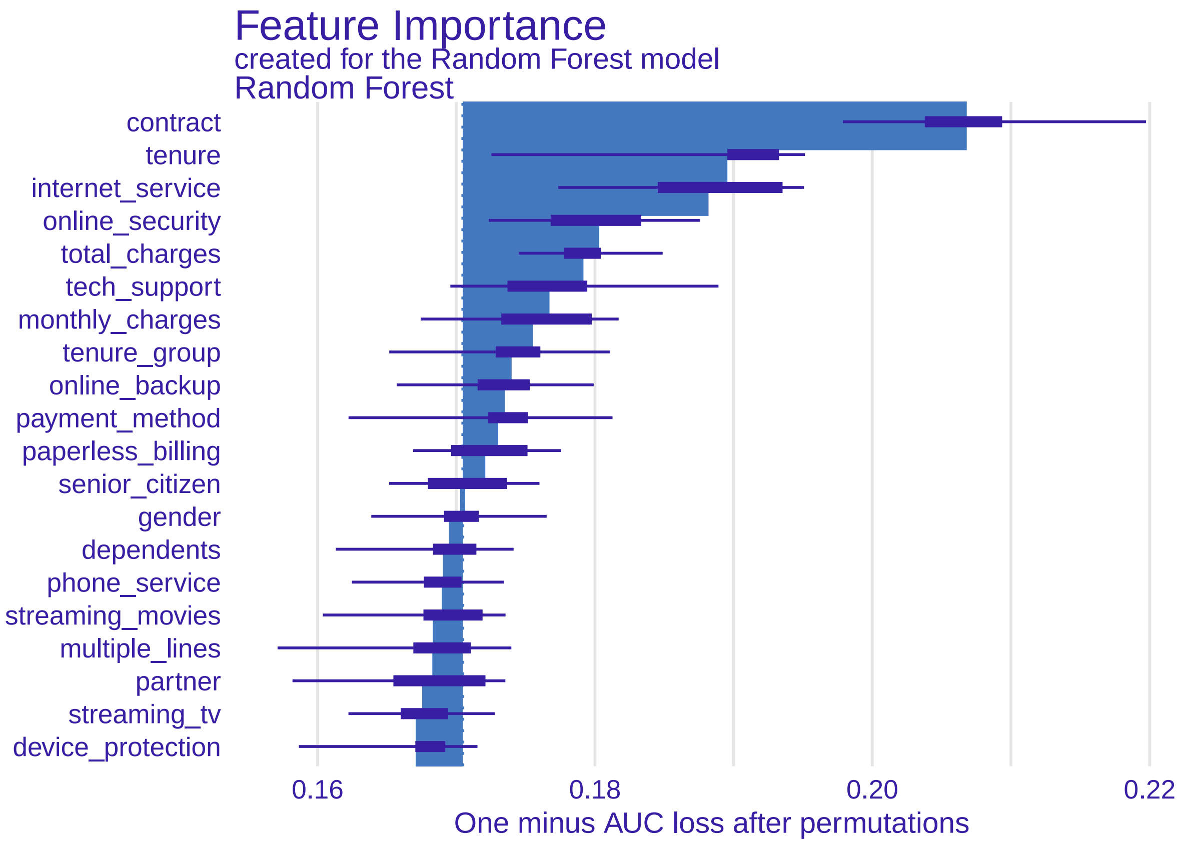
The variable importance plot reveals the most influential features in predicting churn. The plot underlines both customer longevity and billing structure are key drivers of retention. Additionally, service-related variables such as internet_service and tech_support also plays a substantial role. This indicates that the type and quality of services subscribed to may impact a customer's decision to stay. These findings can inform targeted retention strategies, particularly for customers with short tenure, high charges, and flexible contracts.

6. Explainability with DALEX

```r
# Recode churn variable as numeric (1 = Yes, 0 = No)
y_numeric <- ifelse(test$churn == "Yes", 1, 0)

# Create explainer
explainer <- DALEX::explain(
  model = rf_model,
  data = test[, -which(names(test) == "churn")],
  y = y_numeric,
  label = "Random Forest"
)
```

```
## Preparation of a new explainer is initiated
##   -> model label       :  Random Forest
##   -> data              :  1405  rows  20  cols
##   -> data              :  tibble converted into a data.frame
##   -> target variable   :  1405  values
##   -> predict function  :  yhat.randomForest  will be used (  default  )
##   -> predicted values  :  No value for predict function target column. (  default  )
##   -> model_info        :  package randomForest , ver. 4.7.1.2 , task classification (  default  )
##   -> predicted values  :  numerical, min =  0 , mean =  0.2671559 , max =  0.996
##   -> residual function :  difference between y and yhat (  default  )
##   -> residuals         :  numerical, min =  -0.942 , mean =  -0.001675445 , max =  0.996
##   A new explainer has been created!
```

```r
# Plot feature importance
plot(model_parts(explainer))
```



To enhance interpretability of the Random Forest Model, the DALEX package is used to generate a feature importance plot based on permutation analysis. The explainer was successfully created after converting the churn outcome variable into a numeric binary format which resolves DALEX's requirement for numeric input. The resulting variable important plot reveals the top predictors of churn. These features show the greatest reduction in AUC when permuted which indicates their strong contribution to model performance. Service-related variables such as tech_support and online_backup also had meaningful but comparatively smaller impact on churn. Overall, the DALEX output reinforces previous findings from model-based variable importance plots and provides clear and quantitative evidence of which customer characteristics most affect risk for churn.

7. Business Insights and Recommendations

Several key insights emerge that can help the company reduce customer churn and improve retention based on the analysis and model outputs.

First, short-term contract customers are most at risk. Customers on month-to-month contracts were significantly more likely to churn compared to those with one or two year contracts. Introducing incentives such as discounted long-term plans, loyalty perks, or early renewal bonuses can encourage more customers to commit to longer-term contracts. Second, tenure and billing are strong churn indicators. Customers with short tenure and high monthly charges are more likely to leave. This suggests that new customers may be especially sensitive to pricing. The company should consider implementing a welcome program that includes onboarding support and introduce pricing tiers or satisfaction check-ins during the first few months. Third, internet service type and online security matter. Customers with DSL internet or lacking security features such as tech support or online security were more prone to churn. Bundling value-added services such as security, backup, or 24/7 tech support into existing plans may improve satisfaction and perceived value. Lastly, demographics play a smaller role. Features such as gender, dependents, and senior citizen status had minimal influence on churn in this dataset, suggesting that behavioral and service-level factors are more impactful for predicting retention.

Here are a list of actionable recommendations to address the insights. Retention targeting that uses the trained model to score current customers and identify high-risk segments for proactive outreach. Contract restructing should be used to offer flexible upgrade paths and loyalty discounts for customers currently on month-to month plans. Onboarding support would be beneficial when launching early engagement campaigns within the first 3 months of a new customer's tenure. Service bundling to promote bundled tech support and security features to enhance perceived value and reduce churn for internet service customers. Lastly, monitor key indicators to integrate top predictive features such as tenure, contract, and billing into a real-time dashboard to track churn risk trends.

By acting on these insights, the company can move from reactive to proactive retention strategies and ultimately strengthen its long-term customer relationships.

8. Conclusions and Next Steps

This project successfully developed and evaluated predictive models to identify customers at risk of churning. The random forest model performed slightly better than logistic regression by achieving strong accuracy and balanced performance. Key drivers of churn included contract type, tenure, billing amounts, and service-related features like tech support and internet security. These findings highlight the importance of early engagement and bundling value-added services to reduce churn.

Moving forward, the model can be deployed to flag high-risk customers in real time and support proactive retention efforts. Additional steps include expanding the feature set with usage or customer service data, testing targeted offers through A/B experiments, and exploring more advanced models like XGBoost. Finally, building a Shiny dashboard could enable stakeholders to interact with churn predictions and insights in an intuitive, real-time format.