Survival Analysis

Nane Mambreyan

2024-05-01

Building AFT Models

Basic information on data and covariates (**possible** predictive variables)

```
data <- read.csv('telco.csv')
#head(data)
print(names(data))
```

```
##  [1] "ID"        "region"    "tenure"    "age"       "marital"   "address"
##  [7] "income"    "ed"        "retire"    "gender"    "voice"     "internet"
## [13] "forward"   "custcat"   "churn"
```

Some initialization for later use.

```
surv_obj <- Surv(time = data$tenure, event = ifelse(data$churn == "Yes", 1, 0))
model_info <- list()
covariates <- names(data)[!names(data) %in% c("ID", "tenure", "churn")]
reg_formula <- as.formula(paste("surv_obj ~", paste(covariates, collapse = " + ")))
```

Create survival regression object specifying all the available the distribution families for the survival model being fitted.

```
for ( dist in names(survreg.distributions)) {

  reg_m <- survreg(reg_formula, data = data, dist = dist)

  model_info[[dist]] <- list(
      distribution = dist,
      summary = summary(reg_m),
      model = reg_m
    )
}
```

Comparing AFT Models

Comparison criteria - Log-likelihood: Higher log-likelihood values indicate better model fit to the data. - Model fit statistics: A significant chi-squared value suggests that the predictors collectively contribute to the model's fit. - Convergence: A higher number of iterations required for convergence may indicate convergence issues or difficulty in optimizing the model parameters. - Predictive performance: AIC (Akaike Information Criterion), or BIC (Bayesian Information Criterion) measures provide insights into the models' ability to generalize to new data while penalizing model complexity. - Significant features: significant features with p < 0.05 indicate that the predictor contributes to the model's fit.

```
dframe <- data.frame(Dist=character(), LogLikelihood=numeric(), TestStat_P=numeric(), ConvergenceIter=nu

for (i in names(survreg.distributions)) {
  df <- model_info[[i]]$summary$df
  idf <- model_info[[i]]$summary$idf
  chisq <- model_info[[i]]$summary$chi
  p_value <- pchisq(chisq, df-idf, lower.tail = F)

  significant_features <- names(model_info[[i]]$summary$table[,4][model_info[[i]]$summary$table[,4] < 0
  significant_features_string <- paste(significant_features, collapse = ", ")

  new_row <- data.frame(
               Dist = model_info[[i]]$distribution,
               LogLikelihood=model_info[[i]]$model$loglik[2],
               TestStat_P=p_value,
               ConvergenceIter=model_info[[i]]$model$iter,
               AIC=AIC(model_info[[i]]$model),
               BIC=BIC(model_info[[i]]$model),
               SignFeatures=significant_features_string
               )
  dframe <- rbind(dframe, new_row)
}
dframe[, -ncol(dframe)]
```

```
##            Dist LogLikelihood   TestStat_P ConvergenceIter      AIC      BIC
## 1       extreme     -1571.191 8.737464e-64               7 3182.381 3280.536
## 2      logistic     -1554.948 3.839624e-65               5 3149.896 3248.051
## 3      gaussian     -1547.611 5.275164e-60               5 3135.221 3233.376
## 4       weibull     -1462.172 1.099878e-50               7 2964.343 3062.498
## 5   exponential     -1467.598 1.097498e-48               6 2973.195 3066.442
## 6      rayleigh     -1527.438 6.796588e-79               6 3092.877 3186.124
## 7   loggaussian     -1457.012 3.385549e-51               5 2954.024 3052.179
## 8     lognormal     -1457.012 3.385549e-51               5 2954.024 3052.179
## 9   loglogistic     -1458.103 7.462029e-52               5 2956.206 3054.361
## 10            t     -1562.957 1.455278e-67               5 3165.914 3264.069
```

```
write.csv(dframe, "output.csv", row.names = FALSE)
```

```
# Filter rows where p-value is less than 0.05 to identify significant models
significant_models <- subset(dframe, TestStat_P < 0.05)


# Rank models based on criteria such as log-likelihood, AIC, BIC, etc.
ranked_models <- significant_models[order(significant_models$LogLikelihood, significant_models$AIC, sig


# Choose the model with the highest log-likelihood
best_model_loglik <- ranked_models[which.max(ranked_models$LogLikelihood),]
```

```r
# Choose the model with the lowest AIC
best_model_aic <- ranked_models[which.min(ranked_models$AIC),]


# Choose the model with the lowest BIC
best_model_bic <- ranked_models[which.min(ranked_models$BIC),]


# Choose the model with the fewest convergence iterations
best_model_convergence <- ranked_models[which.min(ranked_models$ConvergenceIter),]


# Choose the model with the most significant features
best_model_significance <- ranked_models[which.max(lengths(strsplit(ranked_models$SignFeatures, ", ")) ]


# Extract distribution names from each best model
best_models <- c(best_model_loglik$Dist,
                 best_model_aic$Dist,
                 best_model_bic$Dist,
                 best_model_convergence$Dist,
                 best_model_significance$Dist)


# Create a frequency table of distribution names
distribution_frequency <- table(best_models)


# Display the frequency table
print(distribution_frequency)
```

```
## best_models
##    gaussian loggaussian           t
##           1           3           1
```

So far, LogGaussian model is the best in terms of 3/5 criteria. The log-Gaussian distribution assumes that the logarithm of the survival time follows a Gaussian (normal) distribution.

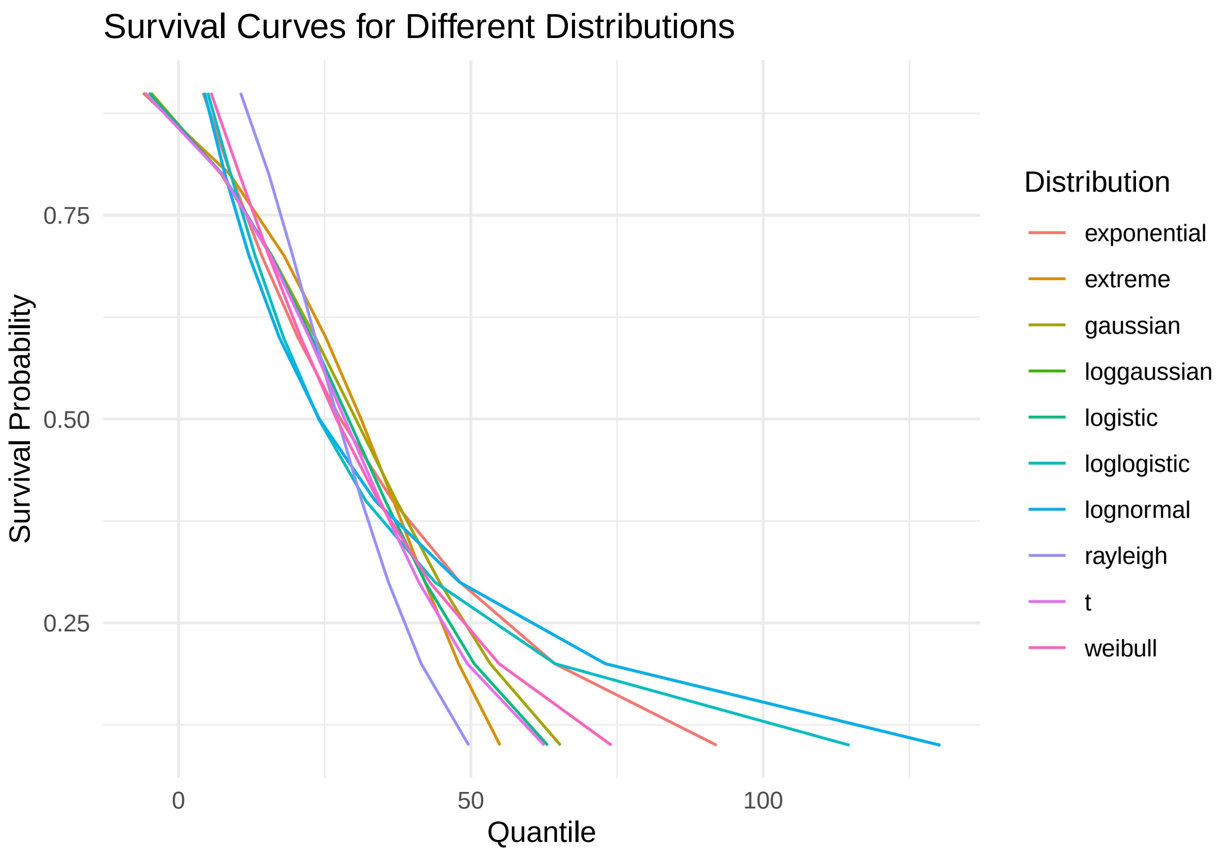
Visualizations

Apart from numeric comparison of statistics of the model, I am now comparing the survival curves visually. For that, I generate a random user and try to see how predicted survival curves differ across considered distribution families.

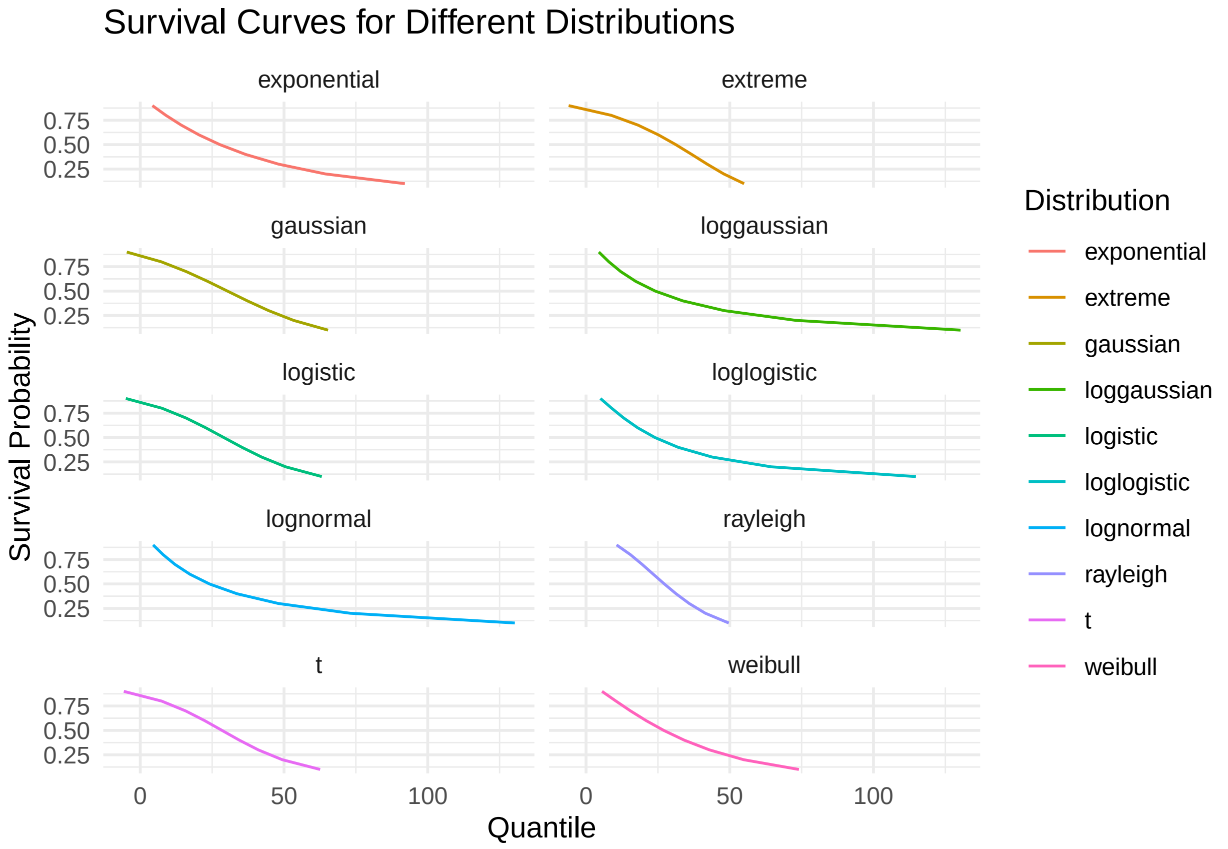
```r
new_data <- data.frame(
  region = sample(data$region, 1),    # Randomly sample one region code from the data
  age = sample(data$age, 1),          # Randomly sample one age value from the data
  marital = sample(data$marital, 1),  # Randomly sample one marital status from the data
  address = sample(data$address, 1),
  income = sample(data$income, 1),
  ed = sample(data$ed, 1),
  retire = sample(data$retire, 1),
  gender = sample(data$gender, 1),
  voice = sample(data$voice, 1),
  internet = sample(data$internet, 1),
  forward = sample(data$forward, 1),
  custcat = sample(data$custcat, 1)
)
```

```r
df_all <- data.frame(Quantile = numeric(),
                     Probability = numeric(),
                     Distribution = character())

# Plot survival curves for each distribution
for (dist in names(model_info)) {

  # Extract the fitted survival model object
  fitted_model <- model_info[[dist]]$model

  probs = seq(.1,.9,length=9)
  pred_quant <- predict(fitted_model, newdata = new_data,
                        type = "quantile", p = 1-probs)

  # Create data frame for current distribution
  df_dist <- data.frame(Quantile = pred_quant,
                        Probability = probs, Distribution = dist)

  # Bind the current distribution data to the overall data frame
  df_all <- rbind(df_all, df_dist)
  }
```

```r
# Plot all survival curves
ggplot(df_all, aes(x = Quantile, y = Probability, color = Distribution)) +
  geom_line() +
  labs(title = "Survival Curves for Different Distributions",
       y = "Survival Probability") +
  theme_minimal()
```



```
# Plot all survival curves facetted by distribution
ggplot(df_all, aes(x = Quantile, y = Probability)) +
  geom_line(aes(color = Distribution)) +
  facet_wrap(~ Distribution, ncol = 2) +
  labs(title = "Survival Curves for Different Distributions",
       y = "Survival Probability") +
  theme_minimal()
```



- Some curves end at lower quantiles (e.g., Quantile $<$ 100) and exhibit a rapid decrease, it suggests that the corresponding distribution assigns lower survival probabilities to individuals over a shorter duration. This could indicate that the model associated with this distribution predicts a higher risk of the event (churn) occurring within a shorter time frame.
- Conversely, other curves extend to higher quantiles (e.g., Quantile $=$ 400+) and decline more gradually, it implies that the associated distribution assigns higher survival probabilities to individuals over a longer duration. This suggests a lower risk of the event occurring within the observed time frame, indicating a more prolonged survival or churn-free period.

The curve produced by the Log-Gaussian model exhibit a smoother and more consistent pattern compared to other distributions. This smoothness indicates that the LogGaussian model captures the underlying variability in the data more effectively, leading to more stable predictions of survival probabilities over time. It predicts higher survival probabilities for longer durations compared to other distributions. This aligns well with the objective of understanding churn behavior over extended periods, as it provides more reliable estimates of survival probabilities for customers with longer lifetimes. Furthermore, this choice is in line with the statistical analysis shown above.

Finalizing model with significant features

```
dframe$SignFeatures[dframe$Dist == "loggaussian"]
```

```
## [1] "(Intercept), age, maritalUnmarried, address, voiceYes, internetYes, custcatE-service, custcatPlu
```

```r
names(data)
```

```
##  [1] "ID"       "region"   "tenure"   "age"      "marital"  "address"
##  [7] "income"   "ed"       "retire"   "gender"   "voice"    "internet"
## [13] "forward"  "custcat"  "churn"
```

```r
predictors <- c("age", "marital", "address", "voice", "internet", "custcat")
reg_formula <- as.formula(paste("surv_obj ~", paste(predictors, collapse = " + ")))
final_m <- survreg(reg_formula, data = data, dist = 'loggaussian')

summary(final_m)
```

```
## 
## Call:
## survreg(formula = reg_formula, data = data, dist = "loggaussian")
##                          Value Std. Error     z       p
## (Intercept)            2.53488    0.24261 10.45 < 2e-16
## age                    0.03683    0.00640  5.75 8.7e-09
## maritalUnmarried      -0.44732    0.11447 -3.91 9.3e-05
## address                0.04282    0.00885  4.84 1.3e-06
## voiceYes              -0.46350    0.16677 -2.78  0.0054
## internetYes           -0.84054    0.13826 -6.08 1.2e-09
## custcatE-service       1.02582    0.16905  6.07 1.3e-09
## custcatPlus service    0.82250    0.16942  4.85 1.2e-06
## custcatTotal service   1.01326    0.20958  4.83 1.3e-06
## Log(scale)             0.28303    0.04602  6.15 7.7e-10
## 
## Scale= 1.33 
## 
## Log Normal distribution
## Loglik(model)= -1462.1   Loglik(intercept only)= -1602.5
##  Chisq= 280.83 on 8 degrees of freedom, p= 4.9e-56 
## Number of Newton-Raphson Iterations: 5 
## n= 1000
```

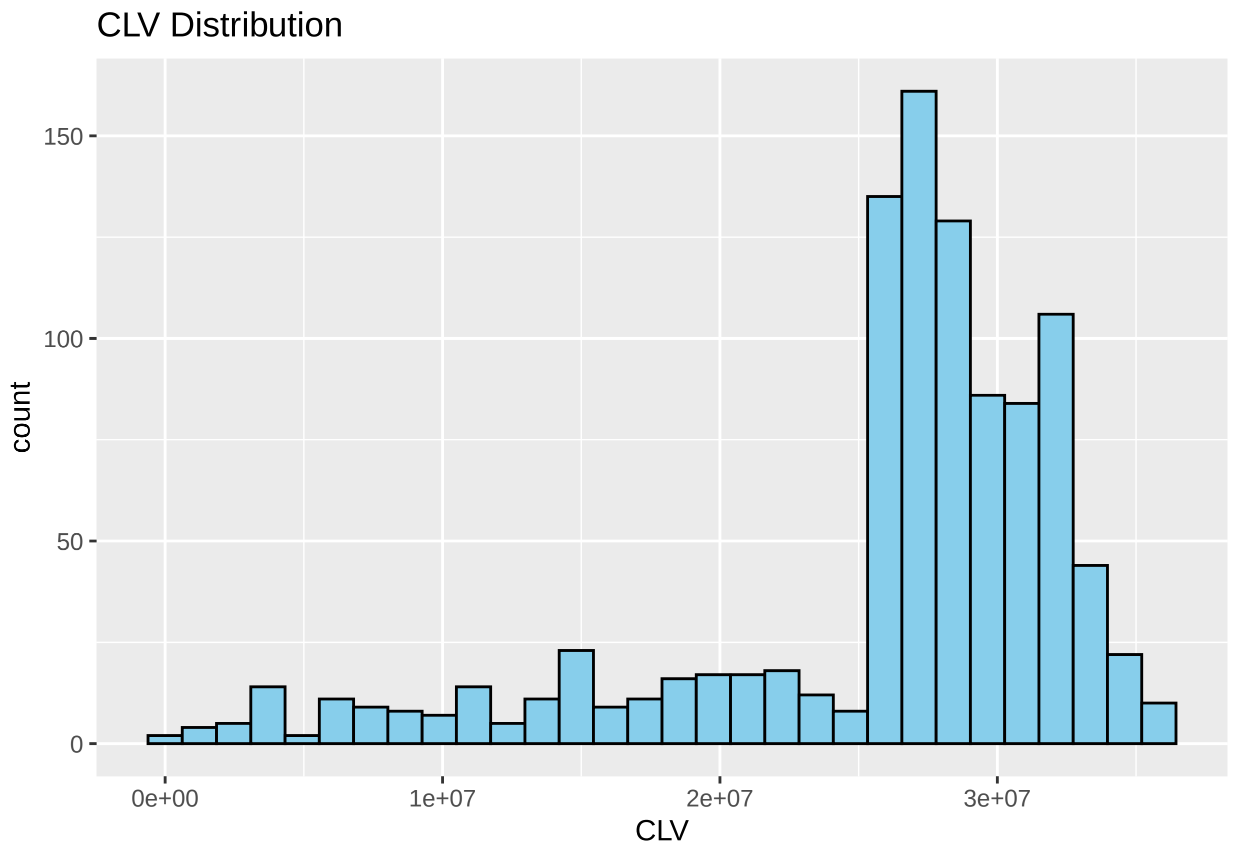
The survival regression model using the reveals several significant predictors influencing churn risk. Those are :

- Age: Older customers tend to have a higher churn risk, as indicated by the positive coefficient (0.03683, $p < 0.001$).
- Marital Status (Unmarried): Unmarried customers are more likely to churn compared to married customers, as suggested by the negative coefficient (-0.44732, $p < 0.001$).
- Address: An increase in the number of years at the current address is associated with a higher churn risk, supported by the positive coefficient (0.04282, $p < 0.001$).
- Voice Service (Yes): Customers using voice services are less likely to churn, as indicated by the negative coefficient (-0.46350, $p = 0.0054$).
- Internet Service (Yes): Similarly, customers with internet service are less likely to churn, supported by the negative coefficient (-0.84054, $p < 0.001$).
- Customer Category (E-service, Plus service, Total service): Customers subscribed to higher-tier service categories (E-service, Plus service, Total service) exhibit lower churn risk compared to other categories, as evidenced by the positive coefficients (ranging from 0.82250 to 1.02582, all with $p < 0.001$).

CLV

```
# Calculate CLV per customer based on the final model
MM <- 1300  # Average monthly margin
r <- 0.1    # Discount rate


# Predict survival probabilities using final_m for individual observations
surv_prob <- predict(final_m, type = "response", newdata = data)

# Discount survival probabilities and calculate CLV for each observation
CLV <- numeric(length(surv_prob))
for (i in seq_along(surv_prob)) {
  CLV[i] <- MM * sum(surv_prob[i:length(surv_prob)] / (1 + r / 12)^(seq(1, length(surv_prob) - i + 1) -
}


# Plot the distribution of CLV
ggplot(data.frame(CLV = CLV), aes(x = CLV)) +
  geom_histogram(fill = "skyblue", color = "black") +
  labs(title = "CLV Distribution")
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```

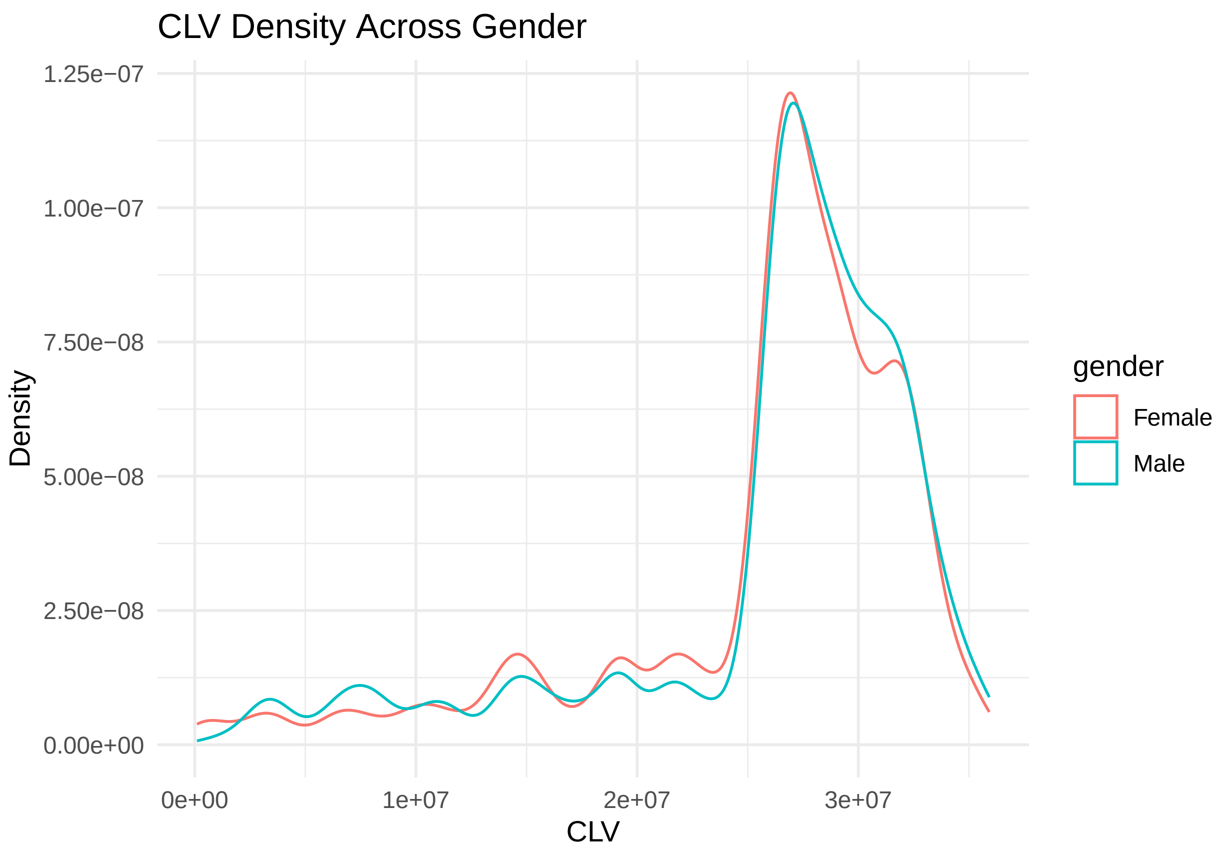


The bars representing the count of customers are notably lower for CLV values less than 25,000,000. However, as CLV increases to around 30,000,000, the count rises to approximately 125-150 customers. This increase suggests a concentration of customers with higher CLV values. Beyond this point, the count starts to decrease again, with a small peak and subsequent decline. This pattern indicates that while a significant portion of customers falls within the lower CLV range, there is a smaller but notable segment of customers with substantially higher CLV. Understanding and targeting this high-value segment can be crucial for maximizing overall profitability and optimizing retention strategies. For that reason, I am going to further explore clv across different segments to find out which segment is responsible for picks of overall clv and define those segments.

```
# Function to generate density plot for CLV across different segments
plot_CLV_density <- function(data, segment_var, title) {
  ggplot(data, aes(x = CLV, color = !!sym(segment_var))) +
    geom_density() +
    labs(title = title, x = "CLV", y = "Density") +
    theme_minimal()
}

# Combine CLV with original data
data_with_CLV <- cbind(data, CLV)

summary(data_with_CLV$CLV)
```

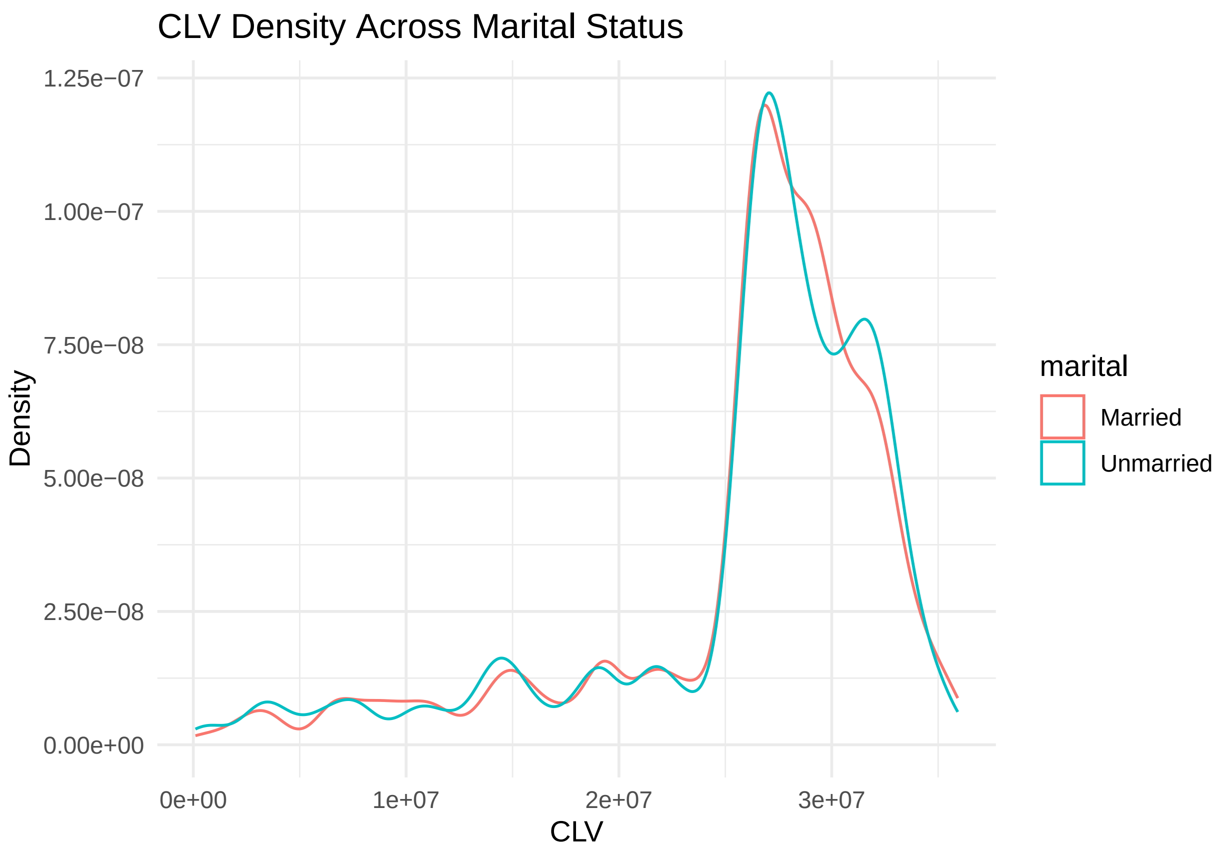
```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##   100288 25838574 27596080 25954175 30446478 35922317
```

```
# Generate density plots for different segments
plot_CLV_density(data_with_CLV, "gender", "CLV Density Across Gender")
```



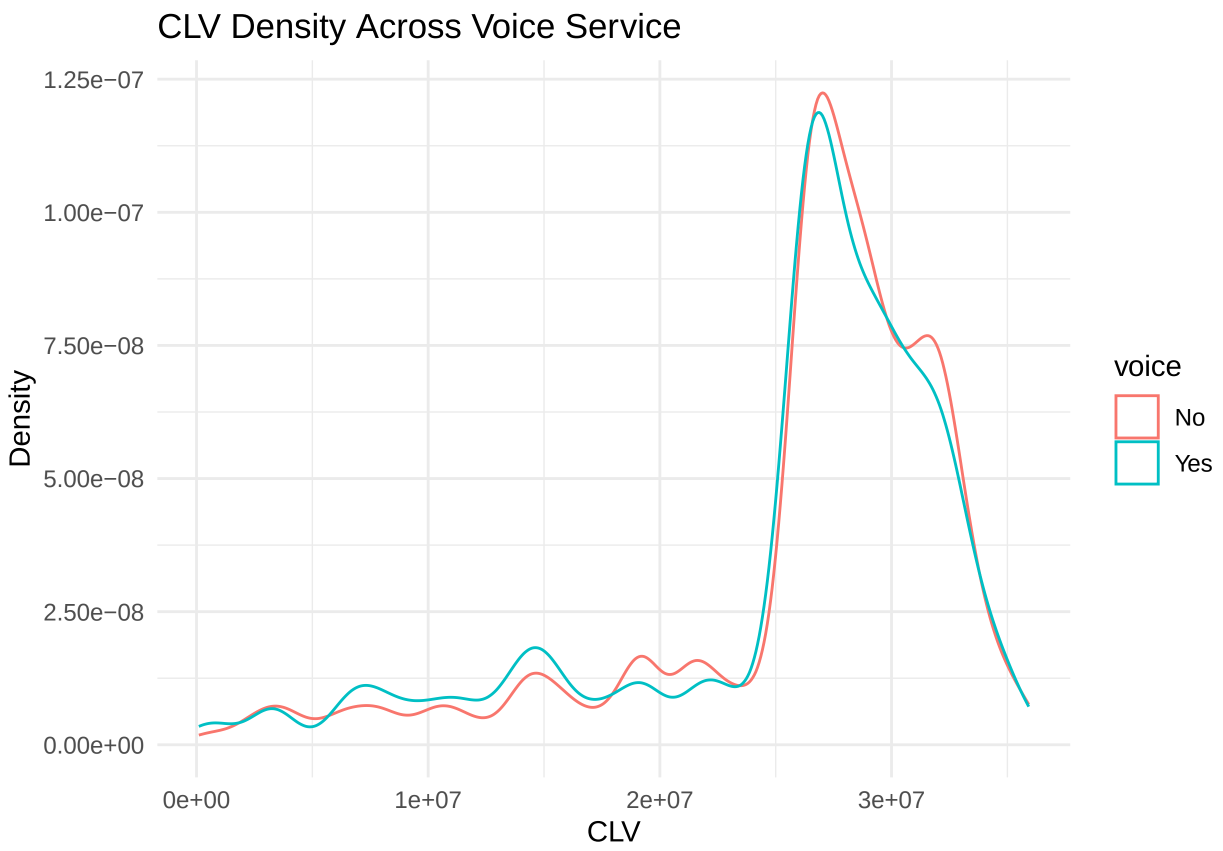
CLV for Male and Female customers is mostly the same, but after CLV=3e+07 there is a difference with higher density for Male. As we have seen in overall CLV plot, there was a pick in the very same place. The initial assumption is that even though the number of customers decrease after CLV=3e+07, still mmale customers contribute to increase(pick) we observe after CLV > 3e+07.

```
plot_CLV_density(data_with_CLV, "marital", "CLV Density Across Marital Status")
```



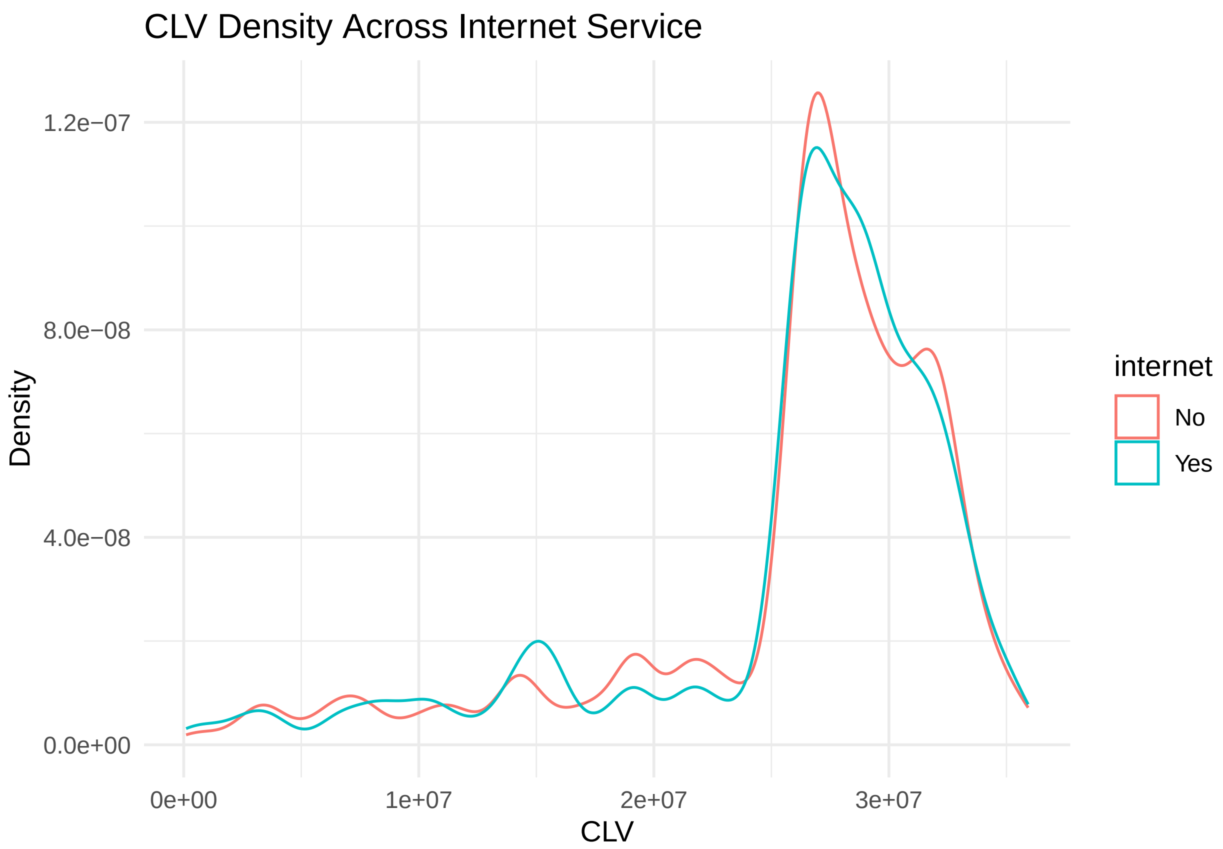
Similarly, customers contributing to already mentioned pick in clv data are likely to be unmarried, hence the pick in blue line after CLV=3e+07.

```
plot_CLV_density(data_with_CLV, "voice", "CLV Density Across Voice Service")
```



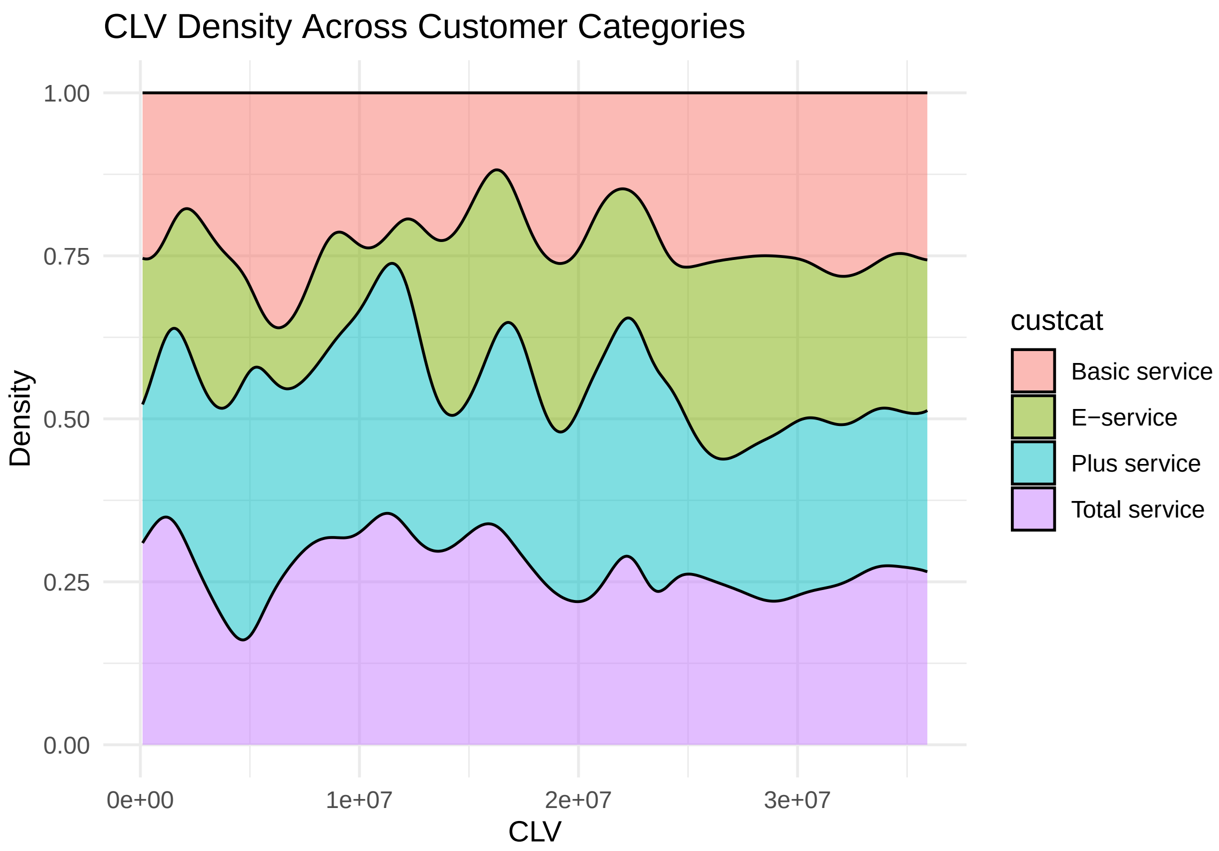
Despite similar density curves among custoers who do and don't use voice services, the pick after CLV=3e+07 is now demonstrated by customers who don't use voice services.

```
plot_CLV_density(data_with_CLV, "internet", "CLV Density Across Internet Service")
```



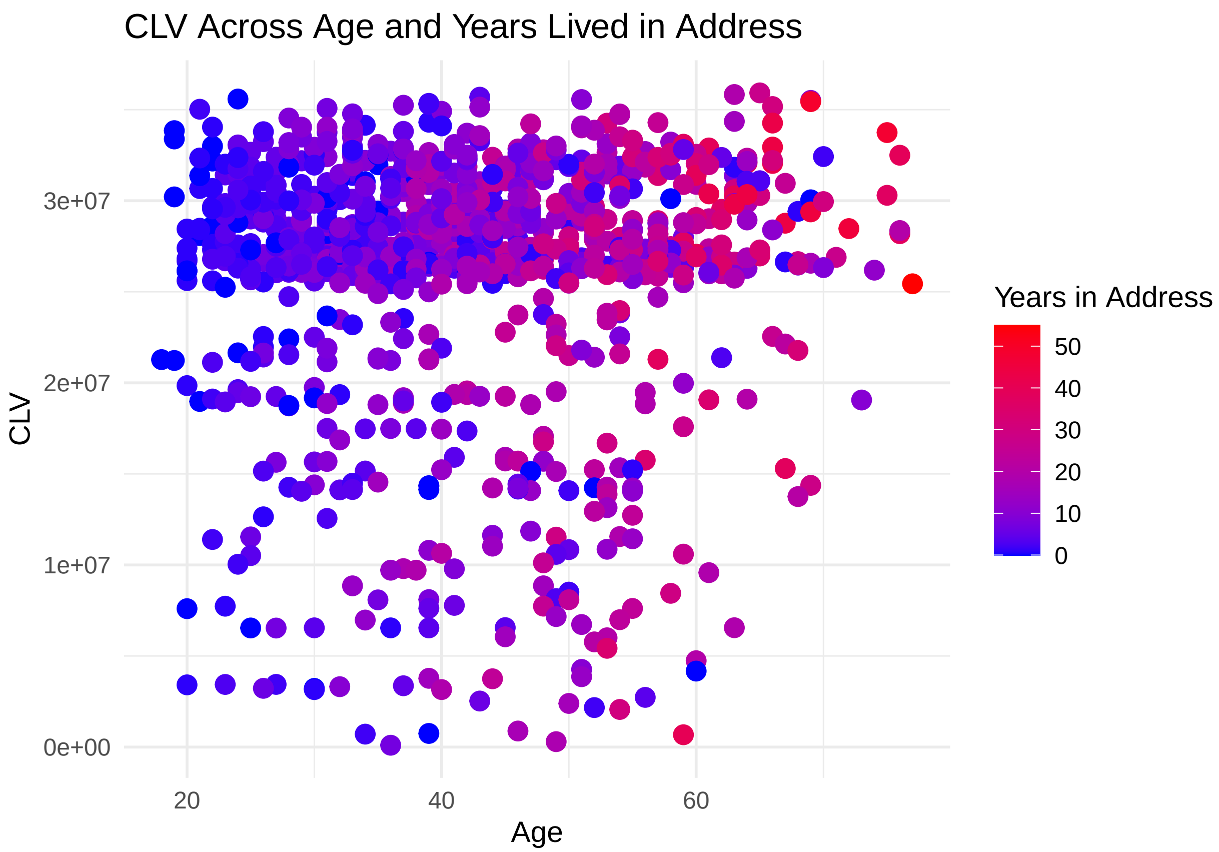
Lastly, those are customers who don't use internet services.

```r
# using a density bar plot
ggplot(data_with_CLV, aes(x = CLV, fill = custcat)) +
  geom_density(alpha = 0.5, position = "fill") +
  labs(title = "CLV Density Across Customer Categories", x = "CLV", y = "Density") +
  theme_minimal()
```



Customers using Basic and Total services behave similarly in terms of their CLV distribution. In the plot, the color representing E-service customers exhibits the lowest density compared to other customer categories. This suggests that the density of CLV values for E-service customers is relatively lower across the entire range of CLV values compared to other customer segments. The majority are Total service users, then Basic, Plus and finally E-service.

```r
# Create a scatter plot of CLV across age and address
ggplot(data_with_CLV, aes(x = age, y = CLV, color = address)) +
  geom_point(size = 3) +
  labs(title = "CLV Across Age and Years Lived in Address",
       x = "Age", y = "CLV") +
  scale_color_gradient(low = "blue", high = "red", name = "Years in Address") +
  theme_minimal()
```



Customers with high values of CLV are most likely 20-40, who live in the particular address up to 20 years.

Based on these observations above, the most valuable segments are: - Male customers, especially those who are unmarried, don't use voice or internet services, and belong to the Total service category.

- Younger customers aged 20-40 who have lived in their current address for a considerable duration.

(the definition of being "valuable" refers to the characteristics or behaviors of customers that lead to a higher contribution to the company's profitability over their lifetime as a customer)

Retenation

Assuming the data is population:

```
# Calculate Total CLV for the Entire Population
total_CLV <- sum(data_with_CLV$CLV)

# Get linear predictors
linear_pred <- predict(final_m, type = "lp", newdata = data)

# Extract shape parameter
shape_param <- final_m$scale

# Compute survival probabilities
surv_prob <- plogis(linear_pred / shape_param)
summary(surv_prob)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.7645  0.9439  0.9674  0.9596  0.9823  0.9976
```

Since 1st Qu=0.9439, assume he minimum survival probability below which customers are considered at risk is 0.94 and find out the number of at risk customers:

```
# Identify the Number of At-Risk Subscribers.
at_risk_threshold <- 0.94

at_risk_customers <- sum(surv_prob < at_risk_threshold)
```

Also let's assume that the cost of retaining each at-risk subscriber for one year is 100 and Calculate the Annual Retention Budget.

```
retention_cost_per_customer <- 100
annual_retention_budget <- at_risk_customers * retention_cost_per_customer

# Output the results
cat("Total CLV for the entire population:", total_CLV, "\n")
```

```
## Total CLV for the entire population: 25954174931
```

```
cat("Number of at-risk subscribers within a year:", at_risk_customers, "\n")
```

```
## Number of at-risk subscribers within a year: 216
```

```
cat("Annual retention budget:", annual_retention_budget, "\n")
```

```
## Annual retention budget: 21600
```

The retention budget can be allocated to:

- Encourage customers to provide feedback on their experiences and reasons for churn. Use this feedback to identify areas for improvement and implement changes to address customer concerns effectively.
- Identify opportunities to cross-sell or up-sell additional products or services to existing customers. Offer personalized recommendations based on their previous purchase history or preferences to increase customer lifetime value.