Marketing Analytics

Elen Sukiasyan

2024-04-30

```r
# Encode 'churn' variable: If 'churn' equals 'Yes', encode as 1, else encode as 0
data$churn<-ifelse(data$churn=='Yes',1,0)
# Categorical Variable Conversion
data$marital <- as.factor(data$marital)
data$ed <- as.factor(data$ed)
data$retire <- as.factor(data$retire)
data$gender <- as.factor(data$gender)
data$voice <- as.factor(data$voice)
data$internet <- as.factor(data$internet)
data$forward <- as.factor(data$forward)
data$custcat <- as.factor(data$custcat)
```

```r
# Creating survival object
surv_obj <- Surv(time = data$tenure, event = data$churn)

# Defining a function to fit accelerated failure time (AFT) model
fit_aft_model <- function(dist) {
  # Fitting AFT model using survreg function
  model <- survreg(
    surv_obj ~ age + marital + address + income + ed + retire + gender + voice + internet + forward + cu
    data = data,
    dist = dist
  )
  return(model)
}

#Get Available Distributions:
distributions <- names(survreg.distributions)

#Fit AFT Models with All Available Distributions
models <- lapply(distributions, fit_aft_model)
```

```r
new_data <- data.frame(
  age = mean(data$age),
  marital = as.factor(names(which.max(table(data$marital)))),
  address = mean(data$address),
  income = mean(data$income),
  ed = as.factor(names(which.max(table(data$ed)))),
  retire = as.factor(names(which.max(table(data$retire)))),
  gender = as.factor(names(which.max(table(data$gender)))),
  voice = as.factor(names(which.max(table(data$voice)))),
```

```r
  internet = as.factor(names(which.max(table(data$internet)))),
  forward = as.factor(names(which.max(table(data$forward)))),
  custcat = as.factor(names(which.max(table(data$custcat)))),
  tenure = median(data$tenure)
)
```

```r
# Define a function to generate survival curves
survival_curves <- function(models, dist) {
  probs <- seq(0.1, 0.9, length = 9)
  all_data <- data.frame()

  # Iterate through models and add survival data to the dataframe
  for (i in seq_along(models)) {
    # Predict survival probabilities using the fitted model
    pred_surv <- predict(models[[i]], type = "quantile", p = 1 - probs, newdata = new_data)
    data <- data.frame(Time = pred_surv, Probabilities = probs, Distribution = dist[i])
    all_data <- rbind(all_data, data)
  }
  return(all_data)
}

survival_curve<-survival_curves(models, distributions)
survival_curve
```

```
##          Time Probabilities Distribution
## 1   98.51839           0.1      extreme
## 2   91.41844           0.2      extreme
## 3   85.66433           0.3      extreme
## 4   80.25139           0.4      extreme
## 5   74.71866           0.5      extreme
## 6   68.66806           0.6      extreme
## 7   61.54718           0.7      extreme
## 8   52.24948           0.8      extreme
## 9   37.37294           0.9      extreme
## 10 107.61308           0.1     logistic
## 11  94.97419           0.2     logistic
## 12  86.57356           0.3     logistic
## 13  79.68730           0.4     logistic
## 14  73.36785           0.5     logistic
## 15  67.04841           0.6     logistic
## 16  60.16215           0.7     logistic
## 17  51.76152           0.8     logistic
## 18  39.12263           0.9     logistic
## 19 109.32823           0.1     gaussian
## 20  97.29046           0.2     gaussian
## 21  88.61039           0.3     gaussian
## 22  81.19358           0.4     gaussian
## 23  74.26127           0.5     gaussian
## 24  67.32897           0.6     gaussian
## 25  59.91216           0.7     gaussian
## 26  51.23208           0.8     gaussian
## 27  39.19432           0.9     gaussian
## 28 371.86412           0.1      weibull
```

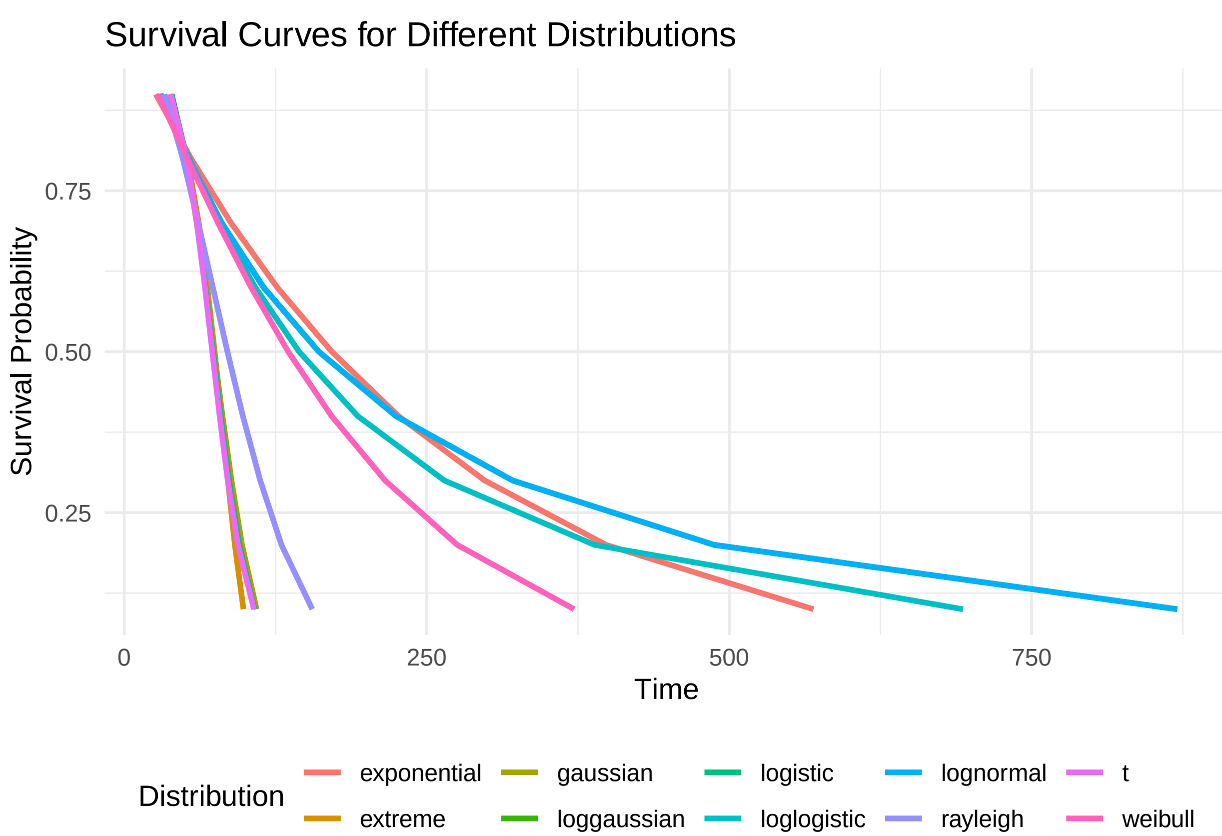
```
## 29 275.17842           0.2      weibull
## 30 215.59207           0.3      weibull
## 31 171.37008           0.4      weibull
## 32 135.52854           0.5      weibull
## 33 104.85481           0.6      weibull
## 34  77.52341           0.7      weibull
## 35  52.26176           0.8      weibull
## 36  27.80875           0.9      weibull
## 37 569.68079           0.1  exponential
## 38 398.18978           0.2  exponential
## 39 297.87398           0.3  exponential
## 40 226.69878           0.4  exponential
## 41 171.49101           0.5  exponential
## 42 126.38297           0.6  exponential
## 43  88.24467           0.7  exponential
## 44  55.20777           0.8  exponential
## 45  26.06716           0.9  exponential
## 46 155.39566           0.1     rayleigh
## 47 129.91765           0.2     rayleigh
## 48 112.36715           0.3     rayleigh
## 49  98.02744           0.4     rayleigh
## 50  85.25969           0.5     rayleigh
## 51  73.19263           0.6     rayleigh
## 52  61.15998           0.7     rayleigh
## 53  48.37525           0.8     rayleigh
## 54  33.24069           0.9     rayleigh
## 55 870.39770           0.1  loggaussian
## 56 487.47787           0.2  loggaussian
## 57 320.93435           0.3  loggaussian
## 58 224.54225           0.4  loggaussian
## 59 160.80994           0.5  loggaussian
## 60 115.16690           0.6  loggaussian
## 61  80.57671           0.7  loggaussian
## 62  53.04822           0.8  loggaussian
## 63  29.71037           0.9  loggaussian
## 64 870.39770           0.1    lognormal
## 65 487.47787           0.2    lognormal
## 66 320.93435           0.3    lognormal
## 67 224.54225           0.4    lognormal
## 68 160.80994           0.5    lognormal
## 69 115.16690           0.6    lognormal
## 70  80.57671           0.7    lognormal
## 71  53.04822           0.8    lognormal
## 72  29.71037           0.9    lognormal
## 73 693.22288           0.1  loglogistic
## 74 388.63569           0.2  loglogistic
## 75 264.54065           0.3  loglogistic
## 76 192.99940           0.4  loglogistic
## 77 144.50765           0.5  loglogistic
## 78 108.19962           0.6  loglogistic
## 79  78.93857           0.7  loglogistic
## 80  53.73274           0.8  loglogistic
## 81  30.12373           0.9  loglogistic
## 82 107.18571           0.1            t
```

```
## 83   93.89165          0.2          t
## 84   85.53428          0.3          t
## 85   78.84671          0.4          t
## 86   72.76980          0.5          t
## 87   66.69289          0.6          t
## 88   60.00533          0.7          t
## 89   51.64795          0.8          t
## 90   38.35389          0.9          t
```

```r
plt <- ggplot(data = survival_curve, aes(x = Time, y = Probabilities, color = Distribution)) +
    geom_line(size = 1) +
    theme_minimal() +
    labs(x = "Time", y = "Survival Probability", title = "Survival Curves for Different Distributions")
    theme(legend.position = "bottom") +
    geom_abline(intercept = 0, slope = 0)
```

```
## Warning: Using 'size' aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use 'linewidth' instead.
## This warning is displayed once every 8 hours.
## Call 'lifecycle::last_lifecycle_warnings()' to see where this warning was
## generated.
```

```r
print(plt)
```



From the results it is obvious that the best survival curve is the lognormal one.

To improve model selection, we can consider additional statistical measures like the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). Lower AIC and BIC values indicate better model performance.

```r
# Create an empty dataframe to store decision data
decision_data <- data.frame()
for (i in seq_along(models)) {
  # Extract log likelihood, AIC, and BIC values for each model
  loglikelihood <- models[[i]]$loglik
  aic <- AIC(models[[i]])
  bic <- BIC(models[[i]])
  data_aic_bic <- data.frame(Loglikelihood = loglikelihood, AIC = aic, BIC = bic, Distribution = distri
  # Append data to decision_data dataframe
  decision_data <- rbind(decision_data, data_aic_bic)
}

min_bic <- min(decision_data$BIC)
min_aic <- min(decision_data$AIC)

decision_data
```

```
##    Loglikelihood      AIC      BIC Distribution
## 1      -1747.194 3181.130 3269.470      extreme
## 2      -1572.565 3181.130 3269.470      extreme
## 3      -1734.223 3149.168 3237.507     logistic
## 4      -1556.584 3149.168 3237.507     logistic
## 5      -1714.485 3133.226 3221.565     gaussian
## 6      -1548.613 3133.226 3221.565     gaussian
## 7      -1606.431 2962.382 3050.721      weibull
## 8      -1463.191 2962.382 3050.721      weibull
## 9      -1606.980 2971.078 3054.510  exponential
## 10     -1468.539 2971.078 3054.510  exponential
## 11     -1739.723 3091.719 3175.151     rayleigh
## 12     -1528.859 3091.719 3175.151     rayleigh
## 13     -1602.518 2951.151 3039.491  loggaussian
## 14     -1457.576 2951.151 3039.491  loggaussian
## 15     -1602.518 2951.151 3039.491    lognormal
## 16     -1457.576 2951.151 3039.491    lognormal
## 17     -1605.208 2953.691 3042.030  loglogistic
## 18     -1458.845 2953.691 3042.030  loglogistic
## 19     -1748.062 3165.973 3254.312            t
## 20     -1564.986 3165.973 3254.312            t
```

In our analysis, we observe that the model with a lognormal distribution yields the minimum AIC (2951.151) and BIC (3039.491). Therefore, based on these criteria, we again select the model with a lognormal distribution as our final choice.

#Feature Signnificance Then which features are influential for the model. Initially, we'll incorporate all available features into the model and evaluate their significance. (Alpha = 0.1)

```r
# Fitting a model with all features and examining their significance
feauture_testing_model <- survreg(surv_obj ~ age + marital + address + income + ed + retire + gender + 
summary_results <- summary(feauture_testing_model)
summary_results
```

```
##
## Call:
## survreg(formula = surv_obj ~ age + marital + address + income +
##     ed + retire + gender + voice + internet + forward + custcat,
##     data = data, dist = "lognormal")
##                                    Value Std. Error     z       p
## (Intercept)                     2.338870   0.281279  8.32 < 2e-16
## age                             0.032795   0.007247  4.53 6.0e-06
## maritalUnmarried               -0.459424   0.114720 -4.00 6.2e-05
## address                         0.042153   0.008882  4.75 2.1e-06
## income                          0.001387   0.000918  1.51   0.131
## edDid not complete high school  0.379168   0.200877  1.89   0.059
## edHigh school degree            0.315976   0.162495  1.94   0.052
## edPost-undergraduate degree    -0.019815   0.222366 -0.09   0.929
## edSome college                  0.285140   0.164846  1.73   0.084
## retireYes                       0.031781   0.444440  0.07   0.943
## genderMale                      0.051108   0.114237  0.45   0.655
## voiceYes                       -0.424370   0.168551 -2.52   0.012
## internetYes                    -0.758597   0.142814 -5.31 1.1e-07
## forwardYes                     -0.196353   0.179535 -1.09   0.274
## custcatE-service                1.059925   0.170244  6.23 4.8e-10
## custcatPlus service             0.923373   0.214843  4.30 1.7e-05
## custcatTotal service            1.182016   0.249736  4.73 2.2e-06
## Log(scale)                      0.275904   0.045997  6.00 2.0e-09
##
## Scale= 1.32
##
## Log Normal distribution
## Loglik(model)= -1457.6   Loglik(intercept only)= -1602.5
##  Chisq= 289.88 on 16 degrees of freedom, p= 3.2e-52
## Number of Newton-Raphson Iterations: 5
## n= 1000
```

```
# Checking features with p-values less than 0.1
significant_features <- summary_results$table[, 4] < 0.10
significant_features
```

```
##                    (Intercept)                            age
##                           TRUE                           TRUE
##               maritalUnmarried                        address
##                           TRUE                           TRUE
##                         income edDid not complete high school
##                          FALSE                           TRUE
##           edHigh school degree    edPost-undergraduate degree
##                           TRUE                          FALSE
##                 edSome college                      retireYes
##                           TRUE                          FALSE
##                     genderMale                       voiceYes
##                          FALSE                           TRUE
##                    internetYes                     forwardYes
##                           TRUE                          FALSE
##               custcatE-service            custcatPlus service
##                           TRUE                           TRUE
##           custcatTotal service                     Log(scale)
```

```
##                                TRUE                               TRUE
```

As some features had p values > 1, hence we need to exclude them from the model.

```
# Building the final model with selected features
final_model <- survreg(surv_obj ~ age + marital + address + ed + voice + internet + custcat, data = data
summary_final <- summary(final_model)
summary_final
```

```
##
## Call:
## survreg(formula = surv_obj ~ age + marital + address + ed + voice +
##     internet + custcat, data = data, dist = "lognormal")
##                                    Value Std. Error     z       p
## (Intercept)                      2.30040    0.26658  8.63 < 2e-16
## age                              0.03672    0.00642  5.72 1.1e-08
## maritalUnmarried                -0.45111    0.11455 -3.94 8.2e-05
## address                          0.04228    0.00884  4.78 1.7e-06
## edDid not complete high school   0.32318    0.19886  1.63    0.10
## edHigh school degree             0.28346    0.16202  1.75    0.08
## edPost-undergraduate degree     -0.00704    0.22287 -0.03    0.97
## edSome college                   0.26066    0.16435  1.59    0.11
## voiceYes                        -0.43112    0.16788 -2.57    0.01
## internetYes                     -0.76976    0.14268 -5.40 6.8e-08
## custcatE-service                 1.06378    0.17072  6.23 4.6e-10
## custcatPlus service              0.80252    0.16934  4.74 2.1e-06
## custcatTotal service             1.05892    0.21074  5.02 5.0e-07
## Log(scale)                       0.28004    0.04601  6.09 1.1e-09
##
## Scale= 1.32
##
## Log Normal distribution
## Loglik(model)= -1459.7   Loglik(intercept only)= -1602.5
##  Chisq= 285.71 on 12 degrees of freedom, p= 4.7e-54
## Number of Newton-Raphson Iterations: 5
## n= 1000
```

```
exp_coefs <- exp(coef(final_model))
exp_coefs
```

```
##                    (Intercept)                            age
##                      9.9781819                      1.0374031
##               maritalUnmarried                        address
##                      0.6369217                      1.0431842
## edDid not complete high school           edHigh school degree
##                      1.3815083                      1.3277135
##    edPost-undergraduate degree                 edSome college
##                      0.9929849                      1.2977840
##                       voiceYes                    internetYes
##                      0.6497821                      0.4631241
##               custcatE-service            custcatPlus service
##                      2.8972934                      2.2311654
##           custcatTotal service
##                      2.8832641
```

.For each additional year of a customer's age, there's a 3% increase in hazard. .Unmarried individuals have roughly a 36% lower hazard compared to married ones. .Education levels are compared to the "College Degree" target group: .Individuals who did not complete high school have a 38% higher hazard. .Individuals with a high school education have a 32% higher hazard. .Individuals with a post-Undergrad degree have approximately a 1% lower hazard. .Individuals who did some college have a 29% higher hazard. .Having "Voice yes" results in approximately a 35% lower hazard compared to the "Voice No" group. .Having "Internet yes" leads to roughly a 55% lower hazard compared to the "Internet No" group. .Customer categories are compared to the "Basic service" target group: ."E-service" customers have a 189% higher hazard. ."Plus Service" customers have a 123% higher hazard. ."Total Service" customers have a 188% higher hazard.

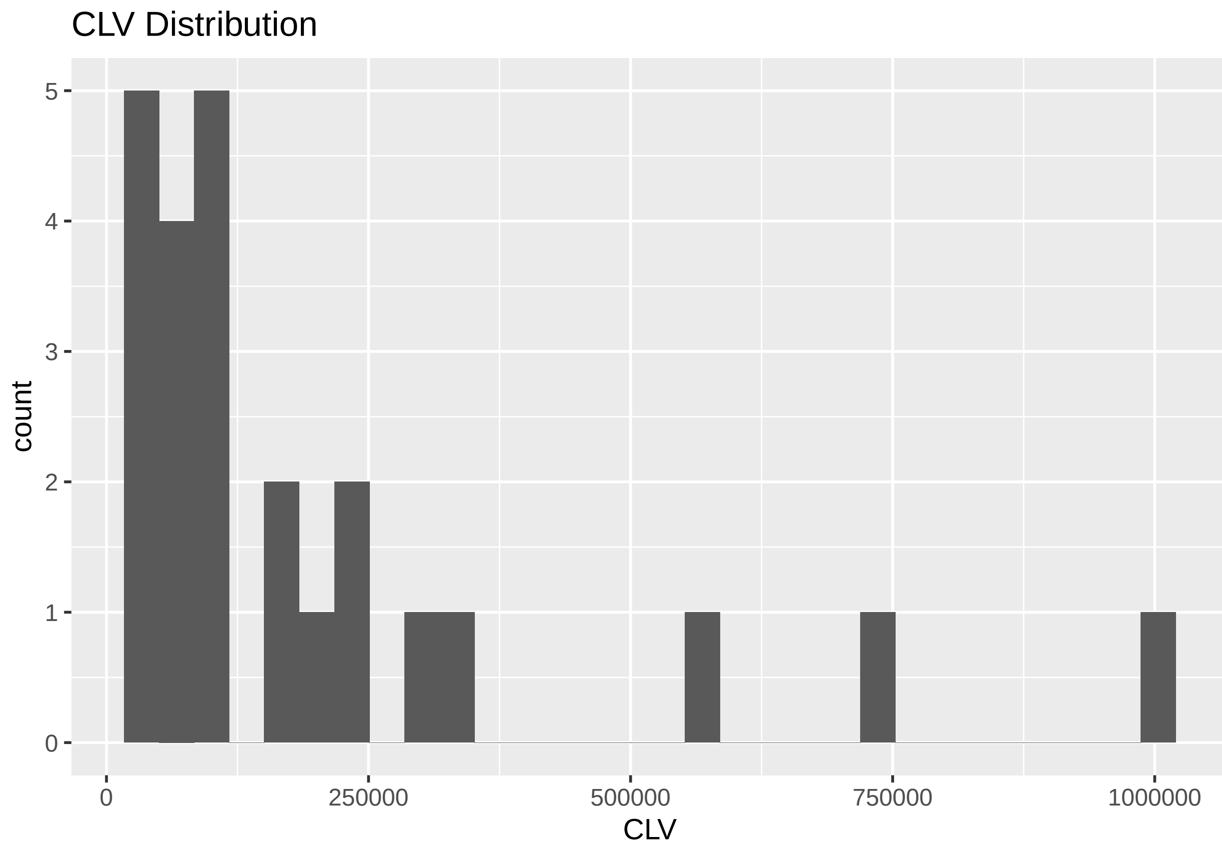
```
new_data <- data.frame(
  age = mean(data$age),
  marital = as.factor(names(which.max(table(data$marital)))),
  address = mean(data$address),
  income = mean(data$income),
  ed = as.factor(names(which.max(table(data$ed)))),
  retire = as.factor(names(which.max(table(data$retire)))),
  gender = as.factor(names(which.max(table(data$gender)))),
  voice = as.factor(names(which.max(table(data$voice)))),
  internet = as.factor(names(which.max(table(data$internet)))),
  forward = as.factor(names(which.max(table(data$forward)))),
  custcat = as.factor(names(which.max(table(data$custcat)))),
  tenure = median(data$tenure) # Median tenure value for prediction
)
```

```
# Making predictions using the final model
predictions <- predict(final_model, type = "response", newdata = data)

# Creating a dataframe with predictions
predictions_data <- data.frame(predictions)

# Adjusting predictions for CLV calculation
sequence <- seq(1, length(colnames(predictions_data)), 1)
MM <- 1300  # Monthly margin assumption
r <- 0.1  # Discount rate assumption
for (num in sequence) {
  predictions_data[, num] <- predictions_data[, num] / (1 + r / 12) ^ (sequence[num] - 1)
}

# Calculating CLV
predictions_data$CLV <- MM * rowSums(predictions_data)

# Summary statistics of CLV
summary(predictions_data$CLV)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6531   55138  117200  246071  266528 3843252
```

```
# Plotting CLV distribution
examine_data <- head(predictions_data, 24)
ggplot(examine_data, aes(x = CLV)) +
  labs(title = "CLV Distribution") +
  geom_histogram()
```

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



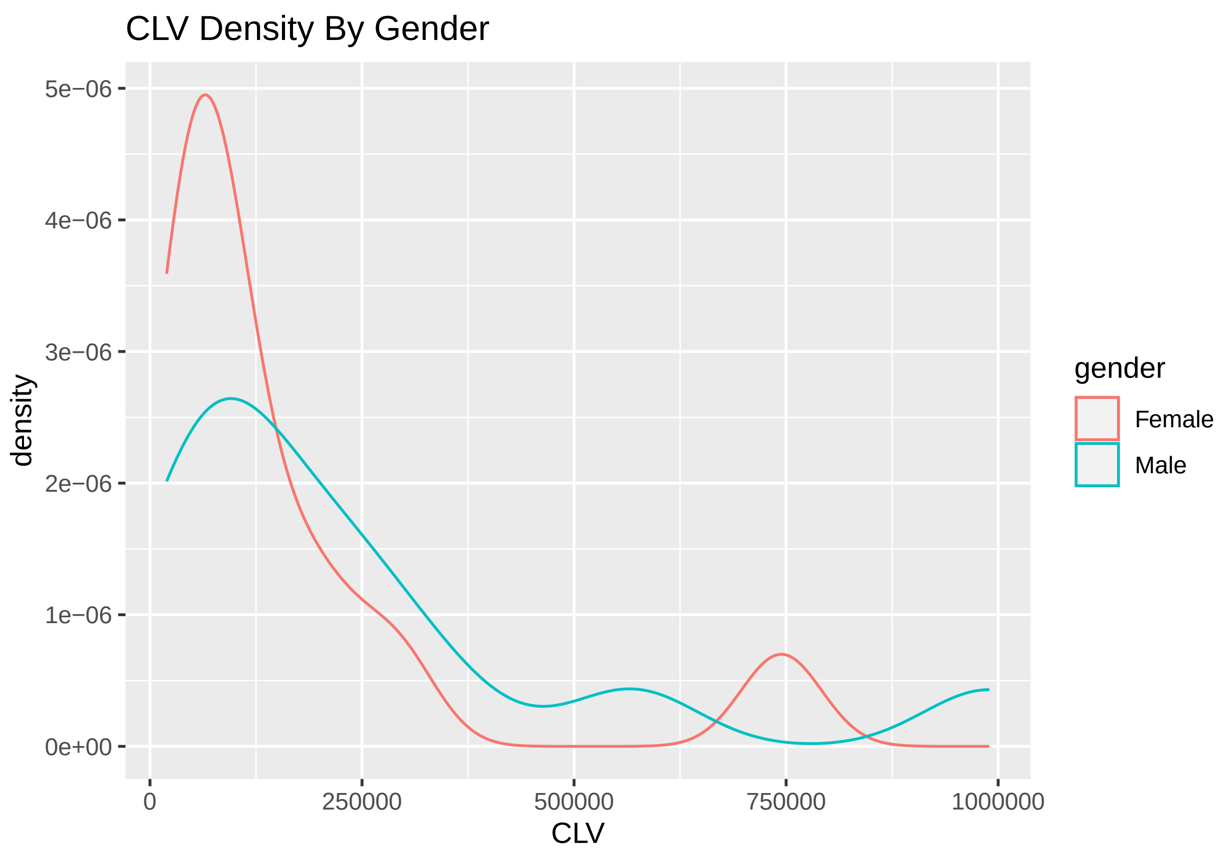
##CLV Comparison: Female and Male customers

```r
# Adding CLV to telco dataframe
data$CLV <- predictions_data$CLV

# Subset of data for examination
examine_data <- head(data, 24)
examine_data
```

```
##    ID region tenure age   marital address income                        ed
## 1   1 Zone 2     13  44   Married       9     64            College degree
## 2   2 Zone 3     11  33   Married       7    136    Post-undergraduate degree
## 3   3 Zone 3     68  52   Married      24    116 Did not complete high school
## 4   4 Zone 2     33  33 Unmarried      12     33          High school degree
## 5   5 Zone 2     23  30   Married       9     30 Did not complete high school
## 6   6 Zone 2     41  39 Unmarried      17     78          High school degree
## 7   7 Zone 3     45  22   Married       2     19          High school degree
## 8   8 Zone 2     38  35 Unmarried       5     76          High school degree
## 9   9 Zone 3     45  59   Married       7    166            College degree
## 10 10 Zone 1     68  41   Married      21     72 Did not complete high school
## 11 11 Zone 2      5  33 Unmarried      10    125            College degree
## 12 12 Zone 3      7  35 Unmarried      14     80          High school degree
## 13 13 Zone 1     41  38   Married       8     37          High school degree
## 14 14 Zone 2     57  54   Married      30    115            College degree
```

```
## 15 15 Zone 2      9  46 Unmarried       3     25 Did not complete high school
## 16 16 Zone 1     29  38   Married      12     75    Post-undergraduate degree
## 17 17 Zone 3     60  57 Unmarried      38    162           High school degree
## 18 18 Zone 3     34  48 Unmarried       3     49           High school degree
## 19 19 Zone 2      1  24 Unmarried       3     20 Did not complete high school
## 20 20 Zone 1     26  29   Married       3     77               College degree
## 21 21 Zone 3      6  30 Unmarried       7     16                 Some college
## 22 22 Zone 1     68  52   Married      17    120 Did not complete high school
## 23 23 Zone 3     53  33 Unmarried      10    101    Post-undergraduate degree
## 24 24 Zone 3     55  48   Married      19     67 Did not complete high school
##    retire gender voice internet forward       custcat churn       CLV
## 1      No   Male    No       No     Yes Basic service     1  95484.99
## 2      No   Male   Yes       No     Yes Total service     1 108989.61
## 3      No Female    No       No      No  Plus service     0 744415.47
## 4      No Female    No       No      No Basic service     1  61204.82
## 5      No   Male    No       No     Yes  Plus service     0 176017.12
## 6      No Female    No       No      No  Plus service     0 210284.60
## 7      No Female    No      Yes      No     E-service     1  56410.08
## 8      No   Male   Yes      Yes     Yes Total service     0  42511.20
## 9      No   Male    No       No     Yes  Plus service     0 339592.36
## 10     No   Male    No       No      No     E-service     0 568552.13
## 11     No Female    No      Yes      No Basic service     1  19618.09
## 12     No Female   Yes       No     Yes  Plus service     0 103920.72
## 13     No Female    No       No      No Basic service     0  97497.46
## 14     No Female   Yes      Yes     Yes Total service     1 290630.13
## 15     No Female    No       No      No Basic service     0  70162.30
## 16     No   Male    No      Yes      No     E-service     0 115868.49
## 17     No   Male    No       No     Yes  Plus service     0 989563.61
## 18     No Female    No       No     Yes  Plus service     0 161912.93
## 19     No   Male    No       No      No Basic service     0  31279.38
## 20     No   Male   Yes      Yes     Yes Total service     1  37060.03
## 21     No Female    No      Yes      No     E-service     1  58200.40
## 22     No   Male    No       No     Yes Basic service     0 248174.27
## 23     No Female   Yes      Yes     Yes Total service     0  36496.52
## 24     No   Male    No       No      No Basic service     0 233178.54
```

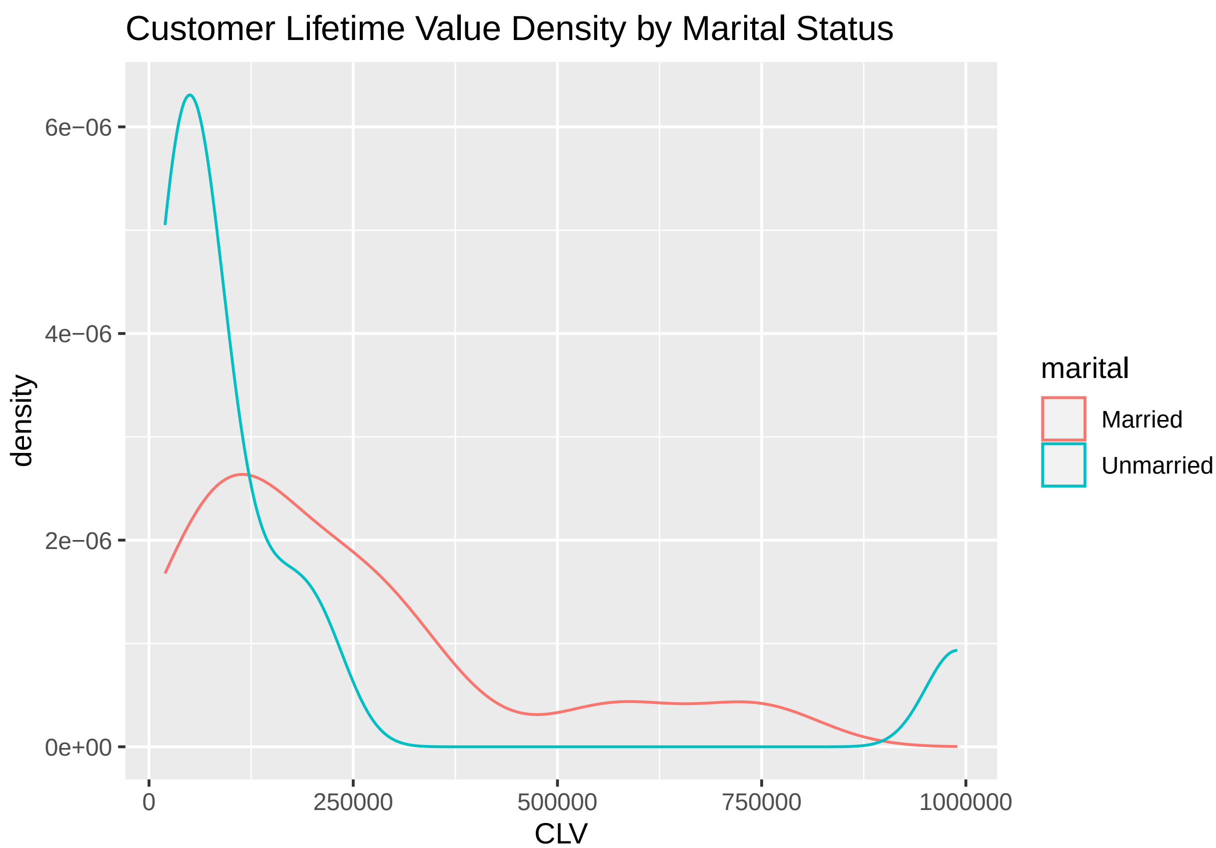
```r
# Comparing CLVs by gender
ggplot(examine_data, aes(x = CLV, color = gender)) +
  labs(title = "CLV Density By Gender") +
  geom_density()
```



From this graph we can see variations in CLV between males and females focused on the first 24 months for simplification. It's apparent that males tend to exhibit lower initial spending compared to females, but as time progresses, males make more consistent and higher-value purchases. Interestingly, both genders typically make a single substantial purchase at the outset, followed by consistent smaller purchases over time.

##CLV Comparison: Married and Unmarried customers

```
ggplot(examine_data, aes(x = CLV, color = marital)) +
  labs(title = "Customer Lifetime Value Density by Marital Status") +
  geom_density()
```



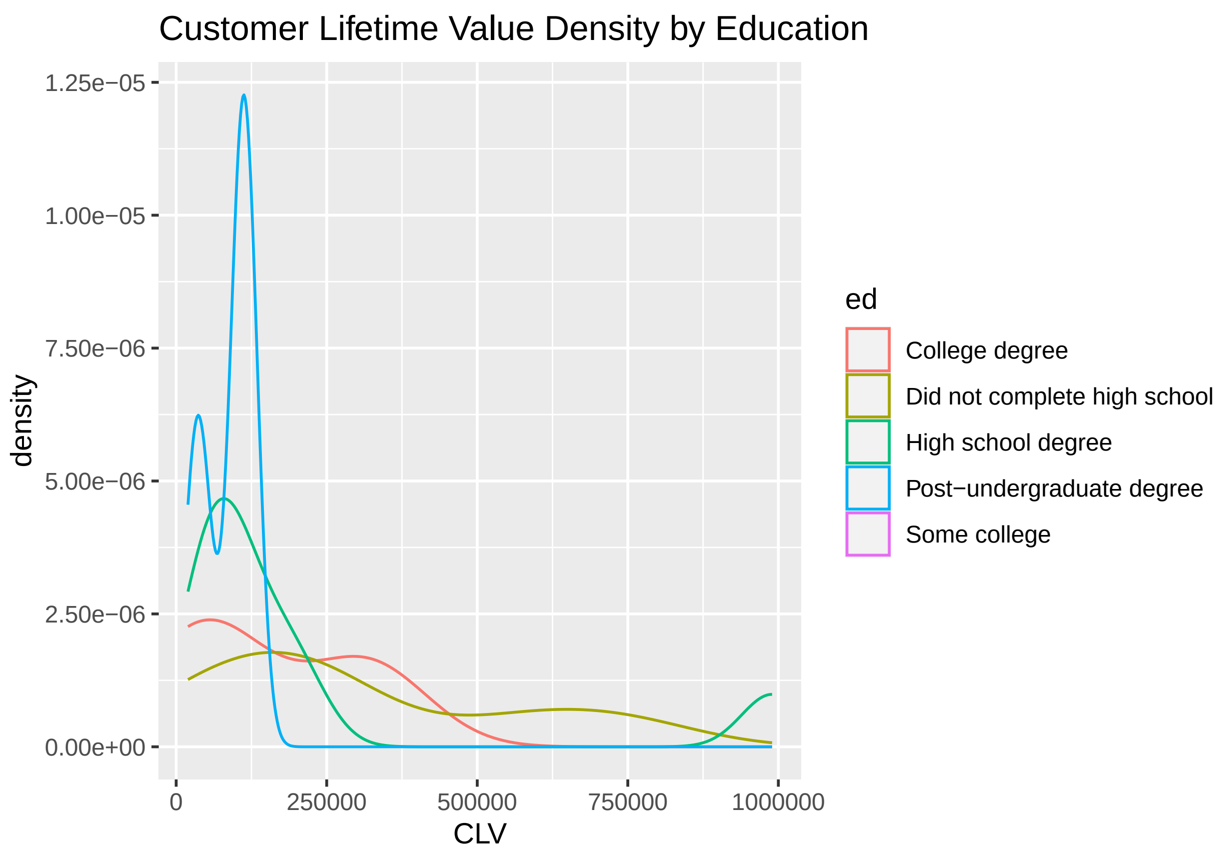
From this comparison of CLV s ov married and unmarried customers we can see that singles typically start with significant purchases but then show inconsistency over time. Married individuals, however, make smaller but consistent purchases after an initial large one. At the end of the graph we can also see that unmarried individuals can start soing purchases after long time not showing any activity.

##CLV Comparison: Education

```
ggplot(examine_data, aes(x = CLV, color = ed)) +
  labs(title = "Customer Lifetime Value Density by Education") +
  geom_density()
```

```
## Warning: Groups with fewer than two data points have been dropped.
```

```
## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning
## -Inf
```



This graph shows analyzation of customers' education levels. Those without high school diplomas tend to make consistent purchases over time. Customers with post-undergraduate degrees initially make high-value purchases but then decrease fastly. This could be because they opt for premium products early on. Customers with only college degree show inconsistent purchasing behavior, likely due to experimenting with different products. High school graduates behave similarly to those with post-undergraduate degrees but start with lower-priced purchases. Overall, both groups demonstrate consistency in their purchasing patterns.

Final conclusions

Based on the findings, the most valuable clients for long-term business success appear to be married individuals. They demonstrate consistent purchasing behavior over time, which is a positive indicator for the business. Next in line are male customers, who also exhibit a consistent purchasing pattern. Regarding education, those who did not complete high school tend to make frequent purchases. Additionally, customers with post-undergraduate degrees make high-value purchases, contributing significantly to the business. Overall, considering consistency and high-value purchases, married males emerge as the most valuable clients.

Retention rate

```
# Estimate churn rate for yearly prediction (considering 12 months)
churn_rate <- mean(predictions <= 12)

# Calculate total number of customers
total_customers <- nrow(data)
```

```r
# Determine the count of at-risk customers
at_risk_customers <- total_customers * churn_rate

# Compute the average Customer Lifetime Value (CLV)
average_clv <- mean(data$CLV)

# Calculate the retention budget
retention_budget <- at_risk_customers * average_clv
retention_budget
```

```
## [1] 3937142
```

What else would I suggest for retention?

To improve retention, it's crucial to segment at-risk customers and assess their value to the company.By that we will undesrtand is that customer worthy for spending resources or not (focus resources on retaining high-value customers). For valuable at-risk customers, we can implement targeted promotions and discounts. Another effective strategy is to maintain regular communication with customers through surveys or events to enhance loyalty and ensure continued engagement.