Customer Churn Prediction Model

Sabahattin Alp Kocabaş, Ali Mustafa Bilibil, Ali Aytuğ Tok, Artun Yayla

2024-10-09

Introduction

Our research area focuses on developing customer churn prediction models. Customer churn is of critical importance for the sustainability of companies, especially in highly competitive industries. A customer's churn not only means the loss of potential revenue from that customer, but also increases the cost of acquiring new customers. Therefore, being able to predict customer churn in advance is of great value for companies to optimize their customer relationship management strategies and increase customer loyalty. Customer churn prediction can help companies use their resources more efficiently, create targeted marketing campaigns, and increase customer satisfaction. Therefore, research in this area is of great financial and strategic importance for the business world.

I. Research Area and Research Question:

- Research Area: Customer Churn Prediction in the Telecommunications Industry
- Research Question: What are the significant factors that influence customer churn, and how accurately can customer churn be predicted using these factors?

II. Key Variables:

- Controllable Variables:

1. Customer Value: The calculated value of a customer based on their usage and payments. This can be influenced by offering tailored pricing, discounts, or value-added services.
2. Complaints: Whether a customer has lodged complaints or not. The frequency and resolution of complaints can be controlled through customer service improvements and proactive problem-solving.

- Uncontrollable Variables:

1. Age Group: Customers can be segmented by age group to tailor marketing strategies and service offerings. This allows for personalized experiences that cater to the specific needs and preferences of different age groups.
2. Churn (Dependent Variable): Whether a customer leaves the service or not. While the business can influence factors leading to churn, the final decision lies with the customer

Hypothesis

For each independent variable, we'll establish the following hypotheses:

Null Hypothesis (H0): The independent variable has no significant effect on churn. In other words, the coefficient of the variable is equal to zero.

$H_0 : \beta_i = 0$ for $i = 1, 2, 3$

Alternative Hypothesis (H1): The independent variable has a significant effect on churn. In other words, the coefficient of the variable is not equal to zero.

$H_1 : \beta_i \neq 0$ for $i = 1, 2, 3$

Where β_i represents the coefficient of each independent variable in the logistic regression model.

Main Body

Loading the Dataset

First, we load the dataset and select the relevant columns.

```
# Loading necessary libraries
library(ggplot2)
library(caret)
library(car)

data <- read.csv("Customer Churn.csv")

# Selecting the relevant columns
data_selected <- data[, c('Customer.Value', 'Complains', 'Age.Group', 'Churn')]
```

Building a Logistic Regression Model

```
new_model <- glm(Churn ~ Customer.Value + Complains + Age.Group, data = data_selected, family = binomial)
#For sake of studying with categorical variables, Age.Group value was classified according to the Age value.
View(data[,c("Age","Age.Group")])
#How the Age.Group value is specified according to the Age

model_mtrx <- model.matrix(~Customer.Value + Complains + Age.Group, data = data_selected)
View(model_mtrx)
#In this matrix, it can easily be seen how the Age.Group effects the model
```

Reviewing the Model Summary

```
summary(new_model)
```

```
##
## Call:
## glm(formula = Churn ~ Customer.Value + Complains + Age.Group,
##     family = binomial, data = data_selected)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.43020  -0.52997  -0.22581  -0.00919   2.66935
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     0.2562759  0.2122791   1.207    0.227
## Customer.Value -0.0071666  0.0005709 -12.554  < 2e-16 ***
## Complains       4.3363722  0.2519798  17.209  < 2e-16 ***
## Age.Group      -0.3459969  0.0663562  -5.214 1.85e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2739.9  on 3149  degrees of freedom
## Residual deviance: 1656.0  on 3146  degrees of freedom
## AIC: 1664
##
## Number of Fisher Scoring iterations: 8
```

```
anova(new_model)
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Churn
##
## Terms added sequentially (first to last)
##
##
##                Df Deviance Resid. Df Resid. Dev
## NULL                            3149     2739.9
## Customer.Value  1   492.54      3148     2247.3
## Complains       1   562.92      3147     1684.4
## Age.Group       1    28.43      3146     1656.0
```

Interpretation and Conclusion

Based on the logistic regression model summary and ANOVA table, it is evident that **Customer Value**, **Complains**, and **Age.Group** have a significant impact on churn, as indicated by their p-values being less than 0.05. Specifically, the p-values for these variables are extremely small (less than 2e-16 for Customer Value and Complains, and 1.85e-07 for Age.Group), strongly suggesting that these variables are statistically significant in predicting churn.

Outlier Checking

```
# Checking for outliers using standardized residuals
data_selected$residuals <- rstandard(new_model)
outliers <- data_selected[abs(data_selected$residuals) > 2, ]
```

```
#Removed outliers
cleaned_data <- data_selected[abs(data_selected$residuals)<2,]
cleaned_model <- glm(Churn ~ Customer.Value + Complains + Age.Group, data = cleaned_data, fam
ily = binomial)
```

Leverage and Influence Points

```
# Identifying leverage points
cleaned_data$leverage <- hatvalues(cleaned_model)
leverage_threshold <- 2 * mean(cleaned_data$leverage)
leverage_points <- data_selected[cleaned_data$leverage > leverage_threshold, ]


# Identifying influence points using Cook's distance
cleaned_data$cooksd <- cooks.distance(cleaned_model)
influence_points <- cleaned_data[cleaned_data$cooksd > (4/(nrow(cleaned_data) - ncol(cleaned_
data))), ]
```

```
# Displaying the identified points

#list(outliers = outliers,leverage_points = leverage_points,influence_points = influence_poin
ts)
```

- Since we consider the source of the data to be reliable, we cannot touch the influence points for theoretical reasons, however we clean the leverage outlier points from the data.

```
# Removing high leverage points (outliers) from the dataset
data_filtered <- cleaned_data[!(rownames(cleaned_data) %in% rownames(leverage_points)), ]
View(data_filtered)
```

Transformation

```
#The variance is almost square of the mean, so it is an intuition to make a log transform.
mean(data_filtered$Customer.Value)
```

```
## [1] 477.9687
```

```
var(data_filtered$Customer.Value)
```

```
## [1] 271480.5
```

```r
# Applying log transformation to Customer.Value
data_filtered$log_Customer_Value <- log(data_filtered$Customer.Value + 1)  # +1 to avoid log
(0)

# Refitting the model after removing high leverage points

transformed_model <- glm(Churn ~ log_Customer_Value + Complains + Age.Group, data = data_filt
ered, family = binomial)
#Logistic Regression method is applied because the model contains categorical variable.

# Summary and ANOVA of the transformed model
summary(transformed_model)
```

```
##
## Call:
## glm(formula = Churn ~ log_Customer_Value + Complains + Age.Group,
##     family = binomial, data = data_filtered)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9561  -0.3377  -0.2417  -0.1655   2.4188
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.29672    0.32440   3.997 6.41e-05 ***
## log_Customer_Value -0.72880    0.04309 -16.914  < 2e-16 ***
## Complains           4.99106    0.26221  19.035  < 2e-16 ***
## Age.Group          -0.16286    0.08630  -1.887   0.0591 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2035.7  on 2626  degrees of freedom
## Residual deviance: 1088.7  on 2623  degrees of freedom
## AIC: 1096.7
##
## Number of Fisher Scoring iterations: 6
```

```r
anova(transformed_model)
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Churn
##
## Terms added sequentially (first to last)
##
##
##                     Df Deviance Resid. Df Resid. Dev
## NULL                                 2626     2035.7
## log_Customer_Value  1   321.70      2625     1714.0
## Complains           1   621.70      2624     1092.3
## Age.Group           1     3.61      2623     1088.7
```

Based on the summary and ANOVA output of the transformed model, the results indicate that the variables **log_Customer_Value** and **Complains** are statistically significant predictors of churn. The p-values for these variables are both less than 0.05 (in fact, less than 2e-16 for both), strongly indicating their significance in the model.

- **log_Customer_Value** has a negative coefficient (-0.72880), suggesting that as the log-transformed customer value increases, the likelihood of churn decreases.
- **Complains** has a positive coefficient (4.99106), indicating that more complaints are associated with a higher likelihood of churn.

On the other hand, the variable **Age.Group** has a p-value of 0.0591, which is slightly above the conventional threshold of 0.05, suggesting that it is marginally significant in predicting churn. This implies that its impact on churn may not be as strong or reliable as the other variables.

Calculating R-squared and MSE for the Transformed Model

```
library(Metrics)
predicted <- predict(transformed_model, type="response")
mse_value <- mse(data_filtered$Churn, predicted)
rsquare <- 1 - sum((data_filtered$Churn - predicted)^2) / sum((data_filtered$Churn - mean(dat
a_filtered$Churn))^2)

list(
  R_squared = rsquare,
  MSE = mse_value
)
```

```
## $R_squared
## [1] 0.4579071
##
## $MSE
## [1] 0.06153806
```

- With an R-squared of 0.4579, the model explains about 45.79% of the variability in churn, which is decent but leaves some variance unexplained. The MSE of 0.0615 indicates that the prediction errors

are relatively small, suggesting the model is fairly accurate in its predictions. However, there is still potential for improving the model to capture more of the underlying patterns in the data.

Checking Assumptions

Residual Mean :

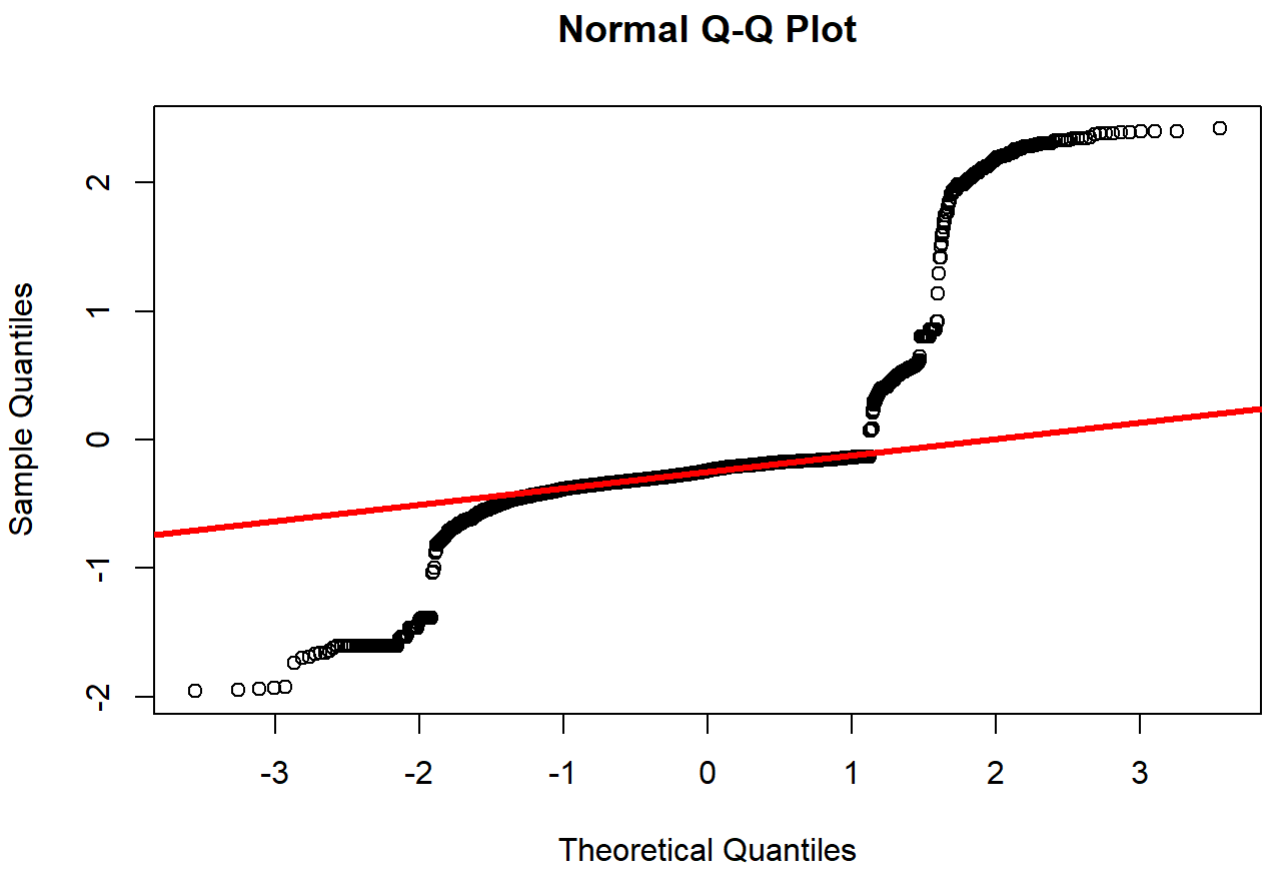
```
#Normality

mean(resid(transformed_model)) #Mean is close to 0
```

```
## [1] -0.1297794
```

- The residual mean is close to zero, as expected. However, the negative value suggests a slight underestimation by the model, although this is typically not a major concern if the value is small, as it is in this case.

Q-Q Plot:

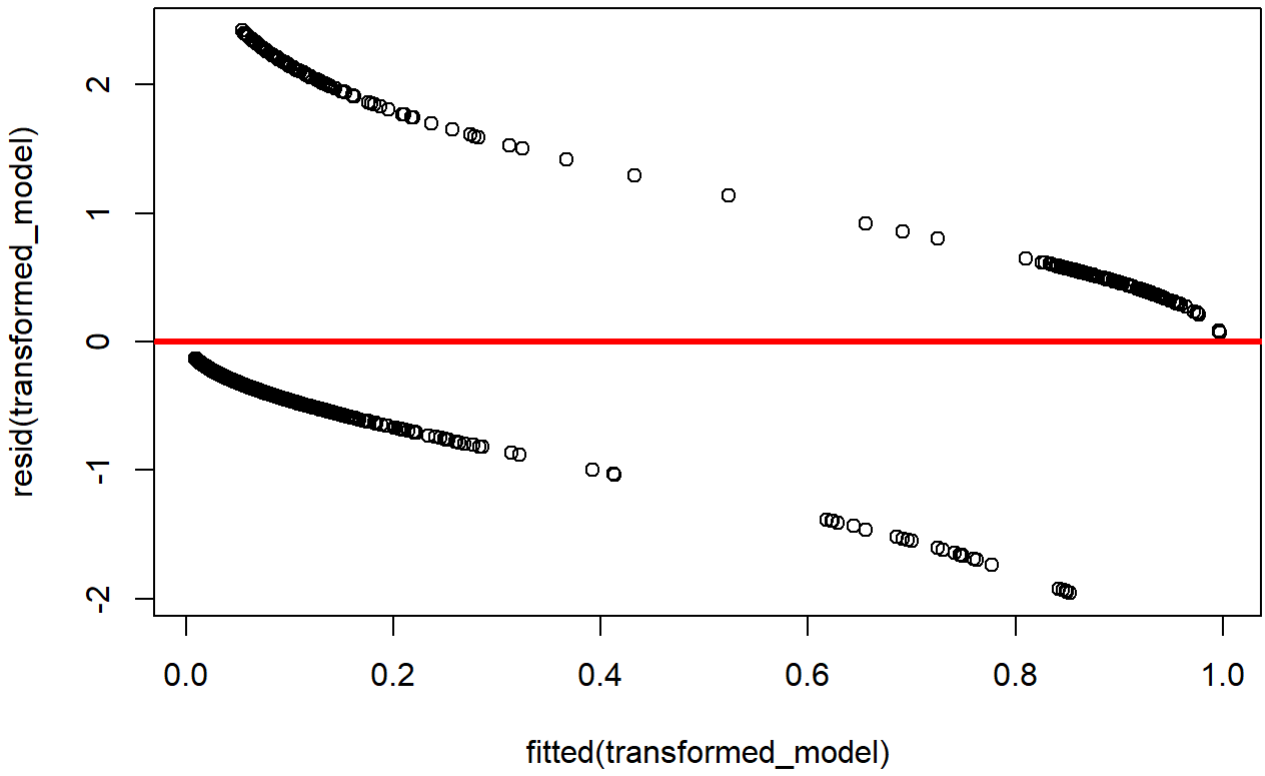
```
#Q-Q Plot to check normality
qqnorm(resid(transformed_model))
qqline(resid(transformed_model), col="red", lwd=3)
```



- The Q-Q plot shows that the residuals deviate from the red diagonal line, especially at the tails, indicating that the residuals do not follow a normal distribution. This suggests that there may be some issues with the model fit, as the normality assumption of the residuals is violated.

Residuals vs. Fitted Values Plot:

```
#Residual vs Fitted Value plot for linearization
plot(fitted(transformed_model), resid(transformed_model))+
abline(h=0, col="red", lwd=3)
```




```
## integer(0)
```

- The residuals vs. fitted values plot shows a clear pattern where the residuals fan out as the fitted values increase, forming a funnel shape. This indicates heteroscedasticity, meaning the variance of the residuals is not constant across levels of the fitted values. This can lead to inefficiency in the estimates and potential bias in the hypothesis tests.

VIF Values:

```
#Checking for multicollinearity
vif(transformed_model)
```

```
## log_Customer_Value          Complains          Age.Group
##           1.070202           1.062276           1.008409
```

```
#VIF values are also suitable to make analysis since being less than 5.
```

- The VIF (Variance Inflation Factor) values for **log_Customer_Value**, **Complains**, and **Age.Group** are 1.0702, 1.0623, and 1.0084, respectively. Since all these values are below 10, multicollinearity does not appear to be a significant issue in this model. This means that the independent variables are not highly correlated with each other, and the model is not adversely affected by multicollinearity.

Overall Interpretation:

While multicollinearity is not an issue (as indicated by low VIF values), the model does exhibit signs of heteroscedasticity and non-normality in the residuals, which could affect the reliability of the model's predictions and the validity of statistical tests. These issues suggest that the model might need further refinement or transformation to meet the assumptions of the regression analysis more closely.

Future Directions

After conducting our analysis, we realized that there's so much more we could explore, especially if we had access to unlimited data sources. If we could tap into any kind of data, here's what we'd add to our study to make it even more insightful.

1. **Usage Patterns:**

- **What We would Add:** It would be super cool to have detailed data on how customers actually use the telecom services. We're talking about things like how often they make calls, how much data they use, when they use it the most, and how their usage changes over time.
- **Future Study:** With this kind of data, we could totally dig deeper into how changes in usage might signal that a customer is about to leave. For example, if someone suddenly starts using way less data, it might be a sign they're thinking about switching to another provider. We could then figure out how to step in early and offer them something to make them stay.

2. **Competitor Campaigns:**

- **What We would Add:** Imagine if we could see exactly what the competitors are doing—like when they launch new promotions, what kinds of discounts they offer, and how they're trying to steal customers away.
- **Future Study:** With that kind of intel, we could analyze how these competitor moves affect churn. For instance, if we notice a spike in customers leaving right after a competitor's big campaign, we could suggest counter-strategies. Maybe the company could launch a special deal or offer better benefits to keep customers from jumping ship.

3. **Market Changes:**

- **What We would Add:** Another interesting angle would be to look at how broader market changes, like new tech developments (hello, 5G!) or economic shifts, influence churn.
- **Future Study:** A future study could explore how these factors play into customer behavior. For example, if 5G becomes the new standard, how does that change what customers expect from their provider? Or, if there's an economic downturn, does that lead to more people canceling their plans? With this kind of data, the company could adapt its strategies on the fly and stay ahead of the curve.

Conclusion:

Overall, with access to more detailed data on **Usage Patterns**, **Competitor Campaigns**, and **Market Changes**, we think we could take churn prediction to the next level. It would allow us to create even smarter, more tailored strategies to keep customers happy and loyal. With the right data, the possibilities are endless, and we'd be excited to see how far we could push this research!