Telecom Churn Project

Brian Buonauro

2023-05-01

Documentation and Exploratory Data Analysis

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.1     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

Through the course of this document, I will explain the process for the extraction of the data into Power BI using R and also perform exploratory data analysis in R to inspect the data before we transform and visualize it in Power BI. There will be minimal cleaning of the data in R since this can be accomplished more efficiently in Power BI. However, for the initial inspection of the data in R, I will utilize the Tidyverse group of packages for data manipulation functions.

Please view the Power BI slide deck, GitHub, and Substack post for additional information about this project.

Extracting the Data Files

In order to execute the data import into Power BI without the problem of dealing with locally based files, I developed an R script that extracts the zip download link into a temporary folder, reads all three of the csv files contained therein into R, and then closes the temporary folder created. The download link for this dataset can be found on the Maven Analytics Data Playground.

```
link <- 'https://maven-datasets.s3.amazonaws.com/Telecom+Customer+Churn/Telecom+Customer+Churn.zip'
temp <- tempfile()
download.file(link, temp)

churn <- read.csv(unz(temp, 'telecom_customer_churn.csv'))
zipcode_pop <- read.csv(unz(temp, 'telecom_zipcode_population.csv'))
data_dict <- read.csv(unz(temp, 'telecom_data_dictionary.csv'))

unlink(temp)
```

Data Summary and Cleaning

To begin, the summary() R function was utilized to inspect the data for anomalies. I discovered one such anomaly that in cases where internet or phone service was not used, columns associated with internet or phone service respectively would be filled with blank strings " " rather than NULLs. This will be adjusted in Power BI during the final analysis using the Table.ReplaceValue() function in the PowerQuery editor. However, a simple solution in R would involve the replace() function to accomplish the same task. Additionally, the Power BI PowerQuery editor was utilized to convert data types (such as currency and percentages) to formats that would be more visually pleasing in the final slide deck, rather than raw integers.

Variables that Strongly Influence Customer Status

In order to judge where to begin with this analysis, I decided to arrange a logistic regression using the Customer.Status column converted to a Boolean expression of whether the customer was churned that quarter as the dependent variable. The independent variables would comprise all numerical variables included in the data, provided they were not closely linked to others that would be included.

To test the influence of the proposed independent variables on each other, the cor() function was used with all numeric variables in the churn dataset. Using a correlation of .6 or greater as a benchmark to evaluate multicollinearity, I found that Total Charges, Total Revenue, and Total Long Distance Charges correlated too closely with Tenure in Months and Monthly Charge. These categories will be discarded from the linear regression since Tenure and Monthly Charge together form an adequate proxy for their presence in the analysis. The cor() test also determined that there were insufficient observations for Average Monthly GB Downloaded and Average Monthly Long Distance Charges to calculate a correlation. These two categories will also be discarded from the regression since it cannot be determined whether they threaten the model's stability.

```r
churn <- churn %>% mutate(is_churned = as.logical(replace(
  churn$Customer.Status, churn$Customer.Status == 'Churned', TRUE)))
churn$is_churned[is.na(churn$is_churned) == TRUE] <- FALSE

churn <- churn %>% inner_join(zipcode_pop, by = 'Zip.Code')
```

The final linear regression model will calculate the raw effect that Age, Number of Dependents, Number of Referrals, Tenure in Months, Monthly Charge, Total Refunds, Population (joined based on a one-to-many relationship from a zip code table provided) and Total Extra Data Charges had on whether customers were churned in the Boolean variable is_churned. Based upon the results of the model, we can conclude that Age, Dependents, Referrals, Tenure, Monthly Charge and Population have a statistically significant impact on whether a customer chooses to leave, or "churns", due to the low p-values observed in column Pr(>|z|). Additionally, we can infer that Total Refunds and Extra Data Charges did not have a significant effect on the likelihood of a customer churning for the opposite reason.

```r
model <- glm(is_churned ~ Age + Number.of.Dependents + Number.of.Referrals +
      Tenure.in.Months + Monthly.Charge + Total.Refunds +
        Total.Extra.Data.Charges + Population, data = churn, family = 'binomial')
summary(model)
```

```
##
## Call:
## glm(formula = is_churned ~ Age + Number.of.Dependents + Number.of.Referrals +
##     Tenure.in.Months + Monthly.Charge + Total.Refunds + Total.Extra.Data.Charges +
##     Population, family = "binomial", data = churn)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9842  -0.6894  -0.3069   0.7016   3.0435
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -1.944e+00  1.248e-01 -15.581  < 2e-16 ***
## Age                       1.189e-02  1.908e-03   6.229 4.70e-10 ***
## Number.of.Dependents     -4.946e-01  5.484e-02  -9.019  < 2e-16 ***
## Number.of.Referrals      -2.557e-01  1.920e-02 -13.316  < 2e-16 ***
## Tenure.in.Months         -4.660e-02  1.751e-03 -26.613  < 2e-16 ***
## Monthly.Charge            2.864e-02  1.299e-03  22.044  < 2e-16 ***
## Total.Refunds            -7.402e-03  4.334e-03  -1.708   0.0877 .
## Total.Extra.Data.Charges  1.091e-03  1.277e-03   0.854   0.3930
## Population                6.060e-06  1.518e-06   3.991 6.58e-05 ***
## ---
## 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: 5910.8  on 7034  degrees of freedom
## AIC: 5928.8
##
## Number of Fisher Scoring iterations: 6
```

Next, I evaluated the same model with all of the independent variables being brought to the same scale via z-scoring. The previous iteration of the model is useful in evaluating the independent variables individually in how they compare to the dependent variable, but to make comparisons between independent variables across units is more prone to inaccuracies based on difference in scale without knowing the relative standing of their effects versus one another.

Below are the same results but done through the z-scoring of the independent variables:

```r
model2 <- glm(is_churned ~ scale(Age) + scale(Number.of.Dependents) +
      scale(Number.of.Referrals) + scale(Tenure.in.Months) +
      scale(Monthly.Charge) + scale(Total.Refunds) +
      scale(Total.Extra.Data.Charges) + scale(Population),
      data = churn, family = 'binomial')
summary(model2)
```

```
##
## Call:
## glm(formula = is_churned ~ scale(Age) + scale(Number.of.Dependents) +
##     scale(Number.of.Referrals) + scale(Tenure.in.Months) + scale(Monthly.Charge) +
##     scale(Total.Refunds) + scale(Total.Extra.Data.Charges) +
##     scale(Population), family = "binomial", data = churn)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9842  -0.6894  -0.3069   0.7016   3.0435
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -1.68307    0.04574 -36.798  < 2e-16 ***
## scale(Age)                        0.19911    0.03196   6.229 4.70e-10 ***
## scale(Number.of.Dependents)      -0.47621    0.05280  -9.019  < 2e-16 ***
## scale(Number.of.Referrals)       -0.76739    0.05763 -13.316  < 2e-16 ***
## scale(Tenure.in.Months)          -1.14358    0.04297 -26.613  < 2e-16 ***
## scale(Monthly.Charge)             0.89371    0.04054  22.044  < 2e-16 ***
## scale(Total.Refunds)             -0.05849    0.03425  -1.708   0.0877 .
## scale(Total.Extra.Data.Charges)   0.02738    0.03205   0.854   0.3930
## scale(Population)                 0.12819    0.03212   3.991 6.58e-05 ***
## ---
## 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: 5910.8  on 7034  degrees of freedom
## AIC: 5928.8
##
## Number of Fisher Scoring iterations: 6
```

In the updated model, we see that the p-values and residuals remain the same since nothing has changed besides the units of measurement now being uniform across independent variables. Besides the two variables that have been discarded previously due to their high p-values, we can observe that Population (of customer zip code) and Age have more minor of an effect on a customer leaving, but are still slightly positive in their effects.

```
churn %>% filter(Customer.Status == 'Churned' & Tenure.in.Months <= 12) %>%
  count() / churn %>% filter(Customer.Status == 'Churned') %>% count()
```

```
##           n
## 1 0.5548422
```

Tenure in Months is the strongest motivating factor causing customers to stay with the telecom company. As can be observed both in the calculation above and in the Power BI slide deck, a large percentage of customers churned (55%) were only customers for 12 months or less. Retaining customers during that crucial first year will be a difficult but necessary task in order to lower churned customer figures in coming quarters. We can now also more clearly observe the strong positive effect that Monthly Charge has on the chances of a customer leaving the service. In the Power BI slide deck, we will explore this by reviewing the results of an exit survey provided to churned customers where customers expressed competition and dissatisfaction with pricing as being their main reasons for leaving.

The Relationship Between Customer Status and Offer

Next, I will evaluate the relationship between churned customers and a significant qualitative variable, what offer the customer received. During this process, I discovered additional NAs that were switched to reflect no special marketing offer being extended to those customers. This was also corrected in the Power BI PowerQuery editor.

```
churn$Offer[is.na(churn$Offer) == TRUE] <- 'None'

churn %>% filter(Customer.Status == 'Churned') %>% group_by(Offer) %>%
  count() %>% full_join(churn %>% group_by(Offer) %>% count(), by = 'Offer') %>%
  mutate(percent_churned = n.x/n.y) %>% arrange(percent_churned)
```

```
## # A tibble: 6 x 4
## # Groups:   Offer [6]
##   Offer     n.x   n.y percent_churned
##   <chr>   <int> <int>           <dbl>
## 1 Offer A    35   520          0.0673
## 2 Offer B   101   824          0.123
## 3 Offer C    95   415          0.229
## 4 Offer D   161   602          0.267
## 5 None     1051  3877          0.271
## 6 Offer E   426   805          0.529
```

What I discovered on the surface level was that Offer E was performing worse than no offer at all. After further investigation, however, I discovered that this was caused in part by a bias within who was receiving which offers.

```
select(churn, Offer, Customer.Status, Tenure.in.Months) %>%
  filter(Offer == 'Offer E') %>% arrange(desc(Tenure.in.Months)) %>% head(5)
```

```
##     Offer Customer.Status Tenure.in.Months
## 1 Offer E          Stayed               10
## 2 Offer E          Stayed               10
## 3 Offer E          Stayed               10
## 4 Offer E          Stayed                9
## 5 Offer E          Stayed                9
```

```
select(churn, Offer, Customer.Status, Tenure.in.Months) %>%
  filter(Offer == 'Offer E') %>% arrange(Tenure.in.Months) %>% head(5)
```

```
##     Offer Customer.Status Tenure.in.Months
## 1 Offer E          Joined                1
## 2 Offer E          Joined                1
## 3 Offer E         Churned                1
## 4 Offer E         Churned                1
## 5 Offer E         Churned                1
```

When looking at the longest standing and newest customers within the group receiving Offer E, I discovered that the entire sampling of customers have been with the company for less than a year. I previously identified that customer segment in this exploratory analysis as containing the bulk of the churned customers and were the worst retained by the company. Without any visibility into what these offers contained, I believe that this may be a new customer introductory offer that expires before the end of the first year of service.

```
select(churn, Offer, Customer.Status, Tenure.in.Months) %>%
  filter(Offer == 'Offer A') %>% arrange(desc(Tenure.in.Months)) %>% head(5)
```

```
##     Offer Customer.Status Tenure.in.Months
## 1 Offer A          Stayed               72
## 2 Offer A          Stayed               72
## 3 Offer A          Stayed               72
## 4 Offer A          Stayed               72
## 5 Offer A          Stayed               72
```

```
select(churn, Offer, Customer.Status, Tenure.in.Months) %>%
  filter(Offer == 'Offer A') %>% arrange(Tenure.in.Months) %>% head(5)
```

```
##     Offer Customer.Status Tenure.in.Months
## 1 Offer A          Stayed               66
## 2 Offer A          Stayed               66
## 3 Offer A          Stayed               66
## 4 Offer A          Stayed               66
## 5 Offer A          Stayed               66
```

The opposite was true when looking into the tenure of customers receiving Offer A, which was only provided to long-standing customers that had 66 months with the service minimum. This was the least churned group out of the dataset provided for Q2 2022. These trends in retention based on tenure and offer do not provide a complete explanation on their own, but can help inform business policy when crafting a new retention strategy for the following quarter.

Please view the Power BI slide deck, GitHub, and Substack post for additional information about this project.