Assignment 6: Churn Management

Günter J. Hitsch

May 22, 2020

Contents

1 Overview

Cell2Cell is a wireless telecom company (the name was changed to disguise the exact identity). As part of its various CRM programs, Cell2Cell actively mitigates customer churn. For this purpose, we develop a model to predict customer churn at Cell2Cell and use the insights from the model to develop a targeted incentive plan to lower the churn rate.

Address these key issues:

1. Is customer churn at Cell2Cell predictable from the customer information that Cell2Cell maintains?
2. What factors drive customer churn? Which factors are particularly important?
3. What incentives should Cell2Cell offer to prevent customer churn?

2 Data

All data are contained in the file `Cell2Cell.RData`. Please consult the file `Cell2Cell-Database-Documentation.xlsx` for a description of the data and some summary statistics.

Note that *calibration sample* is an alternative term for *training* or *estimation* sample. Hence, the variable `calibrat` indicates if a customer belongs to the training (estimation) sample or to the validation sample.

Inspect the data.

Some variables have missing values, which is common and of no concern (unless the missing values indicate some *systematic* flaws or bias in how the data were constructed). Most estimation methods in R will automatically delete rows with missing values before estimating the model. However, the `predict` methods will yield `NA` values if a row in the data used for prediction contains missing values. Hence, in a situation where you don't need to keep the full data I recommend to remove any observations with missing values before you conduct the main analysis.

Recall from the *Introduction to Data Science in R* guide that you can drop all rows in a data frame that contain missing values using the `na.omit()` function.

Oversampling

One important aspect of the data that is that the calibration sample was selected using *oversampling*. Oversampling is a technique to choose a calibration sample such that the values of the outcome (dependent) variable are balanced. In particular, the outcome in one half of the observations takes the value 0 (customer did not churn), and the other half takes the value 1 (the customer churned).

The purpose of oversampling is to obtain more precise estimates (lower standard errors) when estimating a logistic regression model. However, oversampling also results in biased predictions, in our example predicted churn probabilities that are larger than the true churn probabilities. How to correct for this bias is explained further below.

The validation sample, on the other hand, *was not created using oversampling* and represents the *true churn rate* in the data.

Verify that approximately 50% of observations in the training sample take the value 1, and contrast this number with the *true churn rate*, i.e. the average rate or probability at which customers churn, in the validation sample.

3 Model estimation

Estimate a logit model to predict the conditional (on the customer attributes) churn probability.

- Make sure that you estimate the model using the calibration = estimation/training sample only, where `calibrat` takes the value 1. One straightforward way of achieving this is to create two new data frames, `estimation_df` and `validation_df`, from the original data.
- Obviously, `customer` and `calibrat` should not be included as independent variables in the logistic regression.

You can display the regression output using `summary`. However, especially when you have a large number of independent variables in the model, it is often convenient to store the regression estimates in a table (data frame).

A simple way to do this is to install the `broom` and `kable` packages. The `tidy` function in the `broom` package converts the regression output (here, the object `fit`) into a data frame. Furthermore, the `kable` function creates a nicely formatted table.

```
library(broom)
library(knitr)

estimation_results_df = as.data.frame(tidy(fit_logistic))
kable(estimation_results_df, digits = 5,
      col.names = c("variable", "estimate", "se", "z-statistic", "p-value"))
```

Use the help function, `?kable`, to learn more about the options.

Skimming the estimation results, is there evidence that the independent variables may be able to predict differences in churn rates across customers?

Next, predict (`predict`) the churn rates (probabilities) in the validation sample.

Calculate the mean of the predicted churn rates and verify that this mean reflects the average observed churn rate in the estimation sample, not the mean of the observed churn in the validation sample.

Recall that the validation sample was chosen without oversampling and indicates the *true churn rate* in the data. Hence, we need to account for the upward bias in the predicted churn probabilities due to oversampling in the estimation sample. How to do this is explained in the next section.

4 Model estimation and prediction: Accounting for oversampling

4.1 Explanation

Note: Fully understanding this explanation is not very important, and you can simply skim the section.

The idea of oversampling is as follows. If the response rate in the data is small, there is a strong imbalance between observations with a response of $Y = 1$ and a response of $Y = 0$. As a consequence, estimating the model is difficult and the estimates will be imprecise, i.e. they will have large standard errors.

The solution: Create a training sample with one half of observations randomly chosen from the original data with response $Y = 1$, and the other half randomly chosen from the original data with response $Y = 0$. Now estimation is easier and the standard errors will be smaller.

However, when applied to logistic regression, oversampling will result in a biased estimate of the intercept (constant) term, although all other estimates will be unbiased.[1] Hence, if we do not de-bias (adjust) the intercept, the predicted probabilities will be too large, reflecting the artificial response rate of $\frac{1}{2}$ in the over-sampled training data.

In order to de-bias the scale of the predicted response (in this example: churn) in the validation sample we need to supply an *offset variable* to the logistic regression model. An offset is a known number that is added to the right-hand side of the regression when estimating the model, and adding the offset will correspondingly change the estimate of the intercept. The offset takes the form:

$$\text{offset} = (\log(\bar{p}_e) - \log(1 - \bar{p}_e)) - (\log(\bar{p}_v) - \log(1 - \bar{p}_v))$$

Here, $\bar{p}_e$ is the average response rate in the estimation sample and $\bar{p}_v$ is the average response rate in the validation sample.

Then, when we predict the response rate in the validation sample, we set the offset variable to 0.

Why does this work? — Conceptually, logistic regression is a regression model for the log-odds of the response (outcome) probability,

$$\log\left(\frac{p}{1-p}\right) = \log(p) - \log(1-p) = \beta_0 + \beta_1 X_1 + \beta_2 X_1 + \ldots$$

When we add the offset variable to the right hand side of the regression model the estimation algorithm will "incorporate" the offset in the intercept, β_0. The effect of setting the offset to 0 is equivalent to subtracting the offset from the intercept. Subtracting the offset amounts to:

(i) Subtracting $\log(\bar{p}_e) - \log(1 - \bar{p}_e)$, the log-odds of the artificial response rate in the estimation sample, and

(ii) Adding $\log(\bar{p}_v) - \log(1 - \bar{p}_v)$, the log-odds in the validation sample that reflects the true log-odds in the data.

This process de-biases the predicted response, i.e. restores the correct response level.

Note: *Never use over-sampling to create the validation sample*, otherwise the offset variable approach will not work.

[1] To be technically correct, the other estimates will be *inconsistent*.

4.2 Implementation

(1) To create the offset variable, `offset_var`, compute the mean observed churn rates in the estimation and training samples, $\bar{p}_e$ and $\bar{p}_v$.

Then calculate the offset using the formula

$$\texttt{offset\_var} = (\log(\bar{p}_e) - \log(1 - \bar{p}_e)) - (\log(\bar{p}_v) - \log(1 - \bar{p}_v))$$

(2) Add the offset as a new column to the estimation sample. Then estimate the logistic regression model.

To tell `glm` that you want to use `offset_var` you need to use a formula of the form:

```
glm(y ~ <all other variables> + offset(offset_var), ...)
```

Where you place `offset()` on the right-hand side of the formula is irrelevant.

For example, if you use all variables in your data (except `y` and `offset_var`) as independent variables:

```
glm(y ~ . + offset(offset_var), ...)
```

(3) Set the offset variable `offset_var` to 0 and add it to the validation sample. This is simple: In the validation sample, create a column that has the same name as the offset variable column in the estimation sample, and make sure this column takes values `= 0`.

Then predict the churn rates in the validation sample.

Note: When predicting the churn rates you will likely see the warning message, *prediction from a rank-deficient fit may be misleading*. Simply ignore. To suppress the warning message, add `warning = FALSE` to the code chunk header, i.e. `{r, warning = FALSE}`.

Calculate the mean of the predicted churn rates and verify that this mean reflects the observed average churn rate in the validation sample.

5 Model validation

Provide evidence for the validity of the model using a lift table and a lift chart.

Note: All model validation steps are need to be performed in the validation sample.

(1) Create a table that contains both the observed `churn` and the predicted churn rate (probability) from the logistic regression model that you estimated in the previous section. You can simply add the predicted churn rate as a new column to the validation sample data frame.

(2) Add a column, `score`, that assigns a number from 1 to 10 to each customer based on the predicted churn rate. Use the `cut_number` function that is part of the `tidyverse` (to be precise, the function is in the `ggplot2` package):

```
score = cut_number(Pr_churn, n = 10, labels = FALSE)
```

`cut_number` creates `n` equally sized groups, where group 1 represents customers with the lowest predicted churn rates and `n` (here: `n = 10`) represents customers with the highest predicted churn rates.

(3) Create a summary table that, for each of the scores, contains:

(i) Mean predicted churn rate
(ii) Mean observed churn rate
(iii) Lift factor

Recall the definition of the lift:

$$\text{lift in segment } k = 100 \cdot \frac{\text{churn rate in segment } k}{\text{average churn rate across all customers}}$$

Hint: Create the summary table using `group_by` and `summarize`.

(4) Provide a graph that displays score-level mean observed churn on the y-axis and the customer `score` on the x-axis.

(5) Similarly, provide a lift chart (graph).

(6) Display the data in (4) and (5) in the form of a table.

Discuss the results. Do the results provide evidence for the validity of the model?

6 Effect sizes: Why do customers churn?

We would like to understand *why* customers churn, which can help us to propose incentives to prevent customer churn. Hence, we now construct a table that contains comparable effect sizes (changes in the churn probability) for all independent variables, as we discussed in class.

Because logistic regression coefficients are not directly interpretable, we first need to estimate a linear probability model that predicts customer churn.

Note: Do *not* use an offset variable when estimating the linear probability model!

Once you have estimated the model, you can construct the effect size table in Excel, or you can use the code below to construct the table in an automated fashion

6.1 Construct effect size table in Excel

Assuming your linear probability model output is called `fit_linear_prob`, you can convert the estimation results into a data frame and then save the estimation results as a csv file:

```
linear_prob_results_df = as.data.frame(tidy(fit_linear_prob))
write_csv(linear_prob_results_df, path = "<file name>.csv")
```

Now proceed in Excel, using both the `Cell2Cell-Database-Documentation.xlsx` file and the estimation results files.

6.2 Automated approach to construct effect size table

Run all four code chunks below.

For the code to work ensure that:

(i) Your main data frame, `cell2cell_df`, is in memory, and you have removed all rows with missing values.

(ii) Either the estimation results object from the linear probability model is called `fit_linear_prob`, or you modify the object name in the third code chunk below.

(iii) The objects `p_e` and `p_v` that contain the mean observed churn rates in the estimation and validation samples are in memory.

If these conditions are met, the code below will work.

First, we create the function `isDummy` that returns the value 1 if `x` is a dummy variable, and 0 otherwise.

```
isDummy <- function(x) {
   elements = unique(x)
   if (length(elements) == 2L & all(elements %in% c(0L,1L))) is_dummy = TRUE
   else is_dummy = FALSE
   return(is_dummy)
}
```

Second, we create `summary_df`, which contains the standard deviation of all variables and an indicator, `is_dummy = 1`, if the variable is a dummy variable.

```
summary_df = cell2cell_df %>%
   summarize_all(list(sd = sd, is.dummy = isDummy)) %>%
   pivot_longer(cols = everything(),
                names_to = "statistic", values_to = "value") %>%
   separate(statistic, into = c("variable", "statistic"), sep = "_") %>%
   pivot_wider(id_cols = variable, names_from = "statistic", values_from = "value") %>%
   rename(is_dummy = is.dummy)
```

Third, we create the `effect_sizes_df` table that combines the estimation results (in the object `fit_linear_prob`) and the summary statistics.

```
effect_sizes_df = as.data.frame(tidy(fit_linear_prob)) %>%
   transmute(variable = term, estimate = estimate, t_stat = statistic) %>%
   inner_join(summary_df) %>%
   mutate(change_prob = ifelse(is_dummy, estimate, estimate*sd)*(p_v/p_e)*100)
```

Fourth, we sort the variables according to the magnitude of the effect sizes, and print the results table using `kable`.

```
effect_sizes_df = effect_sizes_df %>%
   arrange(desc(abs(change_prob)))

kable(effect_sizes_df, digits = 4)
```

Alternatively, you can write the effect size table to a csv file:

```
write_csv(effect_sizes_df, path = "<file name>.csv")
```

Interpretation of the `change_prob` variable: For a dummy variable, `change_prob` is the change in the churn probability associated with an increase in the dummy value from 0 to 1. For all other variables, `change_prob` is the change in the churn probability associated with an increse in the value of the variable by one standard deviation.

`change_prob` is defined on the scale from 0 to 100. I.e., if `change_prob` takes the value 0.6, the effect size is a change in the churn probability by 0.6%, not by 60%!

7 Develop incentives to prevent churn

Using the effect size table that you constructed in the previous section, identify some factors that are strongly associated with churn. If actionable, propose an incentive that can be targeted to the customers to prevent churn.

8 Summarize your main results

1. Is customer churn at Cell2Cell predictable from the customer information that Cell2Cell maintains?
2. What factors drive customer churn? Which factors are particularly important?
3. What incentives should Cell2Cell offer to prevent customer churn?