Customer Churn Analysis

1. Topic

What is the project about? What problem are you thinking to solve or in which domain you think you can contribute meaningfully?

This project is about finding factors that cause customers to churn. The customer churn analysis contributes to the business domain the most is a severe problem for all kinds of organizations that causes major expenses.

Companies usually increase the customer base by acquiring new customers as well as retaining old ones. In many cases, it is a more cost-efficient way to retain a current customer than to acquire a new one. Therefore, customer retention is a key issue for companies and we aim to investigate it using causal inference tools.

2. Project question

*What specific question are you seeking to answer with this project? For this project, this must be a **causal** question.*

Does a choice of a particular phone plan (e.g., a phone plan with unlimited internet access), lead to a lower probability of churn?

3. Ideal experiment

If *you were a god, what experiment would you run to answer your question? Define both your treatment variable and your outcome of interest.*

We would have two identical copies of each study participant. We would separate them into two groups: one group enrolled in a regular phone plan and the other group enrolled in an upgraded phone plan. Upgraded phone plans could include streaming TV access, free international calls, or, as mentioned above, unlimited internet access, etc. Then we would investigate whether there is a statistically significant difference in churn rates between these two groups.

4. Pick a study context

Where can you get data that (a) measures your outcome variable, and (b) includes variation in your treatment variable?

This data along with churn results can be obtained from phone service providers' open-source datasets.

5. Project design

Given the context you want to study (and data you can find), what design do you think would be feasible?

We will define treatment and control groups: treatment as the upgraded phone plan users, and control as the regular plan users. We then will use propensity score matching to assure that there is no significant difference in the chosen variables (in other words, treatment and control groups are well balanced), except the one indicating phone plan (the treatment variable). After matching the groups, we will use logistic regression to predict the churn rate and analyze the coefficient for the phone plan holding the regular plan users as the baseline.

6. Model results

One of the hardest parts of developing a good data science project is developing a question that is actually answerable. Perhaps the best way to figure out if your question is answerable is to see if you can imagine what an answer to your question would look like. Below, draw the graph, regression table, etc. that you would consider to be an answer to your question. Then draw it again, so you have a model result for if treatment has an effect, and a model result for if your treatment does not have an effect. (If the answer to your question is continuous, not discrete (like: what is the effect of health insurance on life expectancy), draw it for high values (high inequality) and low values (low inequality)).

As discussed in the question above, we aim to have the regression table coefficients after applying a logistic regression and holding the regular plan users as a baseline.
If the coefficient for the odds ratio for the phone plan is less than one, the upgraded plan users are less likely to churn compared to customers with a regular plan.

Results if hypothesis is true:
Churn rates are lower among upgraded plan users in comparison with regular plan users.

Results if hypothesis is false:

Churn rates are higher for upgraded plan users in comparison with regular plan clients, or there is no difference between them in terms of churning out of telephone services.

7. Final variable required

Now that you've specified what an answer to your question looks like, what data do you need to generate that answer?

For each variable, define both the variable you need and the population for which you need the variables to be defined.

For both regular and upgraded plans users, we need the following variables:

1. Demographic variables including age, gender, location, etc. from a representative sample of a population that have a service contract with a tele company. We believe it is important to have information about location since it may impact the available services (for example, rural areas may have fewer services available and therefore lower satisfaction rates influencing customers to churn).
2. Information about the types of services each customer has, contract type, number of months using a service. If service is provided on a monthly basis, we want to have a population that used a service longer than a month so see whether they extended the contract to the next month.
3. Information of customers' level of satisfaction: customers' level of satisfaction is directly related to whether customers will churn or not.
4. Outcome variable: churn or not. We want a population that has all the information we want above but also their churn status is being recorded.

8. Data sources

The data is obtained from IBM Cognos Analytics sample data sets on Telco customer churn. The datasets contain demographic information about clients like gender, age group, dependents, and paying information like monthly charges, types of services each customer has, etc. The churn column indicates whether or not the customer left within the last month.