Report for FoodCorp

Prepared by *ConsultingCorp*

An introduction to churn

Churn in the strictest definition refers to a customer leaving one company and moving to another. This definition is most applicable to businesses where there is a high cost of switching, i.e. mobile phone providers, tv subscriptions etc. However, the term is also used within fast moving consumer goods retailers, with FoodCorp being an instance of this.

In this case *churn* is conceptually linked to concepts of loyalty and regularity. Since regularity is linked to an individual's circumstances (family size, work type, income, time of year etc) there is scope for much more complex models that take this into account. I.e. churn models which define if someone has churned based on a significant (negative for the company) deviation from their regular pattern. An example would be someone who shops every week changing to shopping once a month. Under relatively simple definitions of churn they might not be considered as having churned. However, with respect to their normal behaviour they have certainly become less loyal and are likely shopping elsewhere and can be considered to have undergone some form of churn.

Such approaches, however, are inherently complex. Given the time scales FoodCorp has requested, these approaches have been deemed to be too involved for the upcoming pilot study. As such the proposed methodology will focus on a comparatively simple model, used widely within industry, where churn is defined globally.

This report details the methodology of an initial analysis in determining such a definition in a data driven fashion. The results stop short of a final recommendation, instead providing the evidence required to make an informed choice. It is ConsultingCorp's understanding that this will be subsequently undertaken by a 3rd party.

A methodology for defining churn

The question considered when determining a global definition of churn[1] is that of:

"*what length of observed absence defines a customer as one who has churned?"*

If FoodCorp had data over all time then it would be **known** when someone had churned - if they never came back. There are two issues with this:

1. At any given point the transactional data does not cover all of time - the customer could always come back the day after our last record.
2. Even if the company had data over all time, depending on the business, one typically cares about customers becoming infrequent as well as having churned.

Specifically a decision will need to be made regarding:

Assuming customer return dates are known (for all customers), what length of inactivity (in days) is unacceptable to the company (and therefore willing to attempt interventions)?

This may of course be based on business targets unrelated to the data. However, the data can still give us an indication of which length may be a good length to pick. This analysis presented in this document provides these indications. The code used to generate these will be provided on request, however, understanding and/or modifying the code used to generate this report is not required for this task.

Churn is formally defined in this context as:

If date_today - date_of_last_purchace **>** β **then** churn **otherwise**, not churn

This definition reflects:

1. FoodCorps decision to ignore time of day
2. FoodCorps decision that multiple visits on the same day are not considered
3. Churn is defined as being unobserved for strictly **greater than** the defining inactivity period. This is due to the date difference meaning purchases on consecutive days returns a value of 1. Since β is easiest to interpret as the period of inactivity, if the inequality was not strict the period of inactivity could easily be less than 24 hours due to (1 & 2). Therefore the inequality is required to be strict.

Having provided a formal definition of churn the question becomes how to choose β.

The most simple statistic would be to consider the **distribution of times between visits**.

[1] This is somewhat in contrast when considering a global definition of churn for a KPI where one often considers: *how many days indicates someone has churned*. The questions are, however, clearly related.

A complementary statistic is the **percentage of customers who are considered churners (PCCC)**. Specifically, if the machine learning algorithm got it right, how many people would the company end up targeting? This can help prevent the selection of the inactivity periods that would be considered (un)acceptable to the company as it directly affects the number of people they will have to spend resources targeting.

The **PCCC** has be measured from the transactional data as:
The average (per prediction temporal reference point) percentage of customers who churn in hindsight (i.e. under a perfect predictor) given a churn definition of x days?

A simple implemented would be:

- For each day in the data set up to β days before the last record:
    - For each active customer:
        - Check if they will churn by the definition, compute percentage

This, however, is a little too simplistic. Specifically, FoodCorp has indicated that once the predictive system has labelled someone as churned and places them on a list to be targeted, they will be removed from the list of people to predict. Once they shop again then they will be removed from the list and may again be once again predicted as going to churn.

The final analysis implemented by ConsultingCorp therefore is based on the following:

- For each day in the data set ordered by date ascending, up to β days before the last record:
    - Populate a list of active customers, those that have shopped before the current date and are not on the churned list.
    - Add to the active list any customers that shopped today (they have reactivated if they were on the churned list).
    - For each customer who is active, predict if they are going to churn.
    - Compute the proportion of churned vs. active.
    - Move those predicted as churned to the churned list.

Implementation

The two statistics that ConsultingCorp have implemented are:

1. Distribution of times between visit
2. Target class probability (PCCC)

Each statistic is displayed as an annotated graph. Details for each graph are included below.

Results

Statistic: Distribution of times between visit (Figure 1)

The cumulative distribution of "median per customer days between visits" has been plotted in Figure 1. The x-axis represents the days between visits and the y-axis represents the percentage of customers that have a median days between visits less than or equal to the corresponding x-value.

In context, a y-axis value of 75% and x-axis value of 55 tells us that if we set 55 days as the inactivity period for the global definition of churn 75% of FoodCorp's customers would not, on average (median), be considered to have churned.

For ease of interpretation, the corresponding periods of inactivity for 25%, 50%, 60% and 75% of customers are highlighted by changes in the figure's background colours.

Statistic: Percentage of customers who are considered churners (Figure 2)

The ground truth percentage of customers who churned is plotted against an increasing inactivity period (potential definitions of churn) in Figure 2.

In context, a y-axis value of 9.54% and a x-axis value of 65 tells us that we expect to predict 9.54% of our active customers as having churned on average if we had a perfect predictor and defined churn as being a period of inactivity for at least 65 days.

For ease of interpretation, the churn definitions equal to those in Figure 1 are highlighted. You may select others than those highlighted.

Summary Table (Figure 3)

A summary table is shown in Figure 3, providing an interpretation of the results for the churn definition equal to those in Figure 1. The absolute value of the active people that would be predicted to churn based on the percentage computed for Figure 2 is also reported. You may select others than those highlighted.

Figure 1: Median per customer days between visits

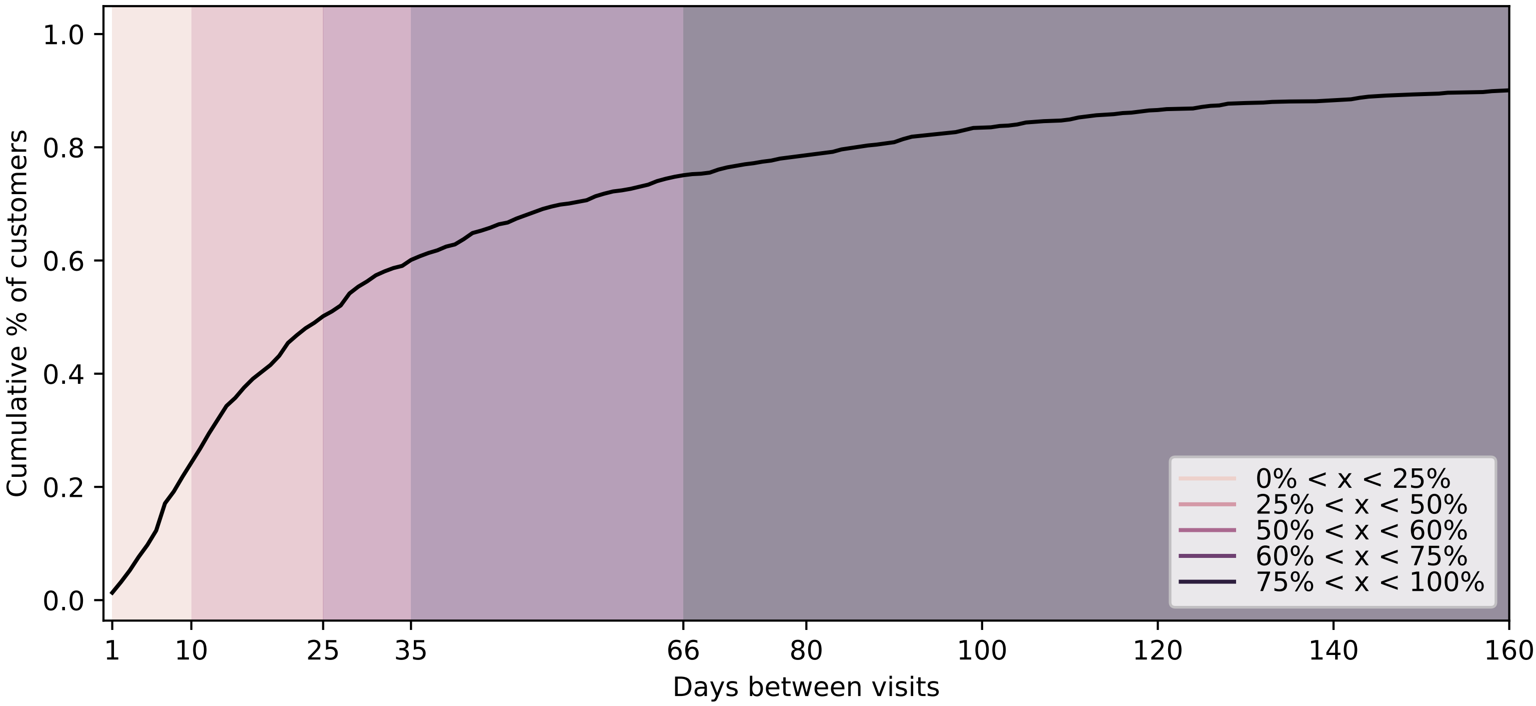

Figure 2: Target class (have churned) proportions vs. churn definition (churn: inactivity >= x)

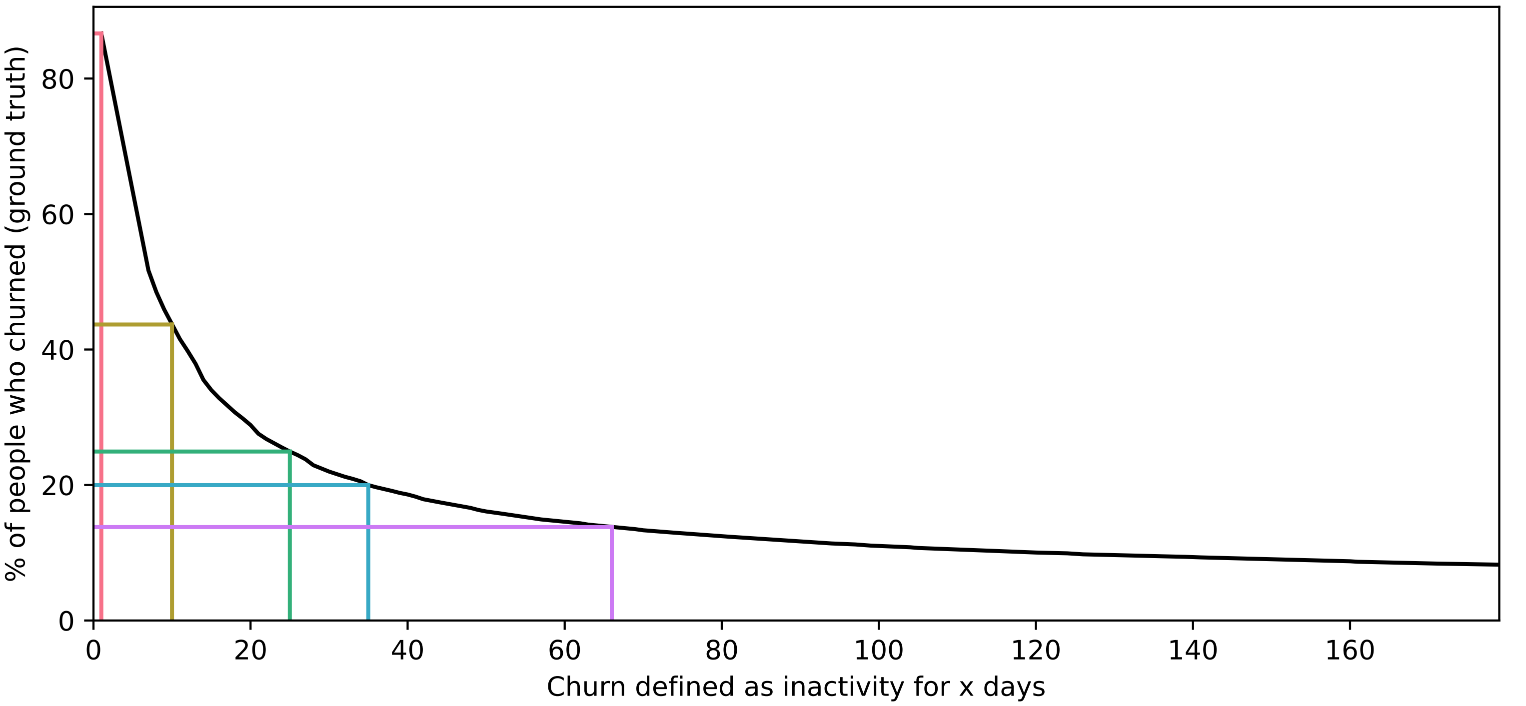

Figure 3: Summary for selected churn definitions

Churn definition: 66 days
75.06% of customers median days between visits is less than this
FoodCorp should expect to target 13.79% of active people (10.68 per prediction day on average) with a perfect classifier.

Churn definition: 35 days
60.09% of customers median days between visits is less than this
FoodCorp should expect to target 19.98% of active people (15.48 per prediction day on average) with a perfect classifier.

Churn definition: 25 days
50.15% of customers median days between visits is less than this
FoodCorp should expect to target 24.94% of active people (19.35 per prediction day on average) with a perfect classifier.

Churn definition: 10 days
24.25% of customers median days between visits is less than this
FoodCorp should expect to target 43.69% of active people (33.86 per prediction day on average) with a perfect classifier.

Churn definition: 1 days
1.31% of customers median days between visits is less than this
FoodCorp should expect to target 86.66% of active people (67.08 per prediction day on average) with a perfect classifier.