Churn Probability

Dylan Zammit

October 2024

Contents

1 Notation

Consider the following list of notation that will be used throughout this document

- x is the unique number of depositing days
- t_x is the day of last deposit since FTD
- T is the day of observation since FTD
- $T_{eol} := max(0, 30 - t_x + T)$ is the day of churn assuming no further deposits
- a, b, α, r are learned model constant hyperparameters
- $L(\cdot|\cdot)$ is the likelihood of the parameters given the customer history
- $B(\cdot, \cdot)$ is the beta function (see here)
- $\Gamma(\cdot)$ is the gamma function (see here)

2 Derivation

To achieve this, we need to set $y = 0$ in Equation (33) of Hardie's notes and set the value of t to T_{eol} at the time of prediction.

$$Pr(Y(t) = 0|x, T, t_x, r, a, b, \alpha) = \frac{A + B}{L(r, a, b, \alpha|x, t_x, T)}. \tag{1}$$

Where

$$A = \frac{B(a+1, b+x-1)\Gamma(r+x)\alpha^r}{B(a,b)\Gamma(r)(\alpha+t_x)^{r+x}}, \tag{2}$$

$$B = \frac{B(a, b+x)\Gamma(r+x)\alpha^r}{B(a,b)\Gamma(r)(\alpha+T+t)^{r+x}} \tag{3}$$

and

$$L(r,a,b,\alpha|x,t_x,T) = \frac{\Gamma(r+x)\alpha^r}{B(a,b)\Gamma(r)} \left(\frac{B(a,b+x)}{(\alpha+T)^{r+x}} + \frac{B(a+1,b+x-1)}{(\alpha+t_x)^{r+x}} \right), \tag{4}$$

the last of which was obtained from Equation 11 of the same notes. Factoring out common terms in $A + B$ we get

$$A + B = \frac{\Gamma(r+x)\alpha^r}{B(a,b)\Gamma(r)} \left(\frac{B(a+1,b+x-1)}{(\alpha+t_x)^{r+x}} + \frac{B(a,b+x)}{(\alpha+T+t)^{r+x}} \right) \tag{5}$$

Plugging the values back in Equation 1, we get

$$\begin{aligned} &Pr(Y(t)=0|x,T,t_x,r,a,b,\alpha) = \\ &\qquad \frac{B(a+1,b+x-1)(\alpha+t_x)^{-(r+x)} + B(a,b+x)(\alpha+T+t)^{-(r+x)}}{B(a,b+x)(\alpha+T)^{-(r+x)} + B(a+1,b+x-1)(\alpha+t_x)^{-(r+x)}}. \end{aligned} \tag{6}$$

3 Underflow Issue

Let us simplify this by writing it as

$$P_t = \frac{B_1 E^\phi + B_2 F_t^\phi}{B_2 M^\phi + B_1 N^\phi}. \tag{7}$$

The issue with computing this is due to exponentiation of the denominators by a large negative factor of $\phi := -(r+x)$, recalling that x is the unique number of customer depositing days. This number could potentially be in the hundreds for an active customer. To solve this issue we will apply the Log-Sum-Exp trick.

Each one of the four terms have the same format, so let us chose first one without loss of generality. We can rewrite this as

$$B_1 E^\phi = \exp(\log(B_1 E^\phi)) = \exp\left[\log(B_1) + \phi \log(E)\right].$$

By applying this trick, ϕ is no longer part of the exponent, and hopefully more manageable. However, we still have to exponentiate with respect to $K_E := (\log(B_1) + \phi \log(E))$, which is too large (negative). Repeating this process we can rewrite Equation 7 as

$$P_t = \frac{e^{K_E} + e^{K_F}}{e^{K_M} + e^{K_E}}. \tag{8}$$

Now, let $K := \max(K_E, K_F)$ and $K' := \max(K_E, K_M)$. Thus we can rewrite P_t once more as

$$P_t = e^{K-K'} \left(\frac{e^{K_E - K} + e^{K_F - K}}{e^{K_M - K'} + e^{K_E - K'}} \right). \tag{9}$$

All exponents of this equation are now small enough to exponentiate without the worry of over/underflow.

4 Final Expression

Writing all terms in full in the order in which they should be computed, we get

$$
\begin{aligned}
E &= \alpha + t_x \\
F_t &= \alpha + T + t \\
M &= \alpha + T \\
\\
K_E &:= \log(B(a+1, b+x-1)) - (r+x)\log(E) \\
K_F &:= \log(B(a, b+x)) - (r+x)\log(F_t) \\
K_M &:= \log(B(a, b+x)) - (r+x)\log(M) \\
\\
K &:= \max(K_E, K_F) \\
K' &:= \max(K_E, K_M) \\
\\
P_t &= e^{K-K'}\left(\frac{e^{K_E-K} + e^{K_F-K}}{e^{K_M-K'} + e^{K_E-K'}}\right).
\end{aligned} \tag{10}
$$