ITC INFOTECH
Business-friendly Solutions

A Prudent Approach in Predicting Time to Churn for Fashion Retail Industry Using Accelerated Failure Time Models in A Non-Contractual Setting

Aditi Raha
Associate Data Scientist
Advanced Analytics,
ITC Infotech Bangalore, India

Atrijit Ghosh
Principal Data Scientist
Advanced Analytics,
ITC Infotech Bangalore, India

Anindya Neogi
Chief Data Scientist
Advanced Analytics,
ITC Infotech Bangalore, India

Keywords

Survival Analysis, Time to churn, CLTV, Accelerated Failure Time Models(AFT), Parametric Survival Regression

Abstract

Customer retention is currently the topmost priority in all businesses in today's world. With rapid growth in the market across all businesses, it is a mandate to focus on retaining the existing customers, failing which might result in massive profitability reduction across major perspectives. A customer churn is referred to as the inclination of a customer to leave a service provider (Chitra Phadke, Huseyin Uzunalioglu et al., 2013; Vivek Bhambri, 2013; Zhen-Yu Chen et al., 2012; Clement Kirui, Li Hong et al., 2013; Yaya Xie a, Xiu Li et al., 2009; Chandar, Laha and Krishna, 2006). Although knowing whether a customer churns in the subsequent period might interest the service provider, simply having an idea of about when he leaves can provide added insights and help the Enterprise take timely necessary retention actions. This paper attempts to predict the time to churn of a customer in a non-contractual setting, thus providing a roadmap to model the said problem using Parametric Survival Analysis. The Prediction accuracy of the deployed framework is around 70%.

1. Introduction:

In an age of increasing customer demand and heightened competition, retaining customers is critical. It's no surprise that the cost of acquiring a customer is far more than retaining one. With churn rates as high as 30% in some markets across the globe, identifying and retaining at-risk customers remains a top priority for all these companies. In a scenario like this, predicting when a customer is likely to churn out can prove to be a greater insight. The point prediction of time to churn aids in many ways:

- It can be used to target the right audience with the right offers.
- Campaigns can be built accordingly to retain the customers, thereby indicating the construction of a long-term relationship with the customer.
- It helps to understand subscription retention pattern and product performances.
- It helps in calculating the customer lifetime value, which in turn can be used to quantify the long term value of customer segments, justifying pricing strategy, etc.
- It helps in creating revenue by transforming customer journey.

2. Methodology:

2.1 Formulation of the Business Problem:

The main objective was to provide an estimated time of customer churn based on covariates and also look for the effects of certain covariates on the survival time of each customer. *(Here, survival time refers to the time till which the customer remains associated with the service provider).*

Defining the event of Churn:

Churn is the event of a customer leaving his/her association with the service provider. In a non-contractual setting, vis-á-vis the retail industry, defining churn is difficult, since the event to be investigated, i.e. churn, is not really encountered at particular time point. Thus for non-contractual businesses, *"if a customer does not complete a critical event on the platform within a fixed window of time, then the customer is considered churned out of the platform".* In view of the retail industry perspective the window is set to 6 months for our problem at hand. Hence, a customer who does not transact for a consecutive 6-month period (after his/her last purchase date), he/she is considered to be a churned out customer.

2.2 Time Frame for Modelling:

April 1, 2017 was the origin time and March 31, 2018 is the observed termination time.

2.3 Exclusions:

Customers who had "0 days" observed time were excluded since they might have been new customers who joined at the end date of the modelling time frame and thus were in the study for any number of days.

Customers with transaction count as low as 1 within the entire period of study were also discarded, keeping in mind to rule out the subset which is supposedly seasonal buyers or one-time buyers. Keeping such observations within the data might affect the modelling procedure.

2.4 Censoring:

The problem setup suggests that those customers whose 6 months i.e.120 days' critical period could not be observed (after their last transaction day) within our follow-up time were marked censored. Only *type I censoring*, i.e. right censoring, is observed in our data. The *type* I censoring specifies that every subject is under study for a specified period C0 or till failure time t_i(essentially $t_i \leq C0$).

The other kinds of censoring will not be encountered here for the following reasons:

a) Only those customers who were active at the start of the study were included (rules out left-censoring).

b) Customers with more than 120 days of average gap between visits were excluded (rules out interval-censoring). Usually seasonal buyers show such behavior, and avoiding them within the modelling data for survival analysis improved prediction accuracy.

2.5 Population:

About 1.6 million customers were eligible for modelling.

About 0.5 million customers were held back as validation set.

Approximately 0.8 million customers were flagged as churned.

2.6 Model Selection:

To get an initial idea about the appropriate survival model, one of the exploratory methods is based on the shape of the *baseline hazard function.* It is the fundamental indicator to identify the best fit survival model. Comparing basic descriptive statistics also provides an indication of the best fit distribution to the data.

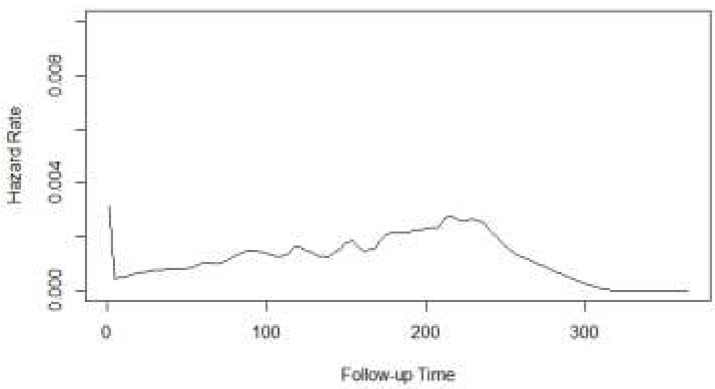

Figure1:Baseline Hazard vs Time

The hazard curve of the data indicates that a possible *Weibull* model shall be a best fit, rather than *lognormal* or *log-logistic (Some of the survival models from Parametric Survival Class of Models).* Further diagnostic plots to account for the data distribution compared to Weibull were looked at (fig. 2).

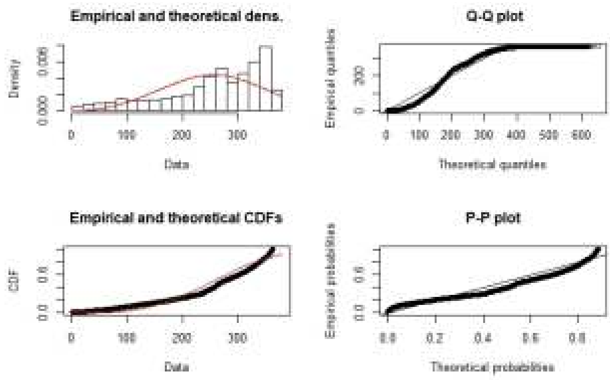

Figure 2: Distribution Fit Plots (Weibull)

However, to validate our results furthermore and to explore other models in view of research perspective, lognormal and log-logistic models were also examined and the best model was selected, based on AIC values.

3. Formulation and Assumptions:

AFT(Accelerated Failure Time) models:

Accelerated Failure Time model can be used to analyze the time to event data. It is generally applied in engineering studies to address the failure rate of a machine. The model works to measure the effect of covariates to 'accelerate' or 'decelerate' survival time. It specifies that predictors act multiplicatively on the failure time (additively on the log of the failure time). The predictor alters the rate at which a subject proceeds along the time axis.

AFT model describes a relationship between the survivor functions of any two individuals. If $S_i(t)$ is the survivor function for individual i, then for any individual j, the AFT holds that,

$$S_i(t) = S_j(\varphi_{ij} t) \text{ for all } t,$$

where φ_{ij} is a constant that is specific to the pair *(i,j)*. Thus, the effect of covariates in an AFT model is multiplicative on the time scale. **φ** is the **acceleration factor** which defines that the ratio of two given survival times is constant for any given survival probability. The survival proportion in one group at any time t is equal to the survival proportion in the second at time ϕt. Therefore, a Quantile–Quantile (Q–Q) plot of the time of survival percentiles would lie on a straight line. AFT models are predominantly fully parametric.

The distribution of T as a function of the covariate x is characterized by the equation:

$$T = e^{x^T \beta} * e^{\sigma\varepsilon}$$

The log linear form of the AFT model represents the relationship between the log of time and the set of covariates as follows:

$$Log_e(T_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \sigma\varepsilon_i \, ,$$

T is the continuous, non-negative random variable representing the survival time, βo is the intercept, $\{\beta_1, \ldots, \beta_p\}$

are the set of coefficients for p covariates(explanatory variables) for the ith individual. σ is the scale parameter and the quantity ε_i is a random variable to model the deviation of the $\log_e(T)$ from the linear part of the model. ε_i follows a particular probability distribution based on the distribution assumed by the time to event under study. The parameters of the regression model are estimated using the maximum likelihood estimation procedure with the help *of iterative Newton-Raphson's process.*

$$Y = ln(T) = x^T \beta + \sigma\varepsilon \Leftrightarrow T = exp(x^T \beta) \times \underbrace{exp(\sigma\varepsilon)}_{T_0} ,$$

where T0 denotes the baseline survival time. Often, the baseline survival time is defined as $T_0 = exp(\beta_0 + \sigma\varepsilon)$. The exponent of the regression coefficients, i.e. $exp(\beta_i)$, (i=1,..,p), are called the **Event Time Ratio(ETR).** An ETR>1 for a respective covariate implies that it slows down or prolongs the time to event, whereas an ETR <1 signifies that occurrence of the event earlier is more likely (compared to the baseline).

The most common survival distributions with the **AFT** class are Exponential, Weibull, Standard Gamma, Log-Normal, Generalized Gamma and Log-Logistic. The following table shows how some of the distributions of the term is based on the distribution assumed by the survival time T:

Distribution of error terms	Distribution of T (survival time)
Standard Gumbel (minimum) with σ ≠ 1	Weibull
Standard Logistic	log-logistic
Standard Normal	log-normal
Standard Gumbel (minimum) with **σ = 1**	Exponential
Log-Gamma	Gamma

Table 1: Different forms of AFT models

3.2 Assumptions

Example

Let $S_F(t)$ and $S_M(t)$ denote the survival functions of females and males, respectively (in fig. 3, G=2 refers to females and G=1 refers to males).

$S_M(t) = S_F(\gamma t)$, where $\gamma > 0$ is a constant named acceleration factor. The AFT assumption can also be expressed as $\gamma T_M = T_F$, where TM is a random variable representing the survival time for males and T_F is the analogous one for females.

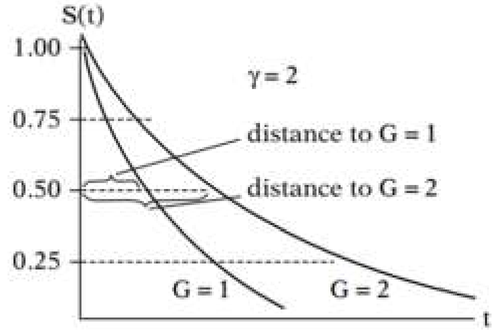

Figure: Acceleration factor γ as ratio of time-quantiles corresponding to any fixed value of $S(t)$. For $\gamma > 1$ ($\gamma < 1$): exposure benefits (is harmful to) survival for Group $G = 2$.

Figure 3: Effects of the Acceleration Factor

The acceleration factor assists in evaluating the effect of the predictor variables on the survival time.

The censoring time is assumed to be independent of the survival time. If the assumptions for the censoring and survival distributions are correct, then a plot of either the censored or the non-censored values (or both together) against time should show no particular patterns (fig. 4).

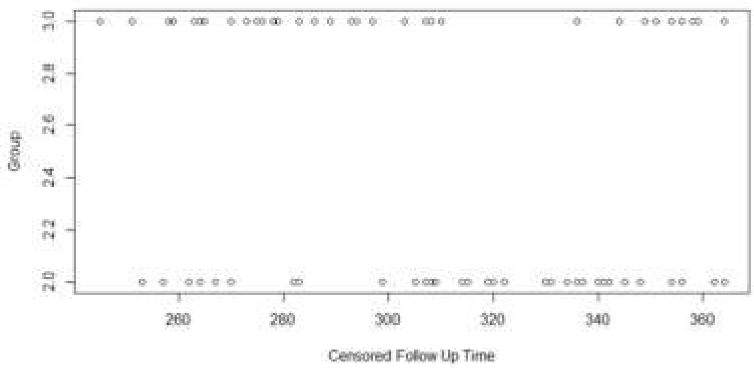

Figure 4: A scatter plot of a stratified sample of the censored time across two groups of customers (Male=3, Female=2) shows random censorings.

Also censoring rates are checked to a proportion of 0.5. The AFT analysis is based on maximizing the likelihood using Newton-Raphson's algorithm. High censoring rates do not usually affect the model results or estimates, but might increase the non-convergence rates(NR) in case of AFT. For a censoring rate as high as 0.9, the NR can be increased to about 6% [Survival Analysis, Mengwei Fei,2012].

3.3 Different forms of the AFT model based on the distributions assumed:

Exponential AF

- TAFT model with σ = 1: $Y = ln(T) = x^T\beta + \varepsilon$ where ε follows the standard Gumbel (minimum) distribution, denoted as G(0, 1), with density

$$f\varepsilon(\varepsilon) = exp(\varepsilon - exp(\varepsilon))\ for\ \varepsilon \in R.$$

- Density of survival time T:

$$f_T(t) = exp(-x^T\beta)\ exp(-(t\ exp(-x^T\beta)))$$

Weibull AFT:

- AFT model with $\sigma \neq 1$: $ln(T) = x^T\beta + \sigma\varepsilon$, where ε follows the standard Gumbel (minimum) distribution.
- The Weibull regression model is an AFT model that has proportional hazards.
- $T \sim W(\rho,\gamma)$ where $\rho = e^{x^T\beta}$ and $\gamma = \frac{1}{\sigma}$. Survival Function of T is thus
-

$$S_T(t) = \exp[-\left(\frac{t}{\exp(x^T\beta)}\right)^{\frac{1}{\sigma}}]$$

- Expected survival time of $W(\exp(x^T\beta), \frac{1}{\sigma})$ is given as

$$E(T) = exp(x^T\beta)\Gamma(\sigma + 1)$$

- **Log-logistic AFT:**
- $ln(T) = x^T\beta + \sigma\varepsilon$, where ε follows the standard logistic distribution with density

$$f_\varepsilon(\varepsilon) = \frac{exp(\varepsilon)}{(1 + exp(\varepsilon))^2}\ for\ \varepsilon \in R.$$

- T has a log-logistic distribution with parameters α & γ.
- In the log-logistic model, the regression coefficients can be expressed in such a way that they can be interpreted as odds ratios.

Log-normal AFT:

- $\varepsilon \sim N(0, 1)$
- $Y = ln(T) \sim N(x^T\beta, \sigma^2)$ with

$$h_Y(y) = \frac{1}{\sigma}\frac{\varphi\left(\frac{y - x^T\beta}{\sigma}\right)}{1 - \Phi\left(\frac{y - x^T\beta}{\sigma}\right)},$$

where $\varphi(\cdot)$ and $\Phi(\cdot)$ denote the pdf and cdf of the standard normal distribution, respectively.

- $T \sim LN(x^T\beta, \sigma^2)$ with $h_T(t) = \frac{1}{t}h_Y(ln(t))$.

Likelihood functions:

$$L(\theta|D_n^t) = \prod_{i=1}[f_i(t_i|\theta)]^{\delta_i}[S_i(t_i|\theta)]^{\delta_i}$$

$$= \prod_{i=1}^n[h_i(t_i|\theta)]^{\delta_i}S_i(t_i|\theta),$$

where $\theta = (\beta, \sigma)$ is the vector of unknown parameters and δ_i is the censoring indicator for the ith observation.

Log-likelihood:

$$l(\theta|D_n^t) = \sum_{i=1}\left[\delta i \ln\left(f_i(t_i|\theta)\right) + (1 - \delta i)\ln\left(S_i(t_i|\theta)\right)\right]$$

$$= \sum_{i=1}^n[\delta i\ ln(h_i(t_i|\theta)) + ln(S_i(t_i|\theta))].$$

3.4 Confidence Intervals:

The inverse of the observed Fisher information matrix provides estimators of the variances and covariances: $\widehat{Cov}(\hat{\beta}) = I(\hat{\beta})^{-1}$. Typically, software packages provide estimates of the standard errors of each of the model coefficients, which are the square roots of the elements on the main diagonal of $\widehat{Cov}(\hat{\beta})$. The endpoints of a $100(1 - \alpha)\%$ confidence interval for the jth coefficient are

$$\hat{\beta}_j \pm z_{1-\frac{\alpha}{2}} - s.e.(\hat{\beta}_j) ,$$

where s.e.($\hat{\beta}_j$) denotes the standard error of the estimator of the coefficient j.

3.5 Assessing The Point Prediction Accuracy:

Parkes has suggested a method for assessing the accuracy of the survival time. Let 't' be the observed survival time and p be the predicted time. If $p/k \leq t \leq kp$ then the point prediction p is defined as "accurate", and any value outside the interval is "inaccurate". Parkes proposes $k = 2$ as suitable. Here the accuracy of the prediction is compared using Parkes' method keeping k at 2. (Appendix A)

4. Data Description:

The historical customer service data was provided by a fashion retail client to investigate the churn rates for the next period and predict when and which customers are likely to churn out. A total of 18 variables were taken into account, namely:

Customer Gender	Units per transaction for four different spend buckets
Total no. of transactions within the observation period	Units per transaction
No. of distinct shops visited (Categorical)	Average Transaction Value
Preferred Merchandise Gender (Categorical)	Recency, meaning no. of days passed after last transaction date (at the end of the study)
Latency, meaning average gap between visits	Season specific variables such as summer sales proportion and winter sales proportion

Table 2: Overview of Variables Used

The usual data preparation including outliers capping for variables like transaction count and invoice value, were carried out. Survival data format usually requires two columns - apart from the explanatory variables 'the time to event' or 'observed time' and 'the status.' The status is primarily an indicator variable that accounts for censored data. It takes the value 1 to denote event occurrence or 0 for right censored.

ID	Gender	Txn	latency	spend 0_200	spend 2_01_400	UPT	ATV	Recency	Winter Sales	Summer Sales	time	status
1	F	4	97.00	0.00	4.30	3	237.68	55	0.60	0.39	309	0
2	M	7	20.00	0.00	1.00	3	1009.6	18	0.00	1.00	346	1
3	M	8	69.80	0.50	1.00	2	771.07	2	0.21	0.79	362	0
4	F	6	35.25	1.00	1.33	1	294.79	217	1.00	0.00	147	1
5	F	35	16.32	0.40	1.80	2	618.17	4	0.36	0.64	360	0
6	F	6	73.25	2.00	1.00	3	755.95	47	0.54	0.46	317	0
…	…	…	…	…	…	…	…	…	….	…	…	…

Table 3: A data snippet

5. Results:

Three different models were examined for reasons stated earlier, with the explanatory variables being the same for all three cases. AIC was used to look for the best model.

Model Dist.	Loglikelihood	Newton-Raphson Iterations	AIC	Prediction Accuracy	Scale Parameter
Weibull	-3615633	13	7231321	70%	0.171
Lognormal	-3954732	10	7909519	57%	0.376
Loglogistic	- 3757704	9	7515462	64.5%	0.147

Table 4: Results of the models used

As it is evident, Weibull model is the best fit for the data with the lowest AIC values (Table 4).

The Weibull model parameters and coefficients are as follows:

	Value	Std.Error	Z	p
(Intercept)	7.398657	0.002945	2512.5	< 2e-16
Gender FEMALE	0.009618	0.001544	6.23	4.719E-10
Gender MALE	0.007615	0.001554	4.9	9.58779E-07
tot_txn	-0.00936	0.000159	-58.98	< 2e-16
distinct_shops2	0.008459	0.000699	12.1	< 2e-16
distinct_shops3	-0.01888	0.001228	-15.37	< 2e-16
distinct_shops4	-0.02549	0.002275	-11.21	< 2e-16
distinct_shops5	-0.02819	0.004137	-6.81	9.5E-12
distinct_shops6	0.038137	0.006845	5.57	2.5298E-08
Pref gender OTHERS	-0.00399	0.000467	-8.54	< 2e-16
Pref gender WOMEN	-0.00079	0.006709	-0.12	0.906
latency	-0.00099	1.45E-05	-68.53	< 2e-16
spend_0_200	0.007816	0.000287	27.21	< 2e-16
spend_201_400	0.002083	0.000213	9.77	< 2e-16
spend_401_800	0.000877	0.000158	5.56	2.65292E-08
spend_ab_800	0.001513	0.000118	12.79	< 2e-16
upt	0.016845	0.000223	75.69	< 2e-16
atv	1.95E-07	3.76E-07	0.52	0.604
recency	-0.01123	6.4E-06	-1754.42	< 2e-16

recency	-0.01123	6.4E-06	-1754.42	< 2e-16
Winter UPT	-0.01718	0.000188	-91.35	< 2e-16
Summer UPT	-0.01695	0.000131	-129.86	< 2e-16
Winter latency	0.001452	3.01E-05	48.2	< 2e-16
Summer latency	0.00106	7.52E-05	14.09	< 2e-16
Winter sales_prop	0.056524	0.001866	30.29	< 2e-16
Summer sales_prop	-0.32136	0.002031	-158.19	< 2e-16
Log(scale)	-1.76772	0.000966	-1830.89	< 2e-16

Table 5: Model coefficients and p values

For further validation the residuals plot for all three models were examined (fig.5)

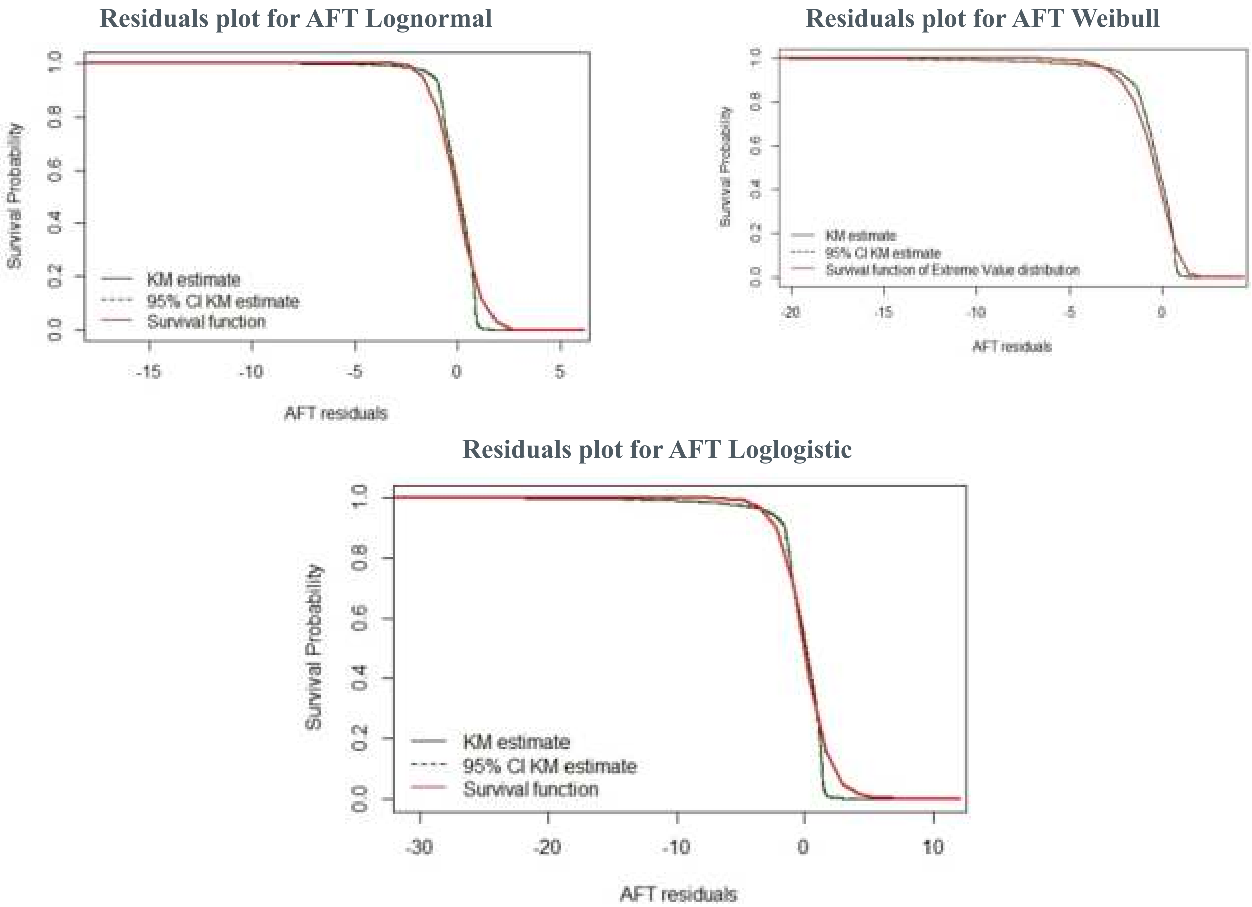

Figure 5: Residuals plot for Weibull, Log-normal, and Log-logistic Models

Diagnostic plot for the Adequacy of the Weibull AFT model:

Weibull model with categorical variables can be checked for its adequacy by stratified Kaplan-Meier curves. A plot of log survival time versus log[–log(KM)] will show linear and parallel lines if the model is adequate (fig. 6).

Interpretation of ETR:

Additionally, the AFT model provides a metric called the Event Time Ratio that helps to understand the covariate effects on survival time. The ETRs, along with the 95% confidence bands for selected few covariates and their interpretations, are presented in Table 6.

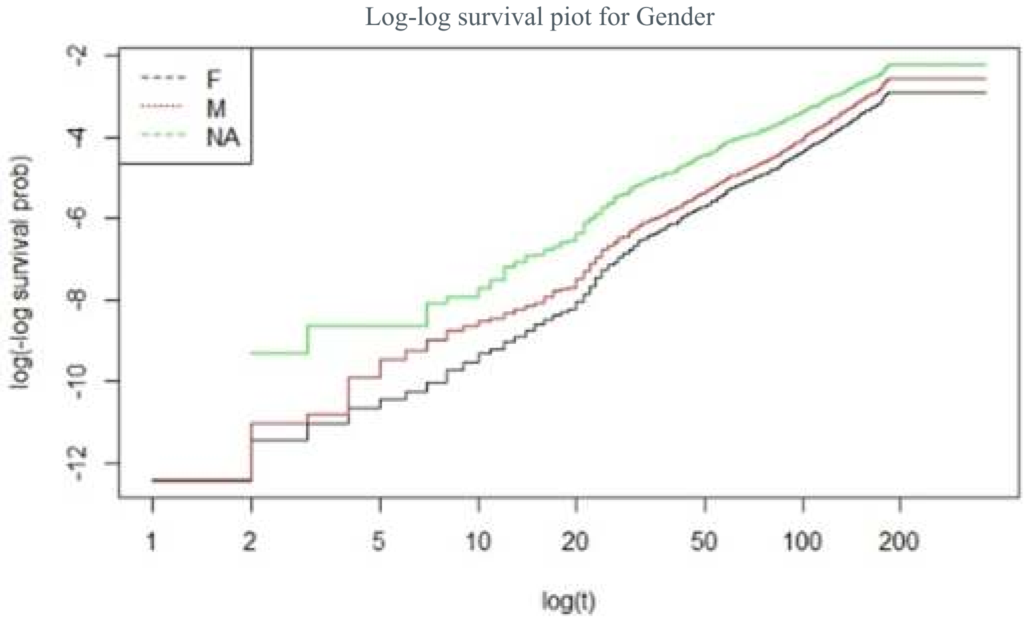

Figure 6: Log-log Survival plot for the explanatory variable Gender

	ETR	LB	UB
Gender FEMALE	1.009664	1.006613	1.012725
Gender MALE	1.007644	1.00458	1.010718
Latency	0.999009	0.99898	0.999037
Recency	0.988831	0.988818	0.988843

Table 6: ETR and confidence bands of covariates

As suggested by the table, female customers tend to show longer survival time than male customers by a small factor of 0.2%. Consequently, an increase in the latency and recency variables tend to decrease the survival time by around 1%.

6. Summary:

Time to churn can be modeled using various Machine Learning models, viz. Random Survival Forest (Random Forest Model applied to Survival Analysis Data), Neural Nets, and several others. Due to sound theoretical roots, here we have used Survival Models (Advanced Statistical Modeling)-both the Semi Parametric and the Parametric version-to model the Time to Churn which would not only provide the estimate of Time to Churn but also the important contributing factors influencing the Time to Churn. The model developed provides a road map to building much more personalized profiles of individuals and helps gain insight into the customer behaviors. With the model, more precise target customers are gained, thereby reducing the overall cost of retaining campaigns, and helps to deliver targeted, optimized offers at the point of impact.

There is a scope of development of the model by introducing more relevant variables and there is a scope of examining the effect of certain covariate groups on the survival time which would eventually make campaigns easier in targeting the right audience. The model can also be further improved if the censored observation proportions could be reduced in a different setup, or a more sophisticated but complex modelling setup might be introduced, taking into account the other types of censoring. The model can potentially be used in predicting a customer's lifetime value (CLTV).

Appendix A:

Prediction Results and Accuracy of a sample of 50 members from the validation set, Mean Time To Failure is the predicted time, and Parkes' Bands give the accuracy measure band.

ID	Mean Time To Failure	Time Observed	Accuracy	Parkes' bands
1	302.9	241	1	(151.5,605.9)
2	426.1	267	1	(213.0,852.1)
3	378.7	244	1	(189.4,757.4)
4	36.8	28	1	(18.4,73.5)
5	232.5	226	1	(116.3,465.1)
6	175.2	166	1	(87.6,350.5)
7	635.7	309	0	(317.8,1271.3)
8	786.6	320	0	(393.3,1573.1)
9	348.9	239	1	(174.5,697.8)
10	560.2	299	1	(280.1,1120.4)
11	532.5	331	1	(266.3,1065.0)
12	203.0	192	1	(101.5,406.1)
13	1028.6	362	0	(514.3,2057.1)
14	574.8	310	1	(287.4,1149.7)
15	983.8	357	0	(491.9,1967.7)
16	583.8	300	1	(291.9,1167.7)
17	781.7	334	0	(390.9,1563.5)
18	335.1	258	1	(167.5,670.2)
19	307.9	231	1	(153.9,615.8)
20	89.1	109	1	(44.6,178.3)
21	418.0	262	1	(209.0,836.1)
22	57.8	69	1	(28.9,115.7)
23	876.8	336	0	(438.4,1753.6)
24	76.7	92	1	(38.3,153.4)
25	94.4	112	1	(47.2,188.9)
26	42.9	43	1	(21.5,85.9)
27	690.6	312	0	(345.3,1381.2)
28	29.1	7	0	(14.5,58.2)
29	354.5	255	1	(177.3,709.0)
30	522.5	323	1	(261.2,1045.0)
31	466.6	288	1	(233.3,933.2)
32	57.5	68	1	(28.8,115.0)
33	682.7	329	0	(341.3,1365.3)
34	404.8	255	1	(202.4,809.5)
35	179.5	188	1	(89.7,358.9)
36	943.0	354	0	(471.5,1885.9)

37	372.6	254	1	(186.3,745.2)
38	515.1	293	1	(257.5,1030.2)
39	260.1	211	1	(130.0,520.2)
40	657.3	309	0	(328.7,1314.7)
41	1274.4	357	0	(637.2,2548.9)
42	205.8	218	1	(102.9,411.7)
43	605.3	298	0	(302.7,1210.7)
44	599.3	289	0	(299.6,1198.5)
45	692.1	337	0	(346.0,1384.1)
46	297.5	249	1	(148.8,595.1)
47	30.8	12	0	(15.4,61.5)
48	802.2	324	0	(401.1,1604.4)
49	1038.2	364	0	(519.1,2076.5)
50	55.6	65	1	(27.8,111.1)

References:

[1] Enwu Liu and Karen Lim (2018), Using the Weibull accelerated failure time regression model to predict time to health events

[2] MingweiFei (2006), A Study Of The Robustness Of Cox's Proportional Hazards Model Used In Testing For Covariate Effects, China

[3] TorbenMartinussen and Thomas H.Scheike (2006),Dynamic Regression Models For Survival Data

[4] Shankar Prasad Khanal, V. Sreenivas and Subrat K. Acharya (2014), Accelerated Failure Time Models: An Application in the Survival of Acute Liver Failure Patients in India, International Journal of Science And Research, Volume 3 issue 6

[5] Steffen Unkel (2018), Analysis of Time to Event Data: Parametric Regression Models, Chapter4

[6] Zhongheng Zhang (2016), Parametric regression model for survival data: Weibull regression model as an example