ATTRITION MODELING FOR ONLINE MEDIA USERS BY COX PROPORTIONAL HAZARDS

STUDENT:

MICHAEL OCHOLA(I56/8387/2017)

SUPERVISOR:

DR. IDAH OROWE

Introduction

- Attrition or churn is the reduction in the number of subscribers of a particular service ranging from banking, telecommunication or online services.
  It is easy to define churn for banking and telecommunication industries.
  It is much harder for online web aplications

Introduction

- Attrition or churn is the reduction in the number of subscribers of a particular service ranging from banking, telecommunication or online services.
  It is easy to define churn for banking and telecommunication industries.
  It is much harder for online web aplications
- A User refers to a writer with an active account on the platform having published at-least an article for the beginning of study period.

Introduction

- Attrition or churn is the reduction in the number of subscribers of a particular service ranging from banking, telecommunication or online services.
  It is easy to define churn for banking and telecommunication industries.
  It is much harder for online web aplications
- A User refers to a writer with an active account on the platform having published at-least an article for the beginning of study period.
- Online media refers to digital journalism anchored on the internet and avails news using smart phones, tablets and computers

Objectives

General objective is to identify various factors that are contributing to customer attrition related to survival time

- To develop a user attrition model using Cox regression.

Objectives

General objective is to identify various factors that are contributing to customer attrition related to survival time

- To develop a user attrition model using Cox regression.
- Estimate retention probability of a user

Objectives

General objective is to identify various factors that are contributing to customer attrition related to survival time

- To develop a user attrition model using Cox regression.
- Estimate retention probability of a user
- Estimate relative risk of churn using significant covariates

Significance of the Study

- Defining churn in a web applications is very dynamic process since the duration of users inactivity varies from one user to another as well as from one application to another.

Significance of the Study

- Defining churn in a web applications is very dynamic process since the duration of users inactivity varies from one user to another as well as from one application to another.
- It is difficult to forecast the future user behavior for such websites by relying on a single estimate by GA although,by using various user retention deter- minants, it is possible to develop a model that can help predict user attrition rates in the future.

Significance of the Study

- Defining churn in a web applications is very dynamic process since the duration of users inactivity varies from one user to another as well as from one application to another.
- It is difficult to forecast the future user behavior for such websites by relying on a single estimate by GA although,by using various user retention deter- minants, it is possible to develop a model that can help predict user attrition rates in the future.
- Operational versions of this model would support user retention campeigns and strategies.

Review of Litrature

The various studies around churn, modeling techniques and findings are summarised.

- James(2012) investigated churn on safaricom subscribers using Cox regression and Decision trees to determine churn determinants. Study found competitor activity as a major determinant of subscriber churn.

Review of Litrature

The various studies around churn, modeling techniques and findings are summarised.

- Shyam (2010) studied customer churn in the wireless telecommunication industry using Naive Bayes algorithm. Data mining helped the researcher in pulling and making use of the fifty thousand records without sampling. Determinants were poor coverage, dropped calls, competitor marketing activities. The model validation test produced 68% accuracy.

Review of Litrature

The various studies around churn, modeling techniques and findings are summarised.

- John Hadden et al (2007) researched on most popular algorithms for building customer churn models, the pros and cons of the various techniques, assumptions and model validation. LR and DT are among the popular.

Review of Litrature

The various studies around churn, modeling techniques and findings are summarised.

- Godsway R (2012) studied churn in Vodafone mobile telcos using Cox regression. Users segmented as High middle and low value and tested significant difference of the survival curves using Log rank test. Cox regression to measure the magnitude of hazard risk for the three groups.

Review of Litrature

The various studies around churn, modeling techniques and findings are summarised.

- Alain et al 2017 in studied churn prediction in mobile social games using survival analysis. Study settled on survival to manage censoring problem. Model used to find when(time) player churned and associated factors. Failure to connect to the game for 10 consecutive days qualified as churn. Cox regression was fitted with a ROC curve having AUC of 0.96.

Review of Litrature

The various studies around churn, modeling techniques and findings are summarised.

- Ali T.J (2009) studied churn in telcos targeting pre-paid subscribers.Three clusters first-class, business and economy based on spending independently modeled using DT with economy experiencing high churn.

Methodology

Survival Models - Used to analyze data with response variable being time until the occurence of an event. Capable of managing censoring unlike other regression techniques.

- Survival function :

$$S(t) = 1 - F(T \leq t) \tag{1}$$

$$\mathrm{F(t)} = Pr(T \leq t) = \int_{x=0}^{t} f(x)\,dx$$

$$S(t) = 1 - \int_{x=0}^{t} f(x)\,dx$$
$$S(t) = \int_{t}^{\infty} f(x)\,dx \tag{2}$$

- Hazard function:

$$h(t) = \lim_{\delta t \to 0} \frac{P\{t < T \leq t + \delta t / T > t\}}{\delta t}$$

- Hazard function:

$$h(t) = \lim_{\delta t \to 0} \frac{P\{t < T \leq t + \delta t / T > t\}}{\delta t}$$

- $$h(t) = \frac{f(t)}{S(t)} \tag{3}$$

- Hazard function:

$$h(t) = \lim_{\delta t \to 0} \frac{P\{t < T \leq t + \delta t / T > t\}}{\delta t}$$

- $$h(t) = \frac{f(t)}{S(t)} \tag{3}$$

- Therefore $h(t)$ becomes

$$h(t) = -\frac{d}{dt} log S(t) \tag{4}$$

Kaplan-Meir Estimator of $s(t)$

- $$S(\hat{t}) = \Pi_{j=1}^{k}(\frac{n_j - d_j}{n_j}) \tag{5}$$

Kaplan-Meir Estimator of $s(t)$

- $$S(\hat{t}) = \Pi_{j=1}^{k}(\frac{n_j - d_j}{n_j}) \quad (5)$$
- Median survival time $S(t) = 0.5$

Kaplan-Meir Estimator of $s(t)$

- $$\hat{S(t)} = \Pi_{j=1}^{k}\left(\frac{n_j - d_j}{n_j}\right) \tag{5}$$
- Median survival time $S(t) = 0.5$
- Log-Rank test $H_0 : h_1(t) = h_2(t)$ vs $H_1 : h_1(t) \neq h2(t)$
  Under H_0 each group $i = 1, 2$ follows a hypergeometric distribution with parameters N_j,N_{1j} and O_j

The distribution has expected value E_{ij} as

$$E_{i,j} = O_j \frac{N_{i,j}}{N_j}$$

Variance as

$$V_{i,j} = E_{i,j} \left(\frac{N_j - N_{i,j}}{N_j} \right) \left(\frac{N_j - O_j}{N_j - 1} \right)$$

Finally Log rank test compares O_{ij} to its expectation E_{ij} under H_0

$$Z_i = \frac{\sum_{j=1}^{J} (O_{i,j} - E_{i,j})}{\sqrt{\sum_{j=1}^{J} V_{i,j}}} \xrightarrow{d} \mathcal{N}(0, 1)$$

Cox Proportional Hazard regression

The one major assumption for Cox regression is that the hazard of death of an individual at any given time in one group is proportional to the same time point in another group. This proportionality on the hazard function of two ensures that survival functions do not cross one another.

- $$h(t, X) = h_o(t).exp(B'X)$$

  Where

  - $h(t, X)$ represents the hazard of users churn or attrition with characteristic X
  - $h_o(t)$ User hazard function at $X = 0$ also referred to as baseline hazard function.
  - $B'[B_1, B_2, ...B_K]$ is the regression coefficient vector.

Cox Proportional Hazard regression

The one major assumption for Cox regression is that the hazard of death of an individual at any given time in one group is proportional to the same time point in another group. This proportionality on the hazard function of two ensures that survival functions do not cross one another.

- $$h(t, X) = h_o(t).exp(B'X)$$

  Where

  - $h(t, X)$ represents the hazard of users churn or attrition with characteristic X
  - $h_o(t)$ User hazard function at $X = 0$ also referred to as baseline hazard function.
  - $B'[B_1, B_2, ...B_K]$ is the regression coefficient vector.

- $$h(t, X)) = h_o(t) + exp(B_1X_{i1} + B_2X_{i2} + ... + B_kX_{ik}) \tag{6}$$

Data Extraction and Transformation from the database

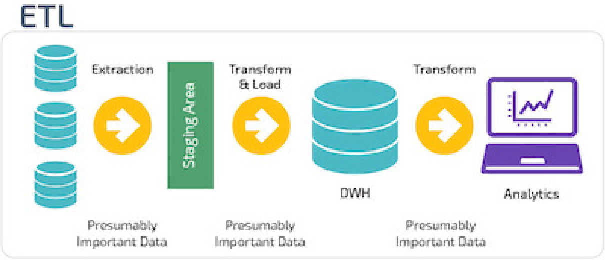
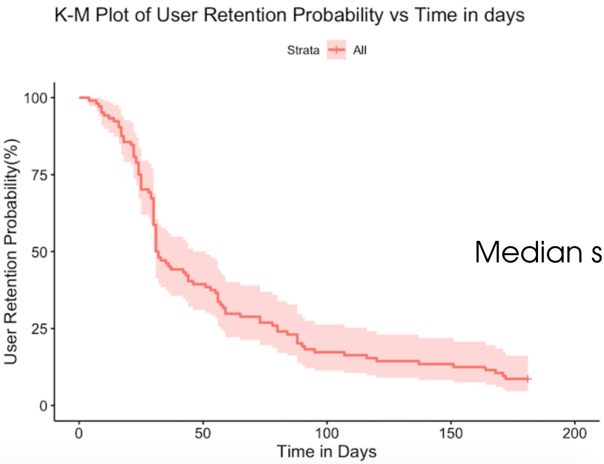


Figure: An ETL Process Diagram

Analysis and Results



Median survival time = 30 days

Figure: Kaplan Mier graph on user retention probabilities over time

Significant Covariates

Hypothesis Testing using Log Rank

- Gender

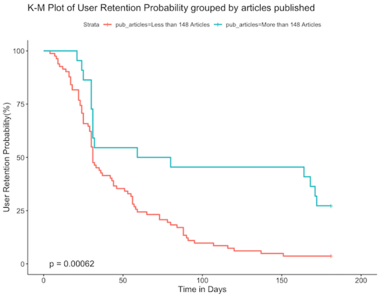
$$H_0 : h(\text{Female}) = h(\text{Male}) \;\; \textbf{vs} \;\; H_1 : h(\text{Female}) \neq h(\text{Male})$$

```
Call:
survdiff(formula = Surv(time = churn_status2$churn_by, event = churn_status2$status) ~
    Gender, data = churn_status2)

          N Observed Expected (O-E)^2/E (O-E)^2/V
Gender=F 28       27     19.6      2.81      3.76
Gender=M 76       68     75.4      0.73      3.76

 Chisq= 3.8  on 1 degrees of freedom, p= 0.05
```

Significant Covariates

Hypothesis Testing using Log Rank

- Number of articles published



Significant Covariates

Hypothesis Testing using Log Rank

- Category of articles published by writer

$$H_0 : h(\text{Political}) = h(\text{Non-political}) \;\; \textbf{vs} \;\; H_1 : h(\text{Political}) \neq h(\text{Non-p}$$

```
Call:
survdiff(formula = Surv(time = churn_status2$churn_by, event = churn_status2$status) ~
    category, data = churn_status2)

                               N Observed Expected (O-E)^2/E (O-E)^2/V
category=Lifestyle and Others 64       61     49.9      2.47      5.57
category=Politics             40       34     45.1      2.73      5.57

 Chisq= 5.6  on 1 degrees of freedom, p= 0.02
```

Table: Log rank test statistic on articles categories

Significant Covariates

Hypothesis Testing using Log Rank

- Other covariates were not significantly different using LR test

All covariates

```
Call:
coxph(formula = surv_object ~ time_spent_category + Gender +
    pub_articles + rej_articles + category + location_category +
    level_of_educ, data = churn_status2)

                                              coef exp(coef) se(coef)      z      p
time_spent_categoryMore than 250 days     -0.21397   0.80737  0.23945 -0.894 0.3716
GenderM                                   -0.46385   0.62886  0.25843 -1.795 0.0727
pub_articlesMore than 148 Articles        -0.81703   0.44174  0.31967 -2.556 0.0106
rej_articlesMore than 10 Articles rejected -0.41056   0.66328  0.29698 -1.382 0.1668
categoryPolitics                          -0.14005   0.86931  0.24919 -0.562 0.5741
location_categoryNational                  0.05779   1.05949  0.22241  0.260 0.7950
level_of_educUniversity                   -0.12378   0.88358  0.24626 -0.503 0.6152

Likelihood ratio test=21.48  on 7 df, p=0.003119
n= 104, number of events= 95
```

Table: Cox regression and coefficients of various covariates.

Conclusion and Recommendation

- Conclusion Survival analysis was able to explain various covariates and their effects on writer attrition at hivisasa.com. Main determinants being gender, category of an article and number published by a writer.

Conclusion and Recommendation

- Conclusion Survival analysis was able to explain various covariates and their effects on writer attrition at hivisasa.com. Main determinants being gender, category of an article and number published by a writer.
- Recommendation The median survival time of 30 days indicates that more can be done to increase the number to at-least 90 days. More research around churn in web applications since this area has not been fully exploited.