University of Nairobi

School of Mathematics

ISSN: 2410-1397

Master Project in Social Statistics

Attrition Modelling for Online Media Users By Cox Proportional Hazards

Research Report in Mathematics, Number 44, 2019

Michael O Ochola

Dec 2019

Submitted to the School of Mathematics in partial fulfilment for a degree in Master of Science in Social Statistics

Attrition Modelling for Online Media Users By Cox Proportional Hazards

Research Report in Mathematics, Number 44, 2019

Michael O Ochola

School of Mathematics
College of Biological and Physical sciences
Chiromo, off Riverside Drive
30197-00100 Nairobi, Kenya

Master Thesis

Submitted to the School of Mathematics in partial fulfilment for a degree in Master of Science in Social Statistics

Submitted to: The Graduate School, University of Nairobi, Kenya

Abstract

In this work survival analysis was used to analyse propensity to churn for online writers of a news website in Kenya known as hivisasa.com. The study sought answers on which covariates were major determinants of writer attrition from the online platform, their statistical significance and magnitude. A total of one hundred and four writers who had at-least one publication for January 2019 formed part of the study sample. The sample historical data for January to July was used to determine writers who churned within the period and those that were retained. Previous literature on attrition research was reviewed and the study settled on survival methods in order to address time to event and manage censored data. Descriptive analysis was handled by fitting Kaplan-Meier curves to visualize the retention curves of various categories of the covariates. Log-Rank test was then used to test the statistical significance of the various differences observed. Cox proportional hazard was fitted on the data including all covariates to determine the magnitude of hazard risk. Three of the covariates that is gender, number of articles published by a writer and category of articles done by the writer were significant in explaining writer attrition risk and magnitude. The results showed a high churn rate among female writers, writers publishing non political content on the site as well as publishing less than 148 articles for the study period. On the other hand three covariates; time spent on the platform from subscription, location of a writer and level of education were not statistically significant in explaining writer attrition. Even though these covariates lacked statistical significance Cox regression coefficients revealed that the magnitude of risk varied across them. Level of education graduate and time spent on the platform of more than 250 days reduced the chances of a writer churning 12% and 19% respectively in comparison to the reference variable holding for the effect of other covariates. The model performance was validated by fitting a ROC curve to ascertain how best the model was able to fit the data. The ROC curve had an AUC of 87% which means the model had a 87% chance of predicting a churned writer as so.

Master Thesis in Mathematics at the University of Nairobi, Kenya.
ISSN 2410-1397: Research Report in Mathematics

DISTRIBUTOR: School of Mathematics, University of Nairobi, Kenya

Declaration and Approval

I the undersigned declare that this dissertation is my original work and to the best of my knowledge, it has not been submitted in support of an award of a degree in any other university or institution of learning.

Signature Date

MICHAEL O OCHOLA
Reg No. I86/8387/2017

In my capacity as a supervisor of the candidate's dissertation, I certify that this dissertation has my approval for submission.

Signature Date

Dr. IDAH OROWE
School of Mathematics,
University of Nairobi,
Box 30197, 00100 Nairobi, Kenya.
E-mail: orowe@uonbi.ac.ke

Dedication

I dedicate this work to my biggest supporter friend and partner Lora. Your encouragements kept me going on during the long hours of writing this work, my mother and my first class room teacher Grace for her belief in education as a solid differentiator in life.

Contents

Acknowledgments

A big thank you to God the creator for giving me strength and endurance to complete this nobble work.
Special thanks to my supervisor Dr. Idah Orowe for her consistent guidance from introduction to the last chapter indeed i have learned a lot through the journey. I also thank Hivisasa.com for availing data used for this study.....

Michale O Ochola

Nairobi, 2019.

1 Introduction

1.1 Background

Attrition or churn is the reduction in the number of subscribers of a particular service ranging from banking, telecommunication or online services.
Chiu and Young (2006) mentioned that there are three categories of churn:

1. Voluntary churn
   Customers changes or moves to a more prospective providers or all together resigning from both.
2. Involuntary Churn
   Service providers stops availing their services for instance a web application domain unreachable or telecommunication or banking app out of service.
3. Unavoidable churn
   A customer died or a migration to unserviceable localities.

User retention for an on-line site, blog and website is measured by its returning users or visitors a term coined by Google and can be defined as the number of users of a particular website or blog that upon their first interaction with the platform returns within a certain period of time. When a user fails to return to the platform for some pre-defined duration then such a user is considered to have churned.

Customer retention for the online media business is seen as a function of services being offered to the end user who then feels the need to continuously trust the website for future searches. This trust will result into a frequent website visitor.

There has been a shift on how we obtain and receive information from the year 2000 onwards which before then was heavily monopolised by the newspaper, radio and television companies. This shift has led to the emergence of new ways of broadcasting news including on-line media websites blogs and even social media platforms creating the on-line media industry

There are many companies operating on the on-line media ecosystem a factor that has led to competition for the eyeballs, It is only prudent that upon customer or user acquisition much is done to retain such users from the virtual competitor. One such thing would be to understand determinants of user retention and use mathematical modelling to predict the ability to retain new and existing users.

The on-line media industry in Kenya includes Hivisasa.com, Tuko news, Mpasho, Nation media, citizen, star etc. All of this companies wish to control a huge chunk of users who are at liberty to read news from their website of choice. The competition for users is vital for the existence of such businesses that heavily rely on advertising for their operations cost.

Hivisasa.com an on-line media company based in Kenya that was established in 2014 with a weekly reader base of two million on average. It has experienced an exponential growth on its user population over a five year period since its launch. Most of the readers visit the site to read an array of stories ranging from politics, lifestyle, news and business. It has a good number of writers who channel out on average 700 stories weekly targeting various groups of readers. The writers form an important part of the business model since they are the worker bee who tirelessly write relevant content to ensure that readers are fully engaged on the site.

1.2 Problem Statement

Hivisasa ecosystem dictates that user retention is not a negotiable factor for its continued existence since, the process of reader or user acquisition is an expensive one. Reader acquisition cost may include advertising, maintaining a good writer army together with other hidden costs on website development and maintenance as well as user research.

User retention is monitored using in-house metrics such as percentage of returning users calculated as the number of weekly returning visitors divided by total number of weekly visitors times one hundred percent.

Despite having this weekly visual model which can only show a dip or a spike on percentage of returning users there is need to improve on the existing metric to incorporate the major determinants of such a dip or a spike on the returning users by developing a working model.

1.3 Research Question

The question to answer is whether user attrition can be modelled using its determinants. To be able to answer this a sub question is formulated:

1. How Cox regression estimates user's attrition rate over time.

1.4 Objectives

General objective is to identify various factors that are contributing to customer attrition related to survival time.
Specific objectives :

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

1.5 Significance of the Study

Currently, the user retention strategy heavily rely on weekly percentage of returning users estimate obtained from in-house dashboard and Google analytics.

It is difficult to forecast the future user behaviour for such websites by relying on a single estimate by GA although, by using various user retention determinants, it is possible to develop a model that can help predict user attrition rates in the future.

Cox Regression can provide proportions of the online user attrition determinants to help predict future user attrition rates by determining the major predictor variables and their significant levels for user attrition modelling. The organization will be well placed in ensuring low user churn rates and high user retention rates over time with operational versions of these models.

2 Literature Review

2.1 Introduction

This chapter basically looks at various studies that have been previously done around user churn or attrition with a keen eye on the data used as well as the modelling techniques employed by the researchers and possible limitations encountered.

2.2 Attrition Modelling

James Kairanga (2012) investigated churn on mobile telecommunication users for Safaricom with an aim to determine factors that predispose a mobile phone subscriber to churn from the service and possibly move to a competitor. The study used Cox proportional hazard model and decision tree to try and determine factors that predispose mobile subscribers to churn. Proportional hazard model is based on survival analysis in statistics while decision tree is popularly common in data mining. The researcher also explored how better the two models;Cox proportional hazard and Decision tree compares. The study found that Cox Hazard model and Decision trees performed much better in churn prediction compared to the standard models used by the company to predict churn. Comparing the performance of Cox and Decision tree models, Decision tree was reported to have performed slightly better than Cox model in this churn prediction case. The decision tree gave out the probability of churn while the standard model only gave whether one is likely to churn or not. In summary the study recommended the use of the two models in determining propensity to churn with one of independent variables being competitor monthly activities that could probably contribute to churn.

Shah Yuan et al (2017) studied users on a popular online learning app in China to determine factors that predispose users to churn so that products' team can optimise their services and prevent churn before it happens taking into account that new user acquisition cost is six times compared to retention. The study sampled a population of 100 thousand users and defined churn as inactivity on the platform for registered users for a period of one week. Due to the nature of heterogeneity in the data the study opted for male and female separate homogeneous datasets as well as new and existing users datasets. The churn prediction model applied was logistic regression on the four datasets. The study found two columns to be more popular than the rest of the columns and the model is reported to be capable of accurately predicting 75 percent of churners.In the column of film is reported to attract more males while the column on beauty attracted more females. In terms of new

and existing users the study reported no significance difference in mean reading time for the various columns. The study used a naives Bayes classification to compare the results obtained using the binary logistic and calculates an F score for the two algorithms. The binary logistic model performs better in predicting user churn compared to Naive Bayes because of the high F score registered. F score for evaluating the model is calculated as $F = 2PR/P + R$ where Precision aims to get what proportion of positive identification was correct against Recall which looks at proportion of actual positives identified correctly.

Shyam (2010) studied customer churn in the wireless telecommunication industry using data mining approach. Data source for the study was an oracle database for 50 thousand users. The researcher modelled churn using Naive Bayes Algorithm, data mining helped the researcher in pulling and making use of the fifty thousand records without sampling. Data mining is useful when dealing with data that in most cases is non structured and grows in velocity and volume. Naive Bayes uses previous data to make predictions on the new dataset. The study categorised input variables for the churn prediction as demographics which categorised data as population and geographic instances, secondly the level of usage included data on dates, timestamps, duration and place of usage. Thirdly service quality variables quantified number of dropped calls, poor coverage and interference and Lastly marketing and features such as recent entry of competitor to advertising campaign and instant messaging respectively
The study used oracle database to administer churn prediction using Naives Bayes algorithm availed by the Oracle database by importing data with above described variables. The model was tested among 2000 users and produced a 68% accurate prediction. The study recommended that a remedy is put in place for customers with higher probabilities of churn in order to ensure customer retention.

John Hadden et al (2007) researched on most popular algorithms for building customer churn management models. The study then reviewed the pros and cons of various modelling techniques for user attrition. The motivation for these work was to allow companies to choose and build for themselves a churn management model. The study had the following core areas. First the steps that can guide one in creating churn management solution. secondly, the study outlined the commonly used techniques like identifying the best dataset with the right variables for churn prediction or management, particularly the need to identify data that suits the type of analysis one intends to perform. An example would be purchasing of products and services being explained best by historical purchase data. All this care and thought process serves to ensure the power and accuracy of the overall model. Thirdly, a key input detailing the less research in the area of churn as well as probable challenges. Feature selection that basically entails choosing the best variables for the model. The section discussed the need to only have variables that are significant in building the model and leave any variable that is noisy through data cleansing and

reduction. The study explains as Sequential forward selection where variables are added to the model based on significance and another approach where all possible variables are added later on removed based on non significant contribution to the overall model this procedure is referred to as Sequential Backward selection. Lastly researchers detailed methods for Validating the results obtained. The study also provided an overview of the various predictive modelling across the industry. Research findings show that decision tree was quite popular. Decision tree model building entails tree building and pruning. Tree pruning involves removing branches that could have the most noise hence it increases the accuracy score of decision tree. Researcher acknowledges that decision tree is just as good as the variable selected and rules used in classification case. Regression especially logistic regression is widely used for predicting customer churn. The one challenge with this modelling approach is that it only provides the probability of occurrence or non occurrence of the event of interest. There are quite a number of papers showing that logistic regression performed much better than neural networks and decision trees for the specific case scenarios. Other researchers used predictive models like Naive Bayes based on Bayes Theory, KNearest Neighbour and Neural networks. The study also reported common churn model validation techniques:-

- Segmenting the data into seventy thirty percent ratio where seventy percent of the data for model training and thirty percent for model validation. This validation method is mostly applicable where data is scarce.
- A different validation approach entails having a different dataset with a different model for instance decision tree against logistic regression to validate the output of the model being tested. This is applicable where data is readily available.
- Cross Validation approach divides the data randomly into the training and validation sets, this is also referred to as Monte Carlo cross validation technique.

Ali Tamaddoni Jahromi (2009) studied churn prediction in telecommunication industry with a target on pre-paid subscribers. Churn definition for pre-paid customer is not very formal as per the post-paid scenario where a contract exists between the subscriber and the firm therefore churn can be identified easily. Anyone can be a subscriber and the next thing a churner. The study used clustering to segment users on homogeneous characteristics upon which prediction modelling was undertaken. The researcher used Decision tree for model building where each of the clusters were modelled separately. The choice for decision tree according to the researcher was in its simplicity in predicting whether a subscriber will be a churner or not.

The prediction phase used 70:30 percent division of training and testing data sets for each of the four clusters. The whole customer base subjected to clustering was thirty four

thousand records which was sufficient data. Each of the clusters seventy percent was subjected to decision tree modelling while the thirty percent reserved for model validation. Results from the study showed a varying churn prediction capability for the four clusters. In conclusion the study was able to help the the telcos predict churners and by so doing improving on retention campaign and consequently reduced marketing cost.

Kojo Abiw (2011) studied churn prediction in mobile telecom industry for MTN in Ghana. The research acknowledges that an industry with low switching cost is more predisposed to subscriber churn as compared to an industry with high switching cost therefore this in itself predisposes telecom industries to high churn from its subscribers. The study sought to find out the major determinants of churn for MTN customers in Ghana and develop a prediction model based on the determinants
A sample of 3333 subscribers was used by the researcher in training the model which used neural network to determine propensity to churn and decision trees to understand the behaviour of churning subscribers. Different from churn prediction model, the researcher sought out major causes of churn by sampling and interviewing 56 respondents using snow ball sampling technique. Study results show major churn causes as poor network quality, competitors promotion, poor customer care activities and other internal factors. In conclusion the research used neural networks to pinpoint users with high propensity to churn while the rule generated by decision tree helped to explain customer churn. Researcher recommends the use of the model in MTN for subscriber retention campaigns.

Alain Saas et al 2017 in studying churn prediction in mobile social games used survival analysis. The researcher mentioned that in mobile social games decrease in churn is equivalent to increasing player retention. Churn understanding help in measuring player loyalty and to estimate when they are likely to stop playing the game with plans initiated to curb churn before it happens. The researcher settled for survival analysis since not all players in the data set would have experienced the event or churn. Censoring is a normal churn problem, the use of regression is limited hence the researcher opted to use survival analysis which already can handle the censoring problem. The study emphasise the fact that churn is not clearly or explicitly defined as other subscription products like banking, telecommunication and e-commerce where churn is rightfully captured by un-subscription from the service but the model used here is a free to play no subscription for a player before hand. Survival analysis was used to find out when players churn and the risk factors associated with player attrition. For these study the researcher defined churn as a player who fails to connect to the game for 10 consecutive days. The researcher used cox regression to model the risk factors to churn and results of the model tested using ROC curve with 0.96 as AUC value with R survival package. Improving on the prediction accuracy of the model could form sections for further studies in the area of mobile game

churn.

Godsway Roland (2012) performed a study to investigate churn in mobile telecommunication industry.The study involved telecoms pre-paid subscriber base for Vodafone. Researcher objectives includes a proposition around churn management for telecoms, use Kaplan Meier to estimate survival probabilities of three subscriber segment; High, Middle and Low value customers, test if there is a significant difference between the survival curves for the three segments and lastly to model the contribution of usage variables to the risk of churning using Cox Proportional Hazards. The study used a sample of 15000 subscribers from the vodafone database and implemented a survival analysis and decision trees models. For survival technique data was analysed using:-

1. Non-parametric Kaplan Meier was used to visualise survival times for various segments
2. Semi-parametric Cox regression was used to explain the effect of various covariates on the likelihood of churning
3. Log rank used to test the difference in survival curves for the three customer segments

In concluding the researcher noted that there was a significant difference among the three segments of subscribers in terms of survival probabilities with the low value customers having a higher churn rate as compared to the high value customers. Median survival time which refers to the time it takes for half of the subscribers to churn was found to be around 12 months in low value group and 28 months in the high value segment. On the regression part of the analysis it was reported that top up amount and total minutes of usage had significant impact on probability of churning.

Dang Van Quynh (2019) studied churn prediction in the computer software security industry using a comparative approach by fitting three models to the data under study due to their comprehensibility and predictability namely:

- Logistic regression
- Decision Tree's
- Random Forest's

The researcher performed preprocessing on the data to deal with missing values and eliminate noise then performed exploratory analysis to reveal the underlying features

on the data prior to modelling. The various levels of churn that is churn and no churn exhibited the problem of class skew where the data had 95.4% as non churners with only 4.6% as churners. To address the problem of class skewness the researcher proposed oversampling and under-sampling comparison to pick the most optimal solution. For the modelling case data was split into 70:30 percentage for training and testing data sets respectively. Evaluation of the model used the standard approaches of ROC curve and AUC, confusion matrix for Precision, recall and F-measure. The logistic regression model performed better in predicting non-churner as opposed to the churner showing a clear bias to the majority class as per the confusion matrix. Class imbalance problem was addressed by reducing the non-churners in the sample, the percentage of model accuracy dropped but it resulted in more number of predicted churners than before. Decision tree and Random forest too exhibited better prediction for the majority class as per the logistic regression case. Researcher constructed a second model for both decision tree and random forest addressing the class imbalance. It was found that for balance class models the precision dropped but F score increased with an increase too on the proportion of predicted churners(the minority level).
The study evaluated the performance of the six models using: ROC curve and AUC, which is visual showing the sensitivity(true positive rate) on the y-axis and 1 minus specificity(false positive rate). AUC is an estimation of area under the ROC curve.
Accuracy is calculated as the sum of TP and TN divided by the total elements(TP+TN+FP+FN)
The study noted that accuracy is good measure of models performance but for churn prediction it should be treated with caution.
Precision is the proportion of predicted churners that do churn. Recall is the proportion of actual churners that are correctly determined as churners F-measure is the harmonic average of precision and recall, the closer to 1 F-measure is the better the harmonic average.

2.3 Online Media

According to Ricco Villanueva (2017), the internet is fourty years old. The internet is a collection of smaller computer networks linked to each other, Tim Berners-Lee who is seen as the father of internet describes the internet as "a postcard with an address on it. If you put the right address on a packet and gave it to any computer which is connected as part of the net, then each computer would figure out which cable to send it down next until it reaches its destination." ARPANET pioneered the internet age. This was developed by US department of defence together with various US universities, the aim of the ARPANET project was to enable university researchers collaborate on Defence department projects, The development in information communication technology has transformed the way news is produced, distributed and consumed. Locally, the use of technology has led to various forms of virtual media houses that have made news available all the time without traditionally depending on the mainstream media bulletin hours. There is a complete shift in journalism currently where events can be updated within

minutes of their occurrence something that would have taken hours or even days to be known to the public. Technology and internet access has made everyone a producer and consumer of news. The digital news platform enables users to subscribe to content they are interested in. International media houses update their news content on hourly basis making the subscribed users get their preferred news when they want or few minutes after occurrence something that would only be a dream if not online media birthed by internet technology. By the year 1971 ARPANET project had thirty participating universities, the general public had a slight feel of the computer network a year later at an International Computer Communication Conference by interacting with an electronic mail application.

Then there was a great need to keep track of all the data flowing from one computer to another hence the development of a communication protocols one being the TCP/IP (Transmission Control Protocol/Internet Protocol) and FTP(File Transfer Protocol) allowing connection and file sharing respectively with computers outside ARPANET. The year 1984 going forward other government wings and non governmental organisations joined the network. The next discovery that made the internet complete is world wide web(www) which often gets confused with internet. Email, video, videoconferencing and www are part of the internet among other features. Tim Berners-Lee an MIT scientist developed a new way of communicating on the internet later on using hyperlinks, these slicing of the internet was later referred to us world wide web. Lee's simplifies the understanding of the internet as a collection of computers each having a collection of documents, videos, pictures and even applications, the various computers connect through cables for communication between them. On the web which is like a collection of various computer contents the connection is through hyperlinks. The internet exist because of programs that enable the communication within the web. The web could not exist without the net but again the web made the net useful because at the end of the day people are much more interested in information gained from the web not the communication behind the computers and the cables.

According to Bovil and Livingstone (1999) in studying the effect of new media to young people, there were basically two options either to listen or watch a specific channel and to decide how much. The current technology allows un-limited use of the media where users can make a choice on what to watch and control the content one is consuming.

It's is important to understand online media, Salaverria Ramon (2019) refers to online media as digital journalism and dates it to be around twenty five years. This is journalism anchored on the internet and therefore its full potential will only become a reality in the near future. Online media can be categorised by the medium of access as:

1. Web based access via computer's web browsers
2. Tablets
3. Smartphones

The online media based on smartphones currently outperforms the other two as a result of smartphone's popularity and ability to access application based systems. Most research studies have focussed on the smartphone medium of access to news giving rise to mobile journalism by citizens for the citizens. Online media has also three other ways of presenting information to the user as opposed to traditional media like magazine, radio and television. Online media content takes the form of text, hypertext, multimedia and interactivity. The combination of this forms of content presentation results into a more better way of presenting online news, hypertext links allows users to move back and forth while multimedia enables users to access videos and pictures around the content while interactivity is enhanced through commenting capabilities for feedback and users engagement with themselves and the platform. Economically journalism mostly get their money in running various advertisements something that has been challenged by the emerging of todays internet giants Google and Facebook. The two companies gets the larger pie of advertising revenue making the online media businesses less lucrative for advertising market. The online media is barely surviving with dwindling revenue necessitating a need for research on better ways to ensure user retention models and profitable business models that can uplift the economic side of things for digital journalism.

3 Methodology

3.1 Introduction

Survival analysis is useful for analysing data sets where time until the occurrence of an event of interest is an outcome variable. The event of interest can be anything for instance birth, death, migration, marriage, user churn etc. A researcher can measure the time to the occurrence of the event in years, months, weeks, days and even hours. In comparison to other modelling techniques like logistic regression where the interest is not on the time to the occurrence of an event but whether the event occurs or not. In research it is somewhat important to determine explanatory variables that not only lead to the occurrence of event of interest but also what makes the occurrence of the event of interest nearer or far of in the space of time.

Traditional models like logistic regression fail to address time to the occurrence of an event. It is worth noting that responses not observed within the set observation time are censored. Censoring distinguishes survival analysis from the traditional regression techniques. The survival time of these individuals are approximated to take the duration of the study. The response variable in survival analysis has two groups:-

- Time to the occurrence of event
- Event status i.e occurrence or not

3.2 Survival function

The function denotes the probability that an individual survives longer than time t. It is denoted by S(t)

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

S(t) gives the probability that an individual will survive past time t or a subject will not experience the event of interest for instance churn past a predefined time t. Assuming T is a continuous random variable with a probability distribution function (pdf) as f(t) and a cumulative distribution function F(t)

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

Therefore We have

$$\begin{aligned} S(t) &= 1 - F(t), \\ F(t) = Pr(T \leq t) &= \int_{x=0}^{t} f(x)\,dx \\ S(t) &= 1 - \int_{x=0}^{t} f(x)\,dx \\ S(t) &= \int_{t}^{\infty} f(x)\,dx \end{aligned} \tag{2}$$

It is therefore shown that equation (3.1) can be written as equation (3.2) which is loosely said as the probability of surviving beyond time t.
Hazard function denoted by h(t) measures the occurrence of an event of interest at a particular time point.

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

The conditional probability in the numerator can be rewritten as a joint probability that T is in (t, t+δt) interval and $T > t$. The first conditional probability by definition is $f(t)\delta t$ for a small change in t while the second conditional probability is $S(t)$ by definition.

Therefore

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

Rate of occurrence of an event at t can be calculated as the density of events at t divide by the probability of surviving to time period t without experiencing event. From equation (3.2) we have f(t) as the derivative of $F(t)$.

Therefore $h(t)$ becomes

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

3.3 Censored Data

Censoring or censored basically is data with incomplete information, the information about the data point is partially known, other studies generally treat such scenario as a case of missing data which is basically handled differently for survival analysis. Censoring is divided into three types :-

1. Right Censoring
2. Left Censoring
3. Interval Censoring

Right censoring is attributed to an individual leaving the study before experiencing the event of interest or an individual lost to follow up before the event of interest is recorded. Individuals are only known the lower limit of time. Left censoring is a situation when the event of interest occurred before a particular time with the exact time of the occurrence unknown to the researcher. In interval censoring the occurrence of an event of interest like birth falls within a time range for instance, a week, a month or a year such an event is said to be interval censored. This study shall concentrate on Right censoring, an individual that doesn't experience churn at the end of the set monitoring period will be right censored.

3.4 Kaplan-Meier estimate of The Survival Function

Denoted by $S\hat{(}t)$ KM is used to estimate the probability of surviving between time intervals

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

At each of the intervals survival probability is calculated as the number of individuals surviving divided by the number at risk. The subjects who have experience the event are dropped out and not included as part of the risk for the next interval. Censored data points or individuals are not included in calculating the survival probabilities of different time intervals. Total probability of survival till time t is a product of all probabilities preceding that time. Kaplan Meier curve maps the probability of survival on the y-axis against time points on the x-axis.
This KM plot can be used to determine median survival time of a group by using a survival probability of 0.5. The plot can as well be used to compare survival probabilities of two groups.

3.5 Log Rank test

Log Rank test is used in survival analysis to ascertain whether differences exists in the survival experience of two groups. An example would be the survival experience for males and females who suffers from a disease and followed within a defined time frame.

The null hypothesis states that the survival experience for the two groups (treatment group, control group) are the same while the alternative hypothesis states a difference in survival experience for the treatment and control groups.
It is computed by obtaining observed and the expected number of events in each category either as treatment or control.
Consider patients belonging to treatment and control groups in clinical study.
Let

J = 1...J be times of observed events,
N_{1j} and N_{2j} be the number of subjects at risk at time j
O_{1j} and O_{2j} represents observed events in the groups at time j
$N_j = N_{1j} + N_{2j}$
$O_j = O_{1j} + O_{2j}$

$H_0 : h_1(t) = h_2(t)$ 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

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

For sufficiently large j Z distribution converges to a standard normal distribution by Central limit theorem. Z above will be estimated from the standard normal distribution.

3.6 Cox Regression Model

A technique of modelling time to event data in presence of censored observation. Cox regression has the ability to handle censored observations and estimate the coefficient of the various covariates. It is useful in determining the effect of continuous covariates. 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.
The hazard function can be expressed as a function dependent on time and a function of the covariates

$$h(t,X) = h(t).G(X,B)$$

$h(t,X)$ and $h(t)$ are positive functions therefore $G(X,B)$ is also positive.
The Cox model replaces $G(X,B)$ with $exp(B'X)$.

Therefore

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

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.

Equation then can be rewritten as

$$\log(h(t,X)) = \log(h_o(t)) + B_1X_{i1} + B_2X_{i2} + ... + B_kX_{ik} \quad \textbf{(6)}$$

Equivalent to

$$h(t,X)) = h_o(t) + exp(B_1X_{i1} + B_2X_{i2} + ... + B_kX_{ik}) \quad (7)$$

The equation 3.6 or 3.7 is the Cox Proportional hazard model.

To check for proportionality of hazard for Cox regression model is of importance. It generally means that the ratio of hazard of two individuals is constant over time, in other words they are proportional. One of the simplest way to check the proportionality of the hazard is using the Kaplan Meier curve and checking for intersection of the two curves. When the two curves cross each other then it is an indication that the data violates the proportional hazard assumption. Non proportionality of hazard can be interpreted as an interaction of our independent variable with time.

3.7 Extract Trasform and Load(ETL)

ETL is a process of extraction of data from the source database, loading to the destination storage platform and performing transformations on the data to make it friendly to the next application. Source of data for research studies for a long time has been a function of surveys.

In the internet age data is so much available in various forms in databases across the internet hence the need to understand databases in order to take advantage of these data sources, how they store data and eventually extraction of the data out of which meaningful information can be obtained. Database is considered as a collection of related data while Database Management System(DBMS) as a software that manages and controls access to the database.

An ETL process for pulling and refining the analytics data for these study uses a couple of open source R libraries and the customers database systems. The goal of extraction is to avail data from various storage sources into a single source which will serve to answer questions around the data without having to pull data from the different sources and merging in order to effectively solve a problem.

When extracting data due diligence is followed to ensure:-

- Source data and extracted data are the same.
- Remove all duplicates and fragmented data and check on the variable types
- Validate the existence of identifying variables or keys.

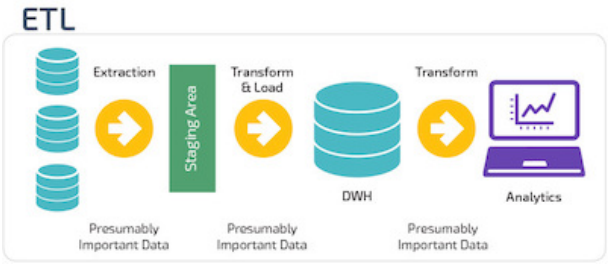


Figure 1. An ETL Process Diagram

The second phase is transformation which is a vital ETL process, the data extracted is in its raw form therefore the need to perform data cleaning, mapping and variable transformations to make it more friendly for visualisation and modelling.

$$\text{Attrition} = f(n : \text{number of weeks}) = \begin{cases} 1 & \text{if } n >= 4 \text{ weeks} \\ 0 & \text{else} \end{cases}$$

Here a set of functions are applied on the data for instance a function returning a variable with a boolean of yes or no that is 1 or 0 for churn or not based on pre-determined conditions set in the function. The number of weeks using a service probably can act as an input to a function to determine whether a user has churned or not based on a condition set to the function. The generated variable is as a result of transformation phase of the ETL process.

Transformation operations would include; filtering of data, variables validation for instance age cannot be 3 digits, cleaning, mapping NULL to 0 or Male and Female to M and F transforming variable types from character to factors if using R for representation of categorical variables, Strategies around handling of missing data.

Loading is the last phase of the ETL process and should be less painful compared to the first two steps. Data warehouse tables are populated either as an initial load which is a one of load of the data frame, incremental load which is periodical and full refresh which erases initial data before writing the new data frame.

3.8 Study Population and Sample

The study population used Hivisasa total writer base as at January 2019. The active writers based on having published an article on the site for January 2019 qualified for inclusion in the sample.

The sample included all users who had at-least a single live publication on the website for the month of January 2019. The number of users who met the above criteria were then followed for a period of 6 months to determine whether they experienced the event of interest(if the writer churned that is he/she stopped publishing articles for the website for a continuous period of 1 month). The group of users who never experienced the event of interest after the 6 months of follow up were right censored. The number of users who met the criteria of at-least a single live publication on the site were 104 users.

3.9 Model Validation

It serves the purpose of knowing how best our model fits the data and how best our model can predict given new data. Most model validation are constructed to determine how best our model is able to predict given a new data. There are a number of methods used in statistical model validation. This study will basically rely on ROC curve to validate how the cox model best fit new data sets

3.9.1 Receiver Operating Characteristic (ROC) Curve

ROC curve serves in providing a graphical connection between sensitivity and specificity. The area below the ROC curve gives an idea on the benefit of using the test in question. Sensitivity is seen as the fraction of true positives to all with the disease in a clinical context. Specificity is the fraction of true negatives to all without disease. In our context with disease would be experiencing the event of interest like churn and without disease not experiencing the event of interest like no churn.

ROC curves are helpful in clinical experiments to determine the most valid cut-off for a test. The higher the true positive rate and lower false positive rate qualifies a test to be better than the other.
The curves were first used during world war II in detection of radio signals in the presence of noise following the Japanese attack of the Pearl Harbour. The motivation for the research was to find out how the US RADAR receiver operators missed the Japanese aircraft.

In constructing a ROC curve one must be familiar with the concepts of true positive, false positive, true negative and false negative. The ROC curve concept is useful when comparing the results of a test with a clinical truth which can be established by diagnostic procedures.

In constructing a contingency table a decision has to made on what the cut-off point is in order to categorise an observation as having or not experienced the event in question. The cut-off evidently separates the sensitivity(proportion of true positives to all observation who have experienced the event of interest) from the specificity(proportion of true negatives to all observations who have not experienced the event of interest) ROC curve therefore is a plot of the true positive rate(TPR) against the false positive rate(FPR) obtained by pre-determined cut-offs.

		Disease	
		+	-
Test	+	True Positive (TP)	False Positive (FP)
	-	False Negative (FN)	True Negative (TN)
		All with disease= TP + FN	All without disease= FP + TN

Table 3.0 : Comparing a method with the clinical truth

Sensitivity, recall, hit rate, or true positive rate (TPR)

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P}} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} = 1 - \mathrm{FNR}$$

Specificity, selectivity or true negative rate (TNR)

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{N}} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} = 1 - \mathrm{FPR}$$

Miss rate or false negative rate (FNR)

$$\mathrm{FNR} = \frac{\mathrm{FN}}{\mathrm{P}} = \frac{\mathrm{FN}}{\mathrm{FN}+\mathrm{TP}} = 1 - \mathrm{TPR}$$

Fall-out or false positive rate (FPR)

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{N}} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}} = 1 - \mathrm{TNR}$$

The level of sensitivity and specificity can be improved by adjusting the cut-off as deemed fit by the researcher since the number of individuals having experienced the event of interest remains constant. For every point on a ROC curve represent a chosen cut-off which is the true positive fraction and false positive fraction.
To construct a ROC curve from experimental data we begin by creating a rank for all values and linking each value to its true state with respect to the event of interest. We obtain TPR and FPR data points using a cumulative distribution function and plotting the TPR against FPR to obtain a ROC curve like the figure below.

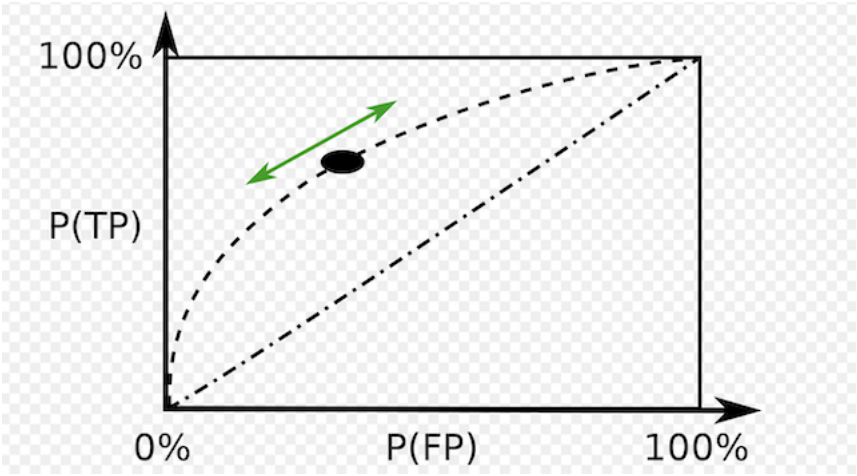


Figure 3.1 ROC Curve Diagram

For survival analysis constructing a ROC curve is slightly different from general linear models. Time dependent ROC curve will be implemented for the cox prediction model. Here sensitivity and specificity are defined at a specific time t and a threshold c as follows

$$\text{Sensitivity}(c,t) = P(M_i < c / T_i \leq t)$$

The cumulative sensitivity considers those who have experienced churn by time t as the denominator and a count of those having a marker value (M_i) higher than c as true positive (positive in churn)

$$\text{Specificity}(c,t) = P(M_i \leq c / T_i > t)$$

The dynamic specificity regards users who have not experienced churn at time t as the denominator (no churn users) and users who have a marker value (M_i) less than or equal to c as true negatives (negative in churn). By varying the threshold c from a lower value to a higher value gives the whole ROC curve at time t.

3.10 Model Building Process

The following steps are generally followed in model building

1. Exploratory Analysis
2. Variable Selection
3. Model Training
4. Model Validation and Testing

3.10.1 Exploratory Analysis(EDA)

Exploratory Analysis entails the initial investigations on the available data in order to discover patterns, anomalies, test various theories and check for assumptions with application of summary statistics and graphical displays.

A rule of the thump is to gain as much insight about the data as possible before getting dirty with it. The data set is scrutinised to determine the number of rows, variables as well as response and predictor variables.
EDA techniques serves to discover the hidden insight about the data like its distribution, discover outliers and devise mechanisms of handling them during modelling stage. EDA also entails testing of various theories around the data.

The study will rely heavily on Kaplan Meier graphs for data visualisation to determine how a group of users churn rates vary and use the curve to estimate median survival time for the groups.

"Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise." — John Turkey.
The above statement by John emphasises the need to ask the right question which is a central part of EDA. The primary goal is to use questions as tools to guide in the investigation. The question will help decide on which visual, transformation or model to try next.

The process is seen as iterative where by one question asked and explored generates a new question to be explored until the researcher feels confident on what to focus on. Simply put the questions have no format but one can decide to ask what variations occur within a variable and covariation among variables. The questions around this study would be

guided mainly by the objectives.

The philosophy of EDA is to postpone the usual assumptions on the data or what kind of model the data follows and to allow the data by itself to reveal its underlying structure and model.

3.10.2 Variable Selection

The process involves selection of best variables to form part of the predictor array(A list of explanatory variables)

The process of variable selection suggest the use of correlation coefficient ρ by examining the value of ρ between the predictors $A, B, C, ..., Z$ with the response variable. The value of $\rho > 0.6$ should then act as a qualifier to include a predictor into a specific model. The use of $\rho > 0.6$ has two major challenges when it comes to scaling it

$$\rho_{X,Y} = \frac{\mathrm{E}[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} \tag{8}$$

Where

- σ_X is the standard deviation of X
- σ_Y is the standard deviation of Y
- μ_X is the mean of X
- μ_Y is the mean of Y

Demerits of Correlation Coefficient

1. The increasing memory requirement for storing the correlation coefficients for a larger data set (space requirement equivalent to the square of the number of variables in the data set)
2. For non linear problems ρ is not a good indicator of correlation

A solution to these problems is to use Chi-square technique where each predictor is checked to see if the chi-square would detect a relationship. Continuous variables can be

categorised using binning technique and then passed through a chi-square test too. Chi-Square test can be represented using the formula below

$$X^2 = \sum_{i=1}^{k} \frac{(x_i - m_i)^2}{m_i} = \sum_{i=1}^{k} \frac{x_i^2}{m_i} - n \tag{9}$$

Where

- x_i observed values of x
- m_i is the expected value of x

The process of variable selection is not entirely left to correlation coefficient for linearly related variables and Chi-Square determination. Researchers at times includes variables into a model based on expert domain knowledge even if the correlation and Chi-Square test are negative meaning predictor variables are not significantly explaining response variables

3.10.3 Model Training or Estimation

After exploratory data analysis(EDA) and variable reduction the reduced set of variables are used to build the model. Here we use Survival analysis specifically Cox Proportional hazard to construct the model.

The reasons why modelling and simulation is vital are:

- We humans are constrained by linear thinking hence it is a challenge to understand the interactions of various parts of a system.
- We lack the ability to imagine all the viable possibilities of a real system
- We lack the ability to foresee the complete effects of various treatments or factors applied.

The primary software tool for these modelling will be R statistical and programming language version 3.6.0 (2019-04-26). The Survival package in R will be used for model creation with other dependency libraries like ggplot2 and survminer for visualisation and dplyr for data management and transformations. The survival package has the $surv()$ function which provides the core functionality for survival modelling.
We will be particularly concerned with time and status features in the data set where time will represent the number of days from account activation and user submitting articles to the final status based on 6 months observation period. For the survival status we shall consider churn, not churn or sensored as various categories of variable status.

Time and Status variables are used to produce the Kaplan-Meir curve which can be used to estimate the probability of survival at a particular time within the observation time range.
$Coxph()$ function within the survival package is used to fit a Cox Proportional Hazard model. Whereas the log-rank test compares two Kaplan-Meier survival curves, which might be derived from splitting a study population into treatment subgroups, Cox proportional hazards models are derived from the underlying baseline hazard functions of the study populations in question and an arbitrary number of dichotomised covariates. Again, it does not assume an underlying probability distribution but it assumes that the hazards of the patient groups you compare are constant over time. That is why it is called "proportional hazards model".
The following outline will be followed for model estimation

- Create the survival object which combines time and status of an event taking into account whether the event is censored or not
- Dichotomise continuous variables to categories something that will help in comparison either using Kaplan-Meir or log-rank test among various groups within a covariate.
- Fit Kaplan-Meier curve for the general data and for various covariates and use log-rank test for significance of the various categories of covariates. Log rank test for null hypothesis that the survival function for the covariates are constant, with p value less than 0.05, we reject null hypothesis and conclude that survival functions are significantly different from each other.
- Perform a Cox proportional hazard model including all the desired covariates and find the hazard ratio associated with the various covariates. Hazard ratio represent the relative risk of death and in this case churn for a user. HR greater than 1 indicates increased risk of death or churn in this case while on the other hand HR less than 1 indicates reduced risk of churn for the users under observation based on a particular covariate.

3.10.4 Model Verification & Validation

Every model that has to support decision making or go into production must undergo verification and validation as an essential part of building a model. Verification goal is to guarantee that the model is built or programmed the right way with the instructions coded appropriately devoid of bugs. Verification in addition to checking on accurate implementation goes further to check if the implementation is in sync with the conceptual model taking into account the underlying theories and assumption for the model.

Model verification helps to ensure that specification is right and implementation is correct but it doesn't guarantee that the model will provide a solution to the problem being modelled, meet model requirements or reflects the working of some real life system.

It is important to state that no computational model will be fully verified to boast of 100% error free implementation. The trade of is on statistical certainty by continuously testing and correcting for bugs until the model meets the acceptable levels.

Validation serves to ensure the model built serves the intended purpose in terms of the inputs added and the output obtained. The main purpose for validation is to ensure the model answers the right question by accurately modelling the system in question.

Validation mostly work in attempting to invalidate a model by continuously improving on components that fall short of the threshold set for a valid model. The outcome is not a completely valid model but one that has passed all the threshold set for invalidation.

The widely used standard for model validation devised around 1967 by Naylor and Finger uses the steps:

- Build a model that has high face validity : appears to be a reasonable imitation of real world system by people who are experts in the phenomenon being modelled
- Validate model assumptions : Looking at the structural and data assumptions where structural tries to understand the physical working or operations of the actual system being modelled. Data assumptions basically checks the accuracy of the data as well as the assumed distribution.
- Compare the model input output transformations to the real system transformations: The outputs generated by the model are observed with the real system output to find out the percentage of success explained by our model.

4 Data Analysis and Results

4.1 Introduction

This chapter presents research findings from the data analysis. The analysis provide information on the research objectives which include:

- To develop a churn model using Cox regression
- To estimate the retention probability of a user
- To estimate the relative risk of churn using significant covariates

4.2 Survival analysis and test of hypothesis

4.2.1 Survival curves

Churn is loosely defined as a writer who fails to publish an article for a period of one month. The table below shows the number of users who churned as from inception of risk. The inception of risk covers every writer that submitted a publication for January 2019. The risk set comprised 104 writers with the following monthly churn distribution.

The study found a very high attrition at the beginning of the year with January having a 50 % attrition rate. Attrition for the months of February to July decreased sequentially with the month of June and July having the same retention probability since non of the

Month	Risk set	Churn	Retention Prob
January	104	52	0.5
February	52	21	0.3
March	31	11	0.19
April	20	5	0.14
May	15	2	0.13
june	13	4	0.09
July	9	0	0.09

Table 1. Monthly churn distribution

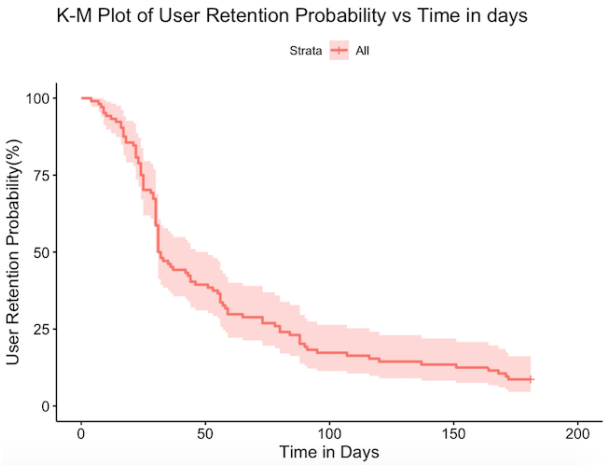


Figure 2. Kaplan Mier graph on user retention probabilities over time

users churned on the month of July.

The Kaplan Meir graph above shows daily retention probabilities for the writers within 180 plus days. There is a sharp drop from 100 % to 50 % retention which is basically within the first 30 days of writer attrition observation. After 30 days the retention probability dropped from 50 % to around 10 % over 130 days. These behaviour will be explained by looking at the length of time spent on the platform prior to inclusion into the study.

The faded shade on both sides of the curve represents the confidence interval for the retention probability estimate based on the standard error. The study ended with 9 users right censored meaning they never experienced churn for the period of follow up. This is shown in the K-M graph towards the end of the curve by the plus sign while every drop on the curve represents a user churn event.

The median survival(retention) time was obtained at an intersection of the survival(retention) probability of 0.5 on the y-axis and time on the x-axis to be 31 days. It took 31 days for half of the website users to churn or stop publishing articles with the news site. The writers' never visited their website profiles for a continuous period of one month based on our initial assumption of attrition for the online writers.

(i) Time spent on the platform for churners

The time spent on the site for a writer was calculated as the difference between the writer accounts creation date and the last date of publication having met the predefined churn assumptions for the study.

Time on platform	n	events	median
Less than 250 days	70	67	32.5
More than 250 days	34	28	31

Table 2. Writer's length of time on the platform & churn distribution

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

                                      N Observed Expected (O-E)^2/E (O-E)^2/V
time_spent_category=Less than 250 days 70       67     63.1     0.244     0.767
time_spent_category=More than 250 days 34       28     31.9     0.482     0.767

 Chisq= 0.8  on 1 degrees of freedom, p= 0.4
```

Table 3. Log rank test statistic on writer time length on the platform.

The average time spent on the platform for the churned writers was 250 days with a standard deviation of 219 days. The number of days spent on the platform was between 31 to 469 days based on the standard deviation of these estimate. Time spent on the platform missing values were replaced by the mean and then categorised as below or above 250 days for purposes of comparison and hypothesis testing.

The table 4.2 shows that writers with less than 250 days on the platform have a median survival(retention) of 32 days compared to more than 250 days at median survival(retention) of 31 days. Both categories seems to be closer to the population median survival time of 30days. The findings are showing an insignificant effect of time spent on the platform as determinant of user retention. There was need to perform a hypothesis test to determine statistical significance of these findings.

Hypothesis Testing using Log Rank Test statistic for time spent on the platform and hazard for the two groups

$$H_0 : h(\text{More than 250 days}) = h(\text{Less than 250 days})$$

$$H_1 : h(\text{More than 250 days}) \neq h(\text{Less than 250 days})$$

At α level of 0.05 The Log rank test statistic gives a p-value of 0.4 which is greater than 0.05. We therefore fail to reject H_0 at 95% CL that the risk of churn is equal for both users of varying length of stay on the platform. Therefore writer total time spent on the platform is not statistically significant determinant of writer attrition(hazard).

This is a mirror of the almost similar median survival times for the two groups based on the time spent on the platform. These results could be so because of the nature of these variable having nothing to do with engagement since it is measured as the difference between subscription date and last publication date as one could have a longer time span

Gender	n	events	median
Female	28	27	30
Male	76	68	44

Table 4. Writer's gender & churn distribution

```
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
```

Table 5. Log rank test statistic on churn risk(hazard) based on gender

but not a constant writer with high reader count articles leading to a descent pay and more satisfaction.

(ii) Gender of writers on platform

The table indicates that majority of the writers were male at 73% while the female stood at 27%. The number of users who did not experience churn by the end of the study were 9 with a distribution of 1 female and 8 males. The median survival time for males is 44 days while that of females is 30 days mirroring the population median survival(retention) time. Males tend to stay longer before churning as compared to female writers.

Hypothesis Testing using Log Rank Test statistic for hazard based on Gender

$$H_0 : h(\text{Gender:Female}) = h(\text{Gender:Male})$$
$$H_1 : h(\text{Gender:Female}) \neq h(\text{Gender:Male})$$

At α level of 0.05 The Log rank test statistic gives a p-value of 0.05 which is equal to α level. We therefore reject H_0 at 95% CL that the risk of churn is equal for both male and female users. Therefore writer's gender is statistically a significant determinant of churn risk, the level of the risk will be explained using a Cox model.

(iii) Number of articles published

The number of articles published by a writer directly depicts an engaged user with the platform and significantly contributes to the number of page views that an article will gather. The more a writer publishes the higher the number of page views which later

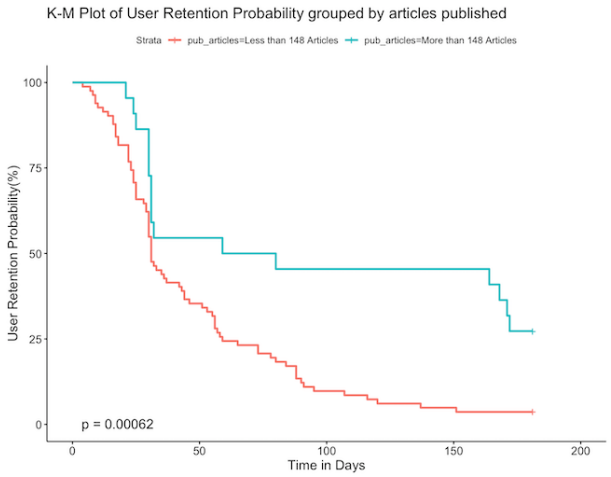


Figure 3. K-M graph on user retention grouped by number of published articles

translates to wages earned. Amount earned by a writer is a pure function of page views generated by their articles. The study found that on average a writer publishes 148 articles over the 6 months period. The number was used to segment users in two categories that is

- Writers who published more than 148 articles over the duration
- Writers who published less than 148 articles over the period.

It is evident from the figure 4.2 that a writer who published more than 148 articles enjoyed a higher survival probability or retention as compared to writers who had published less than 148 articles. At the end of the observation period 27.3% of writers with more than 148 articles did not experience churn while the group publishing less than 148 articles only 3.6% did not churn.
The median survival time for writers who published more articles almost doubled from the population median survival time of 30 days against the median survival time for writer segment with more than 148 published articles at 59 days.

Hypothesis Testing using Log Rank Test statistic for churn risk based on number of articles published

$$H_0 : h(\text{More than 148 articles}) = h(\text{Less than 148 articles})$$
$$H_1 : h(\text{More than 148 articles}) \neq h(\text{Less than 148 articles})$$

At α level of 0.05 the Log rank test statistic gives a p-value of 0.00062 which is less than the α level. We therefore reject H_0 at 95% CL that the risk of churn is equal for both users

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

                                  N Observed Expected (O-E)^2/E (O-E)^2/V
pub_articles=Less than 148 Articles 82       79     64.3      3.34      11.7
pub_articles=More than 148 Articles 22       16     30.7      7.01      11.7

 Chisq= 11.7  on 1 degrees of freedom, p= 0.0006
```

Table 6. Log rank test statistic on number of article published

who published more than 148 articles as well as less than 148 articles.
We conclude that the two groups statistically have different retention or varied risk of churn since the p-value obtained is less than α.

(iv) Number of articles rejected

The site editors police what content goes live therefore an article is subjected to plagiarism as well as grammah check. Any article that fall short the set thresh hold is rejected and never get to be viewed live on the site. The study found an average of 10 articles are rejected from a writer within the follow up period. The writers were then grouped into more than 10 rejected articles and less than 10 rejected articles.

The study found that a writer with more than 10 rejected articles is likely to stay longer on the platform than a writer with fewer rejections. This basically confirms the Fig 4.3 above where more retention is explained by publishing more articles. The median survival time for the segment with more rejected articles is much higher that is 73 days against 25days which is even less than the population survival time of 30days. The follow up period ends with 20% retention for writers with more rejected articles as compared to 5% retention for the less than 10 rejected articles.

Hypothesis Testing using Log Rank Test statistic for articles rejected

$$H_0 : h(\text{More than 10 articles}) = h(\text{Less than 10 articles})$$

$$H_1 : h(\text{More than 10 articles}) \neq h(\text{Less than 10 articles})$$

At α level of 0.05 The Log rank test statistic gives a p-value of 0.012 which is less than the α level
We therefore reject H_0 at 95% CL that the risk of churn is equal for both users who had

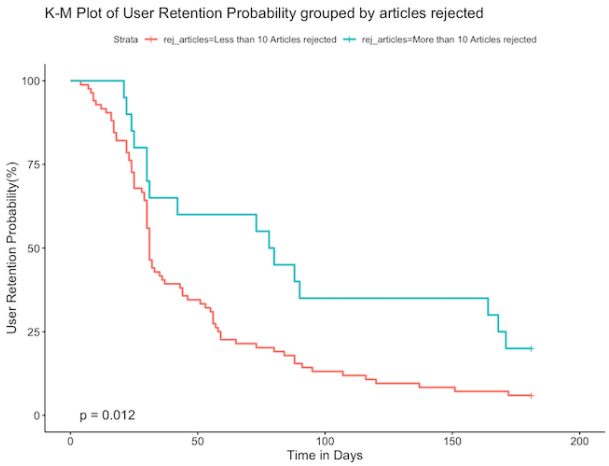


Figure 4. K-M graph on user retention grouped by number of rejected articles

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

                                        N Observed Expected (O-E)^2/E (O-E)^2/V
rej_articles=Less than 10 Articles rejected 84       79     68.4      1.64      6.29
rej_articles=More than 10 Articles rejected 20       16     26.6      4.21      6.29

 Chisq= 6.3  on 1 degrees of freedom, p= 0.01
```

Table 7. Log rank test statistic on number of article rejected

more than 10 articles rejected as well as less than 10 articles rejected.
Therefore the number of articles rejected is statistically significant in explaining retention or risk of churn.

Users having few rejected articles generally shows that they either submit fewer articles hence fewer number of rejections as compared to submitting more articles with majority passing and a good number being rejected. It could also be a signal of specialization that makes them write for a specific category of articles having less errors but not much readership and revenue from the same.

(v) Category of the article published by a writer

The articles pubished had various categories with major categories being politics, entertainment, sports, business and lifestyle. The study grouped article category into two groups for comparison that is Political stories while the rest were grouped as Lifestyle and others category. The risk of attrition was compared for each category by plotting cumulative hazard against time and represented using the figure below.

In general lifestyle and other category of article writers had a higher risk of churn as compared to the writers of political stories. To confirm the statistical significance of these results a log-rank test was performed.

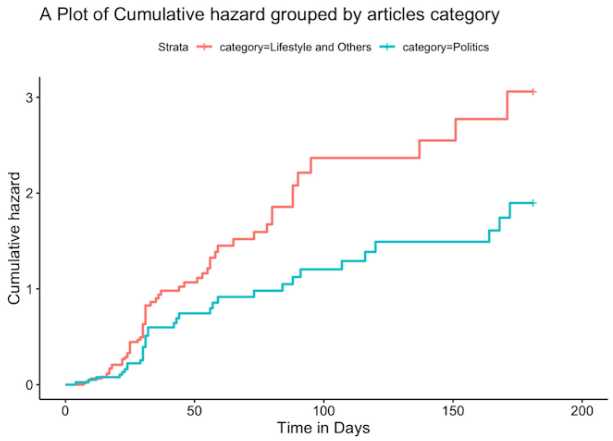


Figure 5. Cumulative hazard grouped by article category.

```
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 8. Log rank test statistic on articles categories

Hypothesis Testing using Log Rank Test statistic on churn risk based on writer's article category

$$H_0 : h(\text{Category:Politics}) = h(\text{Category:Lifestyle \& others})$$
$$H_1 : h(\text{Category:Politics}) \neq h(\text{Category:Lifestyle \& others})$$

At α level of 0.05 The Log rank test statistic gives a p-value of 0.02 which is less than the α level. We therefore reject H_0 at 95% CL that the risk of churn is equal for both users who published aricles with a political inclination or lifestyle and others. Therefore category of articles published by a writer is statistically significant in explaining retention or risk of churn(hazard) .

(vi) Location of a writer

The location category of the article basically segments the writer as National or county with stories majorly biased towards the location segment. The national based writers stood at x% while the county writers were y%

Hypothesis Testing using Log Rank Test statistic for churn risk based writer's location

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

                           N Observed Expected (O-E)^2/E (O-E)^2/V
location_category=County   54       48     51.3     0.214     0.485
location_category=National 50       47     43.7     0.251     0.485

 Chisq= 0.5  on 1 degrees of freedom, p= 0.5
```

Table 9. Log rank test statistic on writer grouped by location.

$$H_0 : h(\text{location:National}) = h(\text{location:county})$$

$$H_1 : h(\text{location:National}) \neq h(\text{location:county})$$

At α level of 0.05 The Log rank test statistic gives a p-value of 0.49 which is greater than the α level
We therefore fail to reject H_0 at 95% CL that the risk of churn is equal for both users who published aricles relating to national as well as county. Therefore the location of a writer publishing is not statistically significant in explaining retention or risk of churn(hazard) within the writer population.

(vii) Education Level of a writer

The study analysed the relationship between a website user's education level and attrition. The table below shows writer highest education attained and attrition within the education segment.

The table above shows that writers with a university education has a median survival(retention) of 41 days compared to the other education levels that average at 31 days close to the population median of 30 days. These result means that it will take a roughly a month to loose non university writer while a university writer will have survived attrition by more than one month. There was need to perform a hypothesis test to determine statistical significance of these findings.

Education level	n	events	median
University	28	27	43
College	29	26	31
College Student	25	22	31
High school	22	20	31

Table 10. Witer education level & churn distribution.

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

                              N Observed Expected (O-E)^2/E (O-E)^2/V
level_of_educ=Non-university 76       68     65.8    0.0731     0.249
level_of_educ=University     28       27     29.2    0.1648     0.249

 Chisq= 0.2  on 1 degrees of freedom, p= 0.6
```

Table 11. Log rank test statistic on writer education level.

Hypothesis Testing using Log Rank Test statistic for writer education level

$$H_0 : h(\text{Education:University}) = h(\text{Education:Non-University})$$

$$H_1 : h(\text{Education:University}) \neq h(\text{Education:Non-University})$$

At α level of 0.05 The Log rank test statistic gives a p-value of 0.6 which is greater than α level

We therefore fail to reject H_0 at 95% CL that the risk of churn is equal for both users of varying education level. Therefore writer level of education is not statistically significant determinant of writer attrition(hazard) despite having seen a higher retention for writers' with university education as well as a higher median survival(retention) time of 43 days as compared to 31 days for writers with no university education. A writer with a university education can probably stay on the site as a side hussle writing job or leave when a better opportunity arrives. These effect would be further tested using Cox regression.

```
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 12. Cox regression and coefficients of various covariates.

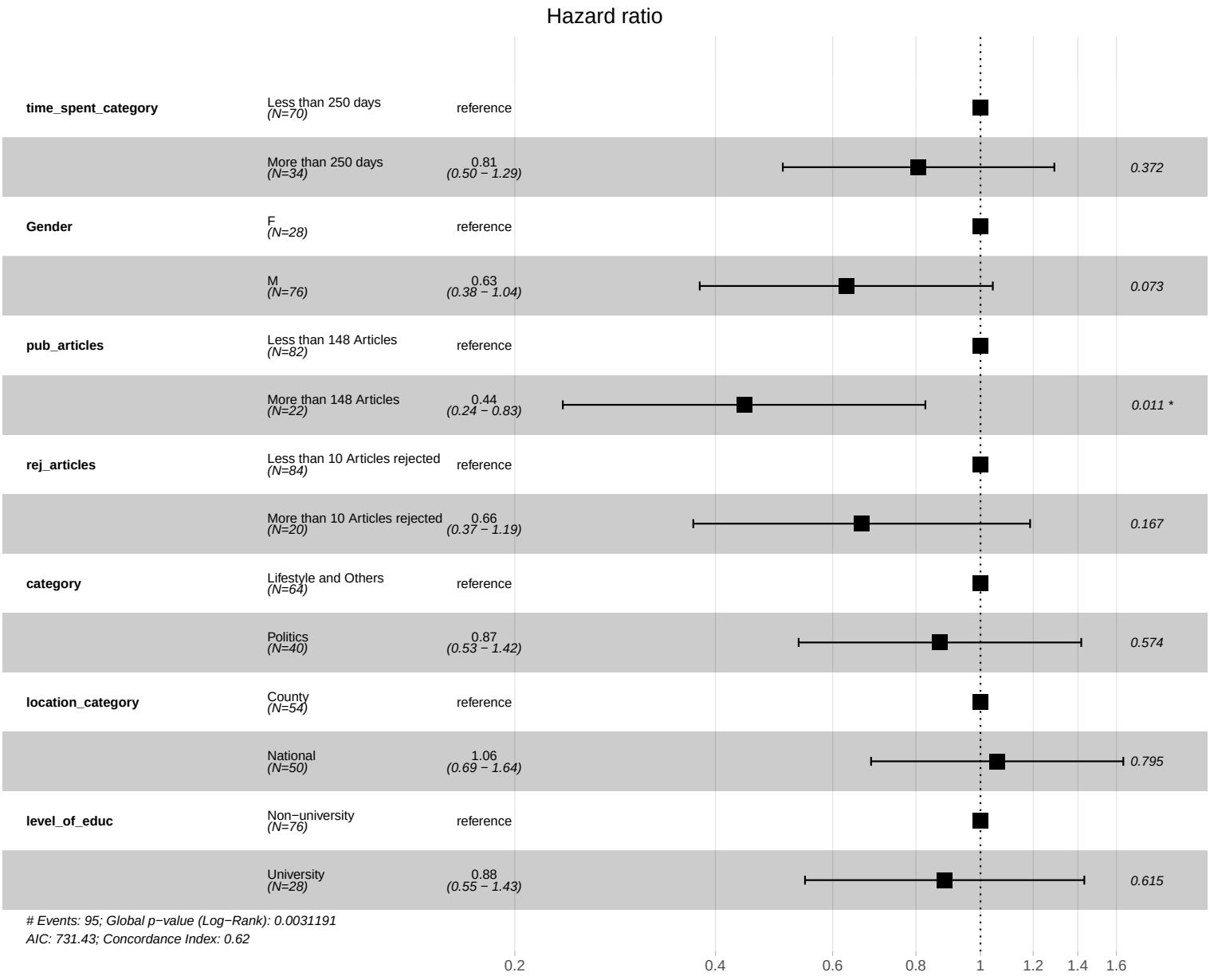
4.2.2 Cox Regression

There is a need to look at the various covariates total effect unlike log-rank and kaplan mier analysis that treat various covariates in isolation. Cox proportional hazards model provide a systematic approach to study the covariates effects in explaining the hazard risk or the risk of churn for these case.

The table 4.12 shows coefficients of the independent variables in explaining the likelihood of churn. They coeffients are hazard ratios for various groups of the covariates in comparison to the reference group.

The hazard ratio less than 1 shows the treatment group is less likely to experience the event of interest while hazard ratio greater than 1 indicate a high risk for the treatment group. The users who write articles for national have a hazard ratio of 1.05 meaning they are likely to churn as compared to the county based writers.

Gender has hazard ratio of 0.62 therefore male writers are less likely to churn than their female counterparts. Articles published has a hazard ratio of 0.44 which is the lowest. Users publishing more than 148 articles are less likely to churn and the result is significant with a p-value of 0.01. Rejected articles has a hazard ratio of 0.66 therefore users with more than 10 articles rejected are less likely to churn. Category of the article has a hazard ratio of 0.86 therefore users writing politically inclined articles are less likely to churn than those writing for lifestyle and other categories. The level of education has a hazard ratio of 0.88 meaning users with university level of education are less likely to churn in comparison to other levels of education. These results are further illustrated below using the forest plot visual.


Hazard ratio
time_spent_category
Less than 250 days (N=70) reference
More than 250 days (N=34) 0.81 (0.50 − 1.29) 0.372
Gender
F (N=28) reference
M (N=76) 0.63 (0.38 − 1.04) 0.073
pub_articles
Less than 148 Articles (N=82) reference
More than 148 Articles (N=22) 0.44 (0.24 − 0.83) 0.011 *
rej_articles
Less than 10 Articles rejected (N=84) reference
More than 10 Articles rejected (N=20) 0.66 (0.37 − 1.19) 0.167
category
Lifestyle and Others (N=64) reference
Politics (N=40) 0.87 (0.53 − 1.42) 0.574
location_category
County (N=54) reference
National (N=50) 1.06 (0.69 − 1.64) 0.795
level_of_educ
Non−university (N=76) reference
University (N=28) 0.88 (0.55 − 1.43) 0.615
Events: 95; Global p−value (Log−Rank): 0.0031191
AIC: 731.43; Concordance Index: 0.62
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6

The above forest plot shows various covariates and hazard ratio as either below 1 or above 1. It also shows the p-value and whether the coefficient is significant or not. It shows the category of the covariate receiving treatment as shaded and the reference category as well with a white background. Covariate categories receiving treatment that have hazard ratio less than 1 are less likely to experience churn as compared to the reference category and the opposite holds for covariate category with the hazard ratio greater than 1.

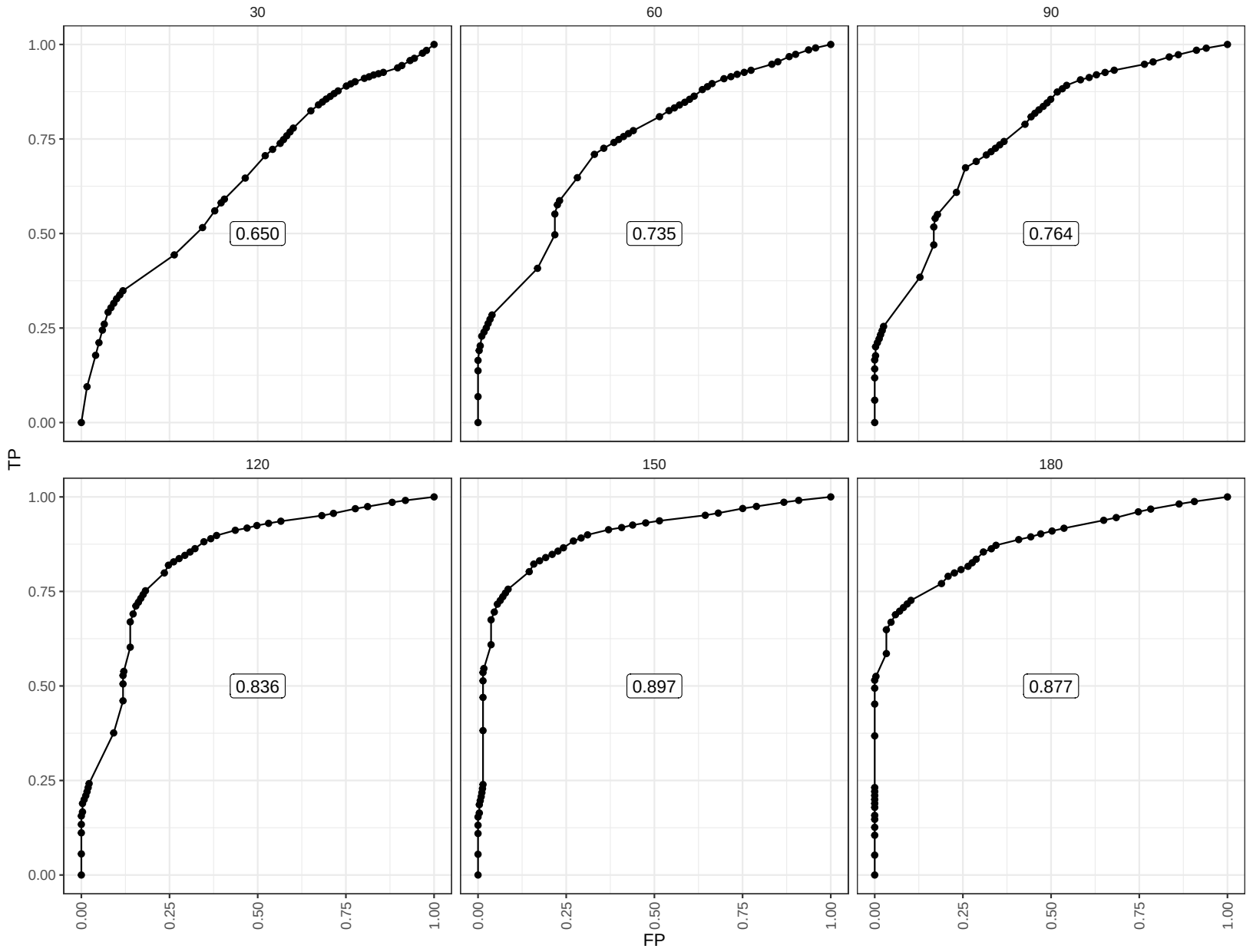
The study findings shows that half the writer population sampled churned after 31 days of observation. Major predictor covariate of writer retention was number of articles published. The more a writer publishes articles the less likelihood of churn. Generalizing these findings to the general online work force means that more engagement with such a platform is a positive contribution to retention while less engagement drives user attrition.

There are a number of covariates that were tested using both log rank test and Cox regression model. Gender of the writer showed a significant difference in terms of attrition using log rank test while for Cox regression it had a hazard ratio less than 1 for males in reference to females meaning less attrition for males but these results were not statistically significant. The number of rejected articles for a writer showed a significant difference in terms of churn risk using log rank, same tested with cox regression showed that the risk of churn is higher for writers with less than 10 rejected articles for the study length. Category of an article as either politics or lifestyle and others too showed a significant difference in terms of churn risk using the log rank test, cox regression coefficient showed that writers who specialize in political stories had less likelihood of churn as compared to the lifestyle category writer. Both rejected articles and category of aricles for a writer were not statistically significant using cox regression.

The covariates that were not statistically significant for both log rank test and cox regression were level of education, location of a writer and time spent on the platform. Although the test were not statistically significant, using Cox regression it was possible to measure the relative risk of churn for the covariates using the hazard ratios for the treatment and reference categories.

4.2.3 ROC curves

Roc curve was used to measure how best the model fit the data used for the analysis. The plot was divided into 6 categories of 30 days of observation. The area under curve was calculated cumulatively starting from 30 days, 60 days, 120 days all the way to 180 days. The results shows that the ability of the Cox model to predict churn improved with time from 0.65 to 0.877. The curves are shown below.


30
60
90
120
150
180
0.650
0.735
0.764
0.836
0.897
0.877
TP
FP
0.00
0.25
0.50
0.75
1.00

5 Summary Conclusions and Recommendations

5.1 Introduction

This chapter provides a summary of the study findings and recommendations within the scope of attrition modelling using online web application data.

5.2 Summary

In this work survival analysis was used to analyze propensity to churn for online writers of a news website in Kenya known as hivisasa.com. The study sought answers on which covariates were major determinants of writer attrition from the online platform, their statistical significance and magnitude. A total of one hundred and four writers who had at-least one publication for January 2019 formed part of the study sample. The sample historical data for January to July was used to determine writers who churned within the period and those that were retained. Previous litrature on attrition research was reviewed and the study settled on survival methods in order to address time to event and manage censored data.

Descriptive analysis was handled by fitting Kaplan-Meier curves to visualize the retention curves of various categories of the covariates. Log-Rank test was then used to test the statistical significance of the various differences observed. Cox proportional hazard was fitted on the data including all covariates to determine the magnitude of hazard risk. Three of the covariates that is gender, number of articles published by a writer and category of articles done by the writer were significant in explaining writer attrition risk and magnitude. The results showed a high churn rate among female writers, writers publishing non political content on the site as well as publishing less than 148 articles for the study period. On the other hand three covariates; time spent on the platform from subscription, location of a writer and level of education were not statistically significant in explaining writer attrition. Even though these covariates lacked statistical significance Cox regression coefficients revealed that the magnitude of risk varied accross them. Level of education graduate and time spent on the platform of more than 250 days reduced the chances of a writer churning 12% and 19% respectively in comparison to the reference variable holding for the effect of other covariates. The model performance was validated by fitting a ROC curve to ascertain how best the model was able to fit the data. The ROC curve had an AUC of 87% which means the model had a 87% chance of predicting a churned writer as so.

5.3 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 the number of articles published by a writer.

5.4 Recommendation

The median survival time of 30 days indicates that more can be done to increase the time to loose half of the subscriber base to at-least 90 days by serious writer engagement within the first month of recruitment Secondly the study proposes specifically more research around churn in web applications and the software as a service industry(SaaS) in general.

Bibliography

[1] Alain Saas, Anna Guitart and Colin Magne: *Churn Prediction in Mobile Social Games*, Siliconstudio company Japan, 2017, pages 1-10.

[2] Alfred Demaris. *Regression with Social Data:Modeling continuos and limited response variable*, John Wiley and Sons Inc Publication, 2001, pages 390-418.

[3] Ali Tamaddoni Jahromi. *Predicting Customer Churn in telecommunications Service Providers*, Lulea University of Technology, 2009, pages 45-55

[4] Dang Van Quynh. *Customer Churn Prediction in Computer Security Sofware*, Researchgate, 2019, pages 15-60

[5] Dirk Van den Poel and Bart Larivie. *Customer Attrition Analysis for Financial Services Using Proportional Hazard Models*, Ghent University, 2003, pages 32-55

[6] Doni Suhartono Asep Saefuddin and Made Sumertajaya. *Survival Analysis Of Customer In Postpaid Telecommunication Industry*, Department of Statistics, Bogor Agricultural University, Indonesia, 2013, pages 1-10

[7] Godsway Roland . *Churn Model in the Mobile Telecommunications Industry in Ghana*, University of Ghana repository, 2012, pages 45-62

[8] James Kairanga. *Churn Prediction Modelling in mobile Telecommunication Industry*, University of Nairobi Repository, 2012, pages 25-60

[9] John Hadden. *A Customer Profiling Methodology for Churn Prediction*, World Academy of Science, Engineering and Technology 16 2008, 2009, pages 30-55

[10] Kojo Abiw. *Predicting customer Churn in the mobile telecommunication industry A case of Ghana MTN*, Knust Space, 2011, pages 20-45

[11] M M Sepehri B Teimourpour and S Choob- dar. *Modeling customer churn in a non-contractual setting: the case of telecommunications service providers*, Journal of Strategic Marketing, 18(7), 2010, pages 587-598

[12] Seo D Ranganathan C and Badad Y. *Two-Level model of customer retention in US mobile Telecommunication Service Market*, Telecommunications Policy, 32, 2008, pages 182-196

[13] Sha Yuan, Shuotian Bai, Mengmeng Song and Zhenyu Zhou. *Customer Churn Prediction in the Online New Media Platform: a Case Study on Juzi Entertainment*, ResearchGate, 2013, pages 10-20

[14] Simon McPhillips. *Media convergence and the evolving media business model: an overview and strategic opportunities*, ResearchGates, 2008, pages 230-250

[15] Sohn, S. and Kim, Y.. *Searching customer patterns of mobile service using clustering and quantitative association rule*, Expert Systems with Applications, 34(2), 2008, pages 1070-1077

[16] Wei C and Chiu I (2002). *Turning Telecommunications call details to churn prediction*, Expert Systems with Applications, 23, 2002, pages 103-112