Latent Dirichlet Allocation for Predicting User Churn Reasons Based on APP Negative Comment Text

Xuan Ding

Renmin University of China, Beijing 100872, China

Abstract: Based on the negative comment text of mobile applications, this research proposes a framework integrating Latent Dirichlet Allocation (LDA) with duration modeling to predict not only when users churn but also the underlying reasons for their departure. The study aims to address a gap in existing literature, which often focuses solely on churn probability or timing, by incorporating thematic analysis of user-generated content from app stores. Methodologically, it employs the Duration Model to estimate the time until churn and extends it into a Competitive Risk Model that accounts for multiple churn reasons (categorized as controllable, uncontrollable, and unknown risks). The LDA algorithm is utilized to extract latent topics from negative reviews, transforming unstructured text into interpretable variables for predictive modeling. These variables, alongside duration data, are integrated into the model to enhance prediction accuracy. Evaluation will involve metrics such as perplexity for LDA performance and log-likelihood, AIC, and BIC for model comparison. The research expects to contribute a novel, text-enhanced approach to churn prediction, offering practical insights for internet companies to better understand churn drivers, improve user experience, and design targeted retention strategies, despite potential limitations in LDA's performance on short text.

***Keywords*:** Latent Dirichlet Allocation (LDA), user churn, negative comment

1. Introduction

User growth is always at the core of the business to customers (B2C) of Internet company. User growth can be decomposed into acquisition of new users plus retention of old users. According to the AARRR model, user retention is a crucial intermediate link in the user lifecycle. According to a survey by Bain Company, a 5% increase in customer retention rate in the business world means a 30% increase in company profits, and the probability of selling products to old customers is three times higher than selling to new customers. So, there is a famous saying in Growth Hacker: retaining existing users is better than expanding new customers, which is commonly known as 'one bird in the hand is better than two birds in the forest.'

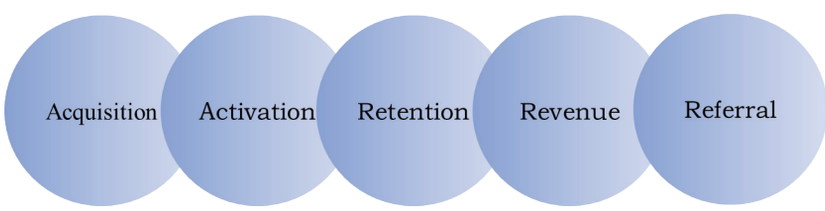

Figure 1. AARRR model of user lifecycle

However, users of the app may churn due to various reasons. The manifestations of user loss in mobile apps include uninstalling the app, not opening the app for a long time, and closing app push notifications (including notification messages, advertising promotions, various types of pop-ups, etc.). Therefore, app suppliers need to use various operational methods and strategies to recall users and reactivate them.

Currently, in Internet companies, there are several methods for user recall:

- EDM email push, also known as Electronic Direct Mail, refers to a digital marketing method that sends business or promotional information to customers through email. The EDM method has advantages such as low cost, more content that can be conveyed by emails, and timely reminders for new emails. However, it also faces the problem of low usage among domestic users.
- APP message push. APP message push is one of the important operational methods for APP operators nowadays, which has a strong effect on user recall and user activity. App message push is a double-edged sword that, when used well, it can help app operators efficiently achieve various operational goals. Conversely, it can be annoying to users

and lead to an increase in app uninstallation rates.

- SMS push. This method has a higher user-reach rate, but its cost is higher than the EDM email push. It can also be blocked by smartphone anti-harassment software.
- WeChat private domain push. Many apps closely connect users with WeChat official account. Many app suppliers use WeChat traffic to develop more product lines in WeChat official account for user precipitation, and then lead users to the APP, which is conducive to cultivating users' brand stickiness.

2. Related work

Recent studies mainly focus on churn probabilities and the prediction of the customer lifetime value[1], effects of information and communication technology service and its complementary strategies on customer loyalty[2], and the formation of financial reports and communication with management regarding churn[3]. Overall, we find that the majority of the research on churn is focused on which customers are going to leave. In addition, researchers also investigate the moment when a customer terminate the use of app and on the duration of use of app.

Another stream of research is focused on why a customer terminates the use of app. The risk of user churn can be divided into three categories: (1) Value risk. Including anything related to the company's emotions, such as reactions to poor service, price of products or services, and quality of products or services. These risks are controlled by the company and are therefore defined as controllable risks[4]. (2) Personal risk. Including all risks within the company's control, such as migration or death. (3) Unpaid risk. Including all risks associated with unpaid fees, such as misuse or theft of products or services. In the event of loss of all unpaid users, the company will decide to terminate the user's contract. Among them, personal risk and non-payment risk combined are considered uncontrollable risks.

Table 1. Description of competing risks

Risk	Description
Controllable	All risks which are in control, such as churn due to high prices, bad services, or a better offer of a competitor.
Uncontrollable	All risks which are out of control, such as migration or death.
Unknown	The risk that contains all customers who did not provide a reason for which they churned.

3. Methodology

To predict the moment *when* a user churns, we choose to use the Duration Model which provides the time spent in a specific state before transitioning to another state. The key concept of the Duration Model is risk probability or risk function, which refers to the probability of users losing after one certain day without loss. The Duration Model is used to simulate the length of time spent transitioning from a given state to another state[5]. Use t to represent a continuous random variable, also known as a duration variable, where t represents the duration of the user's continuous use of the app (in days). This model defines the hazard function $\lambda(t)$ as the instantaneous probability of user churn occurring at time t, assuming that the user has not yet lost at time t, and the formula is as follows:

$$\lambda(t)=\lim_{\delta\to 0}\frac{P\left(t\le T\left\langle t+\delta\right|T\ge t\right)}{\delta}$$

In order to predict the reasons for customer fluctuations, our goal is to establish a predictive model that considers multiple exit states. This feature can be found in the Competitive Risk Model[6], which can be seen as an extension of the Duration Model. The Competitive Risk Model takes multiple exit states (or risks) into account, where different risks race to decide which risk will trigger the eventual churning act.

The risk set $r=\{0, 1, 2, 3\}$, where $j=0$ represents the initial state, and $j=1$, 2, and 3 correspond to the three types of risks mentioned above; T_j represents a continuous random variable with risk $j\in r$, representing the time (in days) during the trial period when users experience loss due to risk j. Therefore, the formula for the non-para metric hazard function $\lambda_j(t)$ is as follows:

$$\lambda_j(t) = \lim_{\Delta t \to 0} \frac{P(t \le T_j \le t + \Delta t \mid T_j \ge t)}{\Delta t}$$

where λ_j is a non-parametric hazard function for risk j with values of 1, 2, and 3.

To capture the content of comments from customers in the app stores within a variable, we suggest detecting the underlying theme of each comment. To this end, we propose the Topic Model, which is a probability model used to reveal the underlying semantic structure of a text set, so that each comment text can be assigned to a potential topic[7].

Our study chose to use the LDA algorithm to process the content of user comments on apps in the app store, and captured the themes of these comment text data into the variables of the prediction model. Two types of variables are expected to be extracted into the prediction model:

- The duration variable

The duration variable represents the time difference between when the user first downloaded the app and when they uninstalled the app.

- The status variable

The state variable depends on the purpose of the model, which can be predicting churn or predicting the reasons for user loss. For the previous purpose, that is, for the Duration Model, the value logic of the state variable is: current user=0, lost user=1. For the latter purpose, that is, for the Competitive Risk Model, the current value of the state variable is 0 until the user's state changes. The value of the state variable depends on the reason for user churn, where each reason is represented as a different integer.

The algorithm model framework diagram of our study is as follows:

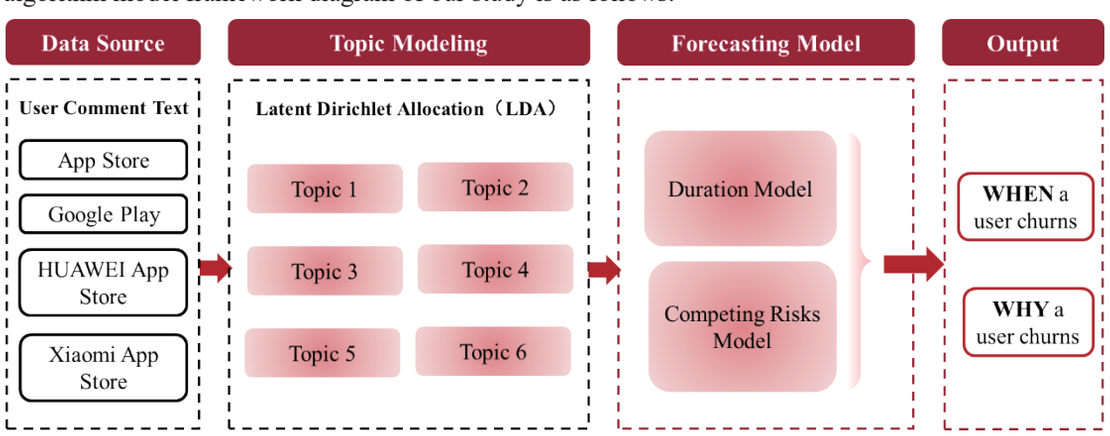

Figure 2. Model framework

4. Evaluation

4.1 Performance evaluation of LDA algorithm

For the evaluation of the performance of the LDA algorithm, some scholar proposed using the complexity of a reserved test set to evaluate the LDA model[8]. Complexity is a measure of the quality of a probability model's predicted samples. A convenient feature of complexity is that as the likelihood of the model increases, complexity decreases. Therefore, the lower the complexity, the higher the performance of the LDA model.

4.2 Comparative evaluation of user churn models

To compare the performance of different user churn models, consider using the logarithmic likelihood (LL) value of the model, the Akashi Information Criterion (AIC), and the Bayesian Information Criterion (BIC). The performance metrics within these samples indicate the goodness of fit of the model to the training data.

$$LL = \ln(L)$$

$$AIC = 2k - 2\ln(L)$$

$$BIC = ln(N)k - 2ln(L)$$

In the formula above, L is the maximum value of the likelihood function of the loss model, k is the number of explanatory variables in the loss model, and N is the number of observations. The evaluation criteria are: the higher the LL value, the higher the model performance; The lower the AIC or BIC , the higher the model performance. Based on this, the model with the highest LL value and the lowest AIC and BIC values was ultimately selected.

5. Innovations of the Study

This study makes significant contributions through its key innovations, which can be categorized into theoretical and practical domains.

5.1 Theoretical Innovation

The primary theoretical innovation of this research lies in the novel methodological integration of unstructured text analytics with advanced duration models for churn prediction. While existing literature predominantly focuses on predicting ***if*** or ***when*** a customer will churn using structured behavioral data, this study breaks new ground by explicitly modeling the reasons for churn through the lens of user-generated content. We achieve this by bridging two distinct methodological streams: Latent Dirichlet Allocation (LDA) from natural language processing and the Competitive Risk Model from survival analysis. This synergy allows for the transformation of qualitative, subjective negative app reviews into quantifiable, interpretable topic variables that serve as direct inputs for competing risks. By doing so, we address a critical gap in the literature, which has largely treated churn reasons as latent or unobserved, and provide a robust framework for inferring the specific drivers — such as dissatisfaction with service, product quality, or technical issues — that precipitate customer departure.

5.2 Applied Innovation

The applied innovation of this work is its provision of an actionable and interpretable diagnostic tool for product and operational teams. Moving beyond traditional churn scores that only indicate the likelihood or timing of attrition, our model delivers granular insights into the "why" behind user disengagement. This empowers businesses to shift from a reactive to a proactive and highly targeted retention strategy. For instance, by identifying that a cohort of users is at high risk of churning due to "poor customer service" (a theme extracted by LDA), a company can deploy personalized intervention campaigns, such as targeted service recovery emails, rather than generic broadcast messages. This capability to diagnose the root cause of churn directly enables the optimization of user experience, the prioritization of product development roadmaps based on user feedback, and the design of personalized recall strategies, ultimately leading to a more efficient allocation of operational resources and a significant improvement in customer retention rates.

6. Conclusion

This research proposal outlines a comprehensive framework for enhancing user churn prediction by integrating the thematic analysis of negative app reviews with advanced statistical duration models. The study is motivated by a critical gap in the extant literature, which often overlooks the nuanced reasons for churn in favor of predicting its binary occurrence or timing. To address this, we propose a novel methodology that synergizes Latent Dirichlet Allocation (LDA) for topic modeling from unstructured text data with a Competitive Risk Model, an extension of the Duration Model. This integration allows for the quantification of subjective user feedback into interpretable risk categories, thereby creating a more granular and explanatory prediction system.

The anticipated workflow involves crawling a substantial dataset of negative reviews from major app stores, employing LDA to distill predominant complaint themes, and using these themes to define the competing risks within the hazard model. The evaluation will be rigorous, utilizing perplexity for the LDA model's performance and established statistical criteria like Log-Likelihood, AIC, and BIC for the comparative assessment of the churn prediction models.

We acknowledge potential limitations, primarily the challenge of applying LDA to short-text user comments, which may affect topic coherence. Future work will therefore focus on data acquisition, model training, and thorough evaluation. The expected contribution of this study is twofold. Theoretically, it advances the field of customer analytics by presenting

a novel methodological fusion of text mining and survival analysis for a deeper causal understanding of churn. Practically, it provides businesses with an actionable diagnostic tool, enabling product and operations teams to move beyond generic churn scores towards targeted, reason-specific interventions. Ultimately, this research aims to equip internet companies with the insights needed to improve user experience, design effective retention strategies, and foster long-term customer loyalty.

References

[1] Peter, S. F., Bruce G. S. H. (2009). Customer-base valuation in a contractual setting: The perils of ignoring heterogeneity. Marketing Science, 29(1):85-93.

[2] Xu, X., Thong, J.Y., Venkatesh, V. (2014). Effects of ict service innovation and complementary strategies on brand equity and customer loyalty in a consumer technology market, Information Systems Research, 25(4): 710–729.

[3] Gupta, S., Lehmann, D. R., & Stuart, J. A. (2004). Valuing customers. Journal of Marketing Research, 41: 7-18.

[4] Keaveney, S. M. (1995). Customer switching behavior in service industries: an exploratory study. Journal of Marketing, 59(2), 71-82.

[5] Cameron, A., Trivedi, P. (2005). Microeconometrics, Cambridge University Press.

[6] Braun, M., & Schweidel, D. A. (2011). Modeling customer lifetimes with multiple causes of churn. Marketing Science, 30(5): 881-902.

[7] Caigny, A. D., Coussement, K., Bock, K. W. D., & Lessmann, S. (2019). Incorporating textual information in customer churn prediction models based on a convolutional neural network. International Journal of Forecasting, 36: 1563–1578.

[8] Blei, D. M., Ng, A., & Jordan, M. I. (2003). Latent dirichlet allocation. The Journal of Machine Learning Research, 3: 993–1022.