Predicting Replacement of Smartphones with Mobile App Usage

Dun Yang[1], Zhiang Wu[1(✉)], Xiaopeng Wang[2], Jie Cao[1], and Guandong Xu[3]

[1] School of Info. Engineering, Nanjing University of Finance and Economics, Nanjing, China
zawuster@gmail.com

[2] Jiangsu Posts & Telecommunications Planning and Designing Institute, Nanjing, China

[3] Advanced Analytics Institute, University of Technology, Sydney, Australia

Abstract. To identify right customers who intend to replace the smartphone can help to perform precision marketing and thus bring significant financial gains to cellphone retailers. In this paper, we provide a study of exploiting mobile app usage for predicting users who will change the phone in the future. We first analyze the characteristics of mobile log data and develop the temporal bag-of-apps model, which can transform the raw data to the app usage vectors. We then formularize the prediction problem, present the hazard based prediction model, and derive the inference procedure. Finally, we evaluate both data model and prediction model on real-world data. The experimental results show that the temporal usage data model can effectively capture the unique characteristics of mobile log data, and the hazard based prediction model is thus much more effective than traditional classification methods. Furthermore, the hazard model is explainable, that is, it can easily show how the replacement of smartphones relate to mobile app usage over time.

Keywords: App usage · Smartphone replacement · Hazard model · Mobile log data

1 Introduction

Recent few years have witnessed a smartphone surge, due in large part to the rapid advances in mobile Internet technologies. Such high-speed yet ubiquitous access to the mobile Internet has given birth to a variety of applications (apps) to facilitate people's daily life. For example, as of the end of June 2015, there are more than 3 million apps and 594 million smartphones in China.

The study on the usage of mobile apps can help to understand users' behavior and preferences, which not only motivates the development of many intelligent services or adaptive user interfaces but also provides new marketing opportunities [6,14]. There are mainly two aspects of this interesting problem that the existing research has been done. On the one hand, the usage prediction and classification of mobile apps themselves [8,10,14], which can help users to search

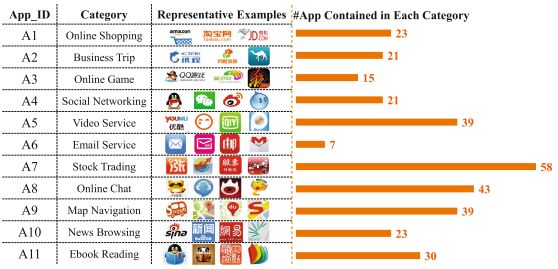

Fig. 1. Selected applications and categorization.

and launch apps efficiently. On the other hand, to exploit the app usage for developing other business intelligence services, such as recommendation, churn controlling, target advertising [13]. The topic of this paper falls into the second aspect, but we address another business problem. That is, we target at predicting the replacement of smartphones by mining mobile app usage data[1]. Since mobile Internet service providers (e.g., China Telecom) usually sell mobile phones simultaneously, pinpointing a set of users who intend to change their phones will improve marketing strategies and thus bring significant financial gains. There is therefore a clear need to apply data mining techniques on mobile Internet log data for the phone-replacement prediction.

This paper investigates the phone-replacement prediction problem by applying a unique data set that consists of individual-level log records. We present the temporal bag-of-apps model for representing these raw data in a convenient format. Then, we propose a hazard model from survival analysis to predict the replacement of phones. There is a varying time interval between a user changes his/her phone and the time point that we censor the app usage of this user. Hence, the hazard based models are preferred over both standard regression models and classification models for this problem, due to their ability to model particular factors of *duration data*. Furthermore, the hazard based prediction model is very simple and explainable, which are the important merits for real-world applications. In other words, with the hazard model, we can easily examine how the replacement of cell-phones relate to mobile app usage over time.

2 Data Description

Our sample consists of 411,331 mobile users who used 3G and 4G Internet service between January 12, and January 14, 2016. There are over 130 million log records, each of which denotes a user has used a specific application. Every record contains the information with the following attributes: `user_ID`, `MEID` (Mobile Equipment IDentifier), `app_ID`, `start_time` and `end_time`. `MEID` is a globally unique number identifying a physical mobile phone, and it is used to judge a

[1] We use the terms "replace phone" and "change phone" interchangeably in this paper. Both terms mean a user changes his/her *physical* mobile device.

user whether replace his/her phone. That is, if we observe a user with same `user_ID` yet different `MEID`, we say this user has changed his/her cell-phone.

We extract a total of 319 typical applications and manually divide them into 11 categories, including online shopping, business trip, online game, social networking, video service, email service, stock trading, online chat, map navigation, news browsing, and ebook reading, denoted by `A1` to `A11` respectively. Figure 1 shows several representative apps of each category, and the number of apps contained in every category. As can be seen, video service and social networking are two most popular categories, including 58 and 43 apps respectively.

3 Data Model

As introduced above, our raw data contains information about at what time and how long a user has used an application. For instance, a user spends averagely one hour on Wechat in evening. To analyze these temporal frequency data, we need to adopt a specific data model for transforming these raw data into a convenient format. To address the behavioral changes with respect to the time of a day, we discretize 24 h into 4 timeslots [5]: *Morning (6 am–12 am)*, *Afternoon (12 am–6 pm)*, *Evening (6 pm–0 am)* and *Night (0 am–6 am)*. These timeslots are denoted by their initial letters (i.e., $m - a - e - n$).

We then quantify the app usage for each timeslot on our observed data. Let T denote the observed time period (e.g., 3 days) and $s \in \{m, a, e, n\}$ denote one of timeslots. Denote the usage time of ith user on jth app in the timeslot s as t_{ij}^s. Thus, the average usage on the observed data is defined as $U_{ij}^s = \frac{1}{|T|} \sum_{s \in T} \frac{t_{ij}^s}{|s|}$. We use the time bucket rather than the frequency to model the app usage. With the temporal bag-of-apps (TBoA) representation, we can represent a user as a vector, each element is a triplet $\langle timeslot, app, usage \rangle$. Since we have categorized all apps into 11 classes and a day into 4 timeslots, a total of 44 features are generated for each user.

4 Prediction Model

In this section, we first formularize the prediction problem mathematically. Then we present the hazard based prediction model and derive its inference procedure.

4.1 Problem Statement

Assume the observed period is $[t_0, t_0 + T]$, where t_0 is the starting time point and T is the interval of observation. Besides the averaged app usage described by TBoA model, we can also observe whether a user has replaced his/her smartphone. It provides class labels denoted as $y_l, l \in \{1, 0\}$ indicating the user has changed his/her phone or not. In the meanwhile, the event that a replacement of phone happens also has a timestamp. Let S be a non-negative random variable representing the waiting time until the occurrence of this event. In what

follows, we call S as *survival time* [4]. For example, we might observe a user replaces his/her phone at $t_0 + \Delta t, \Delta t \leq T$, where this user is labeled as y_1 with survival time $S = \Delta t$. In contrast, if a user did not change his phone during the observation period, the user is labeled as y_0 with survival time $S = T$.

We aim to learn a prediction model based on the observation data during $[t_0, t_0 + T]$, including the usage vector, class label and survival time. Then, given any user represented by a usage vector, we want to predict whether he/she will replace his/her mobilephone in a pre-defined future period denoted by t.

4.2 Hazard Model

The nature of our problem is to predict the probability that an event is going to happen after t units of time. In the literature [4,7], a statistical approach called survival analysis has provided a rich set of methods for handling the time of occurrence of events. Here, we formularize the problem of predicting the replacement of phones by using the survival analysis model.

Let $\mathbf{U}_i = [\mathbf{U}_i^m, \mathbf{U}_i^a, \mathbf{U}_i^e, \mathbf{U}_i^n]$ be the usage vector of ith user, where each element $U_{ij}^s, s \in \{m, a, e, n\}$ is the average usage vector. The vector $\mathbf{U}_i$ can be interpreted as a set of covariates. We then define the *hazard* function $\lambda(t|\mathbf{U}_i)$ to measure the instantaneous rate of occurrence of the event.

$$\lambda(t|\mathbf{U}_i) = \lim_{dt \to 0} \frac{\Pr(t \leq S < t + dt)}{dt}. \tag{1}$$

According to Eq. (1), $\lambda(t|\mathbf{U}_i)$ means the probability that the replacement of phones to be occurred at time t, given an individual with covariates $\mathbf{U}_i$. A large family of models introduced by [4] focus directly on the hazard function. Among them, the *proportional hazard* model is the simplest and most widely-used one. Denote a set of regression coefficients corresponding to covariates as $\boldsymbol{\beta}$. The Cox's proportional hazard model is expressed as

$$\lambda(t|\mathbf{U}_i) = \lambda_0(t)\exp(\mathbf{U}_i\boldsymbol{\beta}^T), \tag{2}$$

where $\lambda_0(t)$ is the baseline hazard function describing the risk for individuals with $\mathbf{U}_i = \mathbf{0}$. Note that $\lambda_0(t)$ does not depend on $\mathbf{U}_i$ but only on t. In contrast, $\exp(\mathbf{U}_i\boldsymbol{\beta}^T)$ is the relative risk, a proportionate increase or decrease in risk, associated with the set of features $\mathbf{U}_i$.

4.3 Model Estimation

The parameter estimation of Cox's proportional hazard model generally consists of two parts: (1) the parameter to determine the baseline hazard function $\lambda_0(t)$, and (2) the set of regression coefficients $\boldsymbol{\beta}$. In the literature [4], various kinds of distributions for modeling $\lambda_0(t)$ are introduced, among which the *Weibull* distribution has gained the particular attention. This is because the Weibull distribution can provide a flexible model to depict the baseline risk varying with time. Meanwhile, the Weibull distribution remains in the exponential family that

coincides the Cox's proportional hazard function. The baseline hazard function in Weibull model is defined as $\lambda_0(t) = \frac{p}{k^p} t^{p-1} e^{-(t/k)^p}$, where $p > 0$ is the shape parameter and $k > 0$ is the scale parameter. We can fit the training data to $\lambda_0(t)$, i.e., to determine the parameter p and k.

As for the estimation of $\boldsymbol{\beta}$, the *partial likelihood* is often employed as the objective function. We define an indicator $\delta_i = 1$ to signify the ith user has changed the phone, otherwise for $\delta_i = 0$. The objective function is

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{N} \left(\frac{\exp(\mathbf{U}_i \boldsymbol{\beta}^T)}{\sum_{j \in R(t_i)} \exp(\mathbf{U}_j \boldsymbol{\beta}^T)} \right)^{\delta_i}, \tag{3}$$

where N is the total number of observed users, t_i is the time unit that ith user changed the phone and $R(t_i)$ is the set of users observed at time t_i. Taking the logarithm, we have

$$\log L(\boldsymbol{\beta}) = \sum_{i=1}^{N} \delta_i (\mathbf{U}_i \boldsymbol{\beta}^T - \log \sum_{j \in R(t_i)} \mathbf{U}_j \boldsymbol{\beta}^T). \tag{4}$$

Similar to the Logistic regression, we can solve the $\log L(\boldsymbol{\beta})$ maximization problem by using the Newton-Raphson method [2]. In fact, we focus on estimation of the regression coefficients $\boldsymbol{\beta}$, yet regarding the estimation of $\lambda_0(t)$ as another independent procedure. This approach is also known as the *non-parametric* strategy. A handsful of standard packages are available for estimating $\boldsymbol{\beta}$, such as R, SAS and Python, among which we choose the Python package to improve the computational efficiency on the large-scale data.

5 Experimental Results

In this section, we evaluate the performance of the hazard model for solving our proposed smartphone replacement prediction problem. To obtain the ground-truth label indicating whether changing the phone, we gather the `MEID` attribute of every user in the log data in the next three weeks. Figure 2 shows statistics of the collected ground-truth information. As can be seen, there are totally 87,346 (21.2 %) users who have replaced the phone in three weeks, where nearly half users changed the phone in the first week.

We select three classification models, including Logistic Regression (LR), Naïve Bayesian (NB) and Random Forest (RF). The usage vector of every users (see Sect. 3) are employed as the input of all competitive classifiers. As the ground-truth is available, we adopted the widely-used precision (P), recall (R) and F-measure (F) as evaluation measures. The precision is the ratio of truly identified users changing the phone and the total users who are predicted to change the phone. The recall is the ratio of truly identified users changing the phone and the number of users who have changed the phone.

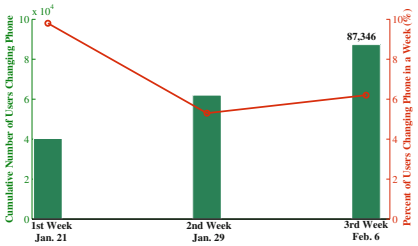

Fig. 2. Statistics of users who change the phone in three weeks.

5.1 Prediction Performance

Here, we evaluate the effectiveness of Cox's proportional hazard model for predicting the smartphone replacement. Firstly, we compare our hazard model with three baseline algorithms via the 10-fold cross validation. The comparative results are shown in Table 1, where the maximal value of each metric is bolded. As can be seen, except the NB model performs poorly, the precision of other three models are very close to each other, i.e., around 0.6. However, the recall value of our hazard model is much higher than those of other models, which leads to the its best overall performance indicated by F. These results state that the traditional classifiers are very strict, that is, they tend to identify a small fraction of users who truely change the phone. By the control of the survival time, our hazard model can seize more potential users with changing behaviors. Interestingly, the RF model was regarded as one of best algorithms for chrun prediction [7,11], and also RF performed best among three baseline methods. Since the RF model used TBoA for modeling mobile log data, this result demonstrate our data model, i.e., TBoA to represent the usage, can effectively fit the prediction for changing phones.

Table 1. Performance comparison (10-fold cross validation)

Algorithm	P	R	F
Logistic Regression	0.614	0.253	0.358
Naïve Bayesian	0.484	0.167	0.248
Random Forest	**0.646**	0.567	0.604
Cox's Prop. Hazard	0.581	**0.762**	**0.660**

5.2 Analysis on Covariates

Besides the satisfactory prediction accuracy, the explainability of a prediction model is crucial to decision makers, who except to know whether there is a positive or negative interdependence between any factor and a given event. With the hazard regression, we can easily observe the importance of every covariate in

Table 2. Several important positive and negative covariates

	No	ID	Coefficient	Standard error	Semantics
Positive	1	n_5	4.93e–02	8.96e–03	Night, Video Service
	2	n_4	4.82e–02	5.26e–03	Night, Social Networking
	3	n_8	3.62e–02	3.18e–03	Night, Online Chat
	4	m_5	1.39e–03	8.96e–03	Morning, Video Service
	5	m_8	1.37e–04	1.78e–04	Morning, Online Chat
Negative	6	m_{10}	–1.47e–03	8.12e–05	Morning, News Browsing
	7	m_7	–8.65e–03	9.26e–05	Morning, Stock Trading
	8	a_7	–3.81e–03	8.81e–05	Afternoon, Stock Trading
	9	a_2	–4.95e–03	2.45e–05	Afternoon, Business Trip
	10	a_{11}	–5.97e–03	1.68e–05	Afternoon, Ebook Reading

terms of their coefficients, i.e., $\boldsymbol{\beta}$. We select a case from 10-fold cross validation and show several covariates with their coefficients. Table 2 shows ten important covariates exerting both positive and negative impacts to the event of replacing smartphones. All covariates listed in Table 2 are highly significant with p-value $< 10^{-4}$, using two-tailed t-test. The regression coefficient tells us how much a unit change in the value of the covariate impacts the user's rate of changing the phone.

Three observations are noteworthy from Table 2. First, none app category in evening has been included in the important covariates, because the usage in evening is much higher than that of other timeslots. Thus, the usage vector in evening has weak discriminative power for different users. Second, as indicated by #1, #2 and #3 covariates, users playing their phones at night tend to change their phones frequently, and the launched apps are very trendy, e.g., video service, social networking and online chat. We can bold guess most users changing their phones are probably teeny-bopper. Third, the negative covariates are in sharp contrast to the positive ones. In detail, many traditional apps exert negative affect to the smartphone replacement behavior, such as news browsing, stock trading, business trip and ebook reading.

6 Related Work

Our work is related to a group of literature about mobile apps usage analysis and churn prediction. Firstly, the studies on the usage analysis focus on understanding when and where apps are used in mobile phones [1]. The contextual usage patterns can then be leveraged for apps prediction (or recommendation) [8,10], which usually guides the development of adaptive user interfaces [5]. Moreover, a number of mobile recommender systems and target advertising engines have been designed based on the app usage analysis [9,13].

Secondly, a related line of research has studied the problem of churn prediction. This problem is often defined as a binary classification problem where users are categorized based on a set of behavioral features into two classes: future churners or non-churners. A lot of classification models have been utilized for churn prediction including logistic regression, neural networks and support vector machines, though the random forest is found to be better in performance [3,11]. Furthermore, the survival analysis is widely used for analyzing or predicting user behavior in online environment, including how user participation patterns affect the lifetime on online knowledge sharing communities [12], and predicting return time for web services [7]. The above work inspires our research a lot when handling this new problem of predicting the replacement of smartphones.

7 Concluding Remarks

This paper provided a study of exploiting mobile app usage for predicting users who will change the phone in the future. In particular, we first analyzed the characteristics of mobile log data that we obtained from a large telecommunications service company in China. We designed the temporal bag-of-apps (TBoA) model for data representation, and presented the hazard based prediction model. The experimental results demonstrated that the hazard base prediction model was thus much more effective than traditional classification methods. Furthermore, we analyzed the important positive and negative covariates to show the good explainability of the proposed hazard prediction model.

Acknowledgments. This research was partially supported by National Natural Science Foundation of China under Grants 71571093, 71372188 and 61502222, National Center for International Joint Research on E-Business Information Processing under Grant 2013B0135, National Key Research and Development Program of China under Grant 2016YFB1000901, and Industry Projects in Jiangsu S&T Pillar Program under Grant BE2014141.

References

1. Böhmer, M., Hecht, B., et al.: Falling asleep with angry birds, Facebook and kindle: a large scale study on mobile application usage. In: MobileHCI, pp. 47–56 (2011)
2. Böhning, D.: Multinomial logistic regression algorithm. Ann. Inst. Stat. Math. **44**(1), 197–200 (1992)
3. Buckinx, W., Van den Poel, D.: Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. EJOR **164**(1), 252–268 (2005)
4. Cox, D.R.: Regression models and life-tables. In: Kotz, S., Johnson, N.L. (eds.) Breakthroughs in Statistics. Springer, New York (1992)
5. Do, T.M.T., Gatica-Perez, D.: By their apps you shall understand them: mining large-scale patterns of mobile phone usage. In: MUM (2010)
6. Ghose, A., Han, S.P.: An empirical analysis of user content generation and usage behavior on the mobile internet. Manag. Sci. **57**(9), 1671–1691 (2011)

7. Kapoor, K., Sun, M., Srivastava, J., Ye, T.: A hazard based approach to user return time prediction. In: KDD, pp. 1719–1728 (2014)
8. Parate, A., Böhmer, M., Chu, D., et al.: Practical prediction and prefetch for faster access to applications on mobile phones. In: UbiComp, pp. 275–284 (2013)
9. Shi, Y., Karatzoglou, A., Baltrunas, L., Larson et al.: TFMAP: optimizing map for top-n context-aware recommendation. In: SIGIR, pp. 155–164 (2012)
10. Shin, C., Hong, J.H., Dey, A.K.: Understanding and prediction of mobile application usage for smart phones. In: UbiComp, pp. 173–182 (2012)
11. Xie, Y., Li, X., Ngai, E., Ying, W.: Customer churn prediction using improved balanced random forests. Expert Syst. Appl. **36**(3), 5445–5449 (2009)
12. Yang, J., Wei, X., et al.: Activity lifespan: an analysis of user survival patterns in online knowledge sharing communities. ICWSM **10**, 186–193 (2010)
13. Yuan, B., Xu, B., Chung, T., Shuai, K., Liu, Y.: Mobile phone recommendation based on phone interest. In: Benatallah, B., Bestavros, A., Manolopoulos, Y., Vakali, A., Zhang, Y. (eds.) WISE 2014. LNCS, vol. 8786, pp. 308–323. Springer, Heidelberg (2014). doi:10.1007/978-3-319-11749-2_24
14. Zhu, H., Chen, E., Xiong, H., Cao, H., Tian, J.: Mobile app. classification with enriched contextual information. IEEE Trans. Mob. Comput. **13**(7), 1550–1563 (2014)