Quantifying and Leveraging User Fatigue for Interventions in Recommender Systems

Hitesh Sagtani
hiteshsagtani@sharechat.co
ShareChat
Bengaluru, India

Madan Gopal Jhawar
madangopal@sharechat.co
ShareChat
Bengaluru, India

Akshat Gupta
akshatgupta@sharechat.co
ShareChat
Bengaluru, India

Rishabh Mehrotra
rishabhmehrotra@sharechat.co
ShareChat
London, UK

ABSTRACT

Predicting churn and designing intervention strategies are crucial for online platforms to maintain user engagement. We hypothesize that predicting churn, i.e. users leaving from the system without further return, is often a delayed act, and it might get too late for the system to intervene. We propose detecting early signs of users losing interest, allowing time for intervention, and introduce a new formulation of *user fatigue* as short-term dissatisfaction, providing early signals to predict long-term churn. We identify behavioral signals predicting fatigue and develop models for fatigue prediction. Furthermore, we leverage the predicted fatigue estimates to develop fatigue-aware ad-load balancing intervention strategy that reduces churn, improving short- and long-term user retention. Results from deployed recommendation system and multiple live A/B tests across over 80 million users generating over 200 million sessions highlight gains for user engagement and platform strategic metrics.

CCS CONCEPTS

• **Information systems** → *Recommender systems*; *Personalization*;

KEYWORDS

user churn, user fatigue, recommendation system, social media

ACM Reference Format:

Hitesh Sagtani, Madan Gopal Jhawar, Akshat Gupta, and Rishabh Mehrotra. 2023. Quantifying and Leveraging User Fatigue for Interventions in Recommender Systems. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3539618.3592044

SIGIR '23, July 23–27, 2023, Taipei, Taiwan

ACM ISBN 978-1-4503-9408-6/23/07…$15.00
https://doi.org/10.1145/3539618.3592044

1 INTRODUCTION

Developing better understanding of how users interact with recommendation systems has become increasingly important for improving user experience on the platform. Predicting user churn is one such mechanism to understand user unhappiness and it helps companies in customer relationship management. Earlier approaches in the industry leveraged logistic regression and decision trees for churn prediction problem because of their ease of explanation and interpretation [3, 5, 9, 11, 21]. Yu *et al.* [26] proposes ESVM, whereas Wu *et al.*[23] leveraged improved SMOTE technique to handle large scale class imbalance to predict churn. Recent approaches have matured to use sequential modeling, with Rakesh *et al* [19] predicting customer churn using user interactions data and attention based LSTM model. Churn prediction has been widely adopted by a number of industries including content based network [6], messaging apps [24], gaming applications [14], Q&A sites [18] and chatbots [1, 2]. Finally, recent work has started adopting survival models for predicting long term survival of the user on the platform [7, 15]. While churn prediction is an important task, by the time churn prediction scores are high, it often gets too late for the system to intervene and improve user experience in an attempt to retain users. To this end, we introduce a formulation *user fatigue* – a score that helps us identify scenarios where the user starts losing interest in the platform and starts to show churn intent.

In addition to user churn, there are various short term implicit user signals from user interaction data have been traditionally used to gauge user satisfaction in search and recommender systems: dwell time [25], proportion of video watched [22], click & scroll data [12] and other interaction signals [13]. Interpreting relationship of such signals to long-term user engagement is an open issue. Dupret and Lalmas[8] proposed to address this by measuring user loyalty using absence time – the time between visits metric. More recently, metrics like time-to-inactivity have been defined and used as part of survival models for estimating long-term survival of the user on the platform [4]. Our findings also support the use of inactive period as a proxy for churn, a concept we build upon to define the proposed concept of user fatigue.

While most of the existing work is on predicting churn, the notion of early churn detection, and intervention to stop long term churn is less studied in the literature. Hatt *et al.* [10] explored the possibility of early detection of user exit and introduced a markov

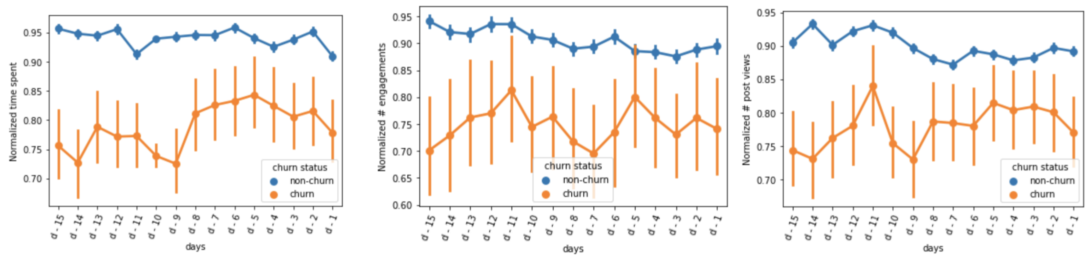

Figure 1: Historical engagements variations for churn vs non churned users

modulated marked point process (M3PP) model for the same. Oskarsdottir *et al.* [17] proposed a dynamic customer behavior model and used similarity forests to detect churn early, allowing organizations to make arrangements for retention in a timely manner and hence using churn estimates for intervention. Other work on intervention strategies to reduce churn include deploying marketing strategies [20] and personalized push notifications in gaming applications [16]. In this work we instead focus on short term user fatigue with the platform and highlight that short term fatigue often leads to long term churn, if not acted upon. Furthermore, we go beyond simply predicting fatigue and demonstrate that leveraging the predicted fatigue can indeed help us devise timely intervention strategies, thereby reducing user churn on the platform.

2 MODELLING USER FATIGUE

We define *user fatigue* as the temporary dissatisfaction of the user with the platform where user starts to lose interest and can possibly leave the platform without further return. We collected user data from ShareChat, a widely popular social media application with over 180M monthly active users in 18 regional languages in India.

User Fatigue vs User Churn While user fatigue is a temporary phenomenon, user churn is a long-term event. In this case, we call users churned when they do not return to the platform for at least 60 days. In this section, we present analysis to understand the differences between fatigue and churn. Figure 1 plots historical user activities like time spent, engagements and posts-viewed against future churn and non-churn users. As we can see, there is a significant difference between user activities in the two sets which indicates that users who are going to churn in the future tend to be less active long before they actually leave. Also, the 95% confidence interval for the non-churned users lie very close to the mean values as compared to much larger variance in the same for the churned users. This indicates that there are large fluctuations in user activities when they are about to churn. Users signal dissatisfaction by these patterns of increasing and decreasing activities over time which can be used to capture user fatigue.

2.1 Problem Formulation

We formulate user fatigue via a strong form of user dissatisfaction, i.e., users not returning back to the platform for a short duration. Since we want user fatigue to give early signals of user churn so that we can intervene, we plot the correlation between users not returning to the platform for next n consecutive days with future user churn. The value of n should be small in order to detect churn as early as possible. We treat n as hyperparameter and present below the formulation to find its value for a dataset.

The first plot in Figure 2 plots the Pearson Correlation between user churn and user not logging-in for next n days. We can see that there is a steep increase in the correlation value for the first 4 days after which the curve starts to flatten out. The same behaviour can be confirmed by the second plot which shows the first derivative (slope) of the curve, and the third plot where the False Positive Rate of user churn prediction decreases sharply for the first 4 days. Therefore, the elbow at n=4 provides an optimal point to define user fatigue. Detecting user fatigue on a given day is defined as a binary classification problem as follows:

$$\phi_{u,d} = \begin{cases} 1, & if\, \delta_{u,d} > \mathrm{d} + 4 \\ 0, & if\, \delta_{u,d} \leq \mathrm{d} + 4 \end{cases}$$

,where $\phi_{u,d}$ represents fatigue value for user u on day d, and $\delta_{u,d}$ represents next login day for user u after day d.

2.2 Prediction Models

Features and Dataset To detect user fatigue, we use a combination of user demographics, historical interactions, login, content and other derived features. In total, we used 165 features, some of which are defined in table 1. While we formulated user fatigue as a binary classification problem, the model outputs a real value probability score between $[0, 1]$, and we define it as the *user fatigue score*. We want to update the user fatigue scores at regular intervals to detect the users who have high probability of churn. Therefore, we calculate user fatigue scores each day for all the users who logged in on the platform at least once in the last 7 days (also known as Weekly Active Users or WAU) with the latest feature values based on user interactions with the platform on that day. For a random sample of 10M data points, we found that only a minority of the entries have fatigue label as '1' and the remaining samples have fatigue label as '0'. The dataset is divided into train, validation and test sets using stratified sampling in 70:15:15 ratio.

We evaluated several baseline and other binary classification models to predict user fatigue given the input features.

2.2.1 *Baseline Models.*

- Interaction Models: In the interaction models, for a given feature f and input number of days d, we compare the average value of the feature f over the last d days with the current feature value and if the current feature value is lower than the last d day's average feature value, the user is classified as fatigued and vice-versa.

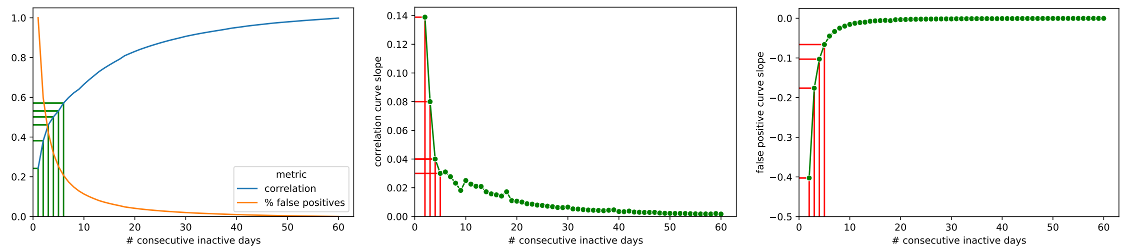

Figure 2: The Pearson Correlation and False Positives Rate between user inactivity and churn flatten out after the first 4 days

Attribute Type	Attribute	Description
Interaction attributes	time spent	total time spent by the user on a particular day
	#views	total post views by the user on a particular day
	#engagement	count of engagements by the user
Login attributes	is today active	whether user logged in today
	num inactive days	number of inactive days passed since user's last login
	was active yesterday	if the user was active yesterday
Historical attributes	fraction of active days	fraction of user login days for the 3, 5 and 7 day period
	historical time spent	average time spent by user for 3, 5 and 7 day period
	historical #views	average post views by user for 3, 5 and 7 day period
	historical #engagement	average engagements by user for 3, 5 and 7 day period
Content attributes	top k genre interactions	interactions with distinct genres in top 10 posts per feed
	top k vbi interactions	ratio of video to image interactions in top 10 posts per feed
	session distinct genre	number of distinct genre interactions within session
	session quit at ad	if user quit the session at ads post
	avg post age	average post age in the feed
	fav genre avg dwell time	average dwell time per feed for the genre with maximum affinity
User attributes	user platform age	number of days since user signed up on the platform
	language	one hot encoding of all the user languages is used

Table 1: Description of features used in the model training

- Login Models: In the login models, we compare the number of time user has logged into the platform over the last d days. If the login count is below a threshold value, the user is classified as fatigued.

2.2.2 Trained Models. We evaluated various supervised machine learning classification models like *Logistic Regression (LR), Multi-layer Perceptron (MLP)* and *XGBoost (XGB)* over a subset and all of feature values to predict user fatigue. To overcome the class imbalance problem in the dataset while training, we use the *weighted-loss function* to avoid over-fitting the models to majority class only. For hyper-parameter tuning of the supervised models, we used *RandomSearchCV* over several hyper-parameter combinations and selected the parameters which yielded best performance over the validation set.

Experiment Results We use *precision, recall* and *f1-scores* of the positive and negative classes on the test set as evaluation metrics, with an emphasis on *f1-score* given the class imbalanced. The model performance results are tabulated in table 2. We can see that the output metrics of *single feature* and *combination thresholding* models are in the similar range, whereas, the trained models produce far superior metrics on average. Among the trained models, we can see that XGB model yields highest average f1-score of 0.65, followed by MLP and LR respectively. Along with the dynamic user attributes like user platform age, we find that both the user feedback features and content features like time spent, last 7 days login/interaction activities like post viewed, engagements, number of sessions, session quit at ads, ads per post viewed have high feature significance. Highly significant content features affecting fatigue score gives us an opportunity to make appropriate intervention and improve user satisfaction. For example, *session quit at ads*, i.e., users quitting the session at an ad post, is one of the significant features indicating that certain users are unsatisfied with current volume of ads on the platform and decreasing the ads volume could improve their platform experience.

3 INTERVENTION BASED ON FATIGUE

The ability to detect user fatigue with the platform can help in avoiding long term user churn by designing timely interventions in order to improve engagement and retention. In this section, we present one such case study where we intervene by personalizing the number of ads the user sees based on their fatigue score.

Offline impact of Ads on User Fatigue: *ads/view* (number of ads per number of organic posts), *feed_quit_at_ad* (whether the user quits the feed at an ad) and *session_quit_at_ad* (whether the user quits the session altogether at an ad) are some of the important features used by the XGB model to predict fatigue. Using two tailed t-test on Pearson correlation coefficient, we see from table 3 that all the ads metrics show statistically significant correlation with p-value less than 0.05.

Ad-Load Policy Design Since ads significantly affect fatigue score, we design the following fatigue-aware ad-load decisioning system. Let ϕ_u represents the user fatigue score, θ represents the default number of ads shown to users in a feed on the platform, and θ_u represents the updated number of ads shown to the user, then the fatigue score-based ad intervention policy can be formulated

	Model	Features	Positive Class			Negative Class		
			Precision	Recall	F1 score	Precision	Recall	F1 score
Single Feature Thresholding	InteractionModel(views, 7)	#views	0.12	0.37	0.18	0.83	0.54	0.66
	InteractionModel(engagement, 7)	#engagement	0.15	**0.58**	0.24	0.86	0.45	0.59
	LoginModel(7, 1)	#login days	**0.37**	0.29	0.33	0.88	**0.91**	**0.89**
Combination Thresholding	LoginModel(7, 1) AND InteractionModel(engagement, 7)	#login days + #engagement	**0.42**	0.21	0.28	0.87	**0.95**	**0.91**
	LoginModel(5, 1) AND InteractionModel(engagement, 7)	#login days + #engagement	0.37	0.28	0.32	**0.88**	0.92	0.9
	LoginModel(7, 2) AND InteractionModel(engagement, 7)	#login days + #engagement	0.34	0.41	**0.37**	0.87	0.83	0.85
Trained models	Logistic Regression	User + Interaction	0.38	0.49	0.43	**0.88**	**0.82**	**0.85**
	Logistic Regression	All	0.49	0.6	0.54	0.87	0.81	0.84
	Multilayer Perceptron	User + Interaction	0.43	0.63	0.51	0.87	0.75	0.8
	Multilayer Perceptron	All	0.52	**0.77**	0.62	0.88	0.7	0.78
	XGBoost	User + Interaction	0.44	0.66	0.53	0.86	0.71	0.78
	XGBoost	All	**0.57**	0.75	**0.65**	0.88	0.76	0.82

Table 2: Experiment results of various baseline and binary classification models

Metric	correlation	p-value
ads/view	0.1353	0.0
feed_quit_at_ad	0.18363	0.0
session_quit_at_ad	0.21363	0.0

Table 3: Significant correlation of ads with user fatigue

Metric (α, β)	variant-1 (0.1, 0.8)	variant-2 (0.05, 0.8)	variant-3 (0.05, 0.7)	variant-4 (0.025, 0.8)	variant-5 (0.025, 0.7)
Time Spent	-0.40%	-0.36%	-0.42%	0.03%+	-0.44%
Post Views	0.01%+	0.08%+	-0.34%	0.15%	0.16%
Engagements	-0.21%	0.08%+	-0.25%	0.54%	0.37%
Ad Impressions	0.75%	0.24%	-0.13%+	0.04%+	0.26%
Ad Clicks	0.11%	0.17%	0.23%	0.17%	0.19%

Table 4: Core metrics for the fatigue policy. All results are statistically significant except for those marked by+

as:

$$\theta_u = \begin{cases} \theta + 1, & \text{if } \phi_u < \alpha \\ \theta, & \text{if } \alpha < \phi_u < \beta \\ \theta - 1, & \text{if } \phi_u > \beta \end{cases}$$

,*where* $0 < \alpha < \beta < 1$ are the thresholds used to determine high and low fatigue users.

Online A/B Test We conducted online A/B tests with several treatment variants with combinations of α and β values as tabulated in Table 4. In the control group, we showed the default number of ads per feed, θ, to all the users. In the treatment group, we showed ads based on the fatigue policy described above. The experiments were conducted on a random sample of traffic comprising of 80 million users across the variants over 7 days.

The treatment vs control metrics for the policy is shown in table 4. As we can see, variant-4 with $\alpha = 0.025$ and $\beta = 0.8$ yields the best results in terms of generating higher number of ad clicks while maintaining user satisfaction metrics like time spent and improving post views and engagements. Most importantly, variant-4 consistently yields statistically significant higher retention for users as compared to control variant for all days with average day-1 retention (users logging-in on the very next day) of +0.07%. This proves that by intervening the high fatigue users with lower number of ads at the right time, we are able to improve their platform retention. Furthermore, we have leveraged the fatigue score to show more ads to active users to generate higher revenue.

We investigated the short and long term change in user fatigue scores and churn data to understand and quantify the impact of the differential treatment. Using 7 days of rolling data, we find that, on average, fatigue score reduces for 0.44% of users with score greater than 0.8 compared to control, thus, proving the impact of the intervention in improving short term user fatigue. For the long term impact, we measure the percent of users who come back at least once in the next 2 months. For users with fatigue score < 0.025, the difference in retention is non-significant between treatment and control, whereas, for users with fatigue score > 0.8, where we reduced the ad load, the retention increased by 0.36% in treatment compared to control. Based on the results, this fatigue-aware policy was subsequently deployed to 100% production traffic.

We developed and evaluated a similar system of calculating user fatigue scores at regular intervals and using it to intervene at the right time for high fatigue score users on Moj - a short-video only social media application by ShareChat with 300M monthly active users in India. In this case, we intervened on the number of Live posts (posts with live-stream content) shown to the users. In an online A/B test of existing system vs fatigue score based intervention system, we found that the average number of users returning on the platform on the next day increased by 0.09% in the test variant. Based on the findings, the system has been deployed to production on Moj. The results solidify our hypothesis that detecting short term user fatigue with platform and designing right intervention strategies can reduce long term user churn and works on multiple platforms and intervention strategies beyond ads.

4 CONCLUSION

We introduced the notion of *user fatigue*, demonstrated the need to predict early signals of permanent churn and established the relationship between short term fatigue and long term churn. An important reason to estimate user fatigue is to be able to design timely interventions to reduce churn. To this end, we demonstrated a fatigue-aware ad-load balancing strategy that is able to dynamically adjust the amount of ads shown to the user based on their real time fatigue score. We conducted a series of A/B tests on a large scale live production system demonstrated that a fatigue-aware policy is able to improve user experience. We envision future extensions to include intervention policies that are able to pin-point the causal factors increasing fatigue and intervene on the recommendation policy to reduce fatigue.

ACKNOWLEDGEMENTS

We are grateful to the Balancer teams on Sharechat and Moj, specifically Suman Singha, Anish Kumar, Sourabh Rana, Ram Vivek and Rahul Aggarwal for their help in insightful discussions, experimentation, and analysis across building and evaluating fatigue intervention policies in Sharechat and Moj products.

REFERENCES

[1] Christian Abbet, Meryem M'hamdi, Athanasios Giannakopoulos, Robert West, Andreea Hossmann, Michael Baeriswyl, and Claudiu Musat. 2018. Churn intent detection in multilingual chatbot conversations and social media. *arXiv preprint arXiv:1808.08432* (2018).

[2] Essam Abou el Kassem, Shereen Ali Hussein, Alaa Mostafa Abdelrahman, and Fahad Kamal Alsheref. 2020. Customer Churn Prediction Model and Identifying Features to Increase Customer Retention based on User Generated Content. *IJACSA) International Journal of Advanced Computer Science and Applications* 11, 5 (2020).

[3] Ionut Brandusoiu and Gavril Toderean. 2013. Churn prediction modeling in mobile telecommunications industry using decision trees. *Journal of Computer Science and Control Systems* 6, 1 (2013), 14.

[4] Praveen Chandar, Brian St. Thomas, Lucas Maystre, Vijay Pappu, Roberto Sanchis-Ojeda, Tiffany Wu, Ben Carterette, Mounia Lalmas, and Tony Jebara. 2022. Using Survival Models to Estimate User Engagement in Online Experiments. In *Proceedings of the ACM Web Conference 2022*. 3186–3195.

[5] Preeti K Dalvi, Siddhi K Khandge, Ashish Deomore, Aditya Bankar, and VA Kanade. 2016. Analysis of customer churn prediction in telecom industry using decision trees and logistic regression. In *2016 Symposium on Colossal Data Analysis and Networking (CDAN)*. IEEE, 1–4.

[6] Kushal S Dave, Vishal Vaingankar, Sumanth Kolar, and Vasudeva Varma. 2013. Timespent based models for predicting user retention. In *Proceedings of the 22nd international conference on World Wide Web*. 331–342.

[7] Simon Demediuk, Alexandra Murrin, David Bulger, Michael Hitchens, Anders Drachen, William L Raffe, and Marco Tamassia. 2018. Player retention in league of legends: a study using survival analysis. In *Proceedings of the Australasian computer science week multiconference*. 1–9.

[8] Georges Dupret and Mounia Lalmas. 2013. Absence time and user engagement: evaluating ranking functions. In *Proceedings of the sixth ACM international conference on Web search and data mining*. 173–182.

[9] Mohammed Hassouna, Ali Tarhini, Tariq Elyas, and Mohammad Saeed AbouTrab. 2016. Customer churn in mobile markets a comparison of techniques. *arXiv preprint arXiv:1607.07792* (2016).

[10] Tobias Hatt and Stefan Feuerriegel. 2020. Early detection of user exits from clickstream data: A markov modulated marked point process model. In *Proceedings of The Web Conference 2020*. 1671–1681.

[11] Sebastiaan Höppner, Eugen Stripling, Bart Baesens, Seppe vanden Broucke, and Tim Verdonck. 2020. Profit driven decision trees for churn prediction. *European journal of operational research* 284, 3 (2020), 920–933.

[12] Dmitry Lagun and Mounia Lalmas. 2016. Understanding and measuring user engagement and attention in online news reading. In *Proceedings of the ACM International Conference on Web Search and Data Mining*. 113–122.

[13] Mounia Lalmas and Liangjie Hong. 2018. Tutorial on metrics of user engagement: Applications to news, search and E-commerce. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. 781–782.

[14] Jiayu Li, Hongyu Lu, Chenyang Wang, Weizhi Ma, Min Zhang, Xiangyu Zhao, Wei Qi, Yiqun Liu, and Shaoping Ma. 2021. A Difficulty-Aware Framework for Churn Prediction and Intervention in Games. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 943–952.

[15] Yunfei Lu, Linyun Yu, Peng Cui, Chengxi Zang, Renzhe Xu, Yihao Liu, Lei Li, and Wenwu Zhu. 2019. Uncovering the co-driven mechanism of social and content links in user churn phenomena. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 3093–3101.

[16] Miloš Milošević, Nenad Živić, and Igor Andjelković. 2017. Early churn prediction with personalized targeting in mobile social games. *Expert Systems with Applications* 83 (2017), 326–332.

[17] María Óskarsdóttir, Tine Van Calster, Bart Baesens, Wilfried Lemahieu, and Jan Vanthienen. 2018. Time series for early churn detection: Using similarity based classification for dynamic networks. *Expert Systems with Applications* 106 (2018), 55–65.

[18] Jagat Sastry Pudipeddi, Leman Akoglu, and Hanghang Tong. 2014. User churn in focused question answering sites: characterizations and prediction. In *Proceedings of the 23rd International conference on world wide web*. 469–474.

[19] Vineeth Rakesh, Ajith Pudiyavitil, and Jaideep Chandrashekar. 2020. User Modeling and Churn Prediction in Over-the-top Media Services.. In *ORSUM@ RecSys*.

[20] Shini Renjith. 2015. An integrated framework to recommend personalized retention actions to control B2C E-commerce customer churn. *arXiv preprint arXiv:1511.06975* (2015).

[21] Petra Posedel Šimović, Davor Horvatic, and Edward W Sun. 2021. Classifying variety of customer's online engagement for churn prediction with mixed-penalty logistic regression. *arXiv preprint arXiv:2105.07671* (2021).

[22] Siqi Wu, Marian-Andrei Rizoiu, and Lexing Xie. 2018. Beyond views: Measuring and predicting engagement in online videos. In *Twelfth international AAAI conference on web and social media*.

[23] Xiaojun Wu and Sufang Meng. 2016. E-commerce customer churn prediction based on improved SMOTE and AdaBoost. In *2016 13th International conference on service systems and service management (ICSSSM)*. IEEE, 1–5.

[24] Carl Yang, Xiaolin Shi, Luo Jie, and Jiawei Han. 2018. I know you'll be back: Interpretable new user clustering and churn prediction on a mobile social application. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 914–922.

[25] Xing Yi, Liangjie Hong, Erheng Zhong, Nanthan Nan Liu, and Suju Rajan. 2014. Beyond clicks: dwell time for personalization. In *Proceedings of the 8th ACM Conference on Recommender systems*. 113–120.

[26] Xiaobing Yu, Shunsheng Guo, Jun Guo, and Xiaorong Huang. 2011. An extended support vector machine forecasting framework for customer churn in e-commerce. *Expert Systems with Applications* 38, 3 (2011), 1425–1430.