Leveraging User Email Actions to Improve Ad-Close Prediction

Oleg Zendel
RMIT University
oleg.zendel@rmit.edu.au

Yaroslav Fyodorov
Yahoo Research
yfyodorov@verizonmedia.com

Fiana Raiber
Yahoo Research
fiana@verizonmedia.com

Natalia Silberstein
Yahoo Research
natalias@verizonmedia.com

Oren Somekh
Yahoo Research
orens@verizonmedia.com

Ali Tabaja
Yahoo Research
atabaja@verizonmedia.com

ABSTRACT

Online advertising systems often provide means for users to close ads and also leave feedback. Although closing ads requires additional user engagement and usually indicates a poor user experience, ad closes are not as scarce as one might expect. Recently it was shown that penalizing ads with high closing likelihood during auctions may substantially reduce the number of ad closes while maintaining a small predefined revenue loss. In this work, we focus on email since this is the property in which most ad closes occur. Using data collected from a major email provider, we present interesting insights about the interplay between ad closes in email and email-related user actions. In particular, we explore the merits of integrating information derived from user actions in email for ad-close prediction. Thorough performance evaluation reveals that incorporating such signals significantly improves ad-close prediction quality over previously reported results.

ACM Reference Format:
Oleg Zendel, Yaroslav Fyodorov, Fiana Raiber, Natalia Silberstein, Oren Somekh, and Ali Tabaja. 2020. Leveraging User Email Actions to Improve Ad-Close Prediction. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20), October 19–23, 2020, Virtual Event, Ireland.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3340531.3412093

1 INTRODUCTION

Online advertising is one of the most influential economic forces driving the Internet. All parties are aware of the high cost of bad ads, and efforts are made to provide the best experience for users [6]. Nevertheless, there will always be users that dislike certain ads in some cases. Therefore, online advertising systems usually provide means for users to close ads and also leave feedback, such as selecting a reason why and even writing free text [14].

It was recently shown that user engagement with the ad-close mechanism is quite significant, especially in email properties, and is in the order of magnitude of ad clicks [14]. Since ad closes usually indicate bad user experience, mostly due to poor ad quality, an effort was made to mitigate the number of ad-close actions [14]. In particular, several features including those derived from past ad closes were used to predict ad-close probability. The predicted probabilities served for estimating the "true"[1] expected revenue to be used during *generalized second price* (GSP) auctions [4]. It was shown that penalizing ads with high closing likelihood provides an inherent incentive for advertisers to improve their ads' quality. The proposed system was tested online serving real Yahoo Gemini traffic and demonstrated a hefty 20% reduction in the number of ad closes while maintaining less than 0.4% revenue loss.

In this work we strive to further improve user experience by providing more accurate ad-close prediction. We focus on email properties where most ad closes occur using data collected from the Yahoo Mail app[2], one of the globally largest email providers. In these properties sponsorship transparent *native ads*[3] are presented to users at the top of their email inboxes. We identify *user email-related actions* (also referred to as *email actions*) occurring when users interact with email services, such as opening, sending and deleting messages, as key features in achieving our goal. Along with the previously reported ad-close predictions (referred to as pClose, see [14]) that are provided with each ad impression and user demographic features, email actions arranged into n-grams are combined and used as features in a *logistic regression* (LR) model. Although the LR model is simple, it still provides insights into the prediction task. In addition to showing improved performances over previously reported results [14], the LR model reveals that certain email actions, such as message deletions are the top contributing features. Interpreting these may suggest that a user that is deleting email messages may be in a "cleaning" mood and also tends to close ads that she dislikes.

In addition to the standard LR model, we apply more advanced *deep learning* (DL) techniques to further improve prediction quality. In particular, we use an architecture that combines two networks. The first network is based only on email actions, while the second uses demographic features and the pClose signal [14]. The performance of the LR and DL models were evaluated using data collected from Yahoo Mail mobile app traffic during a fortnight earlier this year. We show that our approach of combining email actions with the pClose signal greatly improves ad-close prediction quality when compared to the previously reported results [14].

CIKM '20, October 19–23, 2020, Virtual Event, Ireland

ACM ISBN 978-1-4503-6859-9/20/10…$15.00
https://doi.org/10.1145/3340531.3412093

[1] More accurate, which accounts for ad-closes long-term hidden costs [14].

[2] All processes performed as part of our data construction and analysis were conducted under the European and US privacy regulations.

[3] Native ads resemble the surrounding page items, are considered less intrusive to users, and provide a better user experience in general. In contrast to search ads, user intent is usually unknown which makes ad matching more challenging.

The main contributions of this work are as follows.

- Focusing on email properties, we introduce the rather unexplored interplay between email-related user actions and ad closes using data collected from the Yahoo Mail app.
- Applying standard LR and more advanced DL models, we demonstrate the importance of email actions to the ad-close prediction task.
- Conducting a thorough performance evaluation, we show a significant improvement in ad-close prediction quality when compared to previous techniques.

2 RELATED WORK

Several works consider user feedback in the context of online advertising. The studies in [1, 14, 16] present prediction models for ad quality and demonstrate how the predicted values can be used to improve ad ranking. While the main signal used in [16] is the offensive ad feedback, the works in [1, 14] consider a more general type of ad feedback, namely removal of ads to which a user was exposed.

Email action prediction has attracted much research attention recently. The authors in [2] analyzed different email actions and devised a learning framework for predicting them. Prediction of the email reply action was presented in [3, 10, 13, 15]. Lastly, [5, 8] study promotional emails, where the authors in [5] present a framework to predict the unsubscription action, and the authors in [8] consider the task of click-through rate prediction in promotional emails.

Most of the works mentioned above focus on the prediction of ad feedback, in particular, the closure of undesired ads, and on the prediction of specific email related user actions. To the best of our knowledge, there is no prior work that leverages user email actions for ad feedback prediction.

3 ANALYSIS OF AD CLOSES IN EMAIL

In the analysis to follow, we examine the effect of demographic characteristics of users and the actions they performed when interacting with the email service on their disposition to close ads. The analysis is based on data gathered from the Yahoo Mail app during the second week of Apr. 2020. Hereinafter, we consider the single, most popular version of the app since different versions may offer different functionalities. The data includes observations about millions of users that were exposed to hundreds of thousands of unique ads.

We use $\mathcal{A}$ to denote the set of ad instances (impressions) that were examined in our analysis. For a subset of the ad instances $\mathcal{A}_i \subseteq \mathcal{A}$, close rate $CR(\mathcal{A}_i)$ is the ratio between the number of times the ads in $\mathcal{A}_i$ were closed and the number of times these ads were shown. We define the *relative change* in close rate as

$$RCCR(\mathcal{A}_i) \triangleq \frac{CR(\mathcal{A}_i) - CR(\mathcal{A})}{CR(\mathcal{A})} .$$

Demographic characteristics. We partitioned our users into groups based on their age and gender, and computed the relative change in close rate $RCCR(\mathcal{A}_{\mathcal{U}})$ for the set of ad instances $\mathcal{A}_{\mathcal{U}}$ that were presented to the users in each group $\mathcal{U}$. The results are presented in Table 1. We observe a rather clear trend when inspecting the different age groups: the older the users, the more likely they are to interact with the ad-close mechanism. The difference between women and men is less conspicuous.

Table 1: Relative change in ad close rate by age and gender.

≤ 20	21-40	41-60	> 60	women	men
−0.45	−0.49	0.11	1.98	−0.02	0.02

Email actions. We partitioned the ad instances into groups based on the actions the users performed before seeing an ad. An ad group $\mathcal{A}_x$ defined for action x includes all the ad instances in which x was performed before the ad was displayed. We limit ourselves to instances in which the number of actions taken between x and the ad display is at most ten. Note that the groups are overlapping since several different actions may be performed before an ad is shown. We inspected four of the most common actions performed when interacting with email messages: open, delete, edit and send, and two actions performed when interacting with ads: click and close.

Table 2 presents the relative change in close rate for each group. We can see that the close rate of the delete action is higher than the close rates of the other considered message-related actions. This finding suggests that a user deleting email messages might be in a “cleaning” mood and is therefore more likely to also close ads. The highest relative change is observed for the ad close action suggesting that a user who has already closed an ad, will continue to close other ads.

Table 2: Relative change in ad close rate by user actions.

message				ad	
open	delete	edit	send	click	close
−0.35	1.11	−0.56	−0.45	7.94	31.28

4 OUR APPROACH

We next present out data preparation process (Section 4.1), the features we consider (Section 4.2), and the LR (Section 4.3) and DL (Section 4.4) models.

4.1 Data

We collect sequences of email-related actions from the Yahoo Mail app. The actions are combined with data extracted from the native ads and includes the pClose predictions for each ad impression [14]. The action sequences are partitioned into sessions that are delimited by 15 minutes of user inactivity. Since several ads may be shown during a single session and the user can interact differently with each ad, we divided each session into *ad sessions*. An ad session is defined per each ad in the session and includes all the actions taken by the user from the beginning of the session until the ad was displayed. Thus, our ad sessions are overlapping and may contain user interactions with ads previously presented in the same session. An example of an email session that includes two ad sessions is shown in Figure 1. The user closed the ad presented in the first ad session and did not interact with the second one.

4.2 Features

We used three feature types: (i) features derived from the sequence of email-related actions in an ad session performed before the ad was shown; (ii) user demographic features (age and gender); and (iii)

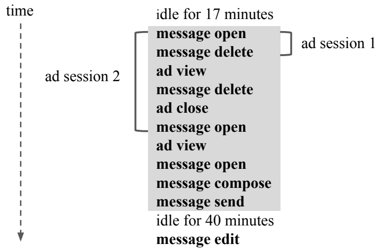

Figure 1: Example of an email session with two ad sessions.

the pClose signal [14]. We note that none of the features consider the content of the email messages.

The sequence of email actions may be viewed as a sequence of terms in a text. Accordingly, for LR, similar to the common practice in text classification, we represent the sequence using n-grams. In our setting, an n-gram is a combination of n adjacent actions, where unigrams ($n = 1$), bigrams ($n = 2$) and trigrams ($n = 3$) are used. The binary values of these features indicate the presence or absence of an n-gram in the sequence. We consider n-grams that occurred at least 2500 times in our ad sessions, which we found to be more effective than using other thresholds. For the DL model, the entire action sequence is passed as input to the network. An additional feature that we consider in LR and DL is the number of actions in the ad session (length).

For the age, gender and pClose features, we used a one-hot encoded representation. The gender feature can take three values ("man", "woman" and "unknown") and is therefore represented using a 3-dimensional vector. For age and pClose, we used quantile discretized (or binned) representation, in which an equal number of values are placed in each bin. After a naive grid search, we ended up using 20 bins for age and 95 bins for pClose.

4.3 Logistic Regression

We used a standard implementation that includes a bias and a weighted sum of the features embedded in a sigmoid function. The bias and weights were learned using stochastic gradient descent (SGD) and L2 regularization using the Spark MLlib [11] package.

4.4 Deep Learning

We trained two separate networks. The input of the first network is the sequence of actions performed by users before seeing an ad. The AWD-LSTM architecture [12] with three stacked layers of LSTM was used to train an action-based language model. The language model then served as the encoder (backbone) in a classifier that included a sequence of batch normalization, dropout, a linear layer, and ReLU activation. The second network uses age, gender, length and pClose as features. For each feature, we learned an embedding matrix that maps the feature values into fixed-size vectors. The feature embeddings were concatenated into a single vector and passed to three similar stacked blocks, consisting of batch normalization, dropout, linear layer and ReLU activation. Batch normalization and dropout were omitted from the first block and ReLU was not used in the third block. Finally, we removed the last three layers in each network, concatenated the outputs into a single vector and fed it into a linear layer to produce the final result.

The models were implemented using the FastAI library [7] with default hyper-parameter values with the following exceptions. The sizes of the last two linear layers used in the second network were set to 100 and 50, respectively. In all cases, the cross-entropy loss function and ADAM optimization [9] were used.

5 EVALUATION

Section 5.1 describes the data extraction process, the evaluation metrics and the baseline. Section 5.2 presents the results.

5.1 Setting

Data. Our experiments are based on a sample of over 2 million email sessions collected from the Yahoo Mail app over a two-week time period (Dec. 21, 2019 to Jan. 4, 2020). As noted in Section 4.1, a session is a sequence of email-related user actions delimited by 15 minutes of inactivity. Considering all the sessions in which at least one ad was shown to a user yields highly imbalanced data because in most of these sessions the presented ads were not closed. Therefore, our data includes all the sessions in which one of the displayed ads was closed. In addition, for the group of users that closed at least one ad, we randomly sampled a similar number of sessions in which an ad was displayed but was not closed. Overall, our data includes over 6.5 million ad sessions. There are on average 38 actions in an ad session with a standard deviation of 37.6. In 13.5% of the ad sessions, the ad was closed.

The data was randomly split into train (60%), validation (20%), and test (20%) sets. The train set was further split for training the models (90%), and setting hyper-parameters (10%). The validation set was used to determine the number of feature bins, and the threshold for converting the output of our models into binary predictions used when computing the F1 metric.

Evaluation metrics. To evaluate the quality of our predictions, we use AUC (area under the ROC curve), LogLoss (logarithmic loss) and F1. AUC served for selecting hyper-parameter values.

Baseline. For our baseline we use the LR model with the binned version of pClose [14]. The binned version is used because our data construction process changes the ad-close rate and requires compensation provided by the learned weights of the LR model.

5.2 Results

Logistic regression. Table 3 presents the performance of the LR model for individual feature types and different feature groups. When considering each feature type separately, we can see that the worst performance in AUC and F1 is attained for gender. This finding is in line with the results reported in Table 1, where small differences between women and men were observed. The length of the action sequence is quite useful and its performance surpasses that of the two demographic features age and gender in terms of AUC and F1. The best performance of an individual feature type is attained for pClose, which also serves as our baseline.

Inspecting the action-based features, we can see that similar performance is attained for unigrams, bigrams and trigrams. However, when all three types of features are combined (denoted actions in Table 3) additional improvements are observed, and the resultant performance is merely 1.7% lower in AUC than the baseline. This

Table 3: Logistic regression performance of different groups of features. The best result in a column is underlined.

	AUC	LogLoss	F1
unigrams	0.826	0.293	0.478
bigrams	0.827	0.298	0.487
trigrams	0.827	0.293	0.474
age	0.606	0.387	0.269
gender	0.545	0.393	0.241
length	0.695	0.406	0.334
actions = unigrams + bigrams + trigrams	0.836	0.282	0.496
actions + age + gender + length	0.846	0.276	0.516
pClose	0.850	0.265	0.585
all = actions + age + gender + length + pClose	0.919	0.203	0.688

Table 4: Top logistic regression features by importance.

Feature name	LR weights	Coverage	Importance
ad view+message delete+message delete	5.00	48.5%	1.25
ad view+list scroll	2.51	23.7%	0.45
list view+message close	1.31	24.7%	0.24
list refresh+list refresh+ad view	2.70	9.5%	0.23

finding attests to the importance of the action-based features and their complementary nature. Finally, the best performance in Table 3 is attained when combining all the features, with a significant improvement of 8.1%, 23.4%, and 17.6% in AUC, LogLoss, and F1 respectively, over the baseline pClose.

Feature importance. To assess the importance of the different email-related actions for prediction, we trained the LR model with all considered features and examined the resulting weights. In addition, we computed a proxy for feature importance by multiplying the absolute value of the LR weight by $C_f \cdot (1 - C_f)$, where C_f is the feature coverage. Note that $(1 - C_f)$ penalizes high coverage binary features that may act like biases when their coverage approaches 1. The importance factors are then ranked where higher values imply higher importance. The top four important features are listed in Table 4. We can see that the highest ranked feature, with more than twice the importance value than that of the runner-up, is the trigram “ad view+message delete+message delete.” This suggests that this action sequence is highly indicative of ad-close events, and that the user tends to close ads more after “cleaning” her email inbox by repeatedly deleting unwanted messages. We note that despite its high relative change value reported in Section 3, the “ad close” action has little importance (rank 53) due to its low coverage.

Deep learning. Table 5 presents the results of the LR and DL models when considering all the features and for the two feature groups used by the two networks in the DL model. We see that for a given set of features, the performance of DL almost always surpasses that of LR. When considering all the features, DL outperforms LR and the baseline for all metrics. In particular, it provides improvements of 8.8%, 24.9%, and 18.8% in AUC, LogLoss, and F1, respectively, over the baseline. These results consist of a small but non-negligible improvement provided by the more advanced DL techniques over the standard LR model (0.7%, 2%, and 1% in AUC, LogLoss and F1, respectively).

Table 5: Comparison of logistic regression (LR) and deep learning (DL) models. Underline: best result in a column.

		AUC	LogLoss	F1
LR	actions	0.836	0.282	0.496
	age + gender + length + pClose	0.878	0.249	0.605
	all	0.919	0.203	0.688
DL	actions	0.845	0.283	0.504
	age + gender + length + pClose	0.882	0.246	0.607
	all	0.925	0.199	0.695

6 CONCLUDING REMARKS

We demonstrated the potential of utilizing email-related actions to provide better ad-close predictions for improved user experience. We combined the actions collected from the Yahoo Mail app with features derived from user demographic characteristics and a previously proposed ad-close predictor. The features were used to train standard LR and more advanced DL models. Our evaluation reveals that our email-action-based predictors provide significant performance lifts over the previously reported results. Future work includes the integration of email actions into online advertising systems, which is a challenging task due to the long delays it takes to log the actions and train the ad ranking models.

Acknowledgements. We thank the reviewers for their comments and Dan Pelleg for helpful discussions.

REFERENCES

[1] M. Bron, K. Zhou, A. Haines, and M. Lalmas. Uncovering bias in ad feedback data analyses & applications. In Proc. of WWW, pages 614–623, 2019.

[2] D. Di Castro, Z. S. Karnin, L. Lewin-Eytan, and Y. Maarek. You've got mail, and here is what you could do with it!: Analyzing and predicting actions on email messages. In Proc. of WSDM, pages 307–316, 2016.

[3] M. Dredze, T. Brooks, J. Carroll, J. Magarick, J. Blitzer, and F. Pereira. Intelligent email: Reply and attachment prediction. In Proc. of IUI, pages 321–324, 2008.

[4] B. Edelman, M. Ostrovsky, and M. Schwarz. Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. The American economic review, 97(1):242–259, 2007.

[5] I. Gamzu, L. Lewin-Eytan, and N. Silberstein. Unsubscription: A simple way to ease overload in email. In Proc. of WSDM, page 189–197, 2018.

[6] D. G. Goldstein, R. P. McAfee, and S. Suri. The cost of annoying ads. In Proc. of WWW, pages 459–470, 2013.

[7] J. Howard et al. fastai. https://github.com/fastai/fastai, 2020.

[8] K. Jaidka, T. Goyal, and N. Chhaya. Predicting email and article clickthroughs with domain-adaptive language models. In Proc. of WebSci, page 177–184, 2018.

[9] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In Proc. of ICLR, 2015.

[10] F. Kooti, L. Maria Aiello, M. Grbovic, K. Lerman, and A. Mantrach. Evolution of conversations in the age of email overload. In Proc. of WWW, pages 603–613, 2015.

[11] X. Meng, J. K. Bradley, B. Yavuz, E. R. Sparks, S. Venkataraman, D. Liu, J. Freeman, D. B. Tsai, M. Amde, S. Owen, D. Xin, R. Xin, M. J. Franklin, R. Zadeh, M. Zaharia, and A. Talwalkar. MLlib: Machine learning in apache spark. Journal of machine learning research, 17:34:1–34:7, 2016.

[12] S. Merity, N. S. Keskar, and R. Socher. Regularizing and optimizing LSTM language models. In Proc. of ICLR, 2018.

[13] B.-W. On, E.-P. Lim, J. Jiang, A. Purandare, and L.-N. Teow. Mining interaction behaviors for email reply order prediction. In Proc. of ASONAM, pages 306–310, 2010.

[14] N. Silberstein, O. Somekh, Y. Koren, M. Aharon, D. Porat, A. Shahar, and T. Wu. Ad close mitigation for improved user experience in native advertisements. In Proc. of WSDM, pages 546–554, 2020.

[15] L. Yang, S. T. Dumais, P. N. Bennett, and A. H. Awadallah. Characterizing and predicting enterprise email reply behavior. In Proc. of SIGIR, pages 235–244, 2017.

[16] K. Zhou, M. Redi, A. Haines, and M. Lalmas. Predicting pre-click quality for native advertisements. In Proc. of WWW, pages 299–310, 2016.