An Effective Ensemble Framework with Multichannel Time Series for User Retention Prediction

The first-place entry for Retention Score Prediction at WSDM CUP 2022

Zhiruo Li, Zhihui Cui, Shunyao Wu*
Qingdao University
Qingdao, China
wushunyao@qdu.edu.cn

ABSTRACT

User retention is an important indicator that can help companies to understand user loyalty. Efficient solutions to predict user retention is essential for Internet companies like iQIYI to attract and keep more users. However, it is challenging to predict user retention due to the diversity of user behaviors and the complexity of multichannel time series. In this paper, we proposed an ensemble framework to address the user retention score prediction challenge in ACM WSDM Cup 2022. In our solution, we extracted several useful features based on multichannel time series and assembled four tree-based models to predict the retention score. Besides, we developed post-processing methods to adjust the predictions which are crucial to improve the performance. Fortunately, our team QDU won 1st place with an evaluation score of 86.4486 on the private leader board. The source code is available at https://github.com/hansu1017/WSDM2022-Retention-Score-Prediction.

KEYWORDS

User Retention, Ensemble Framework, Multichannel Time Series, Post-processing

ACM Reference Format:
Zhiruo Li, Zhihui Cui, Shunyao Wu. 2022. An Effective Ensemble Framework with Multichannel Time Series for User Retention Prediction: The first-place entry for Retention Score Prediction at WSDM CUP 2022. In *Proceedings of WSDM '22: ACM International Conference on Web Search and Data Mining (WSDM '22).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

With the widespread usage of the Internet and mobile devices, hundreds of online systems are being developed every year. As the market matures and the cost of acquiring new users rises, every platform is striving to attract and keep more users. Therefore, improving user retention is of great significance for consumer based internet companies, which is not only beneficial to optimize the existing products or services but is also helpful to eliminate the uncertainty associated with a sudden loss of customer base and reduce operating costs.

In Retention Score Prediction Challenge, which is created by iQIYI, participants were challenged to develop a solution that predicts the 7-day retention score of any user. For example, a user having 7-day retention score 3 means this user would launch the iQIYI app in 3 separate days during the next 7 days. It is non-trivial to predict the retention score as the task has two challenges. On the one hand, the start and end points of sampling time and behavior patterns vary from user to user, and hence how to handle this difference is a major challenge, which is essential to select similar users for the target population and extract appropriate training samples. On the other hand, with a wide variety of data and huge user behavior logs, it is difficult to effectively deal with multichannel time series (e.g. login sequence, playtime sequence, historical retention score sequence) and extract useful features for user retention prediction.

To address the above mentioned challenges, we developed an effective ensemble framework based on multichannel time series to predict user retention prediction. Rather than all users, we only selected users in the test set to construct the training set, which made the distribution of the training set consistent with the test set. For each user, all available dates were exploited to generate samples which contain more information of users' login habits. To effectively integrate the multiple sequences, we extracted statistical features of the historical retention score sequence for tree-based models and adopted deep convolutional neural networks (CNN) to investigate the multichannel time series data. Moreover, a powerful distribution-free independent test, mean variance test [2, 3], was employed to select feasible features. Our approach achieved 86.4338 and 86.4486 scores in the public and private leader boards, respectively.

2 PROPOSED FRAMEWORK

Figure 1 demonstrates the main flow of our proposed framework. Firstly, we extracted statistical features and trained four tree-based regression models including LightGBM [5], Catboost [4], Xgboost [1] and TensorFlow Decision Forests

*Corresponding Author.

WSDM '22, February 21–25, 2022, Phoenix, AZ, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

(TFDF) [7], and assembled the prediction results by the harmonic mean. Secondly, we constructed two CNN models for binary classification by integrating multichannel time series, which were used to determine whether users would log in and whether they would log in every day during the next 7 days, respectively. Finally, the predictions were post-processed by the two CNN models and some rules.

2.1 Preprocessing

When constructing the offline training set in stage A, we found that the behavioral patterns and data distribution among users varied greatly. Therefore, only users from the online test set were selected to construct the training set. In order to fully exploit information, for each user, we utilized all available dates to generate training samples and moved forward 7 days from the end date of the online test set to create offline test samples. Since the percentage of samples with a retention score of 0 in the offline set is obviously larger than that in the online test set, we randomly selected this kind of samples at a ratio of 0.7, which is an important trick to improve the online performance.

2.2 Feature Engineering

Through data analysis and exploration, we extracted hundreds of features based on user playback data, user portrait data, video related data, user launch logs and historical retention scores. Then, we applied the mean variance test (MvTest) [3] to test whether the features are statistically associated with the retention scores. The method was mainly based on mean variance index [2] to test whether a continuous variable and a discrete variable are independent, which has no assumptions on the distribution of variables. Finally, 33 statistical features are selected, and the top 20 features evaluated by the test statistic (Tn) of MvTest are illustrated in Figure 3. In addition, to explore the periodic pattern of user login, we also extracted three sequences including login sequence, video playback duration sequence and historical retention score sequence.

2.2.1 User Playback based Features. Since the retention score was more correlated with the user behaviors close to the end date, we extracted the number of videos watched by users and the total playtime per day in the last week.

2.2.2 User Portrait based Features. For user portrait data, all information except the territory code was utilized, which contained the rom and ram of the device, sex, age, education and occupation status. For the case that some users had multiple devices, we replaced the multiple values with their mean values.

2.2.3 Launch Logs based Features. Based on app launch logs, we extracted several useful features including the login status on the end date (*is_launch* and *launch_type_new*), the total number of historical logins (*launchNum*), the total number of logins in the last week (*NumLastWeek*) and the time difference between the last login and the end date (*diff_near*). Among them, the time difference feature effectively reflected the user's stickiness to the platform and improved the performance to a certain extent.

2.2.4 Week. Through analyzing the number of logins from date 131 to date 160, we found that it exhibited a cyclical pattern which repeated itself every week. According to the cyclical pattern, we can infer the day of the week of any date. For example, as shown in Figure 2, the number of logins significantly increased at the date 143 and 144 and suddenly declined at the date 145. Hence, date 143 and date 144 were weekends since users had more free time for entertainment, while date 145 was Monday because most users were busy working.

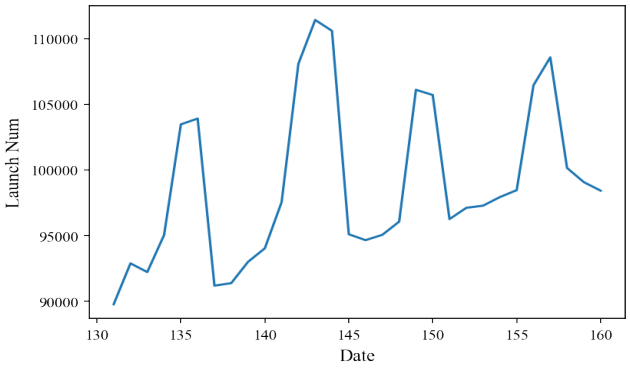

Figure 2: Time plot of the login number.

2.2.5 Historical Retention Score based Features. The historical retention score series was crucial information to forecast the retention score of the next 7 days. We generated four useful features which were the median (*preds_median_30*) and the weighted median (*weighted_median*) of the historical retention score in the last month and the mean (*preds_mean_4*) and the weighted mean (*preds_mean_4_weighted*) of the historical retention score at the corresponding date in the last four weeks. For instance, when the end date of a given user was 160, the corresponding dates in the previous four weeks were 153, 146, 139 and 132. The weighted mean and the weighted median were set larger weights for more recent dates. These features effectively reflected the periodic characteristics of user logins and were important features as they all obtained approximately 85 scores in the private leader board.

2.2.6 Sequences. For the last 64 days, we extracted the login sequence and the user playtime sequence. In addition, the historical retention score sequence was generated from the time range of [end date-71, end date-7].

2.3 Model Averaging

We applied four tree-based regression models including LightGBM, Catboost, XGBoost and TFDF to predict the 7-day retention score. Then, we assembled the predicted results by the harmonic mean.

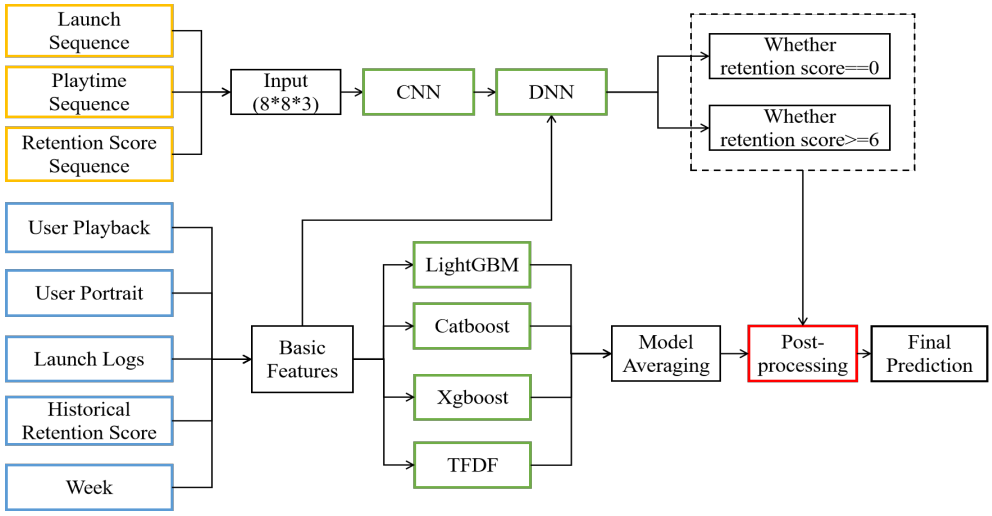

Figure 1: An overall framework and pipeline of our solution.

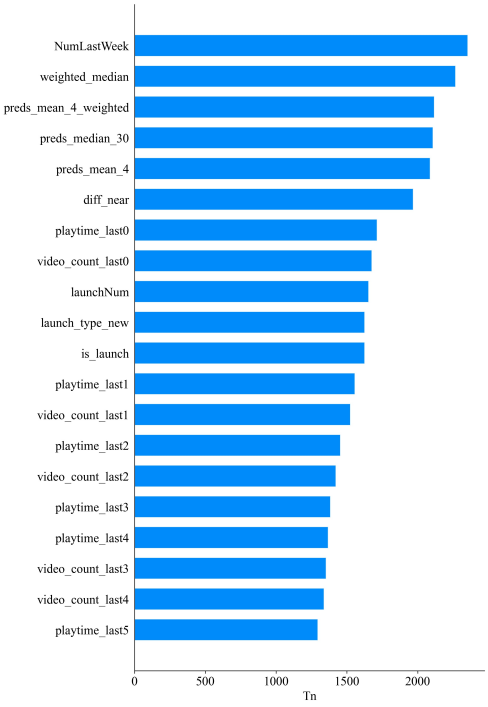

Figure 3: Feature importance of the top 20 features evaluated by MvTest.

2.4 CNN on Multichannel Time Series

In recent years, CNN based on multichannel time series has achieved excellent performance [6]. Therefore, we constructed two CNN models to handle the multichannel time series and found they could outperform other models in post-processing. As shown in Figure 4, we transformed the three sequences into a 3-dimensional array with size $8*8*3$. The CNN model contained two convolutional layers and two max-pooling layers. In convolutional layers, the model learned the periodic features by moving the $3*3$ convolutional filters and generated a feature map which was subsequently input to a fully connected layer. Then, the output of the fully connected layer was merged with the statistical features where the discrete variables were encoded to 32-dimensional embedding vectors. Finally, the probabilities were predicted through a DNN with a sigmoid layer.

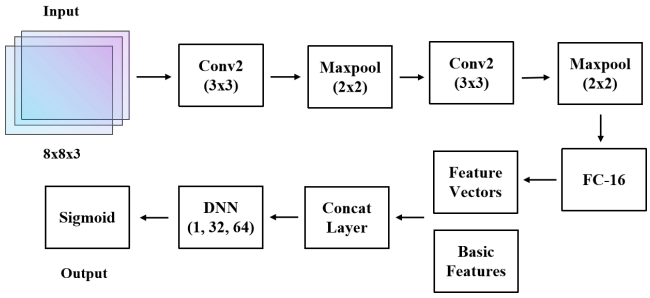

Figure 4: CNN on Multichannel Time Series

2.5 Postprocessing

The evaluation index of the competition was the absolute difference between the actual values and the predicted values. Therefore, through data analysis and exploration, we developed some post-processing methods to make the predictions closer to the true retention score.

Firstly, for users who did not log in for a long time, it was almost impossible to predict the next login date. And the empirical analysis demonstrated that the probability was extremely low for such users to suddenly launch the application during the next days. Therefore, the predictions

Table 1: Intervals for adjusting predictions.

Interval	Adjusted Prediction
(0, 0.5)	0
(0.6, 1.4)	1
(1.55, 2.4)	2
(2.55, 3.4)	3
(3.55, 4.4)	4
(4.55, 5.2)	5
(5.55, 6.2)	6

Table 2: Performance of some statistical features.

Statistical Features	Stage A	Stage B
preds_mean_4	84.522	84.581
preds_mean_4_weighted	84.994	85.013
weighted_median	83.778	83.906

for users that did not log in for more than 30 days were replaced with 0. Secondly, for users who never played a video in the past, the probability to log in again is quite small. Thus, for any user who never played a video and had a small prediction value less than 0.5, the predictions were reset to 0. Thirdly, we utilized the two CNN models to determine whether the 7-day retention score was 0 and whether it was greater than or equal to 6, respectively. With the former CNN, the predictions with top 4500 probabilities were modified to 0; with the latter CNN, the results with top 1200 probabilities were modified to 7. Finally, the predictions were adjusted according to Table 1.

3 EXPERIMENTS

Through data analysis and exploration, we found some statistical features of historical retention scores that could effectively reflect the user login pattern. Table 2 showed the online performance of three important features. Both the mean and weighted mean of the 7-day retention score at the corresponding date in the last four weeks can achieve a score around 85. Besides, the weighted median of the last month was also an effective statistic which can achieve a score close to 84.

Table 3 demonstrates the performance of four tree-based models, the model averaging result and the final post-processed result in Stage B. On the online test, Catboost outperformed the other three tree-based models, while the performance of Xgboost was a little worse to some extent. The model averaging can slightly increase the score by approximately 0.02. Moreover, the post-processing methods played an important role in elevating the final score which raised the score by 0.664.

4 CONCLUSIONS

In this paper, our team proposed an effective ensemble framework to address the user retention score prediction challenge.

Table 3: Performance comparison in Stage B.

Model	Offline	Online
LightGBM	85.819	85.742
Catboost	85.818	85.763
Xgboost	85.809	85.739
TFDF	85.835	85.750
Model Averaging	85.840	85.785
Post-processing	**86.561**	**86.449**

By data analysis and exploration, we extracted 33 statistical features and 3 sequence features. Then, we adopted four tree-based regression models including LightGBM, Catboost, Xgboost and TFDF to forecast the 7-day retention score and assembled the predicted results by the harmonic mean. To make the predictions closer to the actual values, we developed effective post-processing methods which were crucial to enhance the retention score prediction.

ACKNOWLEDGMENTS

This paper is supported by the grant 21BTJ045 from National Social Science Found of China. We thank everyone associated with organizing and sponsoring the WSDM Cup 2022. Dataset was provided by iQIYI. and the challenge was sponsored and managed by the 15th ACM International Conference on Web Search and Data Mining (WSDM 2022). The competition platform was hosted by iQIYI and DataFountain. We are very grateful to the WSDM Cup Chairs and the staff of iQIYI for their great efforts during the challenge.

REFERENCES

[1] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. *CoRR* abs/1603.02754 (2016). arXiv:1603.02754 http://arxiv.org/abs/1603.02754

[2] Hengjian Cui, Runze Li, and Wei Zhong. 2015. Model-Free Feature Screening for Ultrahigh Dimensional Discriminant Analysis. *J. Amer. Statist. Assoc.* 110, 510 (2015), 630–641. https://doi.org/10.1080/01621459.2014.920256

[3] Hengjian Cui and Wei Zhong. 2019. A distribution-free test of independence based on mean variance index. *Computational Statistics & Data Analysis* 139, C (2019), 117–133. https://doi.org/10.1016/j.csda.2019.05.00

[4] Anna Veronika Dorogush, Andrey Gulin, Gleb Gusev, Nikita Kazeev, Liudmila Ostroumova Prokhorenkova, and Aleksandr Vorobev. 2017. Fighting biases with dynamic boosting. *CoRR* abs/1706.09516 (2017). arXiv:1706.09516 http://arxiv.org/abs/1706.09516

[5] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

[6] Jian Bo Yang, Minh Nhut Nguyen, Phyo Phyo San, Xiao Li Li, and Shonali Krishnaswamy. 2015. Deep Convolutional Neural Networks on Multichannel Time Series for Human Activity Recognition. In *Proceedings of the 24th International Conference on Artificial Intelligence* (Buenos Aires, Argentina) *(IJCAI'15)*. AAAI Press, 39954001. https://doi.org/10.5555/2832747.2832806

[7] Zhi-Hua Zhou and Ji Feng. 2017. Deep Forest: Towards An Alternative to Deep Neural Networks. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*. 3553–3559. https://doi.org/10.24963/ijcai.2017/497