Retention Score Prediction via Tabnet

Team:
ASDF

Qian Zhang
Longying Zhida Technology Co., Ltd

Huayuan Sun
Peking University
sunhuayuan@stu.pku.edu.cn

ABSTRACT

Using large-scale data to accurately predict user retention is a common problem faced in current business practice. Due to the large scale of users, large differences between different users, and the impact of time series data, it is challenging to model user retention prediction. In this paper, we explore the application of deep learning on large-scale user datasets with temporal features to create high-accuracy user retention prediction model. The proposed model was submitted in the WSDM 2022 Cup User Retention Score Prediction Challenge and achieved the second place among 991 teams[1].

Keywords: User Retention, Predictive models, TabNet

1 INTRODUCTION

User retention is a key indicator to measure user satisfaction. The accurate prediction of user retention can support the maintenance and operation of existing customers. Combined with models such as precision marketing, churn warning and sleep customer activation, it can improve customer value and promote business growth. Therefore, even slight improvements in accuracy can lead to a significant increase in profits.

With the rapid development of modern machine learning methods, user retention prediction models have been applied in business practice. However, the current mainstream methods are still dominated by GBDT-based models, and deep learning are rarely used. In addition, the challenges of establishing a user retention prediction model lie in the large differences in the preferences and activities of different users, the impact of sampling errors and the processing of time series data.

For the purposes of this paper, we participated in the WSDM 2022 Cup User Retention Score Prediction Challenge to address the complicated task of accurately predicting user retention score using deep learning to explore methods of sample extraction and engineering time series features in large datasets.

The goal for the challenge was to accurately predict a user's 7-day retention score, i.e., the number of days a user will log in in the next 7 days, which is a typical regression problem. The dataset for analysis was extracted from iQIYI, a leading music streaming service in Asia which utilizes a subscription based business model. More than 500 million users enjoy entertainment services on iQIYI every month. Model performance was evaluated using the following function.

$$100 * (1 - \frac{1}{n}\sum_{t=1}^{n}\left|\frac{F_t - A_t}{7}\right|)$$

The final model submissions were scored against a final test set and ranked according to the function.

2 DATASET

The dataset analyzed in this paper came from the WSDM 2022 Cup Challenge and was provided by iQIYI, a music streaming service. The dataset consists of 5 distinct sources: user portrait data, app launch logs, video related data ,user playback data and user interaction data, involving a total of 600,000 users, of which the number of users in the A-list test set is 15,000 and the B-list The number of users in the test set is 35,000.

2.1 Train set

A sample is drawn for each user_id, and the login date of the sampled user is between the earliest login date and a period before the last login. If the time interval between the earliest login and the last login is too short, no sampling will be performed. The target variable is the number of logins 7 days after the login date of the sampled user.

2.2 Validation set

Validation uses KFold for five-fold cross-validation, 4 copies of training, 1 copy of validation.

2.3 Sampling

When constructing the training set, the login date of the sampled users is between the earliest login and the last login date. This sampling method defaults that the sampled users will log in in the future, which will cause the predicted value of the retention score of the low retention users to be enlarged, resulting in sampling error. Therefore, the time interval from the last login should be appropriately relaxed during sampling. In order to obtain a more

[1] Code: https://github.com/shyjin/WSDM

reasonable target variable, sampling should be carried out at least 7 days before the last login.

In this solution, the time interval is relaxed to 10 days, which has a certain restraint effect on the overestimation of retention score caused by sampling error, but it also leads to the reduction of the sample size and the risk of underestimating the retention score for other samples. In short, it is recommended to add a coefficient for sampling adjustment in the train set sampling process.

In addition, we also noticed that there are a large number of users who have not logged in for a long time in the test set, while users in the train set will log in in the future by default, which leads to inconsistent label distributions between the train set and the test set and the predicted value of the retention score of users who have not logged in for a long time in the test set is unreasonable.

Since sampling error cannot be eliminated, we made some adjustments to the data based on actual business conditions. In the actual business scenario, the next login of users who have not logged in for a long time is random and the probability of continuous login in the future is extremely low, so we set a step function to correct the prediction of the retention score of users who have not logged in for a long time.

3 FEATURE ENGINEERING

Features were derived from data within the 5 data sources, and a total of 214 features are constructed, of which the main features are derived from the app launch logs, target_encoding of each classified data, and statistical indicators of historical data.

The feature engineering methods for each data source are as follows:

- User portrait data: perform mean processing on multi-device parameters, and split the territory code for aggregation;
- App launch logs: Extract features such as whether the user has logged in on the current day, the duration of the previous n logins from the current day, the number of consecutive logins, the total number of historical logins, etc.
- Video related data: the unique number of items contained in father_id
- User playback data: Combined with video portraits, the average duration is used as an estimate of the actual duration, as well as the statistical features of video playback duration and quantity.
- User interaction data: Aggregate and extract statistical features for video items and interaction types.

4 MODEL

In this paper, we use a deep learning network structure for tabular data, TabNet[1]. The model implements instance-wise feature selection through a sequential attention mechanism similar to the additive model, and also implements self-supervised learning through the encoder-decoder framework, which well combines the interpretability of tree models with the representational power of DNNs.

The advantages of Tabnet are as follows:

- The model can directly use tabular data without preprocessing; the gradient descent-based optimization method makes it easy to add to the end-to-end model.
- At each decision time step, the sequence attention model is used to select important features and learn the most prominent features, which makes the model interpretable.

TabNet has two obvious advantages. On the one hand, it shows similar or even better model effects than other models in both classification and regression, and it also has local interpretability and global interpretability.

5 CONCLUSION

This paper introduces a user retention score prediction solution based on the TabNet deep learning network. In the process of data preparation, the sampling error is adjusted in combination with the actual business logic to ensure the consistency of the label distribution of the training set and the test set. The features are fully mined in the feature engineering part, and due to the large scale of data, we also optimize the feature processing efficiency and memory control. In the model part, the Lightgbm model was used in the initial scheme, which took a lot of time for feature engineering, and the score is around 85.4. After switching to TabNet, the effect of the model was significantly improved, and the coding of the login situation also played an important role. In the end, the solution we designed in this competition task finally won the second place with a score of 86.3473.

ACKNOWLEDGMENTS

We thank everyone associated with organizing and sponsoring the WSDM Cup 2022. Dataset was provided by IQIYI. Challenge was sponsored and managed by the 15th ACM International Conference on Web Search and Data Mining (WSDM 2022). Competition platform was hosted by IQIYI and DataFountain.

REFERENCES

[1] Sercan O. Arik, Tomas Pfister. TabNet: Attentive Interpretable Tabular Learning. 2019. https://arxiv.org/abs/1908.07442