Age Prediction of News Subscribers Using Machine Learning

Case Study of Hosting Worldwide Data Science Competition "Kaggle"

Shotaro Ishihara†
Business Information Services Unit, Digital Business
Nikkei Inc.
Tokyo Japan
shotaro.ishihara@nex.nikkei.com

Norihiko Sawa
Digital Strategy & Planning Unit, Digital Business
Nikkei Inc.
Tokyo Japan
norihiko.sawa@nex.nikkei.com

ABSTRACT

Attribute information such as age and gender is an important factor in managing news media. This kind of information can be used for providing the best possible experience for readers. Many news media encourage users to become their members and keep track of attribute information.

Nikkei is the world's largest financial newspaper, with a daily circulation around 3 million. There is a digital version service [1] available, and the subscription system with "Nikkei ID"[2] enable us to grasp attribute information of users.

Nikkei explores and utilizes attribute information of users for personalized article distribution, user segmentation for service growth etc. One of the troubles here is the existent of readers who don’t have Nikkei ID. The number of such users has been increasing recently, because the option has been started for users of charging with iOS and Android without creating Nikkei ID.

To address this problem, a system that predicts age from user behaviour using machine learning has been developed. More specifically, it uses a supervised learning framework to learn the relationship between the target and some features. Here, the target is age of users and features are extracted from access logs and article information.

The model has utilized the knowledge of data scientists from around the world through “Kaggle”[3], the world's largest data science community. Nikkei held a competition to compete for the performance of machine learning models at “Kaggle Days Tokyo”[4] on December 12, 2019. Nikkei provided datasets shown in Table 1 and the task to be solved. 149 individuals (88teams) joined the 8 hours competition. The winner had got prize, and Nikkei had got the solutions including source codes.

The performance of models was evaluated by root mean square error, and the score of the winner reached 11.10. Evaluated by the average error, it scored under 9. When considered as a five-level classification problem of "29 or less", "30-44", "45-59", "60-74", and "75 or more", the accuracy was about 50 percent. It's not a great number, but it has some value in distinguishing youngers from elderly.

The winner mainly used gradient boosting decision tree as a machine learning algorithm and created lots of features from the original features including text ones. It seemed that features created from texts were one of the key points for going up in the leaderboard.

While providing datasets in Kaggle has the advantage of using collective intelligence of many data scientists, there are some points to be considered. One of them is the problem of protecting personal information. In order to prevent individuals from being identified, the datasets were preprocessed in some way including anonymization. This paper presents the use case of machine learning in news media and the possibility of publishing datasets.

Table 1: The overview of datasets in Kaggle Days Tokyo

Feature	Description
Age	Target
Access Logs	What articles the user read
	When the user read the article
	Additional information like access place, OS, browser, device, referrer and so on
Article information	Text for headline and body
	Genre
	Tags attached automatically

KEYWORDS

age prediction, machine learning, supervised learning, Kaggle

ACM Reference format:

Shotaro Ishihara, Norihiko Sawa. 2019. Age Prediction of News Subscribers Using Machine Learning: Case Study of Hosting Worldwide Data Analysis Competition "Kaggle". In Computation + Journalism Symposium 2020. Boston, MA, USA.

REFERENCES

[1] Nikkei. https://www.nikkei.com/. Available on December 13, 2019.
[2] Nikkei ID. http://www.nikkei.co.jp/nikkeiinfo/en/news/announcements/471.html. Available on December 13, 2019.
[3] Kaggle. https://www.kaggle.com/. Available on December 13, 2019.
[4] Tokyo - Kaggle Days. https://kaggledays.com/tokyo/. Available on December 1 3, 2019.