

# Machine Learning based analysis of users' online behaviour

Gábor Kőrösi

Supervisor: Dr. Richárd Farkas

A THESIS SUBMITTED FOR THE DEGREE OF DOCTOR OF  
PHILOSOPHY  
OF THE UNIVERSITY OF SZEGED



University of Szeged  
Doctoral School of Computer Science

March 2022



# Introduction

Events and activities of daily life are increasingly often taking place in the online space, including, for example, the purchase of durable goods and education. Both of these areas, shopping and learning, which until a few years ago existed almost exclusively in the traditional offline format, have changed significantly. This change poses new challenges for professionals working in these fields, as most of the methods and methodologies used to date have become completely obsolete and unworkable in the online space. This is particularly true of the expertise of offline shop assistants or the role of teachers in brick-and-mortar educational facilities, roles which were once indispensable, but have now become outdated. The disappearance of these roles has not gone unnoticed, given that many online businesses are struggling with dwindling customer numbers and decreasing effectiveness of online learning systems (such as Massive Open Online Courses - MOOCs) with effectiveness at barely 25-30%. While it is undeniable that the online presence has created considerable challenges for business and education managers, it has also opened up new opportunities that can be exploited, notably by involving data science professionals. The various online platforms have a wealth of log data.

There are three levels to dive into:

- **High-level:** The simplest high-level of access to log data includes users' purchases, the provisional and final contents of the shopping cart, and in the case of educational platforms, interactions with videos, tutorials and quizzes.
- **Middle-level:** More in-depth than the previous category, the middle-level provides information on the time spent on the page and the order of the items involved within the page.
- **Low-level:** In addition, some log systems go deeper into hardware interactions, where mouse clicks and movements, keyboard press habits are stored, which is called the low-level information space.

With this data, one can create support systems and decision support systems that, in addition to aiding the operator, also improve the user experience. The topic of this

dissertation is the development of different Machine Learning methods for webshop and MOOC applications based on log data analysis.

What all applications have in common is the creation of aggregated databases, so-called user profiles, using log data of different widths and depths, which are used for classification, regression or even clustering. For more than fifteen years now there has been active research on the analysis of user log data. Initially, research and development were carried out in isolation on small databases in research teams or on closed internal databases in companies. In recent years, as online business and online educational interfaces have become more common, the number of real business applications and the amount and depth of data generated by each application have increased. Therefore, the previously traditional feature extraction and Machine Learning methods have been replaced by Deep Learning methods, which can provide high-quality solutions for large amounts of data, even starting from low-level data. Altogether the dissertation contains 7 chapters, composed of separate studies implementing the above-mentioned approaches. In the first two chapters, the author presents the special challenges of log data collection and preparation on high-level log databases. He describes forecasting results on a real-life Hungarian Webshop database, and the MOOC course ‘Conscious and safe Internet use’, which was developed and launched as a cooperation of two departments of the University of Szeged. Apart from the data collection and formulation of solutions, this dissertation also proposes application-specific feature sets. Through these training and evaluation databases, The author presents several comparative Machine Learning experiments. Based on the experience of the feature space design work, He focused his efforts on the possibility of building end-to-end systems using Deep Learning algorithms directly from low-level log data. The author subjected the data to minimal data processing and then successfully applied different neural network architectures. For the Deep Learning experiments, He used the log data from the Education-115-Spring-2014 MOOC course at Stanford University consisting of 39.5 million records. The experimental results achieved are primarily determined by the user profiles and data preprocessing techniques employed. The Deep Learning models outperform classical Machine Learning methods based on feature extraction in accuracy, but they are

black-box in nature, which hinders their real-life application (for instance, an instructor who does not trust the prediction of a black box). In the last chapter, He proposed three visualization methods for interpreting deep neural networks learned over sequential log data, which can contribute to human experts' better understanding of the patterns observed from the data.

## **The structure of the dissertation**

In the dissertation, the author was going to introduce online user behaviour modelling techniques and several empirical experiments. In Chapter 2, He summarize the application area and research challenges of webshops and MOOC sites through a literature review. Special attention has been paid to summarizing the different classical Machine Learning and Deep Learning techniques in the field, as well as to investigating data processing and feature extraction solutions.

Using log data from a webshop which is based in Hungary and operating within the borders of the EU, He created an actual business solution to build a reaching model for targeted offers and promotional mailings. In order to achieve this, He implemented and analysed the shopping habits and behavioural patterns of users (high-level log data). This process consisted of feature extraction methods, where new aggregated preprocessed data was created from the log data. Using a combination of so-called combined regression and classification models on a hand-crafted feature space, He carried out various prediction experiments that were successfully applied to real business operations, which are presented in detail in Chapter 3.

Chapter 3 describes the next step of this research. Namely, in collaboration with the Software Engineering Department of the University of Szeged, He developed a MOOC course with the title 'Conscious and safe Internet use' along with the full stack user interface. This system records the task completion (middle-level log data), mouse movements and video viewing log data (low-level log data). He managed a dataset of 'course completed' among students aged 10-21 which yielded nearly 4.5 million log records from 510 students. On the resulting, considerably richer middle- and low-level datasets, He proposed preprocessing and feature extraction methods,

and then investigated the effectiveness of traditional Machine Learning, as conducted for the previous, webshop research. The results highlighted the shortcomings of this dataset and the drawbacks of the implemented methods, but also proved the initial hypothesis that methods using richer and longer time series of middle and high-level log data provided higher efficiency.

In the final two parts of this dissertation, Chapters 4 and 5, the author presents student performance predictive models using 39.5 million log data of 142,395 students enrolled in a MOOC course at Stanford University. Instead of the difficult and time-consuming feature space extraction, He investigated raw, clickstream- level data and Deep Learning methods that could handle the raw sequences. In Chapter 4, He argues that given such a large training dataset, these methods are significantly more accurate than classical methods based on feature extraction. In Chapter 5, the author compares convolutional and recurrent neural network architectures on the Stanford MOOC dataset. Further, He provides insights into numerical and discrete sequential data processing techniques, where He investigated embeddings per variable. Lastly, He also proposed three visualization techniques for deep neural networks trained on sequential log data to help, to aid learning site owners in understanding the patterns learnt by the neural models.

## **User behavior analysis from high-level log data**

In the first thesis point, the author presented the special challenges of log data collection and preparation on high-level log databases. The author dealt with forecasting results on a real-life Hungarian webshop database, which was developed and launched as a collaboration of the University of Szeged and a Hungarian company (Kőrösi and Vinkó, 2021). Apart from data collection and preparation solutions, he proposed application-specific feature sets. Further, He presented several comparative Machine Learning experiments. The proposed problem and its solution to predict acceptance of the sales promotion were significant, since He did not predict the next purchase but, in fact, the buyer's reaction to advertising letters. His goal was to predict whether or not a given user was likely to act upon a received sales promotion, which was a

binary classification problem. The model achieved a considerable level of accuracy based on the first four purchases and high-level sequences data.

Related to his publications (Kőrösi and Vinkó, 2021) author regards the following results as his main contributions to the field:

- The author used the log data of an existing webshop in Hungary to develop a solution that could reliably predict the sales promotion acceptance probability from the high-level user log data. He proposed a specific feature representation that contained cumulative data from the obtained user sequences that effectively supported the operation of the model, which was designed as a combined classification and regression solution.
- In the combined model the classification method aimed to determine whether a user would accept the sales promotion or not. Using the output of this classification model with a regression task, He separately predicted the probability of the promotional package to be accepted. The output of this combined model was not only able to predict user behavior with considerable efficiency, but also to provide a solution that was easy for the client to interpret.
- The author performed empirical measurements with almost a dozen different Machine Learning methods, and ran hyper-parameter tuning to find the optimal solution. He demonstrated that when using high-level log data, the cumulative feature extraction method with a combined classification and regression solution could provide fast and effective results, which was confirmed by the customer's satisfaction.

## **Educational performance prediction from middle-level click-stream data**

The time series structure and analysis of e-commerce and MOOC platforms are very similar. However, during the analysis of the high-level webshop log data used in Chapter 2, the author recognized that the short and high-level log data limited the

accuracy of his model. This meant that, in order to build models with higher accuracy, He first needed to change the depth of the data. To obtain better and deeper data, He designed an e-learning course on a Moodle site (Conscious and Safe Internet Usage - Tudatos és biztonságos internethasználat alapjai TÉBIA) and collaboration of two departments of the University of Szeged developed a middle-level user behavior logging component to collect the users' online activities (Kőrösi and Havasi, 2017; Kőrösi et al., 2018). In that study He analyzed the log data of pupils and students who were motivated by their teachers and schools to attend and complete the short (few-day-long) MOOC course. The main contribution of his investigation was that He managed to confirm that deeper middle-level data can support a more accurate model even when using short MOOC courses. The author also highlighted the features which influenced the classifier results the most, hence providing useful insights for MOOC developers. Based on the implemented methods, He described which the most notable features in his prediction models were. The chapter offers a detailed statistical methodology for predicting student performance based on log data which was created in short MOOCs and led by the teacher. Based on these datasets, the author determined that classic Machine Learning models were successful and they were influenced by several strong features. The accuracy of the models achieved a satisfactory accuracy of more than 80%.

Related to his publications (Kőrösi and Havasi, 2017; Kőrösi et al., 2018) the author regards the following result as his main contributions to the field:

- He was able to add a functioning logging system to a Moodle platform with weak tools of analyses, which would be useful for similar portals to live up to current measuring requirements.
- He designed a data engineering solution that would automatically process input data without human intervention and which could intervene if extreme values emerged. He defined 263 features to describe the middle-level clickstream sequence of short video MOOCs.
- Despite a relatively low sample size, He was able to render clickstream based predictive algorithms. He introduced a Machine Learning methodology for feature



selection and binary classification techniques with leave-one-out cross validation for short video MOOCs based on middle-level sequence data. The primary goal was to make binary prediction of course completion. The created models were capable of predicting who would “Fail” or “Complete” an online course, which would be an immense help for the faculties that provide e-learning courses.

- He implemented and tested more than ten Machine Learning approaches, the most efficient tools were the Random Forest and Bagging achieving approximately 80% accuracy

## **MOOC performance prediction by Deep Learning from raw clickstream data**

In terms of low-level log data collection in the form of clickstream or social network measures, the MOOC systems offer a treasure trove of data. We can design efficient online user (student) models which would then serve as a forecasting tool for estimating how many students were likely to drop out, or preferably, complete the course. This was made possible by extensive research into comprehending and, hopefully, increasing the registration and completion rate, ultimately contributing to a better all-round learning experience in MOOCs.

Several studies showed that low-level data can be used to create more successful prediction models. To demonstrate this, conducted experiments using data from Stanford Lagunita’s datasets to predict learner behavior using Deep Learning models on low-level data, and the author comparatively evaluated traditional and Deep Learning models.

To better understand the obtained results, the author performed a recurrent Neural Network for solving the outcome performance prediction problem in an online learning platform. The main contribution of that research part was the building of a prediction model which could use raw low-level datasets, and obtain the same or better results than regular prediction models. The key advantage of the model was that there was no need for manual feature engineering, because it could be automatically

extracted from the raw log-line level records. Therefore, this approach could save a lot of time and human effort, and ignore the possible inconsistencies introduced by the hand-made process. Experimental results on Stanford Lagunita’s dataset consisting of data by 12015 students showed that the expected model achieved significantly better than the baseline models. The results for the model were sufficient to demonstrate the feasibility of using recurrent Neural Networks when large datasets were available. Related to his publication (Kőrösi and Farkas, 2020) the author regards the following results as his main contributions to the field:

- He proposed a data preparation step for low-level clickstream event data. On the 3D tensor, He was able to effectively run RNN experiments.
- He built a baseline pipeline and Deep Learning model to predict student outcome as a regression and multi class classification problem. He evaluated the Deep Learning based prediction pipeline which outperformed the classic Machine Learning based solution.

## Deep learning models and interpretations for MOOC performance prediction

The majority of the time series Deep Learning models were applied to numerical data, event logs, such as the clickstream-level MOOC data used, consisting of multivariate discrete-valued sequences. Hence, time series Deep Learning techniques could not be directly applied. However, most of the discrete-valued sequence prediction solutions have been published for Natural Language Processing. The raw event logs were significantly longer than natural language sentences, with their varying lengths, thus NLP techniques could not be applied directly. To handle these special characteristics of the given clickstream-level MOOC dataset, He proposed an embedding based Deep Learning model architecture (Kőrösi and Farkas, 2021). In this part of the research He trained state-of-art RNN and CNN models to predict the outcome scores of the students at the MOOC. He conducted experiments using various embedding layers

to represent the multivariate discrete-valued data. Recurrent and Temporal Convolutional Neural Networks provided accurate forecasts without having any access to explicit knowledge about the investigated system. Yet, Deep Learning methods are typically considered ‘black boxes’, where it is almost impossible to fully understand based on what, why, and how RNN and CNN make forecasting decisions. His research aimed to open the black boxes of RNNs and CNNs trained for time series regression. The offered three visualization techniques which could support domain-expert users in interpreting discrete-valued multivariate time series regression neural models. Moreover, He analyzed the online behavior of Massive Online Open Course (MOOC) students and introduced a Deep Learning architecture to predict the outcome score of the students at a MOOC.

Related to his publication (Kőrösi and Farkas, 2021), the author regards the following results as his main contributions to the field:

- Empirical results showed that the embedding method was able to significantly improve his forecasting results and provide an effective aid in the preprocessing of discrete-valued sequence.
- He also comparatively evaluated GRU, LSTM and TCNN architectures along with classic Machine Learning on cumulative features.
- To better understand the results, the author performed three visual inspections of the deep learnt models. His investigation clearly showed the different learning methods between RNN and CNN models, and offered useful visualization for pedagogical analysis.



# Bibliography

- Gábor Kőrösi, Péter Esztelecki, Richard Farkas and Krisztina Tóth. 2018. Clickstream-based outcome prediction in short video moocs. In *2018 International Conference on Computer, Information and Telecommunication Systems (CITS)*, pages 1–5. IEEE.
- Gábor Kőrösi and Richard Farkas. 2020. Mooc performance prediction by deep learning from raw clickstream data. In *International Conference on Advances in Computing and Data Sciences*, pages 474–485. Springer.
- Gábor Kőrösi and Richárd Farkas. 2021. Deep learning models and interpretations for multivariate discrete-valued event sequence prediction. In *International Conference on Artificial Neural Networks*, pages 396–406. Springer.
- Gábor Kőrösi and Ferenc Havasi. 2017. Moodle-based data mining potentials of mooc systems at the university of szeged. In *2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 755–760. IEEE.
- Gábor Kőrösi and Tamás Vinkó. 2021. A practical framework for real life webshop sales promotion targeting. *Informatika*, 45(4).

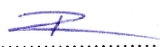
## Társszerzői Nyilatkozat

Alulírott Dr. Farkas Richárd József hozzájárulok ahhoz, hogy Körösi Gábor az alább felsorolt közleményekben bemutatott eredményeket a Szegedi Tudományegyetem Természettudományi és Informatikai Karának, Informatikai Intézetének Doktori Iskolához tartozóan benyújtott „Machine Learning based analysis of users’ online behaviour” című PhD értekezés tézispontjainak alátámasztására önálló eredményként felhasználhatja.

Egyúttal kijelentem, hogy a közös publikációkban és a tézisekben foglalt tudományos eredményeket nem kívánom a Szegedi Tudományegyetem vagy más egyetem doktori iskolájában fokozatszerzés céljából felhasználni.

- Gábor Körösi, Péter Esztelecki, Richard Farkas and Krisztina Tóth. 2018. Clickstream-based outcome prediction in short video moocs. In 2018 International Conference on Computer, Information and Telecommunication Systems (CITS), pages 1–5. IEEE.
- Gábor Körösi and Richard Farkas. 2020. Mooc performance prediction by deep learning from raw clickstream data. In International Conference on Advances in Computing and Data Sciences, pages 474–485. Springer.
- Gábor Körösi and Richárd Farkas. 2021. Deep learning models and interpretations for multivariate discrete-valued event sequence prediction. In International Conference on Artificial Neural Networks, pages 396–406. Springer.

Dátum: 2022. 02. 17.

  
.....  
Dr. Farkas Richárd József

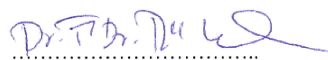
## Társszerzői Nyilatkozat

Alulírott Dr. Farkasné Dr. Tóth Krisztina hozzájárulok ahhoz, hogy Kőrösi Gábor az alább felsorolt közleményekben bemutatott eredményeket a Szegedi Tudományegyetem Természettudományi és Informatikai Karának, Informatikai Intézetének Doktori Iskolához tartozóan benyújtott „Machine Learning based analysis of users’ online behaviour” című PhD értekezés tézispontjainak alátámasztására önálló eredményként felhasználhatja.

Egyúttal kijelentem, hogy a közös publikációkban és a tézisekben foglalt tudományos eredményeket nem kívánom a Szegedi Tudományegyetem vagy más egyetem doktori iskolájában fokozatszerzés céljából felhasználni.

- Gábor Kőrösi, Péter Esztelecki, Richard Farkas and Krisztina Tóth. 2018. Clickstream-based outcome prediction in short video moocs. In 2018 International Conference on Computer, Information and Telecommunication Systems (CITS), pages 1–5. IEEE.

Dátum: 2022. 02. 17.



Dr. Farkasné Dr. Tóth Krisztina

## Társszerzői Nyilatkozat

Alulírott Eszteleki Péter hozzájárulok ahhoz, hogy Körösi Gábor az alább felsorolt közleményekben bemutatott eredményeket a Szegedi Tudományegyetem Természettudományi és Informatikai Karának, Informatikai Intézetének Doktori Iskolához tartozóan benyújtott „Machine Learning based analysis of users’ online behaviour” című PhD értekezés tézispontjainak alátámasztására önálló eredményként felhasználhatja.

Egyúttal kijelentem, hogy a közös publikációkban és a tézisekben foglalt tudományos eredményeket nem kívánom a Szegedi Tudományegyetem vagy más egyetem doktori iskolájában fokozatszerzés céljából felhasználni.

- Gábor Körösi, Péter Eszteleki, Richard Farkas and Krisztina Tóth. 2018. Clickstream-based outcome prediction in short video moocs. In 2018 International Conference on Computer, Information and Telecommunication Systems (CITS), pages 1–5. IEEE.

Dátum: 2022.02.20.



(aláírás)



## Társszerzői Nyilatkozat

Alulírott Havasi Ferenc hozzájárulok ahhoz, hogy Kőrösi Gábor az alább felsorolt közleményekben bemutatott eredményeket a Szegedi Tudományegyetem Természettudományi és Informatikai Karának, Informatikai Intézetének Doktori Iskolához tartozóan benyújtott „Machine Learning based analysis of users’ online behaviour” című PhD értekezés tézispontjainak alátámasztására önálló eredményként felhasználhatja.

Egyúttal kijelentem, hogy a közös publikációkban és a tézisekben foglalt tudományos eredményeket nem kívánom a Szegedi Tudományegyetem vagy más egyetem doktori iskolájában fokozatszerzés céljából felhasználni.

- G Kőrösi and F Havasi. 2017. Moodle-based data mining potentials of mooc systems at the university of szeged. In 2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pages 755–760. IEEE.

Dátum: Szeged, 2022. 02.17.

.....Havasi Ferenc.....

Havasi Ferenc