Clickstream-based outcome prediction in short video MOOCs

Gábor Kőrösi
Institute of Informatics
University of Szeged
korosig@inf.u-szeged.hu

Péter Esztelecki
Institute of Informatics
University of Szeged
epeter@inf.u-szeged.hu

Richard Farkas
HAS Research Group on Artificial Intelligence,
University of Szeged
rfarkas@inf.u-szeged.hu

Krisztina Tóth
TBA21 Ltd.
ktoth@tba21.hu

***Abstract*— In this paper, we present a data mining approach for analysing students' clickstream data logged by an e-learning platform and we propose a machine learning procedure to predict course completion of students. For this, we used data from a short MOOC course which was motivated by the teachers of elementary schools. We show that machine learning approaches can accurately predict the course outcome based on clickstream data and also highlight patterns in data which provide useful insights to MOOC developers.**

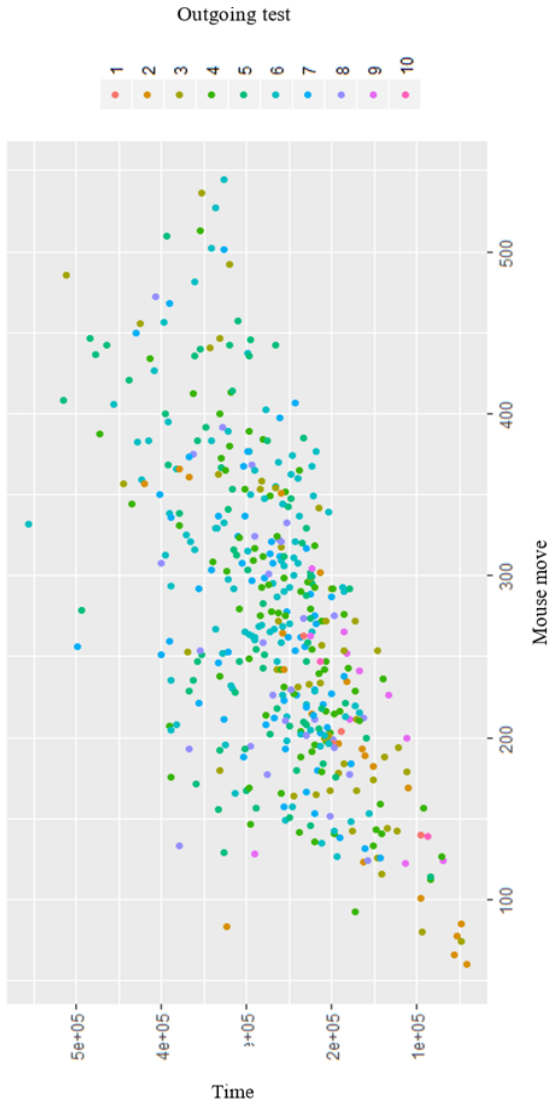
Keywords*— *data mining, clickstream, edm, predicting

I. INTRODUCTION

Over the past decade, technology and the internet have had a huge impact on the ways we learn. One of the most used parts of these innovations is the Massive Open Online Course (MOOC). Despite their early promise, however, MOOCs are still relatively unexplored and poorly understood [2]. Meanwhile, MOOCs often attract an enormous number of registrants, but only a small fraction of them can successfully complete their courses. Even of those students who declare at the start of a course an intent to complete it, 75% do not do so[6]. These high drop-off rates are often attributed to factors such as low teacher-to-student ratios, the asynchronous nature of interaction, and heterogeneous educational backgrounds and motivations, which make it difficult to scale the efficacy of traditional teaching methods with the size of the student body [4] [14]. Several researchers have analysed the server logs associated with these MOOCs to determine the factors associated with students dropping out, such as [5] [14] and conclude that student dropout rates are a major deterrent to the growth and success of MOOCs [6] [7]. [5][8]. As more and more higher education institutions make their courses available for learners through platforms such as MOOCs, the immense amount of data generated make it possible to provide continuous and automated assessment of student progress [9].

Analysing MOOC server log data in order to identify student drop-out patterns is an Educational Data Mining (EDM) task. Data is recorded during the time when learners are interacting with the MOOC platform providing a unique opportunity to learn about the efficacy of different resources, build predictive models that can help develop interventions and propose/recommend strategies for the learner [11]. By detecting whether student behaviour changes in a significant manner over the time-period of a particular term, we could identify students who increase, decrease, or show no changes in their clickstream activities, and whether these changes relate to course performance [3]. Student clickstream data has been the subject of a number of prior studies, such as the investigation of potential predictive relationships between online student activity and student outcomes, such as course grades.

Fig. 1. Student achievement by behavior



There are two main MOOCs which have been investigated. The first group uses a huge log file from edX or Coursera [5,6,10,12,14] which has been generated by big Universities, such as Stanford or MIT. This data has been generated by ten or more thousand self-motivated students. On the other side, there

is a second group which wants to create a successful prediction model from school-class-related e-learning platforms [1,2,37,,12,,,15] based on very shallow data from Moodle or Moodle-based forums. In our research, we analyse the log-data of students who were motivated by their teacher and their school to attend and complete the short (few- day-long) MOOC course. Our work has got similarities with both research avenues. Our logged data is very similar - wide and deep – to the data from edX and Coursera, even though it had been created in a much shorter period than that and is of a school-class nature. This log file gave us an opportunity to study clickstream data and user attitudes in short MOOC's. In this study, we present classification models that utilize data about the activities of students in courses to predict their final exam outcome. We propose a feature space of 263 attributes to describe students' clickstream data. Then, we apply various feature selection and various classification approaches.

The main contributions of our investigation are that our data mining procedure is able to accurately predict the success of students even if using short MOOC courses, and we highlight features which influence the classifier results the most, hence providing useful insights for MOOC developers.

II. DATASET

In this paper, we focus on clickstream data from a course which was recorded by an E-learning platform in the 2016-2017 academic year. In the course of recording, clickstream data was obtained through our course management system in the form of student IDs, time stamps, and activities.

The samples of data are constructed form a course named TÉBIA which involved upper grade pupils form 20 elementary schools. Components of a previously used and tested learning material were taken as the basis of the course content, which included an initial test with a video lesson and 3 further units. Every unit consisted of an obligatory video task and further optional textual learning material. To complete a unit, students had to solve 3 tests with a minimum score of 5 points out of 10. Every unit ended in a test with a maximum score of 10 points, except for the initial test. The structure of the learning material is demonstrated in Table I.

TABLE I. COURSE CONTENTS

Course name	TÉBIA
Content	Basics of Conscious and Safe Internet Usage
Time frame	6 weeks
Parts of the Learning Material	Introduction:Video (3.37 min., Embed);
	Digital footprint: Video (14.04 min, Embed);HTML embedded text;
	Conscious and Safe Internet Usage: Video (13.07 min, Embed); HTML embedded text; External link;
	Online bullying: Video(13.31 min, Embed); HTML embedded text;;Extra video (11.55 min, Embed);

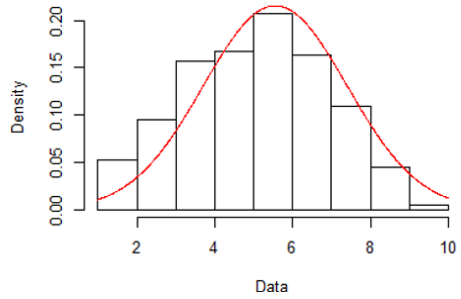


Fig. 1. Distribution of students' final scores (n=603)

The distribution of students' final scores shows a Gaussian distribution (Fig. 2) which supports the validity of the outcome test.

A. Data cleaning

The types of activities recorded are those which correspond to broad categories of student behaviour, such as previewing lectures, mouse behaviours (move, scroll, click), video watching attitudes and text inputs. For instance, the course we examine in this paper had 1370 registered students who generated 2.430.975 click events over a 6-week period. The portal recorded 1370 students and lecturers, out of which only 1077 filled in and completed the initial test (Q0). As Fig 1 shows, the noisy and complex nature of this set of data made it impossible to use simple statistical or clustering methods to create a predictive model. Those students who had an output test but had insufficient amount of activities were eliminated from the measurement. The number of obtained results amounted to 603. According to conditions set to complete the course, we split the group (Q1>=5 and Q2>=5 and, Q3>=5) into two parts, which were labeled as 0 ("Failed") and 1 ("Completed").

B. Preliminary investigation

To investigate the structure of the data and understand user behaviour, we visualized the class-labelwise distributions of several log properties.. Because of the unbalanced nature of the data (n "Failed" = 419, n "Completed" = 184) we present density distribution. The following density figures show the differences between failed and completed students' main attributes. The final diagrams show a significant overlap between the two groups, which makes it harder to adjust the weighting settings. Without striving to present an extensive number of differences between the two groups, we show some of them in the following tables and figures. (Fig. 3, Fig. 4, Fig. 5, Table II, Table III, Table IV) All the following diagrams were constructed using the Ggpolt R package of Wickham, et al.[16]

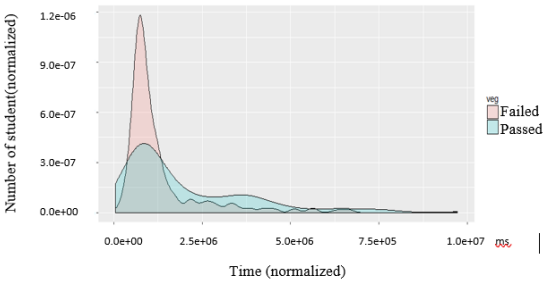


Fig. 1. Time spent in course

TABLE II. TIME SPENT IN COURSE

„Failed”

Min.	1st Qu.	Median	Mean	3rdQu.	Max.	NA's
51000	684300	861900	1294000	1291000	9725000	3

„ Completed

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
123500	750400	1098000	1987000	2970000	9462000	3

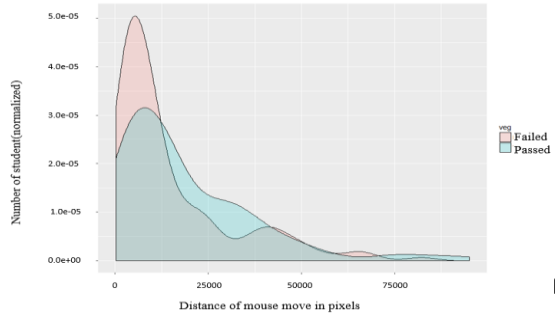


Fig. 4. Average distance of mouse in course contents

TABLE III. AVERAGE DISTANCE OF MOUSE IN COURSE CONTENTS

„Failed”

Min.	1st Qu.	Median	Mean	3rdQu.	Max.	NA's
226	4325	8190	14800	20570	82210	262

„ Completed”

Min.	1st Qu.	Median	Mean	3rdQu.	Max.	NA's
451	5632	12990	19040	28520	95030	76

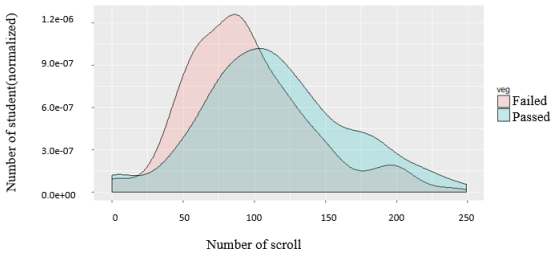


Fig. 5. Average number of scrolls in course contents

TABLE IV. AVERAGE NUMBER OF SRCOLLS IN COURSE CONTENTS

„Failed”

Min.	1st Qu.	Median	Mean	3rdQu.	Max.	NA's
0	64	89.50	95.58	118.80	244.00	9

„ Completed”

Min.	1st Qu.	Median	Mean	3rdQu.	Max.	NA's
0	81	109	115	144	249	9

III. MACHINE LEARNING EXPERIMENTS

We carried out machine learning experiments using clickstream log-data to predict whether a particular student will fail or succeed in the final exam of the MOOC. We employed the Rminer [17] package of R.

A. Feature space

We defined 263 features to describe our clickstream data. There were two types of data. In the first group, there was the data which was collected during the filling process in the incoming test. The second type was the clickstream which was collected during the learning process in the three parts of the curriculum (see Table 1). This collection was divided into 18 main categories: binary and numeral answers provided to input and output tests (28+60), time spent on the quizzes (6) and the sites of the curriculum (7), the number of visits to the site of quizzes (6) and site of curriculum (7), the mouse move distance in pixels (13), the average mouse speed (13), the cumulated data (6), the number of mouse movement on a page (13), the number of clicks on a page (13), the use of test buttons during the input/output testing (4), the number of scrolls on a page (13), the last login date to a page compared to the first login to the site (13), the first login date compared to the first login to the site (13), the days spent on the sites (13), the mean behaviour on the sites (19), the number of calendar days between the output tests (7), binary output results (1), output results (2), user-related data (6).B. Feature selection

We investigated different feature selection methods, and the gain-ration function in the FSelector package [15] proved to be the most effective.

Gain-ratio examines all the parameters one-by-one and creates a hierarchy, which distinguishes weak and strong correlational connections. The FSelector package was designed to handle such problems and the most useful functions of all were the chi.square and gain.ratio filtering algorithms. Between the two, the latter provided accurate calculations so the choice to present the underlying theory.

The information gain method chooses a split based on which attribute provides the greatest information gain. The gain is measured in bits. Although this method provides satisfactory results, it favours splitting on variables that have many attributes. The information gain ratio method incorporates the value of a split to determine what proportion of the information gain is valuable for that split. The split with the greatest information gain ratio is chosen. [13] The information gain calculation starts by determining the information of the training data. The information in a response value, r, is calculated in the following expression:

$$-log_2\left(\frac{freq(r,T)}{|T|}\right)$$

T represents the training data and $|T|$ is the number of observations. To determine the expected information of the training data, sum this expression for every possible response value:

$$I(T) = -\sum_{i=1}^{n} \frac{freq(r_i,T)}{|T|} \times log_2\left(\frac{freq(r_i,T}{|T|}\right)$$

Here, n is the total number of response values. This value is also referred to as the *entropy* of the training data.

Next, consider a split S on a variable X with m possible attributes. The expected information provided by that split is calculated by the following equation:

$$I_s(T) = \sum_{j=1}^{m} \frac{|T_j|}{|T|} \times I(T_j)$$

In this equation, T_j represents the observations that contain the j^{th} attribute.

The information gain of split S is calculated by the following equation:

$$G(S) = I(S) - I_s(T)$$

Information gain ratio attempts to correct the information gain calculation by introducing a split information value. The split information is calculated by the following equation:

$$SI(S) = -\sum_{j=1}^{m} \frac{|T_j|}{|T|} \times log_2\left(\frac{|T_j|}{|T|}\right)$$

As its name suggests, the information gain ratio is the ratio of the information gain to the split information:

$$GR(S) = \frac{G(S)}{SI(S)}$$

B. *Prediction Models*

Classifying whether the student failed or completed the course was the core goal of this study. We train various machine learning models for prediction. Because of the limited size of our dataset, we applied the LEAVE-ONE-OUT cross validation method.

We comparatively experimented with the following classifiers: "lr"- logistic regression, "xgboost" - eXtreme Gradient Boosting ,"mlpe"- multilayer perceptron ensemble, "mlp"- multilayer perceptron with one hidden layer ,"ksvm" - support vector machine, ,"kknn"- k-nearest neighbor,"naiveBayes"- naive bayes,"naive","ctree"- conditional inference tree, "rpart"- – decision tree,"randomForest"- random forest algorithm,"boosting"- boosting ,"bagging"- bagging.

IV. Experimental Results

During the data cleaning process, we reduced the number of students from 1370 to 603. The preliminary results showed that every student-user had a unique click stream pattern, which was very similar and independent of user achievement and final scores. Such a finding underpinned that data saved by the MOOC system is suitable to build prediction models. It will be possible to help educational institutions to fight to lower the drop-out rate. They could also take action to help users whose achievement results fall below the average to prevent negative outcomes.

We carried out binary classification experiments to predict whether a student will successfully complete the MOOC and get the certificate. There were 429 students out of 603 who successfully completed the course, i.e. the most frequent class baseline is 71%.

The gain-ratio feature selection ranked features in an ascending order and the experiment showed that approximately the top 60 features are useful. Results were completely tested in 60 cases and by halving further 30, 15. The following table summarizes accuracies achieved by the 12 classifiers using the top 60 features. Besides accuracy (ACC), we also report the recall, precision, and F-score values of the Completed class.

Figure 6 provides an overview of the classifiers using only the top 15, 30 and 60 features. We can conclude that along the 30 properties, the most accurate results were achieved by the supported vector machine and the random forest function. While in the case of 60 features, the most accurate was the bagging function. In the end, we could say that the most successful model were the bagging (ACC 80.10%) and the random forest (ACC 79.44%) methods (Table V., Fig. 6.).

TABLE V. Performance of Certificate Earner Prediction with Different Methods (%), the most weighted 60 feature

	Bagging	Boosting	Ctree	Kknn	Ksvm	Lr	Mlpe	Random Forest
ACC	80.1	78.11	64.34	71.14	77.61	78.44	73.47	79.44
RECALL	91.14	87.88	79.25	76.92	94.87	87.65	82.05	92.07
PRECISION	82.66	82.49	72.96	81.48	78.27	83	80.92	81.44
F1	86.7	85.1	75.98	79.14	85.77	85.26	81.48	86.43

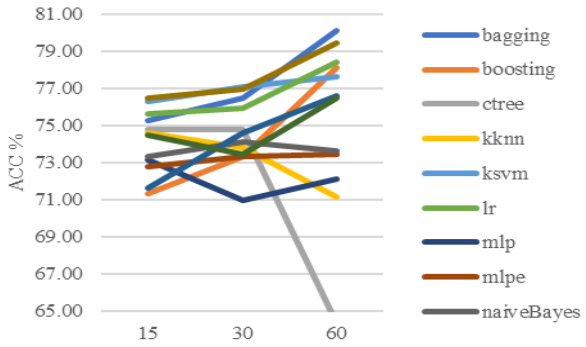


Fig. 6. Average Prediction performance in the function of the number of features for training

V. Discussion

This paper describes a statistical methodology for predicting binary outcome in a set of data which was created in short MOOC and driven by the teacher. Based on these data sets, we found a successful model which was influenced by a couple of strong features. The accuracy of the models has achieved satisfactory accuracy of more than 80%. It confirms the supposition that we are able to efficiently predict learning outcomes. Through more detailed research, the two models show significant differences. The best results were achieved by those features which were connected to the learning material or the average value of cursor distance on a curriculum page. Based on our methods, we could describe which were the most notable features in our prediction models. As we can see in Fig 7, the most highly weighted features in feature selection process over data stream. As we expected, the highest weight got the input test grades, after which followed the average time, mouse speed and mouse distance spent in the whole course. The other important things were the number of clicks, and scrolls, and the number of mouse moves on the page of the curriculum. At the beginning, we expected the amount of time would most influence the outcome because those who spend more time on the system, would learn more. In the end, we realized that taking more time in the course does not have a considerable effect on the outcome of grades. On the other hand, as in ordinary

schools, the number of days spent learning and testing has shown its effect during the evaluation process.

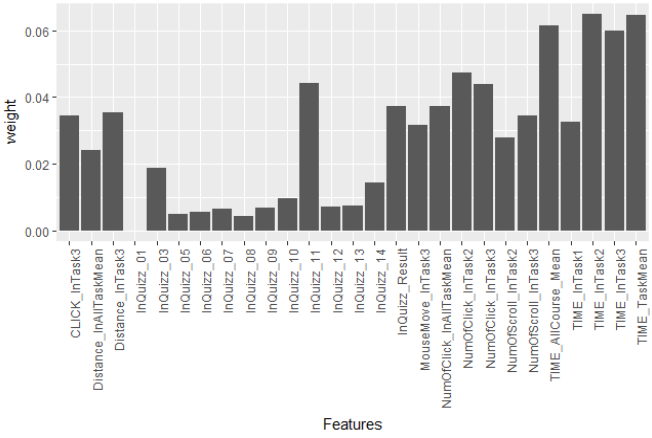


Fig. 7. The highest weighted features

VI. CONCLUSION

Student clickstream data is inherently difficult to work with given its complex and noisy nature [3]. Several data mining applications are focused on educators, where the object is to help create accurate feedback, categorization of learners based on their abilities, course creation, and instructional plans.

This paper introduced a machine learning methodology for outcome classification of short video MOOCs based on clickstream data. Our primary goal was to do binary prediction of course completion and of student engagement. Our models could predict who would "Fail" or "Complete" an online course, which would be an immense help for the faculties that provide e-learning courses. Despite a relatively low sample size, we could still render click stream based predictive algorithms. We proposed 263 features to describe clickstream data of short video MOOCs. We employed feature selection and binary classification techniques in a leave-one-out cross validation evaluation setting. The most efficient tools for our models were the Random Forest and Bagging achieving with approximately 80% accuracy.

While the results in this paper are promising and there are interesting methodological avenues to pursue, the most important future direction from an education research perspective will involve more in-depth investigation of the utility of these types of methods in terms of providing actionable insights that are relevant to the practice of education.

ACKNOWLEDGEMENT

The project was supported by the European Union, co-financed by the European Social Fund (EFOP-3.6.3-VEKOP-16-2017-00002).

REFERENCES

[1] C. Robinson, M. Yeomans, J. Reich, C. Hulleman, and H. Gehlbach, "Forecasting Student Achievement in MOOCs with Natural Language Processing", *Sixth International Conference on Learning Analytics & Knowledge*, Universitz of Edinburhg, Edinburgh, pp. 383–387, 2016.

[2] Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec, "Engaging with Massive Online Courses", *23rd international conference on World wide web*, Seoul, Korea, pp. 687–698, April 7–11, 2014.

[3] J. Park, K. Denaro, F. Rodriguez, P. Smyth, M. Warschauer, "Detecting Changes in Student Behavior from Clickstream Data", *Seventh International Conference on Learning Analytics & Knowledge*, Vancouver, BC, Canada pp. 21–30, 2017.

[4] X. Wang, D. Yang, M. Wen, K. R. Koedinger, and C. P. Rose, "Investigating how student's cognitive behavior in MOOC discussion forum a ect learning gains". *In Proceedings of the EDM Conference*,International Educational Data Mining Society (IEDMS), pp. 226–233, 2015.

[5] Z. Ren, H. Rangwala, and A. Johri, "Predicting Performance on MOOC Assessments using Multi-Regression Models", *Computers and Society*, 2016.

[6] H. Daumé, D. Goldwasser, L. Getoor, B. Huang, and A. Ramesh, "Modeling Learner Engagement in MOOCs using Probabilistic Soft Logic", 2013.

[7] U. Anderson, T. Arvemo, and M. Gellerstedt, "Can Measurements of Online Behavior Predict Course Performance?", *7th International Multi-Conference on Complexity, Informatics and Cybernetics: IMCIC 2016 and the 7th International Multi-Conference onSociety and Information Technologies: ICSIT 2016*: Volume II (Post-Conference Edition), pp. 4–9, 2016.

[8] S. Tang, J. C. Peterson, and Z. A. Pardos, "Modelling Student Behavior using Granular Large Scale Action Data from a MOOC", 2016.

[9] M. M. Ashenafi, G. Riccardi, M. Ronchetti, "Predicting Students' Final Exam Scores from their Course Activities*", Frontiers in Education Conference (FIE),* 2015 pp. 10–22, 2015.

[10] C. G. Brinton, S. Buccapatnam, M. Chiang, and H. V. Poor, "Mining MOOC Clickstreams: On the Relationship Between Learner Behavior and Performance", Cornel University, 2015.

[11] S. Boyer, and K. Veeramachaneni, "Transfer Learning for Predictive Models in Massive Open Online Courses" , In: Conati C., Heffernan N., Mitrovic A., Verdejo M. (eds) *Artificial Intelligence in Education. AIED 2015.* Lecture Notes in Computer Science, vol. 9112. Springer, Cham, 2015.

[12] Ch. G. Brinton, and M. Chiang, "MOOC performance prediction via clickstream data and social learning networks", *Computer Communications (INFOCOM)*, 2015 IEEE Conference , 2015.

[13] Inc. 2015. SAS® Visual Analytics 7.2: User's Guide. Cary, NC: SAS Institute Inc.

[14] Y. Tsung-Yen, Ch. G. Brinton, C. Joe-Wong, "Behavior-Based Grade Prediction for MOOCs via Time Series Neural Networks", IEEE Journal of Selected Topics in Signal Processing ,vol. 11, Issue: 5, Aug. 2017

[15] P. Romanski, L. Kotthoff, Package 'FSelector', 2016 , related:https://CRAN.R-project.org/package=Fselector

[16] H. Wickham, W. Chang, "Create Elegant Data Visualisations Using the Grammar of Graphics" Package 'GGplot', 2016 , related: https://CRAN.R-project.org/package=ggplot2

[17] P. Cortez, "Data Mining Classification and Regression Methods", Rminer package, 2016 , related: https://CRAN.R-project.org/package=rminer