https://doi.org/10.17234/INFUTURE.2019.14 Original scientific paper

Using LMS Activity Logs to Predict Student Failure with Random Forest Algorithm

Dejan Ljubobratović
Department of Informatics, University of Rijeka, Croatia
dejan.ljubobratovic@student.uniri.hr

Maja Matetić
Department of Informatics, University of Rijeka, Croatia
majam@uniri.hr

Summary

The paper presents a Random forest model in the task of predicting student success (grade) on the base of input predictors (lectures, quizzes, labs and videos) extracted from Moodle activity logs. Since 2010. University of Rijeka is using Moodle based Learning Management Systems (LMS) to complement traditional teaching. LMS is used for documents sharing, quizzes, assessments, video lecturing, tracking student progress and much more. When student access an LMS using his personal account, a digital profile is created that is saved in LMS log files. These logs were used to create a dataset with couple of hundreds of observations. However, building a prediction model using Random forest algorithm is relatively easy comparing to explaining the results. Interpreting Random forest and other machine learning black box models is a challenge regarding to complexity of their decision-making mechanisms. There are a number of new techniques allowing us to interpret such models, and couple of them is used in this paper for that purpose.

Another problem a researcher is facing using black box algorithms is GDPR. General Data Protection Regulation has a significant impact on many aspects of EU citizen's data collection and processing. This paper will highlight most challenging GDPR restrictions on data mining including GDPR's "right to explanation".

Key words: LMS system, random forest algorithm, educational data mining, predicting student success, interpretability, interpretable machine learning

Introduction

Mining and interpreting data collected in Massive Online Open Courses (MOOCs) is well researched and popular, giving a researcher huge database of different logs to deal with. For example, only one course "Learning How to Learn: Powerful mental tools to help you master tough subjects" offered by McMaster University at University of California San Diego enrolled more than 1.7 million people using Coursera platform.(Learning How to Learn, 2019)

LMS systems like Moodle are used to complement educational processes in universities and schools, with significantly smaller log database.

In this research we build a model to predict student success (grade) as a function of course activities using Random forest algorithm. Later in this work several methods were used to interpret the given model giving explanations to Random forest algorithm results. For data exploration, prediction model and result explanation in this work R language v. 3.6.1. is used, which is a freely available language and environment for statistical computing and graphics.

This work is divided in three logical parts. In the beginning are presented the basic ideas found in similar researches in order to compare them with our work, after which we highlighted the most challenging GDPR restrictions on data mining. Main part of this research is building a prediction model using Random forest algorithm, and explaining data used in the process. Interpreting results of our Random forest model using four different techniques is the main goal of the third part of this paper.

Related work

Educational data mining

Creating a precise model that can predict student future behaviour or student's final grade based on his activity is very appealing to any educational institution.

In order to classify the dropout student Yukselturk et al. (2014) used four data mining algorithms; k-Nearest neighbour, Decision tree, Naive Bayes and Neural networks. In their final results as the most important factors in predicting the dropouts were three variables; *online technologies self-efficacy*, *online learning readiness*, and *previous online experience* (Yukselturk, Ozekes, Türel, 2014).

In another conducted research, authors examined students' activity by gender, and by log time using LMS Moodle activity logs. They found there significant correlation; the female students were more active and successful in the course than are the male ones and the students were most active in the test weeks, specifically, on the day before the tests (Kadoic, Oreski, 2018).

Mishra et al. (2014) build performance prediction model based on students' social integration, academic integration, and various emotional skills. The key influencers to the *semester results* were *previous semester results,* followed by *good academic performance*. Out of all emotional attributes the *semester performance* was affected only by *leadership* and *drive of the students* (Mishra, Kumar, Gupta, 2014).

Using data mining methodology based on CRISP-DM methodology, Chalaris et al. (2014) found out that in the theoretical courses *student understanding* relates mainly with the instructor and teaching effectiveness, while in the laboratory practice courses, *lab facilities* are found to be the most correlated with the *achievement of learning objectives* (Chalaris et al., 2014).

Predicting student failure or revealing dropout factors in MOOC (Gupta, Sabitha, 2019) can help educators to redesign MOOC features (Xing, 2019), personalise teaching processes (Zhang et al., 2019), increase student performance (Ajibade, Ahmad, Shamsuddin, 2019) and finally keep the students from leaving the course. Of course, researching student data must be done in ethical way, respecting their privacy.

GDPR and data mining

The EU General Data Protection Regulation (2018), known as GDPR, is the most important change in data privacy regulations in 21st century. It has a significant impact on many aspects of EU citizen's data collection and processing, and affects not only EU companies but also multinationals which operate in EU. Machine learning models are fuelled by large amount of personal data. This means we need to respect the privacy of the individual in ethical way in order to overcome privacy risks (Ashford, 2019).

"Right to explanation" is another significant effect of GDPR on Machine Learning. According to Gregory Piatetsky GDPR doesn't really require an explanation of Machine Learning (ML) algorithms. Author distinguishes two explanations on those matters: Global explanation and Local explanation (Piatetsky-Shapiro, 2018).

Global explanation is mainly focused on how ML algorithm works. Some deep learning algorithms, so called black box algorithms, are almost impossible to interpret. Their complexity makes very challenging to understand exactly why, and how, a machine learning model has made a particular decision. On the other part, Local explanation deals with a question of factors contributed to a particular decision impacting a specific person. It is difficult to see how the meaningful explanation about the logic involved in some black box algorithms can be satisfied, especially in cases where a machine learning process involves multiple data sources, and elements that are not transparent or intuitive, whether for technological or proprietary reasons.

Revealing the full algorithm code and detailed technical descriptions of machine learning processes is unlikely to help. On the other hand a simple, non-technical, description of the process is more likely to be meaningful (Kuner et al., 2017).

Data set description

Database used in this research has 408 records collected from 5 generations of student activity in course "Programming 2". Dataset contains 6 variables: ID, lectures, quizzes, labs, videos and grade. Variable ID represents a student; although dataset is anonymized this variable was removed.

Variables *lectures*, *quizzes* and *labs* are total number of scores students received within the corresponding domain. Variable *videos* represent number of views of the video lectures and *grade* represents student grade on final exam. Research data was collected as described in previous research by Matetic using Interpretable neural networks in predicting student failure (Matetic, 2019). Sample of data used in this research is shown in Table 1.

Table 1. Sample of dataset used in the research

lectures	quizzes	Labs	videos	grade
0	19,33	32	15	D
5	22	27	7	D
5	15	7	10	F
5	27,66	27,5	13	C
3	28,66	0	50	F

Data exploration

First step, which precedes building a prediction model, is data exploration.

We're trying to predict grade, so we must pay attention to variables *labs* and *quizzes* which has the strongest relationship with grade. But, as the heatmap (Figure 1) suggests *labs* and *quizzes* has the strongest correlation between each other, while variables *videos* and *lectures* have the weakest correlation.

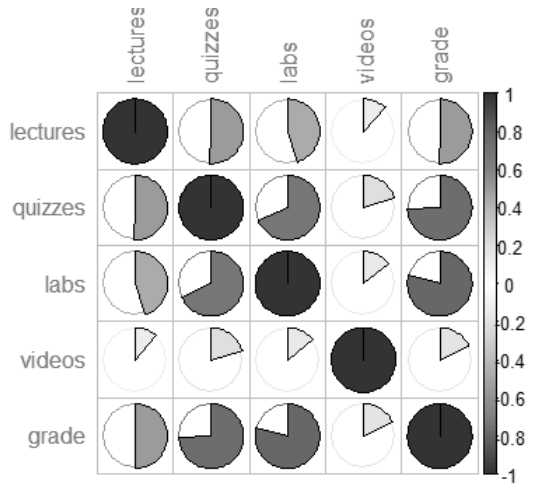

Figure 1. Data heatmap, showing correlation between variables

Plot in Figure 2 shows that FAIL grades were outnumbered by PASS ones. That means, for better results, data needs to be normalized.

Analysing plots on Figure 3, from distributions of student's grades (FAIL or PASS); we can see that lower scores in labs and quizzes mostly results in fail, giving us right skewed normal distribution. This is something that we expected.

Interesting fact to notice on quizzes plot is that FAIL distribution is slightly bimodal, showing us that certain number of students with relatively high scores on quizzes still manage to fail.

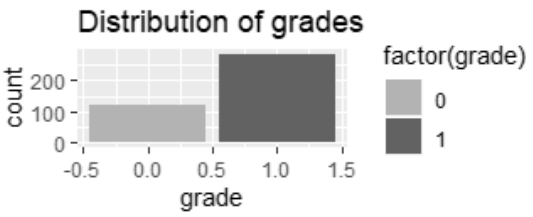

Figure 2. Distribution of grades (0-FAIL, 1-PASS)

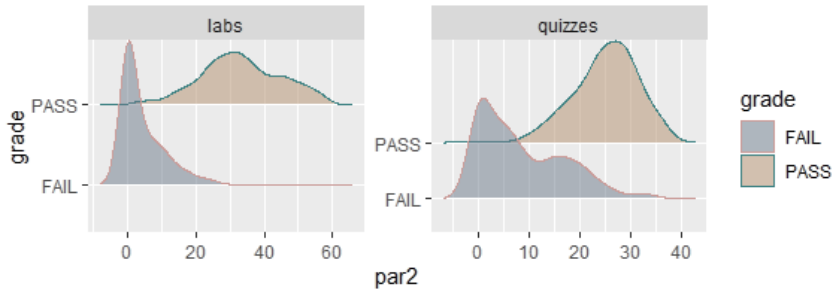

Figure 3. Distributions of students' grade (FAIL and PASS) by labs and quizzes variables; x axis (par2) shows student activity points

Building a prediction model using Random forest algorithm

Random forests algorithm constructs each tree using a different bootstrap sample of the data, and change how the classification or regression trees are constructed (Liaw, Wiener, 2003).

While in standard trees, each node is split using the best split among all variables, Random forest splits each node using the best among a subset of predictors randomly chosen at that node. This method performs very well compared to many other classifiers, including Discriminant analysis, Support vector machines and Neural networks, while it is robust against overfitting. It is very user-friendly method in the sense that it has only two parameters - the number of variables in the random subset at each node and the number of trees in the forest, and is usually not very sensitive to their values (Breiman, 2001).

For creating a Random forest model in this research, we used R language v. 3.6.1. with Caret package installed.

First step in our process was splitting our data into two sets: training data (80%) and test data (20%). We used 3 fold cross validation repeated 5 times, and then we build Random forest model (rf_model) with centered and scaled data. After the model was build, it was tested on test data, and model accuracy was 96.3%.

So, we build a model that predicts student failure using Random forest algorithm with high accuracy, but we have no clue on how this model makes prediction. Random forest algorithm is so called *black box* algorithm. Black box models, such as Random forest or Neural networks, give us little information regarding their decision-making processes, so we need an extra effort to explain it (Grigg, 2019).

Interpretation of Random forest model

Algorithms that hide their internal logic to the user, so called black boxes give us little information regarding their decision-making processes. This lack of explanation presents practical and an ethical issue. There are many approaches aimed at overcoming this weakness sometimes at the cost of reducing accuracy in benefit of interpretability (Guidotti et al., 2018).

On the other hand, models that are easy to interpret (whitebox) such as linear regression and decision trees tend to be inaccurate, as they often fail to capture complicated relationships within a dataset. In this work several methods to interpret the results of our Random forest model was used.

Variable importance

When training a Random forest model, it is normal to ask which variables have the most predictive power. High-importance variables are essential to model making and their values significantly affect the outcome values. On the other hand, variables with low importance can be left out from a model, and make it simpler and faster to fit and predict (Hoare, 2019).

The prediction error rate for Random forest classification model is calculated for permuted out-of-bag data of each tree and permutations of every feature. These two measures are averaged and normalized.

As we can see on variable importance plots (Figure 4), variable *labs* is the most important variable in decision making process. Predictor *videos* is relatively important for class value PASS, but it's overall irrelevant.

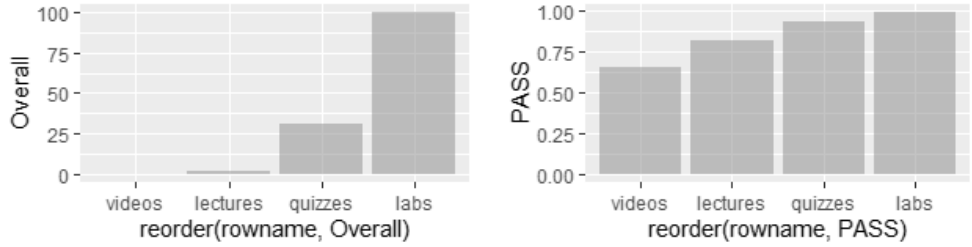

Figure 4. Overall variable importance and variable importance for PASS

Break Down model

The Break Down is a model agnostic method for decomposition of predictions from black boxes such as Random forest, Xgboost, Support vector machine (SVM) or Neural networks. As a result we get decomposition of model prediction that can be attributed to particular variables. Break down plot presents their contributions in graphical way (Figure 5).

Using R code with package breakDown, we detected variables that contributed the most to our final prediction. This method gives us the same variable *labs* as a most valuable predictor. That corresponds to result given by Variable importance tool.

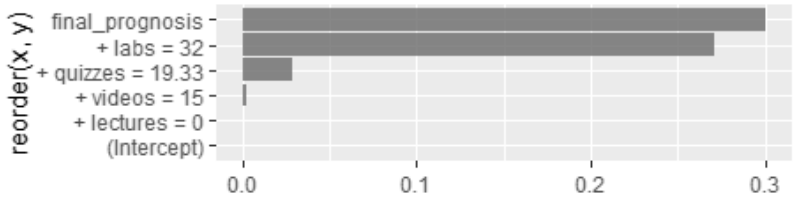

Figure 5. Break down plot visualise variables from Break down table

Tree surrogate

The tree surrogate method uses decision trees on the predictions where conditional inference tree is fitted on the predictions from the machine learning model and data. The R-squared value (variance explained) gives an estimate of the goodness of fit or how well the decision tree approximates the model. Our surrogate model has an R-squared of 0.836 which means it approximates the underlying black box behaviour quite well, but not perfectly. As we can see on Tree surrogate plot (Figure 6) *labs* is the most important predictor again.

On the right side of a plot, predictors *labs* and *quizzes* contributed the most to variable class PASS, while predictors *labs* and *lectures* (left side of a plot) are the most important to variable class FAIL.

The results are given in decision tree form, which is easy to interpret in contrast to Random forest lack of transparency.

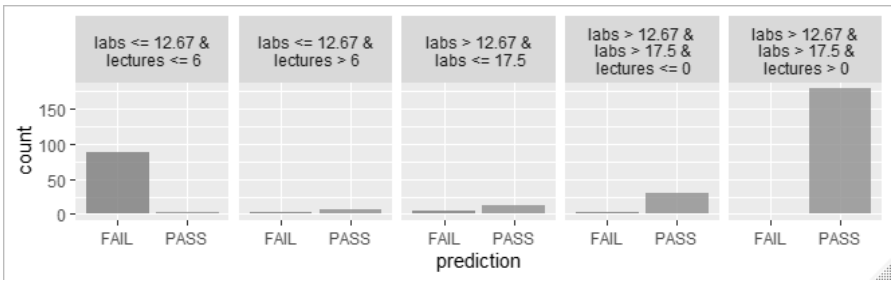

Figure 6. Tree surrogate plot

Local Interpretable Model-agnostic Explanations (LIME)

LIME is explanation technique that learns an interpretable model locally around the prediction, explaining predictions of any classifier in an interpretable and faithful manner (Guidotti et al., 2018).

LIME is explaining the predictions of black box classifiers in a way that for any given prediction and any given classifier it is able to determine a small set of features in the original data that has driven

the outcome of the prediction. It creates a model agnostic locally faithful explanation set which helps us to understand how the original model makes its decision. By creating a representative sample set LIME provides to users' global view of a model's decision boundary.

The R code will give us output (Figure 7) with huge number of single outcomes which are individually explained by predictors in their own surroundings. These explanations can be visualized, but we will end up with enormous list of cases and their plots. Figure 8 show us just a sample of visualized explanation (cases from 3 to 25 out of 172).

model_type	case	label	label_prob	model_r2	model_intercept	model_prediction
<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
1 classific~	3	FAIL	1	0.539	0.193	0.812
2 classific~	3	FAIL	1	0.539	0.193	0.812
3 classific~	6	PASS	0.992	0.206	0.585	1.08
4 classific~	6	PASS	0.992	0.206	0.585	1.08
5 classific~	10	PASS	1	0.0234	0.675	0.673

Figure 7. R output - Sample of individual cases with corresponding predictors and their weights

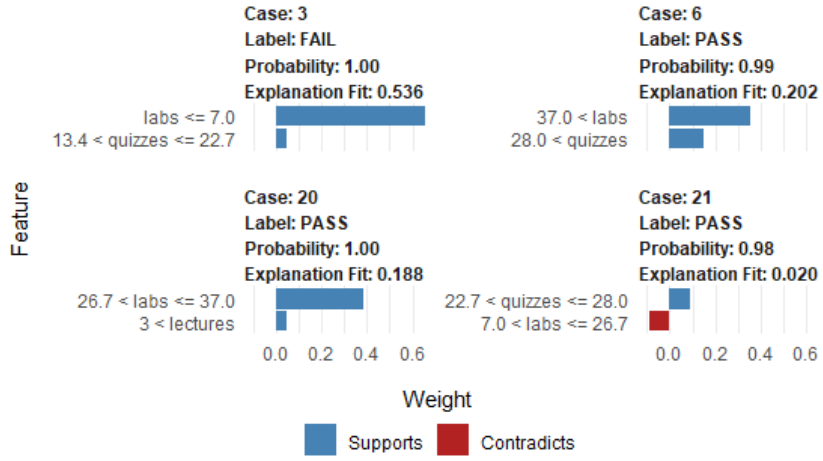

Figure 8. LIME sample of visualized explanation

Conclusion

If we need accuracy in predictions, we are usually forced to use machine learning models that are mostly black boxes. In other words, we cannot understand its learning processes or figure out logic behind its conclusions. But there are tools that explain our model's decision boundary in a human understandable way and for that purpose in this work we used several tools.

If we plan to take actions based on a prediction, or when we choose whether to deploy a new model or not, it is fundamental to understand the reasons behind predictions, and this is very important in assessing trust. Understanding the model, we can transform an untrustworthy model or prediction into a trustworthy one.

In order to create trust in our model, we need to explain the model not only to machine learning experts but also to domain experts which require a human understandable explanation.

In this work we used Random forest algorithm to build a model that can predict student failure with 96.3% accuracy what is quite good, but knowing almost nothing about which inputs contributed to that result. Using model interpreting tools, we revealed two most important variables; *labs* and *quizzes*. Variable *labs* is the strongest predictor in all our interpreting models and that understanding gives us the chance to intervene in educational process and make it better, what was our initial goal. We could use any given model to interpret our model predictions, but achieving same results with several techniques gives us trust in our model.

In our future work we plan to apply also problem domain appropriate time-series models investigating their interpretability.

Acknowledgment
This work has been fully supported by the University of Rijeka under the project number uniri-drustv-18-122.

References

Ajibade, S. S. M., Ahmad, N. B. B., Shamsuddin, S. M. (2019). Educational Data Mining: Enhancement of Student Performance Model Using Ensemble Methods. IOP Conference Series: Materials Science and Engineering 551, 012061. https://doi.org/10.1088/1757-899x/551/1/012061

Ashford, W. (2019). GDPR a Challenge to AI Black Boxes. // ComputerWeekly.Com. 2019. https://www.computerweekly.com/news/252452183/GDPR-a-challenge-to-AI-black-boxes

Breiman, L. (2001). Random Forests. // Machine Learning 45, 1, 5-32. https://doi.org/10.1023/A:1010933404324

Chalaris, M., .Gritzalis, S., Maragoudakis, M., Sgouropoulou C., Tsolakidis, A. (2014). Improving Quality of Educational Processes Providing New Knowledge Using Data Mining Techniques. // Procedia - Social and Behavioral Sciences 147, 390-97. https://doi.org/10.1016/j.sbspro.2014.07.117

Grigg, T. (2019). Interpretability and Random Forests. // Towards Data Science. https://towardsdatascience.com/interpretability-and-random-forests-4fe13a79ae34

Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., Giannotti, F. (2018). Local Rule-Based Explanations of Black Box Decision Systems. http://arxiv.org/abs/1805.10820

Gupta, S., Sabitha, A. S. (2019). Deciphering the Attributes of Student Retention in Massive Open Online Courses Using Data Mining Techniques. // Education and Information Technologies 24, 3, 1973-1994. https://doi.org/10.1007/s10639-018-9829-9

Hoare, J. (2019). How Is Variable Importance Calculated for a Random Forest? DisplayR. https://www.displayr.com/how-is-variable-importance-calculated-for-a-random-forest

Kadoic, N., Oreski, D. (2018). Analysis of Student Behavior and Success Based on Logs in Moodle. // 41st International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2018 - Proceedings, 654-659. https://doi.org/10.23919/MIPRO.2018.8400123

Kuner, C., Svantesson, D. J. B., Cate, F. H., Lynskey, O., Millard, C. (2017). Machine Learning with Personal Data: Is Data Protection Law Smart Enough to Meet the Challenge? // International Data Privacy Law 7, 1, 1-2. https://doi.org/10.1093/idpl/ipx003

Learning How to Learn (2019). Powerful Mental Tools to Help You Master Tough Subjects. https://www.coursera.org/learn/learning-how-to-learn

Liaw, A., Wiener, M. (2003). Classification and Regression by RandomForest. R News 2. // R News 3 (December 2002), 18-22

Matetic, M. (2019). Mining Learning Management System Data Using Interpretable Neural Networks, 1282-1287. https://doi.org/10.23919/mipro.2019.8757113

Mishra, T., Kumar, D., Gupta, S. (2014). Mining Students' Data for Prediction Performance. // International Conference on Advanced Computing and Communication Technologies, ACCT, 255–62. https://doi.org/10.1109/ACCT.2014.105.

Piatetsky-Shapiro, G. (2018). Will GDPR Make Machine Learning Illegal? // KDnuggets. https://www.kdnuggets.com/2018/03/gdpr-machine-learning-illegal.html

Xing, W. (2019). Exploring the Influences of MOOC Design Features on Student Performance and Persistence. // Distance Education 40, 1, 98-113. https://doi.org/10.1080/01587919.2018.1553560

Yukselturk, E., Ozekes, S., Türel, Y. K. (2014). Predicting Dropout Student: An Application of Data Mining Methods in an Online Education Program. // European Journal of Open, Distance and E-Learning 17, 1, 118-133. https://doi.org/10.2478/eurodl-2014-0008

Zhang, M., Zhu, J., Wang, Z., Chen, Y. (2019). Providing Personalized Learning Guidance in MOOCs by Multi-Source Data Analysis. // World Wide Web 22, 3, 1189-1219. https://doi.org/10.1007/s11280-018-0559-0