Statistic Lab

Learning Analytics

Report

De la Cruz Oliver‎

Krishna Vaibhav

Scherrer Michael

Table of Contents

[ABSTRACT 3](#_Toc484637987)

[OBJECTIVE 3](#_Toc484637988)

[INTRODUCTION 3](#_Toc484637989)

[DATA COLLECTION 4](#_Toc484637990)

[EXPLORATORY DATA ANALYSIS 5](#_Toc484637991)

[MODELING 7](#_Toc484637992)

[Assessment 7](#_Toc484637993)

[Variable importance 7](#_Toc484637994)

[Results - Dataset 1 8](#_Toc484637995)

[Interpretation - Dataset 1 8](#_Toc484637996)

[Results – Dataset 2 9](#_Toc484637997)

[Interpretation – Dataset 2 10](#_Toc484637998)

[RECOMMENDATIONS 10](#_Toc484637999)

[Moodle data collection 10](#_Toc484638000)

[Psychological data 11](#_Toc484638001)

[SUMMARY 11](#_Toc484638002)

[CONCLUSION 11](#_Toc484638003)

# 

# ABSTRACT

The research described in this paper aims to study how learning analytics methods can be used to impact on the process of learning. Using the data collected, various analyses and results are presented and discussed, aiming to disclose important learner’s behaviors and regularities during the educational process.

Recommendations are made for further improvements of the learning processes. The paper concludes by enumerating some challenges and further works for creating effective learning analytics tools.

# OBJECTIVE

The objective of this assignment is to identify common patterns from students by applying machine learning methods given the interaction data with the online platform and the exam performance to improve flipped-classroom didactics.

# INTRODUCTION

Education systems worldwide have advanced exponentially coincident with modern technology. Electronic learning complements learner abilities and performance, providing them with increased control over learning hours, pace, and methods via various software packages. In addition, flipped classrooms improve student understanding, concept clarification through increased discussion time and engagement with faculty.

In recent times, the ETH Zurich has taken numerous steps to integrate the use of technology in its educational system.12 E-learning has the potential to enhance quality of education. Very few studies have investigated the application of learning systems in educational institutions in the Europe and the insights gained from data collected with those systems.

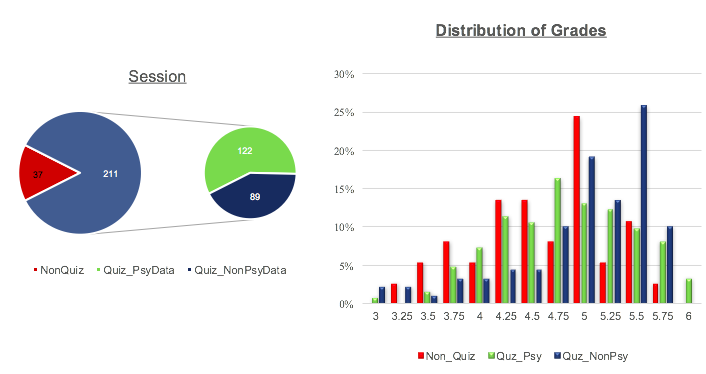
As a brief overview, the following key points need to be considered for our analysis:

* In the course Principles of Economics, a flipped classroom setting is used for teaching.
* Students prepare topics before class with the help of online materials (e.g. videos, texts, quizzes), then deepen their understanding in tutored classes.
* In this project, the data obtained from the learning management system (LMS) and quiz-results from students will be compared with their detailed exam-results in order to find possible correlations.
* The outcomes will be used to optimize materials for the out-of-class phase of the course.

# DATA COLLECTION

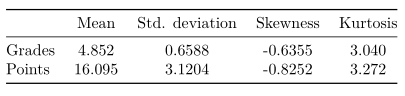
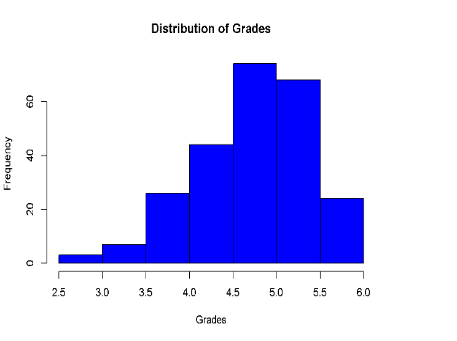
Student interactions regarding their learning experience were collected through the moodle system, a free and open-source software learning management system. Moreover, students were also asked to fill in personality tests at the beginning of the semester and the end of it. The questionnaires were distributed and collected through the system and also the attendance to participative sessions has been recorded and monitored. In addition, we partitioned the whole information gathered into two datasets due to the fact that there is quite a significant number of students who did not participated in the personality tests.

* Total Students: 248
* Students with quiz data: 211
* Students with psychological data available: 89

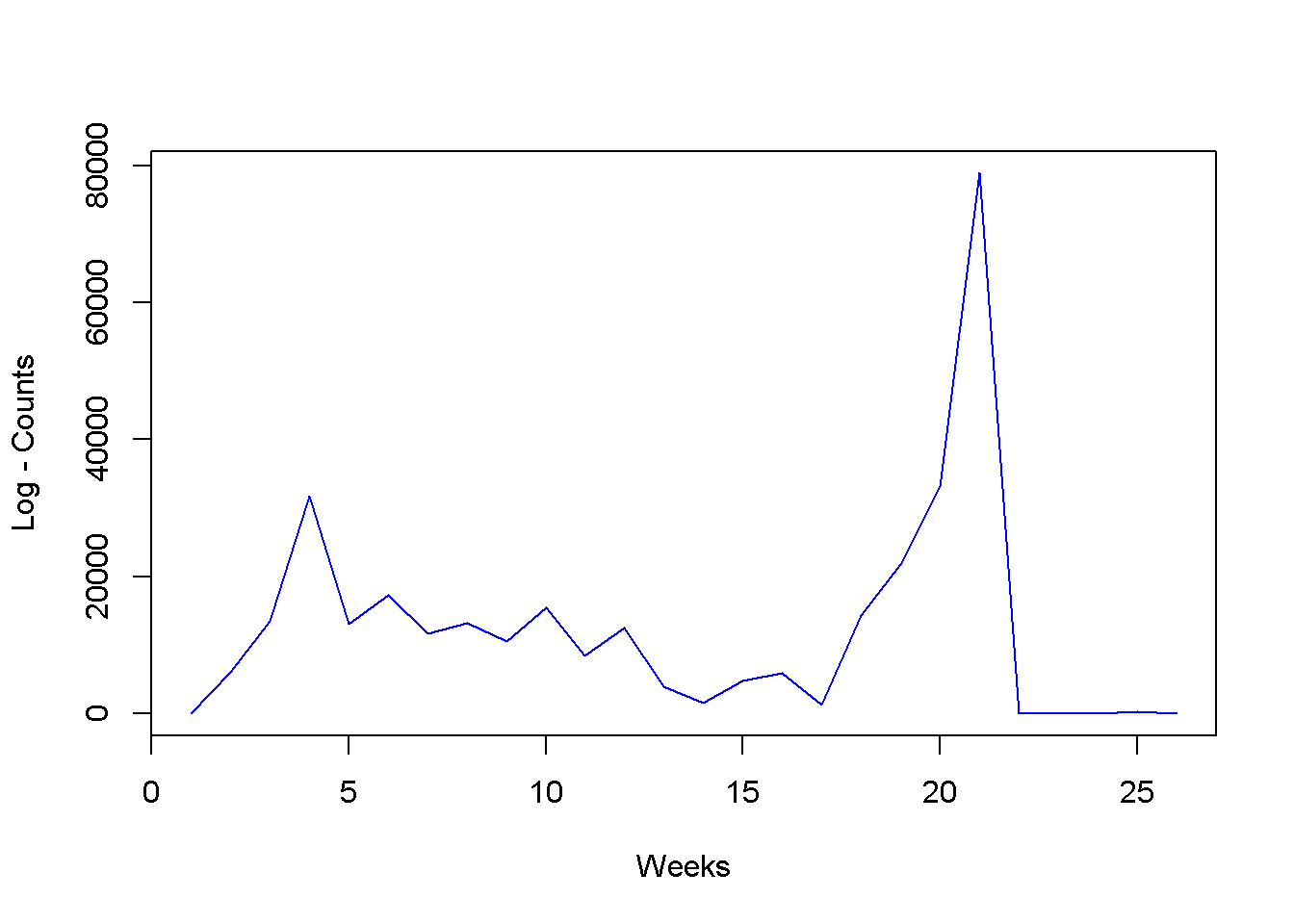
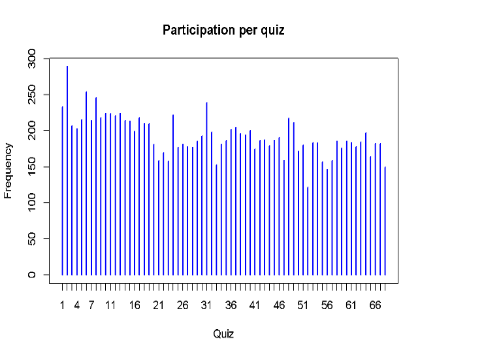


# EXPLORATORY DATA ANALYSIS

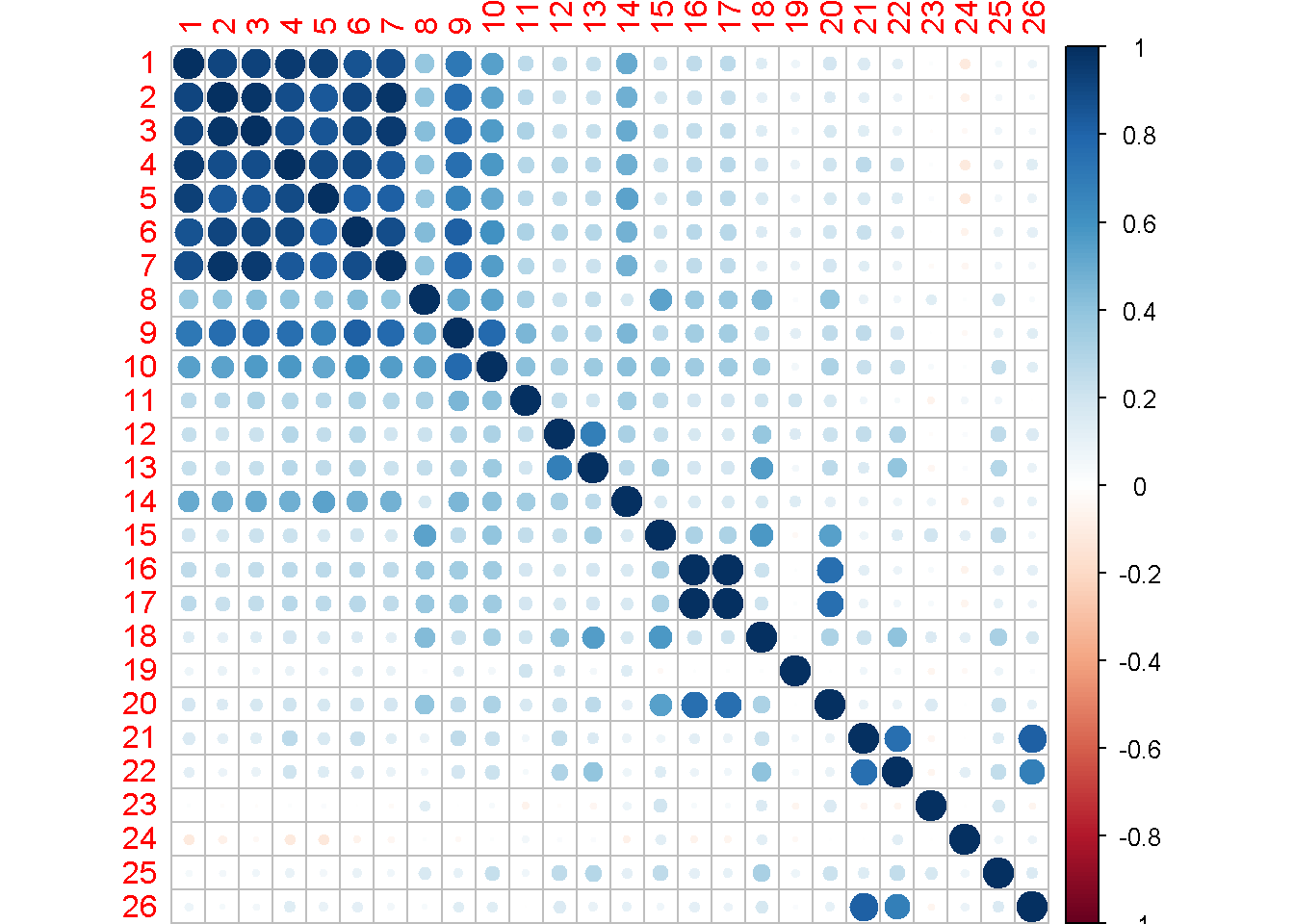
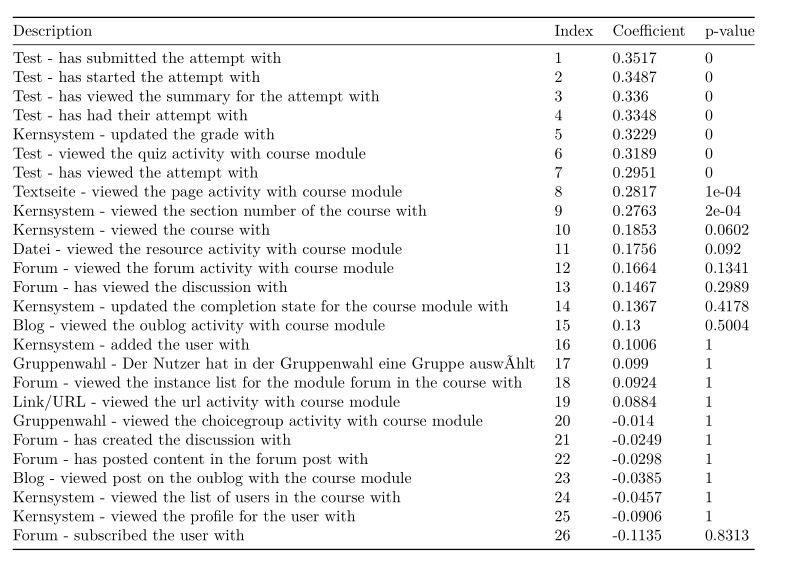
Exploratory data analysis is an approach to analyzing datasets to summarize their main characteristics with visual methods. Generally, a statistical model can be used to find trends in the data and handling missing values or making transformation of variables. As a first step, we depicted the distribution of the grades and total points obtained in the exam. As we can observe, the shape of the histogram is very similar to a normal probability distribution.



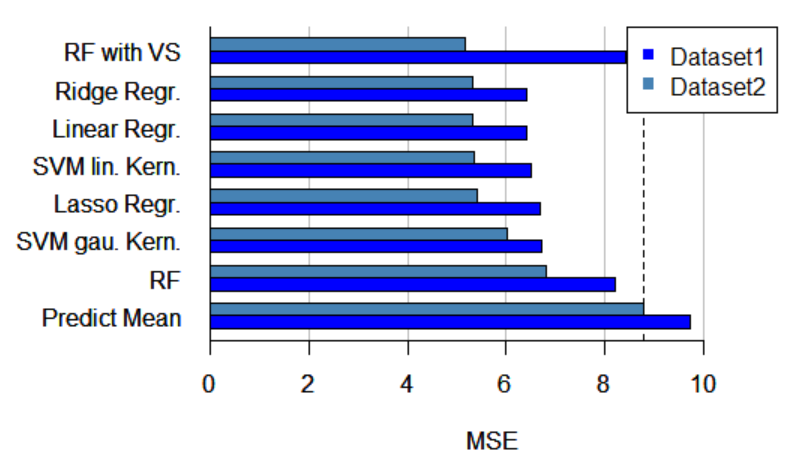
We also analyzed the level of participation for each quiz and the total number of interactions with the system. We see slightly decline in the total number of participations per quiz and an explosion on the number of logins on the weeks before the exam date.



Furthermore, as we see later when using supervised statistical methods the information obtained from the log file is quite useful and has a high predictive power. Below, we determine the correlation between the response variable in our case, the total number of points achieved in the exam and the number of counts per attribute. Note that the components which have the highest correlation are those related to the quizzes (the label test in the table below). Only the first nine variables show in the table are have significant positive correlation. Besides, we also depicted the correlation matrix between the those variables. We might be confronted a multicollinearity problem, a phenomenon in which two or more predictor variables are highly correlated. This will be considered when building our models.



# MODELING

For the two datasets seven different regression models were fitted. The dependent variable is given by the total score in the final exam (max: 21) and the independent ones as described in the in the section XXXX.

## Assessment

As a measure of quality a 10 fold cross-validated mean squared error was used. Random Forest with variable selection on the dataset 2 (with psychological data) performed best among all fitted models (figure XXXX), even though the ridge regression, linear regression, SVM with lin. Kernel and lasso regression achieve almost the same results. However, for the dataset 1 all models performed worse compared to dataset 2, this indicates that the psychological data has some predictive power.

As a benchmark, for each student the mean over all exam scores was predicted. As expected the benchmark performed inferior to the regression models.

## Variable importance

For all the models, it is possible to calculate a measure of importance for each variable.

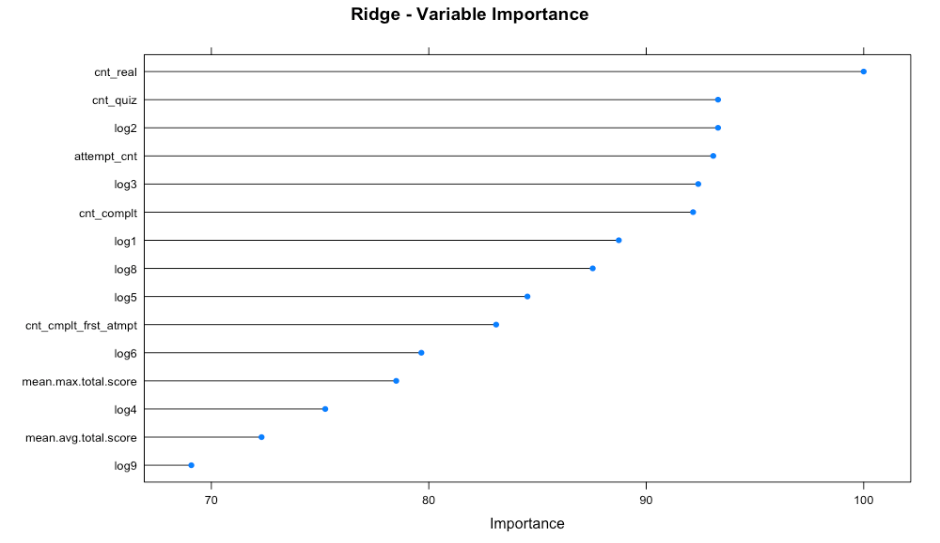
Among the different models the set of the most important variables stays fairly stable. For brevity we only discuss the best model for each dataset (ridge regression for dataset 1, random forest with variable selection for dataset 2.

## Results - Dataset 1

The most important variables according to ridge regression are the following (figure XXXX):

* cnt\_real: Number of quizzes taken with less than 3 min on avg per session
* cnt\_quiz: Number of quizzes started (by Vaibhav’s dataset)
* log2: Test – has started the attempt with
* attempt\_cnt: Number of quizzes started (by Michael’s dataset)
* log3: Test – has viewed the summary for the attempt with
* cnt\_complt: Number of quizzes regularly terminated

## Interpretation - Dataset 1

As one can see, all the variables are very similar. They all describe how often and in which quality the student participated in the Quizzes. Furthermore, those variables are highly correlated with each other as we can see in the correlation plot (figure XXXX).One can see that the participation in the quizzes has a high predictive power compared to the other variables tested, and there for a high correlation exists. However, one cannot say that the participation in the quizzes causes better exam results.

## Results – Dataset 2

The model random forest with variable selection was proposed by Diaz-Uriarte, R., Alvarez de Anders, S., 2006. It aims to reduce the number of variables used in the model by selecting the set of variables with the highest predictive power.

The most important variables according to random forest with variable selection are the following (figure XXXX):

- Role7\_2: Role team player in self-assessment (winter)

- Log7: Test - has viewed attempt with

- Cnt\_complt\_oct: completed quizzes in October

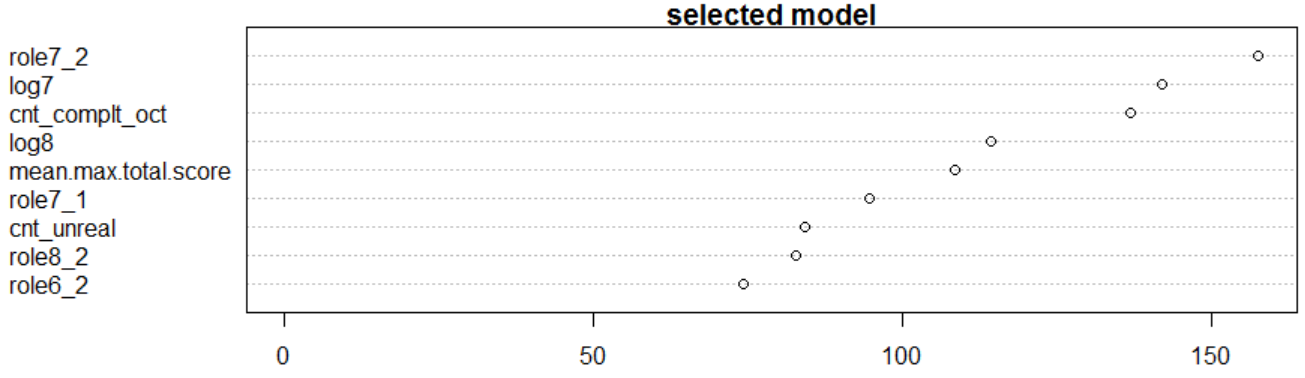
Due to some randomness in the model sometimes 9 variables get selected but mostly only the first 3.

## Interpretation – Dataset 2

The variable with the best predictive power is from the psychological dataset. It means the student describes herself/himself as a team player. It can be that team players more often try to explain the course content to their colleagues what helps themselves to perform better in the exam.

The second variable suggested by the model is from the log file and describes how often the student has submitted his answer and he has looked at the solution once submitted. This emphasizes that better students are more active on moodle (probably in general study more intensive), as we expect.

The last important variable is the quiz activity in October. This can be seen as a measure to describe whether the students are already active in the beginning of the semester or whether they just do all the work right before the exam.

In general, it is interesting that the most important variables suggested by the model are quite different in terms of meaning. However, the model does not say that those variables are the only ones that are important. It could be that another variable is highly correlated with one of the chosen variables. That means they both have almost the same predictive power and the model selects one of the two. The other variable then becomes very unimportant even though they both explain almost the same.

# RECOMMENDATIONS

## Moodle data collection

In the moodle course from the semester A16 it is possible to see the solution before submitting the results. It is even possible to see the solution after each exercise.

To perform an analysis this means that the scores in the quizzes do not indicate anything about the performance of the students. Mostly they achieved the maximum number of points, probably because they just submitted the results they knew where correct.

Also from an educational point of view it is probably better if the students only get a feedback at the end of a quiz.

## Psychological data

If a prediction of the performance for all the students in the course is a future goal one would have to collect psychological data for all the students and not just for the one from the environment department, since the most important variable originates from this dataset.

# SUMMARY

In the setting as it is, causal inference is not possible. All the lessons learned rely on correlation.

Demographic data like sex, age, department, etc. does not correlate with the exam results at all. However, the important variables are whether the students consider themselves as team player, the number of times they watched tutorials and the number of times they participated in quizzes in the beginning of the semester.

# CONCLUSION