FACULTY OF INFORMATION TECHNOLOGY AND ELECTRICAL ENGINEERING

Henna Kokkonen

Effects of Data Cleaning on Machine Learning Model Performance

Bachelor's Thesis
Degree Programme in Computer Science and Engineering
October 2019

Kokkonen H. (2019) Effects of Data Cleaning on Machine Learning Model Performance. University of Oulu, Degree Programme in Computer Science and Engineering, 32 p.

ABSTRACT

This thesis is focused on the preprocessing and challenges of a university student data set and how different levels of data preprocessing affect the performance of a prediction model both in general and in selected groups of interest. The data set comprises the students at the University of Oulu who were admitted to the Faculty of Information Technology and Electrical Engineering during years 2006–2015. This data set was cleaned at three different levels, which resulted in three differently processed data sets: one set is the original data set with only basic cleaning, the second has been cleaned out of the most obvious anomalies and the third has been systematically cleaned out of possible anomalies. Each of these data sets was used to build a Gradient Boosting Machine model that predicted the cumulative number of ECTS the students would achieve by the end of their second-year studies based on their first-year studies and the Matriculation Examination results. The effects of the cleaning on the model performance were examined by comparing the prediction accuracy and the information the models gave of the factors that might indicate a slow ECTS accumulation. The results showed that the prediction accuracy improved after each cleaning stage and the influences of the features altered significantly, becoming more reasonable.

Keywords: Educational Data Mining, Data Preprocessing, Learning Analytics

Kokkonen H. (2019) Datan siivouksen vaikutukset koneoppimismallin suorituskykyyn. Oulun yliopisto, Tietotekniikan tutkinto-ohjelma, 32 s.

TIIVISTELMÄ

Tässä tutkielmassa keskitytään opiskelijadatan esikäsittelyyn ja haasteisiin sekä siihen, kuinka eritasoinen esikäsittely vaikuttaa ennustemallin suorituskykyyn sekä yleisesti että tietyissä kiinnostuksen kohteena olevissa ryhmissä. Opiskelijadata koostuu Oulun yliopiston Tieto- ja sähkötekniikan tiedekuntaan vuosina 2006–2015 valituista opiskelijoista. Tätä opiskelijadataa käsiteltiin kolmella eri tasolla, jolloin saatiin kolme eritasoisesti siivottua versiota alkuperäisestä datajoukosta. Ensimmäinen versio on alkuperäinen datajoukko, jolle on tehty vain perussiivous, toisessa versiossa datasta on poistettu vain ilmeisimmät poikkeavuudet ja kolmannessa versiossa datasta on systemaattisesti poistettu mahdolliset poikkeavuudet. Jokaisella datajoukolla opetettiin Gradient Boosting Machine koneoppismismalli ennustamaan opiskelijoiden opintopistekertymää toisen vuoden loppuun mennessä perustuen heidän ensimmäisen vuoden opintoihinsa ja ylioppilaskirjoitustensa tuloksiin. Datan eritasoisen siivouksen vaikutuksia mallin suorituskykyyn tutkittiin vertailemalla mallien ennustetarkkuutta sekä tietoa, jota mallit antoivat niistä tekijöistä, jotka voivat ennakoida hitaampaa opintopistekertymää. Tulokset osoittivat mallin ennustetarkkuuden parantuneen jokaisen käsittelytason jälkeen sekä mallin ennustajien vaikutusten muuttuneen järjellisemmiksi.

Avainsanat: Tiedonlouhinta, Datan esikäsittely, Oppimisanalytiikka

TABLE OF CONTENTS

FOREWORD

I would like to thank my supervisor Dr. Satu Tamminen for excellent advice and guiding, as well as for interesting conversations.

Oulu, 23rd October, 2019

Henna Kokkonen

LIST OF ABBREVIATIONS AND SYMBOLS

ECTS	European Credit Transfer and Accumulation System, a unit to measure the volume of learning
EDM	Educational Data Mining
DM	Data Mining
GBM	Gradient Boosting Machine
ITEE	Information Technology and Electrical Engineering
CSE	Computer Science and Engineering
ECE	Electronics and Communications Engineering
LA	Learning Analytics
RMSE	Root Mean Squared Error
ME	Matriculation Examination
PTE	Primary Teacher Education
ICE	Individual Conditional Expectation
PDP	Partial Dependence Plot

1. INTRODUCTION

1.1. Literature Review

Educational Data Mining (EDM) refers to the application of Data Mining (DM) methods to educational data sets in order to extract new, useful and tangible information. Prediction is one of the most used methods in EDM, which usually means either to predict future events (e.g. graduation time, student retention or student performance) or to predict variables that cannot be directly collected in real time (e.g. a prediction model that uses student log data to identify the students who misuse intelligent tutoring software) [1, 2]. Predicting student performance has been a very important part of the pioneer and older research in EDM, i.e. a substantial number of studies on the topic can be found in the educational journals and conferences. The practical objective of using student performance prediction in the EDM research is often to develop learning process and guiding, e.g. by grouping learners, finding their regular and irregular patterns and identifying factors that may deteriorate learning. Another objective is usually to enhance decision making in the higher education by finding the most cost-effective ways to improve student retention and grades. [3]

The EDM research on predicting student performance has mainly been focused on the used machine learning models: the comparison of the performance of different models, the development of existing models and the analysis of a model in order to extract information about the factors that have the highest influence on the student performance. For example, Rudian et al. (2019) discussed the reasons why neural networks are used more than support vector machines in prediction [4]; Kumar et al. (2017) reviewed several EDM research papers focusing on the used machine learning models, their lowest and highest accuracies and the important factors that affect the student performance prediction [5]; Francis and Babu (2019) proposed a new hybrid algorithm that combines clustering and classification, arguing that on their data set, it performed better than a decision tree or a neural network model [6].

The preprocessing of data and its effects on the performance of a model have not been a prominent aspect in the EDM research papers apart from feature selection methods, the effects of which on the performance of different algorithms have been compared in some papers [7, 8]. In most studies, preprocessing has usually been only briefly explained or bypassed even though the importance of it is a widely recognized fact. However, the concept of fairness and transparency in machine learning, i.e. the predictions and choices of machine learning models do not discriminate against minority groups and the reasons behind them can be explained, has started to direct some interest into the other aspects of preprocessing as well. For example, Gonzalez Zelaya (2019) defined two simple metrics of volatility to quantify how prone a data point is to change its predicted value under two different preprocessing methods and presented a concrete example of this by applying two different sampling methods on two data sets [9]. This approach, however, appears to be mainly applicable to data transformation methods and is focused on demonstrating which data points are

more likely to be affected by different preprocessing methods rather than on how the performance of a model is affected by different methods.

1.2. Contribution

In this thesis, the data cleaning part of preprocessing is the primary focus. A Gradient Boosting Machine (GBM) model was trained to predict the cumulative number of ECTS the students at the Faculty of Information Technology and Electrical Engineering (ITEE), namely in Computer Science and Engineering (CSE) and Electronics and Communications Engineering (ECE) degree programmes would achieve by the end of their second-year studies. The aim is to present how the exclusion of the students who had uncertainties in their study paths affected the performance of the model. A common practice in big data applications is to use all the data that are available to train a model. Only a basic cleanup is performed, which is usually superficial, e.g. only the observations with invalid values are deleted, and the cleaning process does not go very much onto the individual level. However, in the case of ECTS accumulation prediction, students' varying study paths have to be taken into account as well. If the students whose slow or incredibly fast accumulation of ECTS is due to special circumstances are included in the training data, the results of the model may be biased. Thus, it is important to approach the data cleaning in the context of the purpose of the model.

This thesis was done as a part of a two-year project called AnalytiikkaÄly (AnalyticsIntelligence), which was funded by the Finnish Ministry of Education and Culture. The project aimed to find ways to facilitate study paths by applying Leaning Analytics (LA) methods to the planning and guiding of the studies, teaching and the management of the university. The prediction model presented in the thesis was planned to be an early warning tool for tutor teachers, whose task is to guide and advise a student group. The tool would have been used in a student's first tutor teacher meeting at the beginning of the second year to see if the student had many risk factors that could have led to not achieving the accumulation of 120 ECTS by the end of the second year. The purpose of the tool was to predict accurately ECTS accumulation among the average students who had studied continuously during their first year and intended to continue on the second year. Thus, it was important that the model was trained on the students who were on the presumed study path and not on the students who had exceptional study paths due to the compulsory national military service, for example. It is important to note that this aspect is only explained because it will elucidate the choices made in the cleaning of the data and feature selection. This thesis will not cover the actual use of the model as an early warning tool; the scope of the thesis is limited to examining the effects of different levels of data cleaning on the performance of the model.

In order to examine the effects of the different levels of data cleaning, three different data sets from the original data set were formed: an unprocessed data set that has only gone through the basic cleanup; a partially processed data set where only the students with obviously exceptional study paths have been

removed; a processed data set where the students who have even the slightest possibility of having an exceptional study path have been removed. Each of the data sets was used to train a GBM model. The performances of these models were compared for the whole test data and subgroups that characterize three different accumulation groups: the students who had an accumulation $\leq$ 70 ECTS; the students who had an accumulation $>$ 70 ECTS but $\leq$ 105 ECTS; the students who had an accumulation $>$ 105 ECTS. Thus, the possible effect of the uncleaned data on a specific accumulation group could be detected. Root mean squared error (RMSE) was used as a performance meter. To see how the data cleaning affected the factors that have an effect on the predictions, the feature importances of the models were examined. Finally, to present the model performance variation on an individual level, two students from the test data were selected, one achieving 75 ECTS by the end of the second year and the other 110 ECTS, and a result analysis was conducted by examining visually the Shapley values of the features.

This thesis is structured as follows. Section 2 describes in detail the used data set, challenges associated with it and the three-stage cleaning performed for it. Used machine learning model and its training is described as well. Section 3 presents the results. Section 4 is the discussion and future work. Section 5 concludes the thesis.

2. IMPLEMENTATION

2.1. Data Selection

The student data collected for this study from the whole University of Oulu provide the possibility to examine the factors that affect students' study progress at the university. Combining the Matriculation Examination (ME) grades from the upper secondary school studies with the university studies gives the possibility to examine if the students requiring more support at an early stage could be recognized as soon as possible before the first records from the university studies are available. Hence, as a part of this research project, other prediction models beside the second year ECTS accumulation prediction model for the students in the CSE and ECE degree programmes were created as well. It has to be noted that both the CSE and ECE programmes were selected because they have similar programme structure diagrams for the first two years and thus, selecting both increased the number of the students in the modelling.

For the CSE and ECE students, a model to predict the graduation time to a bachelor's degree was trained based on their first- and second-year studies, ME results and course specific variables indicating whether the course was passed at the recommended point of time. In addition, two respective prediction models (degree graduation time and the second year ECTS accumulation) were trained for the students in the Primary Teacher Education (PTE) degree programme. The second year ECTS accumulation prediction model for the CSE and ECE students was chosen for this thesis over the other three models for several reasons:

1. In the Finnish degree programmes, the aim is to gain 60 ECTS per year and complete a bachelor's degree in three years (180 ECTS). In the CSE and ECE degree programmes, the first and the second year are structured, i.e. by following the curriculum students should achieve at least 120 ECTS by the end of the second year. This also applies to the PTE students; however, in the PTE degree programme, majority of the courses are graded pass or fail, whereas in the CSE and ECE degree programmes, majority of the courses are graded from 1 to 5. This means that for the PTE students, the only possible variables to present their performance on a course are dummy variables indicating the passing of the course.
2. PTE students proceed in their studies excellently; for example, after data cleaning, 75% of the PTE students achieved at least 118 ECTS in two years and 50% of the accumulations were between 118 and 133 ECTS. This led to the result that the prediction model performed badly on those few students who did not achieve the goal of 120 ECTS, because the reasons for their slower ECTS accumulation could not be found in the dummy variables presenting their course performance or in their ME results.
3. The models predicting the graduation time to a bachelor's degree performed awfully in terms of accuracy because the studies during the first and the second year and the ME results could not explain the different graduation times. For example, it is common for the CSE and ECE students to work during their studies, which more often than not prolongs the graduation

time. In addition, there has not previously been any compelling need for the students to graduate in time, which has led to a custom to take out a bachelor's degree just a little before a master's degree.

4. In the CSE and ECE degree programmes, the courses are more connected to each other than in the PTE degree programme, i.e. the knowledge gained from the previous courses helps to pass the upcoming courses. This makes it more plausible that the performance on the first-year courses will explain the accumulation of ECTS by the end of the second year, because the performance directly affects the probability to pass the upcoming second-year courses.

2.2. Data Description

The base data set is composed of two data sets collected from the study registers. The first data set includes information of the students who were admitted to the university during years 2006–2015: anonymized student IDs, all the degree programmes the students have been in, degree programme start dates, possible bachelor's and master's degrees, graduation dates, status at the university in the autumn of 2015 (absent|present), birth year and the ME grades and completion year. The second data set contains all the study records of the aforementioned students till the end of 2018, connected by the anonymized student IDs: name of the passed course, ECTS gained, grade, status (passed|substituted) and the date when the course was passed or substituted. Only the rows where the degree programme was marked as CSE or ECE were selected from the first data frame, leaving 917 rows.

2.2.1. Challenges

There are several challenges associated with the used data which have to be addressed before advancing to the data cleaning.

Time span

The ten-year time span in the data causes difficulties because the programme structure diagrams have changed partially each year: courses have been removed and added, course contents have changed, the recommended time to attend a course has changed from the first year to the second, and vice versa. This had to be considered in the feature selection and ECTS calculation.

ECTS calculation

The cumulative number of ECTS had to be calculated from the study record data for each student. Because the interest is in the progress the students made in their main degree programme (CSE or ECE), only the ECTS from the courses that belong to the main degree programme were counted. For this, the

programme structure diagrams from years 2006–2015 were used: the structure diagram of a student's degree programme start year determines what courses the student is supposed to attend in the first two years. Each student's study records were examined at three points: at the end of the first autumn, the first year and the second year. Every course that belonged to the first two years of the student's respective structure diagram and had been passed or substituted before the aforementioned points were taken into account, resulting in three new columns to the first data frame: the first autumn ECTS, the first year ECTS and the second year ECTS. It should be noted that these cumulative ECTS values were calculated without a lower time bound, i.e. if a student had passed some courses belonging to his main degree programme before the actual start of the studies, those courses were also taken into account. This kind of situation is possible if a student has changed degree programmes, for example.

To make it easier to recognize anomalies in the data, all the ECTS inside five overlapping time frames were calculated regardless of what studies they were from. These time frames are the following: from the start of the studies in CSE or ECE to the end of the first autumn, the first year, the second year, the third year and the fourth year.

Combination of the data frames

In order to use the first-year studies as predictors, each student's grades from the first year had to be collected from the study record data. This was done by creating new columns with course names to the first data frame and collecting the course grades from the students who had passed the course before the end of the first year. The grades having been attained possibly before the actual start of the studies were collected as well.

As an important side note, when referring to the end of an autumn or a year in the previous section and this one, the time frame is in fact a month longer than the actual academic time period of the University of Oulu. This is because some course grades from the previous academic period might be given a little too late, and since it is impossible to attain any grades and ECTS from the courses of the following academic time period in the very first month, this prolongation of time does not cause any problems.

Missing values

Because the data are mostly ME and course grades, every student has missing values. Although several statistical and machine learning based missing value imputation methods have been developed and applied in the literature [10], no method is applicable for this type of data because imputing (e.g. giving students grades through mean or mode imputation) would make the data strongly biased, for example, by changing the shape and location of the data distribution.

Degree programme changes

There are students who have been in another degree programme before starting in CSE or ECE, and students who have changed at some point to another

degree programme. These students are not usually on the same line with the students for whom CSE or ECE is the first degree programme. For example, if a student's previous degree programme was another engineering programme, he has most likely already passed several basic courses that are common for all the engineering programmes. As another example, if a student changes to a new degree programme immediately after the first year, he will not have any studies in the CSE or ECE programmes during the second year and has most likely not been motivated to attend the courses in the CSE or ECE programmes during the first year. It is important to identify these students as their presence in the model's training data may deteriorate the model's performance.

Gaps in the study records

The CSE and ECE degree programmes are highly male-dominated: 94% of the students in the data set are male. Compulsory national military service can be done during the university studies, which causes time intervals of 6 to 12 months in the students' study records when no courses have been passed. Unfortunately, there is no clear information available about the point when a student has done his service, which means that it cannot be said for sure if the gap in the study records is due to the military service. A gap of 6 to 12 months can also be caused by parental leave or other personal reasons that do not affect a multitude of students. Nevertheless, the inclusion of the students having these kinds of special circumstances in the training data may deteriorate the model's performance.

Because the study record data do not contain any information of the failed courses and there is no enrollment status history available, it is also possible that the gap is due to the fact that the student has not been able to pass any courses. The students in this kind of situation should be included in the training data, which causes a significant challenge in the data cleaning because it is impossible to specify the exact reason for the gap.

2.2.2. Cleaning

Unprocessed Data

The unprocessed data set is obtained from the base data set with 917 rows by removing basic exceptions:

- The degree programme start date is not in August, 17 students
- The degree programme start date is unknown, 45 students
- No information about studies in the study record data, 3 students
- Double IDs, i.e. a student has been in both the CSE and ECE degree programmes, 11 students
- Nonnumerical grade in some of the basic first-year courses due to substitution, 7 students

After these removals, the unprocessed data set comprises 823 students.

Partially Processed Data

The partially processed data set is gathered from the unprocessed data set by removing students who obviously have abnormalities in their study records. This cleaning stage is especially focused on removing the students who have not studied at all or have done well in their studies but have a particularly slow ECTS accumulation due to probable military service or the fact that they have been attending a lot of courses in other degree programmes. Removals:

- Students who have changed to another degree programme before earning a bachelor's degree and the time between the starts of the new and the previous degree programmes is less than three years, 111 students
- No studies during the first autumn, 21 students
- The cumulative number of ECTS at the end of the second year is less than or equal to 5 ECTS (equal to only one course in the main degree programme in two years), 49 students
- The cumulative number of ECTS at the end of the first autumn is more than 45 ECTS (a lot of substituted courses because the students have changed from another degree programme), 8 students
- Students who have attended a lot of courses in other degree programmes, identified by examining the differences between the main degree programme specific ECTS and the general ECTS gained by the end of the first and second year, 21 students
- Students who gained more than 45 ECTS during the first year and the number increased only by a few ECTS during the second year but the general ECTS during the third year increased significantly (obvious break during the second year), 13 students

After the removals, the partially processed data set comprises 600 students.

Processed Data

The processed data set is obtained from the partially processed data set by focusing on the removal of everyone who has possible exceptions in their study records. The aim is to leave only the average students who have certainly studied continuously during the first two years. Removals:

- Students who have changed from another engineering degree programme (common basic courses) or from the Physics degree programme (usually postgraduates who do not truly intend to attend CSE or ECE courses), 19 students
- Students who have gaps in their study records, identified by examining the following attributes: the cumulative number of ECTS is the same at the end of the first autumn and the first year or at the end of the first year and the second year or the number has increased only by a few ECTS, 178 students

After the removals, the processed data set comprises 403 students.

The distribution of the cumulative number of ECTS at the end of the second year after each cleaning stage can be seen in Figure 1, showing how significantly the cleaning changed the distribution of the values. As such, this is natural, as the cleaning affects particularly the lower end of the distribution.

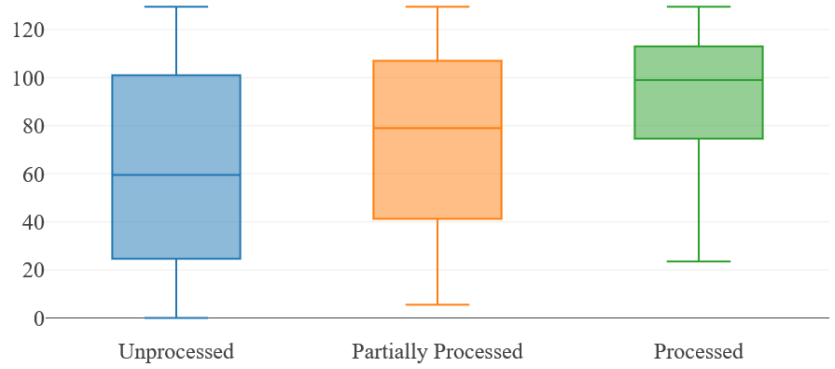

Figure 1. Distribution of the accumulated ECTS by the end of the second year after each cleaning stage.

2.3. Machine Learning Model Training

Machine learning algorithms are used to learn complex relations between a dependent variable (output) and independent variables (inputs). Currently, widely and frequently used machine learning algorithms for regression models capable of handling complex nonlinear relations are artificial neural networks, random forests and GBMs.

Neural networks are perhaps the best-known choice for modelling, but they require large quantities of training observations in order to perform well and are not able to handle missing values. Hence, neural networks are not applicable to situations in which the amount of data for model building is limited and the missing values can be regarded as an innate property of the data type.

Random forests and GBMs are ensemble methods, which build a number of regression trees called weak learners that form a collective strong learner. Both utilize bagging, i.e. a single tree is built by selecting a random subsample without replacement from the training data, which means that less data are required to build a model and missing values can be handled. The main difference is that the random forests build the trees in parallel as the method does not utilize the prediction error of the previous trees in the training, whereas the GBMs build the trees sequentially because the next tree added into the ensemble aims to minimize the error (loss function) of the previous ensemble, which is called boosting [11, 12]. In the case of squared error loss function, boosting practically means the fitting

of the next tree into the residuals of the predictions of the current ensemble and the true values of the dependent variable [12].

GBM was chosen for this thesis because of the boosting feature and because regression trees are capable of handling variables of any type and scale, can handle missing values and automatically model the interactions between variables [13]. Gradient boosting has also shown considerable success in many practical applications and various machine learning and DM challenges [14]. R's gbm package was used to build the models.

2.3.1. Test Data and Training Data Partition

The test and training data partition was done in three stages, starting from the processed data set.

The processed data set was split at random using 85% as the training data (N=343) and 15% as the test data (N=60). As the number of the students is small, even the slightest changes in the test and train partition will affect the results by causing a very different series of splits [15 p.312]. Hence, the random seed for the split was chosen with the aim of having the distributions of the accumulated ECTS by the end of the second year and the predictor values within the test and training data as balanced as possible. Thus, the training data would represent different student groups as well as possible, and a situation in which all the extreme values had been randomly placed in the test data could be avoided, for example. This was conducted by creating different splits with different random seeds. The balance of the split was examined simply with histograms and confirmed with a two-sided Mann-Whitney U tests, and the first adequate split was chosen.

The test and training data sets of the processed data were used as the core test and training data sets for the partially processed data. Firstly, the training data of 343 students and the test data of 60 students were extracted from the partially processed data set, leaving only the students who had been removed. This set of removed students was split at random in 85–15 ratio, and the balance of this split was examined. These students were added to the core test and training data sets, resulting in the training data set of 510 students and the test data set of 90 students for the partially processed data.

The same procedure was applied to the unprocessed data for which the number of the students in the training and test data sets is 700 and 123, respectively. The idea behind using the same core test data set is to ensure the comparability of all the three models. The approach of adding into the same core test data was chosen over removing from the same test data selected from the unprocessed data set because the latter approach could have led to highly imbalanced data sets.

2.3.2. Hyperparameter Tuning

Since the number of the students is quite small, a decision not to sacrifice observations for a separate validation set for hyperparameter tuning was made.

Hence, 5-fold cross validation was used instead. There were five parameters to tune: learning rate, number of trees, bag fraction, interaction depth and minimum observations in a node.

The bag fraction, interaction depth and minimum observations in a node were chosen to be 0.5, 3 and 5, respectively, for all the three models. Bag fraction of 0.5 means that randomly subsampled half of the training data is used to build each weak learner, which enables the stochastic gradient descent [16]. The values of the interaction depth and minimum observations in a node ensure that the weak learner trees are shallow, which in its turn reduces overfitting. The fixed values of these three hyperparameters are based on the previous models built from the same data for which an extensive, time-consuming grid search with 5-fold cross validation was performed in order to find the optimal values. It was noted that the tuning of these parameters from the aforementioned values did not actually improve the model performance considerably, and there is no reason to believe otherwise for the models in this thesis.

Learning rate was chosen from the set of {0.001, 0.005, 0.01, 0.05, 0.1}. The optimal number of trees associated with each learning rate value was the iteration out of 10,000 iterations that had the lowest cross validation error. The optimal learning rate was also chosen based on which learning rate and number of trees value pair had the lowest cross validation error.

2.3.3. Feature Selection

The features in this data set are mainly course and ME grades. Courses were selected based on their existence in both the CSE and ECE programme structure diagrams currently and for most of the ten-year time span with the same recommended time to attend the course. The ME grades were chosen from the exams which had been taken at least by 50% of the students in each of the differently processed data sets. The base feature set is as follows:

- Grades from the courses Calculus I, Calculus II, Matrix Algebra, Probability and Mathematical Statistics, Digital Techniques I and Electrical Measurement Principles (Discrete ordinal values from 1 to 5)
- Dummy variable from the course Elementary Programming, 1 for passed and 0 for not passed or not participated
- ME grades from the exams Native Language (FIN), Advanced English, Swedish, Chemistry, Physics and Advanced Mathematics (Discrete ordinal values from 2 to 7 for passed exam, 0 for failed)
- ME average (Continuous variable)
- Number of exams taken in the ME (Discrete variable)
- Years between the ME and the start of the degree programme; gap years (Continuous variable)
- First autumn ECTS and the first autumn average (Continuous variables)

The chosen feature selection method was backward stepwise selection [17] from the base feature set. The selection procedure was conducted for each model as

follows. The model was trained ten times with the current feature set. At the end of each training round, the RMSE values for the whole test data and for the upper, middle and lower thirds of the test data were recorded, as well as the least significant feature in the model and its relative influence. If the least significant feature stayed the same and its relative influence was less than 1 in the majority of the rounds, it was deleted unless the removal of the previous least significant feature had caused the model performance to deteriorate in the whole test data or in some subgroup of it by more than 0.5 ECTS on average (the limit was based purely on empirical knowledge of the variation of the average accuracy when the model had been trained multiple times with the same feature set). In that case, the previous removed feature was restored and the final feature set for the model was formed. If the performance had not deteriorated but the relative influence of the least significant feature had stayed above 1, the current feature set was chosen as the final feature set. This procedure was conducted to attenuate the stochastic component in the model building.

It has to be noted that using the same test data that is used to test the generalization ability of the model in the feature selection is considered to be an erroneous practice in general. A separate test data set should be used, which is not possible in this case due to the small number of the students. However, because the results are used to assess the effects of the data cleaning on the same model trained with differently processed data sets, and the core test data for each model is the same, choosing the features that perform in the best way in this specific test data is not deemed as severe a mistake as it would be if the purpose was to compare the performance of different machine learning models. On the contrary, if the deletion of a certain feature deteriorates the performance of a model trained on particular data set but has no effect on the other two models trained with differently processed data sets, it provides insight into how the cleaning affects the information the models give about significant features.

3. RESULTS

Henceforth, the different models are referred to as follows: the model trained with the processed data set is called model 1, the model trained with the partially processed data set is model 2 and the model trained with the unprocessed data set is model 3.

3.1. Accuracy

The RMSE values of each model for the whole test data, the core test data and the subgroups representing different accumulation groups can be seen in Table 1. As seen, model 1 clearly outperforms the other two models in the core test data, as well as in the whole test data and every subgroup. Especially in the intermediate subgroup, the performance of model 1 is great, average error being approximately 8.6 ECTS.

It is interesting to note that every model achieves the best accuracy in the intermediate subgroup and the worst accuracy in the top subgroup. The high error of the top subgroup is partially explained by the students who get a low prediction based on their first-year studies as they indicate a risk to fall behind but are able to catch up with the schedule after all. With regard to the purpose of the model, the high error of the top subgroup is not as severe as it would be in the lower subgroups, because giving low predictions for the students who have risk factors but will do well in their studies in the end is not as harmful as giving high predictions to the students who will not do well.

The difference of 8 ECTS in the lowest subgroup between model 1 and the other two models presents the effect of the data cleaning very well, because models 2 and 3 contain a lot of students in the lowest subgroup whose low accumulation is due to inexplicable gaps in the study records. Thus, there are students who have similar study success during the first year, some of whom achieve a high accumulation and others a lower one due to personal reasons. This causes confusion in the modelling and results in deteriorated prediction accuracy.

The RMSE values for the same 60 students in the core test data prove that the data cleaning has helped the model to find more explanatory connections between the features and the outcome as the average error reduces after each cleaning stage, achieving the highest accuracy with the processed data.

3.2. Information of the Most Significant Features

The relative influences of the features for each model can be seen in Table 2. The relative influence of a variable is the square root of the measure based on the number of times the variable is chosen for splitting in the internal nodes of a tree, weighted by the squared improvement to the model's error as a result of each split, and averaged over all the trees in the ensemble. [15 p.368] Thus, relative influences add up to 100 in each model, and, for example, a relative influence of over 50 for a variable means that the variable in question accounts for over 50%

Table 1. RMSE values for each model. The number of students in the group is denoted in the parenthesis after the RMSE value.

RMSE	Processed	Partially Processed	Unprocessed
For the whole test data	13.97330 (60)	18.91363 (90)	20.80097 (123)
For the core test data	13.97330 (60)	18.98766 (60)	21.81183 (60)
Under 70 ECTS	12.52278 (16)	20.08459 (43)	20.29351 (71)
From 70 to 105 ECTS	8.58866 (18)	13.40944 (21)	14.17405 (21)
Over 105 ECTS	17.40722 (26)	20.63735 (26)	25.24782 (31)

reduction to the loss function in the model with the specific set of features. Table 2 presents prominently that models 2 and 3 are mainly basing their predictions on the grade of Calculus II, whereas the influences of the grades of Calculus II and Matrix Algebra, as well as the first autumn ECTS are almost equal in model 1, the first autumn average and the grade of Calculus I following closely behind.

In model 3, the grades of Calculus II and Digital Techniques I along with the first autumn ECTS account for 90% of relative influence. The model is basing its predictions mostly on these features and does not see any significance in any other ME grade than Physics.

Model 2 sees some significance almost in every feature, even though the features Calculus II and the first autumn ECTS are clearly the most significant features, accounting for over 70% of relative influence. It is also notable that the ME grades of Native Language (FIN) and Advanced Mathematics, as well as the ME average have influence in this model besides Physics, showing some connection between the success in the ME and the success in the university studies. It is also very interesting how the model learned to use the feature Years between ME and the degree start as an explanatory factor for the students who have a low ECTS accumulation due to gaps in the study records: deleting the feature from the model deteriorated the model's prediction accuracy in the lowest third of the test data. The centered Individual Conditional Expectation (ICE) curves combined with the Partial Dependence Plot (PDP) of the feature can be seen in Figure 2. It shows that one gap year has a positive impact on the prediction compared to zero gap years, which can be partially explained by the fact that those who have one gap year *before* university studies will most likely do their military service during the gap year and thus be able to study without interruptions.

The influences in model 1 are more evenly distributed across features, i.e. the model does not rely almost completely on one feature as the other two models. It is also significant how the influence of Native Language (FIN) almost quadrupled compared to model 2. In addition, Physics had little influence in this model, replaced by the influence of the grade of Advanced English that had little influence in the other two models. It is also very significant how the influences of Matrix

Algebra, the first autumn average and Calculus I increased in the model compared to the other two models.

All the three models offer similar information of the two most important features, Calculus II and the first autumn ECTS. Figure 3 shows the ICE curves and PDPs of the features for each model. On average, there is a gradual increase in the prediction of each model as the grade of Calculus II raises, but in models 2 and 3, the individual lines are more dispersed. On the feature the first autumn ECTS, all the models see a significant increase in the prediction if at least 15 ECTS are achieved and another smaller increase if over 20 ECTS are achieved.

The features Matrix Algebra and the first autumn average gain a much more significant influence in model 1. Figure 4 shows the ICE curves and PDPs of these features for each model. Model 1 indicates a more distinct increase in the prediction individually and on average if at least a grade of 3 is achieved. The other two models indicate a similar increase if at least a grade of 4 is achieved, but the individual lines are more dispersed, especially in model 2. As for the first autumn average, model 3 does not indicate any clear points after which an increase in the prediction is achieved, whereas model 2 indicates a uniform increase if an average over 3.5 is achieved. Model 2 also indicates a smaller increase if an average of 2 is achieved. The complete cleaning reveals a more significant increase at this very point and indicates a smaller increase as well if an average over 3.5 is achieved.

Figure 5 shows the ICE curves and PDP of Native Language (FIN) for models 1 and 2. Model 2 indicates to some extent that a grade below 3 is a risk factor. The complete cleaning reveals a stronger indicator of this.

Figure 2. The centered ICE curves (black lines) indicate how the prediction changed for individual students when the value of the feature Years between ME and the degree start in model 2 was altered, and the PDP (red line) indicates the average change in the prediction.

Table 2. Relative influence of each variable in each model sorted by the influences of the model trained with the processed data

Feature	Processed	Partially Processed	Unprocessed
Calculus II	22.97686	53.6687587	51.133032
Matrix Algebra	19.402799	4.5329968	4.432226
First Autumn ECTS	16.573725	20.5404828	26.949377
First Autumn Average	10.595998	3.7719194	2.229105
Calculus I	9.963852	1.0423006	1.031182
Digital Techniques I	6.00821	6.824873	11.775142
Native Language (FIN)	3.920231	1.0349105	-
Prob. and Mathematical Stats	2.68019	1.4393533	-
ME average	2.584942	2.0681442	-
Electrical Meas. Principles	2.150251	0.990944	-
Advanced Mathematics	1.85792	1.2057412	-
Advanced English	1.285022	-	-
Physics	-	1.1844722	1.418044
Elem. Programming (passed)	-	1.034338	1.031891
Years between ME and the degree start	-	0.6607654	-

3.3. Examples of Result Analysis for Individual Students

The differences in the models on the individual level are examined with Shapley values, which are a game-theory based approach on how the prediction is distributed among the features: they indicate how each feature value contributed to the prediction of an instance compared to the average prediction in selected comparison group [18]. The comparison group is the training data in these examples.

Figure 6 presents the Shapley values of a random person selected from the core test data who achieved 75 ECTS by the end of the second year. Some values have been altered to ensure anonymity. In every model, the first autumn ECTS and the grade of 3 from the course Digital Techniques I are factors that increase the prediction from the average the most. The missing grade of the course Calculus II is the factor that decreases the prediction from the average the most, especially in models 2 and 3, which emphasize this factor greatly as they

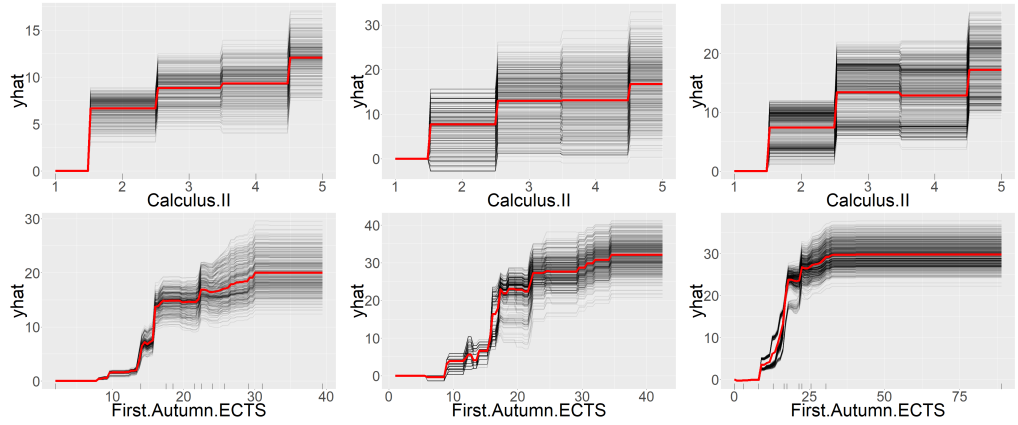

Figure 3. ICE curves and PDPs for the features Calculus II and the first autumn ECTS for the processed data (column 1), the partially processed data (column 2) and the unprocessed data (column 3). Note the different ranges of the y axis, as well as the different ranges of x axis in the feature the first autumn ECTS, which is caused by the fact that the students with particularly fast accumulations due to the substitution of courses have been removed in the cleaning process.

predict the accumulation too low. Model 1 is the most accurate in its prediction, sharing the feature contributions more evenly.

The Shapley values of a random person selected from the core test data who achieved 110 ECTS by the end of the second year are presented in Figure 7. Some values have been altered to ensure anonymity. In every model, the grade of Calculus II increases and the missing grade of Digital Techniques I decreases the prediction the most from the average. As seen, models 1 and 2 are more accurate in their prediction than model 3. In model 1, the grades of Matrix Algebra and Probability and Mathematical Statistics and the first autumn average contribute to the prediction more than the first autumn ECTS as opposed to model 2.

These two examples present the way the model could be used as a tutor teacher's early warning tool. Even though all the three models offer similar information of the features that contribute to the prediction the most, cleaning uncovers more significant information. Model 3 bases the predictions mainly on three features and lacks in accuracy, whereas the other two models have a better accuracy in general and give a more comprehensive picture of the features that contribute to the prediction. Even just partial cleaning indicates much more features that contribute to the prediction and achieves better accuracy in general, even though model 2 appears to be biased to some extent, emphasizing the grade of Calculus II strongly above anything else. The complete cleaning gives the most accurate predictions and a more plausible overall picture of the feature contributions. Of course, it is important to note that Figures 6 and 7 show only two examples from the individual level and do not provide a comprehensive picture of the feature contributions in general. Nevertheless, these two examples highlight the effects of the data cleaning very well.

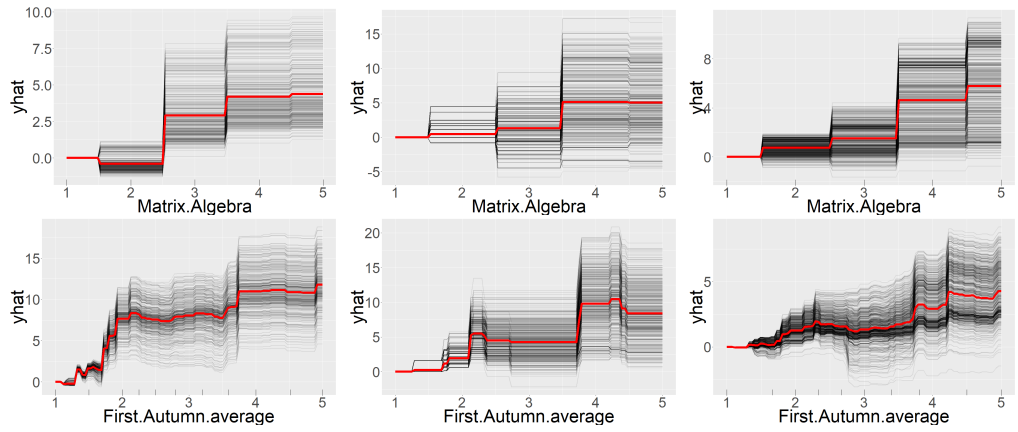

Figure 4. ICE curves and PDPs for the features Matrix Algebra and the first autumn average for the processed data (column 1), the partially processed data (column 2) and the unprocessed data (column 3). Note the different ranges of the y axis.

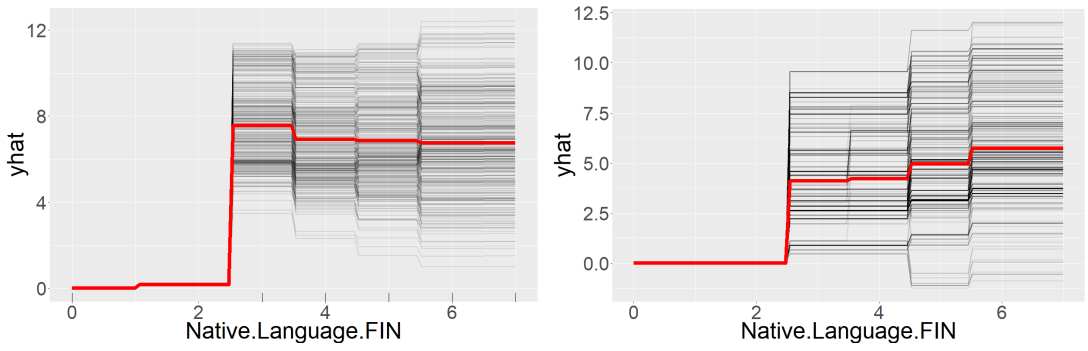

Figure 5. ICE curves and PDPs for the feature Native Language (FIN). The left plot is from the model trained with the processed data and the right plot is from the model trained with the partially processed data.

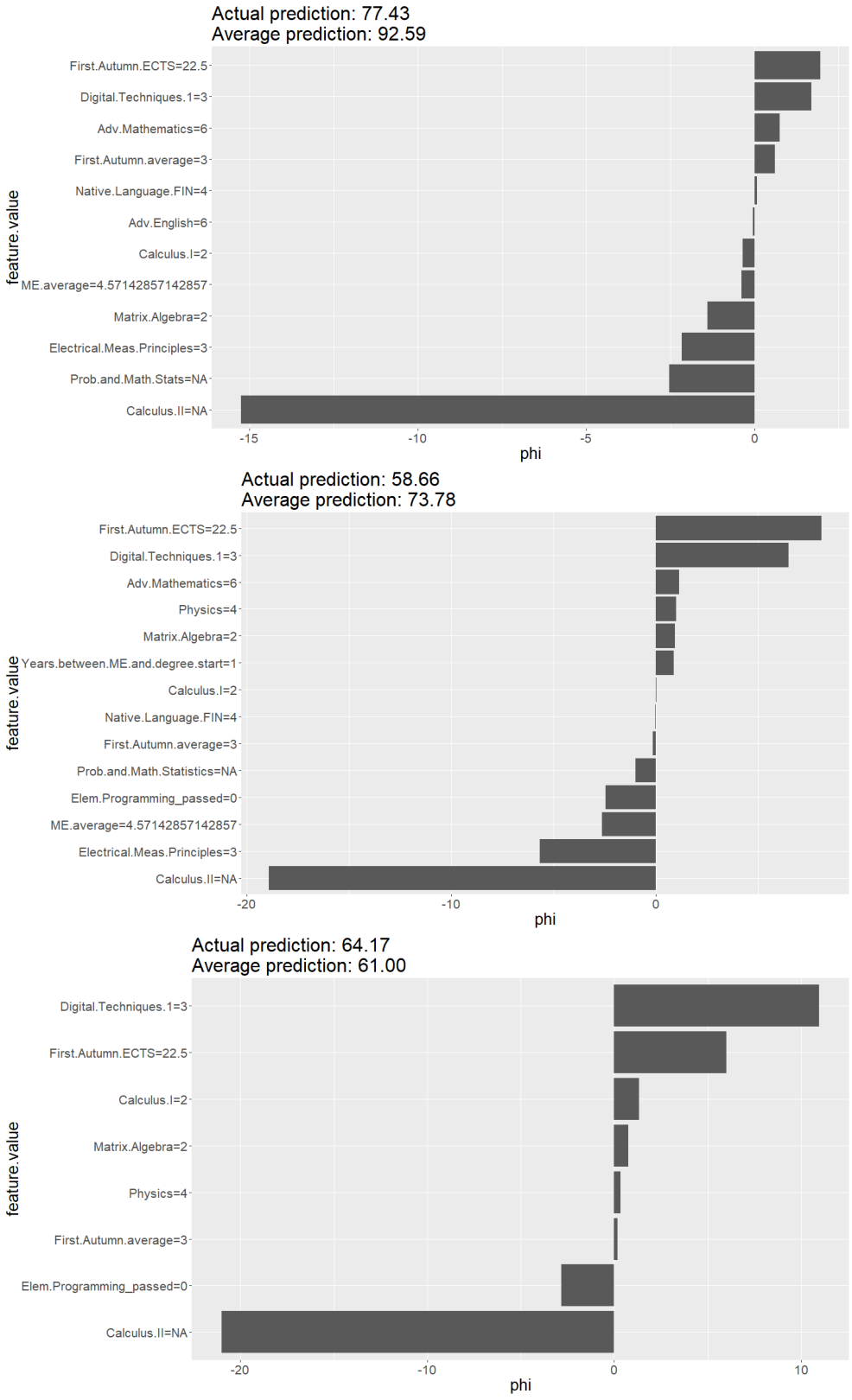

Figure 6. Shapley values for a student who achieved 75 ECTS by the end of the second year. The top plot is from the model trained with the processed data, the middle one from the model trained with the partially processed data and the bottom one from the model trained with the unprocessed data.

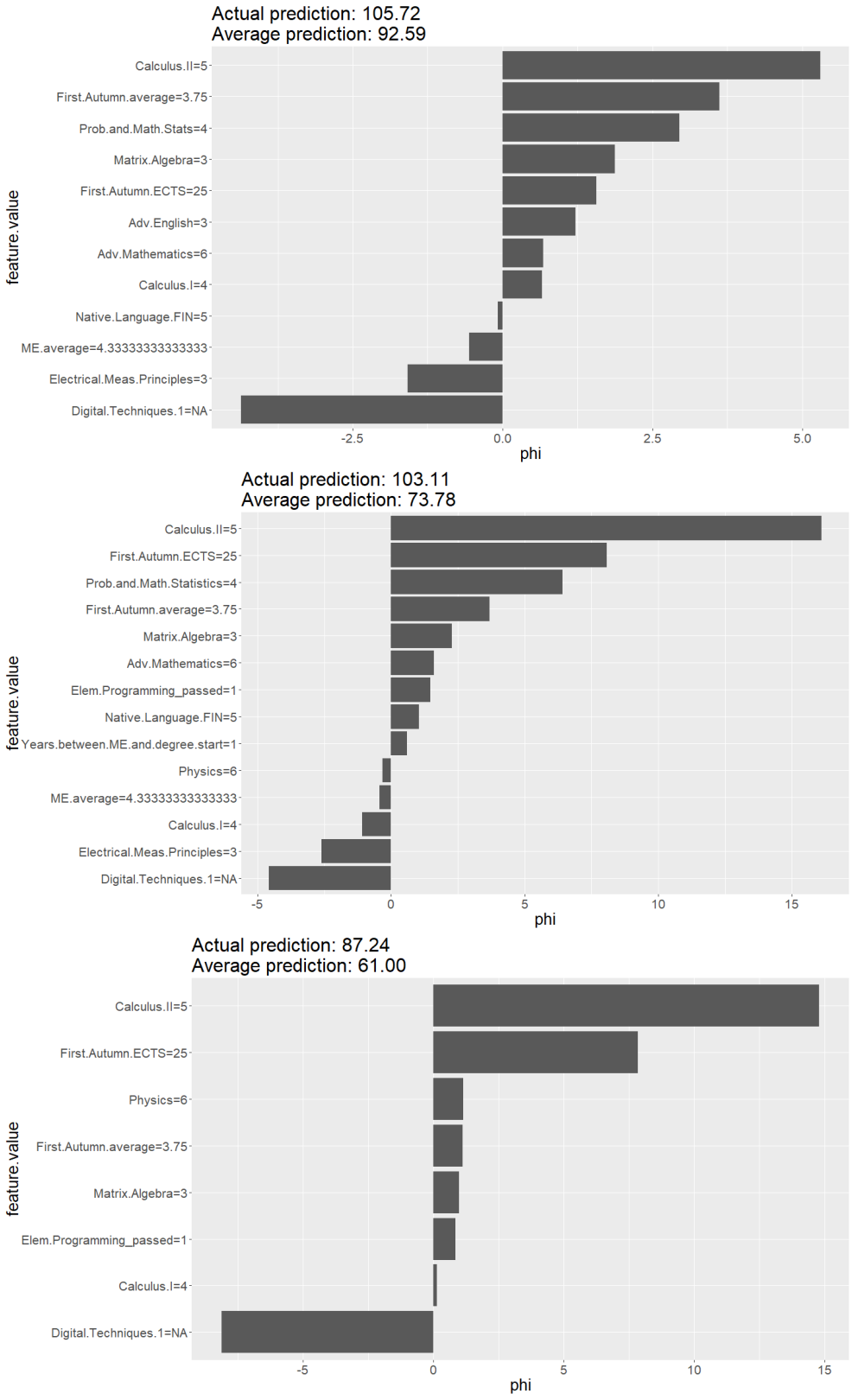

Figure 7. Shapley values for a student who achieved 110 ECTS by the end of the second year. The top plot is from the model trained with the processed data, the middle one from the model trained with the partially processed data and the bottom one from the model trained with the unprocessed data.

4. DISCUSSION

4.1. Outcome of the Cleaning

The aim of this thesis was to demonstrate the significance of proper data cleaning in machine learning model performance. A good understanding of the characteristics of data is extremely important before applying any DM methods. In order to find useful information from a given data set, the data have to be cleaned in such a way that the observations will represent the group of interest in the most comprehensive and accurate way as possible.

In this specific case, using the data set that had only been superficially cleaned out of invalid values led to a model that was very arbitrary in its predictions and provided a very simplistic, improbable picture of the feature influences. When the students with particularly slow or fast accumulations were removed, a more detailed picture of the feature influences was provided. However, the removal of the students with gaps in study records implied that the information of the previous model might have been misleading. For example, the model trained with the partially processed data indicated that changing the value of the gap year feature from zero to one increased the prediction to some extent. One could argue that this indicator is logical as those who have had one gap year might be more motivated in their studies because they might have had to try harder to get admitted, and they might have matured a little more. However, since the removal of students with gaps in the study records diminished the influence of this feature, it would appear that the students with zero gap years are more likely to have a gap in study records due to military service as opposed to the students with one gap year.

The results strongly indicate that the most processed data set represents the average students who study continuously for two years the most accurately out of these three differently processed data sets. However, it is possible that the processed data set still contains some anomalies and some students may have been removed unintentionally even though they can be considered as average students. For example, some of the students who have gaps in their study records might have been attending to courses but have not been able to pass any of them, and that has resulted in them having inexplicable gaps in their study records.

As a result of the systematic cleaning, the number of the students for model building reduced by 50% in the data set, which might appear at first glance that valuable data were lost. In this case, however, this reduction of observations is actually in line with the knowledge of the dropout rate of the students in the CSE programme. According to the education coordinator of the CSE degree programme, around 30 students have earned a bachelor's degree annually for the past four years, which is half of the annual intake of about 60 students. This gives some veracity to the final result of the systematic data cleaning. Furthermore, with this systematically cleaned data set, the prediction model was able to find the most reasonable connections between the dependent variable and the independent variables.

In the beginning of this study, a decision was made to combine the CSE and ECE degree programmes for the training because the amount of the training data

would have been very small otherwise. The decision was based on the fact that the students in these two programmes are known to have similar backgrounds, and the programme structure diagrams for the first two years are almost identical. The cleaning process did not take into account the degree programmes of the students, but it affected both programmes evenly, nevertheless: in the unprocessed data set, 54% of the students are in the ECE programme, and the corresponding percentages for the partially processed and processed data sets are 55% and 57%. Thus, the final result of the cleaning did not indicate that these two programmes would have been treated differently.

4.2. Future Work

Similar prediction models such as the one presented in this thesis have been desired to build also for other degree programmes. However, the manual preprocessing of data for each degree programme would be a very arduous task, requiring a lot of time and effort. Thus, the automation of the preprocessing is a very appealing prospect: could the whole procedure be automated?

In the educational data set used in this thesis, the cleaning was done manually for the CSE and ECE students. The basic cleaning conducted for this specific data set could be easily automated for every degree programme as the removals on that stage were mainly based on general attributes such as unknown degree programme start date or no studies in the study record set. However, the removal of the students who had a nonnumerical grade in some of the numerically graded first year courses due to substitution is not applicable to those degree programmes in which most courses have a nonnumerical grading to begin with.

Advancing beyond the basic cleanup level to more detailed cleaning stages makes automation more difficult as the cleaning becomes more degree programme specific. Different degree programmes often require different approaches to the data cleaning for the removals depend on what is considered as an exceptional study path in the context of the degree programme. For example, in the PTE degree programme, the students with gaps in the study records do not have a significant part in the cleaning as opposed to the CSE and ECE degree programmes, because there are very few of them. On the other hand, the students who have gained a great number of ECTS due to the substitution of many courses or even whole study modules require more attention in the cleaning of the PTE data, because the data contain multiple students who can have even up to 60 ECTS substituted during the first year.

Half automation of the data cleaning process could be a plausible option as the next step. Although a comprehensive cleaning is a degree programme specific process as mentioned above, common attributes, such as gaps in the study records, the number of substituted ECTS and the ECTS limits used to determine whether an accumulation is particularly slow or fast, can be used as a basis for the cleaning. For example, the cleaning processes conducted for the PTE degree programme, as well as the CSE and ECE programmes have been applied to the Economics programme: the cleaning steps of these processes have been a basis that has been used to examine whether similar exceptions can be found in the Economics

data. An automated system that searches given degree programme data for these common attributes that are known to be signs of unusual study paths in other degree programmes could be build. However, the ultimate decision whether the students found by the system are truly exceptional has to be made by a human. In addition, the data have to be re-examined after using the system because it may contain exceptions that have not been present in other data sets. Of course, introducing an artificial intelligence to this half-automated process would make the data cleaning completely automated in the end.

To summarize, the automation of the data cleaning process would be easy if the cleaning was only focused on removing observations based on generic attributes. However, in most applications, proper data cleaning requires knowledge from the application field, and the cleaning must be done with regard to the purpose of the model. In this case, the automation of the data cleaning could be possible as well if an artificial intelligence was used, for example. However, the size of the data set affects greatly the sensibleness of the automated data cleaning. If the size of the data set is small and the application field is limited, manual data cleaning could lead to a better final result as every removal is recorded and justified, and a concrete picture of the nature of the data is gained, which will help in the interpretation of the results of a model.

5. SUMMARY

The effects of data cleaning were examined in this thesis. First, a description of the data set, its challenges and the three-stage cleaning process was given in detail, presenting why this kind of student data was perfect for examining the effects of data cleaning on the performance of a model; that, in this kind of data, it truly mattered who were the students left into the training data in order to get the best and most reasonable results out of the prediction model. In addition, a description of the machine learning model and its training was given, in which the choice of the GBM model was argued and an overall picture of the training process given.

The results showed that the prediction accuracy of the model improved with every cleaning stage, and the best accuracy was achieved with the most processed data set. All the models showed that the grade of the course Calculus II and the accumulated ECTS by the end of the first autumn had the highest effect on the prediction, but the model trained with the most processed data showed the most reasonable and evenly shared feature effects. The model trained with the partially processed data was able to find more reasonable connections between the predictors and outcome and achieve a better accuracy than the model trained with the unprocessed data, which based its predictions only on three features and had the worst accuracy. However, the model trained with the partially processed data was still more biased in its results than the model trained with the most processed data, emphasizing the influence of the grade of Calculus II above anything else and seeing some significance in the predictor presenting the number of gap years, the significance of which disappeared with the deletion of the students with inexplicable gaps in the study records.

The result analysis for individual students with Shapley values indicated that the model trained with the most processed data was able to give the most reasonable results of the feature influences for individual students. The model trained with the partially processed data emphasized too much the influence of the grade of Calculus II, and the model trained with the unprocessed data based its predictions solely on three features, which meant that its predictions appeared to be very arbitrary.

In conclusion, this thesis gave a concrete example of how a model's performance can be improved with proper data cleaning. It also showed that the negligence of proper data cleaning can lead to deteriorated accuracy and abstruse results.

6. REFERENCES

[1] Baker R. & Siemens G. (2014) Educational data mining and learning analytics. In: Keith R. (ed) The Cambridge Handbook of the Learning Sciences, Cambridge University Press, second edition, pp. 253–272. DOI: http://dx.doi.org/10.1017/CBO9781139519526.016.

[2] Baker R., Corbett A. & Koedinger K. (2004) Detecting student misuse of intelligent tutoring systems. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 3220, pp. 531–540.

[3] Romero C. & Ventura S. (2010) Educational data mining: A review of the state of the art. IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews 40(6), pp. 601–618. DOI: http://dx.doi.org/10.1109/TSMCC.2010.2053532.

[4] Rudian S., Lui Z. & Pinkwart N. (2019) Comparison and prospect of two heaven approaches: SVM and ANN for identifying students' learning performance. In: Proceedings - 2018 7th International Conference of Educational Innovation through Technology, Auckland, New Zealand, pp. 156–161. DOI: http://dx.doi.org/10.1109/EITT.2018.00038.

[5] Kumar M., Singh A. & Handa D. (2017) Literature Survey on Student's Performance Prediction in Education using Data Mining Techniques. International Journal of Education and Management Engineering 7(6), pp. 40–49. DOI: http://dx.doi.org/10.5815/ijeme.2017.06.05.

[6] Francis B. & Babu S. (2019) Predicting academic performance of students using a hybrid data mining approach. Journal of Medical Systems 43(6). DOI: http://dx.doi.org/10.1007/s10916-019-1295-4.

[7] Zaffar M., Hashmani M. & Savita K. (2018) Performance analysis of feature selection algorithm for educational data mining. In: 2017 IEEE Conference on Big Data and Analytics, Kuching, Malaysia, pp. 7–12. DOI: http://dx.doi.org/10.1109/ICBDAA.2017.8284099.

[8] Rachburee N. & Punlumjeak W. (2015) A comparison of feature selection approach between greedy, IG-ratio, Chi-square, and mRMR in educational mining. In: Proceedings - 2015 7th International Conference on Information Technology and Electrical Engineering: Envisioning the Trend of Computer, Information and Engineering, Chiang Mai, Thailand, pp. 420–424. DOI: http://dx.doi.org/10.1109/ICITEED.2015.7408983.

[9] Gonzalez Zelaya, C. (2019) Towards explaining the effects of data preprocessing on machine learning. In: Proceedings - International Conference on Data Engineering, Macau, China, pp. 2086–2090. DOI: http://dx.doi.org/10.1109/ICDE.2019.00245.

[10] Lin, WC. & Tsai, CF. (2019) Missing value imputation: a review and analysis of the literature (2006–2017). Artificial Intelligence Review DOI: http://dx.doi.org/10.1007/s10462-019-09709-4.

[11] Breiman L. (2001) Random Forests. Machine Learning 45(1), pp. 5–32. DOI: http://dx.doi.org/10.1023/A:1010933404324.

[12] Friedman J. (2001) Greedy function approximation: A gradient boosting machine. Annals of Statistics 29(5), pp. 1189–1232.

[13] Elith J., Leathwick J. & Hastie T. (2008) A working guide to boosted regression trees. Journal of Animal Ecology 77(4), pp. 802–813. DOI: http://dx.doi.org/10.1111/j.1365-2656.2008.01390.x.

[14] Natekin A. & Knoll A. (2013) Gradient boosting machines, a tutorial. Frontiers in Neurorobotics 7. DOI: http://dx.doi.org/10.3389/fnbot.2013.00021.

[15] Hastie T., Tibshirani R. & Friedman J. (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Series in Statistics, Springer New York, second edition. DOI: http://dx.doi.org/10.1007/978-0-387-84858-7.

[16] Friedman J. (2002) Stochastic gradient boosting. Computational Statistics and Data Analysis 38(4), pp. 367–378. DOI: http://dx.doi.org/10.1016/S0167-9473(01)00065-2.

[17] Kotsiantis S., Kanellopoulos D. & Pintelas P. (2006) Data preprocessing for supervised learning. International Journal of Computer Science 1(1), pp. 111–117.

[18] Molnar C. (2019) Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. `https://christophm.github.io/interpretable-ml-book/`. Accessed August 21, 2019.