PREDICTING HIGHER EDUCATION COMPLETION: COMPARING LOGISTIC REGRESSION AND MACHINE LEARNING ALGORITHMS

WHO WILL GET HIGHER EDUCATION BY THE AGE OF 25?

MIKHAIL BOGDANOV

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

2012960

COMMITTEE

Chris Emmery
dr. Giovanni Cassani

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

June 16, 2022

PREDICTING HIGHER EDUCATION COMPLETION: COMPARING LOGISTIC REGRESSION AND MACHINE LEARNING ALGORITHMS

WHO WILL GET HIGHER EDUCATION BY THE AGE OF 25?

MIKHAIL BOGDANOV

8236 words

Abstract

The role of higher education in social inequality and inequality in access to higher education has been extensively studied in Social Sciences. The large corpus of the studies discovered various factors and their interplay in shaping the educational trajectories of youth. However, it is still unclear to what extent can we predict who will get higher education based on the educational and background information from teenage years. Moreover, it is not yet known what type of models produce better predictive performance in this and similar matters. Previous studies in adjacent domains (life-course studies) have shown that machine learning algorithms can slightly outperform classic statistical models. However, the understanding of what predictors these algorithms rely on and how it compares to well-established statistical models is still opaque. In this study, we made an attempt to answer these questions by predicting higher education completion with fine-tuned machine learning algorithms and comparing it to the classic statistical model – Logistic Regression. We used the data (N = *3743*) from the longitudinal cohort survey "Trajectories of Education and Careers" (TrEC) which is based on the comparative educational study of eighth-graders – "Trends in International Mathematics and Science Study" (TIMSS-2011). The results demonstrate that higher education completion could be predicted relatively accurately (accuracy = 0.72-0.73) from the educational and background information from teenage years. Machine learning algorithms (Random Forest, XGBoost, and Elastic Net) slightly outperform Logistic Regression. The predictors' importance derived from the best-performing machine learning model (Random Forest) and Logistic Regression are moderately correlated but have some intriguing disparities. However,

the predictors' importance of Logistic Regression suffers from multicollinearity in our model specification. We address this problem and suggest possible solutions in the Limitations sections.

1 INTRODUCTION

The *research goal* of this study is to reveal how accurately it is possible to predict who will get higher education based on the socioeconomic background, family composition, school characteristics as well as educational results, aspirations, and attitudes in teenage years.

1.1 *Problem Statement*

In contemporary society, higher education plays a crucial role in intergenerational social mobility. Throughout the history of social sciences, the role of higher education in shaping and maintaining social inequality has been extensively investigated in different contexts and from different perspectives (Ballarino, Bernardi, Requena, & Schadee, 2009; Boudon, 1974; Brown, 2018; Dickert-Conlin & Rubenstien, 2007; Duta, Wielgoszewska, & Iannelli, 2021; Hartas, 2015; Triventi, 2013). Studies revealed the relationship between multiple variables such as educational attainment in school, aspirations, socioeconomic status of the family, school characteristics (and many others), and chances of getting a higher education (Bernardi & Cebolla, 2014; Jerrim, Chmielewski, & Parker, 2015; Morgan, 2012; Yastrebov, Kosyakova, & Kurakin, 2018).

However, most of these studies had explanatory rather than predictive nature (Shmueli, 2010). In other words, the research questions of these studies concentrated on revealing the relationship between various variables and chances of getting higher education rather than assessing the predictive power of these models in predicting who will actually get higher education (Shmueli, 2010). Despite the vast amount of studies that scrutinize complex interactions between socioeconomic background and academic results in education inequality with sophisticated explanatory models, we still do not know how accurately we can predict who will get higher education based on the data from earlier life stages.

Meanwhile, in the last years, more and more social scientists urge to raise research questions of predictive nature, arguing that this will make social research more policy-oriented (Cranmer & Desmarais, 2017; Hofman, Sharma, & Watts, 2017; Hofman et al., 2021; Molina & Garip, 2019; Risi, Sharma, Shah, Connelly, & Watts, 2019; Verhagen, 2022; Yarkoni & Westfall, 2017). Recently, several studies attempted to predict various life outcomes

and latent characteristics of individuals (Joel et al., 2020; Li, Han, Cohen, & Markus, 2021; Salganik et al., 2020).

Nevertheless, it is still unclear to what extent we can predict who will get higher education. Thus, this study will fill this niche by assessing the predictive power of the data about educational results, socioeconomic background, educational aspirations, and attitudes in teenage years in predicting higher education completion by the age of 25. The choice of the age of 25 is driven by the available data. However, it is also reasonable to assume that most people from the cohort who will get higher education during their lifetime will get it by the age of 25. Moreover, research based on the data of older cohorts would have been outdated and could have been less relevant for the current patterns and relations in the investigated phenomena.

1.2 *Research Questions*

Thus, the broad research question of the study is:

> *Given educational and background information from teenage years, how accurately can we predict higher education completion by the age of 25 with machine learning algorithms and classic statistical model, that is, logistic regression?*

This research question could be expanded in the following research sub-questions:

RQ1 *To what extent can we predict higher education completion by the age of 25 based on the educational and background information from teenage years?*

RQ2 *Do the machine learning algorithms outperform classic statistical model, that is, logistic regression in predicting higher education completion?*

RQ3 *What are the differences in predictors' importance between classic statistical model, that is, logistic regression and best-performing machine learning algorithm in predicting higher education completion?*

1.3 *Social and Scientific Relevance*

The proposed research question is important both socially and scientifically for the following reasons. First of all, predictive models of higher education completion on the individual level could be used for more tailored and targeted interventions (Salganik, Lundberg, Kindel, & McLanahan, 2019, p. 1). Secondly, changing focus from explanatory to predictive modeling in

social research could advance social theories by offering novel methods of testing hypotheses (Molina & Garip, 2019; Salganik et al., 2019, p. 1).

2 LITERATURE REVIEW

The research question and design of this research are partly inspired by the aforementioned study by Salganik et al. (2020). The literature review is divided into two subsections. The first one, *predictive modeling in social sciences*, is devoted to reviewing studies in social sciences which were conducted in the framework of predictive modeling while the second, *Educational Studies*, specifically covers theoretical frameworks and empirical studies in the field of sociology of education.

2.1 *Predictive Modeling in Social Sciences*

In the aforementioned study by Salganik et al. (2020), more than one hundred research teams around the world tried to predict various individuals' life outcomes in the format of a machine learning challenge. The study was based on the Fragile Families and Child Wellbeing Study (FFCWS) – a high-quality longitudinal survey of one birth cohort in the USA. In this challenge, teams used all available variables about the socioeconomic background, health, wealth, education, family relationships, and various attitudes measured at ages 1, 3, 5, 9, and 15 to predict six outcomes at age 15: material hardship of the household, GPA, grit, household eviction, participation in job training by the primary caregiver, and caregiver layoff (Salganik et al., 2020, p. 8399). The performance was assessed on the hold-out test data set. The results indicate relatively low predictability of life outcomes at the age of 15 despite the vast amount of available information about the respondents and usage of the broad range of different machine learning algorithms and approaches to data preprocessing (Salganik et al., 2020, 2019). For example, the best performing model for the target variable GPA had R^2 of 0.19 (Salganik et al., 2020). Moreover, the classic statistical models, that is, linear and logistic models, performed only slightly worse than sophisticated fine-tuned machine learning algorithms (Ahearn & Brand, 2019; Salganik et al., 2020, 2019).

Nevertheless, this challenge resulted in several papers describing the best approaches and nuances of data wrangling and predictive modeling based on the longitudinal survey data (Altschul, 2019; Compton, 2019; McKay, 2019; Raes, 2019; Rigobon et al., 2019). One of the best performing models for predicting GPA was gradient boosting as demonstrated by Raes (2019). However, regularized regression models (LASSO, Ridge, Elastic Net) also performed comparably well (Raes, 2019). The paper Rigobon et

al. (2019) describes data preprocessing, feature engineering, and selection as well as winning models for predicting GPA, grit, and layoff of primary caregiver. At first, the authors removed features with low variance and a high fraction of missing values (>80%). Then, they created indicator variables for missing values in each variable. As for the feature selection, the authors tried two approaches: mutual information criterion and LASSO (Rigobon et al., 2019). For the prediction, the authors trained and tuned three types of models: *Elastic Net*, *Random Forest*, and *Gradient Boosting (XGBoost)*. Moreover, they also tried to stack these models. However, for GPA and grit, the best performing model on the hold-out set was *XGBoost*.

In other fields, such as political science, available predictive studies also demonstrate relatively modest predictive capabilities. For example, it was shown in Bach et al. (2021) that the voting behavior could be hardly predicted from the digital traces on the Internet. In this study, the authors employed Gradient Boosting (*XGBoost*) as well as *Random Forests* on the data from mobile device usage and web browsing to predict political preferences and self-reported voting behavior in Germany during elections in 2017 (Bach et al., 2021). The *ROC-AUC* performance measure of the best model varied between 0.6-0.7 for different voting outcomes (Bach et al., 2021, p. 873).

In the Human Resources field, there was an attempt to employ machine learning algorithms and interpretability methods to uncover the relationship between personality traits and leadership (Doornenbal, Spisak, & van der Laken, 2021). Moreover, this study compares the predictor importance of standard linear regression and fine-tuned tree-based machine learning algorithm – *Random Forest*, which is relevant for RQ3 of our research. Random Forest slightly outperformed the linear model with sensitivity on the test set equal to 34.1% versus 27.9%. As for the feature importance, the authors have used a perturbation-based approach which constitutes of adding random noise to values of the particular variable and reassessing the drop in performance of the already estimated model after this (Doornenbal et al., 2021, p. 6). Although, the predictor importance derived from Random Forest and standardized coefficients from Linear Regression mostly aligned it also demonstrated some differences between the two models.

Overall, based on the literature review the following machine learning algorithms were chosen for this research as the most promising: XGboost, Random Forest, and Elastic Net.

2.2 *Educational Studies*

As for the higher education domain, there were a number of studies dedicated to predicting higher education dropout and performance (Nagy & Molontay, 2018; Niessen, Meijer, & Tendeiro, 2016). The systematic literature review of studies on predictors of higher education dropout revealed that previous academic results and educational goals are consistently related to higher education dropout across various countries, contexts, and times (Delnoij, Dirkx, Janssen, & Martens, 2020). Another systematic review of empirical studies demonstrated that the most important factors in predicting students' higher education success are academic performance in school, socio-demographic characteristics, and students' environment (Alyahyan & Düştegör, 2020). However, most of these studies investigated students who have already enrolled in higher education institutions while my study focuses on predicting which schoolchildren will get higher education, thus, considering both enrollment and completion of higher education.

The studies in the sphere of social inequality in education have constantly revealed that the socioeconomic background of the family is an important factor in the educational trajectory of the children even when we control for the academic results (Ballarino et al., 2009; Bernardi & Boado, 2014; Bernardi & Cebolla, 2014; Jerrim et al., 2015; Lucas, 2001; Morgan, 2012; Simonová & Soukup, 2015; Yastrebov et al., 2018).

Nonetheless, the majority of the reviewed studies answered explanatory rather than predictive research questions. Moreover, even though some of these studies collaterally reported predictive performance measures, they were not the primary focus of the papers and usually were assessed using in-sample data. For example, in Simonová and Soukup (2015) it is reported that the model has an accuracy of 90% in predicting transition to higher education. However, the predictive performance was not tested on the hold-out data set nor the limitations of the reported metric were discussed.

In the recent paper by Verhagen (2021), the author used multilevel regression models (MLM) to predict the educational track assigned to the Dutch schoolchildren at the age of 12 by their teachers. However, in contrast to other studies in this field, this paper concentrated on the predictive rather than an explanatory aspect of the MLM which led to the discovery that school effects play an important role in predicting educational tracking by teachers in the Netherlands (Verhagen, 2021). Moreover, adding variability in intercepts on the school level led to a greater increase in predictive performance than adding parental education variables in the models. Thus, we can anticipate that school characteristics and school-level variables might be important predictors of higher education completion as well.

Nevertheless, to my best knowledge, there are no studies regarding the prediction of higher education completion based on the data from individuals' earlier life stages such as teenage years. Thus, this study will fill this niche by estimating how accurately we can predict who will get higher education based on the data from teenage years and by comparing the classic logistic model with fine-tuned machine learning algorithms.

3 METHODOLOGY

3.1 *Dataset Description*

This study is based on the data from the longitudinal cohort study "Trajectories in Education and Careers" – TrEC. TrEC, in its turn, is based on the nationally representative sample (N = 4893) of the cohort of Russian eighth graders of 2011 who participated in the international comparative educational study "Trends in International Mathematics and Science Study" (TIMSS-2011). This study included a set of tests on mathematics, and science as well as a questionnaire about students' attitudes towards school, career aspirations, socioeconomic background, and other contextual factors. Moreover, the study included questionnaires for teachers and school administration that covered such topics as educational practices in the school, bullying, school characteristics, school performance, and other variables. The participants of TIMSS-2011 have been surveyed annually ever since in the TrEC study. New waves of TrEC are conducted every year and cover a broad range of topics from education and employment to family composition and attitudes towards various social phenomena. The precise set of questions could vary from wave to wave but the main core of the questionnaire includes questions regarding education and job occupation (Kurakin, 2014; Malik, 2019). For more details about the design, aim, and history of TrEC see Kurakin (2014); Malik (2019). This dataset is available for researchers upon request. In this research, we will use information about the educational status of the respondents from the last available, 9^{th}, wave of the TrEC (N = 3743). At the moment of 9^{th} wave, the modal age of the cohort was 25.

3.2 *Data Preprocessing*

This section describes the data cleaning and preprocessing pipeline. At first, the TIMSS-2011 and 9^{th} wave of TrEC databases were merged using the student identification variable resulting in a dataset of 3743 observations. Then, we selected variables from TIMSS-2011 related to demographic characteristics, socioeconomic background of the family, educational as-

pirations, attitudes and results, family attitudes towards education, and school characteristics. The selection of the variables was done by filtering out variables that are clearly unrelated to these domains: socioeconomic background, family composition, school characteristics as well as educational results, aspirations, and attitudes. The limitations of this approach are discussed in the Section 5.2. The final dataset (before preprocessing) comprised 91 variables: 1 ID variable, 1 target variable and 89 predictors (the full list of variables is available in Appendix A).

3.2.1 *Variable Constructing and Transformations*

Constructed predictors:

- The "migration status" variable was constructed by coalescing two dependent survey questions about whether the person was born in the country and, if not, at what age they came to the country.
- TIMSS Mathematics and Science scores were computed as arithmetic averages of the respective variables representing 1^{st}-5^{th} plausible values for mathematics and science achievement test scores as proposed in Martin and Mullis (2011).

The target variable was constructed based on the question about the highest completed level of education in the 9^{th} wave of TrEC. Those who answered that they have either bachelor's or master's or specialist degrees were assigned to the category of those who completed higher education while those who selected other options were assigned to the category of those who did not complete higher education.

Table 1: Correspondence of constructed target variable "Higher Education completion" values with values from the initial variable.

Initial values	Higher Education Completion
9 classes of school	Do not have higher education
11 classes of school	Do not have higher education
Secondary vocational education	Do not have higher education
Bachelor's	Have higher education
Specialist (university degree of 5-6 years length)	Have higher education
Master's	Have higher education
Other	Do not have higher education

The distribution of constructed variables about higher education completion is roughly balanced with 49.8% completed higher education and 50.2% did not (see Table 2).

Table 2: Distribution of the target variable – Higher Education completion.

Higher Education completion	*N*	%
Do not have higher education	1878	50.2
Have higher education	1865	49.8
Total	3743	100

3.2.2 *Removing Variables with Missing Values*

Then, we removed all predictors that contained more than 30% of missing values. These included the following variables:

1. "bsbgsls" – student like learning science scale.
2. "bsbgsvs" – student value learning science scale.
3. "bsbgsvs" – student confidence with science scale.
4. "bsbgesl" – student engaged in science lessons scale.

For some reason, all removed variables were related to the "Science" school subject. After investigation, it turned out that these variables are fully empty and do not contain any real values. One of the possible reasons for this is that these scales are available in the more detailed partition by particular scientific subjects: chemistry, biology, physics, and earth science. The plausible reason for this is that in Russia there is no single subject for Science in the academic curriculum of eight-graders but rather four distinct subjects by subfields. Thus, the removed variables contained 100% of the missing values.

3.2.3 *Imputing Missing Values with kNN*

Missing values within the remaining predictors were imputed with *the k-Nearest Neighbours* algorithm. *k-Nearest Neighbours* is an algorithm that calculates a distance between each observation in the space defined by some set of predictors (or all of them) and imputes missing values based on the top-k matching instances in this space. In the case of continuous missing values, the mean value of the respective variable of k most similar instances is used for imputation while in the case of a nominal variable – the mode. In this study, k was set to 5 as it was suggested as a practical default value in the literature (Kuhn & Johnson, 2019). Thus, the imputation of the particular missing value was based on the 5 most similar respondents derived based on the whole set of all available predictors.

One advantage of using kNN for imputation is that it will produce values that are within the observed in-sample range of values for the

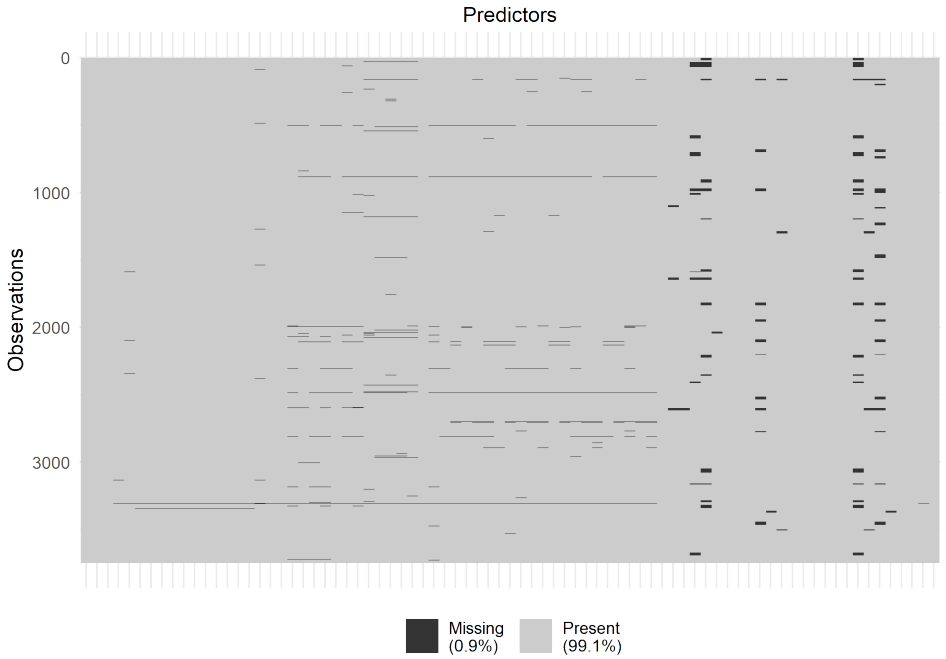

Figure 1: Missing values in the data.

particular variable. Thus, in contrast to, for instance, linear models, it is impossible to get unrealistically low or high values for continuous variables. For example, kNN could not propose negative values for the weight variable or zero for the human height. Moreover, since the share of the missing values in the dataset in general and in most variables, in particular, is neglectable low (see Figure 1) it is unlikely that the choice of the imputation method will substantially affect the results of the analysis. Moreover, the imputation procedure will be conducted separately for the train and test data as well as within folds during cross-validation. Thus, we avoid the problem of information leakage and thoroughly mimic the process of testing models on the fully unseen hold-out data.

3.2.4 *Categorical Predictors Transformation*

In this step, infrequent categories of each nominal variable that constitutes less than 5% of the training data were combined into one artificial category called "other". This step affected 18 nominal predictors out of 42 present in the data.

Then, all categorical predictors were transformed into a set of dummy variables (also known as one-hot encoding in the machine learning community) with the removal of one dummy variable for each categorical predictor to avoid linear dependency between these newly constructed

variables. This approach was implemented only for machine learning algorithms that require a transformation of categorical predictors into a set of dummy variables: *XGBoost* and *Elastic Net*. For other algorithms, namely Random Forest and Logistic Regression, we used categorical features without transformations into dummy format. However, we experimented with dummy variables and Random Forest and Logistic Regression algorithms and received similar results. This is expected because dummy variables are just another way of representing categorical features. It does not produce or extract more information or signal from the data.

3.2.5 *Near-Zero Variance Filter*

At this stage, variables with near-zero variance were removed from the data since they are not informative for the models and, thus, do not bring any added value to predictive performance. Moreover, the presence of non-informative predictors could worsen the performance of some types of models (especially linear models) and increase computational expenses and, consequently, training time (Boehmke & Greenwell, 2019, p. 52). The near-zero variance filtering was done based on the two criteria:

1. Share of the number of unique values in the variable relative to the sample size. This parameter was set to 10 as suggested in Boehmke and Greenwell (2019).

2. "The ratio of the frequency of the most prevalent value to the frequency of the second most prevalent value" (Boehmke & Greenwell, 2019, p. 52-53). This parameter was set to 97/3 as suggested in Boehmke and Greenwell (2019).

This filtering method removed 8 variables.

3.2.6 *Continuous Predictors Normalization*

All continuous predictors, including newly created dummy variables, were normalized (z-standardized), that is, centered and scaled. By centering, it implies subtracting the training set mean value of the particular variable from each value of this variable. Scaling is dividing centered values on the standard deviation of this variable calculated on the training set. This transformation converts all variables to the same measurement scale with mean = 0 and standard deviation = 1. This might be particularly important for specific machine learning algorithms, such as regularized regressions (Elastic Net in our case) (Boehmke & Greenwell, 2019, p. 57). It is important to note that necessary statistics for normalization, mean and standard deviation, were calculated on the training set and for each

training fold inside cross-validation. This prevents data leakage from the test data into the modeling process.

3.2.7 *PCA for Mathematics and Science achievement scores*

As was shown in the literature and previous studies (see Section 2), educational achievement is a strong predictor of the educational trajectory of children. Thus, it might be convenient to have a single variable representing general academic results. This would also ease the interpretation of this variable in further analysis. Moreover, some machine learning methods tailored to regularization might exclude some of the highly correlated predictors while other methods will not. This would potentially cause problems with the interpretation of predictors' importance.

As could be observed from the scatter plot in Figure 2, there is a strong linear relationship between Mathematics and Science achievement scores (Pearson's $r = 0.86$, $p = 0.00$).

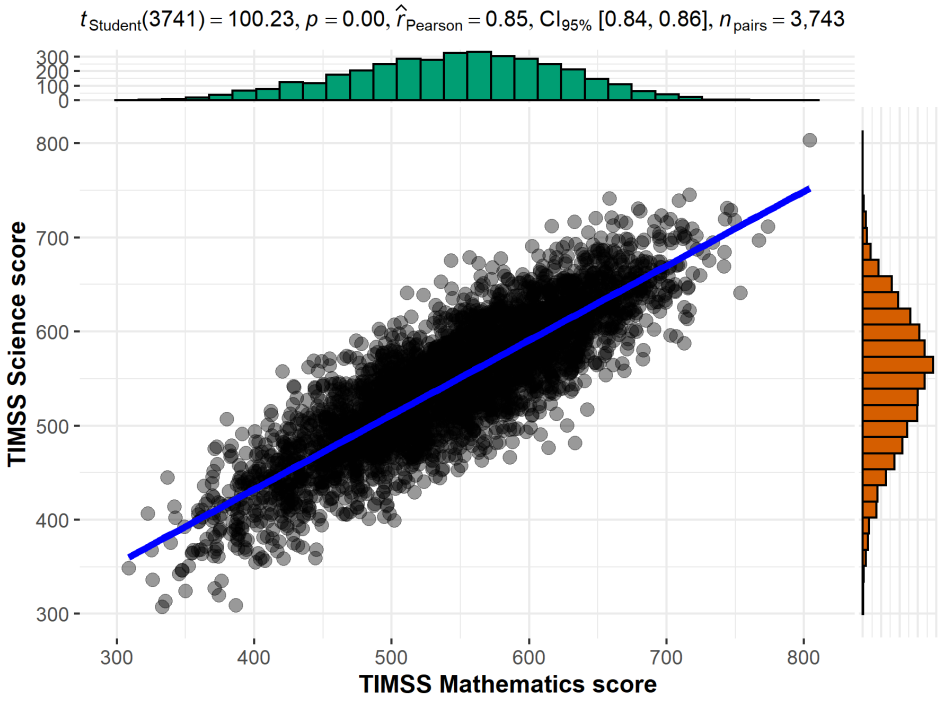

Figure 2: Scatter plot and Pearson correlation coefficient of TIMSS Mathematics and Science scores on the training data.

Therefore, we exploited this relationship by projecting these two variables onto a single dimension using Principal Component Analysis (PCA). Since there are only two highly correlated variables, it is meaningful to

extract one component that would represent general educational achievement based on the Mathematics and Science achievement scores measured in TIMSS.

As could be seen from the Figure 3, the first component explains almost 93% of the variance in the Mathematics and Science achievement scores. In other words, by knowing the values of this component we can reconstruct initial achievement scores in Mathematics and Science with high accuracy.

As well as with other performed preprocessing steps, PCA was also trained only using the training data and/or training folds inside the cross-validation to avoid potential data leakage (Kuhn & Johnson, 2019). Thus, when assessing model performance on the hold-out test data, the extraction of principal components for the test data was done using the pretrained PCA model on the train data.

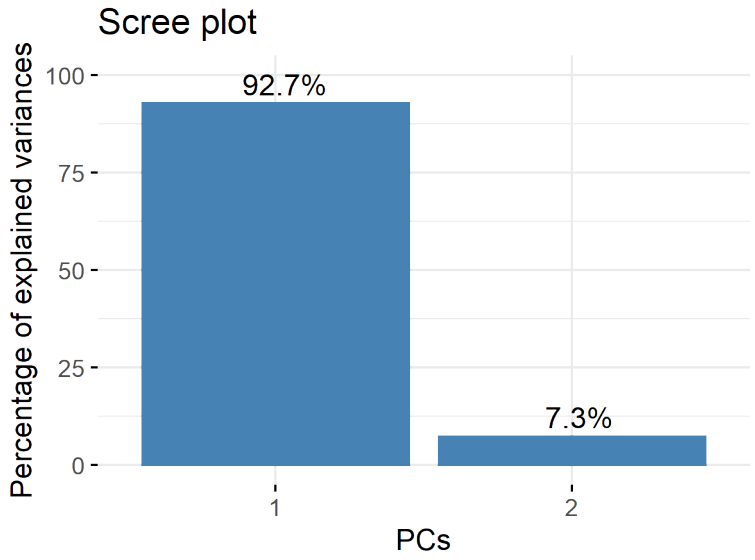

Figure 3: % of explained variance by Principal Components.

3.2.8 *Correlation and Linear Combination Filters*

At this stage, we filtered out variables that were either linearly dependent or highly correlated. In other words, if one variable could be perfectly reconstructed from the others or it highly correlates with others. The threshold for the correlation filter was set to 0.95 to remove only those variables that correlate so strongly that it is reasonable to assume that they might measure the same phenomenon. The application of these filters resulted in the removal of the scale representing the degree to which instruction at schools is affected by mathematics resource shortages.

3.3 *Algorithms*

To answer the main research question, that is, how accurately we can predict who will obtain higher education, we will employ the following machine learning algorithms: *XGboost*, *Random Forest*, *Elastic Net*, and *Logistic regression*. The choice of gradient boosting (*XGboost*) is driven by the fact they have shown superior performance on the tabular data (Ivanov & Prokhorenkova, 2021; Prokhorenkova, Gusev, Vorobev, Dorogush, & Gulin, 2018; Shwartz-Ziv & Armon, 2022). *ElasticNet* and *Random Forests*, as it was shown in section 2.1, have been employed in similar studies and have shown relatively good performance (Ahearn & Brand, 2019; Raes, 2019).

3.3.1 *Random Forest*

Random Forest is a machine learning algorithm that is based on the ensemble of classification trees. There are different implementations of the Random Forest algorithm that was first introduced by Breiman (2001). However, in this study, we describe and employ the version implemented in the *ranger* package in R (Wright & Ziegler, 2017). In more detail, at first, the algorithm takes a sample of the instances and a sample of variables and builds the decision tree on this subset of the data. Usually, the decision trees in Random Forests are grown based on the bootstrapped sample of instances of the original data size. However, in contrast to canonical decision trees, in the Random Forest at each node the algorithm makes a split based on the subsample of predictors of prespecified size rather than considering all variables for splitting.

The random sampling of observations for tree growth and predictors for splitting the nodes inside the trees allows for growing the forest of the *de-correlated* and *independent* trees which usually leads to better predictive performance and good out-of-sample generalization (Boehmke & Greenwell, 2019, p. 203).

The maximum depth of the decision trees could be preset beforehand or could be explicitly constrained by the preset minimum number of instances in the nodes at which the further splitting of these nodes stops. Thus, a minimum number of instances partially controls overfitting by constraining the depth of individual trees in the ensemble.

In this study, we tuned the following hyperparameters of the Random Forest algorithm:

- Number of trees in the ensemble: 800, 1000, 1500.
- Number of sampled variables that will be considered for splitting the nodes: 10, 15, 20.

- Minimum number of instances in the nodes (odd numbers were chosen so that at each terminal node it would be possible for a tree to determine the class by majority vote principle): 3, 7, 11, 15.

The grid of hyperparameters' values was defined based on the previous research which used Random Forests on similar survey data (Bach et al., 2021; Doornenbal et al., 2021; Rigobon et al., 2019) and suggested default values for tuning in (Boehmke & Greenwell, 2019, p. 206).

3.3.2 *Gradient Boosting (XGBoost)*

Gradient boosting is a machine learning algorithm that is based on the principle of sequentially training simple models taking into account the mistakes of previous models (Boehmke & Greenwell, 2019, p. 221). If *Random Forest* is an ensemble of independent and large trees, *Gradient Boosting* is an ensemble of small trees that are trained one after another on the mistakes of previous trees.

There are several variations of *gradient boosting* algorithms. For example, *CatBoost* and *LightGBM* are quite common in contemporary machine learning applications (Ivanov & Prokhorenkova, 2021; Prokhorenkova et al., 2018; Shwartz-Ziv & Armon, 2022). However, in this study, we use XGBoost as it has shown good performance in similar studies that were also based on the survey data of comparable size and nature (Raes, 2019; Rigobon et al., 2019).

We employed the *maximum entropy approach* to create the grid of the hyperparameters for XGBoost. The maximum entropy approach is trying to fill the space of the potential hyperparameter values such that it has low overlapping and is evenly distributed across the hyperparameter space (Kuhn & Silge, 2020). The size of the grid was set to 50. Thus, the grid has 50 unique combinations of hyperparameters' values. Table 3 describes hyperparameters of *XGBoost* that are available in *tidymodels* framework in R and were tuned during the grid search.

Table 3: Descriptive statistics of hyperparameters' values for XGBoost created by maximum entropy approach.

Hyperparameters	Unique values	Min	Q1	Mean	Median	Q3	Max
Number of trees	50	16	508.75	930.32	1001	1333.5	1955
Min N in nodes	50	5	14.25	22.96	22.5	32	40
Tree depth	50	1	3	7.72	8	12	15
Learning rate	50	0.0011	0.0026	0.049	0.023	0.077	0.24
Loss reduction	50	0.0000000001	0.0000001	0.693	0.00002	0.032	20.08
Sample size	50	0.1356	0.3546	0.581	0.590	0.829	0.97
Number of iterations	50	3	6	10.92	10.5	15	20

3.3.3 *Elastic Net*

Elastic Net is an extension of Linear or Logistic Regressions specifically tailored to the improvement of out-of-sample predictive performance by imposing regularization to the loss function during the training of the model. It is a mixture of Ridge regularization ($\lambda \sum_{i=1}^{p} \beta_i^2$) and Lasso regularization ($\lambda \sum_{i=1}^{p} |\beta_i|$) that are imposed on the loss function during the training where β_i is the coefficient for predictor i in the set of predictors p (Boehmke & Greenwell, 2019, p. 126) . There are two hyperparameters that can be tuned in *Elastic Net*: *mixture* and *penalty*. The *mixture* is a degree of balance between Ridge and Lasso regularization. The *penalty* is a magnitude of these regularizations (proportional to the balance between Ridge and Lasso defined by *mixture*) – λ.

Table 4 presents descriptive statistics of the hyperparameters of Elastic Net that were tuned using cross-validation. This grid of hyperparameters was also constructed using a *maximum entropy approach* and a plausible range of values for these hyperparameters available in *tidymodels* framework in R.

Table 4: Descriptive statistics of hyperparameters' values for Elastic Net grid created by maximum entropy approach.

Hyperparameters	Unique values	Min	Q1	Mean	Median	Q3	Max
Penalty	20	1.136E-10	2.232E-08	0.080	1.79E-05	0.002	0.831
Mixture	20	0.061	0.241	0.509	0.527	0.735	0.996

3.3.4 *Predictors' Importance*

We will employ *permutation-based feature importance tests* to substantively compare the importance of the predictors in the classic interpretable algorithm, that is, logistic regression, with the best-performing machine learning algorithm.

The *permutation-based feature importance* work as follows (Biecek & Burzykowski, 2021; Molnar, 2020):

1. Calculate performance metric for the model.
2. Shuffle values within the predictor or set of predictors.
3. Feed the dataset with permuted predictor(s) to the pretrained model and produce predictions.
4. Calculate performance metric for the predictions from the previous step.

5. Feature importance for the predictor(s) is the difference or ratio between the original performance of the model and performance on the data with permuted predictor(s)

The visual schema and toy example of permutation feature importance are presented in the Figure 4.

The idea behind permutation feature importance is that it eliminates all relationships of the particular feature in the dataset, both with the target variable as well as with other features, thus, also removing potential interaction effects (Molnar, 2020). One of the main advantages of permutation feature importance is that it is model-agnostic. In other words, this method could be applied to models of various types (Biecek & Burzykowski, 2021; Molnar, 2020). Despite the fact that Logistic Regression belongs to the class of interpretable models and its' coefficients have straightforward interpretation, we applied permutation feature importance to it as well for the sake of direct comparison with less interpretable algorithms. Moreover, the permutation feature importance allows to directly interpret features in terms of an impact on the predictive power of the model. Another advantage of permutation feature importance is that it is agnostic to variables' measurement. It provides a single approach to measuring and comparing the importance of both categorical and continuous predictors.

In this study, we used a ratio of the model performance on the test data to the model performance on the data with the permuted feature. This approach allows direct comparison of features' importance between different models and interpretation of features' importance in terms of relative change to the original model's performance. For each feature 30 iterations of permutations were conducted and results were then averaged to get more robust estimates of importance.

3.4 *Training, Validation, and Testing Procedure*

80% of the data was used for training, development, and validation of the models while the remaining 20% was used to evaluate the out-of-sample predictive performance of the models. We employed 5-fold cross-validation to tune hyperparameters of machine learning algorithms and pick the best combination of them for each type of model. We then applied these models to the test data to determine the best one.

It is important to note that this setup "will produce slightly optimistic estimates of the performance in new holdout data" (Salganik et al., 2020, p. 8401-8402). However, this bias is assumed to be negligible and the same approach was used in the aforementioned machine learning challenge published in the *Proceedings of the National Academy of Sciences (PNAS)* (Salganik et al., 2020, p. 8401-8402). The same methodological approach

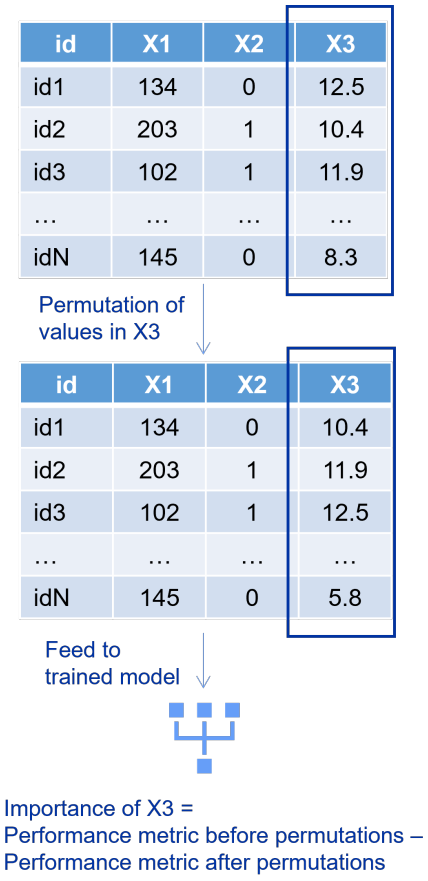

Figure 4: Schema of Permutation Feature Importance calculation.

to tuning, testing, and interpreting the best-performing model on the test data was practiced in the paper describing winning models for the Fragile Families Challenge by Rigobon et al. (2019).

All data splittings implied *stratification* by the target variable to enforce the similar distribution of the target variable in train and test subsets as well as between cross-validation folds. *Stratified* splitting is a variety of random splitting but the instances are sampled randomly within each class of the target variable (Kuhn & Johnson, 2019).

3.5 *Performance Metrics*

Since the distribution of the target variable is roughly balanced (49.8% vs 50.2%) and there is no substantive importance of one class over another in terms of the research questions, *accuracy* and *ROC-AUC* will be used to evaluate and compare the predictive power of the models. *Accuracy* is simply a share of correctly classified instances while *ROC-AUC* presents a more

complex metric that reflects the ability of the models to correctly classify instances using different thresholds for determining class belonging.

Usually, in classification models, the decision of whether a particular sample belongs to one class or another is determined based on the 0.5 threshold. If the estimated probability of a positive class for the set of observations is greater than 0.5 these observations are assigned to a positive class and if it is smaller than 0.5 they are assigned to the negative class. However, this threshold could be adjusted in order to find a compromise between True Positive Rate ($TP/(TP+FP)$) and False Positive Rate ($FP/(FP+TN)$). *ROC-AUC* measures True Positive and False Positive Rates at different thresholds, plot these values (*Receiver Operating Curve*) and calculates the area under this *ROC curve*. Thus, it is possible to compare the performance of classification models by comparing areas under respective *ROC curves*. *ROC-AUC* of the random classifier would be equal to 0.5.

Despite that both *accuracy* and *ROC-AUC* are reported in this study, *ROC-AUC* is the primary metric that hyperparameters were optimized for during grid search on cross-validation. It was also used to determine which model performs better on the test set as well as the primary metric for assessing the drop in performance for the permutation-based feature importance algorithm. *Accuracy* is also reported, however, the primary intent for this is an ease of interpretation of the results and a concise and expressive answer to the question "How accurately can we predict who will get higher education?". The naive baselines for accuracy and ROC-AUC were set by the random guessing principle and are equal to 0.502 and 0.5 respectively.

3.6 *Programming Language and Frameworks*

All data preprocessing and modeling were conducted in R programming language (R version 4.1.2 (2021-11-01) by R Core Team (2021)). The *tidymodels* framework was employed for combining data preprocessing, feature engineering, hyperparameter tuning, model training, validation, and testing into a single workflow (Kuhn & Wickham, 2020). For feature importance analysis and visualization, we used the *DALEX* package since it provides built-in functionality for exploring and explaining machine learning models from *tidymodels* framework (Biecek, 2018).

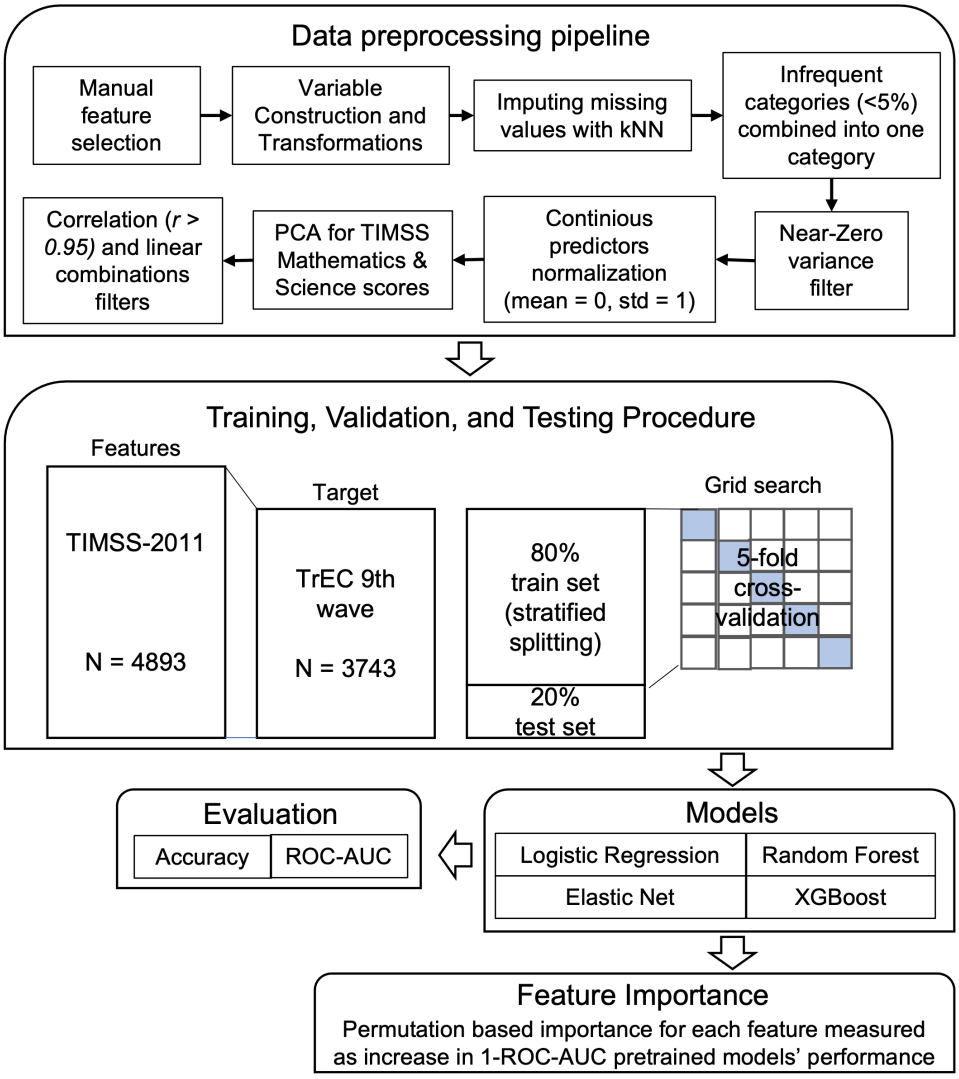

Figure 5: Methodology and modeling pipeline schema.

4 RESULTS

4.1 *Hyperparameter Tuning and Cross-Validation Results*

In this paragraph, we present and discuss the results of the hyperparameter tuning with grid search and cross-validation assessment of the predictive performance of the models. As could be seen from Table 5, the fine-tuned Random Forest model demonstrated the highest performance on cross-validation while Logistic Regression, expectedly, scored the lowest.

Table 5: Models' accuracy and ROC-AUC with best hyperparameters on 5-fold cross-validation. Standard errors calculated over 5-fold cross-validation are presented in the brackets.

Models	Accuracy (s.e.)	ROC-AUC (s.e.)
Elastic Net	0.733 (0.009)	0.804 (0.009)
XGBoost	0.728 (0.007)	0.800 (0.009)
Random Forest	0.737 (0.007)	0.799 (0.008)
Logistic Regression	0.728 (0.01)	0.798 (0.0116)
Baseline	0.502	0.5

4.1.1 *Random Forest*

As could be seen from Figure 6, tuning hyperparameters of Random Forest has not led to substantial gains in the predictive performance on cross-validation. The ROC-AUC values for different combinations of hyperparameters are spanned between 0.7965 and 0.7995. This was expected as Random Forests are known for the good out-of-box performance (Boehmke & Greenwell, 2019, p. 203). The best combination of hyperparameters with 800 trees, 10 randomly selected predictors, and a minimum node size of 11 achieved a ROC-AUC of 0.7995.

Figure 6: Grid search results for Random Forest.

4.1.2 *XGBoost*

The best performing combination of hyperparameters of XGBoost (see Table 6) achieved a ROC-AUC of 0.7998 on 5-fold cross-validation. Moreover, XGBoost demonstrated the largest spread in performance for different

hyperparameter combinations compared to other models – from 0.75 to almost 0.80.

Table 6: Best hyperparameters' values of XGBoost on 5-fold cross-validation.

Hyperparameters	Best values
Number of trees	1208
Min N in nodes	19
Tree depth	5
Learning rate	0.0039
Loss reduction	0.0002
Sample size	0.69
Number of iterations	20

Interestingly, we can see a strong reverse U-shaped dependency of ROC-AUC and log-transformed learning rate (see Figure 7). Moreover, XGBoost specifications with a higher proportion of sampled observations for tree growth on average demonstrated higher performance in terms of ROC-AUC. On the contrary, the models with larger minimum node sizes of the trees demonstrated lower performance. Both of these patterns might happen because of the relatively high dimensionality of the data (N/p ratio). Trees grown during gradient boosting might need more data to detect potential interactions in the multidimensional space.

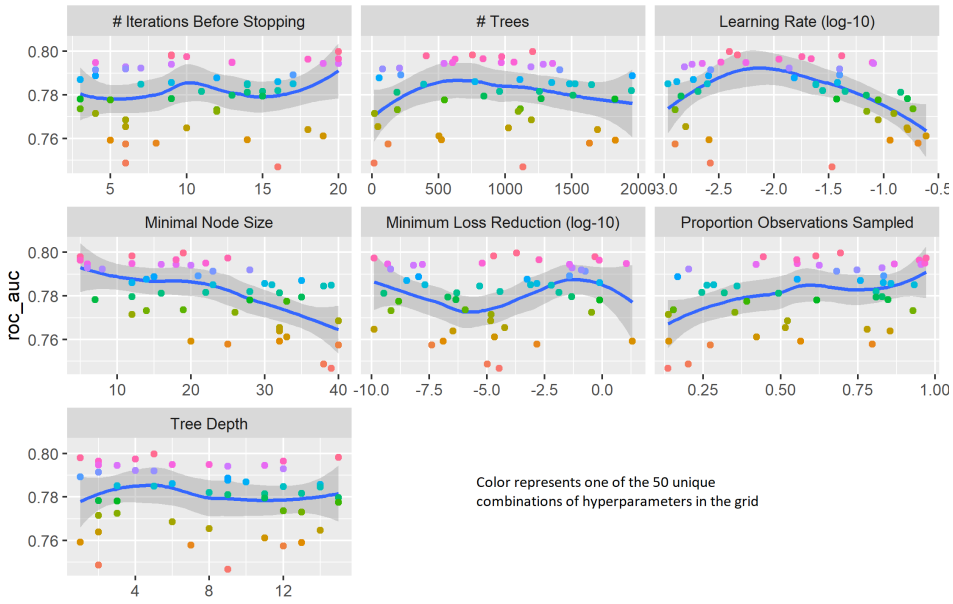

Figure 7: Grid search results for XGBoost. Lines represent locally estimated scatterplot smoothing (loess), and shaded areas around the lines are the 95% confidence intervals.

4.1.3 *Elastic Net*

Elastic Net demonstrated the best performance on cross-validation across all models. The best set of hyperparameters achieved a ROC-AUC of 0.8037 as could be seen from the Figure 8. Interestingly, the best Elastic Net specification is more inclined towards Ridge regularization since the mixture parameter is equal to 0.119. This could be explained by the fact that the Ridge penalty could better handle correlated features (Boehmke & Greenwell, 2019, p. 124). However, other hyperparameter combinations show almost the same performance.

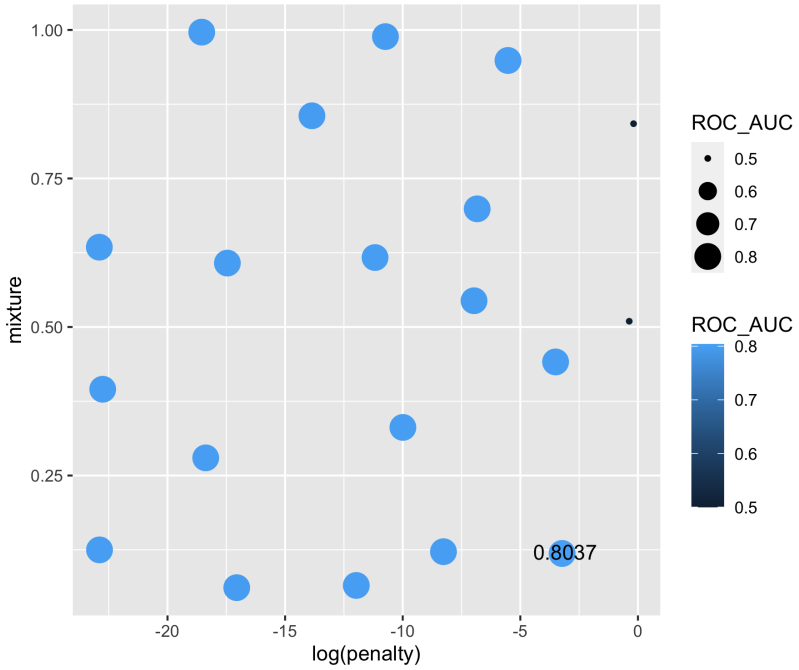

Figure 8: Grid search results for mixture and logarithm of penalty hyperparameters for Elastic Net.

4.1.4 *Logistic Regression*

Since Logistic Regression does not have any hyperparameters to tune its' performance was simply assessed on 5-fold cross-validation. It has demonstrated the lowest performance and the highest standard error estimated on 5-folds compared to other models – ROC-AUC = 0.798, std. error = 0.0116. The higher standard error of the performance metric might be an indicator of the greater degree of the model overfitting.

4.2 *Performance on the Test Data*

In this section, we report the performance of the fine-tuned models on the hold-out test data. After determining the best sets of hyperparameters for each model on the cross-validation, we retrained all models with the best hyperparameters on the training data. Data transformations and preprocessing described in the section 3.2 were also recalculated again using the whole training data.

All models demonstrated slightly worse performance than on the cross-validation (see Table 7, ROC curves are available in Appendix B). Nonetheless, all models substantially outperformed the baseline which was set based on the random guessing principle.

The best performing model on the test data is *Random Forest* with *a ROC-AUC* of 0.798. However, in terms of *accuracy*, the best model is *XGBoost* – 0.729. In other words, with XGBoost we can correctly classify who will get higher education in almost 73% of the cases.

Overall, *Elastic Net*, *Random Forest*, and *XGBoost* demonstrated comparable predictive performance and the differences are negligible. However, Logistic regression demonstrated the worst performance with *ROC-AUC* = 0.787 and *accuracy* = 0.717. Nevertheless, the difference in ROC-AUC between Logistic Regression and the best performing algorithm, that is Random Forest, is small. The ratio between the predictive performance of the best model and the worst model is $\frac{0.798}{0.787} = 1.014$.

Table 7: Performance of the models on the test data.

Model	Accuracy	ROC-AUC
Random Forest	0.725	0.798
XGBoost	0.729	0.793
Elastic Net	0.725	0.792
Logistic Regression	0.717	0.787
Baseline	0.501	0.5

4.3 *Predictors' Importance*

In this section, we explore the differences in features' importance in predicting higher education completion between the best-performing machine learning algorithm, that is, Random Forest and Logistic Regression. For this purpose, we plotted in Figures 9, 10 the top 10 most important predictors for each algorithm derived by the permutation feature importance approach introduced in the section 3.3.4.

There are two predictors that stand out in terms of impact on the predictive performance of the Logistic Regression, both of them measured on the school level: number of instructional days per week in school and type of settlement where the school is located. Next, with almost twice as small measure of predictive importance, is the educational aspirations or, in terms of the survey question, the maximum level of education the student plan to complete.

However, estimates of predictors' importance in Logistic Regression might be partly flawed because of the multicollinearity issue. Some predictors might apparently measure adjacent phenomena which could have provoked the issue of multicollinearity between them in the Logistic Regression model. We assessed the Generalized Variance Inflation Factor (GVIF) for all predictors in the Logistic Regression model and found out that there are predictors with GVIFs larger than 5 which indicates multicollinearity

issues for these predictors as suggested in Fox and Monette (1992) (see Figure 13 in Appendix B). Notably, there is an overlap between predictors with high GVIF values and those that are considered the most important by the features permutation algorithm. Both "number of instructional days per week in school" and "type of settlement of school location" are considered important predictors and at the same time have high GVIF values that are evidence of the multicollinearity issue for these predictors (Fox & Monette, 1992).

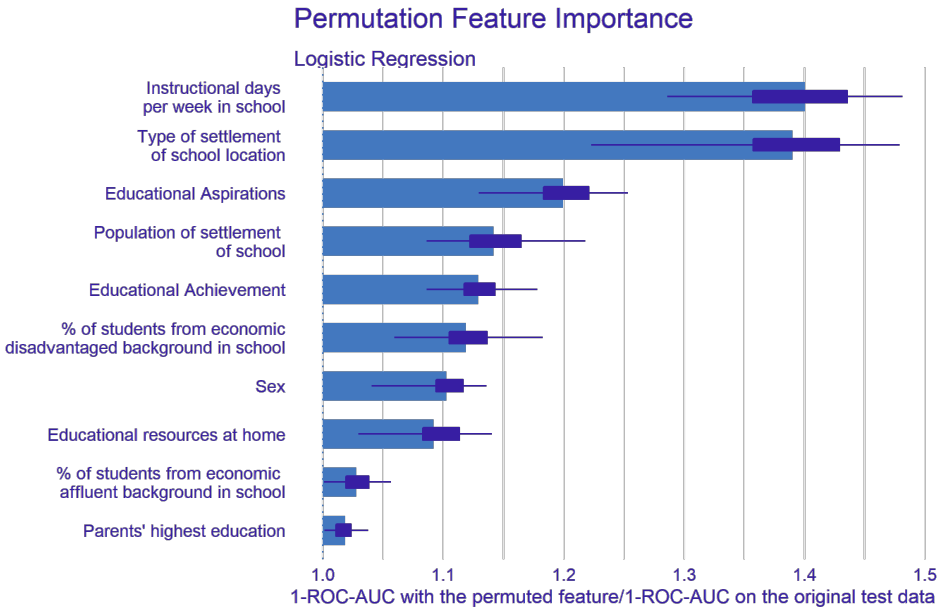

Figure 9: Permutation feature importance for Logistic Regression based on the 30 iterations. Box plots represent the range of predictors' importance over 30 iterations.

As for the Random Forest's features' importance, educational aspirations and educational achievement (the principal component of TIMSS Mathematics and Science scores) play the most important role in predicting higher education completion followed by sex. Interestingly, the predictors' importance of the Random Forest model on average have much higher spread when we look at the box plots in the Figure 10. One possible explanation for this is that Random Forest is an ensemble of large trees and predictors might appear in different trees and on different levels of these trees, thus, interacting with other predictors in this branch, both on higher and lower levels of the trees. Therefore, when the values of the feature are permuted it might also affect the outcome because of the interactions with other predictors. The magnitude of this effect could vastly vary from permutation to permutation.

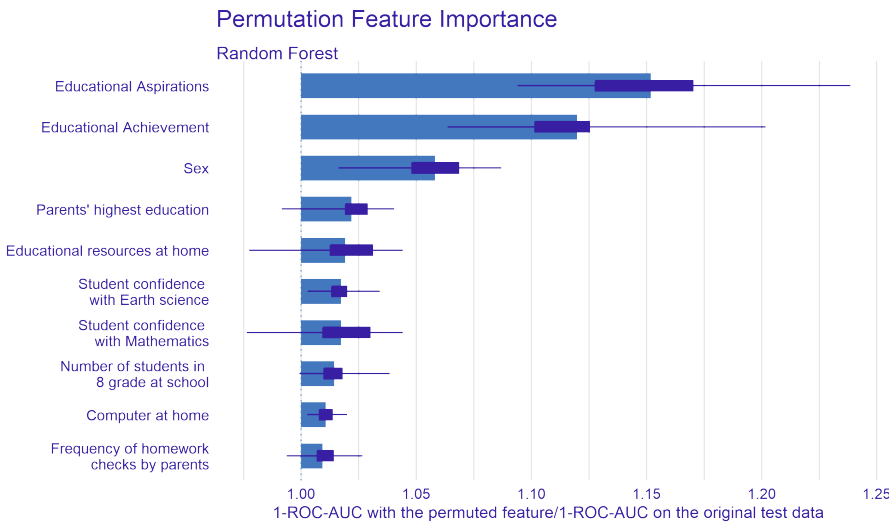

Figure 10: Permutation feature importance for Random Forest based on the 30 iterations. Box plots represent the range of predictors' importance over 30 iterations.

The Figure 11 presents the scatter plot and Pearson's correlation between predictors' importance in Logistic Regression and Random Forest models. Although there is evidence of the moderate and statistically significant correlation between predictors' importance derived from Logistic Regression and Random Forest models there are some discrepancies between them as well.

Most notably, there are features that play a crucial role in predicting higher education completion in Logistic Regression but they are considered much less important by the Random Forest model. These are, as described above, instructional days per week in school and the type of settlement where the school is located. These two predictors have the greatest effect on the predictive performance of the Logistic Regression but do not play such a major role in the Random Forest model. On the other hand, predictors that have the greatest impact on the predictive performance of the Random Forest model are also playing a notable role in Logistic Regression. Those are educational aspirations, educational achievement, and sex. More than that, among the top 10 most crucial predictors in Logistic Regression there are more school characteristics compared to the Random Forest model where individual-level predictors prevail. Moreover, there is only one school-level predictor among the top 10 most important predictors of Random Forest – the number of students in 8 grade in the school.

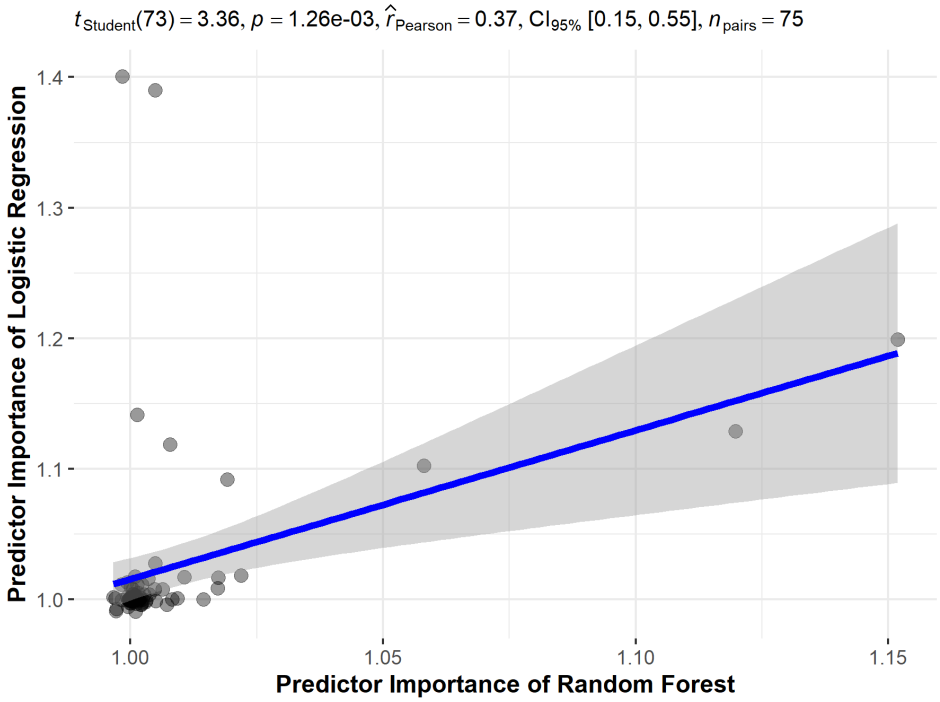

Figure 11: Scatter plot and Pearson's correlation of predictors' importance in Logistic Regression and Random Forest. The line represents linear smoothing, and the shaded area around the line is the 95% confidence interval.

5 DISCUSSION

The *research goal* of this study was to estimate how accurately we can predict who will get higher education by the age of 25 based on the information about socioeconomic background, family composition, school characteristics as well as educational results, aspirations, and attitudes from teenage years. Two other *research sub-questions* implied a comparison of the predictive performance of the machine learning algorithms and Logistic Regression as well as predictors' importance of those models.

5.1 *Results Discussion*

Answering the main research question, we can conclude that with given data it is possible to predict who will get higher education based on the information from teenage years with moderately high *accuracy* – 0.725-729 depending on the algorithm. In other words, in almost 73% of the cases, the machine learning algorithms can correctly predict who will get higher education by the age of 25. Moreover, this result could be considered robust

since all models demonstrated slightly higher but comparable performance on cross-validation (0.728-0.733). Given that the naive baseline for the target variable was 0.502 this result could be considered a significant improvement over the random guessing benchmark. Although this result does not come close to the (nearly) perfect predictions, it is intriguing because presumably demonstrates promising findings on human behavior prediction compared to other studies that made similar attempts (Bach et al., 2021; Salganik et al., 2020). Nevertheless, it is undoubtedly impossible to directly compare the predictive performance of studies that investigated different phenomena and used different metrics.

As for the comparison of machine learning algorithms and Logistic Regression, in our experimental and methodological setup, the fine-tuned machine learning algorithms (Random Forest, Elastic Net, XGBoost) *outperformed* Logistic Regression in predicting higher education completion both on cross-validation and test data. However, the difference between the machine learning algorithms and Logistic Regression is neglectably small. In more detail, the best-performing model, Random Forest, achieved a ROC-AUC of 0.798 while the Logistic Regression scored 0.787. The ratio of performance between two models is $\frac{0.798}{0.787} = 1.014$. In other words, Random Forest's discriminatory ability to predict higher education completion is 1.4% better than the respective ability of the Logistic Regression. The small differences in predictive performance between fine-tuned machine learning algorithms and classic regression models were also reported in the Salganik et al. (2020, 2019). However, in the study by Salganik et al. (2020) the benchmark models were estimated using four expertly preselected predictors with one predictor being the outcome measured in the previous period (survey wave). In our case, even if it would be of research interest, it would be unfeasible to use the information about higher education from the earlier waves because of the cumulative nature of this variable. Once the individuals complete higher education, they will have this status for the rest of their lives. Thus, information about higher education from the previous waves for those individuals who had higher education will deterministically define their outcome in the current, predicted, time period.

Although there is some alignment between predictors' importance in the best-performing algorithm, that is, Random Forest, and the classic Logistic Regression model, there are differences between them in this regard as well. The most striking difference is that Logistic Regression more heavily relies on the school-level predictors such as the number of instructional days per week in school and the type of settlement where the school is located. The importance of school-level features for the predictive performance of the regression models in the domain of educational

inequality was also reported in (Verhagen, 2021). For the Random Forest model, on the other hand, the most important predictors are the individual ones: educational aspirations, educational achievement, sex, and parents' highest education. Interestingly, these predictors perfectly correspond with the results from the explanatory studies discussed in the literature review section 2.2. However, as further discussed in the section 5.2, predictors' importance in the Logistic Regression might be flawed by the multicollinearity issue.

5.2 *Limitations*

One of the limitations of this study is lack of the strict criteria for the inclusion of the variables in the analysis which introduce relative arbitrariness in this process. However, based on the used variables it is possible to reproduce and recreate the research design to some extent using the data from similar longitudinal cohort surveys.

The other limitation is the sample size of the available data and its' multidimensionality. Since the most powerful machine learning algorithms achieve their' superiority by detecting complex interactions and non-linearities in the data, they might require more data to detect such patterns (Shmueli, 2010, p. 295). This might be of special importance when the ratio of the number of predictors to the sample size is large which introduces the curse-of-dimensionality (Friedman, 1997). The fundamental for this study paper by Salganik et al. (2020) also experienced this critique (Garip, 2020, p. 8235). However, our study might be less prone to this problem because of the preselection of the variables in the analysis while in the study by Salganik et al. (2020), most teams used (nearly) all available variables from multiple waves.

As for the predictors' importance analysis, one of the main limitations is the potential problem of multicollinearity in the Logistic Regression model. Although multicollinearity does not affect the overall predictive performance of the Logistic Regression it might affect the interpretation of predictors' importance since it might result in unstable regression coefficients, which in its turn, is a proxy measure of predictors' importance. We believe that this might impose significant restrictions on one of the advantages of using Logistic Regression for predictive modeling – direct interpretability of predictors' importance through regression coefficients. Thus, in some cases, it could require a trade-off between removal of the collinear predictors for the sake of better interpretability and keeping collinear predictors for the sake of better predictive performance. However, given that more complex algorithms only slightly outperformed Logistic Regression in our study and evidence from McKay (2019) that linear

models with few expertly selected predictors can achieve almost the same performance as sophisticated machine learning models with thousand of predictors, we might expect that removal of collinear predictors might not result in a substantial drop in the predictive performance but in the same time can offer clear interpretation for such models.

5.3 *Future Research*

The research design implemented in this study could be extended and applied to data from other countries and sociocultural contexts. The predictability of the higher education completion based on the information from the teenage years might be an intriguing metric itself that could be compared between contexts and time. Moreover, the predictive studies as suggested in (Garip, 2020, p. 8235) could benefit from using large sample sizes.

If such studies would be also interested in employing classical Logistic Regression because of its' interpretability they could employ Lasso regularization to enjoy data-driven feature selection (Boehmke & Greenwell, 2019, p. 125). However, the regression coefficient estimates derived from the regularized models are distorted due to the shrinkage which is imposed on them during regularization and could not be considered an unbiased estimates (Knaus, 2021, p. 287). One of the possible solutions could be the usage of fine-tuned machine learning algorithms, selection of the feasible number of the most important predictors, and further exploration of the relationship and interactions between them with more complex explanatory methods – structural equation modeling, multilevel regressions, causal inference techniques, etc.

Future research interested in gains in predictive performance could also try deep learning approaches crafted for the tabular data – for example, TabNet (Arık & Pfister, 2021). TabNet is the novel architecture of deep neural networks that employ a sequential attention approach and is declared to be a good fit for tabular data (Arık & Pfister, 2021). However, this approach might also require larger sample sizes to take advantage of architectures that allow the detection of more complex, non-linear, and detailed patterns in the data.

6 CONCLUSION

Overall, this study contributes to the growing body of research in predictive modeling in Social Sciences as well as to the Educational Studies by answering the following research questions:

RQ1 *To what extent can we predict higher education completion by the age of 25 based on the educational and background information from teenage years?*

The study revealed that using data about the various educational, background, and family characteristics from teenage years we can correctly predict with an accuracy of 72-73% which eighth-graders will get higher education by the age of 25.

This finding could be useful for social policy and can be used for more tailored educational interventions. Moreover, this result suggests that some life outcomes could be predicted relatively well which collides with the results achieved by previous predictive research in the sphere of life-course studies (Salganik et al., 2019).

RQ2 *Do the machine learning algorithms outperform classic statistical model, that is, logistic regression in predicting higher education completion?*

All tested machine learning algorithms slightly outperformed Logistic Regression. On the one hand, this result might argue for the usage of more sophisticated algorithms in predictive studies. On the other hand, one might consider it as an argument for the usage of linear models as more interpretable and robust which at the same time achieves comparable performance with sophisticated but less interpretable machine learning algorithms.

RQ3 *What are the differences in predictors' importance between classic statistical model, that is, logistic regression and best-performing machine learning algorithm in predicting higher education completion?*

The main difference is in type of predictors that are considered influential for Logistic Regression and Random Forest (best-performing algorithm). The former mostly relies on school-level predictors while the latter gives higher importance to the well-known in the literature individual-level predictors: educational aspirations and results, sex, and parents' education. Nevertheless, there is moderate correlation between predictors' importance of two models. However, the predictors' importance of the Logistic model might be flawed by the multicollinearity issue.

7 DATA SOURCE/CODE/ETHICS STATEMENT

Work on this study did not involve collecting data from human participants or animals. The data is available for researchers upon request from the data operator. The author of this study acknowledges that they do not have any legal claim to the data. The code used in this study is available at GitHub. All figures and tables in this study are made by the author.

REFERENCES

Ahearn, C. E., & Brand, J. E. (2019). Predicting layoff among fragile families. *Socius*, *5*, 2378023118809757.

Altschul, D. M. (2019). Leveraging multiple machine-learning techniques to predict major life outcomes from a small set of psychological and socioeconomic variables: A combined bottom-up/top-down approach. *Socius*, *5*, 2378023118819943.

Alyahyan, E., & Düştegör, D. (2020). Predicting academic success in higher education: literature review and best practices. *International Journal of Educational Technology in Higher Education*, *17*(1), 1–21.

Arık, S. O., & Pfister, T. (2021). Tabnet: Attentive interpretable tabular learning. In *Aaai* (Vol. 35, pp. 6679–6687).

Bach, R. L., Kern, C., Amaya, A., Keusch, F., Kreuter, F., Hecht, J., & Heinemann, J. (2021). Predicting voting behavior using digital trace data. *Social Science Computer Review*, *39*(5), 862–883.

Ballarino, G., Bernardi, F., Requena, M., & Schadee, H. (2009). Persistent inequalities? expansion of education and class inequality in Italy and Spain. *European Sociological Review*, *25*(1), 123–138.

Bernardi, F., & Boado, H.-C. (2014). Previous school results and social background: Compensation and imperfect information in educational transitions. *European Sociological Review*, *30*(2), 207–217.

Bernardi, F., & Cebolla, H. (2014). Social class and school performance as predictors of educational paths in Spain. *Revista Española de investigaciones sociologicas*, *146*(1), 3–22.

Biecek, P. (2018). DALEX: Explainers for complex predictive models in R. *Journal of Machine Learning Research*, *19*(84), 1-5. Retrieved from https://jmlr.org/papers/v19/18-416.html

Biecek, P., & Burzykowski, T. (2021). *Explanatory Model Analysis*. Chapman and Hall/CRC, New York. Retrieved from https://pbiecek.github.io/ema/

Boehmke, B., & Greenwell, B. (2019). *Hands-on machine learning with r*. Chapman and Hall/CRC.

Boudon, R. (1974). *Education, opportunity, and social inequality: Changing prospects in western society.* ERIC.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Brown, R. (2018). Higher education and inequality. *Perspectives: Policy and Practice in Higher Education*, *22*(2), 37–43.

Compton, R. (2019). A data-driven approach to the fragile families challenge: Prediction through principal-components analysis and random forests. *Socius*, *5*, 2378023118818720.

Cranmer, S. J., & Desmarais, B. A. (2017). What can we learn from predictive

modeling? *Political Analysis*, *25*(2), 145–166.

Delnoij, L. E., Dirkx, K. J., Janssen, J. P., & Martens, R. L. (2020). Predicting and resolving non-completion in higher (online) education – a literature review. *Educational Research Review*, *29*, 100313. Retrieved from https://www.sciencedirect.com/science/article/pii/S1747938X1930171X doi: https://doi.org/10.1016/j.edurev.2020.100313

Dickert-Conlin, S., & Rubenstien, R. (2007). *Economic inequality and higher education: Access, persistence, and success*. Russell Sage Foundation.

Doornenbal, B. M., Spisak, B. R., & van der Laken, P. A. (2021). Opening the black box: Uncovering the leader trait paradigm through machine learning. *The Leadership Quarterly*, 101515.

Duta, A., Wielgoszewska, B., & Iannelli, C. (2021). Different degrees of career success: social origin and graduates' education and labour market trajectories. *Advances in Life Course Research*, *47*, 100376.

Fox, J., & Monette, G. (1992). Generalized collinearity diagnostics. *Journal of the American Statistical Association*, *87*(417), 178–183.

Friedman, J. H. (1997). On bias, variance, 0/1—loss, and the curse-of-dimensionality. *Data mining and knowledge discovery*, *1*(1), 55–77.

Garip, F. (2020). What failure to predict life outcomes can teach us. *Proceedings of the National Academy of Sciences*, *117*(15), 8234–8235.

Hartas, D. (2015). Parenting for social mobility? home learning, parental warmth, class and educational outcomes. *Journal of Education Policy*, *30*(1), 21–38.

Hofman, J. M., Sharma, A., & Watts, D. J. (2017). Prediction and explanation in social systems. *Science*, *355*(6324), 486–488.

Hofman, J. M., Watts, D. J., Athey, S., Garip, F., Griffiths, T. L., Kleinberg, J., ... others (2021). Integrating explanation and prediction in computational social science. *Nature*, *595*(7866), 181–188.

Ivanov, S., & Prokhorenkova, L. (2021). Boost then convolve: Gradient boosting meets graph neural networks. In *International conference on learning representations*. Retrieved from https://openreview.net/forum?id=ebS5NUfoMKL

Jerrim, J., Chmielewski, A. K., & Parker, P. (2015). Socioeconomic inequality in access to high-status colleges: A cross-country comparison. *Research in Social Stratification and Mobility*, *42*, 20–32.

Joel, S., Eastwick, P. W., Allison, C. J., Arriaga, X. B., Baker, Z. G., Bar-Kalifa, E., ... others (2020). Machine learning uncovers the most robust self-report predictors of relationship quality across 43 longitudinal couples studies. *Proceedings of the National Academy of Sciences*, *117*(32), 19061–19071.

Knaus, M. C. (2021). A double machine learning approach to estimate

the effects of musical practice on student's skills. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *184*(1), 282–300.

Kuhn, M., & Johnson, K. (2019). *Feature engineering and selection: A practical approach for predictive models*. CRC Press.

Kuhn, M., & Silge, J. (2020). *Tidy modeling with r.* Dostęp (14.07. 2021): https://www. tmwr. org.

Kuhn, M., & Wickham, H. (2020). Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles. [Computer software manual]. Retrieved from https://www.tidymodels.org

Kurakin, D. (2014). Russian longitudinal panel study of educational and occupational trajectories: Building culturally-sensitive research framework. *Series Working papers of Institute of Education "Scientific Reports Institute of Education"*(WP).

Li, X., Han, M., Cohen, G. L., & Markus, H. R. (2021). Passion matters but not equally everywhere: Predicting achievement from interest, enjoyment, and efficacy in 59 societies. *Proceedings of the National Academy of Sciences*, *118*(11).

Lucas, S. R. (2001). Effectively maintained inequality: Education transitions, track mobility, and social background effects. *American journal of sociology*, *106*(6), 1642–1690.

Malik, V. (2019). The russian panel study 'trajectories in education and careers'. *Longitudinal and Life Course Studies*, *10*(1), 125–144.

Martin, M., & Mullis, I. (2011). Timss and pirls achievement scaling methodology. *Methods and procedures in TIMMS and PIRLS. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College*, 1–11.

McKay, S. (2019). When 4 ≈ 10,000: The power of social science knowledge in predictive performance. *Socius*, *5*, 23780231188117 74.

Molina, M., & Garip, F. (2019). Machine learning for sociology. *Annual Review of Sociology*, *45*, 27–45.

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Morgan, S. L. (2012). Models of college entry in the United States and the challenges of estimating primary and secondary effects. *Sociological Methods & Research*, *41*(1), 17–56.

Nagy, M., & Molontay, R. (2018). Predicting dropout in higher education based on secondary school performance. In *2018 IEEE 22nd international conference on intelligent engineering systems (INES)* (pp. 000389–000394).

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2016). Predicting performance in higher education using proximal predictors. *PloS one*, *11*(4), e0153663.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). Catboost: unbiased boosting with categorical features. *Advances in neural information processing systems*, *31*.

R Core Team. (2021). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Raes, L. (2019). Predicting GPA at age 15 in the fragile families and child wellbeing study. *Socius*, *5*, 2378023118824803.

Rigobon, D. E., Jahani, E., Suhara, Y., AlGhoneim, K., Alghunaim, A., Pentland, A. S., & Almaatouq, A. (2019). Winning models for grade point average, grit, and layoff in the fragile families challenge. *Socius*, *5*, 2378023118820418.

Risi, J., Sharma, A., Shah, R., Connelly, M., & Watts, D. J. (2019). Predicting history. *Nature human behaviour*, *3*(9), 906–912.

Salganik, M. J., Lundberg, I., Kindel, A. T., Ahearn, C. E., Al-Ghoneim, K., Almaatouq, A., ... others (2020). Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences*, *117*(15), 8398–8403.

Salganik, M. J., Lundberg, I., Kindel, A. T., & McLanahan, S. (2019). Introduction to the special collection on the fragile families challenge. *Socius*, *5*, 2378023119871580.

Shmueli, G. (2010). To explain or to predict? *Statistical science*, *25*(3), 289–310.

Shwartz-Ziv, R., & Armon, A. (2022). Tabular data: Deep learning is not all you need. *Information Fusion*, *81*, 84–90.

Simonová, N., & Soukup, P. (2015). Impact of primary and secondary social origin factors on the transition to university in the czech republic. *British Journal of Sociology of Education*, *36*(5), 707–728.

Triventi, M. (2013). The role of higher education stratification in the reproduction of social inequality in the labor market. *Research in Social Stratification and Mobility*, *32*, 45–63.

Verhagen, M. D. (2021, Apr). *To predict and explain. how prediction improves our understanding of models: an application to the study of teacher bias.* SocArXiv. Retrieved from osf.io/preprints/socarxiv/y6mnb doi: 10.31235/osf.io/y6mnb

Verhagen, M. D. (2022). A pragmatist's guide to using prediction in the social sciences. *Socius*, *8*, 23780231221081702.

Wright, M. N., & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, *77*(1), 1–17. Retrieved from https://www.jstatsoft.org/index.php/jss/article/view/v077i01 doi: 10.18637/jss.v077.i01

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in

psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122.

Yastrebov, G., Kosyakova, Y., & Kurakin, D. (2018). Slipping past the test: Heterogeneous effects of social background in the context of inconsistent selection mechanisms in higher education. *Sociology of Education*, *91*(3), 224–241.

APPENDIX A

Table 8: Variables selected for the analysis.

Variable	Label in TIMSS data
idstud	TIMSS student id
itbirthy	*date of students birth\year*
itsex	*sex of students*
bsbg03	gen\often speak <lang of test>at home
bsbg04	gen\amount of books in your home
bsbg05a	gen\home possess\computer
bsbg05b	gen\home possess\study desk
bsbg05c	gen\home possess\books
bsbg05d	gen\home possess\own room
bsbg05e	gen\home possess\internet connection
bsbg05f	gen\home possess\<country specific>
bsbg05g	gen\home possess\<country specific>
bsbg05h	gen\home possess\<country specific>
bsbg05i	gen\home possess\<country specific>
bsbg05j	gen\home possess\<country specific>
bsbg05k	gen\home possess\<country specific>
bsbg07	gen\how far in edu do you expect to go
bsbg08a	gen\<stmo or fem guard>born <country>
bsbg08b	gen\<stfa or ma guard>born in <country>
bsbg09a	gen\born in <country>
bsbg09b	gen\born in <country>\age of coming
bsbg11a	gen\how often\home\parents ask learning
bsbg11b	gen\how often\home\talking about school
bsbg11c	gen\how often\home\parents make sure
bsbg11d	gen\how often\home\parents check homewor
bsbg12a	gen\agree\being in school
bsbg12b	gen\agree\safe at school
bsbg12c	gen\agree\belong at school
bsbm33ba	mat\hm\minutes spent on hmwork\mat

Table 8 continued from previous page

Variable	Label in TIMSS data
bsbb33bb	bio\hm\minutes spent on hmwork\bio
bsbe33bc	ear\hm\minutes spent on hmwork\ear
bsbc33bd	che\hm\minutes spent on hmwork\che
bsbp33be	phy\hm\minutes spent on hmwork\phy
bsmmat01	*1st plausible value mathematics*
bsmmat02	*2nd plausible value mathematics*
bsmmat03	*3rd plausible value mathematics*
bsmmat04	*4th plausible value mathematics*
bsmmat05	*5th plausible value mathematics*
bssscio1	*1st plausible value science*
bssscio2	*2nd plausible value science*
bssscio3	*3rd plausible value science*
bssscio4	*4th plausible value science*
bssscio5	*5th plausible value science*
bsbgher	*home educational resources /scl*
bsbgsbs	*students bullied at school/scl*
bsbgslm	*students like learning mathematics/scl*
bsbgsls	*students like learning science/scl*
bsbgslb	*students like learning biology/scl*
bsbgslc	*students like learning chemistry/scl*
bsbgslp	*students like learning physics/scl*
bsbgsle	*students like learning earth science/scl*
bsbgsvm	*students value learning mathematics/scl*
bsbgsvs	*students value learning science/scl*
bsbgsvb	*students value learning biology/scl*
bsbgsvc	*students value learning chemistry/scl*
bsbgsvp	*students value learning physics/scl*
bsbgsve	*students value learning earth sci/scl*
bsbgscm	*student confidence with mathematics/scl*
bsbgscs	*student confidence with science/scl*
bsbgscb	*student confidence with biology/scl*
bsbgscc	*student confidence with chemistry/scl*
bsbgscp	*student confidence with physics/scl*
bsbgsce	*student confidence with earth sci/scl*
bsbgeml	*students engaged in mathematics lessons/scl*
bsbgesl	*students engaged in science lessons/scl*
bsbgebl	*students engaged in biology lessons/scl*
bsbgecl	*students engaged in chemistry lessons/scl*
bsbgepl	*students engaged in physics lessons/scl*
bsbgeel	*students engaged in earth science lessons/scl*

Table 8 continued from previous page

Variable	Label in TIMSS data
bsdgedup	*parents' highest education level*
bcbg01	gen\total enrollment of students
bcbg02	gen\total enroll <eighth grade>std
bcbg03a	gen\students background\economic disadva
bcbg03b	gen\students background\economic affluen
bcbg04	gen\percent of students <lang of test>
bcbg05a	gen\how many people live in area
bcbg05b	gen\immediate area of sch location
bcbg05c	gen\average income level of area
bcbg06a	gen\instructional days per year
bcbg06c	gen\instructional days in 1 calenderweek
bcbg07	gen\total number computers
bcbg08a	gen\existing science laboratory
bcbg08b	gen\existing assistance during exp
bcbgsrs	*instruction affected by science resource shortages/scl*
bcbgmrs	*instruction affected by mathematics resource shortages/scl*
bcbgeas	*school emphasis on academic success - principal reports/scl*
bcbgdas	*school discipline and safety/scl*
bcdg03	*school composition by student backgd*
bcdgcmp	*computer availability for instruction*
bcdg06hy	*total instructional hours per year*

APPENDIX B

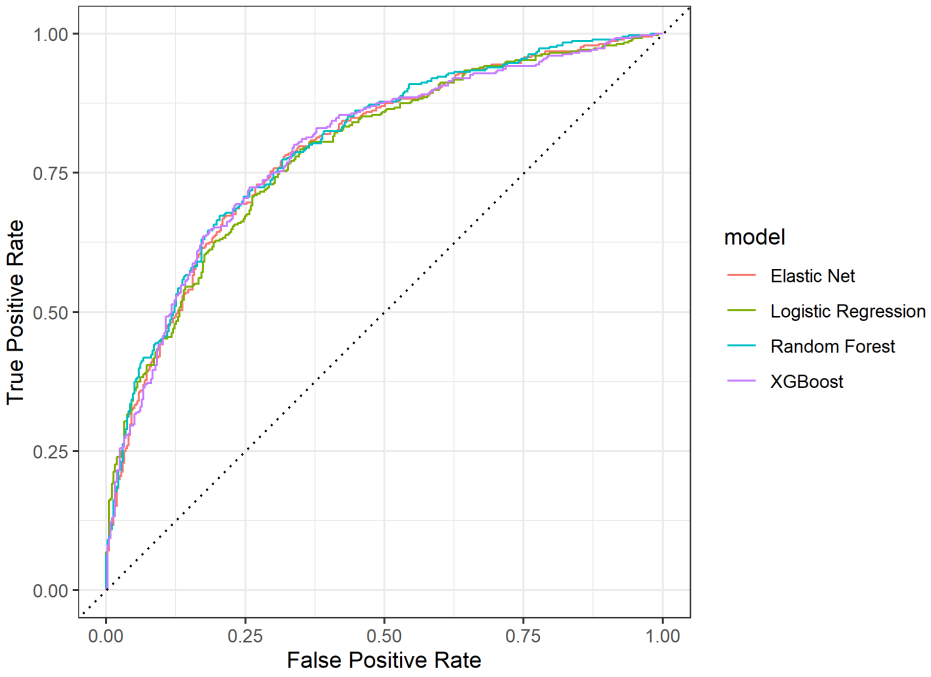

Figure 12: ROC-AUC of the models on the test data.

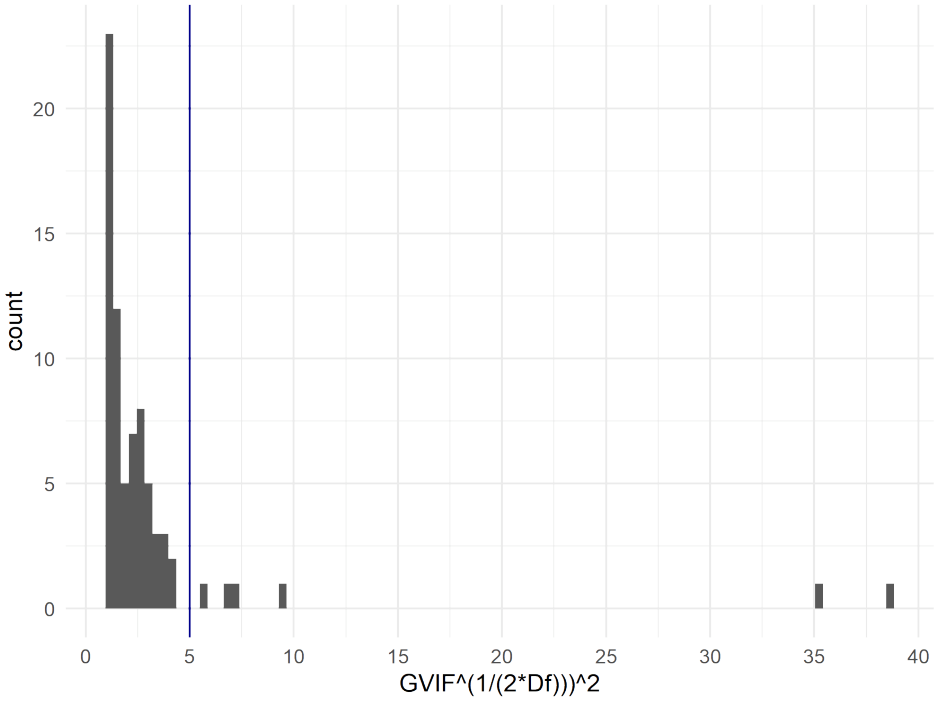

Figure 13: Histogram of the squares of Generalized Variance Inflation Factor adjusted by the degrees of freedom for the particular predictor. The blue vertical line represents the threshold for the rule-of-thumb for multicollinearity.