SZKOŁA GŁÓWNA HANDLOWA W WARSZAWIE
WARSAW SCHOOL OF ECONOMICS

Undergraduate studies

Course: Quantitative methods in economy and information systems

Stanisław Wojciech Sender
Student's number: 114165

Classification Predictive Models of Polish Students' Performance in PISA

Bachelor's thesis
written in the
Institute of Econometrics
under the academic direction of
Katarzyna Bech-Wysocka, PhD

Warsaw, 2024

Contents

Chapter 1

Introduction.

1.1. Research intent and thesis structure.

The aim of the following bachelor's thesis: "Classification Predictive Models of Polish Students' Performance in PISA" is to verify the accuracy of logistic regression models and random forest models as well as the stableness of the patterns across the editions in the prediction of above-average performance in mathematics and reading PISA tests. The research may result in revealing possibly non-obvious factors influencing Polish students' scores.

PISA database is arguably one of the vastest sources of the complex information in an educational field. The influence of education on the functioning of entire societies and differing individuals is immensely essential as it not only greatly impacts productivity, development and advancement but also allows people question the way things are and redefine the status quo. At the very beginning of each scientific revelation, pushing science forward, putting it into practice, and adapting to a continuously changing environment is the learning process itself. This is why so many countries and organizations increasingly emphasize observing the factors that shape the next generation's way of thinking. To predict how the educational environment can prepare students for their future lives, it is crucial to thoroughly analyze it.

The structure of the thesis is as follows. Chapter I introduces organization and history of PISA as well as previously established dependencies regarding educational assessed performance. Chapter II reveals quantitative methods, models and measures applied in the research. Chapter III describes exact process of data analysis, model evaluating and drawing conclusions from empirical research.

1.2. Description of PISA.

PISA (Programme for International Student Assessment) is an international study programme started in the year 2000. Its aim is to evaluate education systems around the world by testing the educational performance of 15-year-old students. Since the year 2000 over 100 countries (including Poland in every edition) have participated in PISA.

Every three years (the gap between 2018 and 2022 is larger because of the Covid-19 pandemic), a group of fifteen-year-olds selected randomly, take tests in three main subjects – reading, mathematics, and science – each assessment edition concentrates on one of them. The main focus was reading in 2018 and mathematics in 2022.

Students took tests lasting two hours, each devoted to one subject. Different students were given nonidentical test questions and varying combinations of subjects (e.g. reading followed by mathematics, or mathematics followed by science, etc.). Test items consisted of multiple-choice questions and questions that required students to make their own responses.

The school principals students and their parents were also asked to fill in background questionnaires in order to provide information on the students' backgrounds and the way their schools are organized. It means that there are three main sources of information regarding possible factors influencing students' PISA performance – school questionnaire, parents' questionnaire, and student questionnaire.

PISA aspires to develop tests which are not linked directly to the schooling curricula and aims to provide context information through the background questionnaires which allows analysts to explore the relationships between student performance and various environmental, social, and educational factors. The tests' main role is to assess the ability of 15-year-olds, often at the end of compulsory education, to apply their school knowledge in real-life situations and participate fully in society.

Poland has witnessed one of the fastest increases in median scores for mathematics and reading among OECD countries since the assessments began. However, the 2022 edition saw a more significant decline in these median scores compared to other OECD nations. This downturn could be linked to multiple changes in the Polish educational system over the years, which may have influenced the results. Additionally, one might argue that the COVID-19 pandemic has taken its toll on educational outcomes all over the world.

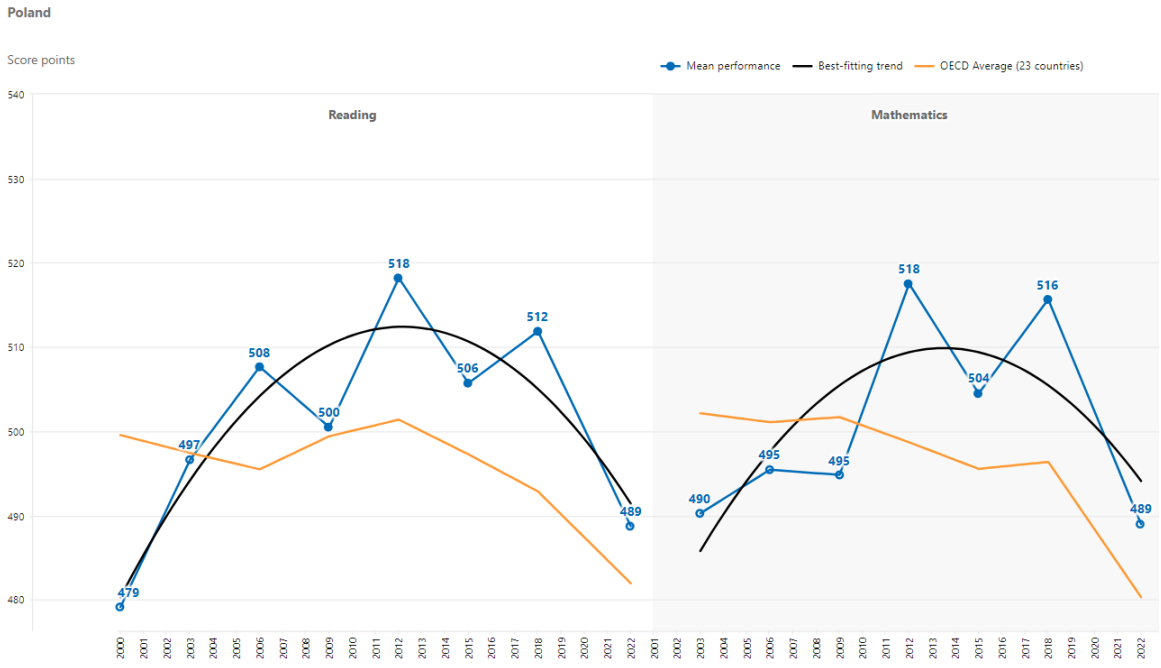

Figure 1.1: Polish students' performance across the years. Source: PISA Database, OECD.

In the PISA assessments, the comparison of median scores across different participant countries is standardized based on a global norm. Specifically, the scores are aligned to a normal distribution, denoted as $N(\mu = 500, \sigma = 100)$. Here, μ represents the mean score, set at 500, and σ, the standard deviation, set at 100. This standardization process ensures that the median score is centered at 500 with a standard deviation of 100, enabling equitable and consistent comparisons across the international student populations.

1.3. Theoretical view on factors influencing educational performance

The Organization for Economic Co-operation and Development (OECD) has consistently conducted comprehensive research since the inception of the Programme for International Student Assessment (PISA), aiming to identify the most universally significant factors influencing test outcomes. These factors are conventionally divided into several crucial domains:

• socioeconomic family status, usually measured by wealth items and the occupational status of parents as they have been proven to be associated with the educational attainments of students;

• cultural family capital, which is often measured by possession of certain cultural assets and by the level of education of parents, on the grounds that this is the cultural equipment with which an actor is provided by the family and which increases his or her chances of attaining

educational success;

• students' well-being, that is directly measured by questions on the characteristics and feelings of students, as mental health is an important precondition for educational success;

• course of students' education, which includes variables such as length of preschool education, whether the student has repeated a school year, which are significant for educational outcomes.

• school organization, including the structural aspects of educational institutions, such as the size and organizational complexity of schools, as well as multiple issues that may impede educational processes.

In addition to these primary factors, other common performance indicators are also considered, including gender, birth month, and the size of the urban area where the school is located, among others.

Created by PISA, aggregated measures of different student environment factors such as the ESCS (Economic, Social, and Cultural Status) index are scrutinized in Avvisati (2020) The ESCS index is an extremely complicated aggregate of hundreds of variables of different types, often created from questions that vary among nations due to the subjective nature of wealth items, unevenly distributed missing values, and more. Its main advantage is its ability to compare students' subjective status across different cultures, simplifying a multi-dimensional problem to one index while preserving its correlative and predictive qualities. In this research, limited to single country, the emphasis was also placed on the interpretability and simplicity of variables, taking into account a rather uniform interpretation of separate items regarding students' environments. The only widely used score-type values that could not be replaced by simpler types of variables and had to be retained in modeling were those describing parental occupational status. This status was measured based on open-ended questions about the father's and mother's job titles, which were converted into an ordinal or interval scale using prestige rankings or income rankings based on international studies approved by OECD.

Not the less essential was to acknowledge the complexity of the phenomena under study and the inevitability of interactions between various variables. As highlighted by Walczak & Walczak (2022), there is a notable potential for a child's gender to significantly influence the level of active parental support in education, as observed in the PISA 2018 assessments in Poland. US studies Rutkowski et al. (2017) examine relationship of poverty with other educational factors. Here, the strong correlation between under-performing impoverished students and their truancy of school, tardiness and other similar factors is proven. Broad Australian research Gabriel et al. (2018) describing machine learning methods notes great potential of those in revealing more complicated relationships and interactions between variables especially with high density of good quality data. It also underlines role of data-cleaning and handling processes as different approaches may influence final conclusions.

Researchers are debating whether different fields of science require varying skill sets. The general factor accounted for about 84% of the common variance in cognitive item responses amongst mathematics, reading, and science tests in OECD countries in PISA 2018 according to psychometric research Pokropek et al. (2022). That implies that there are factors influencing both mathematics and reading performance in a similar fashion, though as two-dimensional distribution does not perfectly align with the linear pattern we may expect some differences.

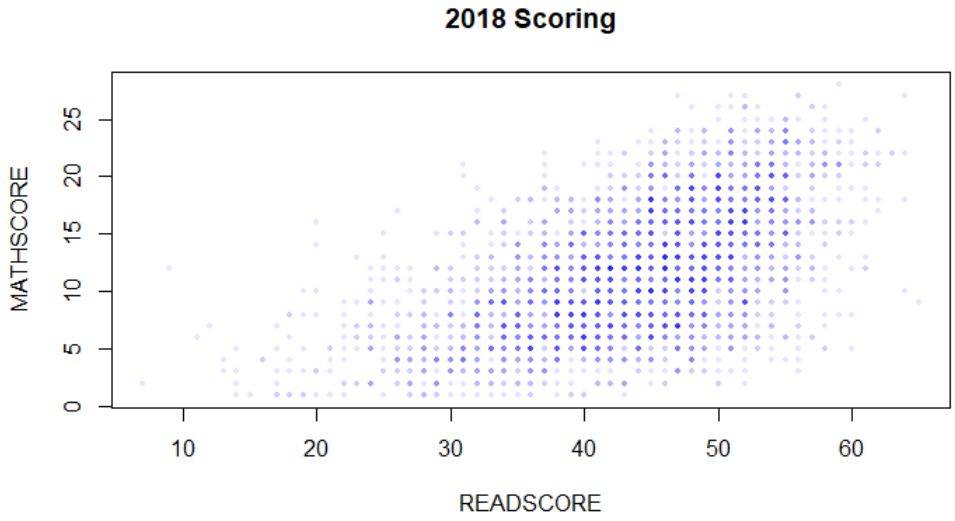

Figure 1.2: Polish students' reading and math scores in 2018. Source: own calculations

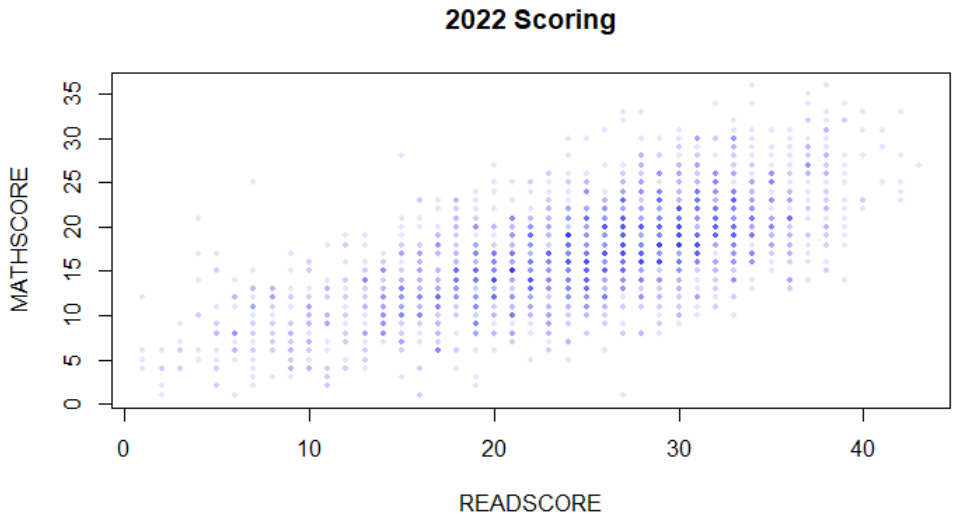

Figure 1.3: Polish students' reading and math scores in 2022. Source: own calculations

Chapter 2

Methodology

2.1. Logistic regression.

Logistic regression is a statistical method enabling analysis of a dataset in which there are some independent variables that determine an outcome as in the case of created PISA dataset. The outcome is measured with a binary variable deciding whether the student performed above or below the population average. It is particularly useful for understanding and interpreting the impact of several independent variables on a binary outcome.

The logistic regression model is described by the following equation:

$$\log\left(\frac{1-P(Y=1\mid X)}{P(Y=1\mid X)}\right)=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_nX_n$$

Where:

- Y is the binary dependent variable (whether a student performs above or below average in math and reading tests).
- $X_1, X_2, \ldots, X_n$ are the predictors (such as GENDER, BOOKSHOME, etc.).
- $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ are the coefficients of the model which describe the relationship between each predictor and the log-odds of the dependent variable.

The very method estimates the probability of the occurrence of an event utilizing a logistic function, which is an S-shaped curve that can take any real number and map it into a value between 0 and 1, but theoretically never exactly at those limits or other way. Thus, logistic regression is extremely useful for predicting the likelihood of the occurrence of an event, which in this case, means scoring by an individual student above the Polish students' average score in mathematics and reading.

The coefficients in logistic regression are usually estimated using Maximum Likelihood Estimation (MLE). The great advantage of MLE used in logistic regression is not requiring a linear relationship between the independent and dependent variables. However, it does assume linearity between independent variables and their log-odds.

Though logistic regression is robust regarding small sample sizes and simple to implement and interpret, it requires that observations be independent of each other. In cases where predictors are not independent of each other meaning multicollinearity, it can lead to issues with estimation. That is one of the reasons behind not using MATHSCORE in order to predict READSCORE and vice versa.

Applying logistic regression, it's generally advisable to limit the number of variables in the model in order to enhance its interpretability and to avoid overfitting, especially as the sample size might not be sufficiently large relative to the number of predictors. What's more, the model with too many variables may become complex and difficult to understand, making it impossible to draw clear conclusions about the relationships within the data.

Given these, the focus was made only on the most important variables. Those were selected by performing a variable selection process, which included techniques such as stepwise selection, analysis of the variables' statistical significance, and their contribution to the model fitting. This approach hopefully simplified the model while retaining its predictive power. By reducing the number of variables, we also decrease the risk of fitting noise present in the training data, thus enhancing the robustness and reliability of the model's predictions allowing us to make general conclusions.

2.2. Random Forests.

Random Forests is a type of machine learning method capable of performing classification tasks amongst many other. It is particularly powerful when it comes to dealing with datasets with high dimensionality, as is the case with the PISA dataset which consists of a wide array of variables potentially influencing the performance of students. Unlike logistic regression, used to predict binary outcomes based on a linear combination of independent variables, Random Forests depend on constructing many decision trees during the training phase and outputting the mode of the classes of the individual trees.

Random Forests manage binary outcomes - such as students performing above or below the population average - through a whole ensemble of decision trees. Each tree in the forest independently considers random subsets of the data and features. While individual trees may exhibit low bias, they also have high variance among them. Random Forests maintain the advantage of low bias from the individual trees and mitigate the risk of overfitting. Simultaneously, ensembling significantly reduces variance, enhancing the stability and accuracy of the model.

The training process randomly select both observations and features in order to create multiple decision trees. The final prediction is made by taking a majority vote when it comes to classification problems:

$$\text{Prediction} = \text{mode}\left(\{\text{Tree}_1(x), \text{Tree}_2(x), \ldots, \text{Tree}_N(x)\}\right)$$

One significant advantage of Random Forests over Logistic Regression is its capability to model complex interactions between variables without the need for explicit specification. Moreover, Random Forests do not assume a linear relationship between the independent variables and the dependent variable as well as do not require the independent variables to be linearly separable from each other.

However, despite some strengths, Random Forests require careful parameter tuning such as the number of trees in the forest or the number of features considered at each split. Additionally, it is rather more intensive computationally than logistic regression, as it requires more resources for prediction or training, especially when dealing with large datasets.

To sum up, while logistic regression provides kind of a robust and straightforwardly interpreted method for analyzing binary outcomes using less predicting variables, Random Forests may offer a more flexible and powerful alternative, possibly capable of capturing more complex patterns in the data. In practice, Random Forests models are expected to perform better than logistic regression regarding the accuracy of prediction.

2.3. Forecast evaluation methods

Forecasting for a similar sample ex-post implies that, based on historical data, predictive models are built, which, after that, are tested on future data that has been observed. This will serve the purpose of validating the model's power of prediction and its ability to generalize across time.

Practically, the dataset is usually divided into two distinct subsets: a training one and a testing one. The training set, consisting of earlier data, is used to train the model; that is, the model parameters are estimated and relationships and patterns are identified among factors.

After training the model on the historical data, the approach is then applied on the test set composed of later data to assess the performance. Data that represents the future values to be predicted is simulated by using the test set that represents the check measures for the future observations.

Ex-post evaluation involves the computation of various metrics such as accuracy, sensitivity, specificity, ROC curves, and AUC to quantify the predictive power of the model. Accuracy measures the ratio of correct predictions. Sensitivity and specificity focus on the model's ability to correctly identify the positive and negative cases, respectively. The ROC curves and AUC may give an integral view of the model's discriminative power.

Accuracy is the simplest performance metric for the classification model. It is the ratio of the number of correctly predicted instances to the total number of instances. For the model

predicting above-average performance in PISA tests, accuracy measures how well the model identifies students who do better or worse than the population average.

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{True Positives} + \text{True Negatives} + \text{False Positives} + \text{False Negatives}}$$

Sensitivity measures the proportion between true positives correctly identified by the model, i.e. the proportion of students who perform above average and are correctly predicted as such.

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Specificity measures the proportion of true negatives correctly identified by the model, i.e., the proportion of students who perform below average and are correctly predicted as such.

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}$$

The Receiver Operating Characteristic curve is a graphical representation that shows model performance with respect to different threshold values. In binary classification, the output of the model is mostly a probability of the instance belonging to that class. To map this probability to the binary outcome class 0 and 1, the threshold value is set. For example, a threshold value of 0.5 would mean that instances predicted with a probability greater than or equal to 0.5 are classified as class 1, and instances predicted with a probability less than 0.5 are classified as class 0.

A model with good discriminative ability will have a ROC curve that approaches the top-left corner of the plot. The Area Under the ROC Curve summarizes the overall performance of the model similarly to accuracy. An AUC of 0.5 is equivalent to random guessing, while an AUC of 1.0 indicates perfect discriminative ability.

Chapter 3

Empirical Analysis

3.1. Dataset properties and operations.

The study involved downloading approximately 20 GB of data from the PISA database, which included SAS Student questionnaire data files for 2018 and 2022, containing detailed responses from students and their parents about their environment, school perception, and well-being. Additional files included SAS School questionnaire data for 2018 and 2022, which provided insights from school principals regarding school parameters and challenges, and SAS Cognitive item data files for the same years, which assessed students' responses to test items. Excel codebooks for 2018 and 2022 were also utilized to identify variables present across all files.

The data were initially opened as dataframes in RStudio. The largest file, the 2022 student questionnaire data, comprised approximately 200,000 observations and around 4,000 variables. The analysis was restricted to Polish students both to maintain focus on a single national context and avoid problems related with handling 'big data' without proper tools. Cognitive item data were aggregated to compute final test scores, interpreting NaN values in test-specific columns as indicative of the student not having participated in that portion of the assessment, while partial NaNs were treated as zero points in the score aggregation.

A merge process was conducted where the student questionnaire data and cognitive item data were joined on student IDs. This combined dataset was then merged with the school questionnaire data based on school IDs, producing two comprehensive datasets for the years 2018 and 2022.

The variables were carefully selected based on their presence in both dataset editions, their response rate (to minimize missing data), and their relevance or potential importance to the study. Observations with any missing values in selected variables were excluded to ensure data quality. The names of the variables were standardized to enhance understandability. What's more many of the ordinal variables were transformed into binary format to facilitate analysis.

The final modeling datasets each contained 33 variables. The training set from 2018 included 2,040 observations, while the testing set from 2022 included 849 observations. The relatively small size of the latter dataset precluded the use of neural networks, prompting a focus on logistic regression and random forests as more suitable analytical methods given the data size constraints.

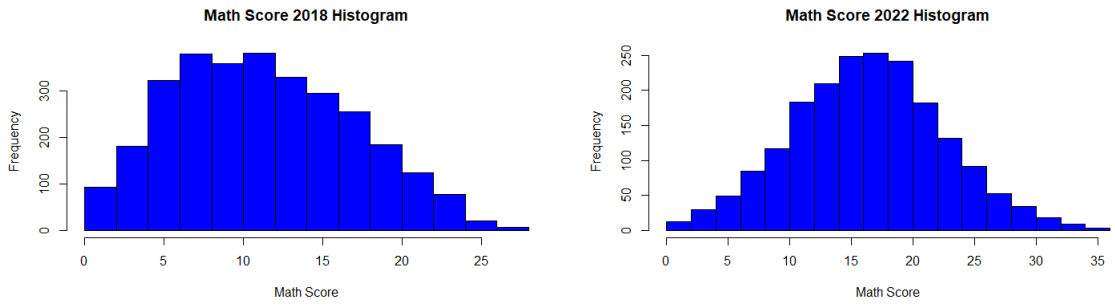

Figure 3.1: Histograms of math scores. Source: own calculations

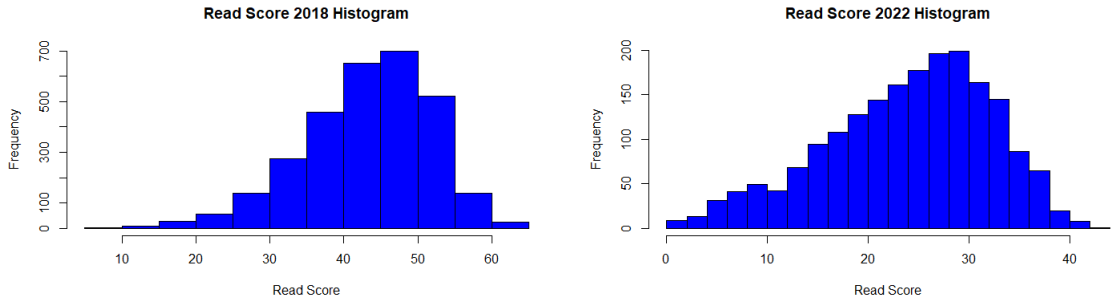

Figure 3.2: Histograms of read scores. Source: own calculations

Histograms of the results for different parts of the study over the years show variability in score ranges due to different focuses made in successive editions. We can also observe left-skewness in the case of mathematics tests - with the median lower than the mean, and right-skewness in the case of reading tests - with the median higher than the mean. However, in all scenarios, the classification proportion of “0” and “1” should be nearly equal.

Table 3.1: Dictionary of Variables. Source: own calculations

Ord.	Variable Name	Type of Value	Description
1	MATHSCORE	binary	Math score over average value in population*
2	READSCORE	binary	Read score over average value in population*
3	GENDER	binary	Male*
4	MISCED	binary	Mother has completed bachelor studies or higher*
5	FISCED	binary	Father has completed bachelor studies or higher*
6	BOOKSHOME	binary	Student declared many (over 100) books in home*
7	OWNROOM	binary	Student declared own room in home*
8	OWNCOMPUTER	binary	Student declared own computer/laptop*
9	INSTRUMENTHOME	binary	Student declared any musical instrument home*
10	CARSHOME	binary	Student declared at least two cars home*
11	TCHLISTEN	binary	Student declared feeling listened by teachers*
12	TCHUNDERSTAND	binary	Student declared feeling understood by teacher*
13	LIFESATISF	binary	On a scale from 1 to 10, a student declared life satisfaction*
14	ALIENATION	binary	Student declared feeling alienated in school*
15	LIKEME	binary	Student declared feeling liked by peers*
16	LATE	binary	Student declared being late at school at least once a month*
17	BULLIED	binary	Student declared feeling threatened or bullied at school*
18	MISEI	score	Mother's measure of socio-economic occupational status
19	FISEI	score	Father's measure of socio-economic occupational status
20	CITY	binary	School in city over 100k inhabitants*
21	PRIVATE	binary	Private school*
22	DRUGS	binary	School reported any drugs-related problems*
23	SCHSIZE	numeric	Total amount of students in school
24	CLSIZE	numeric	Total amount of students in class
25	POORSTU	percentage	Percentage of students from socioeconomically disadvantaged backgrounds
26	TCHLACK	binary	School reported any lack in teaching staff*
27	REPEAT	binary	Student has repeated schooling year*
28	Q1BIRTH	binary	Student born in first year quarter*
29	Q2BIRTH	binary	Student born in second year quarter*
30	Q3BIRTH	binary	Student born in third year quarter*
31	Q4BIRTH	binary	Student born in fourth year quarter*
32	VILLAGE	binary	School in village*
33	TOWN	binary	School in town with less than 100k inhabitants*

* For binary values, the description signifies the "1" or "success" value.

3.2. Models interpretation

3.2.1. Logistic regression read model

Table 3.2: Logistic Regression Results for Reading Performance. Source: own calculations

Variable	Coefficients	Odds Ratio (%)	p-value
GENDER	-0.42	65.8	0.000
BOOKSHOME	0.639	189.5	0.000
VILLAGE	-0.448	63.9	0.000
TCHLISTEN	0.257	129.3	0.010
BULLIED	-0.976	37.7	0.000
MISCED	0.662	187.6	0.000
POORSTU	-0.007	99.3	0.015
REPEAT	-1.945	14.3	0.002

In the analysis of factors affecting reading performance, all variables are statistically significant with POORSTU highest p-value of 1.5 %, confirming their reliable influence on outcomes. The odds ratios were calculated as the exponentials of the coefficients. They reveal the extent to which each factor increases or decreases the likelihood of surpassing the average reading scores, all other things being equal.

Males are significantly less likely to perform above the average, with their odds of achieving higher scores being approximately 34.2% lower than those of females. This substantial disparity trend is confirmed in many countries surveyed in PISA. It is worth noting that the mathematics part is usually written similarly regarding gender.

Students who have access to a large number of books at home are much more likely to excel, with their odds of scoring above average increasing by 89.5%. This suggests the importance of access to cultural wealth in the family.

Living in a village appears to negatively impact reading performance, with these students having 36.1% lower odds of scoring above average compared to their counterparts in more urban settings. It might confirm the still existing gap between cities and villages in Poland.

The feeling of being listened to by teachers has a positive impact, enhancing students' odds by 29.3%. Conversely, being bullied has a profound negative effect on reading scores, with bullied students showing 62.3% lower odds of performing well. Those underscore the importance of positive school attitude of students.

Children of mothers who have higher educational levels show an 87.6% increase in their odds of surpassing the average reading scores. This indicates the influential role of parental, especially maternal education.

Disadvantaged socioeconomic background challenges reduce the likelihood of superior performance, with a mean decrease of 0.7 % odds for 1 percentage point increase in share of poor students in school.

Students who have repeated a year may face significantly tougher odds, with their likelihood of scoring above average plummeting by 85.7%. However, this severe reduction might seem expected, because of lower level of completed education.

3.2.2. Logistic regression math model

Table 3.3: Logistic Regression Results for Mathematics Performance. Source: own calculations

Variable	Coefficient	Odds Ratio (%)	p-value
INSTRUMENTHOME	0.265	130.4	0.008
CITY	0.262	129.9	0.028
BOOKSHOME	0.538	171.3	0.000
OWNCOMPUTER	-0.778	45.9	0.006
LATE	-0.334	71.6	0.003
MISEI	0.013	101.3	0.000
POORSTU	-0.009	99.1	0.001
REPEAT	-1.404	24.6	0.011

As well as in the case of the reading performance model, all modeled variables affecting mathematics performance are statistically significant with CITY highest p-value of 2.8 %.

Access to musical instruments at home shows a positive correlation with mathematics achievement, increasing the odds by 30.4%. This suggests that engaging in music could enhance cognitive functions that are beneficial in mathematical reasoning. Interestingly, this factor is specific only to the mathematics model.

Students in big cities exhibit a better probability of exceeding the mathematics average, with a 29.9% increase in odds. This aligns with findings from the reading model, where village dwellers fared worse, likely reflecting superior educational resources or opportunities available in urban areas compared to rural settings.

The presence of books at home remains a significant positive influence in both mathematics and reading performances, highlighting the universal correlation between educational performance and incentive to read books or cultural capital. For mathematics, the increase in odds is at 71.3%.

Owning a personal computer correlates with a 54.1% reduction in the odds of scoring above average, potentially indicating distractions or ineffective use of technology for educational pur-

poses. Similarly, regular tardiness is negatively associated with mathematical achievement, reducing the odds by 28.4%.

In mathematics, the mother's socioeconomic index has notable positive effect, enhancing performance in mean by 1.3% for 1 score point increase in index value. Disadvantaged socioeconomic background challenges reduce the likelihood of superior performance even more than for reading, with a mean decrease of 0.9 % odds for 1 percentage point increase in share of poor students in school.

Repeating a grade presents a dramatic negative impact on mathematics performance, decreasing the odds by 75.4%. This effect is consistent with the reading model, where academic repetition also leads to significantly lower performance, confirming the detrimental long-term impacts of failing to progress with peers.

3.2.3. Random Forest read model

The random forest reading performance model's sensitivity, enabled by the small node size of 2, allows it to pick up on the effects of variables that might otherwise be overlooked in larger-scale analyses. The limitation on the maximum number of nodes to 30 helps prevent overfitting, ensuring that the model remains generalizable across different sets of data. Meanwhile, employing 500 trees in the forest enhances the stability and accuracy of the predictions, averaging anomalies and reducing variance among the results.

In the analysis of reading performance using a random forest model, it's important to note that while the size or direction of the influence of variables cannot be directly discerned from the mean minimal depth or the frequency of interactions at the root, these metrics do provide valuable insights into the variables' importance and the complexity of their relationships. The mean minimal depth indicates how central an interaction is within the model—the shallower the depth, the more pivotal the variable is in the decision-making process, suggesting its key role in influencing outcomes. Similarly, the frequency with which variables appear at the root nodes across the ensemble of trees reflects their overall significance in the model. A higher frequency means that the variable consistently plays a crucial role across different subsets of the data, highlighting its robustness and relevance.

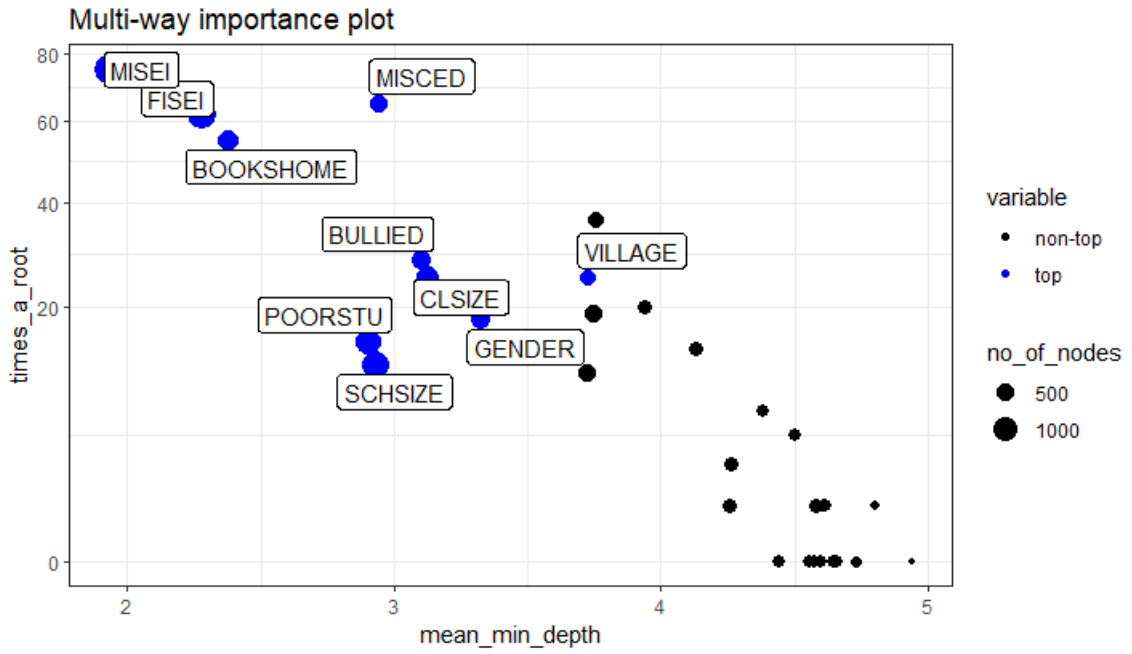

Figure 3.3: Variables importance plot of the read random forest. Source: own calculations

In this analysis, students' access to cultural wealth, such as having a significant number of books at home, emerges as a powerful factor in academic performance. Likewise, the educational attainment of parents, particularly mothers, strongly correlates with reading outcomes, indicating the transmission of educational values and resources across generations. Other influential variables are linked to the rural environment and the social climate of the school.

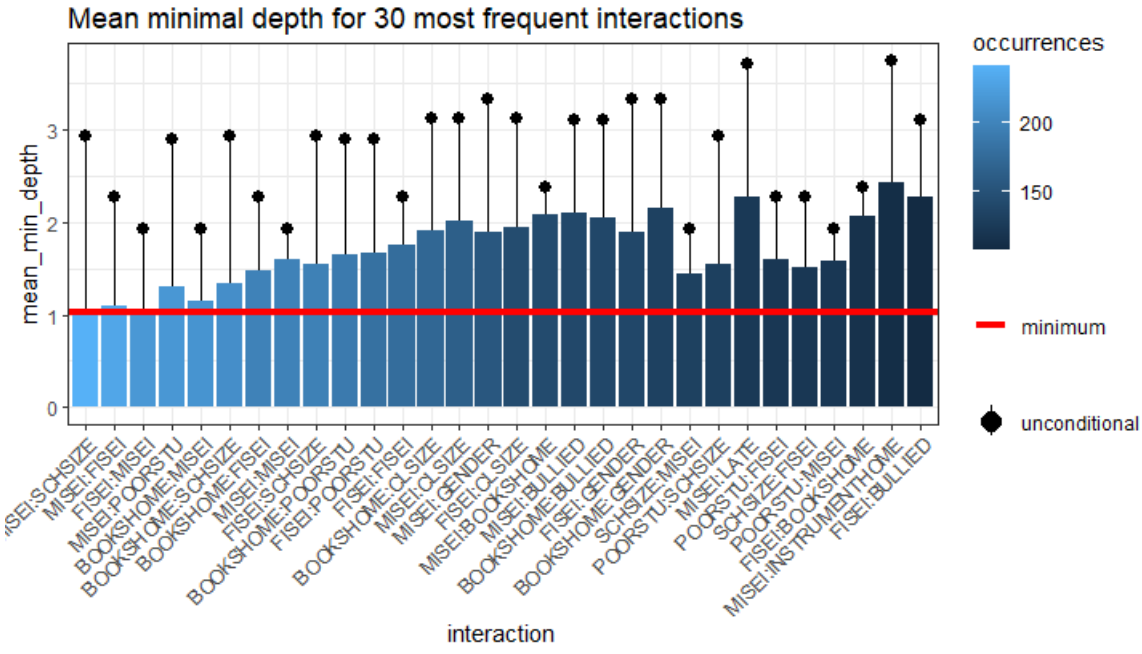

Figure 3.4: Interactions between variables in the read random forest. Source: own calculations

This modeling approach not only confirms the significant direct influences prevalently identified in previous logistic regression analyses but also may uncover intricate interactions among variables. For example, the combined impact of socioeconomic status, parental education, and access to educational resources outlines a complex framework within which students' reading performance can be understood.

Interactions involving parental socioeconomic status variables and home environment variables tend to occur at shallower depths and more frequently. This indicates their strong influence on the model's predictions, underscoring the importance of family background and home resources in influencing reading performance.

Most of the critical interactions appear above the red line, suggesting they occur at a depth that makes them significant but not at the very surface. This implies a balanced complexity in the model, where these factors are important but not overly dominant, allowing for a nuanced understanding of their effects.

Interactions with variables related to school environment highlight personal experiences and school context combination in impacting reading performance. The depth and frequency of these interactions suggest they are crucial but complex, requiring deeper tree levels to resolve.

3.2.4. Random Forest math model

The random forest model for mathematics performance, configured with a node size of 2, a maximum of 30 nodes, and an increased number of 1000 trees is similarly configured to the reading model. The larger number of trees in this model was needed to enhance the precision and stability of the predictions, further averaging anomalies and minimizing variance.

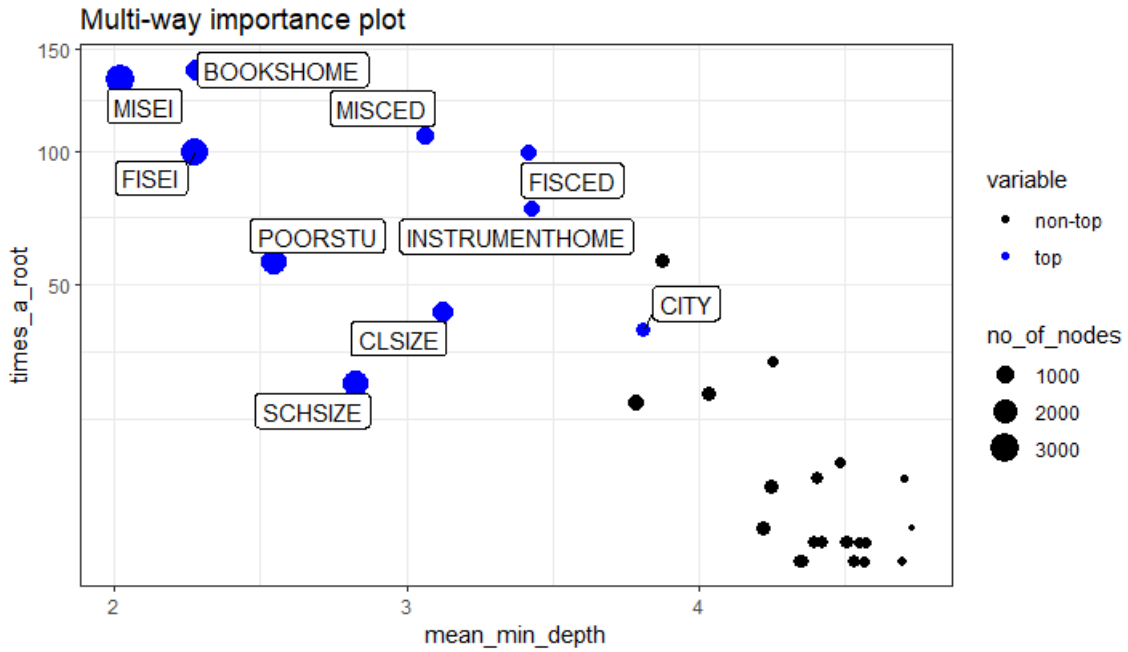

Figure 3.5: Variables importance plot of the math random forest. Source: own calculations

In this mathematical model, variables like parental education and access to learning resources remain significant, similar to the reading model. However, additional factors such as class size or school size emerge as more prominent in their impact on mathematical outcomes, reflecting the specific needs of mathematics education, such as the importance of individual attention and the learning environment. As well as in the logistic regression mathematical model, the correlation between having an instrument home and above-average math test performance remains solid.

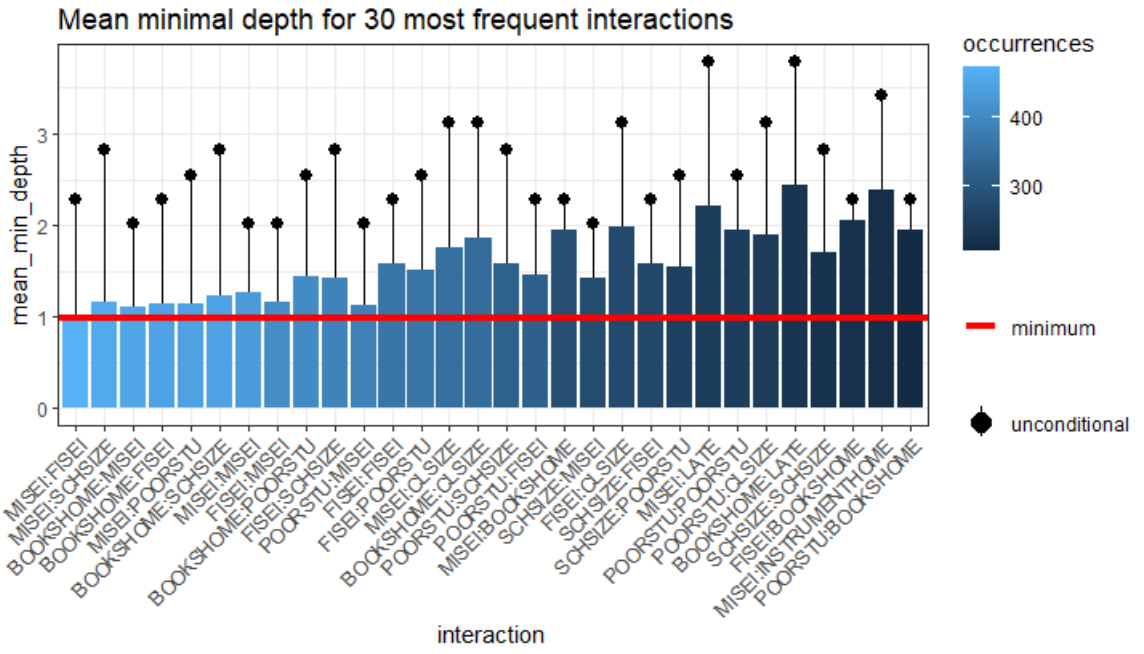

Figure 3.6: Interactions between variables in the math random forest. Source: own calculations

The model does not merely replicate findings from the logistic regression or the reading performance model; it explores and possibly unveils more intricate interactions. For instance, the combined effect of a student's access to technological resources at home and their socioeconomic background may delineate a complex network influencing mathematical proficiency.

3.3. Comparison of models' predictive power.

The comparison of the four models - two random forest models and two logistic regression models - across the subjects of mathematics and reading, allows us to analyze their predictive qualities in a structured manner. Each model was trained on the 2018 dataset and tested on the 2022 dataset, which helps clarify their ability to generalize and effectively predict student performance over time.

Table 3.4: Comparison of Model Performances on Mathematics Performance. Source: own calculations

Model	Accuracy	Sensitivity	Specificity	AUC
Random Forest	63.25%	19.83%	92.69%	0.646
Logistic Regression	62.31%	32.07%	82.81%	0.631

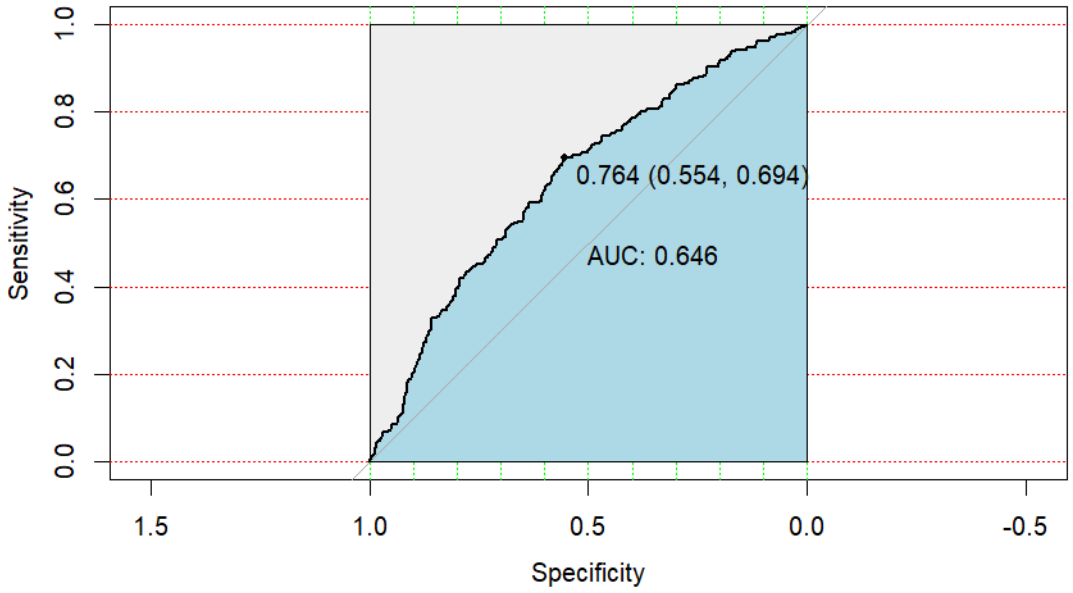

Figure 3.7: ROC curve - math random forest model. Source: own calculations

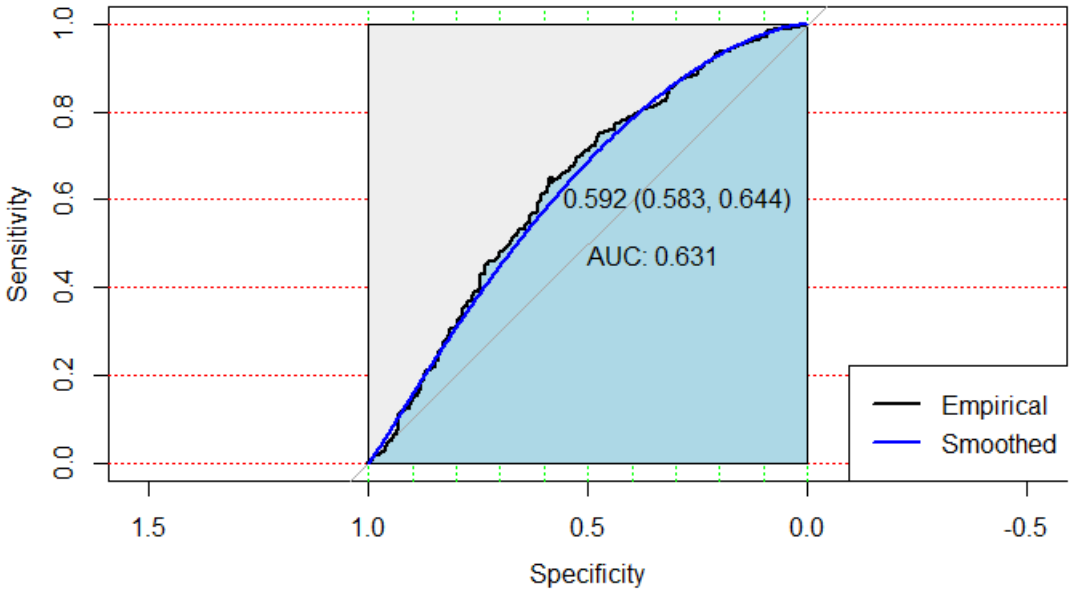

Figure 3.8: ROC curve - math logit model. Source: own calculations

Table 3.5: Comparison of Model Performances on Reading Performance. Source: own calculations

Model	Accuracy	Sensitivity	Specificity	AUC
Random Forest	62.54%	6.87%	98.83%	0.652
Logistic Regression	61.84%	8.36%	96.69%	0.639

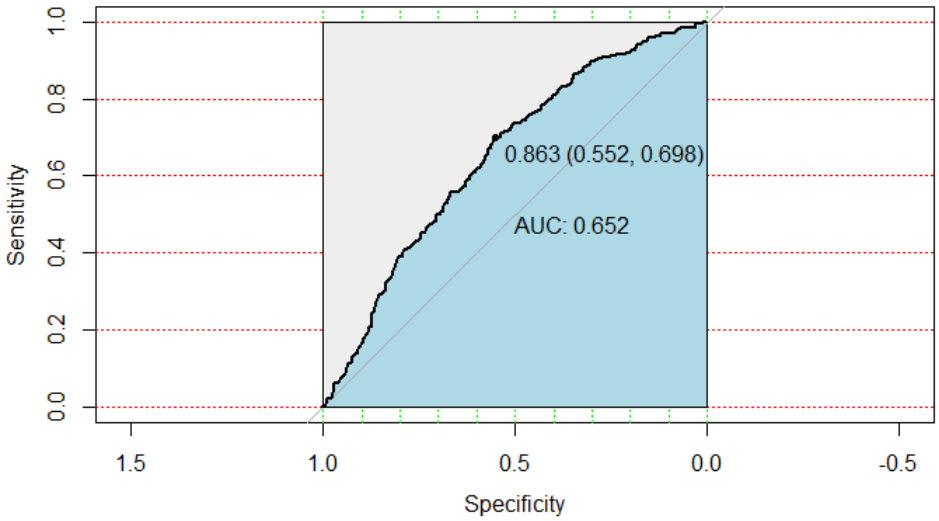

Figure 3.9: ROC curve - read random forest model. Source: own calculations

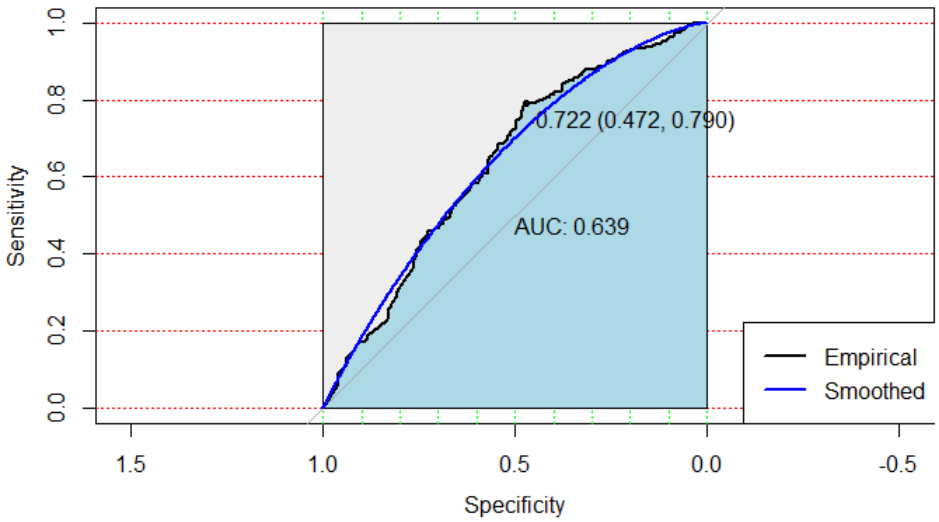

Figure 3.10: ROC curve - read logit model. Source: own calculations

The random forest model for mathematics exhibits an accuracy of 63.25 and an AUC of 0.646. It shows a high specificity of 92.69%, indicating strong performance in correctly iden-

tifying students who are not performing above average. However, its sensitivity is lower at 19.83%, suggesting it may miss identifying some high-performing students.

For reading, the random forest model achieves an accuracy of 62.54% with an AUC of 0.652. It has an extremely high specificity of 98.83%, superior to that of the mathematics model, which points to its efficiency in filtering out non-performers. However, the sensitivity remains severely low at 6.87%, consistent with the pattern observed in the mathematics model.

The logistic regression model for mathematics shows an accuracy of 62.31% and an AUC of 0.631. This model provides a balanced approach with a sensitivity of 32.07% and a specificity of 82.81%. While the sensitivity is higher compared to the random forest model, it achieves this at the cost of lower specificity.

In reading, the logistic regression model scores an accuracy of 61.84% and an AUC of 0.639. It shows a sensitivity of 8.36% and a very high specificity of 96.69%. Although it has better sensitivity than the random forest model for reading, it still falls short of providing a high rate of correct positive identifications.

When comparing these models, it becomes evident that random forest models are particularly strong in specificity, making them more suited for situations where it is more important to correctly identify students who do not meet a performance threshold. However, their lower sensitivity might limit their effectiveness in scenarios where identifying every potential high achiever is crucial.

On the other hand, logistic regression models, while slightly less accurate in overall terms, provide higher sensitivity, especially in the context of mathematics. This makes them suitable for more universal applications, especially taking into consideration their higher interpretability.

3.4. Final conclusions

The primary objective of the research was to assess the accuracy and stability of logistic regression and random forest models in predicting above-average performance in mathematics and reading PISA tests. Additionally, the research aimed to uncover any non-obvious factors that might influence the scores of Polish students.

During the course of the research, some intriguing relationships emerged that might not seem immediately apparent. For instance, the presence of musical instruments at home was found to be positively correlated with mathematics above-average scores, suggesting that engagement in music could be linked to enhanced mathematical ability, possibly due to the cognitive skills both activities share. Other notable dependencies that should be further explored include the impact of cultural capital factors such as the availability of books at home which consistently appeared as significant predictors of student performance. Research also surprisingly shows that owning a computer might hinder mathematics performance.

The findings from the study indicate that random forest models slightly outperform logistic regression models in terms of accuracy in predictions. This seems consistent with common sense as machine learning ensemble methods like random forests generally provide better performance due to their ability to model complex interactions and reduce overfitting through averaging multiple decision trees. Despite this, the performance of both model types was not fully satisfactory, particularly in terms of sensitivity. Both models showed a higher capability to predict under-average rather than above-average performance, suggesting that enhancing the sensitivity of these models could be a valuable direction for future research.

Neural networks with other ML approaches not yet explored extensively in PISA data setting could further elevate accuracy and related metrics as well. For example, neural networks might provide stronger predictive performance —especially having provided training on larger data. Finally, existing methods could always be extended with new and improved capabilities to approach gradient boosting as a special case of random forests.

One more possible explanation for the limited predictive qualities of the models may be the result of instability in pattern recognition across different PISA editions. External factors such as the COVID-19 pandemic likely introduced significant educational disruptions. These environmental disruptions may have altered student performance patterns, making it more challenging to generalize findings across different years.

Bibliography

Avvisati, F. (2020), *The measure of socio-economic status in PISA: a review and some suggested improvements*, Large-scale Assess Educ, 8(8), https://doi.org/10.1186/s40536-020-00086-x.

Gabriel, F., Signolet, J.,& Westwell, M. (2018), *A machine learning approach to investigating the effects of mathematics dispositions on mathematical literacy*, International Journal of Research & Method in Education, 41(3), 306–327, https://doi.org/10.1080/1743727X.2017.1301916.

OECD (2019a), *PISA 2018 Results (Volume I): What Students Know and Can Do*, PISA, OECD Publishing, Paris, https://doi.org/10.1787/5f07c754-en.

OECD (2019b), *PISA 2018 Results (Volume II): Where All Students Can Succeed*, PISA, OECD Publishing, Paris, https://doi.org/10.1787/b5fd1b8f-en.

OECD (2019c), *PISA 2018 Results (Volume III): What School Life Means for Students' Lives*, PISA, OECD Publishing, Paris, https://doi.org/10.1787/acd78851-en.

OECD (2023), *PISA 2022 Results (Volume I): The State of Learning and Equity in Education*, PISA, OECD Publishing, Paris, https://doi.org/10.1787/53f23881-en.

Pokropek, A., Marks, G. N., Borgonovi, F., Koc, P., & Greiff, S. (2022), *General or specific abilities? Evidence from 33 countries participating in the PISA assessments*, Intelligence, 92, https://doi.org/10.1016/j.intell.2022.101653.

PISA Database (2018), *Codebook 2018 and 2022; Student questionnaire data files 2018 and 2022; School questionnaire data file 2018 and 2022; Cognitive item data file 2018 and 2022*.

Rutkowski, D., Rutkowski, L., Wild, J., & Burroughs, N. (2017), *Poverty and educational achievement in the US: A less-biased estimate using PISA 2012 data*, Journal of Children and Poverty, 24(1), 47–67, https://doi.org/10.1080/10796126.2017.1401898.

Walczak, A., & Walczak, B. (2022), *Zaangażowanie rodziców w edukację dzieci. Polska na tle krajów Europy Środkowo-Wschodniej w świetle danych PISA 2018*, Studia z Teorii Wychowania, Tom XIII: 2022 NR 1(38), p. 196.

List of Figures

List of Tables

Summary

This study aimed to analyze the accuracy of logistic regression models and random forest models in the prediction of above-average performance in mathematics and reading PISA tests. Models were built based on data from the cognitive tests and questionnaires in the PISA Database for 2018 as the training period and 2022 as the testing period. The research determined slight advantages in terms of accuracy for random forests. However, logistic regression turned out to be more sensitive and disclosed non-obvious factors influencing the scores of Polish students regarding home possessions and feelings toward the school environment. In conclusion, the research underlines the complexity of factors impacting the above-average performance of Polish students and encourages further research using machine learning methods.

Keywords: education, PISA, random forest, logistic regression, classification