mathematics

MDPI

Article

Portability of Predictive Academic Performance Models: An Empirical Sensitivity Analysis

Jose Luis Arroyo-Barrigüete [1,*], Susana Carabias-López [1], Tomas Curto-González [1] and Adolfo Hernández [2]

[1] Department of Quantitative Methods, Pontifical University of Comillas, 28015 Madrid, Spain; scarabias@icade.comillas.edu (S.C.-L.); tcurto@icade.comillas.edu (T.C.-G.)
[2] Department of Financial and Actuarial Economics and Statistics, Faculty of Commerce and Tourism, Complutense University of Madrid, 28040 Madrid, Spain; adolfher@ucm.es
* Correspondence: jlarroyo@comillas.edu

Abstract: The portability of predictive models of academic performance has been widely studied in the field of learning platforms, but there are few studies in which the results of previous evaluations are used as factors. The aim of this work was to analyze portability precisely in this context, where preceding performance is used as a key predictor. Through a study designed to control the main confounding factors, the results of 170 students evaluated over two academic years were analyzed, developing various predictive models for a base group (BG) of 39 students. After the four best models were selected, they were validated using different statistical techniques. Finally, these models were applied to the remaining groups, controlling the number of different factors with respect to the BG. The results show that the models' performance varies consistently with what was expected: as they move away from the BG (fewer common characteristics), the specificity of the four models tends to decrease.

Keywords: mathematics education; university teaching; academic success; quantitative research; predictive models; portability

Citation: Arroyo-Barrigüete, J.L.; Carabias-López, S.; Curto-González, T.; Hernández, A. Portability of Predictive Academic Performance Models: An Empirical Sensitivity Analysis. *Mathematics* **2021**, *9*, 870. https://doi.org/10.3390/math9080870

Academic Editor: Heui Seok Lim and Kyu Han Koh

Received: 14 March 2021
Accepted: 13 April 2021
Published: 15 April 2021

1. Introduction

Early prediction of academic performance in university studies has been the subject of research in different contexts and from multiple perspectives. This is an area in which a growing amount of research is being developed, especially from the perspective of learning analytics (LA) (as Romero and Ventura [1] pointed out, both the LA and educational data mining (EDM) communities share a common interest in data intensive approaches to education research, LA being focused on the educational challenge and EDM on the technological challenge). The definition of LA, coined by the Society for Learning Analytics Research (SoLAR) and generally accepted in the literature, was given at the 1st International Conference on Learning Analytics and Knowledge: "Learning analytics is the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs". The area of LA has developed especially in the field of higher education, but the number of studies at other educational levels is increasing [2]. Many of the publications in this area involve environments where technology has special prominence and where it is possible to collect a large amount of information, such as technology-enhanced learning (TEL) [3], but the scope of this paper is not limited to a technological context.

There are different research communities with which LA finds common elements, among which so-called educational data mining and academic analytics can be highlighted. Different studies have developed analogies across and explored the differences in these areas of research [4,5]. The fundamental differential factor of the LA approach is the prior objective of supporting teachers in the design of interventions that contribute to improving educational processes; thus, studies are focused on courses, subjects or departments. To

avoid leading to a reductionist approach, these interventions need to be based on the principles of formative assessment [6]. At the current stage, the LA literature highlights the importance of developing studies with a direct impact on teaching practices [2].

Among the different types of analyses developed within LA [7], this work is focused on the identification and evaluation of factors as indicators of performance for the purposes of prediction. The factors taken into consideration in such models mostly fall under three categories [8]: student interactions with their learning management system (LMS), demographic characteristics (such as age and gender) and grades (in previous courses and the results of partial assessments of the subject under study). This last type of factor is precisely the one selected in this work.

Final exam results are taken as a measure of academic performance. Bearing in mind that academic success or failure are constructs that are not straightforward or simple to measure [9,10], grades are widely accepted as indicators of performance [11]. The prediction model is based on observable indicators and not directly on an underlying theoretical model of academic performance, but the implementation of prediction-based interventions should fit the methodology of the course and the discipline to which it belongs [12]. In this sense, the teaching methodology in the subjects under study in this work follows an approach that integrates direct instruction and constructivist theory [13–16]. Neither of these two theories seems satisfactory in isolation [17–19]. From the context of the subjects, in the first stage of the degree, the arguments that support direct teaching are relevant [20]: novice students at elementary levels, in constructivist pedagogical practice, run into problems that hinder learning because they do not have the experience to see things in their full complexity.

However, the university in which the subjects are taught understands that the inclusion of constructivist elements is positive because it boosts student learning-centered teaching [21–24]. This is an objective aligned with the majority of approaches to pedagogy in recent decades [25]. The elements of constructivist theory that are incorporated into the subjects are essentially tasks that students develop outside of the classroom (the design of these tasks includes working with real-world data and information and tries to meet the requirements of authenticity [26]. For instance, in the subject financial mathematics, the students have to study a real loan and develop the mathematical formalization according to the progression of their learning). The objective of this type of task, metaphorically called scaffolding [27–29] in constructivist theory, is the transfer of responsibility to the student for his/her learning process; for this transfer to be efficient, it must be adapted to the needs of the student. The intended early identification of students with a high probability of failure in this study would make it possible to design scaffolding adapted to a context narrower than that of the group as a whole [29,30].

From the identification of students at risk for failure, the instructors can design other interventions tailored to their situation [31]. They can define differentiated communication strategies, which can start with personal notifications of the risk situation and the collection of information on the perceptions of the student [32]. Since the factors taken into account are the results of previous evaluations, the communication derived from the analysis can be interpreted as a complement to the rating, and the formative evaluation literature can be taken into consideration [33,34].

The fundamental contribution of this work, with regard to previous studies that take this approach [35], is the incorporation of the conditions for the application of the model in different groups. First, a methodology was selected that can be implemented by all teachers, as detailed in Section 2. Second, a portability study was developed; this study is the heart of the work. For the model to have predictive utility when teachers intervene, it is essential that it can be applied to groups other than the initial one [36]. The portability of predictive models of academic performance has been extensively studied in the analyses of data related to the use of learning platforms [37–39], but there are few studies in which the results of previous evaluations are taken as factors.

The goal of this study was to analyze the portability of predictive academic performance models based on evaluation results. The study focused on the search for model portability limits that guarantee reasonable performance. To that end, the loss of specificity was evaluated when the difference between the group to which the model was applied and the group for which it was estimated increased. The article is structured as follows: Section 2 contains the details of the considered sample, as well as the design of the study developed, providing insights into the treatment of the main confounding variables and the statistical techniques used. Results are presented in Section 3. Finally, these results and their implications are discussed in Section 4, and the main conclusions of this work are analyzed in Section 5.

2. Materials and Methods

2.1. Participants

This research used panel data on a set of 170 undergraduate students from four different classes of the same degree evaluated over two academic years. Considering the same degree is an important aspect because as pointed out by [37], portability worsens notably when courses of different degrees are included. For reasons outlined later, related to the analysis of sensitivity to portability, these groups were aggregated into three groups: the base group (BG); group 2 (G2), formed by two different classes of students pursuing a degree in business taught in Spanish; and group 3 (G3), formed by students pursuing a degree in business taught in English. The composition of these groups was not random but responded to other criteria. For the business degree taught in Spanish, the classes were formed following the alphabetical order of the students' surnames, since this is the criterion used by the university to assign students. For the business degree taught in English group, it was the students themselves who decided to opt for this group, being accepted or not according to their level of English.

The subjects, on which the predictive models were developed, have a significantly higher number of students with approximately 450 per year, but it was not possible to use them due to the requirements imposed in the study to control the main confounding factors. The first of these factors, which is absolutely critical, is the profile of access to university, not only with regard to the level of previous instruction but also with regard to the university degree, which can affect the degree of interest in certain subjects. For this reason, it was decided to work only with students of the same degree: business administration and management. Another key confounding factor is the exigency level of different groups. Among other reasons, the differences may be due to the particularities of each teacher or to variations between groups taking the subject in different years. This led to the choice of two specific subjects from the quantitative methods department: business statistics and financial mathematics. On the one hand, they are taught by teachers from the same department, which, a priori, should mean greater homogeneity. On the other hand, in these subjects, there is strong coordination, which forces teachers to follow exactly the same syllabus, to use the same teaching materials, to have an identical evaluation system in all groups and to carry out intermediate evaluation tests that are conducted at the same time and that are very similar; these tests are also usually supervised by the coordinator of the subject to guarantee a similar exigency level. In addition, the final exam, prepared by the coordinator, is identical for all groups, with the only difference being the language of writing, in the case of the group taught in English. This is a key element, and it ensures a practically identical exigency level, at least for one particular year. In the case of this paper, by using panel data and, therefore, comparing the same final exam for all groups in each subject, we ensured that we measured performance under practically identical conditions. Both subjects, of 6 European Credit Transfer System (ECTS) credits, are taught in the first semester of the second year in the business administration and management degree program.

2.2. Materials

Once these factors were controlled, different predictive models of academic performance were formulated, under the premise of being relatively simple due to two reasons. The first is that for such models to be useful for practical purposes, they should be able to be run by teachers without specific knowledge of programming or statistical methods. This implies that, apart from the fact that this research was developed entirely in R, a free programming environment oriented toward statistical analysis [40], the resulting algorithms should be able to be implemented in Excel, the only tool that all teachers in different knowledge areas know and can use fluently. More sophisticated algorithms such as TrAdaBoost or AdaBoost [41], which can work better, have the disadvantage that they cannot be implemented by teachers without programming knowledge.

Additionally, the models should require a reduced amount of prior information since on many occasions, the information needed to adjust the predictive models, despite being obtainable for research development, is not available to all teachers, limiting the subsequent use of the models. One of the key aspects when developing predictive models of academic performance is to identify the possible causal factors, which is a rather complex task as one of the latest meta-analyses on this topic [42] collected a total of 105 different variables, encompassing two major areas: those related to the student and those linked to the instructional process. In the case of this paper, as already indicated, the choice of variables presents an important restriction: the variables must contain information that can be obtained relatively easily by any teacher. This implies that certain variables, despite their potential importance for academic performance, must be discarded in the construction of the model. This is the case, for example, in research on parents [43]; although parents have been identified as a relevant variable, information on them is not easily accessible. For this particular study, it would be possible to obtain such information, but this would imply the impossibility of using the model developed, as it would not be possible to access such information regularly in other groups. Fortunately, the variables that the literature identifies as especially relevant can be accessed in a simple way. First, we found information on pre-university performance, which several studies identified as one of the indicators with the greatest predictive capacity in university students [10,44,45]. This is to be expected since it is an indicator that synthesizes, on the one hand, the skills and effort capacity of the student and, on the other hand, his/her initial knowledge. In the specific case of degrees related to the economy and business in Spain, García-Diez [46] concluded that the score obtained for selectivity, currently the EvAU, is useful for the purpose of selecting students with the greatest probability of success in their university studies. In the Spanish university system, the EvAU is the equivalent of the SAT, in the sense that the score obtained is used by public universities to select their future students. The EvAU is not a standardized test but is elaborated by a commission formed by university professors who are experts in each subject, often in collaboration with high school teachers. It is normally carried out over three days in mid-June on the same date for all students in the country, since it consists of as many exams as subjects have been taken during the baccalaureate, so it is a considerably long test. In order to gain access to the university, it is necessary to pass the EvAU and, depending on the grade obtained, the student can choose a university degree with a limited number of places that he/she wants to study according to the cut-off grade (minimum grade is used as a limit to gain access to a certain degree before all the places offered are filled) established for each degree and university.

On the other hand, in this type of degree, good performance in the subjects of mathematics at university positively correlates with good performance in those of economics [47] and finance [48]. Ballard and Johnson [49] pointed out that quantitative skills are a key factor in performance in an introductory course in microeconomics, while Girón Cruz and González Gómez [50] concluded that good results in the area of economics are explained, to a large extent, by previous performance in mathematics. Thus, it seems that performance in mathematical subjects is a good predictor of performance in other subjects related to economics and business. Therefore, the elements mentioned above, together

with the results of early and late assessments in each subject, seem to be not only relevant for the purpose of explaining academic performance but also relatively simple to obtain, meaning that their incorporation into a general model of academic performance is feasible. Thus, the following variables were considered: the EvAU score (EvAU), the first course grade (1st course grade), the grade in the subject of mathematics 1 in the first year (math. 1 grade), the grade in the subject of mathematics 2 in the first year (math. 2 grade), the grade obtained in the early assessment of the considered subject (early assessment) and the grade obtained in the late assessment (late assessment).

Based on the considerations of ease of use and easily accessible prior information, we chose logit-type models, incorporating the interaction between the indicated variables. In this sense, it is necessary to mention that a serious problem of imperfect multicollinearity was detected as a result of incorporating the mentioned interactions, generating very high variance inflation factors (VIFs). Thus, following the recommendations of [51,52], the variables were normalized by subtracting the average value. After this transformation, the problem disappeared, and VIFs significantly lower than 10 were obtained. This aspect is key since the absence of multicollinearity between independent variables is one of the basic assumptions of logistic regression, which is often ignored [53].

2.3. Procedure

Starting from six potentially relevant variables, for the BG in the subject of statistics, logit models resulting from all possible combinations between them, two by two, with the corresponding interactions, three by three, in this case without interaction, were adjusted. Combinations of four or more variables were not incorporated due to the size of the sample since, as Ortega Calvo and Cayuela Domínguez [54] pointed out, an acceptable adjustment requires at least 10 data points for each parameter to be estimated plus one; for this reason, we were forced to limit the model to a maximum of three variables. A total of 60 different models were successively tested, and the 4 best performing models were chosen, which are formulated as follows:

$$Logit = \left[\frac{p}{1-p}\right] = \beta_0 + \beta_1 \cdot \textbf{Early evaluation} + \beta_2 \cdot \textbf{1st course grade}$$

$$Logit = \beta_0 + \beta_1 \cdot \textbf{Early evaluation} + \beta_2 \cdot \textbf{1st course grade} + \beta_3 \cdot \textbf{Early evaluation} \cdot \textbf{1st course grade}$$

$$Logit = \beta_0 + \beta_1 \cdot \textbf{Early evaluation} + \beta_2 \cdot \textbf{Math. 1 grade}$$

$$Logit = \beta_0 + \beta_1 \cdot \textbf{Early evaluation} + \beta_2 \cdot \textbf{Math. 1 grade} + \beta_3 \cdot \textbf{Early evaluation} \cdot \textbf{Math. 1 grade}$$

In each case, to avoid overfitting problems, the k-fold cross-validation algorithm with 10 categories was applied. This method was chosen instead of leave one out cross-validation (LOOCV), which is also very frequently applied when the groups are not very large because it has been empirically demonstrated that it achieves a better bias–variance balance [55]. For each model, precision (correctly predicted cases) was estimated for two threshold values (TVs): 0.5 and 0.9. The reason is that, for the purpose of using performance models, the main application is the early detection of academic failure; thus, it is interesting to obtain high specificity (percentage of correctly predicted failing grades), even partially sacrificing sensitivity (percentage of correctly predicted passes). Therefore, it makes sense to work with a high TV of, in this case, 0.9. Having done this and after comparing all the models with the majority rule, which in this case corresponds to 82% of passes, the four models with the best performance were selected. As a measure of goodness of fit, Nagelkerke's R2 and receiver operating characteristic (ROC) curves were used with R's DescTools [56] and pROC [57] packages, respectively. Additionally, the absence of autocorrelation in the residuals was verified by means of the Durbin–Watson test.

Subsequently, the performance of the four models with respect to the remaining groups was analyzed and divided by their similarity to the BG in statistics, as shown in Table 1. Thus, for example, we observe that the G3 statistics shares two elements with the reference group, the same subject and the same teacher, with the group and the language being different. By stratifying the groups in this way, we can study how the performance of the predictive model evolves according to the common and different elements. A priori, it would be expected that with more elements in common, the performance of the model would be more similar to that presented for the BG. In all cases, and for the different TVs, the precision, sensitivity and specificity were calculated, and the results were subsequently compared.

Table 1. Groups by subject and similarity to the BG.

	Group	**Subject**	**Professor**	**Language**
BG (Statistics)	Reference group			
BG (Fin. Mathematics)	Yes			Yes
G2 (Statistics)		Yes		Yes
G3 (Statistics)		Yes	Yes	
G2 (Fin. Mathematics)				Yes
G3 (Fin. Mathematics)				

3. Results

To verify that all students started from a similar academic level, their university entrance qualification, the EvAU, was studied. The hypothesis of normality of the three groups was evaluated by means of the Shapiro–Wilk test, and the homogeneity of variances was subsequently verified by means of the Brown–Forsythe–Levene test. After checking both hypotheses, an ANOVA was performed to evaluate the possible differences in the access scores. Table 2 shows the results of the analysis, confirming that there are no statistically significant differences in the access profiles. Additionally, it was verified that the percentage of students arriving at university with a baccalaureate in science was similar in the three groups since if there were large differences, a certain bias could be generated [58], given that the subjects selected for the study belong to the area of quantitative methods (in the Spanish university system, high school students can choose different branches of specialization. The science major is for students who wish to pursue a STEM-style degree, while the social sciences and humanities major is recommended for those who will choose a degree in those fields, which includes the business administration degree). In this case, a comparison of proportions was made using the stats package of R [40], which allows a comparison of the equality of proportions between several groups. The null hypothesis in this case is the equality of proportions, and while a p-value of 0.63 was obtained, it was confirmed that there are no significant differences between the three groups.

Table 2. Comparison of EvAU scores for the three groups considered.

		High School		Comparison of EvAU Grades								
				Grade		Shapiro Test		Levene Test		Anova		
	N	**Science**	**Others**	**Mean**	**SD**	**W**	***p*-Value**	**F**	***p*-Value**	**F**	**FG**	***p*-Value**
BG	39	43.6%	56.4%	7.76	0.90	0.98	0.65	1.48	0.22	0.76	3	0.52
G3	52	34.6%	65.4%	7.84	0.86	0.97	0.16					
G2	79	35.4%	64.6%	7.93	0.88	0.98	0.22					

Once these factors had been controlled, different predictive models of academic performance were formulated. Table 3 shows the results of the logit models for the BG, resulting from all the possible combinations between the six variables identified, two by two, with the corresponding interactions, and three by three, in this case without interaction. The results corresponding to the four models with the best performance, that is, a higher average precision for both TVs, are marked in bold. As already indicated, the majority rule is, in this case, 0.82, which corresponds to the percentage of passes. It can be seen that most of the models obtain high levels of precision that are above this limit and that are quite robust to the modification of the TV. This is an interesting result, as only three variables are being used in each of the models. With regard to the variables themselves, there are two effects that should be highlighted. First, early assessment is a good predictor of performance, even better than late assessment. Although it may seem paradoxical, this result is beneficial for the purposes of this paper since it allows for the incorporation of this variable into the model and, therefore, the early detection of students at risk. On the other hand, the average grade in the first year and the grade in mathematics 1 appear to be particularly good predictors. With regard to the average grade, it is probably a measure capable of capturing the previous performance of the student even better than the EvAU scores, both because it is more recent and because it specifically reflects performance at university. With regard to the mathematics 1 grade, as already indicated, the academic literature indicates that performance in mathematics is a good predictor of performance in other subjects. The fact that it performs better than the grade in mathematics 2 may be because this second subject presents a lower level of complexity and exigency, discriminating less and, therefore, being a less relevant variable.

Table 3. Average precision (10-fold cross-validation) of the different models for threshold values (TVs) of 0.5 and 0.9. Variables combined three by three.

		EvAU		1st Course Grade		Math. 1 Grade		Math. 2 Grade		Early Evaluation	
		TV = 0.5	**TV = 0.9**	**TV = 0.5**	**TV = 0.9**	**TV = 0.5**	**TV = 0.9**	**TV = 0.5**	**TV = 0.9**	**TV = 0.5**	**TV = 0.9**
1st course grade	-	0.83	0.78								
	Interaction	0.83	0.78								
	Math. 1 grade	0.86	0.76								
	Math. 2 grade	0.83	0.84								
	Early evaluation	0.88	0.83								
	Late evaluation	0.83	0.81								

Table 3. *Cont.*

		EvAU		1st Course Grade		Math. 1 Grade		Math. 2 Grade		Early Evaluation	
		TV = 0.5	**TV = 0.9**	**TV = 0.5**	**TV = 0.9**	**TV = 0.5**	**TV = 0.9**	**TV = 0.5**	**TV = 0.9**	**TV = 0.5**	**TV = 0.9**
Math. 1 grade	-	0.86	0.76	0.87	0.74						
	Interaction	0.89	0.80	0.92	0.74						
	1st course grade	0.86	0.76								
	Math. 2 grade	0.83	0.84	0.87	0.69						
	Early evaluation	0.94	0.81	0.93	0.83						
	Late evaluation	0.87	0.79	0.90	0.77						
Math. 2 grade	-	0.81	0.78	0.84	0.61	0.88	0.58				
	Interaction	0.81	0.78	0.87	0.58	0.86	0.65				
	1st course grade	0.83	0.84								
	Math. 1 grade	0.81	0.78								
	Early evaluation	0.81	0.83	0.90	0.80	0.91	0.86				
	Late evaluation	0.83	0.81	0.88	0.79	0.88	0.72				
Early evaluation	-	0.86	0.86	**0.93**	**0.85**	**0.93**	**0.88**	0.85	0.83		
	Interaction	0.86	0.86	**0.93**	**0.90**	**0.93**	**0.86**	0.88	0.83		
	1st course grade	0.88	0.83								
	Math. 1 grade	0.94	0.81								
	Math. 2 grade	0.81	0.84								
	Late evaluation	0.84	0.81	0.90	0.86	0.88	0.82	0.83	0.82		
Late evaluation	-	0.81	0.81	0.88	0.78	0.88	0.74	0.78	0.62	0.85	0.84
	Interaction	0.83	0.81	0.85	0.78	0.88	0.77	0.77	0.59	0.88	0.87
	1st course grade	0.83	0.81								
	Math. 1 grade	0.87	0.79								
	Math. 2 grade	0.83	0.81								
	Early evaluation	0.84	0.81								

Table 4 shows the detailed results for the four best performing models, already using a fit with all BG data. In all four cases, the parameter values and their corresponding odds ratio, as well as the p-value, were calculated. Likewise, the VIFs were computed to verify the absence of multicollinearity problems, and the absence of autocorrelation in the residuals was verified by means of the Durbin–Watson test. It can be seen that in all cases, Nagelkerke's R2 has relatively high values, and the ROC curves (Figure 1) also show a high AUC above 0.9, which is usually considered to be a good result. Specifically, values of 0.93, 0.95, 0.92 and 0.93 were obtained in each of the four models. Therefore, we can conclude that all of them present good behavior in the BG.

Table 4. Adjustment of the four selected models, including variance inflation factors (VIFs) and the Durbin–Watson autocorrelation test.

Model 1	Parameter	Odds	z Value	*p*-Value	VIF	DW Test	
						DW	*p*-Value
Constant	4.51	90.92	2.83	4.7×10^{-3}			
1st course grade	3.09	21.91	2.06	0.04	1.02	1.92	0.68
Early evaluation	0.57	1.77	2.30	0.02	1.02		
R2 Nagelkerke	0.68						
Model 2	**Parameter**	**Odds**	**z Value**	***p*-Value**	**VIF**	**DW Test**	
						DW	***p*-Value**
Constant	4.01	55.29	3.12	1.8×10^{-3}			
1st course grade	1.88	6.54	1.50	0.13	1.10	1.94	0.70
Early evaluation	0.13	1.14	0.35	0.73	1.90		
Interaction	−0.79	0.45	−1.45	0.15	1.79		
R2 Nagelkerke	0.72						
Model 3	**Parameter**	**Odds**	**z Value**	***p*-Value**	**VIF**	**DW Test**	
						DW	***p*-Value**
Constant	2.96	19.30	2.90	3.7×10^{-3}			
Math. 1 grade	1.31	3.70	2.05	0.04	1.01	2.15	0.67
Early evaluation	0.51	1.66	2.26	0.02	1.01		
R2 Nagelkerke	0.63						
Model 4	**Parameter**	**Odds**	**z Value**	***p*-Value**	**VIF**	**DW Test**	
						DW	***p*-Value**
Constant	3.51	33.45	2.87	4.1×10^{-3}			
Math. 1 grade	0.98	2.66	1.56	0.12	1.64	2.04	0.98
Early evaluation	0.31	1.37	1.12	0.26	1.32		
Interaction	−0.57	0.57	−1.88	0.06	1.30		
R2 Nagelkerke	0.74						

Table 4 shows the detailed results for the four best performing models, already using a fit with all BG data. In all four cases, the parameter values and their corresponding odds ratios, as well as the p-values, were calculated. Likewise, the VIFs were computed to verify the absence of multicollinearity problems, and the absence of autocorrelation in the residuals was verified by means of the Durbin–Watson test. It can be seen that in all cases, Nagelkerke's R2 has relatively high values, and the ROC curves (Figure 1) also show a high AUC, above 0.9, which is usually considered to be a good result. Specifically, values of 0.93, 0.95, 0.92 and 0.93 were obtained in each of the four models. Therefore, we can conclude that all of them present good behavior in the BG.

In Table 5, the performance of the four models adjusted with respect to the remaining groups is analyzed, indicating the differences and similarities between them and the BG (as in Table 4, for the BG, as well as for the other groups, the models were adjusted using all available data. For this reason, the accuracies of the GB do not match those of Table 3, where the average sensitivity was calculated using k-folds). In all cases, the precision, sensitivity and specificity were calculated, marking in bold those instances in which the majority rule

is not exceeded for the group considered, meaning that they cannot be considered valid models. This rule, corresponding to the percentage of passes, is 82%, 64%, 71%, 77% and 59% for the five groups. As we have already indicated, given that the main interest of these models is the early prediction of students at risk, the most relevant parameter is specificity; thus, the analysis must fundamentally focus on behavior for a high TV. This leads us to discard models 3 and 4, as they do not exceed the majority rule in two of the subjects for a TV of 0.9 (as a test, the results were replicated by incorporating logarithmic transformations in the independent variables to capture possible nonconstant marginal effects, but the results did not improve. Specificity improved in some instances but worsened in others).

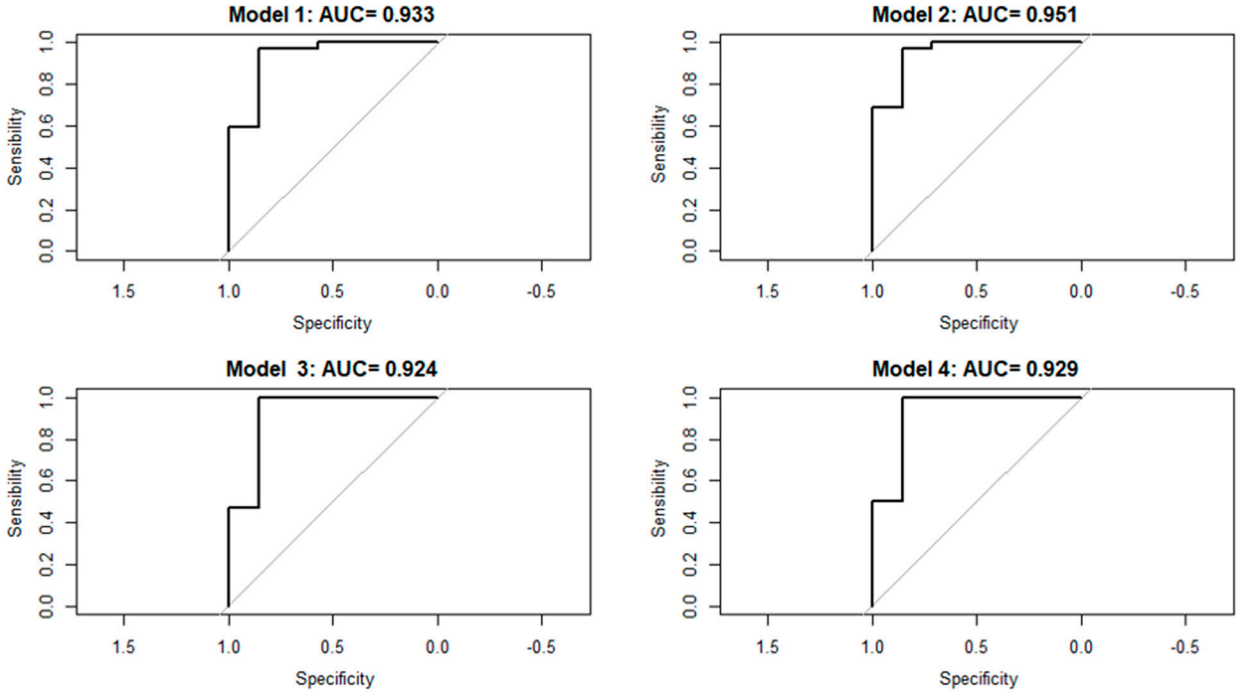

Figure 1. ROC curves for the four selected models.

Table 5. Model behavior (precision, sensitivity and specificity) for different groups using different TVs.

	Group	Subject	Professor	Language		Model 1	Model 2	Model 3	Model 4
BG (Statistics) Pass: 82%					TV: 0.5	(92%/94%/86%)	(95%/97%/86%)	(92%/94%/86%)	(97%/100%/86%)
					TV: 0.9	(87%/88%/86%)	(87%/88%/86%)	(85%/84%/86%)	(87%/88%/86%)
BG (Fin. Math.) Pass: 64%	Yes			Yes	TV: 0.5	**(59%/92%/7%)**	**(61%/96%/7%)**	(72%/100%/27%)	(70%/100%/20%)
					TV: 0.9	(74%/79%/67%)	(74%/83%/60%)	(95%/92%/100%)	(79%/88%/67%)
G2 (Statistics) Pass: 71%		Yes		Yes	TV: 0.5	(78%/96%/40%)	(77%/98%/32%)	(78%/94%/44%)	(76%/94%/36%)
					TV: 0.9	(86%/94%/68%)	(84%/96%/60%)	(78%/81%/72%)	(84%/93%/64%)
G3 (Statistics) Pass: 77%		Yes	Yes		TV: 0.5	(79%/95%/25%)	(77%/95%/17%)	**(75%/83%/50%)**	(83%/95%/42%)
					TV: 0.9	(83%/90%/58%)	(83%/93%/50%)	**(71%/68%/83%)**	**(75%/78%/67%)**
G2 (Fin. Math.) Pass: 59%				Yes	TV: 0.5	(67%/100%/21%)	(66%/100%/18%)	(70%/99%/30%)	(68%/99%/27%)
					TV: 0.9	(77%/98%/48%)	(74%/97%/42%)	(67%/78%/52%)	(68%/89%/39%)
G3 (Fin. Math.) v: 77%					TV: 0.5	(83%/98%/33%)	(81%/100%/17%)	**(73%/88%/25%)**	**(75%/93%/17%)**
					TV: 0.9	(83%/95%/42%)	(81%/93%/42%)	**(75%/80%/58%)**	**(65%/75%/33%)**

Finally, it can be seen that the performance of the models, specifically their specificity, varies consistently with what is expected; this is a particularly relevant result. As they move away from the BG (fewer characteristics in common), the specificity tends to decrease in the two final models. In other words, portability worsens as the groups studied are less similar to the group used to calibrate the models (Figure 2). In this sense, the fact that the best results beyond the BG (statistics) are obtained for the same group in a different subject (BG financial mathematics) is very relevant. It confirms that the group is the most relevant variable for the purposes of portability, even more than the subject, teacher or language of instruction.

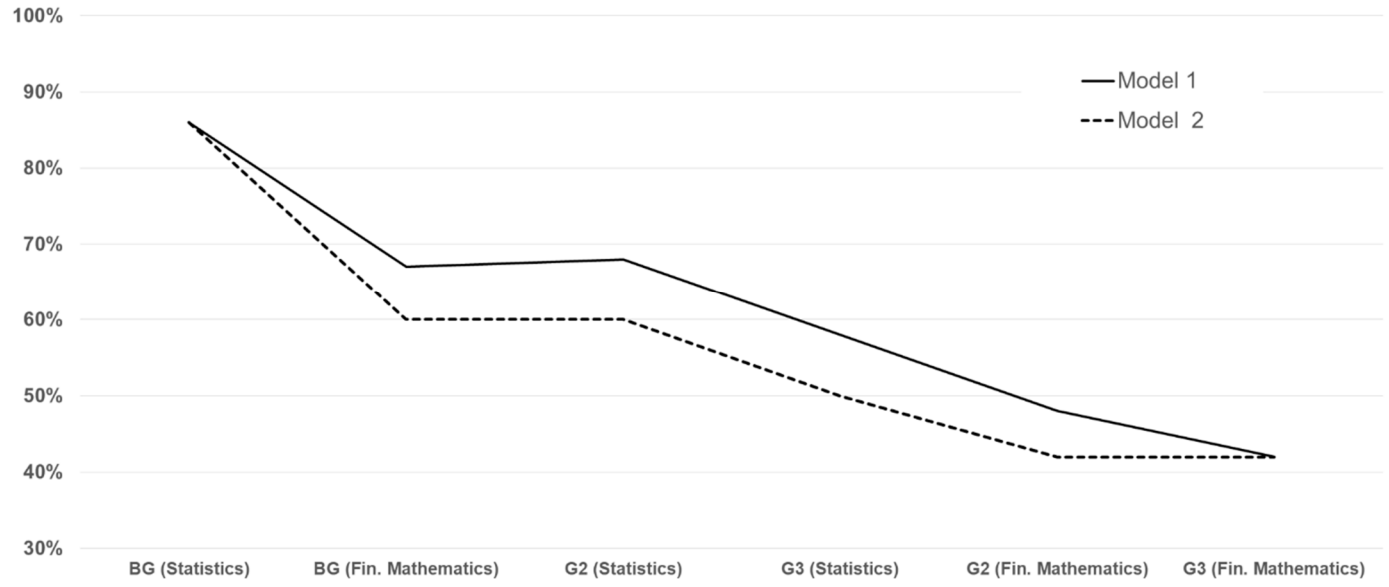

Figure 2. Specificity of models 1 and 2, for a TV of 0.9, in each of the groups.

4. Discussion

As mentioned above, for the model to have predictive utility when teachers intervene, it is essential that it can be applied to groups other than the one in which it was estimated, hence, the relevance of analyzing the portability of predictive models of academic performance. Studies by Widyahastuti and Tjhin [59] and Thakar, Mehta and Manisha [60], analyzing academic publications between 2011 and 2016, and 2002 and 2014, respectively, pointed to the need to search for unified approaches that allow the development of universal models. In the same vein, Muthukrishnan, Govindasamy and Mustapha [61], based on a review of 59 articles on predictive models of student performance, concluded that there is a huge shortage of portable predictive models, something confirmed by, among others, by Gasevic et al. and Gitinabard et al. [12,38]. Even analyzing different courses within the same institution, there are important differences between them, requiring predictive models tailored to each of them.

In this paper and based on a set of 170 students from four different groups pursuing the same degree, evaluated over two academic years, the portability of different predictive models of academic performance was studied. In the study, the main confounding variables were controlled, such as the university entrance qualification, the percentage of students arriving at university with a baccalaureate in science and the exigency level in the groups. After selecting the four models with the best performance in the BG, their performance in the remaining groups was evaluated. The results confirm that the performance of the models tends to decrease as the groups share fewer characteristics with the BG. It is true that the success rate remains at acceptable levels, but for the purpose of early detection of academic failure, the critical variable is specificity, and in regard to this variable, we observe a notable worsening as we move away from the BG. Thus, the enormous

difficulty, and probable impossibility, of developing universal predictive models seems to be confirmed, at least for using the results of previous evaluations as the only factors. The result coincides with that of previous research that, using data from learning platforms, concluded that the models have limited portability [38,39]. Therefore, the conclusion is that the limited portability of predictive models of academic performance is not due to the approach adopted.

5. Conclusions

Our results confirm the enormous difficulty of developing universal predictive models using the results of previous evaluations as the only factors. On the other hand, previous research pointed out a similar problem when considering data from learning platforms. A mixed approach could perhaps lead to better results in terms of portability of universal models. However, our results suggest that even in groups that are remarkably similar in both learner profiles and instructional factors, there appear to be notable differences. Somehow, the internal dynamics of each group, perhaps related to the social interactions among students, generate substantial differences that make the portability of predictive models of academic performance very difficult.

The main weakness of this work is a consequence of the methodological approach employed. Having established such a high level of demand in terms of the confounding factors, it was possible to work with only four groups, as they were the only groups that met all the requirements of the design. This has led to working with only 170 students over two academic years. In this sense, and as a future line of research, we propose replicating this analysis with a greater number of students and using other subjects. A second line of research consists of designing interventions adapted to the students' situations. Currently, the same team that authored this work is designing a first intervention that would use these models of early detection of academic failure to provide students with continuous feedback on their chances of passing or failing. The problem of portability detected in this research requires the use of calibrated models for the same group in previous subjects with similar characteristics or for the same subject in other similar groups, and this is the approach that has been adopted. Over the course of a semester, students will be provided with personalized information on their risk level, and the results of this experimental group will be compared with those of a control group that will not receive such information. Finally, we consider it necessary to complement these works with more qualitative research, which, through in-depth interviews with students, will provide complementary information that will make it possible to enrich the models.

Author Contributions: Conceptualization, J.L.A.-B. and A.H.; methodology, J.L.A.-B. and A.H.; software, J.L.A.-B. and T.C.-G.; validation, A.H.; formal analysis, J.L.A.-B., S.C.-L. and T.C.-G.; investigation, J.L.A.-B., S.C.-L. and T.C.-G.; data curation, J.L.A.-B., S.C.-L. and T.C.-G.; writing—original draft preparation, J.L.A.-B., S.C.-L., T.C.-G. and A.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Research project approved by the Pontifical University of Comillas (approval signed on 25 January 2019), complying with all the ethical and data protection standards appropriate for research of this kind.

Informed Consent Statement: Not applicable.

Data Availability Statement: Spanish data protection law forbids making public data that are not completely anonymized to prevent identification of the subjects. In the case of this study, due to its characteristics (many variables have been included in order to control the most relevant confounding factors), the data do allow identification of the subjects. For this reason, the data cannot be made public.

Acknowledgments: The authors are thankful to the Pontifical University of Comillas and the Faculty of Commerce and Tourism (Complutense University of Madrid) for their support in the development of this research.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Romero, C.; Ventura, S. Educational data mining and learning analytics: An updated survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2020**, *10*, e1355. [CrossRef]
2. Dawson, S.; Joksimovic, S.; Poquet, O.; Siemens, G. Increasing the Impact of Learning Analytics. In Proceedings of the 9th International Conference on Learning Analytics & Knowledge, Tempe, AZ, USA, 4–8 March 2019; pp. 446–455. [CrossRef]
3. Ferguson, R. Learning analytics: Drivers, developments and challenges. *Int. J. Technol. Enhanc. Learn.* **2012**, *4*, 304–317. [CrossRef]
4. Siemens, G.; Baker, R. Learning analytics and educational data mining: Towards communication and collaboration. In *ACM International Conference Proceeding Series*; Association for Computing Machinery (ACM): New York, NY, USA, 2012; pp. 252–254. [CrossRef]
5. Siemens, G.; Long, P. Penetrating the fog: Analytics in learning and education. *EDUCAUSE Rev.* **2011**, *5*, 30–32. [CrossRef]
6. Booth, M. Learning analytics: The new black. *EDUCAUSE Rev.* **2012**, *47*, 52–53.
7. Papamitsiou, Z.; Economides, A. Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Educ. Technol. Soc.* **2014**, *17*, 49–64.
8. Jayaprakash, S.; Moody, E.W.; Lauria, E.J.M.; Regan, J.R.; Baron, J.D. Early alert of academically at-risk students: An open source analytics initiative. *J. Learn. Anal.* **2014**, *1*, 6–47. [CrossRef]
9. Fenollar, P.; Cuestas, P.J.; Román, S. University students' academic performance: An integrative conceptual framework and empirical analysis. *Br. J. Educ. Psychol.* **2007**, *77*, 873–891. [CrossRef]
10. Garbanzo Vargas, G.M. Factores asociados al rendimiento académico en estudiantes universitarios, una reflexión desde la calidad de la educación superior pública. *Rev. Educ.* **2007**, *31*, 43–63. [CrossRef]
11. Kuh, G.D.; Kinzie, J.; Buckley, J.A.; Bridges, B.K.; Hayek, J.C. Commissioned report for the National Symposium on Postsecondary Student Success: Spearheading a Dialog on Student Success. In *What Matters to Student Success: A Review of the Literature*; National Postsecondary Education Cooperative: Washington, DC, USA, 2006.
12. Gasevic, D.; Dawson, S.; Rogers, T.; Gasevic, D. Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. *Internet High. Educ.* **2016**, *28*, 68–84. [CrossRef]
13. Elander, K.; Cronje, J.C. Paradigms revisited: A quantitative investigation into a model to integrate objectivism and constructivism in instructional design. *Educ. Technol. Res. Dev.* **2016**, *64*, 389–405. [CrossRef]
14. Kuhn, D. Is direct instruction an answer to the right question? *Educ. Psychol.* **2007**, *42*, 109–113. [CrossRef]
15. Schwartz, D.L.; Bransford, J.D. A time for telling. *Cognit. Instr.* **1998**, *16*, 475–522. [CrossRef]
16. Schwartz, D.L.; Martin, T. Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction. *Cognit. Instr.* **2004**, *22*, 129–184. [CrossRef]
17. Tobias, S.; Duffy, T.M. The Success or Failure of Constructivist Instruction. An Introduction. In *Constructivist Instruction: Success or Failure?* Tobias, S., Duffy, T.M., Eds.; Routledge/Taylor & Francis Group: New York, NY, USA, 2009; pp. 3–10.
18. Sweller, J.; Kirschner, P.A.; Clark, R.E. Why Minimally Guided Teaching Techniques Do Not Work: A Reply to Commentaries. *Educ. Psychcol.* **2007**, *42*, 115–121. [CrossRef]
19. Kim, J.S. The effects of a constructivist teaching approach on student academic achievement, self-concept, and learning strategies. *Asia Pac. Educ. Rev.* **2005**, *6*, 7–19. [CrossRef]
20. Krahenbuhl, K.S. Student-Centered education and constructivism: Challenges, concerns, and clarity for teachers. *Clear. House* **2016**, *89*, 97–105. [CrossRef]
21. Zain, S.F.H.S.; Rasidi, F.E.M.; Abidin, I.I.Z. Student-centred learning in mathematics–constructivism in the classroom. *J. Int. Educ. Res.* **2012**, *8*, 319–328. [CrossRef]
22. Educational Project Pontifical University of Comillas. Available online: https://www.comillas.edu/Documentos/PROYECTO_EDUCATIVO.pdf (accessed on 1 April 2021).
23. Muñoz San Roque, I.; Martínez Felipe, M. Enfoques de aprendizaje, expectativas de autoeficacia y autorregulación, ¿las metodologías de enseñanza utilizadas en el proyecto piloto del EEES en E2 afectan a la calidad del aprendizaje? In *El Espacio Europeo de Educación Superior, ¿un cambio deseable para la Universidad?: Algunas experiencias de innovación docente en la titulación de Administración y Dirección de Empresas en ICAI-ICADE COMILLAS*; Muñoz San Roque, I., Ed.; Pontifical University of Comillas: Madrid, Spain, 2012; pp. 47–104.
24. Vallejo, P.M. Implicaciones para el profesor de una enseñanza centrada en el alumno. *Misc. Comillas* **2006**, *64*, 11–38.
25. Hoy, A.; Davis, H.; Anderman, E. Theories of learning and teaching in TIP. *Theory Pract.* **2013**, *52*, 9–21. [CrossRef]
26. Garrett, L.; Huang, L.; Charleton, M.C. A framework for authenticity in the mathematics and statistics classroom. *Math. Educ.* **2016**, *25*, 32–55.
27. Wood, D.; Bruner, J.S.; Ross, G. The role of tutoring in problem solving. *J. Child Psychol. Psychiatry* **1976**, *17*, 89–100. [CrossRef] [PubMed]

28. Meyer, D.; Turner, J. Using instructional discourse analysis to study the scaffolding of student self-regulation. *Educ. Psychol.* **2002**, *37*, 17–25. [CrossRef]
29. Belland, B.R.; Kim, C.M.; Hannafin, M.J. A framework for designing scaffolds that improve motivation and cognition. *Educ. Psychol.* **2013**, *48*, 243–270. [CrossRef] [PubMed]
30. Karagiorgi, Y.; Symeou, L. Translating constructivism into instructional design: Potential and limitations. *Educ. Technol. Soc.* **2005**, *8*, 17–27.
31. Mangaroska, K.; Giannakos, M. Learning analytics for learning design: A systematic literature review of analytics-driven design to enhance learning. *IEEE Trans. Learn. Technol.* **2018**, 1–18. [CrossRef]
32. Ortigosa, A.; Carro, R.M.; Bravo-Agapito, J.; Lizcano, D.; Alcolea, J.J.; Blanco, O. From lab to production: Lessons learnt and real-life challenges of an early student-dropout prevention system. *IEEE Trans. Learn. Technol.* **2019**, *12*, 264–271. [CrossRef]
33. Yorke, M. Formative assessment in higher education: Moves towards theory and the enhancement of pedagogic practice. *High. Educ.* **2003**, *45*, 477–501. [CrossRef]
34. Nicol, D.; Macfarlane, D. Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Stud. High. Educ.* **2006**, *31*, 199–218. [CrossRef]
35. Huang, S.; Fang, N. Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. *Comput. Educ.* **2013**, *61*, 133–145. [CrossRef]
36. Gitinabard, N.; Xu, Y.; Heckman, S.; Barnes, T.; Lynch, C.F. How widely can prediction models be generalized? Performance prediction in blended courses. *IEEE Trans. Learn. Technol.* **2019**, *12*, 184–197. [CrossRef]
37. López-Zambrano, J.; Lara, J.A.; Romero, C. Towards Portability of Models for Predicting Students' Final Performance in University Courses Starting from Moodle Logs. *Appl. Sci.* **2020**, *10*, 354. [CrossRef]
38. Conijn, R.; Snijders, C.; Kleingeld, A.; Matzat, U. Predicting student performance from LMS data: A comparison of 17 blended courses using Moodle LMS. *IEEE Trans. Learn. Technol.* **2017**, *10*, 17–29. [CrossRef]
39. Olivé, D.M.; Huynh, D.Q.; Reynolds, M.; Dougiamas, M.; Wiese, D. A quest for a one-size-fits-all neural network: Early prediction of students at risk in online courses. *IEEE Trans. Learn. Technol.* **2019**, *12*, 171–183. [CrossRef]
40. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2020.
41. Hunt, X.J.; Kabul, I.K.; Silva, J. Transfer learning for education data. In Proceedings of the ACM SIGKDD Conference, El Halifax, NS, Canada, 17 August 2017.
42. Schneider, M.; Preckel, F. Variables associated with achievement in higher education: A systematic review of meta-analyses. *Psychol. Bull.* **2017**, *143*, 565–600. [CrossRef] [PubMed]
43. Martínez de Ibarreta, C.; Rua Vieites, A.; Redondo Palomo, R.; Fabra Florit, M.E.; Nuñez Partido, A.; Martín Rodrigo, M.J. Influencia del nivel educativo de los padres en el rendimiento académico de los estudiantes de la ADE. Un enfoque de género. In *Investigaciones de Economía de la Educación*; Mancebón Torrubia, M.J., Pérez Ximénez de Embún, D., Gómez Sancho, J.M., Giménez Esteban, G., Eds.; Asociación de Economía de la Educación: Las Palmas, Spain, 2010; Volume 5, pp. 1273–1296. Available online: http://repec.economicsofeducation.com/2010zaragoza/05-64.pdf (accessed on 4 February 2021).
44. Tejedor, F.J. Poder explicativo de algunos determinantes del rendimiento en los estudios universitarios. *Rev. Esp. Pedag.* **2003**, *224*, 5–32.
45. McKenzie, K.; Schweitzer, R. Who succeeds at university? Factors predicting academic performance in first year Australian university students. *High. Educ. Res. Dev.* **2001**, *20*, 21–33. [CrossRef]
46. García-Diez, M. The effects of curriculum reform on economics education in a Spanish college. *Educ. Econ.* **2000**, *8*, 5–15. [CrossRef]
47. Harbury, C.D.; Szreter, R. The influence upon university performance of the study of economics at school. *J. R. Stat. Soc. Ser. A* **1968**, *131*, 384–409. [CrossRef]
48. Didia, D.; Hasnat, B. The determinants of performance in the university introductory finance course. *Financ. Pract. Educ.* **1998**, *8*, 102–107.
49. Ballard, C.L.; Johnson, M.F. Basic math skills and performance in an introductory economics class. *J. Econ. Educ.* **2004**, *35*, 3–23. [CrossRef]
50. Girón Cruz, L.; González Gómez, D.E. Determinantes del rendimiento académico y la deserción estudiantil, en el programa de economía de la Pontificia Universidad Javeriana de Cali. *Rev. Econ. Gest. Y Desarro.* **2005**, *3*, 173–201.
51. Shieh, G. Clarifying the role of mean centring in multicollinearity of interaction effects. *Br. J. Math. Stat. Psychol.* **2011**, *64*, 462–477. [CrossRef] [PubMed]
52. Robinson, C.; Schumacker, R.E. Interaction effects: Centering, variance inflation factor, and interpretation issues. *Mult. Linear Regres. Viewp.* **2009**, *35*, 6–11.
53. Sarlija, N.; Bilandzic, A.; Stanic, M. Logistic regression modelling: Procedures and pitfalls in developing and interpreting prediction models. *Croat. Oper. Res. Rev.* **2018**, *8*, 631–652. [CrossRef]
54. Ortega Calvo, M.; Cayuela Domínguez, A. Regresión logística no condicionada y tamaño de muestra: Una revisión bibliográfica. *Rev. Esp. Salud Pública* **2002**, *76*, 85–93. [CrossRef] [PubMed]
55. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning*, 8th ed.; Springer: New York, NY, USA, 2018. [CrossRef]

56. Signorell, A.; Aho, K.; Alfons, A.; Anderegg, N.; Aragon, T.; Zeileis, A. DescTools: Tools for Descriptive Statistics. R Package Version 0.99.28. 2019. Available online: https://rdrr.io/cran/DescTools/ (accessed on 4 February 2021).
57. Robin, X.; Turck, N.; Hainard, A.; Tiberti, N.; Lisacek, F.; Müller, M. pROC: An Open-Source Package for R and S+ to Analyze and Compare ROC Curves. *BMC Bioinform.* **2011**, *12*, 77. [CrossRef] [PubMed]
58. Arroyo-Barrigüete, J.L.; Tirado, G.; Mahíllo-Fernández, I.; Ramírez, P.J. Predictors of performance in Business Administration degrees: The effect of the high-school specialty. *Rev. Educ.* **2020**, *390*, 125–148. [CrossRef]
59. Widyahastuti, F.; Tjhin, V.U. Performance Prediction in Online Discussion Forum: State-of-the-art and comparative analysis. *Procedia Comput. Sci.* **2018**, *135*, 302–314. [CrossRef]
60. Thakar, P.; Mehta, A.; Manisha. Performance Analysis and Prediction in Educational Data Mining: A Research Travelogue. *Int. J. Comput. Appl.* **2015**, *110*, 60–68.
61. Muthukrishnan, S.M.; Govindasamy, M.K.; Mustapha, M.N. Systematic mapping review on student's performance analysis using big data predictive model. *J. Fundam. Appl. Sci.* **2017**, *9*, 730–758. [CrossRef]