Predicting the Default Risk for Higher Education Students through Machine Learning

Giovanna Niskier Saadia[1], Jorge Brantes Ferreira[2], Ricardo Rodriguez Whately[3], Fernanda Leão Ramos[4], Jorge Ferreira da Silva[5]

[1] *Researcher – Pontifical Catholic University of Rio de Janeiro – Rua Marquês de São Vicente 225, Gávea, Rio de Janeiro, Brazil – giovanna.niskier@globo.com*
[2] *Associate Professor – Pontifical Catholic University of Rio de Janeiro – Rua Marquês de São Vicente 225, Gávea, Rio de Janeiro, Brazil – jorgebf@gmail.com*
[3] *Ph.D. Candidate – Pontifical Catholic University of Rio de Janeiro – Rua Marquês de São Vicente 225, Gávea, Rio de Janeiro, Brazil – rickwhately@gmail.com*
[4] *Researcher – Pontifical Catholic University of Rio de Janeiro – Rua Marquês de São Vicente 225, Gávea, Rio de Janeiro, Brazil – leaoramos@gmail.com*
[5] *Full Professor – Pontifical Catholic University of Rio de Janeiro – Rua Marquês de São Vicente 225, Gávea, Rio de Janeiro, Brazil – jorge1319@gmail.com*

Abstract

Default is a major problem for private higher education institutions (HEI) and can result in school dropout and loss of revenue. HEI mostly do not use credit scoring techniques to predict the risk of students' default. This work aims to propose and test a credit scoring model using machine learning techniques in the private higher education sector, using financial, academic, and social data from students from a private HEI in Rio de Janeiro. The proposed model evaluates the factors that most influence this kind of behavior and estimates the default risk of each student. The results showed the most relevant variables resulting in default. By using credit scoring methodology, HEI would be able to identify students at higher default risk and could plan specific preventive actions for this group.

Keywords: Machine learning; forecasting; default; higher education; credit scoring.

1. Introduction

From 2010 to 2019, Brazilian higher education recorded a 36.9% increase in private schools' enrollment, reaching 6.52 million enrollments in 2019. However, as significant as the growth curve in the number of enrollments, is the curve of default in private HEIs, with monthly payments with more than 90 days delays rising from 7.8% in 2014 to 9.5% in 2019. In the same period, the number of students with access to reimbursable funding from the federal government fell from 21.3% in 2014 to 2.2% in 2019, while private funding from HEIs rose only from 0.3% to 5.5%. In the US, student loans are the only form of debt balances that virtually sextupled from 2003 to 2018. Student loan borrowing for US higher education has emerged as a top policy concern since its debt now exceeds one trillion dollars, second only to mortgages in consumer debt.

For the HEI, a higher default rate ends up having a burden on students in the form of a tuition increase, sometimes causing more defaults because students cannot afford the increase in the amount of tuition, consolidating a vicious cycle (Lemos, Ribeiro & Siqueira, 2017). In addition, a critical consequence of default is school dropout. Students with difficulties in paying the tuition fees, eventually abandon their courses. The losses of students who start, but do not finish their courses, are social, academic, and economic wastes (Silva Filho, 2007).

Despite the relevance of the theme, the use of credit scoring models, widely disseminated by financial institutions, is less used in the private higher education sector. It would be interesting for HEI to look for a credit analysis process that could predict as accurately as possible each student's default risk, to establish preventive measures. Sobrinho (2007) presents a study of credit scoring models in basic education and Lemos, Ribeiro and Siqueira (2017) study default in HEI, but with a qualitative approach, without statistical deepening due to difficulties in accessing data. Therefore, the purpose of this study is to analyze the behavior of a group of students from a single private HEI in Rio de Janeiro and establish relationships between financial, academic, and social variables and the potential default risk. The theoretical basis to be addressed in the study is CRM, credit scoring, and default in HEI.

This work's relevance is to propose something new for the Brazilian higher education sector, which is the use of credit scoring models with machine learning techniques as a strategic management tool for the HEI so that they can differentiate their students according to their respective default risk and identify social, financial, and academic factors that are most related to this risk. In this way, the HEI will be able to carry out preventive financial and relationship actions aimed at a customer value strategy.

2. Literature Review

2.1. Factors that impact default

According to Lemos, Ribeiro and Siqueira (2017), the main external factors that impact defaults in HEI are the economic situation of the country, the current educational legislation, the poor granting of credit by the HEI, and the lack of adaptation of the HEI to the new reality of the market, which requires an increasingly rapid response, especially concerning the collection of credits to be received.

Regarding the poor granting of credit, the HEI, for the most part, do not perform any kind of prior credit analysis of the student to predict their potential risk of default and thus do not perform preventive actions. Lemos, Ribeiro and Siqueira (2017) propose that default in HEI should be analyzed from a broader perspective, extrapolating the external and financial factors as determinants of the paying capacity of the students. The authors seek to explain default through internal factors of the students themselves, such as analysis of their profiles, their socioeconomic characteristics, and academic situation. The authors present seven factors as being influential in the paying capacity of students: academic origin (public or private school), academic performance, attendance, gender, place of residence (associated with income), marital status (single versus married with children) and professional occupation (e.g., self-employed, unemployed).

2.2. Credit scoring

Credit scoring is an important analytical technique in credit risk evaluation based on customer history and environmental factors (Bhatore et al., 2020). Among the objective techniques of credit risk management, credit scoring stands out. According to Thomas (2000), credit scoring models are systems that attribute scores to a proponent's credit decision variables by applying statistical techniques. These models aim at segregating characteristics that allow the good ones to be distinguished from bad payers. The score can be interpreted as the probability of default when compared to the score established as a cutoff point or minimum acceptable score, which will serve as the basis for the approval or refusal of credit.

Credit scoring models can be divided into two categories, as to their objectives: credit approval models and behavioral anchorage models, also known as behavioral scoring. Credit approval models use data on the personal and professional life of the credit applicant to predict the future behavior of current and new customers about the payment of debts under the credit agreement (Rosenberg & Gleit, 1994). On the other hand, scoring behavioral models also aim at predicting default, but their focus is on the analysis of individuals who are already clients of the institution and have a credit relationship, enabling the incorporation of the payment history of these clients into the set of predictor information of the model (Sobrinho, 2007).

2.3. Data Mining and Machine Learning

Data mining is a term often used to define the process of extracting useful information from a vast customer information database. The main challenge for companies that work with a large amount of their customer's data is to take advantage of this opportunity given by the availability of data and turn it into useful knowledge for the company. In general, data mining can be direct (classification, estimation, and prediction) or indirect (affinity grouping, clustering, description, and visualization), and can be used, for example, algorithms such as logistic regression, decision trees, and neural networks.

The term machine learning deals with machine learning techniques and models that use massive amounts of data, seeking to continuously learn from new data to improve the prediction about a given variable. Prediction models can be implemented using Supervised Learning algorithms, which learn a function that maps a set of independent variables (features) to a dependent variable (target) (Olivé et al., 2020). Modeling with machine learning techniques, the researcher will include variables that are suspected of impacting the target variable. The goal is, considering these attributes, to evaluate what output the model returns for each instance of the database. The main advantage of computer-aided credit risk evaluation is that human work is minimized since it learns from a pre-collected database to make accurate and reliable predictions (Bhatore et al., 2020). During data mining modeling, diverse attributes with unknown relationships are evaluated in search of hidden relationships, which were previously unknown. Using AI, hidden patterns are recognized, and appropriate alerts are raised in a useful and timely manner (Bhatore et al., 2020).

Therefore, considering the objective of this work, algorithms will be used to collect the relevant information of the students of the evaluated HEI and identify those that influence their paying capacity. In the scenario of Brazilian higher education, machine learning techniques find a vast field of application, still little explored (Lemos, Ribeiro & Siqueira, 2017). Default prediction, using machine learning models specifically, has been little studied in this sector, being, therefore, a promising field of research.

3. Methodology

This study addresses default in private higher education as a problem of financial management and relationship marketing. A quantitative approach was adopted, through the application of machine learning models to classify a real database to find models that explain student default in the evaluated HEI. The HEI is located in Rio de Janeiro and has more than 13,000 students on 2 campuses. The profile of the students of the HEI is low income, with a default rate of 20%, with debts of one or more monthly payments. The database used has information regarding a representative sample of 4978 students of the institution who were active in the years 2017 and 2018.

Students who had debts of one to six monthly fees were considered in default. It is also worth mentioning that, given the nature of the service provided and the impossibility of limiting access to educational services when the client/student becomes a debtor, the default condition will be analyzed in this study at the end of each semester.

Classification processes will only indicate the probability of students defaulting at some point, not when they will default. Regarding the proportions between the segments in the database (defaulter/non-defaulter), we have 38% of the students as non-defaulters and 62% as defaulters; thus, it was not necessary to perform oversampling procedures to adjust the proportions between segments.

3.1. Data collection and preparation

With the direct access given by the HEI to the institution students’ data, the following variables were collected, through interaction with the institution’s information technology sector, for the construction of the 4978 students’ database used in this work:

- Sociodemographic: Registration number (identifier), gender, marital status, age, has or not children, people in residence, financial guardian or not, type of paid activity, contribution to family income, where and how he attended high school.
- *Academic*: Amounts (per semester) of subjects enrolled, absences, total disapprovals, disapprovals for misconduct, disapprovals by note, approvals, and locks.
- *Financial*: half-yearly paid amount and outstanding debits amount.

3.2. Data analysis

This study used the free access software WEKA to create the predictive models. This software has several of the algorithms most used in machine learning processes and allows the creation, testing and comparison of models (Witten; Frank & Hall, 2011). Three machine learning algorithms were used for classification: logistic regression (Wilson & Lorentz, 2015), decision trees, and neural networks (Haykin, 2007). We used the supervised training approach since the data were already classified (0 – non-defaulter and 1 - defaulter). The out-of-sample model accuracy and their generalization capabilities were tested using the cross-validation method known as k-fold cross-validation (Witten; Frank & Hall, 2011), with k = 10.

4. Results and Discussion

The risk of default in the evaluated HEI was estimated through three different machine learning models, with the respective coding in the WEKA software: Logistic Regression, Decision D Tree (J48 Tree), and Neural Networks (Multilaver Perceptron). All models used in this study, on average, presented high accuracy rates, evidencing the high capacity to predict the proposed problem. Table 1 summarizes the predictive performance of the evaluated credit scoring models, while Table 2 summarizes the main financial, academic, and sociodemographic variables related to a student's default risk, obtained in the generated model results.

Table 1 - Accuracy, sensitivity, specificity, and accuracy of the estimated models

	Regression Logistics	Decision Tree J48	Neural Networks	Average
Accuracy	81,24%	77,21%	77,04%	78,49%
Sensitivity	80,40%	80,83%	80,67%	80,63%
Specificity	82,57%	71,45%	71,24%	75,08%
Precision	88,05%	81,90%	81,76%	83,90%

Table 2 - Main model results summary

	Financial	Academic	Demographic
Default	Debts in previous semesters 2018.1	Number of subjects attended	Parent, mother, spouse, or another person responsible for family income
			Family income of up to 2 minimum wages
	Debts in previous semesters 2017.1	Variety of courses attended	Students with "divorced" status
			Students with self-employed, unemployed, or state company worker

This work aimed to propose and test a credit scoring model using machine learning techniques in the private higher education sector, using financial, academic, and social data from students from a private HEI in Rio de Janeiro. Among the proposed models, logistic regression presented the best accuracy and accuracy index, indicating the greater capacity of this model to predict students at default risk for the analyzed data. On the other hand, the decision tree model can show the chain of relevant variables presenting relevant information for the understanding of the default process. Possibly, the combination of more than one model can contribute to a better understanding of the phenomenon of students' defaults.

Given the existence of laws that protect students who become defaulters, and limit the HEI effort to receive the debts, the credit scoring models proposed here allow institutions to better manage the educational credit risks. By applying these models, HEI can identify risk factors and students at higher risk of default to seek to mitigate these risks, in addition, to predicting students who can recover their credit from those who will not have any conditions of payment of debts. At a more advanced stage, HEI can seek to select only students who have long-term ability to pay and improve their student finance model.

The relevance of this study lies in the validation of machine learning techniques as a tool to predict the default risk of students from higher education institutions and can be adapted to other segments. Few studies in Brazil have associated the use of machine learning techniques with predicting the risk of student default, by understanding the most relevant factors that influence this behavior and allowing the performance of different preventive actions according to each student's risk. Existing studies don't focus on the private higher education sector, nor perform a quantitative analysis on default, only exploring statistical data superficially.

Finally, regarding the study's limitations, it is important to note that the available database only included students who were active in the four semesters of 2017 and 2018, from a single HEI. Particularly relevant would be to study the default of students after the COVID-19 pandemic, reprocessing the proposed models with more recent data. In addition, this study only considers the financial, academic, and demographic attributes of students of a private HEI, disregarding external factors, such as economic changes and unemployment level, when calculating the probability of default. Future studies could consider external variables that systematically impact the model.

References

Bhatore, S., Mohan, L., & Rheddy, Y. (2020). Machine learning techniques for credit risk evaluation: a systematic literature review. *Journal of Banking and Financial Technology*, 4, 111-138.

Haykin, S. (2007). *Neural networks: principles and practice*. Porto Alegre: Bookman.

Lemos, A., Serra, F., & De Siqueira, E. (2017). Access to higher education and the problem of default: a study on the determining factors of paying capacity of students in a private institution. *International Journal of Professional Business Review*: 2(2), 23-35.

Olivé, D., Huynh, D., Reynolds, M., Dougiams, M., & Wiese, D. (2020). A supervised learning framework: using assessment to identify students at risk of dropping out of a MOOC. *Journal of Computing in Higher Education*, 32(1), 9-26.

Rosenberg, E., & Gleit, A. (1994). Quantitative methods in credit management: a survey. *Operations Research*, 42(4), 589-613.

Silva Filho, R., Montejunas, P. R., Hipólito, O., Lobo, M. B. (2007). The evasion of Brazilian higher education. *Cadernos de Pesquisa*, 37(132), 641-659.

Sobrinho, M. (2007). Um estudo da inadimplência aplicada ao segmento educacional de ensino médio e fundamental, utilizando modelos credit scoring com análise discriminante, regressão logística e redes neurais, In: *Masters Dissertation in Business Administration – Universidade Federal de Pernambuco*, Recife.

Thomas, L. C. (2000). A survey of credit and behavioral scoring: forecasting financial risk of lending to consumers. *International journal of forecasting*, 16(2), 149-172.

Wilson, J. R., & Lorenz, K. A. (2015). Short History of the Logistic Regression Model. In: *Modeling Binary Correlated Responses using SAS, SPSS and R*, Springer International Publishing.

Witten, I., Frank, E., & Hall, M. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. 3 ed., Burlington: Morgan Kaufmann.