Corvinus University of Budapest

Doctoral School of Economics, Business and Informatics

The Business Informatics Doctoral Program

THESIS SUMMARY OF

Doctoral dissertation

Modeling Student's Churn in Higher Education in Bosnia and Herzegovina

Supervisor: Peter Racsko PhD

Candidate: Dragana Preradović Kulovac

August 2024

1 Research background and justification of the topic

Bosnia and Herzegovina, with its strategic geographical location and substantial economic potential and relative small population of less than 3.2 million, possesses all the necessary prerequisites for ensuring the well-being of its citizens. However, the country is the first in the world by brain drain, while the share of highly educated people in working-age population is less than 10 percent, which is almost 4 times lover than average in EU countries. Country is listed as 121st in the world by World Bank in 2022, with 7.2 thousands of USD of GDP per capita, lower than all other bordered countries.

The one of possible ways of increasing the Bosnia and Herzegovina's macroeconomics performances is to have higher number of tertiary educated people. It is confirmed that number of highly educated people have positive influence to the country's GDP, tax revenue, the global competiveness, longevity of citizens, and decrease of crime rates. The studies have showed the numerous positive effects of having highly educated people on society, economy, higher education institutions, and individuals.

Still, since 2010, the number of enrolled students in higher education in Bosnia and Herzegovina decreased by more than 40 percent. The country has 8 public universities, and 13 Ministries of Education and there is no statistical reports, research or estimation of dropout in higher education, or reasons for leaving the higher education.

By turning a blind eye to this problem, the country risks worsening it's already poor performance. One of the ways to achieving the goal of having more highly educated people is by predicting the churn in early stage of education, even before students decide to dropout and preventing its decision to leave the university.

This research seeks to develop an effective machine learning model using a challenging dataset for predicting and preventing student churn in the early stages of higher education and identifying the underlying reasons for churn. To achieve this, the following research questions were formulated:

1) What is the magnitude, structure, and reasons for student dropout at the University of Banja Luka, Bosnia and Herzegovina, between the 2007/08 and 2018/19 academic years?
2) How well do machine learning models perform when trained on a dataset that is predominantly binary, contains missing data, and lacks socio-economic, academic, and secondary education features?
3) How can the explainability and interpretability of a black-box model be effectively enhanced to overcome its inherent limitations and provide a clearer understanding of its outputs?

2 Methodology

This research integrates both qualitative and quantitative methods, utilizing the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology, a well-established approach in the field of Educational Data Mining. This comprehensive, document-intensive methodology was chosen to ensure a systematic, data-driven investigation grounded in robust analysis.

The qualitative component aimed to define churn and understand the reasons behind student attrition. This included interviews with 10 students who had taken a break from their studies and a survey of 96 students who voluntarily discontinued their education, marking the first such study in the country.

The quantitative analysis produced the first comprehensive estimation of higher education churn in the country, based on a sample of 37 thousands students. Additionally, a subset of 20 thousands students, tracked for at least six years from enrollment, was used to predict churn. This analysis employed models such as Histogram-based Gradient Boosting Classifier, Random Forest, Support Vector Machine, and Neural Network.

The data for this research were obtained from the University Computing Centre of the University of Banja Luka, covering academic years from 2007/08 to 2018/19. The dataset had considerable missing data and lacked socio-economic, pre-academic, and academic performance indicators.

Significant effort was devoted to the first two phases of the methodology—problem and data understanding. This included a thorough examination of relevant laws, university regulations, and practices, as well as an in-depth understanding of the dataset.

Data analysis was conducted using Python, with results presented in Python, Excel, and Tableau. Following the selected methodology, detailed reports were generated on the initial data collection, variable descriptions, data exploration by student features, data preprocessing, and missing data. Variable reduction was performed through correlation analysis before modeling, and further refined using Permutation Importance and SHapley Additive exPlanations (SHAP) after modeling.

To ensure accurate modeling at the earliest stage of higher education, each model was trained and tested using a sequentially expanding set of variables: first with the pre-enrollment set of 27 variables, followed by the enrollment set, which included an additional 10 variables, and finally, with the end-of-first-year set, which added 8 more variables, bringing the total to 45 variables.

Furthermore, the best-performing model was rigorously tested by incorporating 3 additional variables with a significant proportion of missing data. The model was also tested under conditions of altered churn definitions, imbalanced data scenarios, and the inclusion of faculty names as variables. These steps were taken to ensure the model's robustness, performance, and validity on the current dataset.

3 The findings of the dissertation

A comprehensive estimation of dropout rates, covering approximately 20 percent of the student population in Bosnia and Herzegovina between the 2007/08 and 2018/19 academic years at the University of Banja Luka (UNIBL), was conducted to examine the magnitude, structure, and reasons for attrition:

- Nearly half of the enrolled students discontinued their studies within the 12-year observation period, with attrition comprising both temporary and permanent churn. The research primarily focuses on permanent churn, which accounts for approximately one-third of the overall enrollment.
- Women exhibited lower dropout rates than men across all study fields and academic years, including those traditionally dominated by men. Notably, half of the dropout cases occurred during the first year of study, and dropout rates were significantly higher in programs of 3-year duration

compared to those of 4-year duration. Thus, efforts to prevent attrition should be concentrated in these areas.

- Survival rate analysis indicates that more recently enrolled cohorts have higher dropout rates, while older generations demonstrated greater persistence.
- The research revealed that the distance from the University and the level of development of a student's place of origin, i.e., their municipality, significantly impact attrition. As the distance from the University increases, the dropout rate decreases, whereas an increase in the municipality's development level corresponds with a rise in dropout rates.
- The findings suggest that one-third of dropouts could potentially be prevented, as they are attributed to institutional factors.
- The leading causes of attrition include disputes or conflicts with lecturers or professors, followed by financial difficulties and challenges in balancing work and study

commitments. In addition to institutional reasons, personal and pedagogical factors were also identified.

- An examination of the trajectories of students who discontinued their studies at their request shows that, in the years following their departure from UNIBL, one-third continued their higher education abroad, another third enrolled in private institutions within the country, and the remaining third returned to UNIBL.

After training the Histogram-based Gradient Boosting Classifier (HGBC), Random Forest, Support Vector Machine (using 7 different kernels), and 2-, 3-, and 4-layer Neural Networks on a challenging dataset—including pre-enrollment data, data at the beginning, and data at the end of the first academic year—the results suggest the following:

- The HGBC was introduced for the first time in Educational Data Mining and proved to be superior in terms of relative performance compared to other widely used models in the field. This was particularly evident on a longitudinal dataset with missing data and limited socio-economic, pre-academic, and academic features. Other models experienced overfitting (Random Forest) or exhibited low performance and quality (some SVM kernels and Neural Networks).

- The HGBC outperformed all other models across all three stages of prediction, demonstrating higher accuracy and faster processing. The overall accuracy reached 82.5% by the end of the academic year using the top 13 variables, while the accuracy at the beginning of the academic year was 75.2%.
- To assess the robustness of the HGBC, the model was tested under varying conditions, including the addition and removal of variables, changes in the churn definition, and the presence of imbalanced data. Under these conditions, the HGBC consistently maintained very high performance.

Permutation Importance (PI) and SHapley Additive exPlanations (SHAP) were employed as tools for visual interpretation and model explainability to understand how our black-box models differentiate between students who drop out and those who persist. These methods also helped analyze the behavior of features over time and across different models:

- The results indicate that some features remain strong predictors throughout the academic year, while others decline in importance over time. A few features, however, show an increase in importance as time progresses.
- One of the strongest predictors of student churn, which maintains its significance over time, is gender, specifically being female, followed by the academic year in which the student was enrolled (i.e., the student's cohort).
- Students who switch faculties, those from highly developed municipalities, and those living within 80 km of the university are at

higher risk of dropping out, as are students who are not scholarship holders. Other important predictors of churn include the duration of the study program, the number of ECTS credits accumulated, age, the number of courses passed, and whether the student attended a gymnasium.

- The use of correlation as a pre-hoc method to establish expectations, combined with the application of PI and SHAP as post-hoc techniques to interpret and validate the decisions of black-box models, enhances the credibility and reliability of the model outcomes.
- Interestingly, models with lower performance exhibited similar variable importance rankings to those with higher performance.

4 The observations relating to the utilization of the dissertation

The findings of this study highlight the urgent need for addressing higher education dropout rates and offer several actionable recommendations:

- Establish a comprehensive dropout prevention policy.
- Introduce annual churn reports or statistics.
- Implement an Early Warning System to identify at-risk students early.
- Improve database management at the University of Banja Luka to accurately address dropout issues.
- Follow other specific recommendations provided in this thesis to reduce churn.

Our analysis shows that the dropout rate at UNIBL is 47.1%, which is significantly higher than Denmark's 30% (using the same methodology) and exceeds the rates in many European countries.

Female students demonstrate greater persistence, and dropout rates are correlated with the university's location and the development level of the students' municipalities, aligning with existing research.

Differences between dropouts and non-dropouts at UNIBL are consistent with the literature regarding age and academic performance but are more pronounced in relation to the study field. While institutional, personal, and external factors such as financial constraints align with known reasons for dropout, conflicts with lecturers—a significant factor at UNIBL—are notably unique. Institutional reasons account for one-third of dropouts at UNIBL, compared to 13% in Europe.

The Histogram-based Gradient Boosting Classifier, applied for the first time in Educational Data Mining, outperformed traditional models like Decision Trees, Random Forests, Support Vector Machines, and Neural Networks on a challenging dataset. In modeling student dropout, HGBC achieved a recall rate of 76%, significantly surpassing the 37% recall rate reported in previous studies using richer datasets.

The strongest predictors of higher education dropout identified in this study are largely consistent with existing literature. However, our study uniquely highlighted the importance of student cohort, study program duration, and an artificial ID variable. Unlike other studies that emphasize socioeconomic factors, ethnicity, and high school GPA, our analysis was constrained by missing data, and lack of those variables.

The recommendation is that future research in this field builds upon those findings by introducing a prediction model at the end of the first semester, incorporating in-semester data to prevent attrition in the second semester. Prevention efforts should focus on the first year, particularly at the faculty level, including the use of exam records. To make the model's results more actionable for students, additional visualization techniques such as Partial Dependence Plots, Anchors, Counterfactuals, and Deletion Diagnostics should be incorporated. Furthermore, a nationwide survey using a systematic approach could help identify factors that can be addressed in both the short and long term.

An interesting finding was that recent student cohorts are more likely to discontinue their studies. We recommend that future researchers explore this trend using unsupervised learning models to identify significant patterns within student cohorts over time. Understanding the causes of attrition and how they evolve from one academic year to another, can lead to more effective prevention strategies.

5 Author's publications in the given topic

The conferences' Book of Abstracts, includes attendance and presented papers that are not published but related to the topic:

D. Preradović Kulovac, (2022): *Analysis of STEM Student Dropout at the University of Banja Luka*. Book of Abstracts, International Conference on Applied Sciences (ICAS 2022), 25-28 May, Banja Luka, 2022, p. 38.

D. Preradović Kulovac, Lj. Mićić, (2022): *Modeling students drop out in social sciences using machine learning.* 9th Researching Entrepreneurship and Economic Development (REDETE 2022), Marche Polytechnic University in cooperation with the University of Banja Luka and the AII Permanent Secretariat, September 15-16, 2022, Ancona, Italy – presented.

D. Preradović Kulovac, Lj. Mićić, (2022): *Challenge of real-life dataset: modeling the university student dropout.* 19th International Conference on Operational Research (KOI 2022), Šibenik, Croatia, September 28-30, 2022, Book of Abstracts, Croatian Operational Research Society in collaboration with Faculty of Economics, Business and Tourism, University of Split, ISSN 1849-5154, UDC 519.8, p. 78

D. Preradović Kulovac, Lj. Mićić, (2022): *The Approach to Reducing the Economic Consequences of Higher Education Attrition in Bosnia And Herzegovina*, 10th International Conference of the School of Economics and Business, ICES 2022, October 14, 2022, Book of Abstracts and Conference Proceedings, ISSN: 2490-3620, p. 329.