

Applied Data Science Master Thesis

**Long-term Prediction of Master's Student Influx in
Computing Science**

TingSyuan Hsu, Sarah

8264937

First Supervisor: Prof. Dr. Gerard Barkema

Second Supervisor: Dr. Albert Gatt



Department of Information and Computing Science

Utrecht University

Netherlands

July, 2022

Abstract

The influx of students is not only an important source of academic deployment but also influences national development. Higher education in the Netherlands has expanded rapidly in the last two decades, and the Dutch labour market also needs these talented people in the fastest-growing ICT industry. It is essential for the university to forecast the incoming students and avoid adverse effects in response to the rising student population. In this thesis, the XGBoost Regression algorithm is proposed to predict the Master's student influx in the following five years and investigate the feature importance of the student population. The model is trained by the background information of Bachelor's students to estimate the long-term growth of the Master's student influx in Computing Science. The findings and analysis can provide objective references to initiate the educational strategy.

1 Introduction

Over the past twenty years, the number of students at Dutch universities has almost doubled from 178,000 students in 2002 to 340,000 students in 2022 [1]. The growth is expected to continue in the coming years which can be attributed to two aspects. One notable reason derives from the rising number of Bachelor's students who are progressing to a Master's programme in a university after they graduated. Besides, the popularity of international students from the European Economic Area (EEA) and outside the European regions around the world are increasing and their share in the growth are rising over the years [2]. The awareness of higher education and English-language education in the Netherlands accelerates the incoming student mobility and total population.

On the other hand, high-educated people and international talent are essential, both for the success of research at universities and for the Dutch labour market. The Research Centre for Education and the Labour Market (ROA) shows that the demand for university graduates will remain high in the Dutch labor market from 2021 to 2026, especially in the ICT sector which is already struggling with significant shortages in the Netherlands [3][4]. The incoming students in science education not only play an important role in academic development but also contribute to Dutch technology advancement and economic growth.

However, the ever-rising number of incoming students gives rise to concerns about the quality of education. The universities are under huge pressure to keep up with the growing student population in terms of limited facilities, accommodation, employee workload, and government funding [5]. Therefore, it has been a pressing issue for Dutch universities to forecast student influx in order to strike a balance between the demand and resources in education. In light of this, Machine Learning techniques have brought remarkable benefits to several fields, but educational implications have not been widely proposed. With regards to the bottleneck in maintaining high-quality education and capability, it will be valuable work to help long-term academic planning and education deployment.

This research aims to predict the number of students enrolled in the Master's programme of Computing Science at Utrecht University in the coming five years as well as find the dominant features which have a significant influence on the student influx. The machine learning process is divided into two parts as **Figure 1**. In the first phase, it applies the slide window method to stimulate the pre-prepared data within the scope of feature engineering for the upcoming five years. In the second phase, the XGBoost regression model is designed to predict the Master's student influx as a predictive model. The model is trained by the selected features in the historical data and Gain mechanism will rank the feature importance towards the total population. After the data was prepared and the model was established, the simulated

data will be tested with the machine learning algorithm to predict the student influx in the near future. As the result, the facts and figures can provide insight into the high-impact features and forecast the long-term student population.

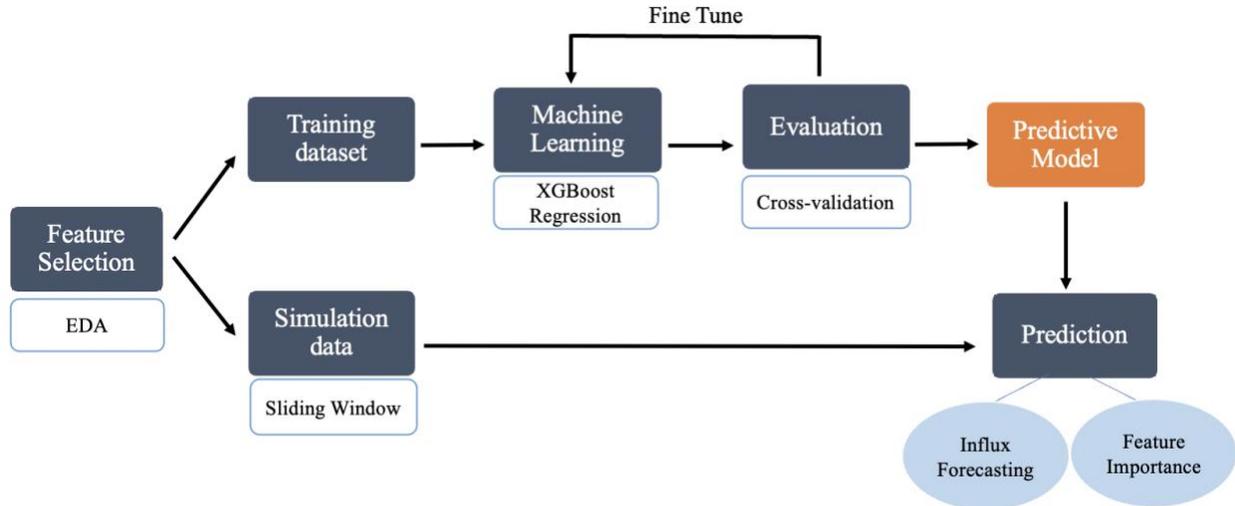


Figure 1. The workflow of Machine Learning in this research.

Literature Reviews

A classic forecasting technique to handle time-based problems is time-series methods, which identify the underlying patterns in a period of time of the past data to predict the future points. ARIMA, an acronym for Autoregressive Integrated Moving Average, is a mathematical model to perform time series forecasting based on statistics and econometrics [6]. The approach has been widely used for various forecasting tasks of market demand, productivity, and energy consumption [7][8]. However, there are some theoretical limitations of ARIMA due to the framework of seasonality and stationarity. Firstly, the ARIMA model is mainly applied to estimate univariable time-series data to generalize the forecast in the assumption of the same mean, variance, and lagged pattern over time. It is challenging for ARIMA to model the correlations across different features in multi-variable problems. Thus, ARIMA is not good at long-term forecasts and handling turning points due to complex regressor settings. Besides, the ARIMA model suffers from fitting short series with fewer than fifty observations [9]. As the reduced form of a vector process, ARIMA fails to inspect the terms in the series of data points when there is not enough data to be withheld for testing purposes [10]. To overcome these problems, a structural regression model trained by more than one variable can bring more information to the model, which has shown better results than conventional time-series methodology in several empirical research [11][12][13].

Machine Learning approaches have tremendous success in prediction, especially the deep learning models that make outstanding achievements in extracting high-level features by incremental learning. The Long Short Term Memory (LSTM), an extension of Recurrent Neural Networks (RNNs), is another popular approach to presenting time-series forecasting [14]. This deep learning network can remember the previous information and uses it for processing the present task. Additionally, the memory units of LSTM are designed to address the long-term dependency problem of gradient vanishing in regular RNNs [15]. The algorithm of LSTM can discover the underlying patterns in the sequential data and have competitive results across various domains [8][16][17]. However, one of the most challenging aspects of deep learning is the reliance on abundant data since the model requires a huge amount of training dataset to

learn the latent patterns behind the data. Moreover, artificial neuron networks are criticized as a “black box” because it lacks transparency and is not interpretable to humans [18]. In this case, the LSTM networks cannot represent the connection between the inputs and outputs whereas it is essential in decision-making.

By contrast, Gradient boosting decision tree (GBDT) techniques, like XGBoost which stands for Extreme Gradient Boosting, combined with the boosting and bagging mechanism is another form of deep learning structure because it optimizes the prediction tree based on fitting the residuals of the previous tree [19]. The strengths of parallelized implementation are recommended to handle problems with small data sizes and avoid overfitting [20]. Besides, the tree-based algorithm can intuitively estimate the feature importance by accumulating the information Gain. Therefore, when it comes to small-to-medium structured data in classification or regression problems, XGBoost algorithms usually outperform artificial neural networks with minimal effort to feature engineering [21][22].

2 Data and Methods

The data consolidation merged two tabular data. One is the individual student data from 2010 to 2019 in all Dutch universities which were collected by the Association of Cooperating Universities in the Netherlands (VSNU), Vereniging van Universiteiten in Dutch. Another one is the key reference table including the unique code of each university, program, and nationality. The metadata provides the background information of graduated students and first-year enrollment students from each program at different universities that are indexed in cohort year order.

2.1 Feature Selection

Feature selection is the process of extracting meaningful features when developing a predictive model. It helps to reduce the computational cost of modeling and improve the performance of the model. In this part, the Exploratory Data Analysis (EDA) is implemented to investigate the data and discover patterns to make reasonable assumptions and hypotheses about how certain factors would affect the target variable in the research question.

According to our primary research and the target output of the Master’s student population in Computer Science at Utrecht University, it is critical to investigate the student composition from two aspects: the student's previous education and nationality. Firstly, **Figure 2** demonstrates that over half of the new enrollment students graduated from Utrecht University and a majority of these bachelor's graduates are from the science faculty in Math (BA_Math), Informatica (BA_Informatica), Artificial Intelligence (BA_AI), and Physics and Astronomy (BA_Physics_Astronomy). It indicates that the number of bachelor’s graduates from the four programmes and the proportion of students who decide to have a further Master's study at the same university (BAtoMA) can be the feasible factors for the student influx in Master’s of Computing Science.

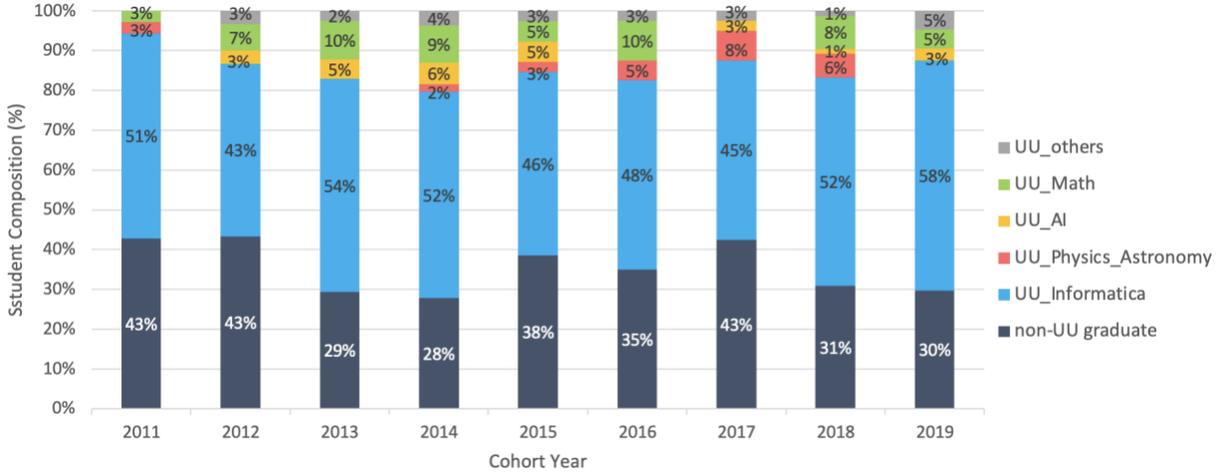


Figure 2. The previous education of Master’s students in Computing Science at Utrecht University from 2011 to 2019. (Source: UU_ GSNS _Ingeschreven)

Secondly, by analyzing the student mobility, all of the student's backgrounds can be summarized into three features based on their nationalities: students from the Netherlands (NL), European Economic Area (EEA), and outside the EEA (others). In **Figure 3**, the trends among these three groups are presented relatively along with the total number of enrollment students. Given the terms, it assumes the student's mobility can be the predictor of the student influx. Overall, there are eight notable attributes, five based on the students' previous education and three based on their nationalities, which are considered as input data to predict the target outcome.

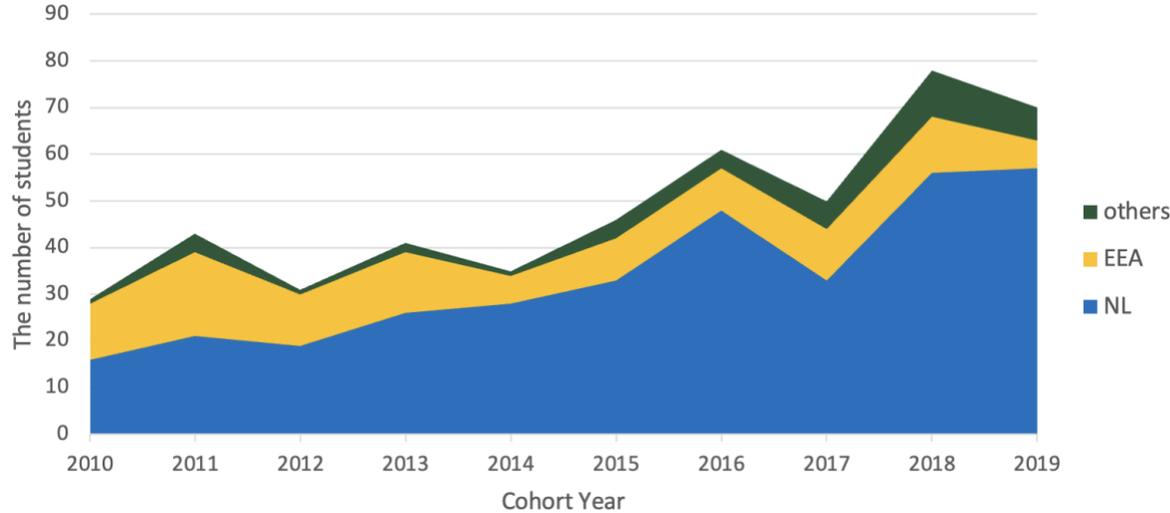


Figure 3. The nationality of Master’s students in Computing Science at Utrecht University from 2010 to 2019. (Source: 1CijferHO)

However, the one-decade data from Utrecht University is somehow insufficient to train a model and suffers from poor accuracy. To optimize the performance, this study collects the training data from eight Dutch universities – Radboud Universiteit Nijmegen, Rijksuniversiteit Groningen, Technische Universiteit Eindhoven, Universiteit van Amsterdam, Universiteit Twente, Technische Universiteit Delft, Universiteit Leiden, Universiteit Utrecht to train and validate the model. Regarding these programs being

developed at each university in different years, the missing values in cohort years are removed and we only extract the data in recent years 2015-2019 to retain the same time frame among all universities and focus on the up-to-date data. Eventually, there are five yearly data for each university. In total, forty annual data from eight universities includes the eight features as input variables and the number of students in Computing Science as targeted prediction (MA_CS) as described in **Table 1**.

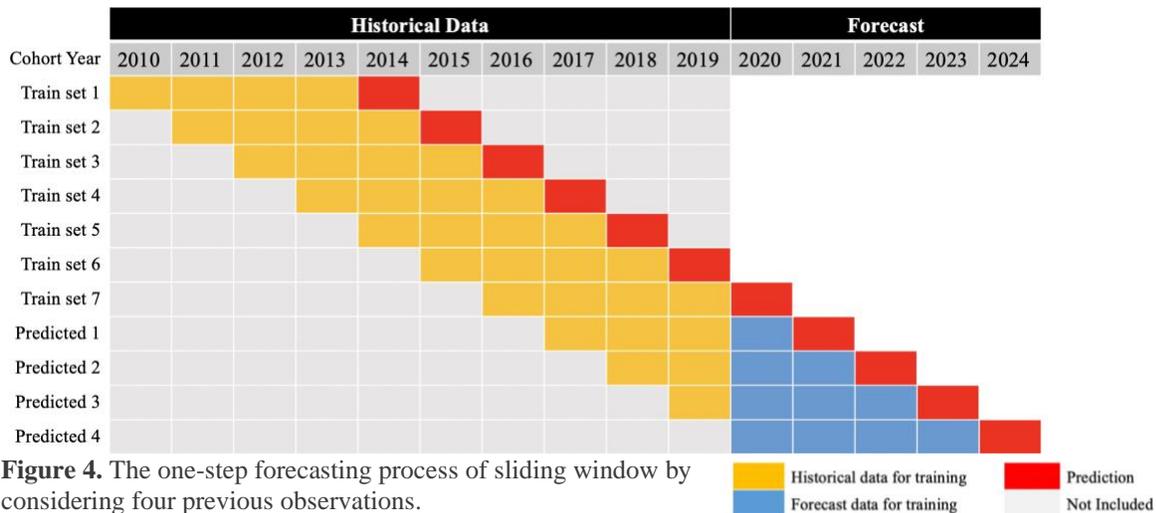
	Variables	Description
Features	BA_Math	The number of Bachelor's graduates in Math
	BA_Informatica	The number of Bachelor's graduates in Informatica
	BA_Physics_Astronomy	The number of Bachelor's graduates in Physics and Astronomy
	BA_AI	The number of Bachelor's graduates in Artificial Intelligence
	BAtoma	The number of Bachelor's graduates continue to study a Master's degree at the same University in the same year
	BA_NL	The number of Bachelor's graduates from the Netherlands
	BA_EEA	The number of Bachelor's graduates from EEA
Output	BA_others	The number of Bachelor's graduates from outside the EEA
	CS_MA	The number of Master's enrollments in Computing Science

Table 1. The data description.

2.2 Data Simulation

Sliding Window

In order to predict the number of Master's students in Computing Science in the coming five years, the simulated data of each feature from 2020 to 2024 is required. However, the observed student data is only from 2010 to 2019, the amount of data is not sufficient to divide into the training set and testing set in the typical machine learning methods. To overcome this technical difficulty, this simulation applies the sliding window to complete the modeling [23][24], which samples a specified segment of data as a window then moves forward step by step along with the timestamps and accumulates the historical time series data as input to predict next timestamp. A window size of n indicates the segment contains $n - 1$ lagged values as inputs at t and one prediction for $t+1$ as shown in **Figure 4**.



The regressor is trained on a collection of training samples and the accuracy is evaluated by Mean Absolute Percentage Error (MAPE) by comparing the predicted results with actual values. To find the optimal size of a window with the smallest approximation error, the data are fitted with the time ranges from three to five years of training data. The lagged range in years with the lowest MAPE is selected as the best window size for each feature (Table 2). For example, the feature of “BA_others” has the lowest MAPE in lagged value of 4, which means the ten-year annual data can be split into seven training sets and have the best prediction by considering the previous four-year observations. Eventually, the sliding window works among these eight features based on its best time range of input data to make corresponding outputs till 2024. The simulation experiment can estimate the trend of each feature and prepare the data for forecasting the Master’s student influx in the near future.

	BA_Math	BA_ Informatica	BA_Physics_ Astronomy	BA_AI	BA_NL	BA_EEA	BA_others	BAtoMA
Lag_3 years	0.05484	0.12335	0.13238	0.54028	0.13767	0.39068	0.45290	0.11389
Lag_4 years	0.08159	0.09374	0.15843	0.52896	0.14643	0.47697	0.23228	0.13961
Lag_5 years	0.07842	0.09106	0.11526	0.47250	0.09410	0.49376	0.23791	0.09668

Table 2. The average MAPE of each feature of the sliding window with lagged year ranges from 3 to 5.

2.3 Predictive Model

XGBoost Regression Model

XGBoost is a novel decision-tree-based ensemble Machine Learning algorithm using a gradient boosting framework. The gradient descent algorithm optimizes the performance by sequentially minimizing the errors from previous models while boosting the influence of high-performing models. In each iteration, it fits a new base learner designed to correct the loss of the previous sequence of learners. As the result, the final ensemble sequence can achieve prominent performance on the training set and computational efficiency. Moreover, the mechanism comes with algorithmic enhancements of regularization and built-in cross-validation method at each iteration which yields robust performance and prevents overfitting.

Given the continuous variables of output in the regression problem, the XGBoost Regression Model calculates the Similarity Score of the leaf to determine the optimal split node as Eq. (1). The algorithm is designed to compute the residual between the observed and predicted values and understand how important the feature is.

$$S = \frac{\sum_{k=1}^{n_\varepsilon} \varepsilon^2}{n_\varepsilon + \lambda}, \quad (1)$$

where S is the Similarity Score, ε denotes the residual of each datapoint in the split point, n_ε denotes the number of residuals on the split point, and λ is a regularization parameter that reduces the prediction’s sensitivity to individual observations which helps to handle the overfitting problems. The larger the similarity score, the more similar the leaves. Based on the average score, observations can be split into two nodes.

Based on the Similarity Score, Eq. (2) calculates the Gain by splitting the residuals into two groups and evaluating each possible split loss reduction. The best splitting position refers to the largest Gain value. It can be an indicator to quantify how great the leaves classify similar residuals compared to the root.

$$G = \frac{G_L^2}{n_L} + \frac{G_R^2}{n_R} - \frac{(G_L + G_R)^2}{n_L + n_R}, \quad (2)$$

where G is the Gain value, G_L and n_L denote respectively the sum of gradient Similarity Score and the number of samples on the left branch, G_R and n_R are the same statistics but for the right branch. Gain helps the XGBoost model to examine high-impact features and perform dimensionality reduction by measuring the information Gain in a leaf. Eventually, the important features corresponding to higher Gain are evaluated by the algorithm to generate a prediction.

In this research, there are eight features based on the student's background as input variables and the number of Master's students in Computing Science as the dependent variable to evaluate feature Gain. The XGBoosting Regression model aims to discover the correlation between the student's previous education and nationality toward the Master's student influx in Computing Science. Therefore, a feature with a higher value of this Gain metric implies it is an important factor to the student population.

Lastly, to find the optimal hyperparameter for the model, the cross-validation approach is applied to fine-tune the model and evaluate the performance. The learning curve in **Figure 5** demonstrates the Root Mean Squared Error (RMSE) of training and testing results as increasing the number of iterations in validation. It shows the best performance stopped by 19 training iterations with the lowest RMSE in the testing set and without overfitting problems by comparing to training loss. As the result, the accuracy reaches around 87% and the results are reliable by visualizing the differences between predicted outcomes and the actual values in the testing set in **Figure 6**.

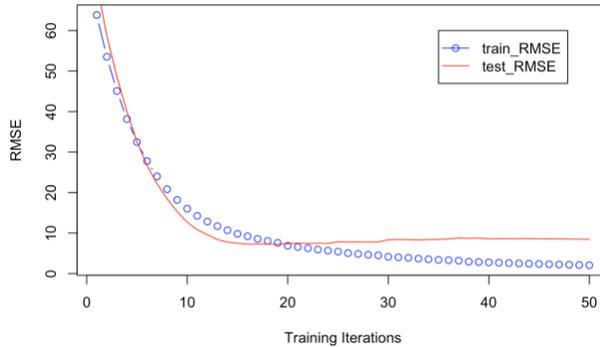


Figure 5. The learning curve of XGBoost Regression Model by calculating the RMSE loss.

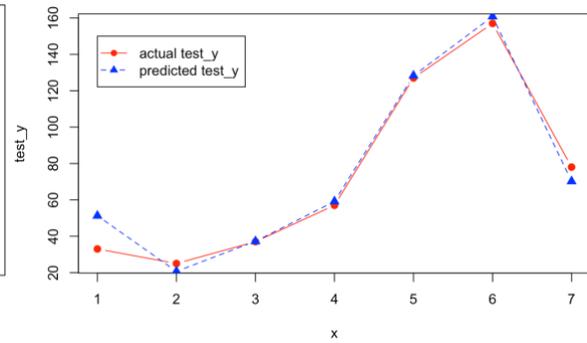


Figure 6. Visualize the model performance by comparing the predicted values and actual values in the testing set.

3 Results

The predictive XGBoost regression model is implemented to forecast the long-term student influx based on the simulated data from 2020 to 2024. **Figure 7** demonstrates the forecasting trend and numbers of Master's student influx in Computing Science at Utrecht University. In respect of the result, there is a slight decrease in 2023 but generally, the incoming student population follows an upward development in the period 2020–2024.

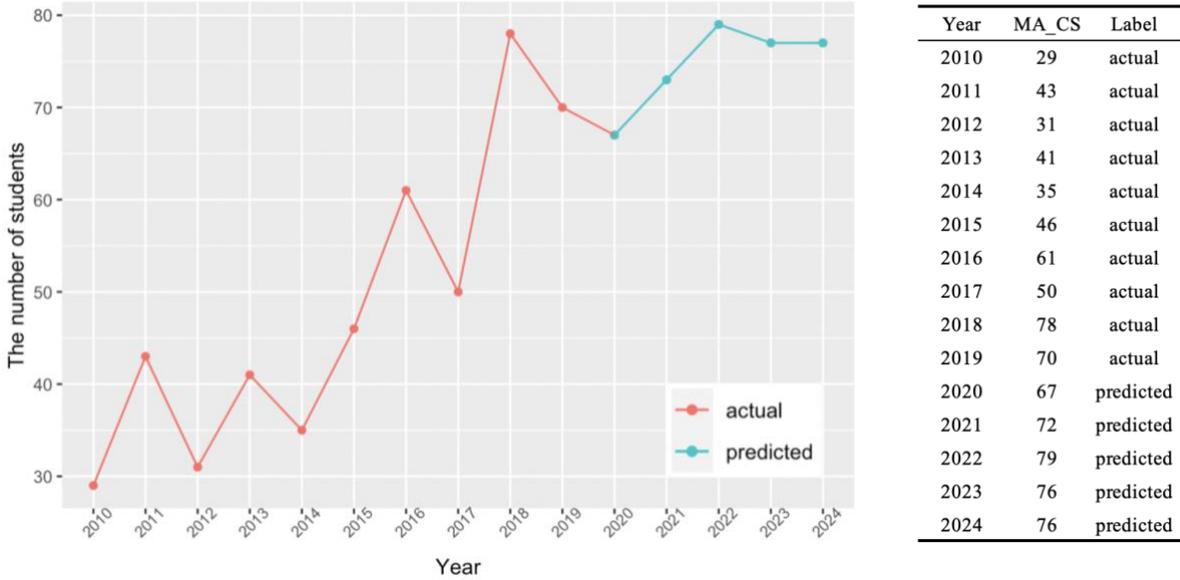


Figure 7. The long-term prediction of Master's student influx in Computing Science at Utrecht University. The red line is the observed data, and the blue line is the prediction by XGBoost Regression (left). The actual value of the student population from 2010 to 2019 and the forecasted results in the coming five years 2020–2024 (right).

The average Gain represents the feature importance across all splits where features are accumulated in the model in **Figure 8**. It indicates the correlation between the student's previous education and region composition toward the prediction. In this case, the top three high-impact features of the student influx are the number of Bachelor's graduates who continue to study for Master's in the same university (BAtoMA), the number of Bachelor's students in Physics and Astronomy (BA_Physics_Astronomy), and the number of students from EEA (BA_EEA). The analysis answers the research question of important factors and specifies that the Master's student's influx in Computing Science at Utrecht University is mainly influenced by the intention of Bachelor's graduates to pursue Master's degrees, especially the students from Physics and Astronomy programme. Furthermore, even if the number of Dutch students reaches a natural ceiling in population increment, the international intake from EEA could intervene more severely in the student influx at a university and affect the general growth.

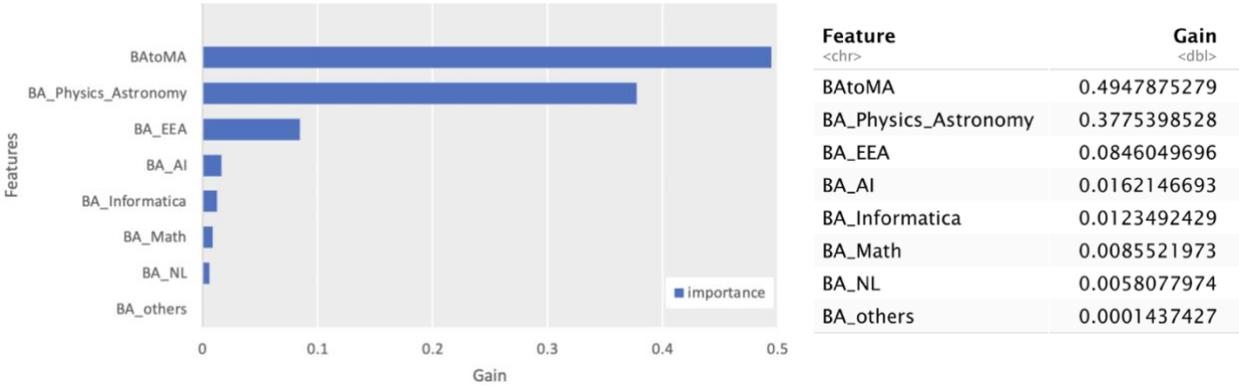


Figure 8. The visualization (left) and values (right) of feature importance which calculated by the Gain algorithm of XGBoost Regression.

4 Conclusion and Discussion

The XGBoost regression model has remarkable success in predicting the long-term Master’s student influx of Computing Science at Utrecht University with regard to the student’s previous education and nationality. The results not only contribute to answering the research question of the incoming student population but also provide insight into influential features for educational institutions to initiate a long-term strategy.

However, there are some challenges and technical limitations worth discussing in this research. Firstly, the training dataset is extracted from the student data at eight different universities to overcome the insufficient amount of data. Although the data preprocessing proposes the identical relation and timestamps among these universities, there are still different patterns between the eight universities under the hood. If the targeted predicted university has enough historical data to train, it will be more effective and perfect for the model to learn solely based on a single source. Secondly, the macro-environment can be considered as a significant effect but unpredictable inference to the student influx which is hard to measure in this study. For example, the policy statement of Brexit decreased the number of students from the UK studying in Dutch universities in 2020; the global outbreak of coronavirus cause a decrease in the number of international students but encouraged more local students to study in the Netherlands [2]. These policy decisions and crises are impossible to predict and it is hard to quantify their impacts. It becomes a critical challenge for the model to understand the niche value or sudden change in quantitative analysis. Therefore, the prediction of the model has limitations to interpret the external factors because Machine Learning can only learn from the quantified observed data [25]. Overall, this research shows that the XGBoost model can provide reliable results in prediction and feature engineering for the near future, but these unexpected incidents need to take into account when people make decisions in real-life applications.

Acknowledgments: The data is provided by Dr. Marc Coemans from Utrecht University. Special thanks for his support in the research.

References

- [1] Universities of the Netherlands, “Quality of education and research under pressure from drop in funding per student,” Vereniging van Universiteiten. 2022. [Online]. Available: <https://www.universiteitenvannederland.nl/dalende-rijksbijdrage.html>
- [2] A. Elfferich, S. Favier, and F. Snethlage, “Incoming degree mobility in Dutch higher education 2021-22,” the Dutch organisation for internationalisation in education, Nuffic, April, 2022.
- [3] D. Fouarge, J. Bakens, I. Bijlsma, S. Dijkman, S. Steens, and R. Goedhart, “De arbeidsmarkt naar opleiding en beroep tot 2026,” Research Centre for Education and the Labour Market (ROA), Maastricht University, May, 2021. [Online]. Available: <https://roa.nl/project-onderwijs-arbeidsmarkt-poa>
- [4] ERUES, “Labour market information: Netherlands,” European Labour Authority, European Commission. [Online]. Available: https://ec.europa.eu/eures/public/living-and-working/labour-market-information/labour-market-information-netherlands_en
- [5] V. Séveno, “Dutch universities struggle to keep up with growing student population,” I AM EXPAT, November, 2021. [Online]. Available: <https://www.iamexpat.nl/education/education-news/dutch-universities-struggle-keep-growing-student-population>
- [6] S.L. Ho, M. Xie, and T.N. Goh, “A comparative study of neural network and Box-Jenkins ARIMA modeling in time series predictions,” *Computers and Industrial Engineering*, Vol. 42, 2002, pp. 371–375.
- [7] C. Nichiforov, I. Stamatescu, I. Făgărășan and G. Stamatescu, “Energy consumption forecasting using ARIMA and neural network models,” 2017 5th International Symposium on Electrical and Electronics Engineering (ISEEE), 2017, pp. 1-4.
- [8] T.J. Mbah, H. Ye, J. Zhang, and M. Long, “Using LSTM and ARIMA to Simulate and Predict Limestone Price Variations,” *Mining, Metallurgy & Exploration*, Vol 38, 2021, pp. 913–926.
- [9] F.V. Şahinarslan, A.T. Tekin, and F. Çebi, “Machine Learning Algorithms to Forecast Population Turkey Example,” *İstanbul Technical University, International Engineering And Technology Management*. 2019.
- [10] V. Ediger, S. Akar, and B. Ugurlu, “Forecasting production of fossil fuel sources in Turkey using a comparative regression and ARIMA model,” *J. Energy Policy*, Vol. 34, No. 18, 2006, pp. 3836-3846.
- [11] T. Abeysinghe, U. Balasooriya, and A. Tsui, “Small-Sample Forecasting Regression or Arima Models?” *Journal of Quantitative Economics*, Vol. 1, 2003, pp. 103–113.
- [12] C.R. Nelson, “Rational expectations and the predictive efficiency of economic Models,” *The Journal of Business*, Vol. 48, 1975, pp. 331-343.
- [13] D.A. Pierce, “Forecasting in dynamic models with stochastic regressors,” *Journal of Econometrics*, Vol. 3, No.4, 1975, pp. 349-374.
- [14] B. Lindemann, T. Müller, H. Vietz, N. Jazdi, and M. Weyrich, “A survey on long short-term memory networks for time series prediction”, *Procedia CIRP*, Vol. 99, 2021, pp. 650-655
- [15] S. Hochreiter and J. Schmidhuber, “LSTM can solve hard long time lag problems,” *Advances in Neural Information Processing Systems*, MIT Press: Cambridge, MA, USA, 1997, pp. 473–479.
- [16] V. Buhrmester, D. Münch, and M. Arens, “Analysis of Explainers of Black Box Deep Neural Networks for Computer Vision: A Survey,” *Machine Learning and Knowledge Extraction*, Vol. 3, No. 4, 2021, pp. 966–989.

- [17] D. Soutner and L. Müller, "Application of LSTM Neural Networks in Language Modelling," Habernal, I., Matoušek, V. Text, Speech, and Dialogue. Springer, Berlin, Heidelberg, TSD 2013. Lecture Notes in Computer Science, Vol. 8082, 2013.
- [18] K. S. Jodha and K. Srinivasan, "Application of LSTM neural network for consumer electronics stock market," 2019 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-TW), 2019, pp. 1-2.
- [19] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2016, pp. 785–794.
- [20] J. Jiang, R. Wang, M. Wang, K. Gao, D.D Nguyen, and G.W. Wei, "Boosting tree-assisted multitask deep learning for small scientific datasets" Journal of Chemical Information and Modeling, 2020, pp. 1235–1244.
- [21] R. Shwartz-Ziv and A. Armon, "Tabular Data: Deep Learning is Not All You Need," IT AI Group, Intel, 2021.
- [22] C. Wang and S. Lee, "Regional Population Forecast and Analysis Based on Machine Learning Strategy," Entropy 23, No. 6: 656, 2021.
- [23] H.S. Hota¹, R. Handa, and A.K. Shrivastava, "Time Series Data Prediction Using Sliding Window Based RBF Neural Network" International Journal of Computational Intelligence Research. Vol. 13, No. 5, 2017, pp. 1145-1156.
- [24] L. Cui, Q. Zhang, L. Yang, and C. Bai, "A Performance Prediction Method Based on Sliding Window Grey Neural Network for Inertial Platform," Remote Sensing, Vol. 13, No.23, November, 2021.
- [25] W.J. Gonzalez, "Artificial Intelligence in a New Context: "Internal" and "External" Factors," Minds & Machines, Vol. 27, September, 2017, pp. 393–396.