Ideal Student Life – Predicting the Students' Satisfaction Levels

Luka Klinčarević, July 2021

Contents

Project Motivation

On June 2nd 1968, youth protests erupted in Belgrade, Yugoslavia - the first mass protest since WWII in this country. They were caused by chronic unemployment and a generally negative outlook on the future they were facing. On the following night, the police retaliated brutally, which resulted in the university blockade. As the number of protestors grew and started including faculty staff, president Josip Broz Tito eventually gave in to their demands. The protest was officially over on June 9th with Tito's appearance on national television, proclaiming that "the students are right" and that "[he is] happy to have such a mature youth". Unfortunately, Tito had no intention of following through with his promise. As the worldwide trend of students' protests ended, he began to prosecute the leading figures of the protest.[1]

Higher education is the basis of the modern world and its economy. The students are a globally-growing demographic, whose welfare will have a tremendous impact on our future production capacities and our ability to tackle the challenges we will inevitably face.[2] We must, unlike Tito, have an understanding ear

for their needs and demands. But in order to understand modern students, we must utilize modern techniques and technologies deployed in the field of data science.

The primary goal of this assessment will be to determine the contributing factors in predicting the students' satisfaction levels. Our analysis will be split into research questions, denoted with QQN, to avoid confusion with the columns of the dataset, which are denoted as QN. The variables for this analysis will be selected or derived from the dataset provided and analyzed in the "Model Development" chapter. The role of the key contributors will be discussed based on the knowledge gathered from the preliminary investigation, conducted in the "Exploratory Data Analysis" chapter.

The Dataset

The data can be obtained from the DataCamp GitHub repository. The origin of the dataset is unknown. However, given the fact that most of the students are from Singapore, it is safe to assume that the survey was conducted there.

The data consists of 19 columns and 2958 rows. The columns of the dataset with their respective data types and missing values are listed below:

Column	Type	NA
Career	object	0
Citizenship	object	0
Nationality	object	0
Year since Matriculation	int64	0
Year of Study	int64	0
Primary Programme	object	0
Gender	object	0
Department	object	0
Housing Type	object	0
Q1-How many events have you Volunteered in ?	int64	0
Q2-How many events have you Participated in ?	int64	0
Q3-How many activities are you Interested in ?	int64	0
Q4-How many activities are you Passionate about ?	int64	0
Q5-What are your levels of stress ?	int64	0
Q6-How Satisfied You are with your Student Life ?	int64	0
Q7-How much effort do you make to interact with others ?	float64	18
Q8-About How events are you aware about ?	float64	18
response_id	int64	0
Q9-What is an ideal student life ?	object	587

dtype: object

The ambiguity of the Q8 column (whether the question is "How many events are you aware of?" or "How do you become aware of events?") is resolved through the inspection of the data type, suggesting that the former is the case.

Analysis Plan

This is a multivariable regression problem, with mixed variable types. Categorical values will be converted into dummies using the pandas.get_dummies() method, then added to the list of continuous variables. A full list of variables will be obtained after the data preprocessing step, and provided in the "Model development" chapter of this analysis. The model will be built using the Extreme Gradient Booster (objective ='reg:squarederror', n_estimators = 10, , seed = 123). The same random seed will be used in the train_test_split part of the analysis (test_size = 0.3). To assess feature importances, the gain metric will be used. A 5-fold Random Search cross-validation will be utilized for hyperparameter tuning (n_iter = 25, scoring = 'neg_mean_squared_error', error_score = 0, verbose = 3, n_jobs = -1). A full list of hyperparameters is provided below:

n_estimators	stats.randomint(150, 500)
learning_rate	stats.uniform(0.01, 0.07)
subsample	stats.uniform(0.3, 0.7)
max_depth	range(3, 10)
colsample_bytree	stats.uniform(0.5, 0.45)
min_child_weight	range(1, 4)

Table 1 -The list of ranges of hyperparameters used in RandomSearchCV

Performance Metrics

RMSE

RMSE is a common evaluation metric for regression problems and is suitable for our analysis. However, given that students report their satisfaction levels on an integer scale, this metric can be somewhat misleading and a supplementary one will be introduced.

Accuracy

Accuracy is not a metric used in regression problems, because we are not only interested in whether our prediction was wrong, but also how off it was. But since the range of values of the target variables is small and discrete, we will obtain the accuracy by rounding our predictions, giving a more intuitive assessment of the validity of our model.

Exploratory Data Analysis

Data Preprocessing

Before we begin our analysis, we must rectify the data by checking whether all the inputs are correct and resolving the missing values.

The missing values from Q7 and Q8 columns appear to be the last 18 entries of this column, suggesting that the columns are simply incomplete, rather than that any systematical errors are present. Given that XGB has built-in imputational mechanisms[3], we will not be performing any manually, but should our model suffer from inadequate accuracy, this option will be taken into consideration. The rows will simply be dropped otherwise, if we need to perform additional analysis on these columns.

As for the missing values in the Q9 column, they will be left untouched, on the premise that silence is also a way of expressing one's opinion.

Each column was then inspected for its unique values, to avoid any errors that arose as a result of typos. This investigation yielded the following results:

- The "Career" column has three unique values (UGRD, GRAD, NGRD), and it was suspected that NGRD is a false entry of UGRD, based on the fact that there are only 46 such entries. Later analysis revealed that this is indeed a separate category, because all of these rows have the value "School of Exchange Students" in the "Department" column.
- The range of possible values for Q1 and Q4 are greater than those for Q2 and Q3, respectively. This is very counterintuitive. How is it possible that students volunteered in more events than they participated in? Likewise, how is it possible that students are passionate about more events than they are interested in? We will assume that the values are correct, but this conundrum will require deeper investigation.
- The column "Primary Programme" has 68 unique entries. This could prove to be a hinderance, as it would be preferable if we could reduce the dimensionality of our analysis. The information about the students' titles was derived from this column and separated into four categories ["PHD", "Master", "Bachelor", "Other"], then stored in a new column "Title". For any future reference regarding the students' professional choices, the column "Department" will be used instead, which has only 21 unique entries. The column "Primary Programme" will not be used further in our analysis.
- Students for whom Year of study and Year since Matriculation differ would be an interesting subgroup to analyze. However, there are only 3 such entries in this dataset, so no conclusions of any statistical significance can be drawn. For that reason, the column "Year of study" will be used in our analysis and the other one will be dropped.

Obtaining any meaningful insight from raw comments, as is the case with the column Q9, can be challenging. It would be preferable if we could derive a measurable quantity from the responses that students provide.[1] For that purpose, the comments were run through NLTK Vader Sentimenter[4] and the compound sentiment component was stored in a new column "Sentiment". The inputs of this column are real number values ranging from -1 to 1, representing the sentimental outlook of the students towards the prospect of an ideal students' life.

Finally, the columns ["Career", "Citizenship", "Nationality", "Year of Study", "Gender", "Department", "Housing Type", "Title"] should be converted to nominal categories.

[1] "Measure what is measurable, and make measurable what is not so." – Galileo Galilei

Intelligence Gathering

QQ1: Is any gender underrepresented?

There are 1899 female and 1059 male students in the dataset. Given the expected uniformity of this distribution, we can conclude that male students are underrepresented. It is unclear whether this holds for the entirety of the college student population. However, the extremely low p-value of the chi-square test performed on this distribution (p-value = 8.19e-54) suggests that it is highly unlikely that this distribution was randomly drawn from an otherwise uniform distribution. Moreover, unequal distribution does not prevail for some subgroups of students. Below is the count plot of the "Title" column, grouped by gender:

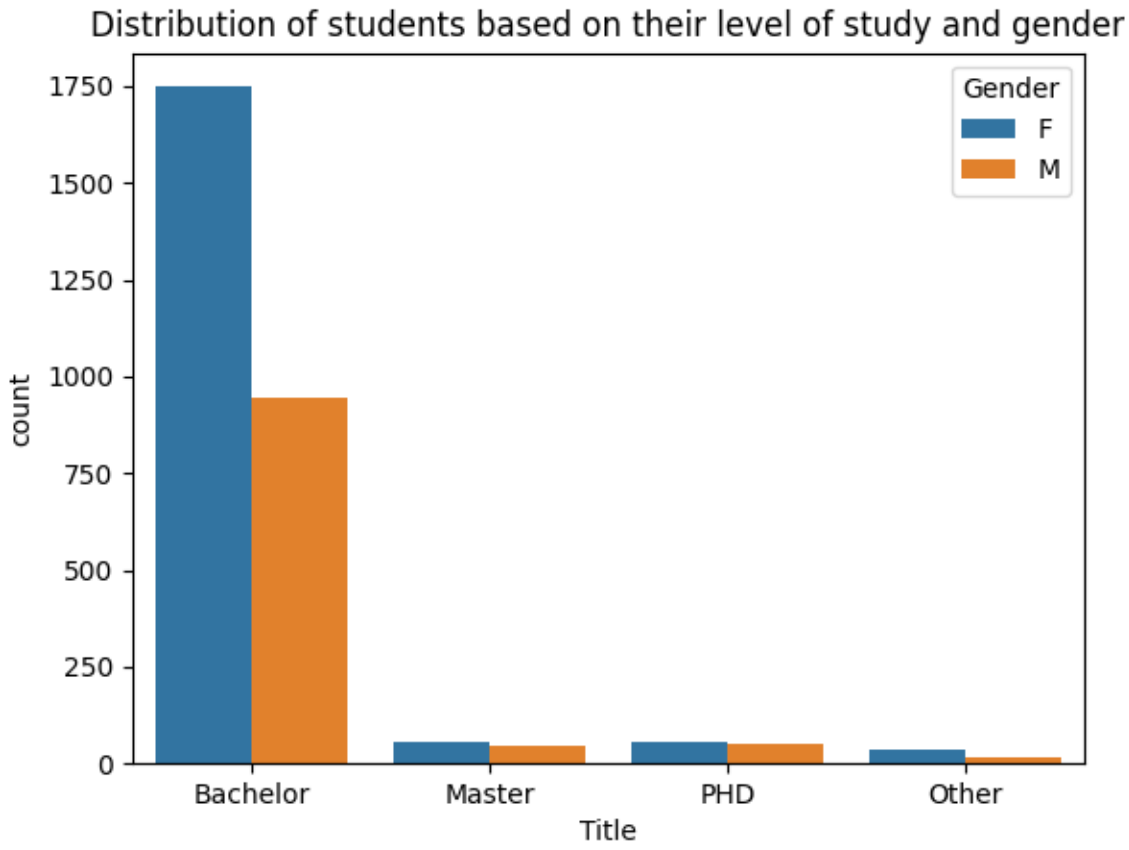

Figure 1 - Students' distribution based on their level of study and gender: F_Bach = 1749, M_Bach = 945, F_Mast = 56, M_Mast = 48, F_PHD = 58, M_PHD = 52, F_Other = 36, M_Other = 14

As we can see, although the Bachelor and Other (mostly exchange) male students are underrepresented, the distribution regresses back to uniformity for Master and PHD students. Several conclusions can be drawn from these findings:

1. Women are favoured during the selection process or display a greater inclination towards engaging in academic studies;
2. Women are favoured or display a greater inclination towards exchange programs;
3. Men are favoured during the Bachelor period, or display a greater inclination towards pursuing post-graduate studies.

QQ2: How much do students attend events?

Columns Q1-Q4 are all related to the question about the general event attendance. While we would like to include this information in our final analysis, it is safe to assume that these columns are not mutually independent. We would therefore like to condense this information to a single column, to reduce the information noise. Below is a triangulated heatmap of the correlation coefficients between these columns:

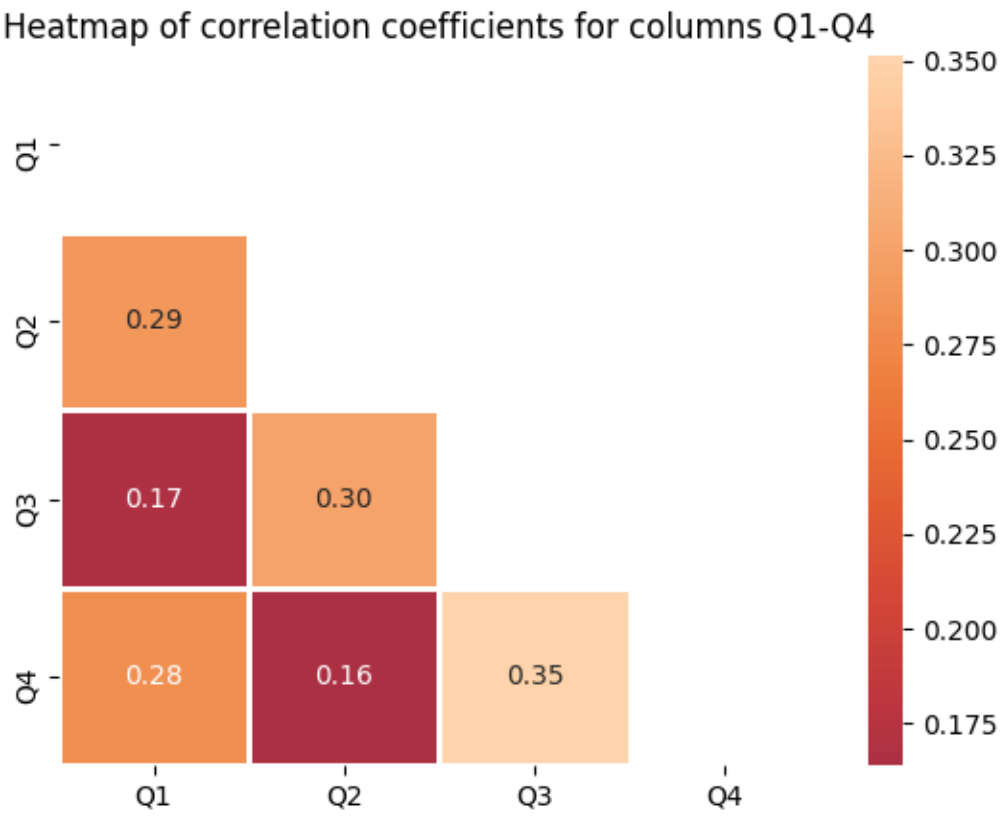

Figure 2 – Heatmap of the correlation coefficients for columns Q1-Q4: The numbers represent the degree of correlation between the columns, with 1 being the maximum value representing absolutely correlated variables.

As we can see, our initial hypothesis was false and the columns do not appear to be sufficiently correlated. Our initial concern about dimensionality still remains however, and should our model suffer from over-fitting, we will resort to feature extraction for these columns. But before we do that, let's focus our attention on extracting as much information from the raw data of these columns before it gets filtered

out in later analysis. The table of the most common distributional parameters of these columns is provided below:

column	Q1	Q2	Q3	Q4
mean	1.248141	1.493577	2.733942	3.184246
std	1.480097	0.958274	1.269236	1.969894
min	0	0	1	0
25%	0	1	2	2
50%	1	1	3	3
75%	2	2	3	4
max	11	5	8	11

Table 2 - The distributional properties of columns Q1-Q4. Columns denoted with N% represent the N^{th} percentile.

The distribution of the Q1 column is indicative of the possible presence of outliers in this column. We can confirm this by doing the violin plot of the said column:

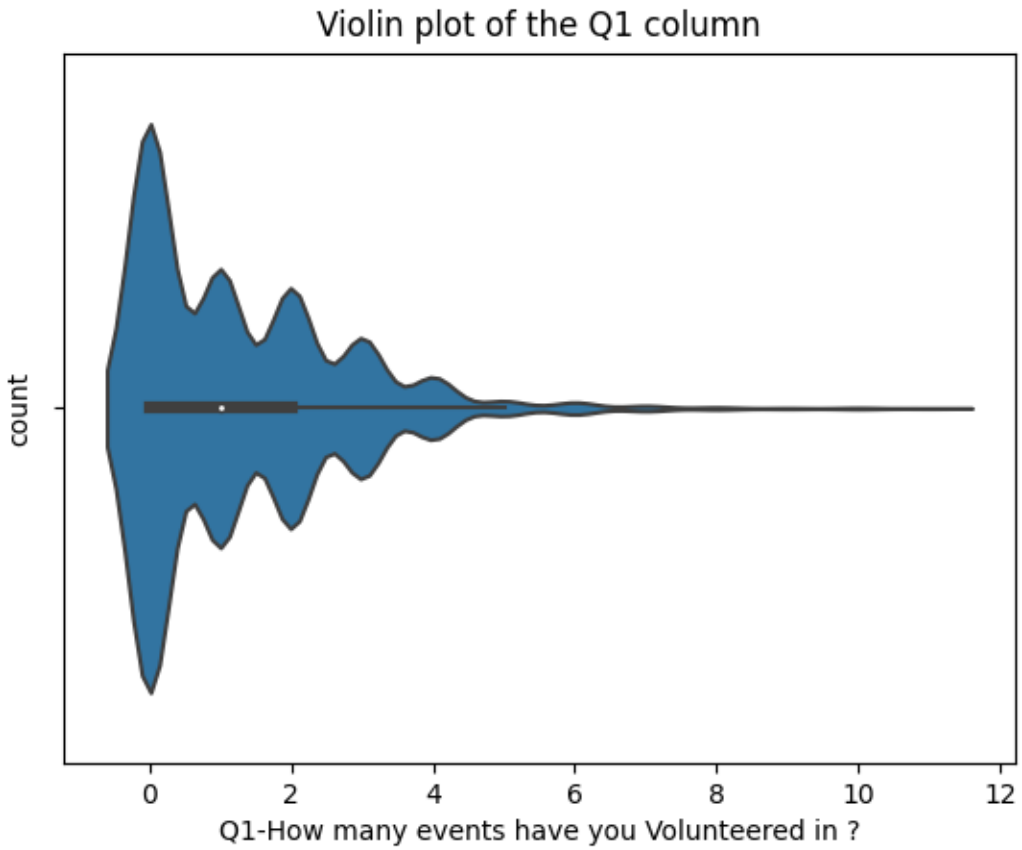

Figure 3 – Violin plot of the Q1 column. The long tail of the violin is indicative of outliers.

Given this strong presence of outliers, the column Q1 will not be used in our further analysis. The column Q4 will not be used either for the same reason. Finally, the fact that the mean of the Q3 column is greater

than the mean of the Q2 column suggests that there is room for expansion – the policymakers should introduce approximately 1.4 (the difference of the means) new events to meet the students' needs.

QQ3: How do students learn about events?

The correlation coefficient between the Q2 and Q8 columns ($corr\_coess_{2/8} = -0.11$), suggests that knowing about events and actually participating in them are not only unrelated, but are slightly anti-correlated. This phenomenon is even more pronounced when we examine the correlation coefficient between the Q7 and Q8 columns ($corr\_coess_{7/8} = -0.43$), suggesting that making an effort to interact with others leads to less knowledge about events transpiring. This may seem counterintuitive, but an explanation can be provided.

We conjecture that the main outlet for promoting events is through social medias and that students who use them the most (and are hence aware about more events than those that don't), do not make as much effort to make meaningful interactions with other people. By the extension, they are also least likely to be interested in such events. Event organizers should shift their focus from promoting their events on social media (to people who are likely not going to participate anyway) to mouth-to-mouth promotion, targeting students on the extraverted side of the spectrum.[2]

QQ4: Do students have a positive outlook on the prospect of an ideal student life?

The table of distributional properties of the column "Sentiment" are displayed below:

mean	0.328635
std	0.354879
min	-0.9403
25%	0
50%	0.34
75%	0.6486
max	0.9907

Table 3 – Distributional properties of the "Sentiment" column

The mean value of 0.33 suggests that, on average, students share a somewhat positive sentiment towards the prospect of an ideal student life. The standard deviation of 0.35 suggests that this topic is somewhat polarized and that the students don't hold this sentiment uniformly.

Model Development

QQ5: What are the predictors of the students' satisfaction levels?

The model to answer this question will be developed based on the instructions provided in the earlier chapters. Using the Q6 column as the target variable and the columns ['Career', 'Citizenship', 'Nationality',

[2] Curious readers are guided to Malcolm Gladwell's "The Tipping Point" for a phenomenal insight into the importance of mouth-to-mouth marketing.

'Year of Study', 'Gender', 'Department', 'Housing Type', 'Title', 'Q2-How many events have you Participated in ?', 'Q5-What are your levels of stress ?', 'Q7-How much effort do you make to interact with others ?'] as the input variables, we obtain the following performance metrics:

RMSE: 0.55 Accuracy: 72.63%

When we take into account the fact that rounding real numbers to integers has a definitional error of 0.5, we conclude that our model has an accuracy that is nearly on par with that of self-reporting, which is a satisfactory outcome. The top 3 contributors to our model are displayed below:

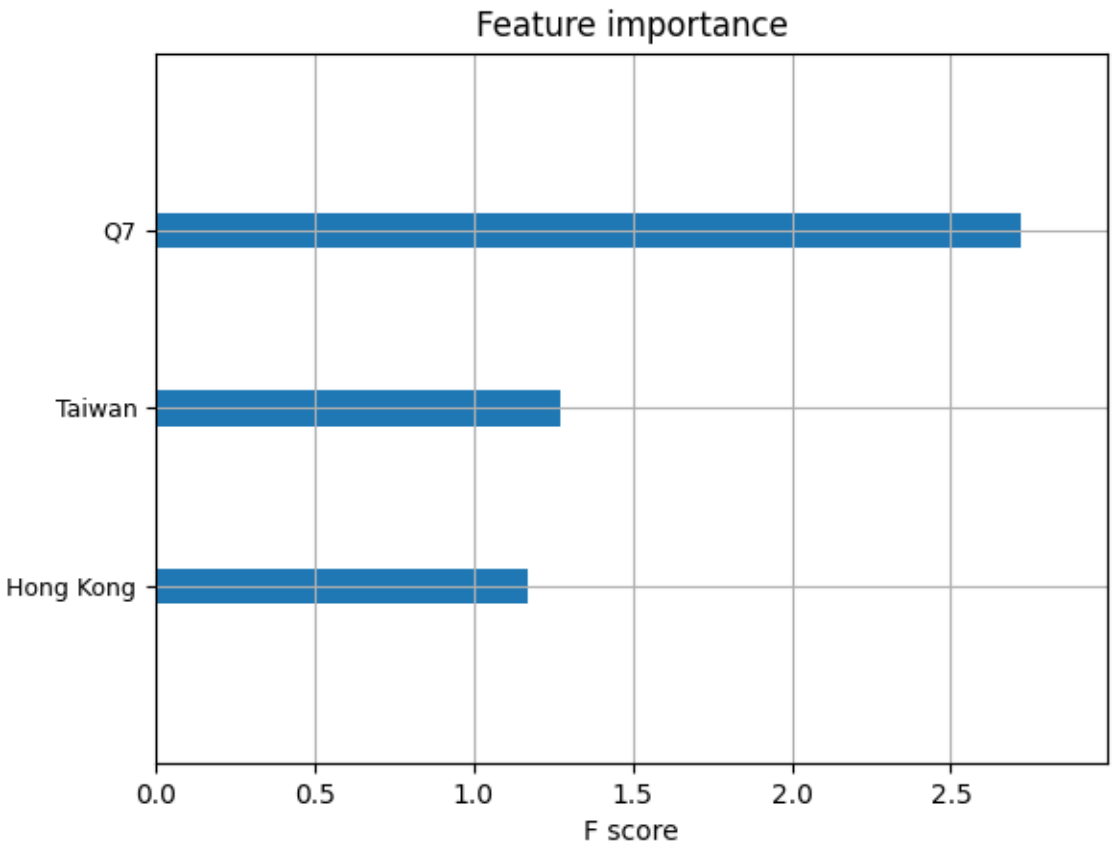

Figure 4 – A graph of the 3 most important features of our model and their F scores

The correlation coefficient between the Q6 and Q7 columns ($corr\_coeff_{6/7}$ = 0.26) suggests that these columns are (unsurprisingly) positively correlated. We conjecture that the effort made to interact with other people is the cause of the student's satisfaction level, but we leave the discussion about the causal relationship between these items to experts in their respective fields.

In order to assess how being from Taiwan and Hong Kong impacts one's satisfaction level, we will compare the mean satisfaction levels of these groups to the overall mean for this column. This data is provided in the table below:

Overall	1.92
Taiwan	1.62
Hong Kong	1.25

Table 4 – A list of mean satisfaction levels by subgroups

As we can see, both of these groups display a lower mean satisfaction level. This is probably due to a large presence of mainland China in Singapore and its current political relationship with these countries. We advise the policymakers to introduce policies that would make the study environment feel less hostile to these students.

Conclusions and Recommendations

Based on the information gathered, interacting with others plays a crucial role in the overall wellbeing of students. In line with conclusions from QQ3 however, simply increasing awareness about events does not appear to contribute to this factor. The policymakers should shift their strategy to, for example, assigning the role of event promoters to charismatic students or individuals. Alternatively, they should consider increasing the possibility of students interacting outside of events, for example, by increasing the campus dormitory capacities.

Given the nature of the survey, it can be assumed that it was conducted in the pre-pandemic environment. A repeated study could provide interesting insight into how the circumstances and the satisfaction impact factors have changed. Furthermore, the columns that contain the information about the students' GPA and different event types would be a welcome inclusion. Lastly, it is possible that the students' answers in columns Q5 and above were primed by previous questions. Meaning, since students were initially asked about event participation, it is possible that their later responses were biased towards this parameter more than they should be. Switching the order of the questions in future surveys could provide another dimension to explore.

We extend our cooperation for any future research that might emerge.

References

[1] Black & Red correspondents. 1968. "Down with the Red Bourgeoisie of Yugoslavia." *Black & Red Number 3* 31.

[2] Calderon, Angel. 2018. *Massification of higher education revisited.* Melbourne, Australia: RMIT University.

[3] D.A. Rusdah, H. Murfi. 2020. "XGBoost in handling missing values for life insurance risk prediction." *SN Appl. Sci. 2* 1336.

[4] Gilbert, C.J. Hutto & E.E. 2014. "VADER: A Parsimonious Rule-based Model for." *Eighth International Conference on.* Ann Harbor.