Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/21/20231145

Analysis of intention to major in computer science in post-secondary education based on random forest

Yu Qiao

J. N. Burnett Secondary School, Richmond V7C5P6, Canada

rq000016@sd38.bc.ca

Abstract. Computer science and artificial intelligence have become an important sector of the world's economy and society. It is essential to determine what factors cause students to study and pursue a career in computer science-related fields to stimulate employment in the industry. In the past, researchers attempted to analyze the effect of different factors on students' intention to study computer science using traditional statistical inference methods. This research, however, used a new way to assess the effects: the random forest model. This study collected two hundred ten responses from a survey posted on online forums. Using the random forest model, the study determined that career satisfaction and personality have the strongest influence on students' intention to study computer science, and family background is less influential. Factors of the three listed categories averaged to have feature importance of 0.1551, 0.1528, and 0.1147 respectively. These results disagreed with prior studies and shed light on further exploration of this topic.

Keywords: random forest, intention to study computer science, prediction, family background.

1. Introduction

In recent years, computer science and artificial intelligence advancements have profoundly impacted our society. Many researchers concluded that the rise of AI might lead to a dramatic change in the world economy because it can potentially increase productivity–which is beneficial to advanced economies with slowing economic growth [1]. It is essential to know what factors influence students to major in computer science in post-secondary education. If strong interests are identified in high school, it would be easier for counsellors to recommend appropriate courses for students. It was shown in previous studies that better course scheduling negatively correlates with grade retention rates but positively correlates with overall academic performances [2]. Additionally, students' interest in STEM subjects (Science, Technology, Engineering, Mathematics, etc,) tends to decline during high school [3]. At the same time, nevertheless, in countries like Canada, computer science related occupations are expected to be facing labor shortages over the 2022-2031 period [4]. Better interventions can be made to promote interest in the labor-demanding field of computer science if factors of interest are identified.

Prior studies on this topic were rare and often old. Giannakos assessed the correlation between motivating factors such as performance expectancy, satisfaction, social influence, self-efficacy, and perceived behavioral control and the intention to study computer science [5]. Similarly, Sathapornvajana and Watanapa assessed the effect of nine factors from two different categories–attitudes toward IT programs and subjective Norms–on students' intention to choose IT programs [6]. The authors eventually rejected all their hypotheses because both their results do not provide any convincing evidence at any reasonable significance level. Another study by Bettina Finzel, Hannah Deininer, and Ute Schmid examined the impact of similar factors on female students' intention to enrol in computer science programs [7].

These studies often assess factors with moderate correlations to the intention to study Computer Science. However, they all attempted to use traditional statistical inference methods such as F-test, T-test, and linear regression to explore the effect of their listed factors. In contrast, this article attempts to predict students' intention to study computer science using random forest–an algorithm with higher classification accuracy and tolerance to outliers and noise [8]. Exploring the relationship between the intention to study computer science and various factors using machine learning methods is significant for scientific research as it opens a new avenue for this research field.

This study analyses the effect of personality, satisfaction, and family background on students' intention to study computer science based on the data of a Google form survey posted on online forums. Based on the data obtained from two hundred and five volunteers, this study constructed and fitted a random forest model to predict the intention of students studying computer science based on their answers to the questionnaire. In section 2, the article introduces the design of the survey and prediction model of the study. In section 3, the text presents the results, analysis, and suggestions. In section 4, the article acknowledges the limitations of this study and advises future studies on this research topic. Conclusions are given in Sec. 5.

2. Methodology

The study designed and posted a survey questionnaire consisting of nine questions. Volunteers were used, and the survey was completely anonymous to reduce potential response bias. The study employed an online survey designed using Google Forms. Indeed, the online survey makes under-coverage bias more likely to occur; however, it is low-cost and efficient (especially for data storage, as all responses are stored in Google Sheets in real-time) [9], which makes it a good fit for this study. The questionnaire was posted on three Reddit forums (r/highschool, r/usaco, r/APStudents) and the Arts of Problem Solving Forum. Based on where the survey was posted, it is likely that the majority of the volunteers are high school students. The survey collected two hundred and ten responses within three days; after that, the survey stopped accepting new responses.

Table 1. Survey design and individual items.

Item 1	Dependent variable	How inclined are you to major in computer science in college/university?
Item 2	Personality	How much do you enjoy STEM-related academic subjects (math, physics, etc,)?
Item 3	Personality	How much do you enjoy NON-STEM related academic subjects (Language arts, History, etc,)?
Item 4	Personality	Would you describe yourself as introverted or extroverted?
Item 5	Personality	How much are you involved in sports activities?
Item 6	Career satisfaction	Do you believe that the advancement of computer science and AI technology will benefit humanity?
Item 7	Family backgrounds	What is the education level of ***the parent that spent the most time with you*** during your childhood?

Table 1. （**Continued**）

Item 8	Career satisfaction	Do the income/working conditions/work style of programmers meet your expectation for your future career?
Item 9	Family backgrounds	What is the estimated yearly household income of your family?

The questionnaire comprised nine distinct questions–eight independent variables and one dependent variable; the random forest model is adaptive to sparsity and should be well-fitted as long as some strong features are selected [10]. The majority of the question was answered using a linear scale labelled from one to ten, with the exception of two questions regarding family background, which had four and five options for the respondents to choose respectively. The categorical data obtained from those two questions were quantified by converting each of the options to a number. The nine questions can be divided into three categories: Personality, Career satisfaction, and Family Background. In the past, Researchers determined that parents influence working conditions, and interests are all factors in career choice [11]. Previous studies also indicated that the influence of family background has a direct relationship with the income trajectories of young adults [12]. In recent years, computer science related occupations are often ranked among the highest in terms of average salaries. As a result, family backgrounds may have a significant relationship with the intention to study computer science. Please refer to Table 1 for specific survey designs

Out of the two hundred ten collected responses. Forty of them are randomly selected to be in the prediction set while the others are put in the train set. A Java program was employed to repeatedly randomly generate integers between two and two hundred eleven; it only terminated when forty distinct integers were generated. All responses that have a row number that corresponds to the forty integers were placed in the prediction set. A standard random forest algorithm was used in the study, and optimal inputs for this study were determined using GridsearchCV at the cross-validation level of 10. Please refer to Table 2 for parameters used for fitting the random forest model. After the random forest model output its prediction for each response, the predictions are compared to the actual values. Another function then takes in the actual values as the y-axis and the predicted values as the x-axis. The R-squared score of the prediction was then calculated.

Table 2. Parameters description.

Parameters	values
n_estimators	57
bootstraps	True
max_depth	2
max_features	auto
min_sample_leaf	2
min_sample_split	5

3. Results & discussion

Seen from Fig. 1, almost all the individual factors expressed weak correlations with the response variables (the absolute value of correlation coefficients ranges from 0.15 to 0.28) except for items two, four, and seven. Item two has a moderate and positive correlation with students' intention to study computer science (r=0.42), and item eight has a strong positive linear relationship (r=0.73). In comparison, item four does not correlate with the response variable (r=-0.024); the responses are almost entirely randomly distributed on the scatterplot with no particular pattern. Thus, item four was not used to construct the random forest model. This result challenges the stereotype that programmers are socially isolated. Item eight's strong positive linear relationship confirmed the results of Kuechler's prior conclusion; many students do not want to major in information systems (IS) is their dissatisfaction regarding the amount of work necessary to obtain and maintain an IS career [13]. Another notable finding is that students' level of enjoyment in non-stem related subjects such as language arts and history have a negative correlation with their intention to study computer science. Given that item four is not used, students' level of enjoyment in non-stem related subjects is the only factor in this study that negatively correlates with students' intention to study computer science. An unintended finding of this study is the relationships between the selected factors. Many factors have moderate or weak correlations with other factors.

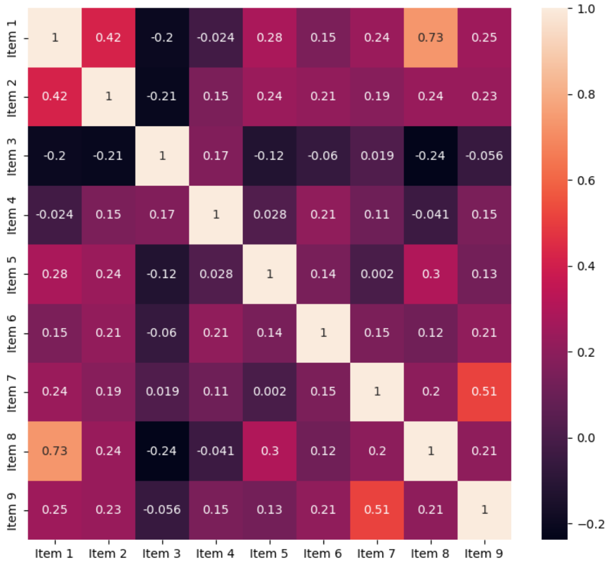

Figure 1. Correlation heatmap.

Table 3. Feature importance and standard deviation.

	Feature importance	SD of feature importance
Item 2	0.1455	0.0036
Item 3	0.1678	0.0033
Item 5	0.1522	0.0032
Item 6	0.1191	0.0028
Item 7	0.1046	0.0027
Item 8	0.1865	0.0032
Item 9	0.1248	0.0033

The process of fitting the model and predicting the response variable using the given factors was repeated one hundred times; the model yielded an average R^2 score of 0.393 with the standard deviation of the R^2 scores being 0.078. The R^2 score is high enough to be considered as acceptable for this research field [14]. A ninety-five percent T-Interval was constructed. The interval ranges from 0.37752 to

0.40848–indicating convincing evidence that the factors selected have a moderate effect on the variability of students' intention to study computer science. It is not plausible that the true R^2 score of the prediction model is lower than 0.3 or higher than 0.5. The model is again refitted one hundred times and the feature importance is printed each time. The mean of the printed feature importance is calculated and recorded. Even though some factors such as item two and item eight have significantly higher correlation coefficients with the dependent variable than other factors, surprisingly according to the random forest model all the feature importance lies between 0.1 and 0.2. The results are summarized in Table. 3 and Fig. 2. Although that item eight and item two have significantly higher correlation coefficients with students' intention to study computer science, they are not much more predictive than other factors. It is certain that the mean feature importance is very close to their actual values because the standard deviation is tiny, and it was previously proven that the random forest model is very consistent [10]. In fact, item two is slightly less predictive of the response variable than items three and five. This indicates that there may exist a moderate or even strong non-linear relationship between item three and item five with the dependent variable.

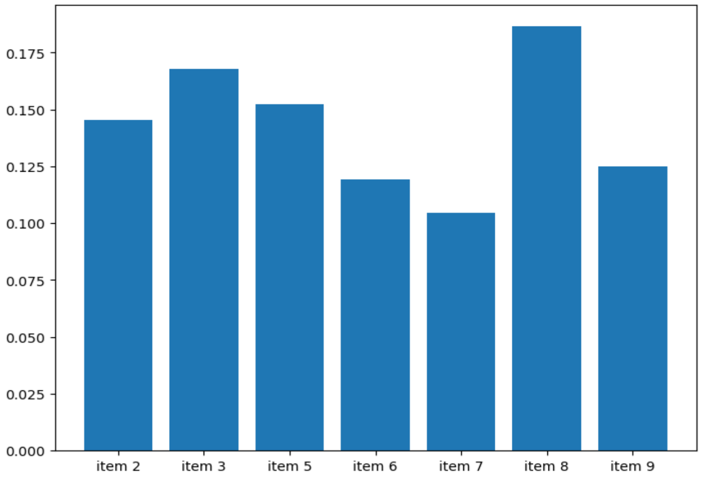

Figure 2. Feature importance bar graph.

Factors of the personality category have a mean feature importance of 0.1551 and a combined standard deviation of 0.0058. Factors of the career satisfaction category have a mean feature importance of 0.1528 and a combined standard deviation of 0.0038. Factors of the family background category have a mean feature importance of 0.1147 and a combined standard deviation of 0.0043. Using two-sample T-tests, it is determined that there are no significant differences between the average feature importance of factors of the personality category and factors of the career satisfaction category at any reasonable significance level (p=0.6299). However, two-sample T-tests revealed that there exists convincing evidence that both the average feature importance of factors of the personality category and career satisfaction category is greater than the family background category at a one percent significance level (p-values equal 0.0019 and 0.0062 respectively). This disagrees with Kazi and Akhlaq's prior study which stated that parental influence is the most significant factor in a student's choice of career [11]. Their conclusion may still be correct, but it does not generalize to computer science-related careers.

Even though the personality and career satisfaction categories are proven to be more predictive of students' intention to study computer science, this article still encourages future studies to utilize these factors in their model. The feature importance of items seven and nine is still significantly greater than zero, meaning this predictor is still relevant [15]. Additionally, this article encourages future studies to explore factors used in prior studies of this field like self-efficacy [16] because none of the factors used in this study have very high feature importance; there may be unused factors with a strong relationship.

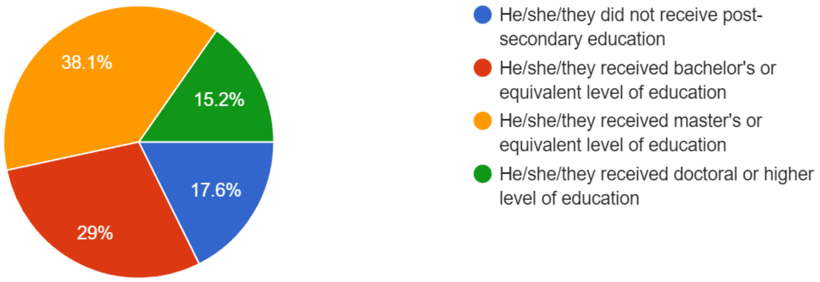

Figure 3. Distribution of responses to item seven.

4. Limitations & prospects

One of the greatest limitations of this study is its sampling method. Due to financial limitations, a simple random sample was unable to be used in this study. This caused under-coverage bias to occur. The bias can be easily spotted in the responses to item seven (seen in Fig. 3). According to Forbes News [17], approximately fifty-three point seven percent of Americans obtained a college or higher level of education. A chi-squared test of goodness of fit is conducted based on results given in Table. 4 as $\chi^2 = (178 - 112)^2/112 + (32 - 98)^2/98$.

Table 4. Chi-squared statistics calculation.

	Observed	Expected
Received bachelor's or higher level of education	178	112
Did not receive bachelor's or higher level of education	32	98

This study obtained a chi-squared statistic of 83.34 and a p-value of 6.9 times ten to the power of negative twenty. It is safe to conclude that there is convincing evidence that the sample is different from the population. This can be problematic; the results of this study may not be able to be safely generalized to the whole population. More accurate results can be achieved using the same methods with randomized samples. In addition, it may also be potentially beneficial to use a blocking design when constructing the random forest model. Sylvia Beyer's research revealed that there exist gender differences in computer self-efficacy and personality [18]. Therefore, it is reasonable for a blocking design to be used when training the random forest model; such a design will minimize the effect of confounding variables. This study may be the first study that attempted to analyse the effects of different factors on students' intention to study computer science. In the future, as more relevant factors are assessed and better sampling methods are used, it is almost certain that future models will be more comprehensive and better serve to analyse the effects of different factors. This study obtained significantly different results from prior studies that employed only traditional statistical inference methods. Future studies should assess the findings of this study using the random forest model and prove whether this study or prior studies are correct.

5. Conclusion

To sum up, this research assessed the effect of eight factors from three different categories on students' intention to study computer science using the random forest model. The results of this study revealed that personality and career satisfaction are approximately equally predictive of students' intention to study computer science. However, family background is less predictive. The finding of this study conflicts with prior studies' results. Future studies should construct random forest models with better sampling methods and more factors to confirm this study's findings. Overall, this study provides guidance for future studies that attempts to analyze factors that affect students' intention to study computer science.

References

[1] Furman J and Seamans R 2018 Ai and the economy. The University of Chicago Press Journals vol 19(1), pp 161-191.

[2] Shi Q, and Brown M H 2020 School counselors' impact on school-level academic outcomes: Caseload and use of Time. Professional School Counseling, vol 23, p 1.

[3] Brophy J 2008 Developing students' appreciation for what is taught in school. Educational Psychologist vol 43(3) pp 132–141.

[4] Government of Canada / Gouvernement du Canada 2023 Web programmer in Canada: Job prospects - the job bank. Web Programmer in Canada | Job prospects - Job Bank. https://www.jobbank.gc.ca/marketreport/outlook-occupation/22539/ca

[5] Giannakos M 2014 Exploring Student's Intentions to Study Computer Science and Identifying the Differences among ICT and Programming Based Courses. The Turkish Online Journal of Educational Technology, vol 13(3) pp 68-78.

[6] Sathapornvajana S and Watanapa B 2012 Factors affecting students' intention to choose it program. Procedia Computer Science, vol 13 pp 60–67.

[7] Finzel B, Deininger H and Schmid U 2018 From beliefs to intention. Proceedings of the 4th Conference on Gender & IT - GenderIT p 18.

[8] Liu Y, Wang Y and Zhang J 2012 New Machine Learning Algorithm: Random forest. Information Computing and Applications, vol 3 pp 246–252.

[9] Siva Durg Prasad Nayak M and Narayan K A 2019 Strengths and weaknesses of online surveys. IOSR Journal of Humanities and Social Sciences vol 24 p 5.

[10] Biau, G 2012 Analysis of a Random Forest Model. Journal of Machine Learning Research vol 3 pp 1063–1063.

[11] Kazi A S and Akhlaq A 2017 Factors Affecting Students' Career Choice. Journal of Research and Reflections in Education vol 11 pp 187–196.

[12] Mooyaart J 2021 The persistent influence of socio-economic background on family formation pathways and disadvantage in young adulthood. Social Background and the Demographic Life Course: Cross-National Comparisons vol 4 pp 139–139.

[13] Kuechler W L, McLeod A and Simkin M G 2009 Why don't more students major in is? Decision Sciences Journal of Innovative Education vol 7(2) pp 463–488.

[14] Ozili P K 2023 The acceptable R-square in Empirical Modelling for Social Science Research. Social Research Methodology and Publishing Results, vol 9 pp 134–143.

[15] Louppe G, Wehenkel L, Antonio S and Geurts P 2013 Understanding variable importances in forests of randomized trees. Advances in Neural Information Processing Systems 26 (NIPS 2013), 1, 431–439.

[16] Shih H P 2008 Using a cognition-motivation-control view to assess the adoption intention for web-based learning. Computers & Education vol 50(1) pp 327–337.

[17] Nietzel M T 2023 Percentage of U.S. adults with college degrees or postsecondary credential reaches new high, according to lumina report. Forbes. https://www.forbes.com/sites/michaeltnietzel/2023/02/01/percentage-of-us-adults-with-a-college-degree-postsecondary-credential-reaches-new-high-according-to-lumina/#:~:text=A%20new%20report%20from%20the,credential%20reached%2053.7%25%20in%202021.

[18] Beyer S 2014 Why are women underrepresented in computer science? gender differences in stereotypes, self-efficacy, values, and interests and predictors of future CS Course-taking and grades. Computer Science Education vol 24(2–3) pp 153–192.