Studying Factors Influencing the Prediction of Student STEM and Non-STEM Career Choice

Varun Mandalapu
Department of Information Systems
University of Maryland Baltimore County
Baltimore, Maryland, USA 21250
varunm1@umbc.edu

Jiaqi Gong
Department of Information Systems
University of Maryland Baltimore County
Baltimore, Maryland, USA 21250
jgong@umbc.edu

ABSTRACT

The increasing capabilities of intelligent tutoring systems (ITS) that collect student interaction data while learning mathematics at the middle school level have enabled researchers in educational data mining (EDM) to develop models that predict student career choice. Current research focuses on feature selection techniques that provide essential features in predicting the target variable. However, the factors that affect the prediction performance of an algorithm at a sample level could be studied in depth as they influence the overall performance of an algorithm. In this study, we analyze the influence of various attributes collected by the ASSISTments online learning platform on the performance of machine learning algorithms in predicting student career fields. Initially, we adopt a feature selection technique based on correlation, ID-ness, stability, and Missing values to determine useful attributes and then apply machine learning algorithms to classify student field. The trained models will be used to extract the supporting and contradicting attributes that influence the prediction performance of an algorithm. The results showed that the affect state confused played a significant role in supporting Non-STEM prediction, and boredom played a substantial role in contradicting STEM predictions while gaming the system influences both STEM and Non-STEM predictions. This proposed study facilitates researches in the field of EDM with factors that influence the development of efficient models in predicting STEM and Non-STEM careers.

Keywords

STEM Career, Factor Analysis, Feature Selection, Educational Data Mining, Affect State.

1. INTRODUCTION

Investigating factors that influence student interest in STEM fields at middle school level supports researchers to develop methods that help to focus on areas that empowers their interest in STEM as a career choice. The increasing adaptation of learning technologies like Intelligent tutoring services (ITS) and Massive open online courses (MOOC) at school level supports researches in the field of educational data mining (EDM) to predict student career choices based on their interaction [6]. With the advancement in the design and capabilities of these systems to collect student interaction data, affect data and knowledge data, different models were developed to understand and predict factors that influence student interest in the STEM field [7]. Social cognitive and career theory (SCCT) show evidence that learning and knowledge pattern at a younger age influences student STEM career [5]. The student interest in mathematics during middle and high school years improve their self-efficacy and performance which can be an influential factor for STEM major enrollment [6,10]. Affective engagement and behavioral models developed earlier showed relationship with choice of majors and college attendance.

An earlier study on predicting student career choice from interaction and affect state data showed a negligible effect of affect state in predicting student career [11]. This study extracts features related to knowledge states based on student problem-solving abilities and skills that were used to predict their fields. Although this study discusses the influence of various predictors on predicting student knowledge states, their approach is to average samples in student log instead of utilizing available comprehensive data. These predictors were then subject to feature engineering and feature selection techniques to incorporate them in algorithm training and testing that improves prediction performance. However, even with the careful selection of predictors following reliable methods the performance of trained algorithms in predicting new student samples is not high. In our study, we incorporate feature engineering and feature selection methods to train and test multiple machine learning algorithms and analyze predictors that support and contradict prediction made by these algorithms. This type of factor analysis will develop an understanding of predictors on classification algorithms.

In this study, we adopt a feature selection technique based on stability, ID-ness, and correlation (Pearson) measures of attributes [4]. We developed three categories (Safe, Moderate and Unsafe) of features based on these measures. Features that fall in the safe and moderate categories were then used to evaluate different machine learning algorithms. The highly supporting and contradicting features for each sample in the dataset were identified based on neighboring attribute weights that utilizes correlation as an identification factor. The local linear relation between attributes is highly influential in prediction compared to the non-linear global relationship [2]. Analysis of features that support and contradict predictions related to STEM, Non-STEM predictions facilitates to understand the importance of each feature on individual class prediction.

Varun Mandalapu and Jiaqi Gong "Studying Factors Influencing the Prediction of Student STEM and Non-STEM Career Choice" In: *The 12th International Conference on Educational Data Mining*, Michel Desmarais, Collin F. Lynch, Agathe Merceron, & Roger Nkambou (eds.) 2019, pp. 607 - 610

2. DATASET

This study adopts the ASSISTments dataset provided during Educational Data Mining (EDM) competition in 2017. ASSISTments platform captured US middle school student interaction data from 2004 to 2007 school years [7,9]. This dataset consists of 1709 student's system interaction data. These students were requested to participate in a survey conducted to record post-high school career achievement. This survey provided the career choices of 591 students. In this, 466 students belong to Non-STEM field, and 125 students belong to STEM field.

3. METHODOLOGY

This study mainly focuses on three aspects; initially, we perform feature selection then evaluate machine learning models and extract predictors that support and contradict predictions. Figure 1. shows the complete methodology of this study.

3.1 Feature Selection

Feature selection technique based on correlation, stability, ID-ness and missing value of an attribute was adopted [4]. Correlation in this study considers the linear correlation between attribute and target column. The percentage of ID-ness implies the percentage of different values present in a column. For instance, an attribute with incremental values can have an ID-ness of 100%. Stability is the measure of constant values in an attribute. Stability is zero if there are no similar values where stability is 100 percent of all the values are the same in an attribute. Missing value measure is the percentage of values missing in a attribute. We categorized these attributes into three categories based on the measures mentioned above.

The unsafe category consists of attributes that have more than 70% missing values, or the column is an ID column which is decided based on the ID-ness value, or stability more significant than 90 percent or correlation less than 0.0001 percent or higher than 95 percent. This study removes the attributes from the dataset as they diminish the performance of algorithms. The moderate category consists of attributes that have an ID-ness value of 85%, or correlation less than 0.01%, or correlation more significant than 40%. Attributes that fall in this category will have minimal impact on the predictions. This study included these attributes for analysis. This category consists of attributes that have low ID-ness, the correlation between 0.01 and 40 %, no missing values and stability less than 90%. Attributes in this category profoundly positively impact prediction.

3.2 Model Validation

We adopt the RapidMiner data science platform to train and test chosen predictive models [3]. In this study, we evaluated five machine learning models that differ based on their principles. We chose gradient boosted tree (GBT), Deep neural network (DL), AutoMLP(Multilayer perceptron) random forest (RF) and logistic regression (LR). We discuss model hyperparameters in below subsection. All the algorithms were evaluated using five-fold cross-validation method on features selected from the above method.

3.2.1 Models and Hyperparameters

1. *Gradient Boosted Tree:* Gradient boosted tree algorithm is a sequential learning algorithm in which a subsequent tree learns from the weak predictors of a previously built tree. The tree adopted in this study has a maximum of 20 trees, maximal tree depth of 20 and a learning rate of 0.1.
2. *Random Forest:* A random forest is an algorithm that works based on ensemble learning principle. This algorithm can combine different models developed based on the bagging method. We obtained optimal settings for this algorithm with a maximum of 100 trees and a maximal depth of 10 per tree.
3. *Logistic Regression:* Logistic regression method is for classification problems as it predicts the probability of each class and classifies based on the probability values. We adopt the standard settings for this model.
4. *AutoMLP:* A multilayer perceptron is a feed-forward neural network that consists of multiple hidden layers in training a neural network. The AutoMLP algorithm can set the optimal learning rate and hidden layers during training. This algorithm works on stochastic optimization and genetic algorithms. This algorithm trains small ensemble methods in parallel with different hyperparameter settings like hidden units and learning rate which are validated to find the best setting.
5. *Deep Neural Network:* A deep neural network is an algorithm that can work with different activation layers, learning rates and optimizers. In this study, we adopt a four-layer (input, hidden_1, hidden_2, and output) fully connected deep learning network that has 250 hidden units in each layer. We set the learning rate at 1.0E-5 and use rectifier activation function. The regularization parameters were auto-adjusted based on the training performance of the algorithm

3.3 Confidence Calculation

In this study, we adopted a confidence-based method that calculates the confidence value ranges between 0 and 1 of student prediction based on the actual and predicted label over all their samples. We extract the cross-validation predictions of each algorithm to calculate the confidence of each student and then label their choice of field as STEM or Non-STEM. If the student prediction confidence over all samples is greater than 50 percent, then the prediction is the same as the actual label. If the student prediction confidence is less than 50 percent, then the prediction is opposite to the actual label of the student.

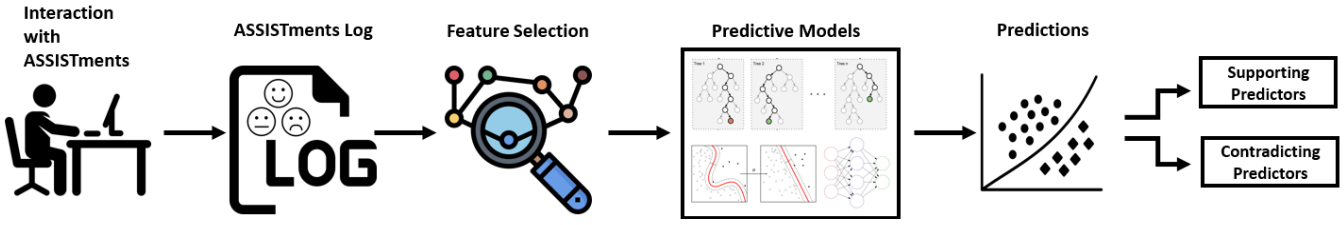

Figure 1: Architectural flow of student career prediction

3.4 Predictor Explanation

This purpose of this study is to understand the factors that influence STEM and Non-STEM predictions. For this purpose, we utilize "explain predictor" operator from RapidMiner to understand the purpose as mentioned above. This method creates neighboring data points for each sample in a dataset and calculates local correlation values to identify the weights of each attribute. The predictors that support and contradict are classified based on the local correlations and weights calculated for each attribute for every sample. The words "supporting" and "contradicting" refer only to the predicted value which might be a accurate or inaccurate prediction. The linear relationship between attribute and prediction locally is highly influential; even the attributes are nonlinear globally.

4. RESULTS

4.1 Cross Validation Performance

This study adopts a cross-validation method to evaluate five machine learning algorithms. Table 1 below shows the cross-validation performance of predictive models built on safe and moderate features categorized by feature selection method. Gradient boosted tree that learns sequentially from weak learners performed better compared to other complex models like deep learning. The performance (AUC, Kappa, and RMSE) of predictive models evaluated on with safe and moderate features show a slight improvement compared to the performance of models based on High impact features.

Table 1: Cross Validation performance of machine learning models on feature selected data with safe and moderate features.

Algorithm	AUC	Accuracy (%)	Kappa	RMSE
Gradient Boosted Tree	0.999	98.83	0.964	0.116+/- 0.002
Deep Learning	0.674	58.19	0.150	0.388 +/- 0.001
AutoMLP	0.623	79.89	0.048	0.396 +/- 0.002
Random Forest	0.635	79.63	0.004	0.398 +/- 0.000
Logistic Regression	0.588	79.58	0	0.400 +/- 0.000

The confusion matrices for STEM and Non-STEM predictions for 591 students were developed based on a confidence cutoff value at 0.5. The below-mentioned table 2 are confusion matrices with Recall and precision scores calculated. As observed earlier GBT and DL does better with high class precision and recall values compared to AutoMLP, RF and LR which were unable to predict STEM classes.

4.2 Explain Predictions

The main focus of this study is to understand the predictions made by the adopted machine learning algorithms. For this purpose, we extract all the supporting and contradicting attributes and present top six in below Tables 3 and 4 for both accurate and inaccurate predictions. These supporting and contradicting algorithms were classified based on the local Pearson correlation values obtained by calculating the correlation between the attribute and prediction made. One should be careful in interpreting support and contradict predictors. For instance, a supporting predictor for a sample with accurate prediction (Actual Label = Predicted Label) means that this predictor acted positively on predicting actual label whereas a supporting predictor for inaccurate prediction (Actual Label ≠ Predicted Label) means that this predictor acted negatively for this prediction. This explanation is similar for contradicting predictors, where the contradicting predictor has negative effect on accurate predictions and positive effect on inaccurate predictions.

Table 2: The below tables shows the confusion matrices with their class recall and precision values for all five machine learning algorithms adopted in this study.

Gradient Boosted Tree	True ST	True NS	Class Precision (%)
Pred. ST	124	0	100
Pred. NS	1	466	99.79
Class Recall (%)	99.20	100	

Deep Learning	True ST	True NS	Class Precision (%)
Pred. ST	95	226	29.60
Pred. NS	29	240	89.21
Class Recall (%)	76.61	51.50	

Table 3: Supporting and Contradicting predictors related to GBT model

Accurate Prediction		**Inaccurate Prediction**	
Supporting	Contradicting	Supporting	Contradicting
NumActions	sumRight	RES_GAMING	NumActions
timeGreater10SecAndNextActionRight	totalFrAttempted	totalFrAttempted	timeTaken
original	sumTimePerSkill	frPast8WrongCount	sumTimePerSkill
frPast5HelpRequest	frPast8WrongCount	hintCount	sumRight
correct	totalFrSkillOpportunities	totalFrSkillOpportunities	totalFrPastWrongCount
manywrong	RES_GAMING	totalTimeByPercentCorrectForskill	Ln

Table 4: Supporting and Contradicting predictors related to Deep Learning model

Accurate Prediction		Inaccurate Prediction	
Supporting	Contradicting	Supporting	Contradicting
totalTimeByPercentCorrectForskill	NumActions	NumActions	timeTaken
timeTaken	totalFrAttempted	totalFrAttempted	sumRight
endsWithScaffolding	attemptCount	frPast8WrongCount	hintCount
sumRight	totalFrSkillOpportunities	frTotalSkillOpportunitiesScaffolding	endsWithScaffolding
correct	frPast8WrongCount	attemptCount	hint
sumTimePerSkill	frTotalSkillOpportunitiesScaffolding	totalFrSkillOpportunities	frPast5HelpRequest

5. DISCUSSION

This study explores the importance of feature selection and investigates the local linear correlation of predictors on predictions. Affect states, knowledge traces, and clickstream records were studied extensively to understand their impact on model predictions. We observe that the "NumActions" has a high impact on overall accurate predictions of GBT but adversely effects Deep Learning algorithm, this might be due to the differences in the statistical background of algorithms and their regularizations functions. Now in case of accurate STEM prediction made by GBT model, attempts and clickstream records support the prediction whereas affect state boredom and disengaged behavior off-task acts negatively on accurate STEM predictions. Affect state confused has a high positive influence in predicting Non-STEM class and disengaged behavior gaming also supports an accurate prediction of this class. Affects states impact on deep learning algorithm seems to be negligible as most of the predictions depend on knowledge states and clickstream records. Gaming the system has negative impact on STEM career prediction and overall predictions. A previous study by San Pedro et al. also found this relationship between gaming the system and Non-Stem students during their major selection [9]. One reason for the pattern mentioned above might be related to students turning from boredom to off-task which negatively impacts STEM choice [1].

Previous studies suggested a high correlation between carelessness and STEM students [6,8]. In this study, the impact of Average carelessness attribute available in this dataset is investigated to check the model performance based on its presence and absence. With the inclusion of average carelessness, the performance metrics of the GBT model increased. From the predictor explanation, we observe that the Average carelessness has a high impact on accurate STEM predictions. This predictor importance is in line with previous studies that proved the importance of carelessness in case of students opting STEM fields [6,10]. One limitation of this study is related to the use of clickstream data which depends on multiple factors like time spent on the system, the number of questions answered which may vary when considering different sets of students that work on the platform during different periods. Another limitation is the generalizability of this study as the dataset analyzed is from a single platform (ASSISTments), and the predictor relevance are model specific.

In our future work, we focus on developing feature selection techniques based on the useful predictors and develop models that efficiently and effectively predict their choice based on their middle school year data.

6. REFERENCES

[1] Bolkan, S., & Griffin, D. J. (2017). Students' use of cell phones in class for off-task behaviors: The indirect impact of instructors' teaching behaviors through boredom and students' attitudes. Communication Education, 66(3), 313-329.

[2] Hall, M. A. (2000). Correlation-based feature selection of discrete and numeric class machine learning.

[3] Mierswa, I., & Klinkenberg, R. (2019). RapidMiner Studio (9.2) [Data science, machine learning, predictive analytics]. Retrieved from https://rapidminer.com/

[4] Mierswa, I., & Wurst, M. (2006, July). Information preserving multi-objective feature selection for unsupervised learning. In Proceedings of the 8th annual conference on Genetic and evolutionary computation (pp. 1545-1552). ACM.

[5] Nugent, G., Barker, B., Welch, G., Grandgenett, N., Wu, C., & Nelson, C. (2015). A model of factors contributing to STEM learning and career orientation. International Journal of Science Education, 37(7), 1067-1088.

[6] Ocumpaugh, J., San Pedro, M. O., Lai, H. Y., Baker, R. S., & Borgen, F. (2016). Middle school engagement with mathematics software and later interest and self-efficacy for STEM careers. Journal of Science Education and Technology, 25(6), 877-887.

[7] Pedro, M. O., Baker, R., Bowers, A., & Heffernan, N. (2013, July). Predicting college enrollment from student interaction with an intelligent tutoring system in middle school. In Educational Data Mining 2013.

[8] Razzaq, L., Feng, M., Nuzzo-Jones, G., Heffernan, N. T., Koedinger, K. R., Junker, B., ... & Livak, T. (2005). The Assistment project: Blending assessment and assisting. In Proceedings of the 12th Annual Conference on Artificial Intelligence in Education (pp. 555-562).

[9] San Pedro, M. O., Baker, R. S., Heffernan, N. T., & Ocumpaugh, J. L. (2015, March). Exploring college major choice and middle school student behavior, affect and learning: what happens to students who game the system?. In Proceedings of the Fifth International Conference on Learning Analytics And Knowledge (pp. 36-40). ACM.

[10] San Pedro, M. O., Ocumpaugh, J., Baker, R. S., & Heffernan, N. T. (2014). Predicting STEM and Non-STEM College Major Enrollment from Middle School Interaction with Mathematics Educational Software. In EDM (pp. 276-279).

[11] Yeung, C. K., Lin, Z., Yang, K., & Yeung, D. Y. (2018). Incorporating Features Learned by an Enhanced Deep Knowledge Tracing Model for STEM/Non-STEM Job Prediction. arXiv preprint arXiv:1806.03256.