A method to predict whether middle school students will enter STEM careers in the future based on FC-Wide&Deep model

Sidan Liu†, Peng Peng and Lei Cao
1. College of Education, China West Normal University, Nanchong, Sichuan, 637000, China

Abstract

STEM education is a hot issue in modern education, and it is important to study whether middle school students enter STEM careers in the future in the early stage of career planning. In this paper, we collected students’ behavioral data through the online tutoring platform ASSISTments, divided the raw log data into five types: single-valued, binary-valued, multi-valued, continuous-valued and cumulative, and aggregated them using different data reconstruction methods. Then, a width & depth prediction model based on feature crossover is proposed to perform feature crossover on the aggregated data, and then the depth and width models are jointly trained using. During the training process, the AUC of the FC-Wide&Deep model improved rapidly from 0.800 to 0.845 in the 1st to 16th training rounds, and then slowly climbed with the increase of training rounds. By averaging the results of the three tests, the AUC index of the FC-Wide&Deep model test results improved by 1.29% compared to the DNN model, and the RMSE index improved by 2.08% compared to the BSN-FM model. The FC-Wide&Deep model is generalizable and generalizable, and can be applied to predict whether students will enter STEM careers in the future, thus contributing to the cultivation and leadership of STEM talents in the field of education.

Keywords: FC-Wide&Deep model; STEM careers; ASSISTments platform; AUC; RMSE
AMS 2020 codes: 68M01
1 Introduction

With the development of the times, advanced science and technology have penetrated into various fields such as national economic construction, cultural prosperity and social development, human existence has undergone radical changes, and the structure of social industries has undergone tremendous changes. The demand for talents in the future society has also changed, and scientists, engineers and high-level technical talents with critical thinking, adaptability, collaborative communication, problem solving and creativity have become the leading human resources in the era of knowledge economy [1]-[3]. However, countries not only face a shortage of STEM talents, but also lack the backup force to promote the sustainable development of STEM talents. STEM education, as a recognized strategic initiative for talents in the 21st century, has become a key direction of educational reform, and countries have introduced relevant supportive policies [4]-[5].

In order to enable more people to enter STEM-related careers in their future career choices, many schools have adopted approaches and measures to induce students to take STEM-related majors and enter STEM careers at work [6]-[7]. In previous studies, STEM development programs adopted by schools can be most effective when the intervention program is adopted before high school or even middle school [8]-[9]. Secondary school has been shown to be a critical period for developing students’ motivation and interest in STEM career choices. During this time, students develop work habits, self-efficacy, and career exploration ideas. Issues such as identifying subtle online learning behaviors for the purpose of cultivation have been addressed by recent advances in computer science.

In response to research on STEM careers, the literature [10] argues that the national focus on rural education highlights a specific need for counseling in the STEM career field and uses a phenomenological approach to examine rural school counselors’ beliefs and experiences regarding STEM career counseling. The literature [11] examined the relationship between STEM interests and career intentions among secondary school students, using the MSOSW program to survey secondary school students and found that 46.6% expressed a desire to pursue STEM careers after testing and that secondary school boys were typically more likely to pursue STEM careers than girls. The literature [12] examined the impact of communities in benign environments on broadening participation in STEM career paths, demonstrating that social context variables can influence the long-term underrepresentation of certain groups on STEM career paths.

The literature [13], on the other hand, examined the impact of high student motivation in math and science on STEM career choices in Swiss high schools, demonstrating that programs that support student motivation increase the intrinsic value of math and science for students and the likelihood of STEM career choices. The literature [14] examined the impact of college- and university-run high school STEM summer programs on high school career choices, using logistic regression modeling and propensity weighting to address differences in group characteristics to model the impact of summer programs. The literature [15] developed an instrument to assess high school students’ STEM career motivation and examined the validity of the program in four dimensions, namely content, substance, structure, and generalizability of validity.

In this paper, we first collect longitudinal datasets of students through the online tutoring platform ASSISTments, where each record records one action and emotional state of a student, and divide the datasets into three parts: training set, validation set, and test set, and use different ways to reconstruct and aggregate for different types of datasets. Next, the transformed features are constructed for the aggregated data using the feature intersection technique, and the width & depth neural networks are formed by constructing width components and depth components. Then, the transformation relationships among the five components of the STEM transformation model were studied, and the objectives, student activities and teacher activities of the “6E” teaching model were analyzed. Finally,
the processed data were put into the FC-Wide&Deep model for training and testing, and the AUC and RMSE metrics were compared with DNN, DeepFM, and BSN-FM models for analysis.

2 Prediction method based on FC-Wide&Deep model

2.1 Data Sample

For the purposes of this paper, the initial logbook dataset is a longitudinal dataset of students collected through the online tutoring platform ASSISTments, a free online formative assessment and tutoring system for secondary school students. Although ASSISTments can be used in multiple domains, it is primarily used in mathematics. Teachers use ASSISTments to assess students’ knowledge of mathematical concepts and skills while facilitating their learning of these concepts. The raw log data on ASSISTments is very redundant in content and will be pre-processed and feature engineered for better experimental results, resulting in a clean sample set of data.

2.1.1 Data sources

The ASSISTments team tracked students longitudinally to see who graduated from high school, those who went on to college, what their majors were, and finally, whether they chose a STEM (science, technology, engineering, and math) track as their first job after college. The dataset contains interaction information from 591 students who used the system during high school and whether they pursued a career in a STEM field after college, and the dataset is divided into three parts in total: a training set, a validation set, and a test set. Although there were only 591 students, each student had hundreds of interactions with the system, which made the dataset redundant, making a total of 316,974 records in the final dataset. Each record recorded one action and emotional state of a student.

2.1.2 Data Reconfiguration

To achieve effective prediction of STEM career choices for each student in the dataset, the first step in this paper is to format and aggregate the original log data set into a tabular data with a total of 591 rows, one row per student. Specifically, for a given variable of interaction information in the raw log data, multiple rows of data for a single student are aggregated and reconstructed into a new variable. First, for the 76 original feature variables in the original dataset, the features with zero variance and those whose meaning could not be understood were removed. After initial screening, the final remaining 69 original feature variables. According to the different format types of these data, they were classified into five types: single-valued, binary-valued, multi-valued, continuous-valued, and cumulative. For these five types of data, different data reconstruction methods were used [16].

For single-valued variables, the aggregation method that preserves the original values is chosen. In the original dataset, the publisher has already collected and organized a part of the data. Although there are hundreds of rows of interaction data for each student, those with the same values under these characteristics in the same student are defined as single-valued variables for this type of characteristic variables.

For binary-valued variables, the aggregation method of summation is used. Some feature variables have data below them all consisting of zeros and ones, for which we define it as a binary type variable.

For multi-valued variables, aggregation is performed based on the number of different kinds of values that occur. Some characteristic variables under which the data are composed of different values or nouns define this type of data as multi-valued variables.
For continuous type variables, two aggregation methods of mean and standard deviation are used for each variable. The mean and standard deviation indicate the trend of concentration of the data and the degree of dispersion among individuals within the group. Most of the values below the characteristic variables in the data set are composed of continuous-type numbers, and these characteristics are defined as continuous-type variables.

For cumulative variables, the aggregation method of final values is used. Some feature variables in the dataset have values that are updated once for each student interaction and are continuously added up by the number of student interactions, which is defined as cumulative data.

After aggregating and reconstructing all the feature variables in the original data, a new dataset containing 101 aggregated variables, one column of student IDs and one column of corresponding target values, with 591 rows and 103 columns, was obtained. Unlike the natural sciences, there are few scientific theories in the social sciences and education that correlate target variables with features for analysis. Some of these 101 characteristics may help predict the target variable in different linear or nonlinear ways, and they may have some meaningful effects on their own or through interaction.

2.2 Feature intersection

2.2.1 Feature intersection principle

Typically, the machine learning (ML) task determines how the ML system processes the sample, which in this paper refers to a student’s sentiment and behavior data in ASSISTments, defined as the set of quantitative features of the object or event to be processed by the ML algorithm.

Define the student example as a vector containing \( n \) different student terms:

\[
M = (m_1, m_2, \ldots, m_n)
\]

where each element of \( M \) is a student in an ASSISTments system \( m_i \). The eigenvector is defined as:

\[
f_i = (f_{i1}, f_{i2}, \ldots, f_{ij}) \in X
\]

It consists of \( j \) unique behavioral and emotional traits \( m_i \) of different characteristics.

The ML algorithm is an algorithmic solution to the mathematical mapping between the feature vectors and the target values. The dataset applied in the learning algorithm is defined as the set of feature vectors \( X \) of \( n \times j \) and target value matrix \( Y \) of \( n \times 1 \), where element \( y_i \in Y \) is the given target value of feature \( m_i \). Then the dataset can be represented as:

\[
T = \{ (x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i) \}
\]

It specifically defines the details of the student samples.

The 101 basic features obtained in this paper and the target values of each sample are sparse for the feature latitude of the subsequent study, and using only these 101 features leads to nonlinear problems in the ML model. The current challenge is to achieve effective generalization and support for nonlinear features that have dealt with the effect of sparse inputs. One solution is to perform a fork
product transformation on each feature to take into account the intrinsic relationship between the features.

In this study, a general linear model had to be developed. The objective is to create a regression mapping between the univariate response variable \( Y \) and the input vector \( X \). Definition \( \hat{y} \) is the desired output of the model with the model parameters:

\[
W = [w_0, w_1, ..., w_{j-1}] \tag{4}
\]

Thus, the general linear model is defined as:

\[
\hat{y}(W, f_i) = w_0 + w_1 \times f_{i1} + w_2 \times f_{i2} + \cdots + w_{j-1} \times f_{ij-1} = W^T f_i \tag{5}
\]

In general, \( f_{i0} = 1 \), the constant term is the deviation. \( \hat{y} \) was obtained using available information and therefore may not fully explain the relationship between \( W \) and \( f_i \).

The method to obtain the transformed features is to embed the feature intersection technique into the basic description. The core part of the feature intersection technique is defined as follows:

\[
\phi_k(f) = \prod_{i=1}^{j-1} f_i^{c_i}, c_i \in \{0,1\} \tag{6}
\]

The constraints are satisfied:

\[
\sum c_i = k, k \in N \tag{7}
\]

where \( c_i \) is a Boolean variable constrained by the summation number \( k \). The Boolean variable \( c_i \) is 1 when the \( i^{th} \) features are used as part of the feature cross transformation and 0 otherwise. This process captures the interactions between features and adds nonlinearity to the generalized linear model.

### 2.2.2 Constructing the transformed features by feature intersection method

In this paper, a dual feature construction method is used to form the transformed features. First, the 101 sentiment and behavior features of the original dataset are retained as they can be directly applied to the statistical learning model. At the same time, a set of large synthetic features constructed in a controlled manner to achieve feature-level nonlinearity is generated to change the data distribution of the dataset. The synthetic feature set is constructed as follows: 8 prototype functions are selected, where \( x \) refers to one of the 101 sentiment and behavior features. By simply applying the prototype functions to feature generation, 808 nonlinearly transformed features are obtained. Then, the feature intersection method is applied to the combination of 101 element descriptors and 808 nonlinear transformation descriptors. Setting \( k = 2 \) to the feature crossover component of the above function results in an additional 413595 crossover features generated. In total, \( 101 + 808 + 413595 = 414504 \) descriptors constitute the overall feature space for the following training model phase. The overall process of constructing the transformed features is shown in Figure 1.
2.3 Wide&Deep model

The width & depth neural network is a network model proposed by Google H.T Cheng et al. The model combines the advantages of both broadband neural networks and deep neural networks, and is able to utilize both manually selected features and depth features learned by the network. The width & depth model has good prediction results due to the combination of the memory capability of the linear model and the generalization capability of the neural network. The network composition of the width & depth model is shown in Figure 2.

![Figure 2. Wide&Deep model network composition](image)

The width & depth neural network consists of two main components, namely the width component and the depth component. The width component is generated based on the width neural network, which is essentially a linear model and can be expressed by the following equation:

$$y = w^T x + b$$  \hspace{1cm} (8)

where $y$ is the predicted value and $x$ represents the selected feature vector:

$$x = [x_1, x_2, \ldots, x_d]$$  \hspace{1cm} (9)
$w$ represents the model parameters:

$$w = [w_1, w_2, \ldots, w_d]$$  \hspace{1cm} (10)

$b$ represents the linear deviation.

When using broadband neural networks, both original and transformed features of the data can be selected as input. The transformed features are the new feature vectors obtained by cross-producting the original features, and the process of feature transformation is as follows:

$$z_i(x) = \prod_{i=1}^{d} x_i^{c_{ki}}, c_{ki} \in \{0, 1\}$$  \hspace{1cm} (11)

Where, $c_{ki}$ is 1 when and only when the $i$th column features belong to the selected transformed features $z_i$. By using the cross product for feature transformation, the linear width network gains some nonlinearity while maintaining the sparsity of the input feature vector, thus enabling the prediction accuracy to be improved. However, the limitation of using only the width neural network is that it can only use the existing data features for combination, and cannot generate some features by itself. To address this drawback, deep neural networks are used to learn some deep features from the data.

In width & depth neural networks, the depth component refers to a forward-feedback neural network, which consists of three parts, i.e., the input layer, the implicit layer, and the output layer. For the input category features, the original features of the input are usually some words, and if these words are directly represented as some independent discrete symbols, it will result in generating a large amount of sparse data, which may cause the model training not to converge. To solve this problem, we need to word embedding the input features, so as to map the sparse features to a new low-dimensional dense space and get a word vector. Word embedding usually chooses a low-dimensional space in the dimension range up to. After word embedding, the depth component will pass the obtained word vector features to the implicit unit for forward propagation, and the output will be computed in each implicit unit according to the following equation:

$$a_{l+1}^{(l)} = f(W^{(l)}a^{(l)} + b^{(l)})$$  \hspace{1cm} (12)

Where $l$ represents the layer number of the deep neural network, $a^{(l)}, b^{(l)}, W(l)$ represents the output, deviation and model weights of the layer $l$ network, respectively. $f$ is the activation function used in the network, and ReLU is usually chosen as the activation function, whose function expression is shown below:

$$F(x) = \max(0, x)$$  \hspace{1cm} (13)

After constructing the width component and depth component, the width & depth neural network is formed. During training, the parameters of both components weights will be optimized, and for the classification problem, the prediction value of the width & depth neural network can be expressed as:

$$P = \sigma(w_{\text{wide}}^T[x, z(x)] + w_{\text{deep}}^T a^*_j + b)$$  \hspace{1cm} (14)
where $P$ is the network output, $\sigma$ is the Sigmoid function, $x$ is the original features of the broadband component input, $z(x)$ represents the transformed features obtained by cross-product transformation of the broadband component input, $a_d^j$ represents the output of the last layer of the depth component, $w_{\text{wide}}$ and $w_{\text{deep}}$ represent the model weights of the broadband component and the depth component, respectively, and $b$ is the bias.

3 STEM education and career choice

3.1 STEM transformation model

Glancy et al. proposed an efficient STEM learning environment based on theories of interdisciplinary problem solving, teamwork and collaborative skills, integration of learning and personal experience, multiple representations, and representational fluency. And based on the idea of integrative STEM education and Lesh’s Representational Transformation Model (LTM), they proposed a transformational model of STEM that emphasizes the connection and transformation between disciplines. Lesh’s representational transformation model is shown in Figure 3.

![Figure 3. Lesh’s Characterization Transformation Model](image)

A STEM learning environment refers specifically to a classroom or school where teachers and students make a conscious effort to reconcile the learning objectives or learning activities of two or more STEM disciplines. It may be that students complete the related learning objectives of each discipline in the same classroom, or it may be that teachers of each discipline collaborate in different classrooms to complete the same instructional objectives in separate disciplines. It follows that integration is the most essential feature of STEM courses. Subdisciplinary learning weakens the connections between disciplines and hinders students’ awareness of the unity of knowledge. In contrast, our lived experiences are inherently holistic in nature, and it is only through reflection that we can identify the different disciplinary knowledge embedded in them. For example, mathematics subject knowledge is more complex, contextual and multidisciplinary, but students are unable to connect what they have learned to contexts outside of school. Thus, a prerequisite for an effective STEM learning environment is the meaningful integration of knowledge across STEM disciplines.
3.2 “6E” teaching model

The American Biology Curriculum Research Team has proposed a “5E” cycle teaching model based on constructivist learning theory. It consists of five instructional components: introduction, inquiry, explanation, refinement, and evaluation, which provide a specific teaching process. The purpose of the constructs is to provide students with experiences that allow them to re-understand the concepts through communication and reflection. “Inquiry is the central component of the 5E model, encouraging students to actively explore and construct their own understanding of scientific concepts and make connections to other concepts.

In the actual teaching practice, the “5E” teaching model tends to make students weaken the design process. Therefore, on the basis of the “5E” teaching model, the “6E” circular teaching model, which integrates design and inquiry and takes “engineering” as the core of teaching, has been proposed. Table 1 is a detailed description of the “6E” teaching model.

<table>
<thead>
<tr>
<th>Sector</th>
<th>Objective</th>
<th>Student activities</th>
<th>Teacher activities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Engage</td>
<td>Stimulate students’ interest in learning</td>
<td>Clarify learning objectives</td>
<td>Introduces the background and concepts related to the topic</td>
</tr>
<tr>
<td>Explore</td>
<td>Allow students to construct their own understanding of the research topic</td>
<td>Concepts are formed in the process of inquiry and questioning</td>
<td>Introduce students to concepts related to modeling</td>
</tr>
<tr>
<td>Explain</td>
<td>Give students the opportunity to refine what they have learned</td>
<td>Comb through and choose the best solution</td>
<td>Introduction to systematic concepts and the interactions between concepts</td>
</tr>
<tr>
<td>Engineer</td>
<td>Deepen your understanding of key issues</td>
<td>Apply concepts or theories related to design</td>
<td>Introduces concepts and resources related to design</td>
</tr>
<tr>
<td>Enrich</td>
<td>Dive into what you’ve learned</td>
<td>Enrich your understanding of engineering concepts</td>
<td>Provide new situations and problems for the application of design concepts</td>
</tr>
<tr>
<td>Evaluate</td>
<td>Determine the degree of student learning and understanding</td>
<td>Conduct formative and summative evaluations</td>
<td>Explain evaluation criteria and assessment tools</td>
</tr>
</tbody>
</table>

4 4. Results and analysis

4.1 4.1 Evaluation index design

The evaluation metrics of commonly used prediction models are mainly root mean square error (RMSE) and AUC, with the former mainly evaluating the accuracy of model prediction and the latter focusing on evaluating the ability of the model to rank positive and negative samples.

The RMSE is calculated as shown in equation (15) below. Where $N$ represents the number of samples and $\tilde{y}_i$ is the true value of the $i$th sample. RMSE is the predicted value of the model for the sample labels. RMSE represents the square root of the difference between the predicted and true values of the model and the square root of the sample size ratio, which can be used to evaluate the accuracy of the prediction model.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (y_i - \tilde{y}_i)^2}{N}}$$ (15)
The prediction results are often used to rank the candidate items, so this paper also chooses the index that focuses on the reasonableness of the response model ranking results, and based on this consideration, AUC is chosen as one of the experimental evaluation indexes in this paper.

The AUC represents the area under the subject operating characteristic curve (ROC). To better illustrate the meaning of the AUC, the concepts of confusion matrix and ROC curve are first introduced.

Each row of the confusion matrix represents the predicted value, and the total number of each row represents the number of samples predicted to be in that category. Each column represents the true category of the data, and the total number of each column represents the true value of the samples in that category. Table 2 shows the confusion matrix. The false positive rate (FPR) and true positive rate (TPR) can be calculated based on the data of the confusion matrix.

<table>
<thead>
<tr>
<th>Confusion matrix</th>
<th>True value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive</td>
</tr>
<tr>
<td>Predicted value</td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>TP</td>
</tr>
<tr>
<td>Negative</td>
<td>FN</td>
</tr>
</tbody>
</table>

AUC is the area under the ROC curve, which considers the classifier’s ability to classify both positive and negative cases. A larger value indicates a more reasonable probability distribution of the model’s predicted results, and a more reasonable result of the ranking. It is still able to make a reasonable evaluation of the classifier in the case of sample imbalance.

In order to evaluate the predicted effect of the FC-Wide&Deep model proposed in this paper more objectively and comprehensively, multi-round training experiments on the model are constructed in this paper to test the performance of the model under real scenarios. A single-round training control experiment is constructed to simulate the training of massive data in real business scenarios. The training models compared in this paper include DNN, DeepFM, and BSN-FM.

### 4.2 Training analysis of the prediction model

The AUCs of the different prediction models for whether junior high school students enter STEM careers in the future are shown in Figure 4. During the training process, the AUC of the FC-Wide&Deep model increased rapidly from 0.800 to 0.845 from the 1st to the 16th training round, and then slowly climbed with the increase of training rounds. The four models performed relatively similarly, with the AUC of training and the AUC of testing increasing with the number of training rounds and finally reaching a level higher than 0.780, and the AUC of training was only slightly higher than the AUC of testing, indicating that the models did not show serious overfitting. When the AUC stabilized, the FC-Wide&Deep model was 2.35% higher compared to the DNN model, 4.84% higher compared to the DeepFM model, and 7.54% higher compared to the BSN-FM. It can be seen that the FC-Wide&Deep model has a small advantage over other prediction models.
4.3 Test analysis of the predictive model

Since random factors are introduced in each model at the stage of parameter initialization and so on. In order to reduce the influence of the random factors, with reference to the previous practice, after several times of tuning the parameters to determine the best parameters for each model, three experiments were conducted for each model, and the average values of AUC and RMSE obtained from the three experiments were taken as the final results. The final test results of the four models are shown in Fig. 5. The AUC metrics of the FC-Wide&Deep model test results showed a 1.29% improvement compared to the DNN model, a 3.96% improvement compared to the DeepFM model, and a 0.865% improvement compared to the BSN-FM model. In terms of RMSE metrics, the FC-Wide&Deep model has a 7.84% improvement compared to the DNN model, a 4.08% improvement compared to the DeepFM model, and a 2.08% improvement compared to the BSN-FM model. Thus, it can be seen that the FC-Wide&Deep model proposed in this paper has different degrees of improvement compared to other models.

Figure 4. AUC of test for different predictive models

Figure 5. Final test results for four models
5 Conclusion

In this paper, a machine learning model is developed to predict whether students can enter STEM careers in the future by using student interaction data obtained from the learning system, and a novel Wide&Deep model is combined, resulting in an FC-Wide&Deep prediction model with a large number of nonlinear features. The following conclusions were drawn from this study:

1) By effectively predicting students at the middle school level, we can learn earlier whether students can enter STEM careers, and intervene with students when teachers teach them so that more students can enter STEM careers that society needs in the future.

2) The key to predicting STEM career choices for middle school students is to develop a good set of learning behaviors and a sufficiently rich set of characteristics on top of the underlying characteristics. Despite advances in automated machine learning tools, these initial steps of forming meaningful features and interactions are best accomplished by a combination of domain experts and data analysts.

3) Middle school students have little awareness of STEM careers, have thin channels of understanding, lack career thinking that reflects the discipline, and are more likely to neglect this aspect in an increasingly demanding curriculum.

References


