Who choose stem career?

Data preprocessing

This 11-year longitudinal data set has 1,242 students and 942,816 transactions, providing 76 variables including combinations of id, knowledge component, time, students’ responses, hint usage, scaffolding, affections and others. The given training data set has 467 students and test dataset has 172 students.

Feature engineering

84 features were made using the dataset to predict students’ stem career choice. Some of features were created by having the maximum of original features, such as ‘Max\_past8BottomOut’, ‘Max\_timeSinceSkill’, and ‘Max\_hintcount’. Some of other features were constructed by obtaining the average of original features, such as ‘Average\_sumTimePerSkill’ and ‘Average\_totalFrTimeonSkill’. Variables regarding students’ responses were also invented such as ‘FirstConCorrect’, ‘IncorrectToCorrect’, and ‘IncorrectToFirstcorrect’. ‘FirstConCorrect’ represents the number of consecutive correct responses in a row before a student make his/her first incorrect (or hint) response. ‘IncorrectToCorrect’ indicates the number of consecutive incorrect responses until a student has his/her correct response. ‘IncorrectToFirstcorrect’ means the number of consecutive incorrect response until a student has his/her first correct response. These features using the sequence of students’ responses were made based on the hypothesis that a student who choose a stem career might have higher number of consecutive correct response before they make incorrect answer and lower number of consecutive incorrect answers than those who have non-stem career.

We built these all features in three different dimensions; student-skill pair level (ID-KC), student-problem type level (ID-PT), and skill-problem type level (KC-PT). We generated a new concept, ‘problem type (PT)’ using ‘step name’ and ‘KC’. It is hypothesized that variables at the ‘problem type’ level can measure students’ general math competency in a different dimension from students’ personal problem-solving ability (ID level) and the difficulty of each skill (KC level).

Table 2. The process for creating Problem Type (PT)

|  |  |  |  |
| --- | --- | --- | --- |
| Problem Name | Step Name | KC(KTracedskills) | Problem Type (PT) |
| ac-arrow-ca-rt-1-p3 | Q1-area-operation | Find added area | rectangle\_  triangle\_area |
|  | rectangle-1-Q1-area | Find individual area  Find individual area in context |  |
|  | rectangle-1-Q1-base | Enter given measurement |  |
|  | rectangle-1-Q1-height | Enter given measurement |  |
|  | triangle-2-Q1-area | Find individual area  Find individual area in context |  |
|  | triangle-2-Q1-base | Enter given measurement |  |
|  | triangle-2-Q1-height | Enter given measurement |  |

Table 2 shows how we built problem type. We arranged the steps and skills based on each problem, extracted keywords from ‘Step Name’ or ‘KC’, and created PT for each problem. The example problem (ac-arrow-ca-rt-1-p3) in Table 2 is assumed to be relevant to ‘the area of rectangle and triangle’ based on its ‘Step Name’ and ‘KC’. The way to invent problem type is not robust and still needs to be investigated further to grasp students’ learning better. Consequently, we have 160 variables on our dataset including 84 variables which were invented (28 variables at the level of student-skill pair, student-problem type pair, and skill-problem type pair respectively) and 76 variables in the original dataset.

Data Analysis

We adopted two different approaches to select significant variables on predicting who choose stem career, forward selection and t-test statistics.

|  |  |  |
| --- | --- | --- |
|  | Forward selection by r | T value over 1 sd (over 2sd are yellow marked) |
| 1 | AveCarelessness | AveCarelessness |
| 2 | Max\_Firstconcorrect | AveKnow |
| 3 | Max\_IncorrecttoCorrect | AveCorrect |
| 4 | Max\_IncorrecttoFirstCorrect | IDPT\_Ave\_Ln.1 |
| 5 | IDPT\_Max\_sumRight | AveResGaming |
| 6 | NumActions | HT/AC |
| 7 | IDPT\_Max\_past8BottomOut | Max\_IncorrecttoCorrect |
| 8 | AveResConf | Max\_IncorrecttoFirstCorrect |
| 9 | IDPT\_Ave\_totalTimeByPercentCorrectForskill | IDKC\_Ave\_Ln.1 |
| 10 | IDPT\_Ave\_timeSinceSkill | IDPT\_Max\_hintCount |
| 11 | IDPT\_Max\_totalFrSkillOpportunitiesByScaffolding | IDKC\_Max\_RES\_GAMING |
| 12 | IDPT\_Max\_RES\_OFFTASK | BOH/AC |
| 13 | ITC/AC` | NumActions |
| 14 | assistmentId | IDPT\_Max\_past8BottomOut |
| 15 | responseIsFillIn | Ln |
| 16 | IDPT\_Max\_conseErrorsInRow | IDPT\_Ave\_totalTimeByPercentCorrectForskill |
| 17 | IDPT\_Max\_RES\_GAMING | IDKC\_Max\_past8BottomOut |
| 18 | AveResGaming | IDPT\_Max\_attemptCount |
| 19 | AveCorrect | IDPT\_Ave\_timeSinceSkill |
| 20 | IDPT\_Ave\_frTimeTakenOnScaffolding | IDKC\_Max\_hintCount |
| 21 | `HT/AC` | Ln.1 |

And other features n independent samples t-test From Table 1, an independent samples t-test on our data shows that there exist significant differences for average carelessness, average student knowledge, average gaming, average correctness, and average number of actions between those who later enroll in STEM majors and those who enroll in non-STEM majors.

there is a strong positive relationship between taking a STEM Major and average correct answers, indicating that success in mathematics using ASSISTments lead to higher probability of pursuing a STEM major. The same strong positive relationship is seen between STEM major enrollment and student knowledge estimate as the student learns with ASSISTments

Taken by itself, the more a student becomes careless or commits more slips, the more likely the student is to choose a STEM major, evidence in keeping with past results that careless errors are characteristic of more successful students [16]. The second concerns the amount of interaction the student has had with the system. Our results show that the number of actions per student is negatively related to majoring in a STEM program, perhaps indicative of struggling students whose actions consist mostly of help requests and scaffolds (which themselves indicate that the student got many problems wrong on the first try).

Conversely, the more a student is bored, the less likely that student is to enroll in a STEM major and more likely to enroll in a non-STEM major – a result compatible with past evidence that gaming is associated with poorer learning in mathematics [17]

In data preprocessing

Comparison balanced data vs skewed data

In methods

Comparison aic and t value

Comparison nostem and isstem

Neural net.

In significance