Decision tree modeling of wheel-spinning and productive persistence in Skill Builders

Article · May 2018

CITATIONS
27

READS
270

5 authors, including:

Shimin Kai
Teachers College
8 PUBLICATIONS **112** CITATIONS
SEE PROFILE

Ma. Victoria Almeda
TERC
27 PUBLICATIONS **203** CITATIONS
SEE PROFILE

Neil T. Heffernan
Worcester Polytechnic Institute
235 PUBLICATIONS **4,502** CITATIONS
SEE PROFILE

Ryan Baker
University of Pennsylvania
428 PUBLICATIONS **11,623** CITATIONS
SEE PROFILE

Decision tree modeling of wheel-spinning and productive persistence in Skill Builders

Shimin Kai
Teachers College
Columbia University
smk2184@tc.columbia.edu

Ma. Victoria Almeda
Teachers College
Columbia University
mqa2000@tc.columbia.edu

Ryan S. Baker
University of Pennsylvania
ryanshaunbaker@gmail.com

Cristina Heffernan
Worcester Polytechnic Institute
cristina.heffernan@gmail.com

Neil Heffernan
Worcester Polytechnic Institute
nth@wpi.edu

Research on non-cognitive factors has shown that persistence in the face of challenges plays an important role in learning. However, recent work on wheel-spinning, a type of unproductive persistence where students spend too much time struggling without achieving mastery of skills, show that not all persistence is uniformly beneficial for learning. For this reason, it becomes increasingly pertinent to identify the key differences between unproductive and productive persistence toward informing interventions in computer-based learning environments. In this study, we attempt to address this by using a classification model to distinguish between productive persistence and wheel-spinning in ASSISTments, an online math learning platform. Our results indicate that there are two types of students who wheel-spin: first, students who do not request any hints in at least one problem but request more than one bottom-out hint across any 8 problems in the problem set; second, students who never request two or more bottom out hints across any 8 problems, do not request any hints in at least one problem, but who engage in relatively short delays between solving problems of the same skill. These findings suggest that encouraging students to engage in spaced practice and use bottom-out hints sparingly is likely helpful to reduce their wheel-spinning and improve learning. These findings also provide insight on which students are struggling and how to make students' persistence more productive.

Key Words: Predictive Modeling, Wheel-spinning, Productive Persistence, Decision Tree, Intelligent Tutoring system

1. INTRODUCTION

Studies outside of the educational domain have long documented the links between "non-cognitive factors" or skills and success in work and personal health (see for e.g., Barrick & Mount, 1991; Bowles, Gintis & Osborne, 2001; Nyhus & Pons, 2005; Chiteji, 2010). While there are various theoretical definitions and terms that describe non-cognitive skills within personality and social psychology, these can generally be viewed as personality and behavioral traits or "patterns of thoughts, feelings and behavior" (Borghans, Duckworth, Heckman & Ter Weel, 2008) that individuals develop through the course of their lives (Garcia, 2014). A broadly accepted model of personality traits is the Five-Factor model (FF), which consists of the overarching traits of agreeableness, conscientiousness, emotional stability, extraversion and autonomy (McCrae & Costa, 1987; Digman, 1990).

Some common examples of non-cognitive skills include traits like conscientiousness, persistence, grit, and self-control (Laursen, 2015; Heckman, Hsee & Rubinstein, 2001; Brunello & Schlotter, 2011). Non-cognitive factors have long been considered important among educators (Montessori, 1912), with increasing support for this belief based on evidence for the importance of these factors for students' achievement in school (Bowles & Gintis, 1976; Klein, Spady & Weiss, 1991; Heckman et al., 2001). With academic interest in measuring the role of such non-cognitive factors in academic achievement, researchers are increasingly exploring the development of alternative methods and instruments that measure them both quantitatively and qualitatively.

One factor that is increasingly thought to be important is whether a student is able to persist or persevere during learning. Persistence is defined as the ability to maintain an action or complete a task regardless of the person's inclination towards the task (Cloninger, Svrakic, & Przybeck, 1993; Duckworth et al., 2007), and is one of the main facets of conscientiousness. As one of the overarching personality traits in the FF model, conscientiousness has been found to predict success (Barrick & Mount, 1991; Goldberg, 1990; John & Srivastava, 1999; McCrae & Costa, 1987) more strongly than the other traits in the FF model. While there has been research on persistence and the larger trait of conscientiousness in the context of the workplace for quite some time (Barrick & Mount, 1991; Tett, Jackson & Rothstein, 1991; Bowles, Gintis & Osborne, 2001; Nyhus & Pons, 2005; Carneiro, Crawford and Goodman, 2007), recent studies have shown that persistence and conscientiousness in educational or academic settings is associated with academic achievement (Borghans, Meijers & ter Weel, 2006; Paunonen & Ashton, 2001; Poropat, 2009), creativity (Prabhu, Sutton & Sauser, 2008) and long-term academic outcomes such as later schooling and future earnings (Heckman et al., 2001; Deke & Haimson, 2006; Poropat, 2009).

However, recent research has suggested that not all persistence is positive. Beck & Gong (2013) suggest that some persistence may be "wheel-spinning", defined as when a student spends too much time struggling to learn a topic without achieving mastery. A student who persists unsuccessfully may face eventual reduced motivation (Sedek & Kofta, 1990).
Ultimately, wheel-spinning may be associated with the failure to seek help when it is needed, often called "help avoidance" (Dillon, 1988), a behavior generally thought to be associated with poorer learning outcomes (Nelson-Le Gall, 1981), though in some situations it is

actually associated with more robust learning (Baker, Gowda, & Corbett, 2011), likely due to incomplete understanding of when help is actually needed.

As such, we have two related phenomena -- productive persistence and unproductive persistence (also called wheel-spinning). We want to encourage students to persist productively, but may want to prevent them from persisting when the eventual outcome will be negative. We may instead wish to encourage these students to stop and seek help or take other corrective actions. The problem is that, in current models, these two situations are indistinguishable. We do not know if a student is persisting productively or unproductively until they have reached a point of failing or succeeding -- and in many cases, the indicator of eventual success may be far in the future.

In this paper, we attempt to use educational data mining (Baker & Yacef, 2009) to distinguish productive and unproductive persistence from each other, early enough in the learning process that intervention is still feasible. We do so by identifying which persistence is productive or unproductive, and then attempting to differentiate these students by their behavior (or the context of their learning) midway through the learning process. We develop this model in the context of the online mathematics learning platform ASSISTments, and discuss the possible implications of our models on future classroom interventions in mathematics learning.

2. Literature Review

2.1. Mathematics Learning in Computer-based Environments

Computer-based and online learning platforms are increasingly being used to create customizable and personalized learning environments for students' individual needs and abilities, particularly in mathematics learning. Despite variations in the types and formats of these computer-based environments, the use of educational technology in mathematics learning has generally been found to have small but significant effects on improving students' academic achievement, particularly with intelligent tutoring systems (see reviews by Steenbergen-Hu & Cooper, 2013; Cheung & Slavin, 2011; Ma, Adesope, Nesbit & Liu, 2014). For example, studies conducted around specific intelligent tutoring systems in mathematics have found significant effects on student achievement in math, such as the SimCalc program (Roschelle, Tatar, Schectman, Hegedus, Hopkins, Knudson & Stroter, 2007), the ASSISTments platform (Feng, Roschelle, Heffernan, Fairman & Murphy, 2014; Roschelle, Feng, Murphy & Mason, 2016) and the Cognitive Tutors in Mathematics (Aleven & Koedinger, 2002; Ritter, Anderson & Koedinger, 2007; Pane, Griffin, McCaffrey & Karam, 2013). Other intelligent tutoring systems that have been found to improve student achievement in mathematics include Wayang Outpost (Arroyo, Woolf, Royer, Tai, & English, 2010) and the ALEKS system, which has also been found to be effective in reducing knowledge gaps in mathematics learning (Huang, Craig, Xie, Graesser & Hu, 2015). Meta-analyses (e.g. Steenbergen-Hu & Cooper, 2013; Cheung & Slavin, 2013; Ma, Adesope, Nesbit & Liu, 2014) have found evidence that despite differences in platforms and formats, intelligent tutoring systems, in general, have been shown to produce positive effects on student achievement in mathematics. These effects have also been found across different populations and length of use of these intelligent tutoring systems.

2.2. Assessing Student Persistence and Wheel-spinning in Computer-based Environments

With a recent resurgence of interest in the role of personality traits on academic achievement and success, more work has been done to create validated measures of non-cognitive skills such as persistence and self-control. Most notably, Duckworth et al. (2007) coined a personality construct named grit, which overlaps with some traditionally measured facets of the Big Five conscientiousness, but was posited to be distinct in other aspects (Duckworth et al., 2007; Duckworth & Quinn, 2009). Duckworth et al. (2007) developed a self-report measure of grit, which has been found to have positive relationships with overall achievement and long-term goal success (Duckworth et al., 2007; Strayhorn, 2014). Other personality researchers, however, have conceptualized conscientiousness as an over-arching personality trait that encompasses other traits like grit, persistence, and industriousness (MacCann & Roberts, 2010; Crede, Tynan & Harms, 2016). Within personality psychology, persistence is considered a personality trait and one of four dimensions of temperament (Cloninger, Svrakic, & Przybeck, 1993; De Fruyt, Van De Wiele & Van Heeringen, 2000). Persistence is defined as the act of persevering in a task despite fatigue or frustration, and in the face of obstacles (Cloninger et al., 1993; Rank, Pace & Frese, 2004). Like grit, it has been found to be strongly correlated with conscientiousness (Crede et al., 2016; De Fruyt et al., 2000; Roberts, Chernyshenko, Stark & Goldberg, 2005), with scales/sub-scales assessing persistence included within self-report and survey measures of conscientiousness in personality and social psychology (De Fruyt et al., 2000; Roberts et al., 2005) Additionally, alternate types of instruments have been employed to complement these self-report measures that assess student persistence and perseverance during learning tasks. For example, Eisenberger & Leonard (1980) created a performance-based measure of persistence, using a perceptual comparison task, which involves the individual detecting as many differences as possible between two pictures. The amount of time spent on each difference was used to measure persistence (Eisenberger & Leonard, 1980).

With the evolution and development of computer-based and online learning environments, recent educational research has delved into the assessment of non-cognitive skills on these platforms as well. One notable contribution to the research in this area is the creation of performance-based assessments of student persistence within the context of video games (Ventura & Shute, 2013). Ventura and colleagues created a measure of persistence within an educational game known as Physics Playground, an open-ended game that encourages the applications of qualitative Physics concepts to achieve game objectives. In their study, 154 students were recruited to play the Physics Playground game over 4 hours, split into five 45-minute sessions, and a game-based assessment (GBA) of player persistence was measured with data logs of these gameplay sessions, using an evidence-centered design (ECD) process (Kim, Almond & Shute, 2016). The measure of persistence was based on the amount of time spent on unsolved trials among all the playgrounds attempted within a player's log file, across all five sessions of gameplay. To validate this game-based measure of persistence, a performance-based measure was adapted from Eisenberger and Leonard's (1980) measure, an anagram riddle task (ART; Ventura & Shute, 2013), in conjunction with a student self-report survey of persistence using a 5-point Likert scale.

The results of this study showed that the amount of time on unsolved trials within Physics Playground was significantly correlated with the performance-based measure of persistence in the form of an ART task, and this relationship was more pronounced among players who struggled in Physics Playground. In addition, both the game-based assessment and ART task

measure of persistence predicted student learning within Physics Playground, even after controlling for other factors such as gender, video game experience, enjoyment of the game and prior knowledge of related Physics concepts. This was in contrast to student self-reports of persistence, which were not found to significantly predict learning within Physics Playground, suggesting that behavioral measures of persistence within the game may be more useful than self-report measures.

Persistence despite failure may be beneficial or detrimental to learning, however, as there may be instances of both productive and non-productive persistence. While productive persistence that leads to mastery is a desirable trait during learning, non-productive persistence could instead lead to poorer learning (Beck & Rodrigo, 2014). Beck & Gong (2013) termed such instances of non-productive persistence as 'wheel-spinning', defined as when a student spends too much time struggling to learn a topic without achieving mastery. Specifically, Beck & Gong (2013) identified wheel-spinning as occurring among students who had 10 or more skill opportunities without mastering a given mathematics concept. Given this definition, Beck & Gong (2013) found that approximately 38% of students did not master a given skill after 10 practice opportunities, and were thus potentially wheel- spinning. With this definition, the researchers developed a wheel-spinning detector for the ASSISTments online learning platform, which identified specific student actions and features occurring during learning that indicate wheel-spinning (Beck & Gong, 2013). One limitation of this wheel-spinning detector was that it was built based on students' mastery of a given skill after more than 10 skill opportunities. However, this definition does not account for students who may have achieved mastery by eliciting external help or obtained the correct answers from a teacher or fellow student. It also does not consider the possibility that some difficult skills may legitimately need more than 10 practice opportunities for students to master. To ensure that student mastery of a given concept or skill is identified accurately, we base our definition of wheel-spinning not just on initial performance but also on retention of the skill learned over time using a delayed test.

Wheel-spinning detectors have also been developed recently in two other studies within other online learning environments. A detector of wheel-spinning was developed on data from 122 students who used Cognitive Tutor Geometry (Matsuda et al., 2016). In this study, Matsuda and colleagues only retained student-skill response sequences that had 5 or more practice opportunities, with the rationale that there was not enough data to identify wheel-spinning with 5 or fewer practice opportunities; this left a total of approximately 3000 student-skill sequences. The researchers first asked human coders to identify students who were wheel-spinning based on students' response data. Using this human coding, they developed a neural network-based model to compute the likelihood of wheel-spinning, attempting to differentiate wheel-spinning students from non-wheel-spinning students, among the students who did not master the skill. While this detector had high recall values, its precision values were relatively low, at around 25%. As such, this detector tended to identify most unsuccessful students as wheel-spinning, even when the human coders did not agree. Additionally, this work by Matsuda and colleagues (2016) differs from the work presented in the current paper in that it did not attempt to distinguish wheel-spinning students from successfully persistent students but instead distinguished wheel-spinning students from non-wheel-spinning students who failed to master the skill (presumably for other reasons).

In addition to intelligent tutoring systems, work was conducted to identify wheel-spinning within Physics Playground. In this context, mastery was defined as having achieved the game level objectives, which manifested in the form of a silver badge (for completing a level) or a

gold badge (for completing a level with an elegant solution). Given the open-ended nature of the Physics Playground environment, wheel-spinning was operationalized in various ways. For example, a student was defined to be wheel-spinning if he/she made attempts within a playground level that took more than 15 minutes, or attempts after achieving a silver badge that did not lead to a gold badge. Based on these operationalizations, researchers made use of machine learning models to create detectors of student wheel-spinning within the game environment, with a predictive accuracy of 82.9% (Palaoag et al., 2016), which was higher than the baseline of 73.4%, and an AUC of 0.853. The baseline of 73.4% was taken from predicting that every data point was the majority class, as 73.4% of data points were originally labeled as not wheel-spinning. However, as with previous efforts, this work did not specifically distinguish wheel-spinning students from successfully persistent students.

In our current study, we focus on the problem of distinguishing wheel-spinning from productive persistence. Differentiating between whether a persistent student will eventually learn the skill or not is valuable for selecting interventions; a wheel-spinning student should probably be encouraged to seek help, whether from the system or their teacher, whereas a productively persistent student should probably be left alone. We do this in the context of the ASSISTments system which is discussed below. ASSISTments has several advantages for studying this problem. One of the foremost is the presence of a functionality that tests whether the student can retain their knowledge over time, known as the ARRS system and discussed below. This enables us to distinguish between students who appear to have achieved mastery but have simply obtained shallow knowledge (or perhaps answers from a classmate) and students whose hard-won mastery is genuine. In the following sections we discuss the ASSISTments platform, and then the detector developed to make this differentiation.

3. ASSISTMENTS

The ASSISTments platform is a free online formative assessment and tutoring system for middle school students. While ASSISTments can be used in a range of domains, it is primarily used for mathematics. Teachers use ASSISTments to assess students' knowledge of mathematical concepts and skills while facilitating their learning of these concepts. The system provides teachers with formative assessments of the students' learning progress in their acquisition of specific knowledge components within the mathematics subject.
Currently, the ASSISTments platform has been adopted by 650 teachers across the United States, with an average of over 5,000 student users a day and over 50,000 student users a year.

In this study, we make use of data obtained from students learning within Skill Builders (Heffernan & Heffernan, 2014), a type of math problem set where students complete several problems related to the same skill. The problems in a problem set may be based on more than one problem template. Problem templates are a problem design for multiple problems where the same problem cover story or design is used but the values change within each problem. In our current sample, a given student encounters between 1 and 340 problems within the same problem set throughout the course of a year. However, the median number of problems that students encounter in a problem set is 11, and the 75th percentile is 22. Additionally, for each problem (i.e., original problem), students have the opportunity to access hints or scaffolding questions. Hints provide a sequence of clues that explain to students how to solve an original problem. The last hint in each hint sequence (called the bottom-out hint) provides students with

the answer to the original problem. ASSISTments also contain scaffolding questions, which break down the original problem into individual steps. These scaffolds are answered in a linear progression, where students must correctly answer the first scaffolding question in order to proceed to the next one. Once all scaffolding questions are completed, students may be prompted to answer the original question again.

Each Skill Builder is based on a single skill mapped to the U.S. Mathematics Common Core standards. Students need to answer 3 questions correctly in a row within a Skill Builder, in order to be considered as having 'mastered' the specific skill that the Skill Builder covers. When the student has completed a Skill Builder, a single-item test is administered after a pre-defined period of time, with a gradually increasing space in between reassessments. This test, consisting of one randomly selected item from a problem template included in the problem set, is delivered through the Automatic Reassessment and Relearning System (ARRS) (Heffernan Heffernan, Dietz, Soffer, Pellegrino, Goldman & Dailey, 2012), and aims to assess the student's retention of the particular skill over time. If the student does not answer this test correctly, he or she will be assigned the Skill Builder to re-learn the forgotten material.

4. METHODOLOGY

In these analyses, we make use of student data from the ASSISTments Skill Builders in the school year of 2014-2015. Data from a total of 23,896 students from 298 different middle schools were used in this set of analyses. These 23,896 students attempted 619 Skill Builder problem sets in 2014.

The objective of the analyses was to build machine learning models that differentiate between students who were persisting productively from those who were wheel-spinning during their work on the Skill Builder problem sets. Both of these groups are considered to be persistent, but one group's persistence appears to produce positive results, while the other group's persistence does not.

First, we operationally defined students who are 'persistent' to be those who worked on 10 or more problems within a single problem set, regardless of mastery. This cut-off was selected in part based on the design of the system, where students are stopped from working on the problem set after they have attempted 10 problems in a problem set within a day. For a student to continue onto the 11^{th} problem in the problem set, he/she must return to the same problem set on a subsequent day.

From the set of persistent student-problem set pairs, we then identified instances of productive persistence and wheel-spinning. Given that our data was collected over the span of a whole year, there have been instances where students were able to attempt the same problem set more than once throughout the year. Students were thus able to achieve the mastery criteria of answering 3 problems correctly in a row - and then attempt the corresponding ARRS test - more than once throughout the year within a single problem set. For tractability, we therefore limit our measure of whether a student is 'productive' or 'wheel-spinning' to the student's first ARRS test outcome and its corresponding set of 3 problems answered correctly in a row.

As shown in Table 1, our operational definitions of 'productive persistence' and 'wheel-

spinning' were based on two measures of learning: mastery (3 correct problems in a row) and retention of knowledge (ARRS test). Specifically, students are classified as 'productively persistent', if they answered 3 problems correctly in a row on or after the 10th problem in a problem set and passed the ARRS test. Conversely, students are classified as 'unproductively persistent', if they answered 3 problems correctly in a row on or after the 10th problem but did not pass the ARRS test. These students demonstrated correct performance immediately on a problem set but not in the longer-term. Similarly, students who completed 10 problems but did not answer 3 problems correctly in a row on or after the 10th problem and, as a result, never received an ARRS test administered to them, were also considered to be 'unproductively persistent'.

Since the administration of the ARRS test could be customized by teachers, data from students who successfully answered 3 problems right in a row but were not given an ARRS test due to teacher customizations were considered to be missing data. It is worth noting that success on the ARRS test can be noisy – it is possible to get an incorrect answer by slipping or a correct answer by guessing (Baker, Corbett, & Aleven, 2008). As with all operational measures, this measure is therefore imperfect, but can still provide a basis for attempting to predict whether a student's persistence will be productive.

In sum, the original dataset consisted of 287,093 student-problem set pairs in total. Of this initial dataset, however, 211,612 student-problem set pairs were removed because they achieved 3 correct problems in a row at any point in time, but did not attempt a corresponding ARRS test. Of the remaining 75,481 student-problem set pairs, only 8,948 student problem set pairs were considered persistent, i.e., having attempted 10 or more problems in a problem set, regardless of mastery. In other words, 66,533 students were excluded from the dataset because they either achieved mastery before attempting 10 or more problems, or they quit the problem set before attempting 10 problems and without achieving mastery. The majority of these non-persistent students (57.3%) mastered the problem set before reaching their tenth problem. The remaining 42.7% of the non-persistent student-problem set pairs quit the problem set without achieving mastery, before attempting their tenth problem.

The final dataset used in our analyses thus consisted of 8,948 student-problem set pairs that were defined as persistent student-problem set pairs, and used to build the final models. Within this final dataset, 2,093 student-problem set pairs were instances defined as productive persistence, while 6,855 student-problem set pairs were defined as wheel-spinning based on our criteria.

According to our definition, then, 9.1% of the total 75,481 student-problem set pairs that had ARRS information involved wheel-spinning, a much lower proportion of wheel-spinning than reported in earlier papers defining wheel-spinning as whether the student took a large number of problems to learn a skill (Beck & Gong, 2013). According to Beck and Gong's (2013) definition, where any student who completes more than 10 problems without getting 3 correct in a row and achieving mastery is wheel-spinning, 46.9% of the total 75,481 student problem set pairs that had ARRS information would have been considered wheel-spinning. When considering the original dataset containing all student-problem set pairs, 12.3% of the original 287,093 student-problem set pairs would have been considered wheel-spinning based on Beck and Gong's definition. Since many of these students do indeed eventually get three correct in a row, and are then able to pass a retention test, we would argue that Beck and Gong's definition contains a great deal of productive persistence. Not all students who struggle are spinning their wheels.

Table 1: Criteria of productive persistence and wheel-spinning, using performance metrics in the Skill Builder system.

10 or more problems	**3 correct in a row (mastery) on or after the 10th problem**	**First ARRS test**	**Definition**
Yes	Yes	Passed	Productive persistence (this paper)
Yes	No	N/A	Wheel-spinning (this paper)
Yes	Yes	Failed	
Yes	Any	Any	Wheel-spinning (Beck & Gong)
No	Any	Any	Neither Wheel-Spinning nor Productive Persistence

5. DATA ANALYSES

5.1. FEATURE ENGINEERING AND FEATURE OPTIMIZATION

In this study, we leveraged a feature set previously created for another analysis (Baker, Goldstein, & Heffernan, 2011). Appendix A provides a list of 25 initially generated core features, which are related to student use of hints in a problem set, the number of student attempts, the number of skill opportunities, and features involving the time between these student actions. These features were identified from within the ASSISTments Skill Builder platform and provided evidence regarding student persistence and learning. Some of these core features which will end up playing important roles in the final model included:

1. Total number of unique problems student has attempted relative to each skill (totalSkillOpportunities)
2. Number of wrong attempts made in the last 5 problems (past5WrongCount)
3. Number of attempts made to solve each problem within a problem set (attemptCount)
4. Amount of time since the current problem set was last seen by the student (timeBetweenProblems)
5. Total number of hints requested within a problem set (hintTotal)
6. Number of bottom-out hints requested in the last 8 problems (past8BottomOut)

The table below shows some descriptive statistics for these core features, while similar statistics for the rest of the core features used can be found in Appendix B.

Table 2: Descriptive statistics for selected core features used during feature engineering

Selected Core Features	Minimum	Maximum	Average	Standard Deviation
totalSkillOpportunities	1	10	4.993	2.736
past5WrongCount	0	5	1.490	1.032
attemptCount (including only hint, scaffolding and answers)	0	151	2.318	2.195
timeBetweenProblems (in seconds)	0	3,538,408,440	878,281.196	2,393,403.523
hintTotal	0	19	1.529	0.984
past8BottomOut	0	8	0.343	0.408

A total of 125 problem-set-level features were then generated from these core features based on their minimum, maximum, average, sum and standard deviations across the problem set prior to the student having reached the 10-problem threshold for being persistent.

Given that the dataset included student data for a whole year, where students could attempt each problem set multiple times throughout the year, the range of values for some of these core features varied quite widely. For instance, the amount of time since the current problem set was last seen by the student (timeBetweenProblems) ranged from a few minutes to several weeks. For most of the core features, however, the range of values obtained were more constrained. Core features such as the number of bottom-out hints requested in the last 8 problems (past8BottomOut), or the number of wrong attempts made in the last 5 problems (past5WrongCount) could only range from 0 to 8 and 0 to 5 respectively.

5.2. Machine Learning

After developing the feature set, we built a set of models predicting a binary variable: whether the student persisted productively or unproductively (i.e., wheel-spinning). We built the model using RapidMiner 5.3 data-mining software (Mierswa et al., 2006), using the Weka J48 decision tree algorithm, an algorithm that has been previously used in building detectors of engagement, affect, and meta-cognitive constructs. Feature selection was conducted using an outer-loop forward selection process, attempting to determine the cross-validated goodness of specific sets of features. It is worth noting that conducting outer-loop forward selection tends to have an upward bias relative to true training- test splits.

The multi-feature models were validated using 10-fold student-level batch cross-validation, using AUC ROC as the primary measure of model goodness. The AUC ROC metric was computed using the A' implementation (Baker, 2008) (rather than computing the integral of the area under the curve) to avoid having artificially high AUC ROC estimates due to having multiple data points with the same goodness, a feature of the integration-based estimates

currently available in most packages (Baker, 2015). A model with AUC ROC of 0.5 performs at chance, and a model with AUC ROC of 1.0 performs perfectly. It is worth noting that AUC ROC takes model confidence into consideration.

We also created a precision-recall curve to identify the tradeoffs between precision and recall at different confidence thresholds of the model of unproductive persistence. Using a precision-recall curve facilitates understanding how well the model functions across its predictive range, rather than at just one point, and can also be used to choose an optimal threshold for interventions with different costs or benefits (Davis & Goadrich, 2016). Precision represents the proportion of instances identified as wheel-spinning that are true instances of wheel-spinning, while recall represents the proportion of instances of true wheel-spinning, which were identified as wheel-spinning. To put it another way, precision indicates how good the model is at avoiding false positives, while recall indicates how good the model is at avoiding false negatives. Together, precision and recall provide an indication of the model's balance between these two types of errors (Davis & Goadrich, 2006).

After creating the J48 decision tree with the initial set of features, we conducted further analysis of its structure. We specifically focused on the features selected in the top nodes in the J48 decision tree, as these play a particularly important role in the tree's process of evaluating specific data points. Presenting a complete analysis of the decision tree structure is outside the scope of a journal paper, as the tree structure was very large (see Appendix C).

The J48 decision tree was built using the WEKA implementation of the C4.5 algorithm, referred to as J48, in the RapidMiner software (Quinlan, 1993). We used the default parameters in the RapidMiner: the confidence threshold for pruning: 0.25, and the minimum number of instances per leaf: 2. The pruning approach of the C4.5 algorithm is based on an estimated error rate of misclassifications at every node of the decision tree. If estimates indicate the tree will be more accurate if subtrees are removed, the children of a particular node can be replaced with a single node, simplifying the classification structure and improving error rate.

6. RESULTS

6.1. MULTI-FEATURE MODEL

The J48 model achieved an AUC ROC value of 0.684. Upon visual inspection, J48 also produced good precision-recall curves. The standard error for the AUC ROC metric for J48 was computed to be 0.003 (using the approach in Hanley & McNeil, 1982). This J48-based model used a combination of 15 features (see Appendix D for a description of each of these features). The final decision tree generated using this algorithm had 95 leaf nodes and 189 decision nodes. The precision-recall curve generated for the J48 model of wheel-spinning is shown in Figure 1.

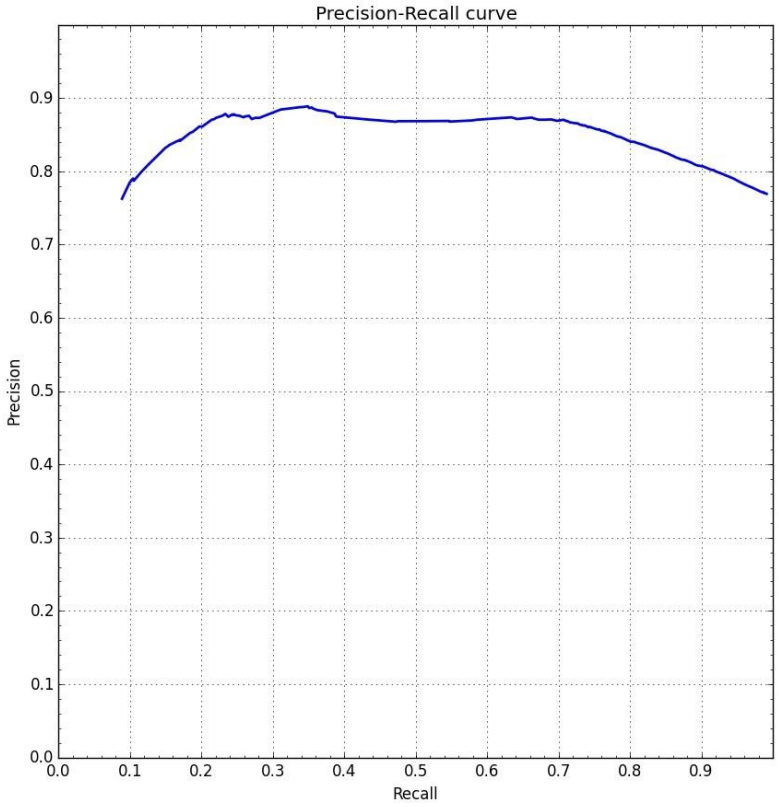

Figure 1: A precision-recall curve for the J48 model's predictions of wheel-spinning

From the precision-recall curve shown in Figure 1, there appears to be a clear tradeoff between precision and recall across thresholds, though the relationship is non-monotonic. The highest precision (nearly 0.90) is seen for relatively low values of recall (between 0.20 and 0.40). These high precision values remain stable as recall increases to 0.7, and only drop slightly afterwards. Overall, it can be seen that the precision of the J48 model (the proportion of true positive predictions out of all positive predictions) remains high (above 0.75) even at high recall values – up to 100% recall.

When analyzing the features in the tree individually, only one feature from the J48 multi-feature model performed at a goodness substantially above chance, the standard deviation of the amount of time since the current problem set was last encountered by the student (std-timeBetweenProblems: AUC ROC = 0.554). As the 15-feature model achieved considerably better predictive performance than the performance of each single feature, this suggests that it is in the interaction of our features that wheel-spinning and productive persistence can be differentiated.

6.2. TOP 3 FEATURES IN J48 DECISION TREE

We can better understand the pattern of relationships that distinguish wheel-spinning from productive persistence by examining the top nodes of the J48 decision tree. This tree has only leaves on one side of the first- and second-level nodes. As such, we can focus on the features in the three top nodes of the tree, which form a single branch. The tree spreads out below these levels. These features are as follows:

- Minimum number of hints requested in any problem in the problem set (min-hintTotal)
- Maximum number of bottom-out hints requested in the last 8 problems, as a rolling average across sets of 8 problems (max-past8BottomOut)
- Standard deviation of the amount of time since the current problem set was last seen by the student (std-timeBetweenProblems)

The second and third of these features deserves a bit of further examination. Both of these features only includes actions in the current problem set – neither feature cuts across problems sets.

The feature, max-past8Bottomout, refers to the maximum number of bottom-out hints requested in the most recent 8 problems within the sequence of 10 problems in this problem set. The possible values for this variable may hence range from 0 to 8 in any given problem set. A value of 0 is obtained when no bottom out hints were requested at all within the most recent 8 problems, whereas a value of 8 is obtained when a bottom-out hint is requested within each of the most recent 8 problems. It is worth noting that because this variable only takes into account the most recent 8 problems within a sequence of 10 problems in the problem set, it is possible to request more than 1 bottom-out hint in the whole 10-problem sequence and still produce a max-past8BottomOut value of 1; specifically when the student requests 1 bottom-out hint right at the beginning of the 10-problem sequence, and 1 more at the end of the sequence. Figure 2 illustrates some of these scenarios, where ‘1’ refers to a problem in which a student requested a bottom-out while ‘0’ refers to when he/she does not.

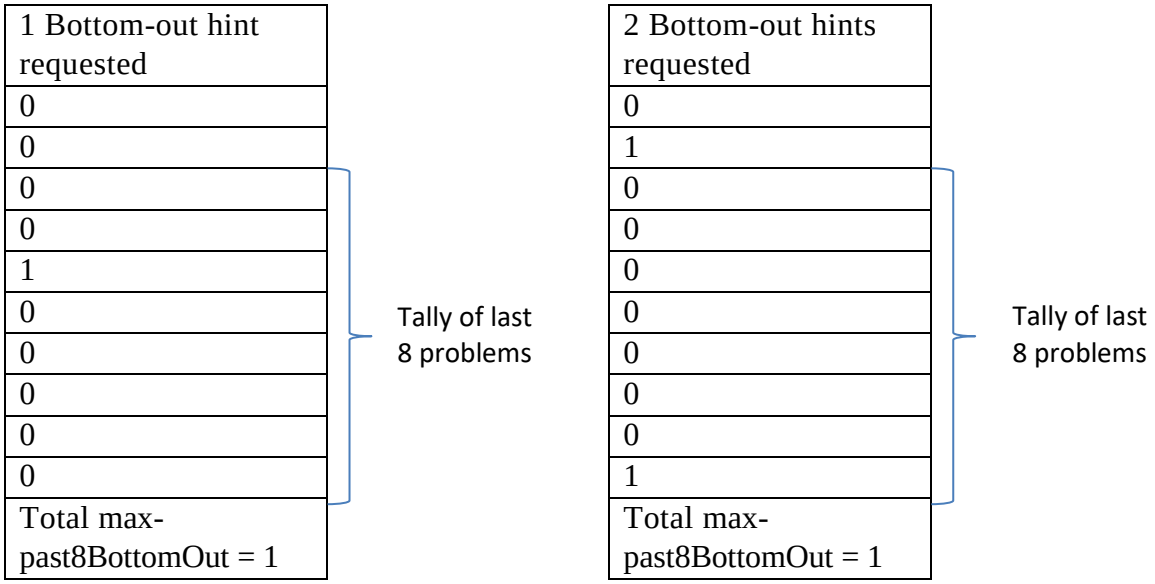

Figure 2: Two scenarios where the value max-past8Bottomout is equal to 1.

The std-timeBetweenProblems variable gives the amount of variation in how much time the student takes between each problem encountered in this problem set. Figure 3 shows examples of student-skill pairs that represent low and high values of this feature in the Distributive Property problem set. Specifically, a low value for std-timeBetweenProblems indicates that the amount of time spent between problems in this problem set are relatively similar. The first and second students in Figure 3 show the patterns in the first 10 problems attempted of student-skill pairs that are defined as wheel-spinning. The first student, for example, illustrates a sequence where a few problems were attempted in a row in the Distributive Property problem set, followed by a break of around 12 days before several more problems were attempted in the same problem set. This pattern would result in a relatively low std-timeBetweenProblems value (compared to other students in our data set). Similarly, the student-skill pair in the second example encounters short periods between problem-solving within the problem set, alternating between a 12-day delay and 2-day delay across the first 10 problems. While both the second and third students have the same values for mean-timeBetweenProblems, their differing patterns in how the first 10 problems were attempted over time led to contrasting values of std-timeBetweenProblems. In the third example, the student attempted several problems within the first 10 problems on the same day, followed by a break of nearly two weeks, before attempting a few more problems in the problem set. This is then followed by a long delay of over a month before more problems in this set were attempted – resulting in a higher std-timeBetweenProblems value relative to the other two student-skill pairs.

As shown in Figure 3, the second and third examples have the same mean-timeBetweenProblems value but differing std-timeBetweenProblems values, suggesting that the same average length of time between solving problems can be associated with different variations of time since a problem set is last seen by a student. When comparing the individual effects between these two features on wheel-spinning, the performance of the J48 model with only std-timeBetweenProblems (cross-validated AUC ROC = 0.544) was very similar to the performance of a similar model with only mean-timeBetweenProblems (cross-validated AUC

ROC= 0.538). These findings thus suggest that std-timeBetweenProblems is about equally as predictive of wheel-spinning as mean-timeBetweenProblems. As such, a multi-feature model with mean-timeBetweenProblems will likely yield a similar AUC performance to our current model. We focus on std-timeBetweenProblems in our analysis solely because it was the feature found in the top levels of the automatically-discovered decision tree.

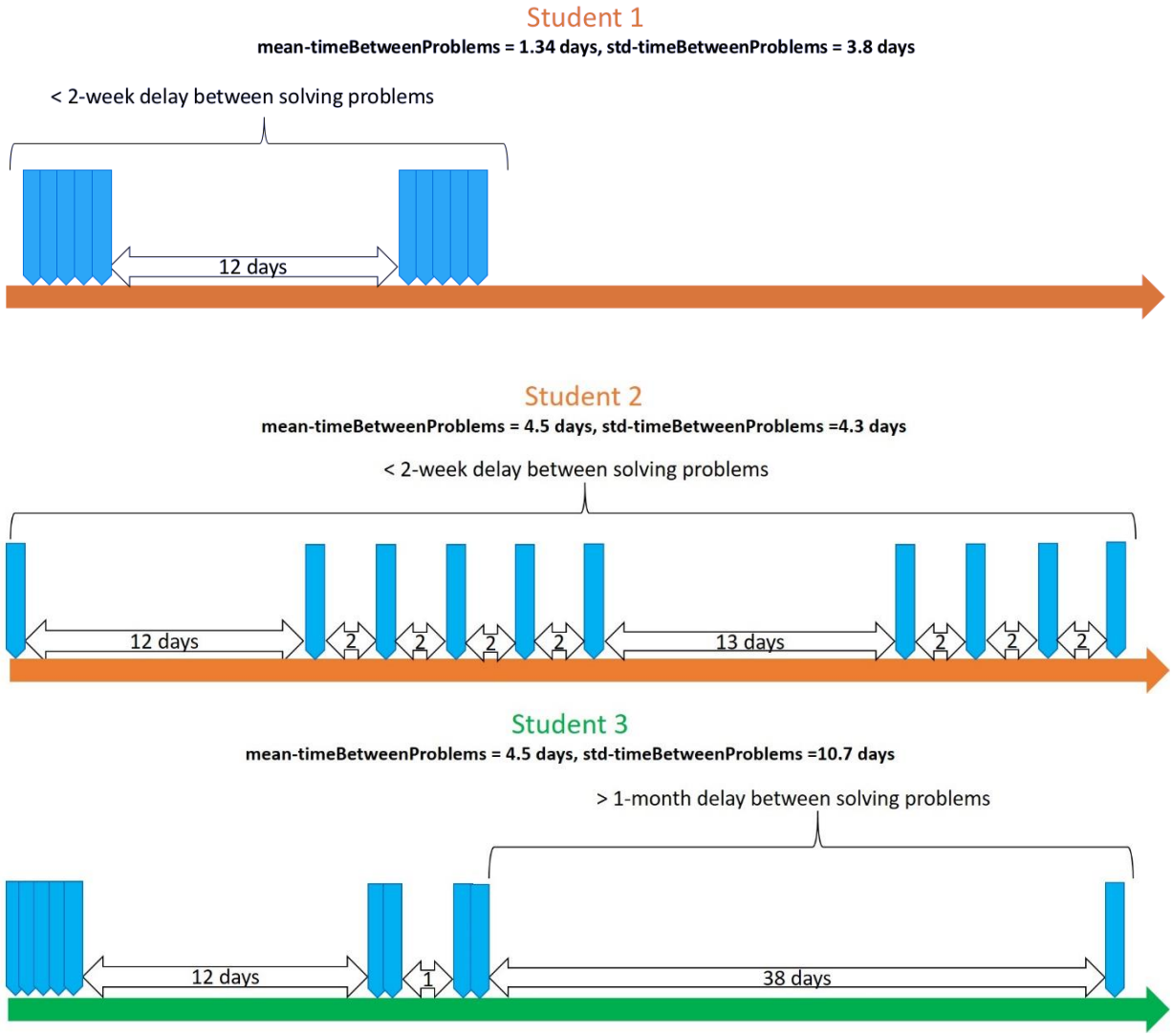

Figure 3: Example of low and high values of std-timeBetweenProblems.

Based on analyses of the top three nodes described above, we found that a considerable proportion of the data is explained by two feature combinations of these three features. Both of feature combinations were associated with high probabilities of wheel-spinning, indicating that there are two distinct types of students who are likely to unproductively persist in the tutoring system.

Based on the two feature combinations, students are more likely to be wheel-spinning in a problem set if they don't request hints for one problem and if, on other problems, they request at least one bottom out hint. On the other hand, students in this group who request less bottom-out hints are also likely to be wheel-spinning if the students spend consistently short amounts of time between the problems they encounter in a problem set.

The first feature combination indicates that students are likely to be wheel-spinning when they do not request any hints in at least one problem within the problem set but request more bottom-out hints in the last 8 problems within the sequence of 10.

Table 3: Relationships between the first combination of features at the top of the tree and wheel-spinning within the J48 model.

Selected Features	**Feature Descriptions**	**Likely to be Wheel-Spinning When**	**Number of instances labeled as wheel-spinning**
min-hintTotal	Minimum number of hints requested in any problem in the problem set	= 0	2544 out of 2856 student-problem set pairs (89.07%)
max-past8BottomOut	Maximum number of bottom-out hints requested in the last 8 problems within the current problem set	> 1	

While the first feature combination indicates that more bottom-out hint requests are associated with more wheel-spinning, wheel-spinning can still occur with fewer bottom-out hint requests. The second feature combination indicates that, even when the maximum number of bottom-out hint requests is 1 or 0, students can still wheel-spin if they do not request any hints in at least one problem and the standard deviation value of the amount of time since the current problem set was last seen is less than or equal to 2.53 days. The probability of wheel-spinning for the second feature combination is lower than the first feature combination.

Table 4: Relationships between the second combination of features at the top of the tree and wheel- spinning within the J48 model

Selected Features	**Feature Descriptions**	**Likely to be Wheel-Spinning When**	**Number of instances labelled as wheel-spinning**
min-hintTotal	Minimum number of hints requested in any problem in the problem set	= 0	5027 out of 6457 student-problem set pairs (77.85%)
max-past8BottomOut	Maximum number of bottom-out hints requested in the last 8 problems within the current problem set	<= 1	
std-timeBetweenProblems	Standard deviation of the amount of time since the current problem set was last seen by the student	< = 2.53 days	

We discuss these three selected features (min-hintTotal, max-past8BottomOut, and std-timeBetweenProblems) and their relationships with student wheel-spinning in the Skill Builders in greater detail in the next section. We will also discuss the goodness of our full 15-feature model in comparison to a smaller model made up of a combination of these three features alone.

6.3. EXPLORING MODEL FEATURES

To gain a better understanding of the range of values for the features at the top nodes of our prediction model, we present graphs showing the frequencies of student-problem set pairs at their respective values for the 3 features of min-hintTotal, max-past8BottomOut, and std-timeBetweenProblems. Included in these graphs are also the corresponding proportions of wheel-spinning students across the value ranges for these features (Figures 4, 5 and 6 respectively).

In the case of min-hintTotal, Figure 4 shows a unimodal distribution. In general, most students do not request any hints in at least one problem. The corresponding proportion of student-problem set pairs that were wheel-spinning based on our operational definition, for each range of values in min-hintTotal, is represented by the line graph in Figure 4. Here we can see that the proportion of student wheel-spinning is much lower if the minimum number of hints requested in any problem increases beyond 1.

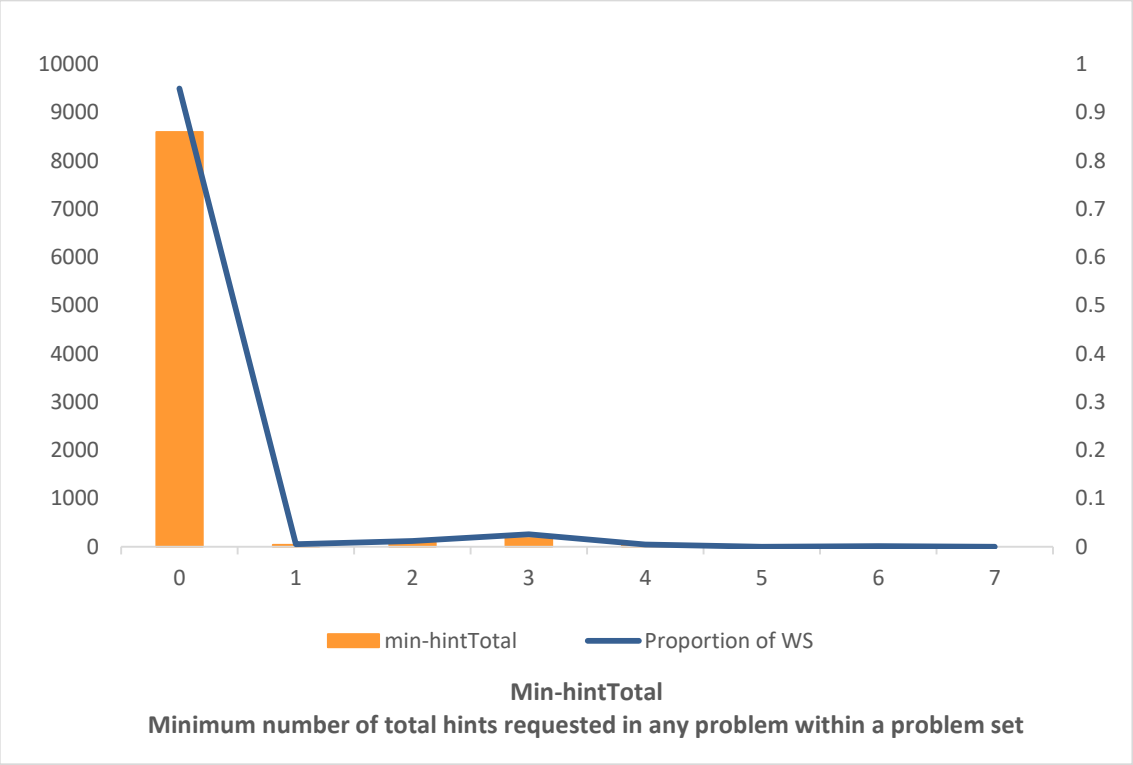

Figure 4: Figure showing a) a histogram of the minimum number of hints requested in any problem across student-problem set pairs (min-hintTotal), and b) the proportion of student-problem set pairs identified as wheel-spinning for each min-hintTotal range of values.

As shown in the histogram for max-past8BottomOut in Figure 5, the distribution for this

feature is relatively skewed to the right. Most students do not make any bottom-out hint requests in the last 8 problems within the sequence of 10 problems. The second most frequent number of bottom-out hint requests is 1, with 2 as the third most frequent. The proportion of students who were wheel-spinning across the values of bottom-out hints requested is represented with a line graph in the same figure (Figure 5). This line graph shows a steady decrease in the proportion of student wheel-spinning as the maximum number of bottom-out hints requested in the past 8 problems increases.

Figure 5: Figure showing a) a histogram of the maximum number of bottom-out hints requested in the last 8 problems across student-problem set pairs, and b) the proportion of student-problem set pairs identified as wheel-spinning for each max-past8BottomOut range of values.

From Figure 6, the most frequent range of standard deviation values for the amount of time between problems of the same skill is 0 – this means that the student started the next problem immediately after completing the previous problem in the system, for every case within the problem set. However, many student-problem set pairs have standard deviation values of 100,000 minutes or more (approximately 69 days or more). The line graph in Figure 6 thus shows the proportion of students defined as wheel-spinning to be highest at very low standard deviation values of the length of time spent between encountering problems in a problem set. This implies that students who have consistent short delays between problems in a problem set are more likely to be wheel-spinning.

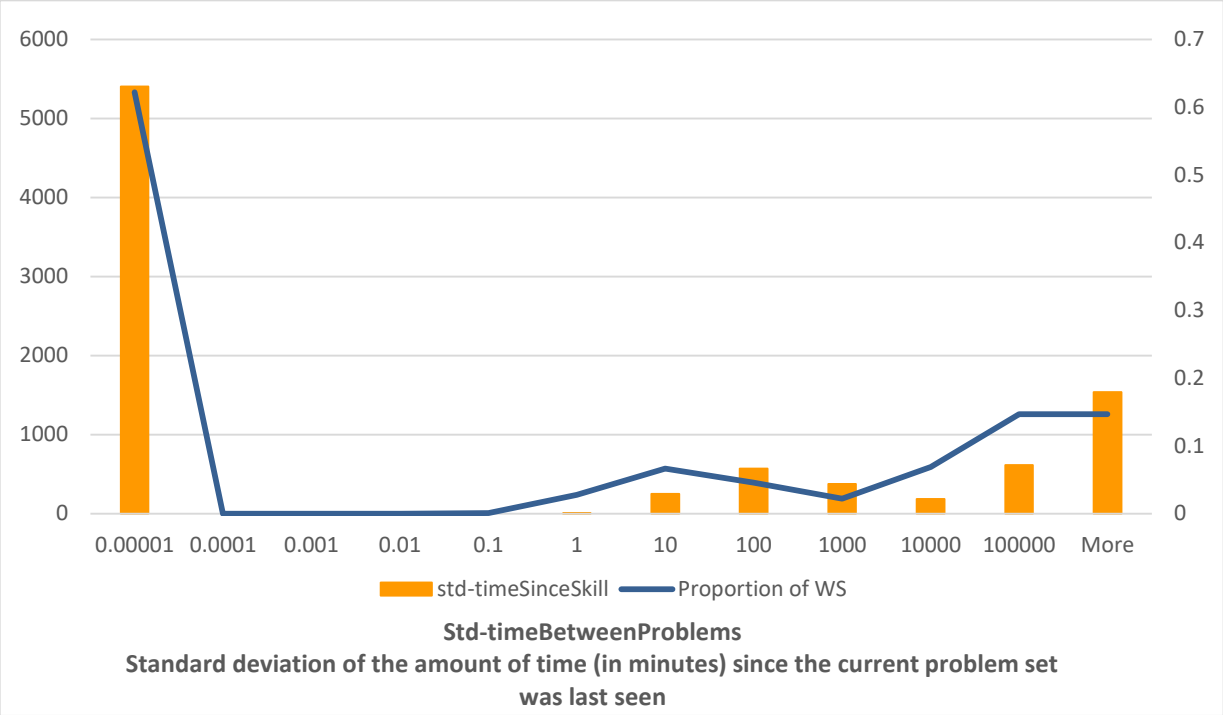

Figure 6: Graph showing a) a histogram of the standard deviation for the amount of time since the current set was last seen (std-timeBetweenProblems) across student problem set pairs, and b) the proportion of student-problem sets pairs identified as wheel-spinning for each std-timeBetweenProblems range of values.

To further explore the performance of these three features selected in the W-J48 decision tree model, we computed the performance of a prediction model using the W-J48 algorithm, but using only the top three selected features: min-hintTotal, max- past8BottomOut, and std-timeBetweenProblems. The J48 decision tree generated with this pared-down model contains 11 decision nodes (see Appendix E). Wheel-spinning is likely to occur when students do not request any hints in at least one problem, never request two or more bottom out hints across any problems, and there is less variation in the delay between solving problems within the same problem set. 76.49% of instances (5958 out of 7789 student-problem set pairs) with this feature combination were labeled as wheel-spinning.

However, the highest probability of wheel-spinning was associated with a branch that included a singleton node, max-past8BottomOut. Students with more bottom-out hints requested in the last 8 problems are more likely to wheel-spin. About 89.74% of instances (2781 out of 3099 student- problem set pairs) were included in this branch. These results match the findings from the 15-feature model, which indicate a nuanced relationship between wheel-spinning and the use of bottom-out hints.

A comparison of the performances between the 3-feature model (Cross-validated AUC ROC = 0.628) and the original 15-feature model (Cross-validated AUC ROC = 0.684) shows that the pared down model performed somewhat worse than the original one. This difference in goodness suggests that there are interactions present between variables in the 15-feature model that contribute to the model goodness, and suggests that wheel-spinning and productive persistence are better differentiated with a combination of the 15 selected features, compared to a combination of the top three selected features.

7. DISCUSSION

Persistence plays a significant role in learning, and it is important for teachers and students to understand when sustained effort is productive or unproductive. This is an increasingly pertinent issue to investigate in the context of online learning systems such as intelligent tutoring systems, where it has been argued that a large percentage of students engage in wheel-spinning or unproductive persistence (Beck & Gong, 2013). In this paper, we attempt to address this issue by creating models to differentiate between wheel-spinning and productive persistence in the ASSISTments tutoring platform, using educational data mining techniques.

In general, our findings indicate (unsurprisingly) that no single-feature model performed as well as our 15-feature model. This multi-feature model was reasonably effective in predicting whether or not a student will engage in wheel-spinning, achieving AUC ROC of 0.684. By comparison, the AUC performance of gaming detectors has been found to be around 0.80 (Baker, Corbett, Koedinger & Roll, 2006; Pardos et al., 2014), which is only slightly better than the current performance. Sensor-free affect detectors of student affective states in ASSISTments, found to be effective in several-year longitudinal prediction (San Pedro et al., 2013), have had AUC values ranging between from 0.63-0.74 (Pardos et al., 2014), slightly lower than the current value of our multi-feature model. AUC values in the 0.74-0.81 range are used in medical decision-making with real-world impact, such as the choice of which anti-retroviral therapy to use for HIV patients (Revell et al., 2013). As such, while there is considerable room for improvement in the models presented here, they are at a level of goodness where they can be used for basic research and intervention, given appropriate caution.

The findings presented here suggest that there is more predictive power in the combination of features, rather than each of the features alone. Upon further analysis of the J48 decision tree, two specific combinations of features were found to be associated with wheel-spinning, indicating two primary types of wheel-spinning students. First, students are likely to wheel-spin when they do not request any hints in at least one problem but request more bottom-out hints. However, a second group of wheel-spinning students request fewer bottom-out hints. These students are likely to wheel-spin when they have less variation in the amount of time spent between problems of the same skill.

Despite the slightly worse performance of the 3-feature model compared to our original 15 feature model, our analysis of the J48 decision tree suggests that each of the selected 3 features are particularly important for providing implications that can help guide intervention. First, we found that less variation in the number of days since the current problem set was last seen was associated with greater wheel-spinning. Specifically, students are more likely to wheel-spin if they have consistently shorter periods of time between each problem attempt. For instance, students are more likely to wheel-spin when they encounter less than 2-week delays across the first 10 problems. In contrast, wheel-spinning is less likely to occur if, for example, if a student has already attempted several problems but does not return to attempt more problems until after a month. Such differences in student behavior may have implications on the discussion whether massed practice is beneficial for students who are at risk of being unable to achieve mastery in a problem set. Previous research has shown the benefits of spacing in the domains of math problem solving (Rohrer & Taylor, 2006; Rohrer, Dedrick, & Burgess, 2014),

and second language acquisition (see review by Cepeda et al., 2006). Additionally, our finding about time between encountering problems also relates to empirical evidence showing that interleaved practice, where the opportunities to learn a different skill is spaced (abcbcacab), is more advantageous than blocked practice, where the opportunities to learn a given skill are massed (aaabbbccc). Specifically, Rohrer, Dedrick, and Burgess (2014) showed that students in the interleaved group significantly outperformed the students in the blocked group on a final test about volume, with math test performance of 63% for the interleaved group compared to 20% for the blocked group. As a whole, our findings along with prior literature suggest that practicing math problems in a spaced fashion can potentially help learners improve learning and reduce their wheel-spinning. By spacing work to a greater degree within ASSISTments (and related platforms), we may be able to reduce wheel-spinning.

Our findings also indicate that not requesting any hints in at least one problem was related to greater wheel-spinning. This result aligns with what previous literature has shown about the relationship between students' help avoidance and their learning. In particular, Aleven et al. (2006) found that avoiding the use of help was negatively correlated with post-test scores, while controlling for pre-test scores. As such, student who wheel-spin and avoid hints could potentially be struggling to realize they need support to learn the math content. This information may be useful in helping teachers identify students who may need additional support towards achieving content mastery and retention.

Finally, our results show that the relationship between the number of bottom-out hints requested and wheel-spinning is more nuanced than the other two features. On one hand, when at least one problem has no hint requests, heavy use of bottom-out hints in the last 8 problems leads to more wheel-spinning. Previous studies indicate that the incautious use of bottom-out hints is associated with gaming (Baker, Walonski, Heffernan, Roll, Corbett & Koedinger, 2008). As such, the type of student represented by this feature combination may be gaming the system. It is possible that students who game the system are less likely to read the intermediate hints (i.e., clues about how to solve the problem) because requesting bottom-out hints could provide them with the answer (Aleven & Koedinger, 2000). On the other hand, fewer bottom-out hint requests in the last 8 problems is associated with more wheel-spinning, when there are no hint requests in at least one problem and there is less variation in the length of time spent between attempting problems in the same problem set. It is possible that the type of student represented by this feature combination may not be checking the correct answers from the bottom-out hints to effectively learn the math content. It is hence worthy of teachers' time to pay attention to students' use of the bottom-out hints in Skill Builders problem sets, to try and identify possible patterns in students' use of this type of hints, as this type of behavior may be indicative of a student at-risk of wheel-spinning later on. Future work is also needed to further investigate how bottom-out hint requests relate to student wheel-spinning in various circumstances and conditions.

From our results, it is also worth noting that among the top three features found to predict student wheel-spinning, min-hintTotal and max-past8BottomOut are features specific to the student, while std-timeBetweenProblems may also be contingent on the teacher's choices around when to assign specific skill builders to students. As such, further analyses and randomized controlled trials (RCTs) may be conducted to investigate the specific conditions in which student wheel-spinning occurs. In general, it is important to note that each of the findings obtained here is correlational. Conducting small-scale randomized controlled trials (RCTs) based on our findings will have the benefit of helping to establish which of these findings is

causal at the same time as possibly enhancing student learning outcomes.

8. CONCLUSIONS AND FUTURE WORK

In this paper, we developed automated models that can differentiate between productive and unproductive persistence, in order to understand the differences between these two modes of engagement. As shown in our results, we found that a combination of features –the minimum number of hints requested in any problem, the maximum number of bottom-out hints requested in the last 8 problems, and the standard deviation of the amount of time since the current problem set was last seen by the student – distinguishes students who productively persist from those who wheel-spin.

These findings make a potentially important contribution in unbundling the concept of persistence, by encouraging students to exercise sustained effort where it's beneficial and supporting them when it is not. For instance, productively persistent students are likely to benefit from learning environments that promote tenacity. Determining which students are productively persisting could help indicate when research findings of how to enhance grit are most appropriate and helpful. Previous research suggests that developing students' growth mindset can be an effective strategy to promote grit (Laursen, 2015). As such, students who are already being productive when they are persistent -- but are not being persistent frequently enough -- may likely stick with challenging problems, if teachers emphasize and praise their efforts over ability.

While most studies have investigated the benefits of grit, less work has examined how to identify and prevent unproductive persistence. Our findings contribute to this gap in the literature, in helping identify students who unproductively persist, and distinguishing this behavior from productive persistence that should be encouraged. As indicated in our results, when students avoid hints in at least one problem but request more bottom-out hints in the last 8 problems, they are more likely to engage in unproductive persistence.

Our results also indicate that wheel-spinning is likely to occur with fewer bottom-out hint requests, when students avoid hints in at least one problem and study according to a massed schedule, with less than a 2-week delay between problem solving. These findings imply that studying math content in a more spaced fashion may likely reduce students' likelihood of wheel-spinning and scaffold them towards a productive path to learning. Additionally, the findings on the minimum number of hint requests in any problem and the number of bottom-out hint requests in the last 8 problems can provide actionable information on discriminating between unproductive and productive persistence, enabling teachers to identify which students are in most need of additional support. In this way, modeling wheel-spinning can help fine-tune interventions based on each student's needs, supporting learners who persist within and beyond tutoring systems.

While the findings in this study provide valuable insight into the student and program factors that are associated with unproductive persistence, several improvements may be made to our model in future work. For example, the thresholds used in our definition of wheel-spinning, despite extending beyond the amount of work completed to include indicators of eventual success at the skill, remain somewhat arbitrary. Although the cut-off of 10 problems maps to

current practice within the ASSISTments system, other thresholds may better capture student persistence. As such, we plan to examine the threshold of wheel-spinning further, by creating a range of possible cut-offs for persistence (e.g., 8 through 14 problems) and determining whether the predictors of wheel-spinning differ significantly across these thresholds. Given that our operationalization of persistence was largely based on the context of the ASSISTments platform, where 10 problems is a meaningful transition point (since the system asks the learner to take a break if 10 problems are completed without mastery), further research should investigate whether other cut-offs for persistence are relevant when distinguishing wheel-spinning from productive persistence in other platforms. In addition, there are other indicators of student success in the long-term that may be relevant (cf. Pardos et al., 2014). It may also be worth using probabilistic labels of some sort, instead of treating students who complete only 9 problems as non-persistent.

Relatedly, it is an open question whether the behavioral predictors of wheel-spinning and productive persistence differ across platforms. We may expect spaced practice to increase the effectiveness of persistence in other systems as well, based on the extensive literature on the benefits of spaced practice for retention. The heavy use of bottom-out hints has similarly been found to be detrimental to robust learning in other platforms. However, other factors may be expected to emerge as important as well. Future work in this area is critically important to better understand how persistence, wheel-spinning, and productive persistence should be defined and understood across learning platforms.

Further work is also needed to understand the differences and similarities of the constructs studied here to broader understanding of related constructs. To this end, we plan to correlate our predictions of productive persistence to other well-known measures of tenacity, such as grit. We also plan to correlate the incidence of students' wheel- spinning to their affective states (cf. Beck & Rodrigo, 2014). Further work in this area can help us determine the potential role of affect in encouraging persistence through adversity.

In investigating how these constructs relate to one another, we can better understand the different trajectories of student persistence and identify potential protective factors to support successful learning. Our findings represent a first step in this direction by differentiating between unproductive and productive persistence - towards developing research-based practice that enhances productive struggle and minimizes wheel-spinning.

9. ACKNOWLEDGEMENTS

We would like to thank Sergio Salmeron for his help in building the features used in this study, Seth Adjei for distilling data from the ASSISTments database for our use, as well as Nicole Shechtman and Mingyu Feng for helpful suggestions and feedback. We would also like to acknowledge NSF grant #DRL-1535340 for making this work possible.

10. REFERENCES

Aleven, V., & Koedinger, K. R. (2000, June). Limitations of student control: Do students know when they need help? In *International Conference on Intelligent Tutoring Systems* (pp.292-303). Springer Berlin Heidelberg.

Aleven, V. A., & Koedinger, K. R. (2002). An effective metacognitive strategy: Learning by doing and explaining with a computer-based Cognitive Tutor. *Cognitive Science*, *26*(2), 147-179.

Aleven, V., Mclaren, B., Roll, I., & Koedinger, K. (2006). Toward meta-cognitive tutoring: A model of help seeking with a Cognitive Tutor. *International Journal of Artificial Intelligence in Education*, *16*(2), 101-128.

Arroyo, I., Woolf, B.P., Royer, J.M., Tai, M., & English, S. (2010). Improving math learning through intelligent tutoring and basic skills training. In V. Alenven, J. Kay, & J. Mostow (Eds.), *Intelligent Tutoring Systems* (pp. 423–432). Berlin Heidelberg: Springer- Verlag.

Baker, R.S. (2008). A' implementation [Sourcecode]. http://www.upenn.edu/learning analytics/ryanbaker/computeAPrime.zip

Baker, R.S. (2015). Chapter 2: Model Goodness and Validation. *Big Data and Education*. 2nd Edition. New York, NY: Teachers College, Columbia University.

Baker, R.S.J.d., Corbett, A.T., Aleven, V. (2008) Improving Contextual Models of Guessing and Slipping with a Truncated Training Set. *Proceedings of the 1st International Conference on Educational Data Mining*, 67-76.

Baker, R. S., Corbett, A. T., Koedinger, K. R., & Roll, I. (2006, June). Generalizing detection of gaming the system across a tutoring curriculum. In *International Conference on Intelligent Tutoring Systems* (pp. 402-411). Springer Berlin Heidelberg.

Baker, R., Walonoski, J., Heffernan, N., Roll, I., Corbett, A., & Koedinger, K. (2008). Why students engage in" gaming the system" behavior in interactive learning environments. *Journal of Interactive Learning Research*, *19*(2), 185.

Baker, R. S., Goldstein, A. B., & Heffernan, N. T. (2011). Detecting learning moment- by-moment. *International Journal of Artificial Intelligence in Education*, *21*(1-2), 5- 25.

Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *JEDM-Journal of Educational Data Mining*,*1*(1), 3-17.

Baker, R.S.J.d., Gowda, S.M., & Corbett, A.T. (2011). Automatically detecting a student's preparation for future learning: Help use is key. In *Proceedings of the 4th International Conference on Educational Data Mining*, 179–188.

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*, 1–26.

Beck, J. E., & Gong, Y. (2013). Wheel-spinning: Students who fail to master a skill. In *International Conference on Artificial Intelligence in Education* (pp. 431-440). Springer Berlin Heidelberg.

Beck, J., & Rodrigo, M. M. T. (2014). Understanding wheel-spinning in the context of affective factors. In *International Conference on Intelligent Tutoring Systems* (pp. 162- 167). Springer International Publishing.

Borghans, L., Meijers, F. and B. ter Weel (2006). The Role of Noncognitive Skills in Explaining Cognitive Test Scores. *Economic Inquiry* 46 (1), 2-12.

Borghans, L., Duckworth, A. L., Heckman, J. J., & Ter Weel, B. (2008). The economics and psychology of personality traits. *Journal of Human Resources*, *43*(4), 972-1059.

Bowles, S., & Gintis, H. (1976). *Schooling in capitalist America*. New York: Basic Books.

Bowles, S., Gintis, H. and M. Osborne (2001). The Determinants of Earnings: A Behavioral Approach. *Journal of Economic Literature, 39 (4),* 1137-1176.

Brunello, G., & Schlotter, M. (2011). Non-cognitive skills and personality traits: Labour market relevance and their development in education & training systems. IZA Discussion Paper No 5743.

Carneiro, P., Crawford, C. and A. Goodman (2007). The Impact of Early Cognitive and Non-Cognitive Skills on Later Outcomes. CEE Discussion Paper 0092.

Cheung, A. C., & Slavin, R. E. (2013). The effectiveness of educational technology applications for enhancing mathematics achievement in K-12 classrooms: A meta- analysis. *Educational Research Review*, *9*, 88-113.

Chiteji, N. (2010). Time-preference, non-cognitive skills and well-being across the life course: Do non-cognitive skills encourage healthy behavior? *The American Economic Review*, *100*(2), 200.

Cloninger, C. R., Svrakic, D. M., & Przybeck, T. R. (1993). A psychobiological model of temperament and character. *Archives of General Psychiatry*, *50*(12), 975-990.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in

verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, *132*(3), 354.

Credé, M., Tynan, M. C., & Harms, P. D. (2016). Much Ado About Grit: A Meta-Analytic Synthesis of the Grit Literature. *Journal of Personality and Social Psychology*.

Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning* (pp. 233-240). ACM.

Dillon, J. T. (1988). Questioning and Teaching: A Manual of Practice, New York Teachers College.

De Fruyt, F., Van De Wiele, L., & Van Heeringen, C. (2000). Cloninger's psychobiological model of temperament and character and the five-factor model of personality. *Personality and Individual Differences*, *29*(3), 441-452.

Deke, J. and Haimson, J. (2006). Valuing Student Competencies: Which Ones Predict Postsecondary Educational Attainment and Earnings, and for Whom? Final Report. *Mathematica Policy Research, Inc.*

Digman, John M. (1990). Personality Structure: Emergence of the Five-Factor Model. *Annual review of psychology*, *41*(1), 417-440.Dillon, J. T. (1988). Questioning and Teaching: A Manual of Practice, New York Teachers College. *Annual Review of Psychology*, 41, 417-40.

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: perseverance and passion for long-term goals. *Journal of Personality and Social Psychology*, *92*(6), 1087.

Duckworth, A. L., & Quinn, P. D. (2009). Development and validation of the Short Grit Scale (GRIT–S). *Journal of Personality Assessment*, *91*(2), 166-174.

Eisenberger, R., & Leonard, J. M. (1980). Effects of conceptual task difficulty on generalized persistence. *American Journal of Psychology*, 93, 285–298.

Feng, M., Roschelle, J., Heffernan, N., Fairman, J., & Murphy, R. (2014). Implementation of an intelligent tutoring system for online homework support in an efficacy trial. In *International Conference on Intelligent Tutoring Systems* (pp. 561-566). Springer International Publishing.

Garcia, E. (2014). The Need to Address Noncognitive Skills in the Education Policy Agenda. Briefing Paper# 386. *Economic Policy Institute.*

Goldberg, L. R. (1990). An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology, 59,* 1216–1229.

Hanley, J.A. and McNeil, B.J. (1982) The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve. *Radiology, 143*, 29-36.

Heckman, J., Hsee, J. and Y. Rubinstein (2001). The GED is a Mixed Signal: the Effect of Cognitive and Noncognitive skills on Human Capital and Labor Market. Working Paper, University of Chicago.

Heffernan, N., Heffernan, C., Dietz, K., Soffer, D., Pellegrino, J. W., Goldman, S. R., & Dailey, M. (2012). Improving mathematical learning outcomes through automatic reassessment and relearning. In *Annual Meeting of American Educational Research Association,* Vancouver, British Columbia.

Heffernan, N. T., & Heffernan, C. L. (2014). The ASSISTments ecosystem: building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education, 24*(4), 470-497.

Huang, X., Craig, S. D., Xie, J., Graesser, A., & Hu, X. (2016). Intelligent tutoring systems work as a math gap reducer in 6th grade after-school program. *Learning and Individual Differences*, *47*, 258-265.

Kim, Y. J., Almond, R. G., & Shute, V. J. (2016). Applying Evidence-Centered Design for the Development of Game-Based Assessments in Physics Playground. *International Journal of Testing*, *16*(2), 142-163.

Klein, R., Spady, R., & Weiss, A. (1991). Factors affecting the output and quit propensities of production workers. *The Review of Economic Studies*, *58*(5), 929-953.

Laursen, E. K. (2015). The power of grit, perseverance, and tenacity. *Reclaiming Children and Youth*, *23*(4), 19.

John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research*, *2*(1999), 102- 138.

Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology*, *106*(4), 901.

MacCann, C., & Roberts, R. D. (2010). Do time management, grit, and self-control relate to academic achievement independently of conscientiousness? In R. Hicks (Ed.), *Personality and Individual Differences: Current Directions* (pp. 79–90). Queensland, Australia: Australian Academic Press.

McCrae, R. R., & Costa, P. T., Jr. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology, 52,* 81–90.

Matsuda, N., Chandrasekaran, S., & Stamper, J. (2016). How quickly can wheel-spinning be detected? In *Proceedings of the International Conference on Educational Data Mining (under review).*

Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., & Euler, T. (2006). Yale: Rapid prototyping for complex data mining tasks. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 935-940.

Montessori, M. (1912). The Discovery of the Child. India: Kalakshetra Publications.

Nelson-Le Gall, S. (1981). Help-seeking: An understudied problem-solving skill in children.

Developmental Review, 1(3), 224-246.

Nyhus, E. and Pons, E. (2005). The effect of personality on earnings. *Journal of Economic Psychology,* 26, 363–384.

Palaoag, T. D., Rodrigo, M. M. T., Andres, J. M. L., Andres, J. M. A. L., & Beck, J. E. (2016). Wheel-Spinning in a Game-Based Learning Environment for Physics. In *International Conference on Intelligent Tutoring Systems* (pp. 234-239). Springer International Publishing.

Pane, J. F., Griffin, B. A., McCaffrey, D. F., & Karam, R. (2013). Effectiveness of cognitive tutor algebra I at scale. *Educational Evaluation and Policy Analysis*, 36(2), 127- 144.

Pardos, Z.A., Baker, R.S., San Pedro, M.O.C.Z., Gowda, S.M., Gowda, S.M. (2014) Affective states and state tests: Investigating how affect and engagement during the school year predict end of year learning outcomes. *Journal of Learning Analytics, 1* (1), 107-128.

Paunonen, S. V., & Ashton, M. C. (2001). Big five predictors of academic achievement. *Journal of Research in Personality, 35*(1), 78–90.

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin, 135,* 322–338.

Prabhu, V., Sutton, C., & Sauser, W. (2008). Creativity and certain personality traits: Understanding the mediating effect of intrinsic motivation. *Creativity Research Journal*, *20*(1), 53-66.

Quinlan, J. R. (1993). C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann.

Rank, J., Pace, V. L., & Frese, M. (2004). Three avenues for future research on creativity, innovation, and initiative. *Applied Psychology: An International Review, 53,* 518–528.

Revell, A. D., Wang, D., Wood, R., Morrow, C., Tempelman, H., Hamers, R. L., Alvarez-Uria, G., Streinu-Cercel, A., Ene, L., Wensing, A.M.J., DeWolf, F., Nelson, M., Montaner, J.S., Lane, H.C., Larder, B.A. (2013). Computational models can predict response to HIV therapy without a genotype and may reduce treatment failure in different resource-limited settings. *Journal of Antimicrobial Chemotherapy, 68* (6), 1406-1414.

Ritter, S., Anderson, J. R., Koedinger, K. R., & Corbett, A. (2007). Cognitive Tutor: Applied research in mathematics education. *Psychonomic Bulletin & Review*, *14*(2), 249- 255.

Roberts, B. W., Chernyshenko, O. S., Stark, S., & Goldberg, L. R. (2005). The structure of conscientiousness: An empirical investigation based on seven major personality questionnaires. *Personnel Psychology*, *58*(1), 103-139.

Rohrer, D., Dedrick, R. F., & Burgess, K. (2014). The benefit of interleaved mathematics practice is not limited to superficially similar kinds of problems. *Psychonomic Bulletin & Review, 21*(5), 1323-1330.

Rohrer, D., & Taylor, K. (2007). The shuffling of mathematics problems improves learning. *Instructional Science*, *35*(6), 481-498.

Roschelle, J., Feng, M., Murphy, R. F., & Mason, C. A. (2016). Online mathematics homework increases student achievement. *AERA Open*, *2*(4).

Roschelle, J., Tatar, D., Schechtman,N., Hegedus, S., Hopkins, W., Knudson, J. & Stroter, A. (2007). Scaling up SimCalc project: Can a Technology Enhanced Curriculum Improve Student Learning of Important Mathematics? Menlo Park, CA: SRI international.

San Pedro, M.O.Z., Baker, R.S.J.d., Bowers, A.J., Heffernan, N.T. (2013) Predicting College Enrollment from Student Interaction with an Intelligent Tutoring System in Middle School. *Proceedings of the 6th International Conference on Educational Data Mining*, 177-184.

Sedek, G., & Kofta, M. (1990). When cognitive exertion does not yield cognitive gain: Toward an informational explanation of learned helplessness. *Journal of Personality and Social Psychology*, *58*(4), 729.

Settles, B., & Meeder, B. (2016). A Trainable Spaced Repetition Model for Language Learning. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 1848–1858.

Steenbergen-Hu, S., & Cooper, H. (2013). A meta-analysis of the effectiveness of intelligent tutoring systems on K–12 students' mathematical learning. *Journal of Educational Psychology*, *105*(4), 970.

Strayhorn, Terrell L. (2014). What role does grit play in the academic success of black male collegians at predominantly White institutions? *Journal of African American Studies*, 18(1), 1-10.

Tett, R. P., Jackson, D. N., & Rothstein, M. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology, 44*(4), 703–742.

Ventura, M., & Shute, V. (2013). The validity of a game-based assessment of persistence. *Computers in Human Behavior*, *29*(6), 2568-2572.

11. APPENDICES

Appendix A: List of core features before feature engineering

Core Features	Description
attemptCount	The number of attempts a student made within problem (including only answer, hint and scaffold actions).
endsWithAutoScaffolding	Problem ends with automatic scaffolding
endsWithScaffolding	Problem ends with scaffolding
IsHelpRequest	First response is a help request
isHelpRequestScaffolding	Whether or not the first response to a scaffolding problem is a help request
past5HelpRequest	Number of last 5 first responses that included a help request
past5WrongCount	Number of last 5 first responses that were wrong
past8HelpRequest	Number of last 8 first responses that included a help request
past8WrongCount	Cumulative count of the number of first responses to a problem that were wrong answers in the past 8 problems
timeTakenOnScaffolding	First response time taken on scaffolding problems
totalSkillOpportunitiesScaffolding	The total number of scaffolding problems divided by the unique problems the user has encountered relevant to the current problem set
workingInSchool	Whether or not the first response was made during school hours (between 7:00 am and 3:00 pm)
hintTotal	Total number of hints requested in any problem in the problem set
past8BottomOut	Number of bottom-out hints requested within the last 8 problems in a given problem set
responseIsChosen	Whether or not a problem requires the correct answer to be chosen from a list of answers (e.g., multiple choice)
responseIsFillIn	Response is filled in (No list of answers available)

stlHintUsed	Second to last hint is used – indicates a hint that gives considerable detail but is not quite bottom-out
timeBetweenProblems	Length of time since a problem involving this skill type was last seen
timeTaken	Time spent on the current step
totalAttempted	The total number of problems attempted in the tutor so far
totalPastWrongCount	Total first responses wrong attempts in the tutor so far.
totalPercentPastWrong	The percentage of all past problems that were incorrect on a given problem set
totalSkillOpportunities	The total number of unique problems the user has encountered relevant to the current problem set
totalSkillOpportunitiesByScaffolding	Total scaffolding opportunities for this problem set so far
totalTimeOnSkill	Total time spent on this problem set across all problems

Appendix B: Descriptive statistics of core features

Core feature	**Min**	**Max**	**Average**	**Standard deviation**
attemptCount	0	151	2.3178	2.1952
endsWithAutoScaffolding	0	1	0.0471	0.0778
endsWithScaffolding	0	1	0.0078	0.0223
IsHelpRequest	0	1	0.0192	0.0752
isHelpRequestScaffolding	0	1	0.0125	0.0226
past5HelpRequest	0	5	0.2414	0.2187
past5WrongCount	0	5	1.4902	1.0316
past8HelpRequest	0	8	0.2908	0.2351
past8WrongCount	0	8	1.7783	1.2986
timeTakenOnScaffolding (in seconds)	0	10000	4.7448	16.7044
totalSkillOpportunitiesScaffolding	0	9	0.2602	0.2896
workingInSchool	0	1	0.5610	0.1127
hintTotal	0	19	1.5291	0.9838
past8BottomOut	0	8	0.3433	0.4084
responseIsChosen	0	1	0.1623	0.0771
responseIsFillIn	0	1	0.2715	0.0719
stlHintUsed	0	1	0.0024	0.0067
timeBetweenProblems (in seconds)	0	3538408440	878281.1959	2393403.5228
timeTaken (in seconds)	0	10000	279.4849	1044.1036
totalAttempted	0	929	83.2187	2.7345
totalPastWrongCount	0	23	2.6548	1.5706
totalPercentPastWrong (in percent)	0	100	0.0728	0.1742
totalSkillOpportunities	1	10	4.9928	2.7359
totalSkillOpportunitiesByScaffolding	0	10	0.1867	0.2470
totalTimeOnSkill (in seconds)	0	102746	3301.1382	18410.8700

Appendix C: J48 decision tree of the 15-feature model

```
amin-hintTotal <= 0
|  amax-past8BottomOut <= 1
|  |  std-timeBetweenProblems <= 218115.7837
|  |  |  sum-responseIsChosen <= 26
|  |  |  |  amax-totalAttempted <= 18: 1 (898.0/172.0)
|  |  |  |  amax-totalAttempted > 18
|  |  |  |  |  amax-totalAttempted <= 491
|  |  |  |  |  |  sum-IsHelpRequestScaffolding <= 2
|  |  |  |  |  |  |  amax-totalSkillOpportunities <= 6
|  |  |  |  |  |  |  |  amax-totalSkillOpportunities <= 1
|  |  |  |  |  |  |  |  |  amax-responseIsChosen <= 0
|  |  |  |  |  |  |  |  |  |  sum-totalAttempted <= 126: 0 (7.0/2.0)
|  |  |  |  |  |  |  |  |  |  sum-totalAttempted > 126: 1 (5.0)
|  |  |  |  |  |  |  |  |  amax-responseIsChosen > 0
|  |  |  |  |  |  |  |  |  |  sum-responseIsChosen <= 4: 0 (20.0/6.0)
|  |  |  |  |  |  |  |  |  |  sum-responseIsChosen > 4: 1 (2.0)
|  |  |  |  |  |  |  |  amax-totalSkillOpportunities > 1
|  |  |  |  |  |  |  |  |  sum-Past8WrongCount <= 34
|  |  |  |  |  |  |  |  |  |  std-timeBetweenProblems <= 6.988398
|  |  |  |  |  |  |  |  |  |  |  amax-totalPercentPastWrong <= 0.9
|  |  |  |  |  |  |  |  |  |  |  |  sum-totalAttempted <= 422
|  |  |  |  |  |  |  |  |  |  |  |  |  sum-totalAttempted <= 209: 1 (2.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  sum-totalAttempted > 209: 0 (5.0)
|  |  |  |  |  |  |  |  |  |  |  |  sum-totalAttempted > 422: 1 (20.0)
|  |  |  |  |  |  |  |  |  |  |  amax-totalPercentPastWrong > 0.9: 0 (4.0)
|  |  |  |  |  |  |  |  |  |  std-timeBetweenProblems > 6.988398: 1 (161.0/12.0)
|  |  |  |  |  |  |  |  |  sum-Past8WrongCount > 34
|  |  |  |  |  |  |  |  |  |  std-totalSkillOpportunitiesByScaffolding <= 0
|  |  |  |  |  |  |  |  |  |  |  amax-totalPercentPastWrong <= 0.363636: 1 (5.0/1.0)
|  |  |  |  |  |  |  |  |  |  |  amax-totalPercentPastWrong > 0.363636
|  |  |  |  |  |  |  |  |  |  |  |  std-WorkingInSchool <= 0.278325: 1 (3.0/1.0)
|  |  |  |  |  |  |  |  |  |  |  |  std-WorkingInSchool > 0.278325: 0 (4.0)
|  |  |  |  |  |  |  |  |  |  std-totalSkillOpportunitiesByScaffolding > 0: 0 (3.0)
|  |  |  |  |  |  |  |  amax-totalSkillOpportunities > 6
|  |  |  |  |  |  |  |  |  amax-past8BottomOut <= 0
|  |  |  |  |  |  |  |  |  |  std-totalSkillOpportunitiesByScaffolding <= 2.18526
|  |  |  |  |  |  |  |  |  |  |  amax-responseIsChosen <= 0
|  |  |  |  |  |  |  |  |  |  |  |  std-totalSkillOpportunitiesByScaffolding <= 0.964126: 1 (992.0/344.0)
|  |  |  |  |  |  |  |  |  |  |  |  std-totalSkillOpportunitiesByScaffolding > 0.964126
|  |  |  |  |  |  |  |  |  |  |  |  |  std-timeBetweenProblems <= 23.786277: 0 (86.0/31.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  std-timeBetweenProblems > 23.786277
|  |  |  |  |  |  |  |  |  |  |  |  |  |  std-timeBetweenProblems <= 13108.93518: 1 (15.0/1.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  std-timeBetweenProblems > 13108.93518
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  sum-Past8WrongCount <= 30: 1 (3.0/1.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  sum-Past8WrongCount > 30: 0 (5.0/1.0)
|  |  |  |  |  |  |  |  |  |  |  amax-responseIsChosen > 0
|  |  |  |  |  |  |  |  |  |  |  |  amax-totalSkillOpportunities <= 9
|  |  |  |  |  |  |  |  |  |  |  |  |  amax-totalPercentPastWrong <= 0.9: 1 (26.0/3.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  amax-totalPercentPastWrong > 0.9
|  |  |  |  |  |  |  |  |  |  |  |  |  |  std-totalSkillOpportunitiesByScaffolding <= 1.022268
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  mean-totalAttempted <= 105.538462
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  mean-totalAttempted <= 67.8: 0 (2.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  mean-totalAttempted > 67.8: 1 (2.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  mean-totalAttempted > 105.538462: 0 (5.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  std-totalSkillOpportunitiesByScaffolding > 1.022268
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  sum-totalAttempted <= 756: 0 (4.0/1.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  sum-totalAttempted > 756: 1 (8.0)
|  |  |  |  |  |  |  |  |  |  |  |  amax-totalSkillOpportunities > 9: 1 (603.0/196.0)
|  |  |  |  |  |  |  |  |  |  std-totalSkillOpportunitiesByScaffolding > 2.18526
|  |  |  |  |  |  |  |  |  |  |  amax-totalPercentPastWrong <= 0.363636
|  |  |  |  |  |  |  |  |  |  |  |  sum-responseIsChosen <= 3: 1 (9.0/2.0)
|  |  |  |  |  |  |  |  |  |  |  |  sum-responseIsChosen > 3: 0 (2.0)
|  |  |  |  |  |  |  |  |  |  |  amax-totalPercentPastWrong > 0.363636: 0 (22.0/1.0)
|  |  |  |  |  |  |  |  amax-past8BottomOut > 0
|  |  |  |  |  |  |  |  |  amax-totalAttempted <= 121: 1 (1060.0/281.0)
|  |  |  |  |  |  |  |  |  amax-totalAttempted > 121
|  |  |  |  |  |  |  |  |  |  amax-totalAttempted <= 302
|  |  |  |  |  |  |  |  |  |  |  amax-responseIsChosen <= 0: 1 (207.0/86.0)
|  |  |  |  |  |  |  |  |  |  |  amax-responseIsChosen > 0
|  |  |  |  |  |  |  |  |  |  |  |  sum-Past8WrongCount <= 92
|  |  |  |  |  |  |  |  |  |  |  |  |  sum-responseIsChosen <= 10
|  |  |  |  |  |  |  |  |  |  |  |  |  |  std-WorkingInSchool <= 0.438019
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  sum-IsHelpRequestScaffolding <= 1
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  std-timeBetweenProblems <= 37505.97416
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  sum-Past8WrongCount <= 35
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  sum-Past8WrongCount <= 28
```

```
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  std-totaSkillOpportunitiesByScaffolding <= 0.943037: 0 (3.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  std-totalSkillOpportunitiesByScaffolding > 0.943037: 1 (3.0/1.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  sum-Past8WrongCount > 28: 1 (6.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  sum-Past8WrongCount > 35
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  sum-Past8WrongCount <= 56: 0 (9.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  sum-Past8WrongCount > 56
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  std-WorkingInSchool <= 0.158114: 1 (4.0/1.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  std-WorkingInSchool > 0.158114: 0 (3.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  std-timeBetweenProblems > 37505.97416: 1 (4.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  sum-IsHelpRequestScaffolding > 1
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  mean-totalAttempted <= 187.12766: 1 (2.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  mean-totalAttempted > 187.12766: 0 (3.0/1.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  std-WorkingInSchool > 0.438019: 1 (6.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  sum-responseIsChosen > 10: 1 (20.0/2.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  sum-Past8WrongCount > 92
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  std-timeBetweenProblems <= 28605.3143
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  amax-totalPercentPastWrong <= 0.96: 0 (6.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  amax-totalPercentPastWrong > 0.96
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  sum-responseIsChosen <= 13: 0 (6.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  sum-responseIsChosen > 13: 1 (2.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  std-timeBetweenProblems > 28605.3143: 1 (2.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  amax-totalAttempted > 302
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  std-WorkingInSchool <= 0.390205: 1 (30.0/2.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  std-WorkingInSchool > 0.390205
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  sum-Past8WrongCount <= 37: 0 (2.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  sum-Past8WrongCount > 37: 1 (3.0/1.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  sum-IsHelpRequestScaffolding > 2
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  amax-totalAttempted <= 307
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  amax-responseIsChosen <= 0
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  std-WorkingInSchool <= 0.162221: 1 (70.0/1.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  std-WorkingInSchool > 0.162221
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  sum-IsHelpRequestScaffolding <= 7: 1 (15.0/1.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  sum-IsHelpRequestScaffolding > 7
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  std-timeBetweenProblems <= 92.603468: 1 (4.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  std-timeBetweenProblems > 92.603468
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  sum-IsHelpRequestScaffolding <= 21: 0 (6.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  sum-IsHelpRequestScaffolding > 21: 1 (2.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  amax-responseIsChosen > 0: 1 (86.0/18.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  amax-totalAttempted > 307
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  sum-IsHelpRequestScaffolding <= 5: 1 (3.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  sum-IsHelpRequestScaffolding > 5
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  std-totalSkillOpportunitiesByScaffolding <= 1.542118: 0 (10.0/1.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  std-totalSkillOpportunitiesByScaffolding > 1.542118: 1 (3.0/1.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  amax-totalAttempted > 491: 1 (32.0)
|  |  |  |  |  |  |  sum-responseIsChosen > 26: 1 (675.0/90.0)
|  |  |  |  |  |  std-timeBetweenProblems > 218115.7837
|  |  |  |  |  |  |  std-timeBetweenProblems <= 12799929.2
|  |  |  |  |  |  |  |  amax-totalSkillOpportunities <= 5
|  |  |  |  |  |  |  |  |  std-timeBetweenProblems <= 1820838.272
|  |  |  |  |  |  |  |  |  |  std-totalSkillOpportunitiesByScaffolding <= 0
|  |  |  |  |  |  |  |  |  |  |  sum-frPast8WrongCount <= 48: 1 (76.0/9.0)
|  |  |  |  |  |  |  |  |  |  |  sum-frPast8WrongCount > 48: 0 (4.0/1.0)
|  |  |  |  |  |  |  |  |  |  std-totalSkillOpportunitiesByScaffolding > 0
|  |  |  |  |  |  |  |  |  |  |  std-timeBetweenProblems <= 684764.2974: 0 (3.0)
|  |  |  |  |  |  |  |  |  |  |  std-timeBetweenProblems > 684764.2974: 1 (3.0)
|  |  |  |  |  |  |  |  |  std-timeBetweenProblems > 1820838.272
|  |  |  |  |  |  |  |  |  |  sum-Past8WrongCount <= 19
|  |  |  |  |  |  |  |  |  |  |  std-WorkingInSchool <= 0.357935
|  |  |  |  |  |  |  |  |  |  |  |  std-timeBetweenProblems <= 3294065.607: 0 (2.0)
|  |  |  |  |  |  |  |  |  |  |  |  std-timeBetweenProblems > 3294065.607: 1 (3.0)
|  |  |  |  |  |  |  |  |  |  |  std-WorkingInSchool > 0.357935: 0 (7.0)
|  |  |  |  |  |  |  |  |  |  sum-Past8WrongCount > 19: 1 (4.0)
|  |  |  |  |  |  |  |  amax-totalSkillOpportunities > 5
|  |  |  |  |  |  |  |  |  amax-past8BottomOut <= 0
|  |  |  |  |  |  |  |  |  |  std-timeBetweenProblems <= 2773779.763
|  |  |  |  |  |  |  |  |  |  |  amax-responseIsChosen <= 0: 0 (228.0/107.0)
|  |  |  |  |  |  |  |  |  |  |  amax-responseIsChosen > 0
|  |  |  |  |  |  |  |  |  |  |  |  amax-totalPercentPastWrong <= 0.363636
|  |  |  |  |  |  |  |  |  |  |  |  |  sum-responseIsChosen <= 23
|  |  |  |  |  |  |  |  |  |  |  |  |  |  std-WorkingInSchool <= 0.347839
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  mean-totalAttempted <= 47.181818: 0 (2.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  mean-totalAttempted > 47.181818: 1 (5.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  std-WorkingInSchool > 0.347839: 0 (2.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  sum-responseIsChosen > 23: 1 (10.0)
|  |  |  |  |  |  |  |  |  |  |  |  amax-totalPercentPastWrong > 0.363636: 0 (85.0/41.0)
|  |  |  |  |  |  |  |  |  |  std-timeBetweenProblems > 2773779.763
|  |  |  |  |  |  |  |  |  |  |  amax-totalAttempted <= 502
|  |  |  |  |  |  |  |  |  |  |  |  std-timeBetweenProblems <= 8717502.941: 0 (65.0/11.0)
|  |  |  |  |  |  |  |  |  |  |  |  std-timeBetweenProblems > 8717502.941: 1 (3.0)
```

```
|  |  |  |  |  |  |  |  amax-totalAttempted > 502: 1 (6.0/1.0)
|  |  |  |  |  amax-past8BottomOut > 0
|  |  |  |  |  |  amax-totalSkillOpportunities <= 8: 1 (20.0/4.0)
|  |  |  |  |  |  amax-totalSkillOpportunities > 8
|  |  |  |  |  |  |  sum-IsHelpRequestScaffolding <= 0
|  |  |  |  |  |  |  |  amax-responseIsChosen <= 0
|  |  |  |  |  |  |  |  |  amax-totalSkillOpportunities <= 9
|  |  |  |  |  |  |  |  |  |  std-WorkingInSchool <= 0.142857: 0 (5.0/1.0)
|  |  |  |  |  |  |  |  |  |  std-WorkingInSchool > 0.142857: 1 (8.0/1.0)
|  |  |  |  |  |  |  |  |  amax-totalSkillOpportunities > 9: 1 (191.0/85.0)
|  |  |  |  |  |  |  |  amax-responseIsChosen > 0
|  |  |  |  |  |  |  |  |  std-timeBetweenProblems <= 1505822.527
|  |  |  |  |  |  |  |  |  |  amax-totalPercentPastWrong <= 0.888889: 1 (8.0)
|  |  |  |  |  |  |  |  |  |  amax-totalPercentPastWrong > 0.888889
|  |  |  |  |  |  |  |  |  |  |  amax-totalPercentPastWrong <= 0.96: 0 (2.0)
|  |  |  |  |  |  |  |  |  |  |  amax-totalPercentPastWrong > 0.96
|  |  |  |  |  |  |  |  |  |  |  |  amax-totalSkillOpportunities <= 9: 0 (3.0/1.0)
|  |  |  |  |  |  |  |  |  |  |  amax-totalSkillOpportunities > 9
|  |  |  |  |  |  |  |  |  |  |  |  sum-totalAttempted <= 2845
|  |  |  |  |  |  |  |  |  |  |  |  |  std-totalSkillOpportunitiesByScaffolding <= 0.57735
|  |  |  |  |  |  |  |  |  |  |  |  |  |  std-WorkingInSchool <= 0.156174: 0 (5.0/1.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  std-WorkingInSchool > 0.156174: 1 (5.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  std-totalSkillOpportunitiesByScaffolding > 0.57735: 0 (2.0)
|  |  |  |  |  |  |  |  |  |  |  |  sum-totalAttempted > 2845: 1 (10.0)
|  |  |  |  |  |  |  |  |  std-timeBetweenProblems > 1505822.527
|  |  |  |  |  |  |  |  |  |  amax-totalAttempted <= 17: 1 (2.0)
|  |  |  |  |  |  |  |  |  |  amax-totalAttempted > 17: 0 (25.0/4.0)
|  |  |  |  |  |  |  sum-IsHelpRequestScaffolding > 0
|  |  |  |  |  |  |  |  sum-Past8WrongCount <= 28: 0 (5.0/1.0)
|  |  |  |  |  |  |  |  sum-Past8WrongCount > 28: 1 (10.0)
|  |  |  std-timeBetweenProblems > 12799929.2: 1 (30.0/1.0)
|  amax-past8BottomOut > 1: 1 (2544.0/312.0)
amin-hintTotal > 0: 1 (355.0/12.0)

Number of Leaves:  95  Size of the tree:  189
```

Appendix D: Features selected with J48 decision tree algorithm

Features	Description
amin-hintTotal	(Minimum) Number of hints requested in any problem in the problem set
amax-past8BottomOut	(Maximum) Number of bottom-out hints requested within the last 8 problems in a given problem set
amax-responseIsChosen	(Maximum) Whether or not a problem requires the correct answer to be chosen from a list of answers (e.g., multiple choice)
amax-totalAttempted	(Maximum) The total number of problems attempted in the tutor so far
amax- totalPercentPastWrong	(Maximum) The percentage of all past problems that were incorrect on a given problem set
amax- totalSkillOpportunities	(Maximum) The total number of unique problems the user has encountered relevant to the current problem set
mean-totalAttempted	(Mean) The total number of problems attempted in the tutor so far
std- IsHelpRequestScaffolding	(Standard deviation) Whether or not the first response to a scaffolding problem is a help request
std-workingInSchool	(Standard deviation) Whether or not the first response was made during or after school hours (between 7:00 a.m. and 3:00 p.m.)
std-timeBetweenProblems	(Standard Deviation) Length of time since a problem involving this skill type was last seen
std- totalSkillOpportunitiesByScaffolding	(Standard Deviation) The total number of scaffolding problems divided by the unique problems the user has encountered relevant to the current problem set
sum- isHelpRequestScaffolding	(Sum) Whether or not the first response to a scaffolding problem is a help request
sum-past8WrongCount	(Sum) Cumulative count of the number of first responses to a problem that were wrong answers in the past 8 problems
sum-responseIsChosen	(Sum) Whether or not a problem requires the correct answer to be chosen from a list of answers (e.g., multiple choice)
sum-totalAttempted	(Standard Deviation) The total number of problems attempted in the tutor so far

Appendix E: J48 decision tree of the 3-feature model

```
amax-past8BottomOut = 1
|   amin-hintTotal = 1
|   |   std-timeBetweenProblems = 218115.7837: 1 (5224.0/1393.0)
|   |   std-timeBetweenProblems > 218115.7837
|   |   |   std-timeBetweenProblems = 12799929.2
|   |   |   |   std-timeBetweenProblems = 1820838.272: 1 (645.0/269.0)
|   |   |   |   std-timeBetweenProblems > 1820838.272: 0 (168.0/59.0)
|   |   |   std-timeBetweenProblems > 12799929.2: 1 (30.0/1.0)
|   amin-hintTotal > 1: 1 (100.0/3.0)
amax-past8BottomOut > 1: 1 (2781.0/318.0)

Number of Leaves  : 6  Size of the tree :    11
```