Enhancing Student outcomes through predictive modeling and recommender systems

Monish Erode Sridhar
North Carolina State University
Raleigh, North Carolina
merodes@ncsu.edu

Sanjaey Shunmuga Sundaram
North Carolina State University
Raleigh, North Carolina
sshunmu@ncsu.edu

Ayush Gupta
North Carolina State University
Raleigh, North Carolina
agupta86@ncsu.edu

Dharani Guda
North Carolina State University
Raleigh, North Carolina
dguda@ncsu.edu

Abstract

Education platforms have become a source of big data to enhance student learning experiences.They include variable skill sets and concepts that differ in multiple aspects and leading to inconsistent performance.This project aims at resolving these issues with the use of a predictive model based on machine learning and a content-based recommender system that assists students. A predictive model is an intelligent measurement based on attempts, hints requested, and time taken that assesses whether a student will correctly answer a question, while the recommender system provides personalized practice exercises with which students can get help in improving weak skills.This project uses the ASSISTments Skill Builder dataset, applying a number of machine learning models, including a Random Forest , SVM, Linear regression, ANN for predicting if a student is giving correct answer and TF-IDF vectorization for the recommender system.Evaluation of system effectiveness is performed using quantitative metrics and simulated user feedback.Results show that combining prediction-based and personalized recommendation-based approaches could tremendously enhance student learning performance, in part by giving teachers valuable insight into areas where students struggle.

Keywords

Online learning platforms, Machine Learning, Random Forest Classification, Predictive Analysis, Feature Engineering, Student Performance Prediction, Exploratory Data Analysis, Online learning Analytics, Imbalanced Dataset Handling, GridSearch, Decision Tree Classificaiton, ANN Classification, Recommender System

ACM Reference Format:
Monish Erode Sridhar, Sanjaey Shunmuga Sundaram, Ayush Gupta, and Dharani Guda. 2024. Enhancing Student outcomes through predictive modeling and recommender systems . In . ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Online learning platforms have become increasingly popular over the past few years, giving a huge amount of data for understanding and improving student learning experiences. Even with this amount of material at their disposal, students frequently struggle with particular concepts and abilities, which results in inconsistent performance. Improving student outcomes requires early detection of difficulties and more individualized treatments. Such early identification of struggles and further personalized interventions are key in enhancing the outcomes of students.The aim of the project is to address these challenges by the development of a machine learning based predictive model and a content based recommender system that can support students during their learning. The major objective is to forecast the probability that a student will properly answer a given question based on the number of attempts, time spent, hints asked, and other contextual factors.Additionally, the project advises students on areas that require focus based on their performance.Therefore, to enhance overall learning results, predictive modeling and recommendation strategies can be used in tandem.

1.1 Related Works

Different methods have been studied to predict if a student can answer a question. One of such methods [1] uses logistic regression, decision trees and neural networks to predict if a student can answer the question using the Assistments datasets. Although the study predicts the performance of the students it lacks in giving rightful advice to them in order to improve their performance. Thus it lacks any future enhancements on personalized interventions like recommender systems.

One of the studies [2] explores feature engineering for predicting the performance of MOOC (Massive Open Online Courses) students, utilizing the Deep Feature Synthesis (DFS) method to automate the creation of features. It focuses on constructing base and time-indexed features to generate new features, which are shown to be effective, as evidenced by Principal Component Analysis (PCA) selecting them as the best features. The study also aims to reduce the workload of manual data pre-processing by supporting automated feature engineering, with plans to test the scalability of the approach on larger datasets.

Recommender System discussed in [3] presents a state-of-the-art analysis of these systems that can be used in educational applications and further compares different techniques, like hybrid recommendation systems and experimental evaluations. It also emphasizes how the ability of personalized content recommendations is particularly relevant for students and teaching staff which connects to the project purpose of developing a content-based recommender system to help struggling students.

Yağcı, M in 2021 [4] uses machine learning models to predict student performance based on attributes such as department, professor, and midterm exam data. It proves the importance of predictive modeling in improving decision-making in educational environments.This project furthers this research by going beyond the mere prediction of student outcomes to actively providing recommendations to improve those outcomes.

2 Methodology

This project encompasses the integration of predictive modeling with a personalized recommendation system which addresses two of the main objectives: predicting if a student will be able to solve a particular problem and then recommending that particular student a set of problems they can solve in order to improve their efficacy. Key novelty features include:

- Prediction alongside recommendation system: Unlike prior work that predicts weak skills and associates them to treatment, which addresses the gap between performance prediction and intervention.
- Feature selection: Random forest was implemented in order to get the most relevant features.
- Scalable Data Processing: The system processes 6.1 million records and thus can be effectively applied in realistic educational scenarios.
- End-to-end Pipeline: The pipes ranging from preprocessing right to the prediction and recommendation make it easy to obtain a one-stop solution for increasing the probability of student learning outcomes.

2.1 Data Preprocessing

In order to increase the data quality, several preprocessing steps were used. The data were then cleaned for missing values, which further completed using the median values for numerical value parameters (e.g., time_taken) and the mode values for categorical value parameters (e.g., first_action). Numerical features were normalized using StandardScaler since it scales their distribution to make them near normal distribution while categorical data were encoded using label encoder. Stratified sampling was used to split the dataset in order to minimize the bias by ensuring that the ratio of the classes in the segmented training and testing sets was similar to the proportions in the original data set.

2.2 Feature Selection

In the feature selection phase, Random Forest emerged as the most effective model for identifying the most pertinent predictors of student performance due to its robust capability to rank features based on their importance.Examining the feature importance it was observed that variables as attempt_count, hint_count and time_taken

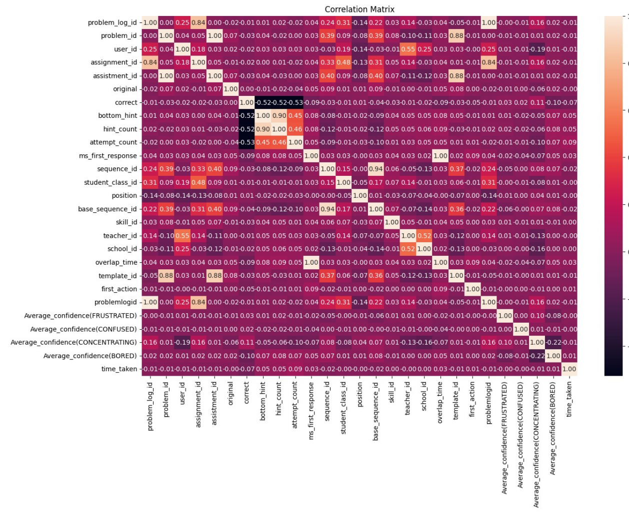

Figure 1: Heatmap showing the correlation between the features

were the most important predictors. This step increases the interpretability of the Feature Importance algorithm and removes the probe features used in the training.

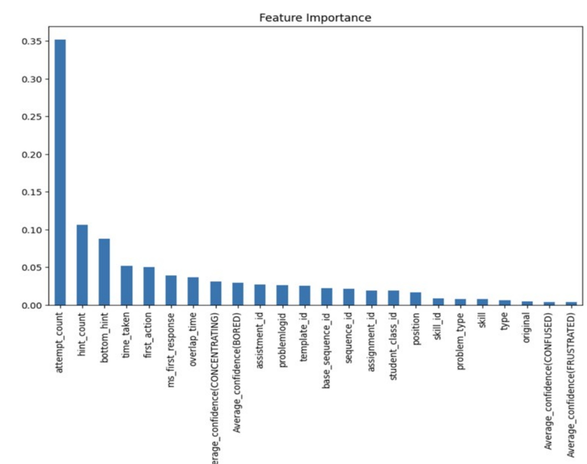

Figure 2: Relevancy of features

Studies have shown that integrating features like time spent, hint usage, and attempts significantly improves the predictive power of models, highlighting the importance of behavioral metrics in understanding student performancening in Educational Data Mining** "Deep learning approaches, such as those utilizing Transformer-based frameworks, have demonstrated the ability to capture complex interactions between student behavior and learning outcomes, enhancing both accuracy and interpretability of predictive models [5]

2.3 Machine learning Models

We tested multiple ML models in order to get the most efficacy. These models include Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Decision Trees, Random Forest, and Artificial Neural Networks (ANN). Our choice of hyperparameters was achieved using grid search which included the number of neighbors in KNN, kernel types in SVM, and tree depth in Decision Trees. For the purpose of increasing the reliability of the results 3-fold cross-validation was used and for comparing the models the set of metrics including accuracy, precision, recall and F1 score was employed.

2.4 Recommender System

The recommendation system concentrates on students' weak skills according to their performance as recognized by the recommender system. TF-IDF vectorization is performed to derive the importance of a given group of terms that is here skill-related, from the dataset and cosine similarity to match these, often so weak, skills, with listed, available exercises. It then presents a student specific learning needs from the knowledge face map then provides a set of learning recommendations to the student. This makes the recommendations as accurate as they are actionable for implementation in the supply chain.

2.5 Rationale

The choice of the methods was dictated by the fact that the dataset typically presents with educational related features. Random Forest was chosen for performing feature selection for its interpretability and efficiency in dealing with large features variables. Thus, TF-IDF and cosine similarity were used in the recommender system so that students will only be suggested resources that will be most relevant to their academic area of interest. Bearing in mind the preprocessing, feature selection, machine learning and recommendation strategies, this work successfully provides an overall solution to improve the student's results.

3 Plan and Experiment

3.1 Dataset

The ASSISTments dataset for the academic year 2012–2013 serves as the basis for this study. It comprises over 10 million rows, each representing an interaction between a student and the platform. The primary goal is to predict whether a student will correctly answer a question, represented by the binary target variable correct (1 for success, 0 for failure). The dataset includes a variety of features that capture student behavior, problem characteristics, and affective states, offering rich contextual information for predictive modeling. Key features include behavioral metrics such as time_taken, the total time spent solving a problem, and attempt_count, the number of attempts made. Additional features like hint_count and ms_first_response offer insights into how students interact with hints and respond initially. Categorical attributes such as skill_id, which links problems to specific skills, and problem_type, which defines the format of a question, provide further context. Emotional state predictions—such as levels of frustration, concentration, and boredom—are included as derived attributes and add a unique dimension to understanding student behavior. Data Preprocessing: Handling such a large dataset required strategic preprocessing:

- Missing Values: Numerical features were imputed using the median, while categorical features were imputed using the mode or clustering methods.
- Sampling: Stratified sampling reduced the dataset to 6.1 million rows, ensuring the target variable's distribution remained balanced.
- Scaling and Encoding: Numerical features were scaled using StandardScaler, and categorical variables, such as skill_id, were encoded using LabelEncoder to prepare the data for machine learning models.

The dataset was then split into training (10%) and testing (90%) subsets to ensure a large evaluation set, crucial for generalizing model performance.

3.2 Hypothesis

This study investigates three key hypotheses:

- Behavioral features such as time_taken, hint_rate, and attempt_count significantly enhance the ability to predict student performance. By quantifying how students allocate time and interact with hints, these features are expected to serve as strong predictors for binary outcomes.
- Emotional state features, including frustration, concentration, and boredom, can add interpretability and contextual relevance to predictive models. While their direct predictive contribution may be modest, they are hypothesized to improve the understanding of underlying behavioral patterns.
- Random Forest will outperform both simpler models like Logistic Regression and more complex models like Artificial Neural Networks (ANN) in terms of accuracy, computational efficiency, and interpretability. This hypothesis reflects the balance Random Forest offers between performance and explainability.

These hypotheses aim to address fundamental questions about the relative importance of features, the role of emotional context, and the trade-offs between different modeling approaches.

3.3 Experimental Design

To test these hypotheses, a series of experiments were conducted, focusing on feature contributions, model performance, and the role of emotional states.

3.3.1 Feature Contribution Analysis. The contribution of each feature to prediction accuracy was assessed using Random Forest's built-in feature importance metric. The following steps were performed:

- A Random Forest model was trained on the full dataset, and feature importance scores were extracted.
- Ablation tests were conducted by sequentially removing features like time_taken and hint_rate to evaluate their impact on accuracy.

This experiment quantified the relative importance of features, addressing potential correlations by measuring performance changes when key features were excluded.

Table 1: Hyperparameters for Various Machine Learning Models

Model	Hyperparameter	Values
Linear Regression	None	None
KNN	`n_neighbors`	3, 5, 7, 9
	`weights`	`uniform`, `distance`
SVM	`C`	0.1, 1, 10
	`kernel`	`linear`, `rbf`, `poly`
Random Forest	`n_estimators`	50, 100, 200
	`max_depth`	None, 10, 20, 30, 50
Decision Tree	`max_depth`	None, 5, 10, 15
	`min_samples_split`	2, 5, 10
ANN	`model_optimizer`	adam, sgd
	`model_activation`	relu, tanh
	`model_neurons`	16, 32, 64
	`model_dropout_rate`	0.2, 0.3
	`model_batch_size`	32, 64
	`model_epochs`	10, 20

3.3.2 Model Comparison. Several machine learning algorithms were evaluated to identify the best-performing model: Logistic Regression, Random Forest, KNN, SVM, Decision Tree, ANN. Hyperparameter Tuning: Each model was optimized using grid search. For example as given in Table 1

3.3.3 Evaluation Metrics. Accuracy, F1-score, precision, and recall were used to assess model performance. Cross-validation with three folds ensured that model evaluation was robust. This approach addressed overfitting risks and provided a comprehensive understanding of each model's capabilities.

3.4 Challenges and Mitigation

Handling the imbalanced nature of the target variable was a significant challenge. Stratified sampling ensured balanced representation during data splitting, while SMOTE was applied during training to address class imbalance further. Additionally, the large dataset size required efficient sampling and preprocessing to make experiments computationally feasible.

4 Results

The project produced significant findings in both predictive modeling and personalized recommendation systems. For the predictive task, the Random Forest model emerged as the most effective classifier in predicting whether students could answer specific questions correctly. Features such as the number of attempts (attempt_count), hints requested (hint_count), time taken (time_taken), and first actions (first_action) showed strong correlations with the target variable. These results were validated using confusion matrices, which highlighted the model's performance in terms of accuracy and error rates. The choice of features and preprocessing steps, including stratified sampling and standard scaling, proved crucial for achieving reliable predictions. In parallel, the recommender system demonstrated its ability to personalize learning by identifying weak skills of students. By analyzing incorrect responses, the system used TF-IDF vectorization and cosine similarity to match students' weaknesses with exercises that could help improve their skills. This novel integration of predictive modeling with recommendation strategies highlights the potential for more personalized educational tools. When compared to prior work, the project built on existing methods for feature engineering and prediction models, such as those discussed in "Feature Engineering for Predicting MOOC Performance." However, it extended these approaches by addressing a critical gap: the lack of personalized intervention systems. Unlike earlier studies that focused exclusively on predictive accuracy, this work successfully combined prediction with actionable recommendations, providing a comprehensive solution to improve student learning outcomes.

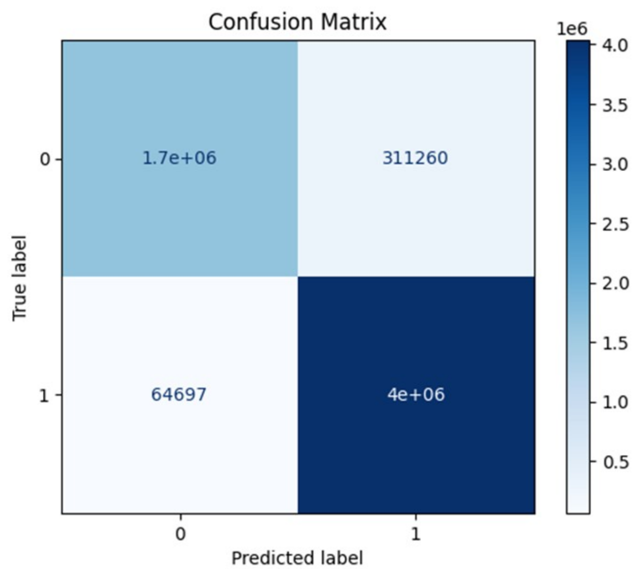

Figure 3: Confusion matrix of Random Forest Classifier with hyper-parameters `max_depth: 20` and `n_estimators: 200`, achieving an F1-Score of 0.94.

5 Conclusion

The project presented a successful implementation of machine learning in the educational domain, addressing both prediction and personalization. The predictive model demonstrated its efficacy in identifying whether students could correctly answer questions

	problem_id	skill_name
2910817	338844	Algebra
1225515	523501	Algebra
904483	338858	Algebra
1755654	338871	Algebra
5553767	540837	Algebra

Figure 4: Figure 4: Questions Recommended by our Recommender system for a student who was weak in Algebra. Our Recommender system returned 5 questions tagged with skill algebra for the student to improve this skill

based on specific features, offering insights into factors influencing student performance. At the same time, the recommender system bridged the gap between performance analysis and actionable feedback, ensuring that students received tailored recommendations to improve their skills. This dual approach marks a significant advancement in the use of automated learning tools for education. By focusing on both prediction and intervention, the project offers a framework that can enhance student outcomes on a large scale. Future iterations of the study could incorporate more advanced techniques, such as deep learning models, to further improve prediction accuracy and the quality of recommendations. Additionally, expanding the recommender system to include multimedia resources and adaptive learning pathways could enrich the student experience and provide even more targeted support.

6 Meeting Attendance

Over the past two weeks, the group conducted five meetings to coordinate efforts and progress. All members were actively involved throughout this period.

Monish Erode Sridhar consistently contributed to developing the machine learning model, ensuring that the preprocessing and feature selection stages aligned with the project’s goals.

Sanjaey Shunmuga Sundaram focused on the preprocessing and feature engineering tasks, preparing the data for effective model training.

Ayush Gupta played a key role in designing and implementing the recommender system, leveraging algorithms to personalize student feedback.

Dharani Guda contributed through data exploration and visualization, ensuring that the project’s results were well-represented and interpretable.

The Table 2 shows number meeting have been conducted and when the team members meet during these meetings.

Table 2: Meeting Attendance Record

Member Name	**1**	**2**	**3**	**4**	**5**
Monish Erode Sridhar	P	P	P	P	P
Sanjaey Shunmuga Sundaram	P	P	P	P	P
Ayush Gupta	P	P	P	P	P
Dharani Guda	P	P	P	P	P

References

1. Bresciani Ludvik, M.J., Bresciani Ludvik, M.J. (2019). Outcomes-Based Program Review: Closing Achievement Gaps In- and Outside the Classroom With Alignment to Predictive Analytics and Performance Metrics (2nd ed.). Routledge. https://doi.org/10.4324/9781003446231/

2. Mohamad, Nadirah Ahmad, N. Jawawi, Dayang Mohd Hashim, Siti. (2020). Feature Engineering for Predicting MOOC Performance. IOP Conference Series: Materials Science and Engineering. 884. 012070. 10.1088/1757-899X/884/1/012070.

3. A REVIEW OF EDUCATIONAL RECOMMENDER SYSTEMS FOR TEACHERS Mbarek Dhahri1 and Mohamed Koutheair Khribi Higher School of Sciences and Technologies of Tunis, University of Tunis, Tunisia Technologies of Information and Communication Lab

4. Yağcı, M. (n.d.). Educational Data Mining: Prediction of students’ academic performance using machine learning algorithms. Smart Learning Environments, 9(1). https://doi.org/10.1186/s40561-022-00192-z

5. Chun-kit Yeung, Zizheng Lin, Kai Y Incorporating Features Learned by an Enhanced Deep Knowledge Tracing Model for STEM/Non-STEM Job Prediction.