The Future of Engineering Education
2024 Annual Conference & Exposition
Oregon Convention Center
Portland, OR . June 23 - 26, 2024
ASEE
Paper ID #43609

Predicting Student Performance Using Discussion Forums' Participation Data

Mac Joseph Gray, Duke University

Mac Gray is currently a second-year Master of Science student in Electrical and Computer Engineering at Duke University. With an interest in the intersection of machine learning and software engineering. Mac is specifically passionate about advancing natural language processing (NLP) technologies.

Dr. Rabih Younes, Duke University

Rabih Younes is an Assistant Professor of the Practice in the Department of Electrical and Computer Engineering at Duke University. He received his PhD in Computer Engineering from Virginia Tech in 2018 after receiving his BE and MSE in Computer Engineering from the Lebanese American University in 2011 and 2013, respectively. Rabih speaks nine languages (fluent in three) and holds a number of certificates in education, networking, IT, and skydiving. He is also a member of several honor societies, including Tau Beta Pi, Eta Kappa Nu, Phi Kappa Phi, and Golden Key. Rabih has a passion for both teaching and research; he has been teaching since he was a teenager, and his research interests include wearable computing, activity recognition, context awareness, machine learning, engineering education, and Middle Eastern politics. As a professor, Rabih is committed to helping his students achieve their goals and providing them with opportunities to realize that. He also focuses on their personal development and on improving their abilities to be critical thinkers, better communicators, and active members of their community and the world. More information can be found on his personal website: www.rabihyounes.com.

Predicting Student Performance Using Discussion Forums' Participation Data

Abstract

A significant gap in education lies in the need for mechanisms that enable early detection of potentially at-risk students. Through access to an earlier prediction of student performance, instructors are given ample time to meet with and assist under-achieving students. As with any prediction modeling problem, there are many predictors to choose from when formulating a model. Previous related works have shown limited success in predicting course performance using students' personal and socioeconomic traits. Students learn by asking clarifying questions. Therefore, discussion boards have been a staple of learning at the university level for years. This paper aims to utilize participation in discussion forums to predict final student performance. Using students' course grades at roughly the halfway point in the term and various discussion forum predictors, our model predicts the students' final percentage score. Using the model's prediction, instructors can speak with at-risk students and discuss ways to improve. The student grades and discussion board participation datasets are gathered from a graduate-level Electrical and Computer Engineering (ECE) course at Duke University. Various classical machine learning models are explored, with random forest yielding the highest accuracy. This random forest model, trained on discussion forum participation data, surpasses other similarly trained state-of-the-art models. Furthermore, related research attempts the classification problem of predicting what discrete letter grade a student will earn [1]. This is not an accurate representation of a student's performance, and therefore, we attempt the regression problem of predicting the exact percentage a student will earn. A significant finding of this paper is that our random forest model can predict student performance with an average error of approximately 2.3%. Additionally, our random forest model can generalize to a different graduate-level course and make performance predictions with an average error of 3.3%. The final important finding is that a model including discussion board predictors outperforms another whose sole predictor is the students' halfway point grade. This indicates that discussion forums hold significant value in determining final performance. We envision that the knowledge from our findings and our optimal random forest model can enable instructors to identify and support potentially at-risk students preemptively.

1 Introduction

Early and accurate student performance predictions are important because they allow instructors to intervene before struggling students stray too far from success. Unfortunately, many variables contribute to a student's performance in a course. Predicting student performance is challenging because most of these variables are difficult to both model accurately and obtain data for. Student participation and student grades at the halfway point of the term (checkpoint) constitute a set of measurable and impactful variables that help predict a student's final grade. Based on the authors observations, one efficient way of measuring student participation is through discussion boards. Previously, researchers' use of discussion board participation to predict student performance has yielded varying levels of success [1, 2, 3].

Ed Discussion is an anonymous discussion board that allows students and instructors to discuss class material. This paper uses Ed Discussion data in combination with student checkpoint (CP) grades to predict final course performance. The following experiments use data that originates from multiple graduate-level Electrical and Computer Engineering (ECE) courses at Duke University. As with all temporal-based prediction tasks, more accurate predictions come with the cost of becoming available later. Therefore, this paper replicates multiple experiments with varying time frames. Fortunately, the discussion forum we used, Ed Discussion, allows for easy data collection and the ability to download data within a desired time range [4]. During the time of this study, the Learning Management System (LMS) that was used for the analyzed courses was Sakai. Similar to Ed Discussion, Sakai allows for easy data collection, which was used to extract checkpoint grades and ground truth (i.e., actual final performance scores) [5].

All initial experiments were performed using one graduate course while the other was held out for generalization experiments. Prior to optimization, various classical machine learning models were compared to find the optimal model. Once the best-performing vanilla (i.e., without any optimization) model was selected, an extensive optimization process resulted in even better accuracy. The following experiments were to determine the trade-off between prediction strength and temporal availability (i.e., how early-on instructors can know about potential at-risk students with good confidence). As a final test of the model's generalizability, it predicted final performance on a separate held-out course's data.

The following is organized as follows. Section 2 discusses related state-of-the-art studies and evaluation metrics they used. Section 3 describes the methods employed to set up experiments. Section 4 presents the results of these experiments and analyzes them. Finally, Section 5 summarizes our work and Section 6 lays out directions for future work.

2 Background

Machine learning engineers have heavily studied Educational Data Mining (EDM) for more than a decade [6]. Historically, experiments involving EDM have fallen into two main categories. The first type of EDM research is course-level, where researchers attempt to predict student performance in a course before the final examination period [1, 2]. These student performance predictions allow instructors to reach out to at-risk students before it is too late. The other category of EDM research is department-level, where predictions are made regarding whether a student will graduate from their curriculum based on their current behavior [7].

This paper will perform course-level EDM experiments to predict student performance. Most course-level EDM research uses external parameters related to the student's environment outside of the classroom. Examples of these parameters include religion, age, gender, etc. [8, 9]. Although models using these predictors yield somewhat accurate results, they don't consider the students' work ethic or study habits. Therefore, we plan to factor in students' efforts when predicting their course performance.

One of the best ways to measure how much a student cares about their academic performance is to analyze their participation in the class [1, 10, 11]. A discussion forum is a platform that enables students to seek help from their peers and instructors. Multiple studies have focused on producing and analyzing the statistical correlation between discussion forum data and student course performance [11, 12, 13]. While statistical correlations can benefit inference, student performance predictions allow instructors to assist at-risk students. Therefore, this paper hopes to address this gap by contributing a model capable of early and accurate student performance predictions.

One major issue prevalent across many course-level EDM papers is the reliance on Data Mining tools such as Waikato Environment for Knowledge Analysis (WEKA) [14]. WEKA is a tool that automates the machine learning pipeline, only requiring users to provide the task and the input data. The issue with these tools is that there is little human interaction at any stage. This becomes clear when noting that most results were either trivial or exaggerated. Two examples are seen in the work of Al-Shehri et al. and Hashim et al., where trivial conclusions are made, and all data preparation and model selection are done via WEKA [15, 16]. Our results outperform these studies through ensuring that all experiment steps are done with care and human interaction.

Qualitative metrics curated from discussion forums can also be used to analyze performance. The work of Lee et al. separated forum posts into different categories based on qualitative attributes [10]. These attributes included details such as whether posts were on topic, length of posts, quality of posts, and quantity of posts. The key takeaways were that high-performing students posted reflective posts and kept their posts short. Furthermore, students who did not perform well read many posts and made posts featuring non-academic topics. The paper also analyzed whether the type and quality of instructor feedback impacted the overall class performance. It was found that the amount of instructor participation, the ratio of feedback used to provide the correct answer, and whether it was encouraging feedback all made no impact on student performance.

In a paper published by Carceller et al., authors conducted a study on one remote and one hybrid group of students. It found that in both groups, the students who participated more in the provided discussion forum did better than those who didn't. Furthermore, there was a more significant variance in blended students' final performance than their fully remote counterparts. Another important conclusion was that in the field of education, small correlations between features are very important [12]. This is because many untraceable factors contribute to student course performance. Some of these factors include mental health, financial status, family issues, and other highly volatile circumstances. Therefore, finding a positive correlation between a predictor and the prediction is worthy of further analysis. However, it is essential to note that discussion forum participation was optional in this paper, possibly leading to skewed results.

It is possible that some researchers believe that discussion forums do not provide enough benefit to students to warrant their use. This conclusion is likely formed because past related research did not use the correct discussion board statistics as its predictors. A study by Palmer et al. found that the number of views that students make does not affect their performance [13]. This implies that students known as ’lurkers’ are putting themselves at a disadvantage. A ’lurker’ is a student who commonly views other students’ posts but does not reply, ask follow-up questions, or make their own posts. If the number of views made by a student was the only discussion board predictor analyzed, then the research would conclude that discussion boards did not help determine student performance. However, the authors found that the total number of posts a student made was positively correlated to final performance. Therefore, the number of posts made by a student could be used to generate student performance predictions.

3 Approach

This paper used data from an anonymous discussion board called Ed Discussion to predict student performance. The data was generated from graduate ECE students at Duke University. Using the students’ Ed Discussion data and checkpoint grades, their final scores in Fundamentals of Computer Systems were predicted. Once the data was cleaned and preprocessed, it was fed into multiple regression models to see which had the highest performance. Prediction results were then compared against the ground truth data, from which average error metrics were analyzed. We then formed plots that compared average error metrics vs. the number of weeks after the checkpoint used for data collection. Next, the model whose data was collected the optimal number of weeks past the checkpoint was used to predict performance for an unseen graduate course, Systems Programming. This was a reasonable experiment because Systems Programming also used Ed Discussion as its platform for learning and communication. The excellent performance on this unseen course’s data showed that the best-achieving model generalized well and was not overfit to the Fundamentals of Computer Systems (ECE 550) weights. There were 124 available samples for the Computer Systems course and 87 for the Systems Programming (ECE 650) course.

3.1 Discussion Forum Student Engagement Categories

Ed Discussion logs statistics about each student’s engagement with others and content consumption on the platform. In this paper, we focus on using these statistics rather than the posts themselves. It is important to note that participation in Ed Discussion was optional for all courses used in this study. Furthermore, students can post and comment anonymously to their classmates. This feature is important because it encourages shy students to seek help when struggling. Moreover, it remedies many students’ fears of not wanting to ask what they believe to be a ”stupid question” in front of their peers. The Ed Discussion data included statistics on Views, Questions, Posts, Days Active, etc.

3.2 Collecting Temporal Discussion Forum Data

This paper aims to predict student performance in a graduate-level course using a progress checkpoint grade and Ed Discussion participation data. Since this prediction can identify potentially at-risk students, the earlier it is available, the better. Conversely, there needs to be enough content in the grade book in the selected time frame so that the students’ checkpoint grade is representative of their final course grade. The optimal data collection time range of Ed Discussion was everything up to and including the midterm exam. Additional datasets with more significant time ranges, such as one week, two weeks, and three weeks after the midterm, were curated. These datasets determined whether adding more data to the model enabled it to produce more accurate and robust predictions. Fortunately, Ed Discussion had a feature to extract data that fell between two dates. The four datasets were collected for Systems Programming and Fundamentals of Computer Systems.

3.3 Collecting LMS Data

The LMS the students used in the courses we studied was Sakai. We considered many factors when deciding the optimal point in the semester to collect the Sakai and Ed Discussion data. The cutoff point at the midterm was ideal because it was a good balance between giving teachers ample time to intervene and having enough data to predict student performance accurately. With more Ed Discussion data and a later checkpoint grade comes a more representative course standing and higher prediction results. However, the earlier the predictions of at-risk students are available, the earlier instructors can begin to assist.

3.4 Comparing Results to Ground Truth

The most popular regression metrics include R-squared, Mean Squared Error (MSE), and Mean Absolute Error (MAE) [17, 18]. It was decided that MAE would be the best metric to use going forward. The reasoning was that since the model predicted a final grade and compared it to ground truth, it could either be too high or too low. The prediction's distance from the ground truth could be recorded for each sample, but ultimately, the absolute aspect of MAE accounted for both overshoots and undershoots.

3.5 Histogram Plot of Error and Accuracy Logging

To better visualize a model's performance, we created a histogram that comprised each prediction's distance from its corresponding truth value. Each output prediction was produced via the leave-one-out (LOO) method to ensure that the maximum amount of data was being used for training and that there would be enough points to have a significant visualization. Each point in the histogram was calculated by subtracting the ground truth value from the prediction. It is important to note that only the model with the lowest MAE was selected to visualize its error histogram. It was redundant to perform post-processing analysis on models that were not as strong as the optimal model. The information logged for each histogram was the corresponding MAE and the standard deviation of the distribution of errors.

3.6 Comparing and Visualizing all Time Ranges vs. Error

As previously mentioned, analyzing how the model performed when trained on varying amounts of temporal data is essential. Therefore, for the optimal algorithm, four different models were introduced. The difference between these models was the amount of Ed Discussion data collected. These models had data up to and including the midterm cutoff, one week, two weeks, and three weeks after. Additionally, a few other models that didn't include any Ed Discussion data were trained as a sanity check. One such model had the checkpoint grade as its sole predictor. Comparing these models' metrics and errors to the Ed Discussion-based optimal model was crucial in understanding whether the Ed Discussion data improved the predictions.

3.7 Using Optimal Model to Predict on Unseen Data

The final experiment was to see how well the optimal model could generalize to unseen data. Generalizability was vital because it opened the possibility of using this model to predict scores for other classes in the future (assuming they also used Ed Discussion). Fortunately, there were available samples from a similar yet different course, Systems Programming, to replicate this scenario. The model that performed best on the Computer Systems course was trained on all available Computer Systems samples and then tested on all the Systems Programming samples. Next, the same error histograms and accuracy statistics were generated for this generalized experiment. The performance of this model will be analyzed in future sections. However, its performance implications can be discussed. If the model performed well on unseen data from a uniquely structured course, it could be used in other classes from different departments. This would transition our findings into a potential stand-alone application that could be used to better education across all fields that participate in online discussion forums.

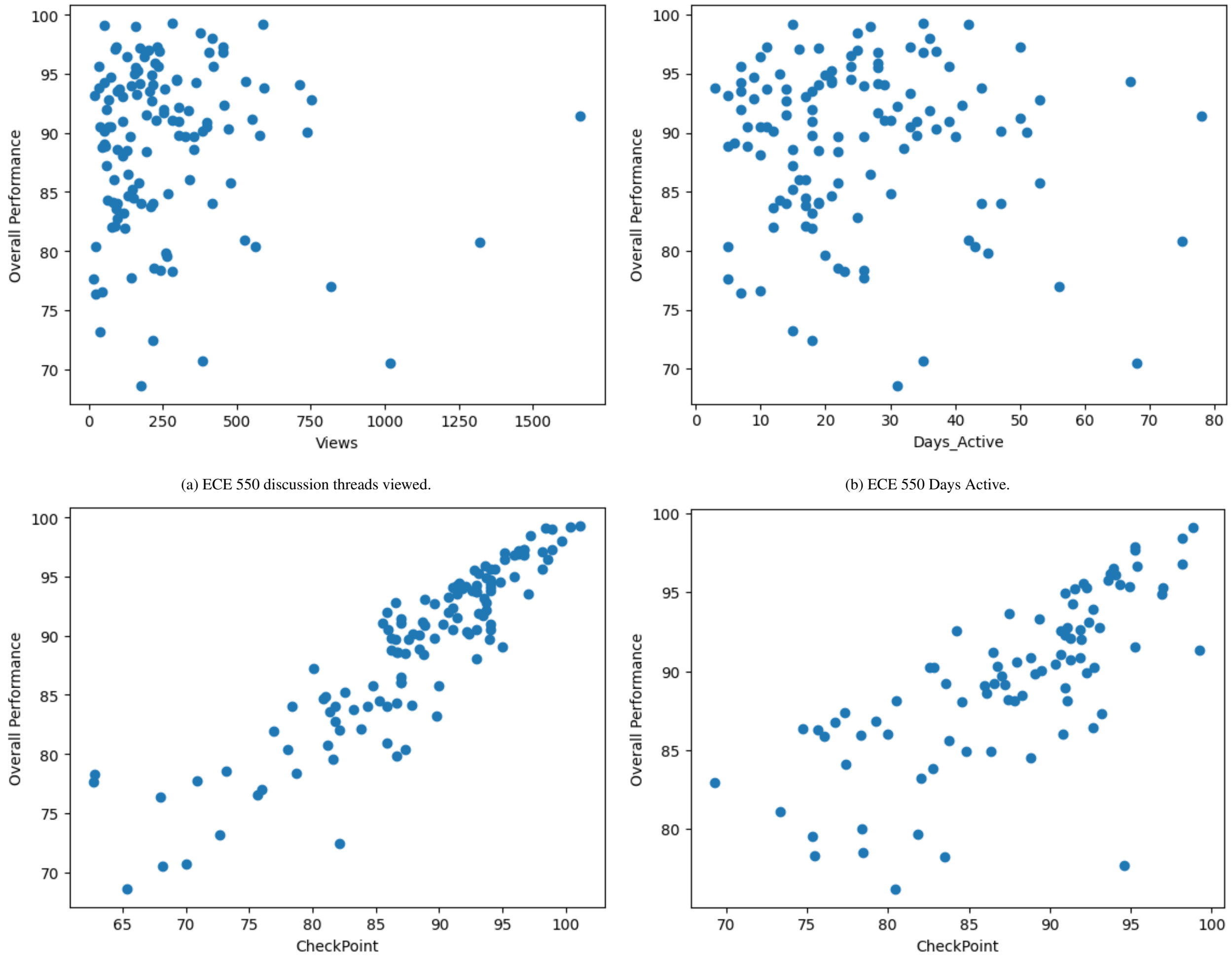
4 Results

This section aims to discuss our findings and accomplishments. Data exploration, model performance metrics, and conclusions inferred from our findings will be analyzed. Comparing input variables and ground truth scores before training models helps showcase any underlying relationships in the data. Visualizations of such comparisons will be available in section 4.1. Next, we will visualize how each model performs on the student performance prediction problem. The benefits of hyperparameter tuning will be highlighted by examining the performance of the out-of-the-box models and their tuned counterparts. Furthermore, the effects of regularization will be studied, displayed, and criticized. We will cross-analyze models with varying Sakai and Ed Discussion data time ranges. The ability of these models to generalize to unseen data will also be visualized. Finally, results from various experiments that were not the focus of this paper will also be discussed.

4.1 Input Parameter vs Output Data Exploration

Statistical analysis uses underlying patterns in the data to extract relationships between input and output features [17]. We believe machine learning is necessary for this type of research to make deeper connections between input and output features. As seen in Figure 1a, there is no relationship between the number of times students viewed discussion posts and their final course performance. A closer look reveals that the student who viewed discussion threads over 1500 times scored decently well in the course. Furthermore, many students who did not perform well seem to be on the lower side of views made. Despite this, the results are inconclusive enough to claim that high participation guarantees success.

Unlike Figure 1a, Figure 1b has a slightly positive linear relationship between the number of days students are active on Ed Discussion and their final performance. It is interesting to note that the



(a) ECE 550 discussion threads viewed.

(b) ECE 550 Days Active.

(c) ECE 550 grade at Checkpoint.

(d) ECE 650 grade at Checkpoint.

Figure 1: Various input parameters vs overall performance.

data points on the very far right of the independent axis are likely the same students across Figures 1a and 1b.

Figure 1c shows a significant positive relationship between the students' grades at the checkpoint and their final performance. Since this course has a lot of roughly equally weighted assignments, this relationship is expected. A closer look at Figure 1c reveals that some outliers do not follow the general linear trend. The hope is that Ed Discussion features will remedy these outliers. Finally, although the relationship is roughly linear, the band surrounding a y = x relationship is not perfectly thin. This indicates that there are still some imperfections when using checkpoint grades as the sole predictor for final performance.

Figure 1d depicts the relationship between final performance and checkpoint grade for ECE 650 students. Checkpoint grade and final performance in ECE 650 are not as tight of a linear relationship as in ECE 550. This finding is important because it shows the model needs the help of Ed Discussion to achieve accurate performance predictions. Since a ECE 550 trained model can generalize well to ECE 650 data, this shows the strength of Ed Discussion features. This is because the ECE 550 model, which depends less on Ed Discussion predictors, can still generalize well to data that relies heavily on Ed Discussion (ECE 650 data). This shows that not every class's checkpoint grade is equally important to final performance.

4.2 Importance of Leave-One-Out Accuracy Metric

Due to the limited number of samples and volatility of our data, using a LOO cross-validation metric is necessary for our results to be significant and replicable. Without using this metric, the 85-15 split includes samples so different that each trial's reported accuracy varies heavily.

Table 1: Models up to Checkpoint Ed Discussion 85-15 split accuracy.

Model	Mean Absolute Error	Mean Squared Error
Catboost	2.527	9.896
Random Forest	2.082	5.926
XGBoost	2.394	10.048
Linear Regression	2.034	5.025
Lasso Regression	2.015	4.968
Ridge Regression	2.137	5.384

Table 2: Models up to Checkpoint Ed Discussion Leave-One-Out accuracy.

Model	Mean Absolute Error	Mean Squared Error
Catboost	2.658	13.243
Random Forest	2.568	10.506
XGBoost	2.673	12.101
Linear Regression	2.499	11.166
Lasso Regression	2.390	9.738
Ridge Regression	2.522	10.970

Multiple models in Table 1 have impressive performance, especially Random Forest and Lasso Regression. The models' impressive results are not representative of subsequent trials where they could perform better or worse, depending on how the training and testing data was split. LOO validation ensures that less randomness is associated with the accuracy of the models. Table 2 depicts the more robust LOO accuracy across varying models. While the accuracies may not be as impressive in Table 2 compared to Table 1, they are deterministic and will always be the same. Furthermore, the LOO metric represents how the model does on average training on everything except the held-out test sample for every sample. Some trials will use an outlier as the validation sample, likely resulting in a poor prediction and reducing the average LOO score. This means the LOO score will already account for real-world outliers appearing during inference. Said another way, there is a trade-off between accuracy and robustness. LOO validation ensures fewer predictions will stray from the advertised model accuracy during inference [17].

4.3 Benefits of Hyperparameter Tuning Visualizations

Although classical machine learning models are typically interpretable, choosing the optimal parameters based on the available training data is difficult. Let's consider a model such as Random Forest, which has many tuneable hyper-parameters. While a set of parameters may work well on one dataset, this does not mean they will be suited for all datasets. A solution to these

Table 3: Models with increasing amounts of Ed Discussion data past Checkpoint.

	CP + 1 week		**CP + 2 weeks**		**CP + 3 weeks**	
Model	MAE	MSE	MAE	MSE	MAE	MSE
Catboost	2.641	12.215	2.791	13.225	2.791	13.515
Random Forest	**2.459**	**10.068**	2.508	11.015	2.471	10.526
XGBoost	2.758	13.145	2.875	14.355	2.710	12.778
Linear Regression	2.468	11.104	2.459	11.191	2.467	11.242
Lasso Regression	2.393	9.744	2.390	9.693	2.393	9.733
Ridge Regression	2.507	11.007	2.496	10.974	2.502	11.042

hurdles is the concept of hyper-parameter tuning. Using cross-fold validation on the training data, we obtain a set of optimal training parameters tailored to our dataset.

Table 4: Hyperparameter tuned vs vanilla Random Forest model accuracies trained on 1 week + Checkpoint.

	Vanilla LOO		**Tuned LOO**		**Tuned 85-15**	
Model	MAE	MSE	MAE	MSE	MAE	MSE
Random Forest	2.459	10.068	**2.329**	**9.736**	1.822	6.109

As expected, hyperparameter tuning increases performance, which can be seen in Table 4. While some may argue that this increase is insignificant, the tuning price is very low and, therefore, worth it. Additionally, considering that future work includes gathering more data to train these models, the price of hyperparameter tuning becomes negligible. Another interesting point is that, once again, the 85-15 split accuracy is extremely impressive. This metric is included to emphasize how volatile this metric can be when performing predictions on a small dataset [19].

Once again, due to data limitations, we use LOO accuracy to compare model performance. However, the hyperparameter tuned model uses an 85-15 split to search for optimal parameters. It then uses these parameters consistently for every LOO trial. This distinction is important because it would be less generalizable if we used all the data to find the optimal parameters at every iteration of LOO validation. Furthermore, if the tuning process occurred for every LOO trial, this would no longer be a LOO metric since there would be 124 slightly different models.

4.4 Regularization Improvements and Pitfalls

Another common machine learning practice is regularizing a model's weights to combat overfitting. Complex models with many weights memorize patterns in the training data, as this minimizes their loss function the quickest. This is problematic because the model fails to generalize well to unseen data [19, 20, 21].

Figure 2 shows the effect of regularizing feature weights during regression. Linear regression is included as a baseline for how a linear model weighs the input features' importance. All three models value the checkpoint grade most when predicting a student's final score. This finding makes sense because students typically perform equally for a course's first and second half. Interestingly, the comments feature is weighted as important in all three models and is the only other significant feature for Lasso Regression. Since these models aren't paying as much attention to the Days Active or Views features, our findings align closely with the work of Palmer et al. [13]. We see that lurking is not as effective as contributing with comments or answers.

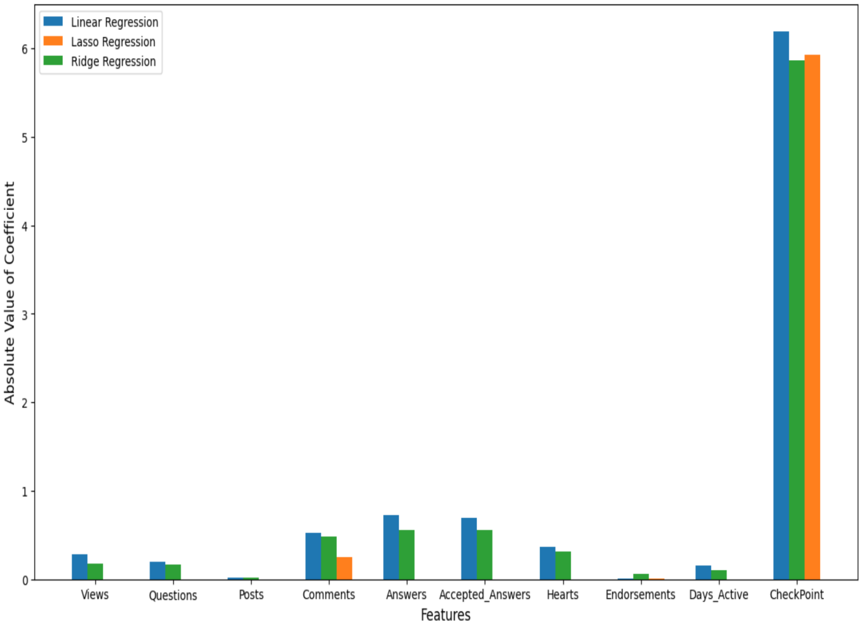
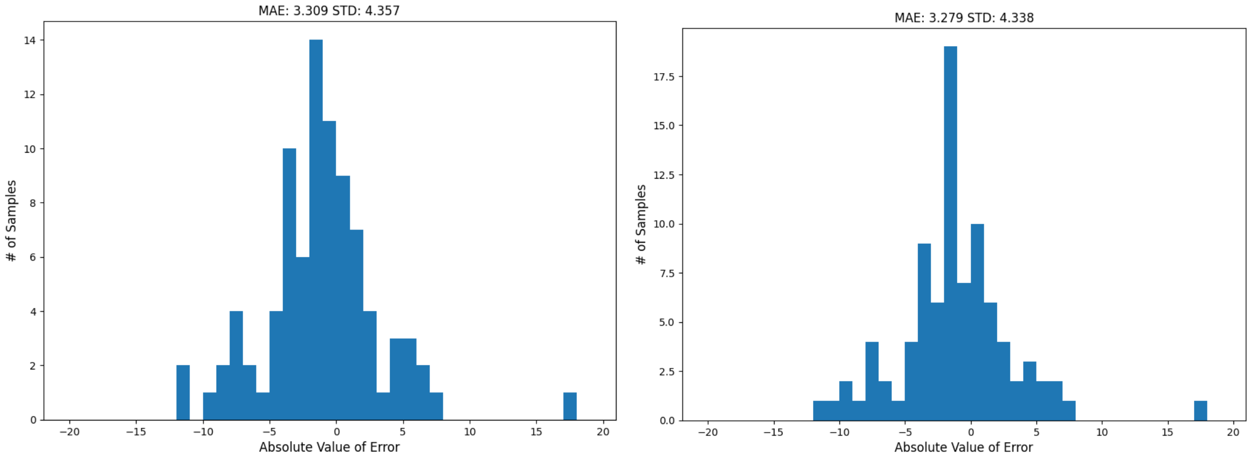


Figure 2: Various models' feature coefficient values.

As seen in Table 3, applying Lasso and Ridge regularization both increase performance. However, this performance enhancement cannot justify using these models for future experiments. This is because this paper aims to see how Ed Discussion can assist in predicting student performance. Regularization techniques zero out or make certain features extremely small to generalize better. In the case of our data, and as seen in Figure 2, these techniques eliminate most of the Ed Discussion-related parameters. Determining which parameters are most important is a data exploration task and is therefore out of the scope of this paper.

4.5 Various Time Range vs Accuracy Model Visualizations



(a) Checkpoint + 0 weeks.

(b) Checkpoint + 2 weeks.

Figure 3: Histogram of ECE 550 train - ECE 650 test prediction error with data from a varying amount of weeks after the checkpoint.

In most cases, adding more data increases model performance. For most of the experiments in this paper, multiple trials are held for varying time frames of data. For example, a model trained on Ed Discussion data until the checkpoint (usually about halfway through the term) is compared to a model trained on data until one week after the checkpoint. This is repeated until three weeks after the checkpoint. One might ask why we do not include as many weeks as possible since more data increases performance. The main reason is that the instructor will have access to predictions late into the term, only for the model to make slightly better predictions. For example, imagine an instructor using this model to predict their students' final scores as a preventative measure for those on track to do poorly. If they have to wait more than 1, 2, or even three weeks after the checkpoint to make a prediction, the term would already nearly be over. In a way, the sooner the model can make accurate predictions, the more value it holds. With earlier access to performance predictions, instructors have more time to assist students on track to do poorly.

Table 3 depicts how adding more data can increase model performance. Oddly enough, there are some cases where an extra week of data hurts the model, which might indicate that Ed Discussion data is not worth using. One potential explanation for this is that Ed Discussion data taken closer to the midterm checkpoint cutoff is more relevant than 2 or 3 weeks past this point, causing the model to misuse this additional information. Regardless, as discussed in section 4.7, using Ed Discussion is justified and leads to better performance instead of not including it.

Figures 3a and 3b show that when generalizing to an entirely different course, adding more data creates more accurate models. However, this trend is not perfectly consistent, and the optimal amount of ECE 550 training data for ECE 650 generalization is two weeks past the checkpoint instead of three. This shows that, in general, Ed Discussion data serves a significant role in student performance prediction.

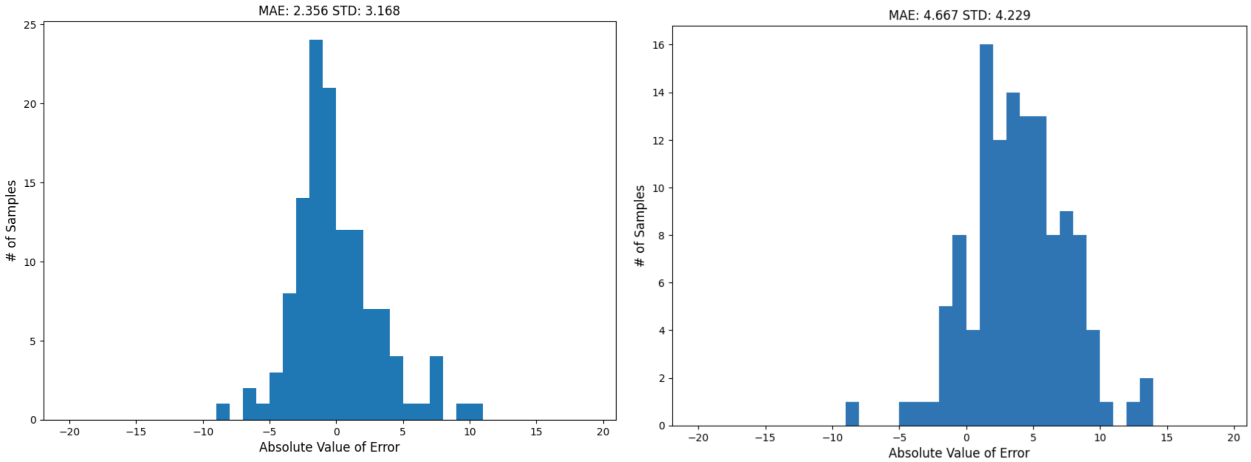
4.6 Optimal Model's Ability to Generalize to Unseen Data

We believe that a model should be able to train on training data and perform well on testing data if both datasets came from the same domain. However, training on a given course, such as ECE 550, and making accurate predictions on another course, such as ECE 650, is much more impressive. This indicates that Ed Discussion can make accurate student performance predictions across a wide variety of courses.

ECE 550 and ECE 650 only have two commonalities, one of which is the instructor, and the other is the use of Ed Discussion. This means the content, number of assignments, and grade breakdown are different. Figures 3a and 3b show that despite these differences, the models trained on ECE 550 still accurately predict student performance in ECE 650. This is important because it shows that the model does not memorize specific patterns in the training data. The final interesting point regarding the model's generalization ability is that the batch of students is not in the same academic year. ECE 650 (the second course in the series) took place in Spring 2022, while ECE 550 took place in Fall 2022. This means that the model cannot cheat by learning from the behavior of students who appear in both datasets.

4.7 Miscellaneous Experiments such as 'Is the Ed Discussion Data Worth Using?'

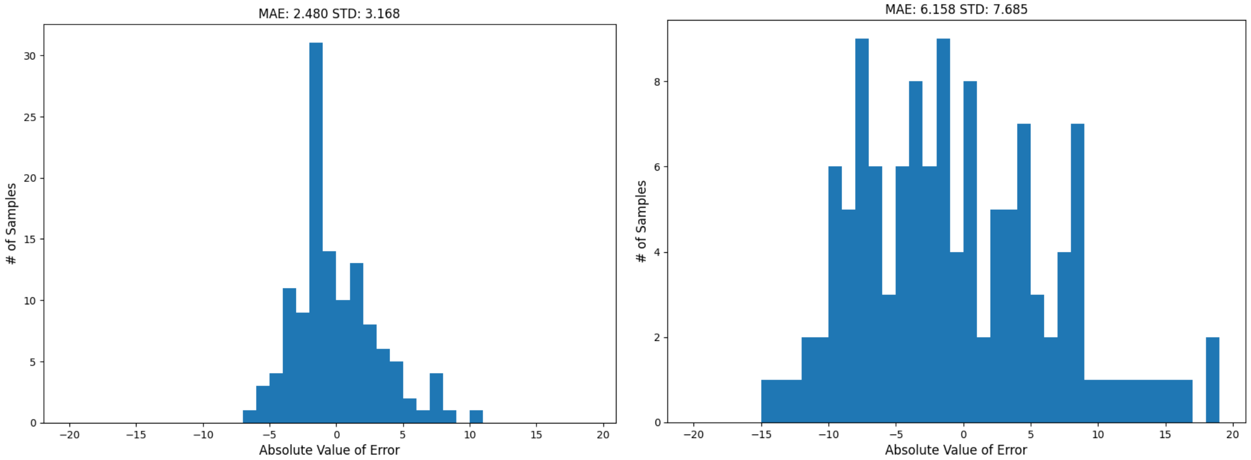
Figure 4 contains the benchmark and baseline for the following experiments. Figure 4a is the best-performing model, and Figure 4b shows results for simply extrapolating final performance from checkpoint grade. The improvement over the baseline (Figure 4b) in Figure 4a proves that



(a) Checkpoint + 1w trained model.

(b) Baseline (using Checkpoint grade to predict final performance).

Figure 4: Histogram of LOO prediction error for hyperparameter tuned ECE 550 model trained with 1 week of data past checkpoint and baseline.



(a) Only Checkpoint grade.

(b) Only Ed Discussion statistics.

Figure 5: Histogram of LOO prediction error for hyperparameter tuned ECE 550 model trained with only Checkpoint grade and with only Ed Discussion statistics.

machine learning and Ed Discussion data are well worth the effort. The performance for each sample is predicted via LOO cross-validation because of the low amount of data available. Once again, this ensures consistent results that are less susceptible to irregularities caused by train-test-split.

Figure 5a results from training a model using checkpoint grade as its sole predictor. The results are very accurate, which is expected due to the high correlation between final performance and student progress halfway through the course. However, it is essential to note that the LOO average error is worse than when Ed Discussion is included (Figure 4a). At first glance, the difference between Figure 4a and Figure 5a may seem negligible. However, as supported by the work of Carceller et al., minor improvements in EDM are statistically significant due to the plethora of factors that go into student performance. Further, experiments in Section 4.6 motivate why training a model using discussion forum statistics makes the model more robust and better at generalizing to unseen data.

Finally, we discuss Figure 5b, which only uses Ed Discussion data as its input features. While the accuracy is lower than previous figures, it is still impressive, considering the model uses discussion forum data to predict student performance. The ability to predict a student's performance within half of a letter grade based on their discussion forum participation is impressive, even more so when considering there are only 124 training samples. The primary purpose of including this model is to show that the impressive results of the Ed Discussion and checkpoint grade model are not solely due to the checkpoint grade.

5 Conclusion

In this paper, we applied machine learning techniques on student discussion forum data to predict student performance in graduate-level courses, and we were able to obtain a very good prediction accuracy. We also showed that our method could also generalize to other courses by training it with the data of one course and obtaining good prediction results with the data of another course. Out of many evaluated out-of-the-box models, Random Forest proved to be the best balance between bias and variance with an MAE of 2.459. After an extensive hyperparameter search, our optimized RF model achieved an MAE of 2.329.

The relationship between temporal prediction availability and accuracy is expected to be inversely related. Experiments confirm that as more data is added, meaning the prediction is available later, the model makes more accurate predictions. While the model trained on data up to one week after the checkpoint is optimal for its evaluation, training two weeks past the checkpoint is ideal for generalizing to the other course. The best RF model trained on ECE 550 data and generalized to ECE 650 yields an MAE of 3.309. This is roughly one percentage point lower than an ECE 550-trained model predicting its own students' final performance.

Another interesting finding is that a model trained on only the checkpoint grade (no Ed Discussion data) can make performance predictions with an MAE of 2.480. This shows there is value in including Ed Discussion data, even if only a slight performance increase exists. Finally, a model trained only on Ed Discussion data can make predictions with an MAE of 6.158. While this is less impressive than the checkpoint grade-only model, its accuracy is impressive considering the number of variables that go into a student's final performance.

The ability to work with at-risk students at the checkpoint outweighs using indicators such as poor exam scores, or lack of attendance. This is shown through analyzing the increase in model performance when adding in Ed Discussion data.

6 Future Work

Future work involves further experimentation to squeeze out better performance. The first objective is to add text-based predictors from the students' posts to our model. Another objective is that as more courses are taught using Ed Discussion, their data can be used as additional training data to improve the model's performance. The final objective is to evaluate our model on undergraduate-level and graduate-level courses in potentially different departments which also use Ed Discussion. This would make our model available to a larger group of instructors and students and isn't a very far stretch from the current model's capabilities.

References

[1] Cristóbal Romero, Manuel-Ignacio López, Jose-María Luna, and Sebastián Ventura. Predicting students' final performance from participation in on-line discussion forums. *Computers & Education*, 68:458–472, 2013. ISSN 0360-1315. doi: https://doi.org/10.1016/j.compedu.2013.06.009. URL `https://www.sciencedirect.com/science/article/pii/S0360131513001607`.

[2] Yutong Liu, Si Fan, Shuxiang Xu, Atul Sajjanhar, Soonja Yeom, and Yuchen Wei. Predicting student performance using clickstream data and machine learning. *Education Sciences*, 13(1), 2023. ISSN 2227-7102. doi: 10.3390/educsci13010017. URL `https://www.mdpi.com/2227-7102/13/1/17`.

[3] Febrianti Widyahastuti and Viany Utami Tjhin. Performance prediction in online discussion forum: state-of-the-art and comparative analysis. *Procedia Computer Science*, 135:302–314, 2018. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2018.08.178. URL `https://www.sciencedirect.com/science/article/pii/S1877050918314674`. The 3rd International Conference on Computer Science and Computational Intelligence (ICCSCI 2018) : Empowering Smart Technology in Digital Era for a Better Life.

[4] KK Ashraf. Ed discussion, 2012. URL `https://edstem.org/`.

[5] Charles Severance. Sakai learning management system, 2005. URL `https://www.sakailms.org/`.

[6] C. Romero and S. Ventura. Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, 33(1):135–146, 2007. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2006.04.005. URL `https://www.sciencedirect.com/science/article/pii/S0957417406001266`.

[7] Concepción Burgos, María L. Campanario, David de la Peña, Juan A. Lara, David Lizcano, and María A. Martínez. Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. *Computers & Electrical Engineering*, 66:541–556, 2018. ISSN 0045-7906. doi: https://doi.org/10.1016/j.compeleceng.2017.03.005. URL `https://www.sciencedirect.com/science/article/pii/S0045790617305220`.

[8] Siti Dianah Abdul Bujang, Ali Selamat, Roliana Ibrahim, Ondrej Krejcar, Enrique Herrera-Viedma, Hamido Fujita, and Nor Azura Md. Ghani. Multiclass prediction model for student grade prediction using machine learning. *IEEE Access*, 9:95608–95621, 2021. doi: 10.1109/ACCESS.2021.3093563.

[9] Anal Acharya and Devadatta Sinha. Early prediction of students performance using machine learning techniques. *International Journal of Computer Applications*, 107:37–43, 12 2014. doi: 10.5120/18717-9939.

[10] Ji-Eun Lee and Mimi Recker. The effects of instructors' use of online discussions strategies on student participation and performance in university online introductory mathematics courses. *Computers & Education*, 162:104084, 2021. ISSN 0360-1315. doi: https://doi.org/10.1016/j.compedu.2020.104084. URL `https://www.sciencedirect.com/science/article/pii/S0360131520302827`.

[11] Cho Kin Cheng, Dwayne E. Paré, Lisa-Marie Collimore, and Steve Joordens. Assessing the effectiveness of a voluntary online discussion forum on improving students' course performance. *Computers & Education*, 56(1): 253–261, 2011. ISSN 0360-1315. doi: https://doi.org/10.1016/j.compedu.2010.07.024. URL `https://www.sciencedirect.com/science/article/pii/S0360131510002198`. Serious Games.

[12] Charles Carceller, Shane Dawson, and L. Lockyer. Improving academic outcomes: Does participating in online discussion forums payoff? *International Journal of Technology Enhanced Learning*, 5:117–132, 02 2013. doi: 10.1504/IJTEL.2013.059087.

[13] Stuart Palmer, Dale Holt, and Sharyn Bray. Does the discussion help? the impact of a formally assessed discussion on final student results. *British Journal of Educational Technology*, 39:847 – 858, 10 2007. doi: 10.1111/j.1467-8535.2007.00780.x.

[14] Liran Zvibel. Home, 2005. URL `https://www.weka.io/`.

[15] Huda Al-Shehri, Amani Al-Qarni, Leena Al-Saati, Arwa Batoaq, Haifa Badukhen, Saleh Alrashed, Jamal Alhiyafi, and Sunday O. Olatunji. Student performance prediction using support vector machine and k-nearest neighbor. In *2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE)*, pages 1–4, 2017. doi: 10.1109/CCECE.2017.7946847.

[16] Ali Salah Hashim, Wid Akeel Awadh, and Alaa Khalaf Hamoud. Student performance prediction model based on supervised machine learning algorithms. *IOP Conference Series: Materials Science and Engineering*, 928 (3):032019, nov 2020. doi: 10.1088/1757-899X/928/3/032019. URL `https://dx.doi.org/10.1088/1757-899X/928/3/032019`.

[17] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

[18] Colin Lewis-Beck and Michael Lewis-Beck. *Applied regression: An introduction*, volume 22. Sage publications, 2015.

[19] Roger Grosse. Generalization, Jan 2024. URL `https://www.cs.toronto.edu/ lczhang/321/notes/notes09.pdf`.

[20] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[21] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970.