**Final Report for Deep Dive 3**

**1 Motivation and Introduction**

The objective of our deep dive model is to predict whether a course with the same course number and department will be easy or hard next year. Alongside with the model, we also find a bunch of interesting behavior which is crucial for students, professors and the university to make informed decisions.

**1.1 Motivation**

Many students suffer from their GPA. Although GPA cannot really value people, a low GPA might be very harmful for college students, including drop from schools, peer pressures etc. Thus, through our model, students can choose and find challenging or easy courses for next semester as they wish. For those who want to have a more relaxed semester, easy courses are recommended. And for those who want to spend more time, those courses which are challenging under our model prediction might be a good choice.

Within our university, there are many courses that have similar contents. For example, there are many introductory statistics and probability courses on campus, including but not limited to MATH 463, STAT 400, CS361, IE310, ASRM401. In some cases, students are provided with the choice to choose whichever introductory statistics course they would like to take. Alongside subtle differences between courses, whether the course is challenging or not is also important for students to make decisions.

**1.2 Assumption**

We assume that whether a course is easy or difficult can be determined by the percentage of students who got an A or A+ in the course. Further, we define easy courses to be those courses that more than half of the students get an A or A+.

**2 Dataset Overview & Visualization**

**2.1 Data sample**

The following graph shows our sample data:

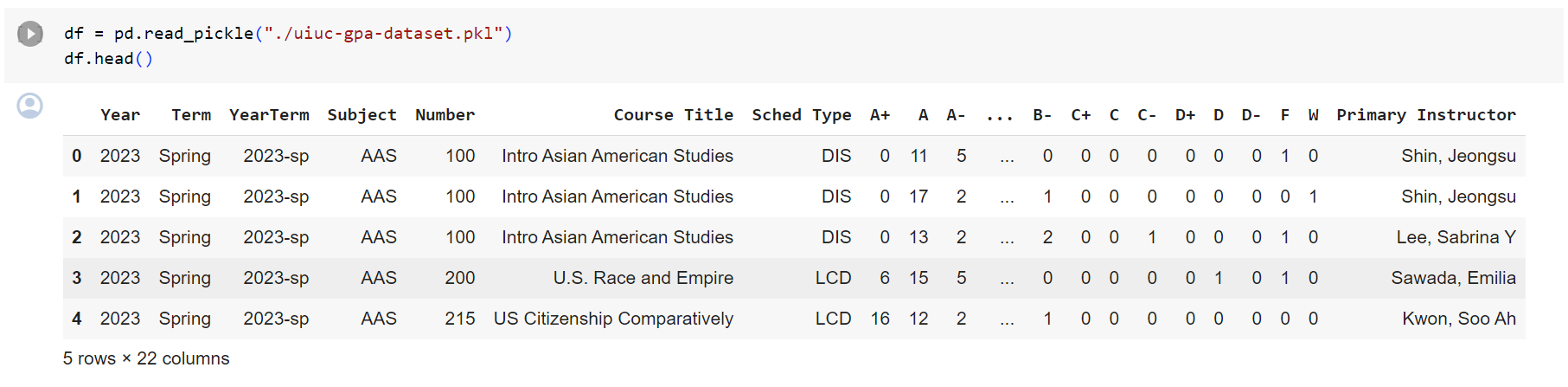


Image 2.1.1. Sample data.

We print out the head of our data and find that features appeared in our dataset: Year, Term, YearTerm, Subject, Number,Course Title, Schedule Type, letter grades and primary instructor.

**2.2 Handle NaN**

Nan value exists in column Schedule Type and Primary Instructor. We avoid nan value in training as we don’t use Primary Instructor in our model as the constant change in professor might be hard for the model to learn pattern from this category.

**2.2 Handle imbalance data**

Imbalanced data might occur in a classification problem setting.This data was collected systematically without any bias or over/under-sampling of certain categories, it's more likely to be balanced. As shown, the percentage of 4.0 for each section as majority is 0.38, which is a signal for very slightly imbalanced data, which can be ignored.

**2.3 Visualizations**

Here are some data discovery graphs we have made through our deep dive.

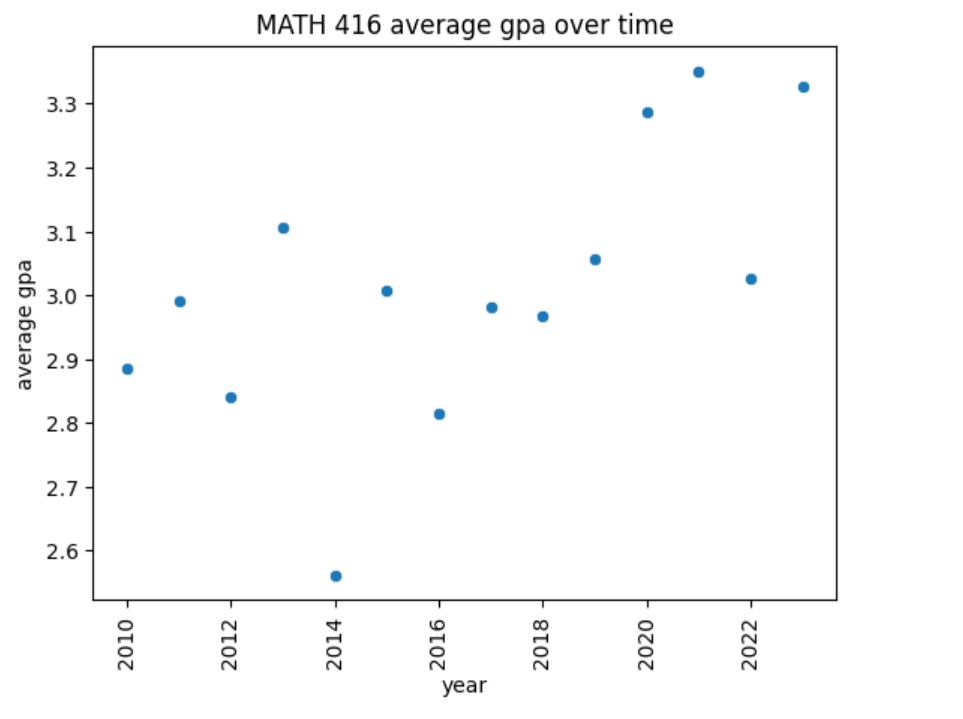


Image 2.3.1. Average GPA of MATH416 over time.

We realize year might be an important factor that can affect average GPA. For example, the average GPA of Math 416 is about 2.5 in 2014 where the average GPA of MATH 416 in 2021 is about 3.4.

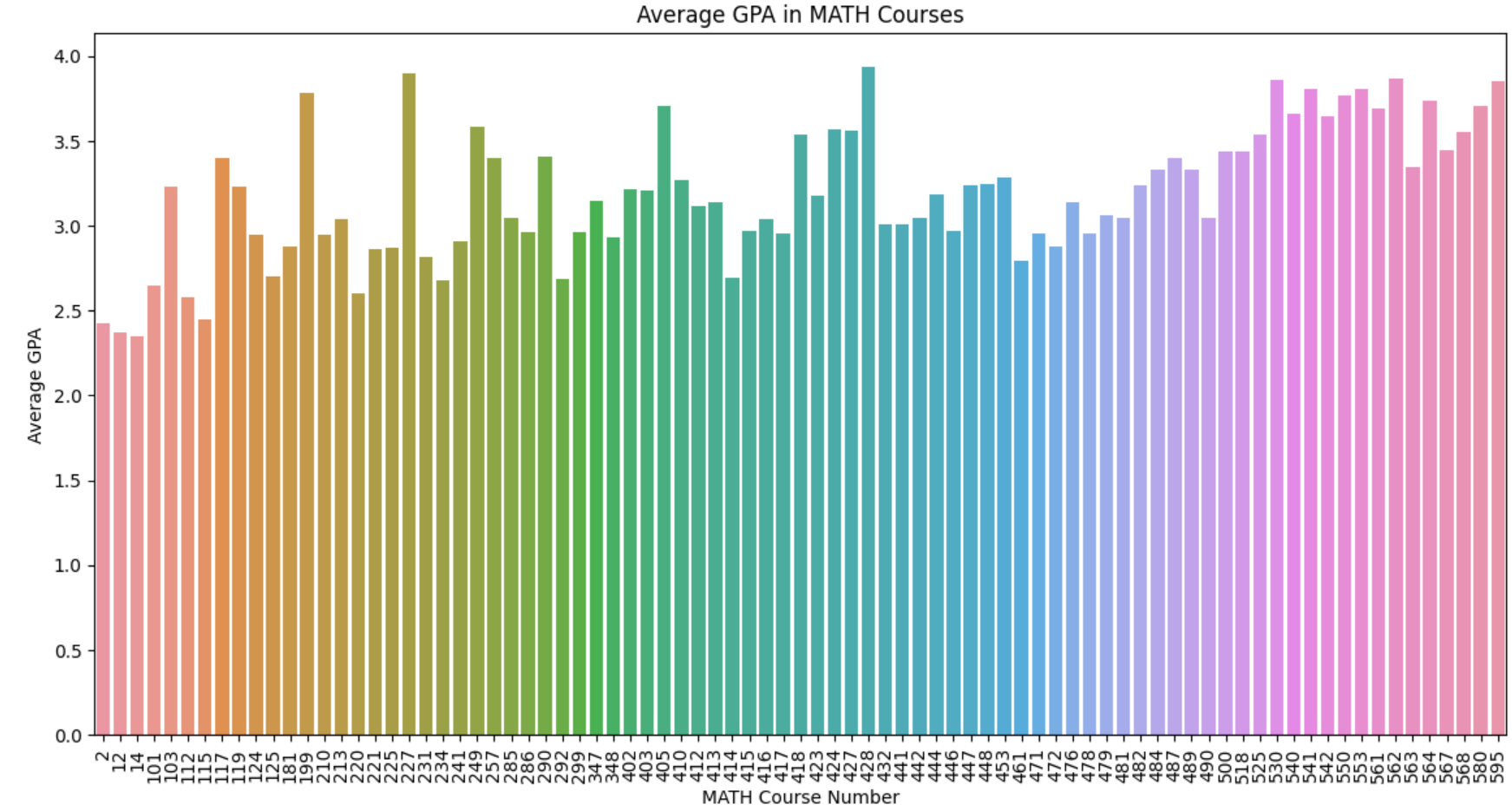


Image 2.3.2. How GPA changes with respect to years.

To see the general trend of GPA as course number increases, we make a graph to show average GPA across 2010-2022 for each math course. The general trend is that average GPA increases as GPA increases.

**3 Model**

**3.1 Objective**

The objective of our deep dive model is to predict whether a course with the same course number and department will be easy or hard next year.

In order to get the train and test data, for each course in Math in each year, we calculated whether the same course is easy or not next year, and use the value we calculated as the expected model output to train the model. With this, we can utilize all 12 years of data to train our model.

Most features presented in the original data were used in our model. However, for the Primary Instructor parameter, we choose to not use it in our model as the constant change in professor might be hard for the model to learn pattern from this category. Also, for other categorical data, we use one hot encoding to process it for the model.

**3.2 Model Training**

**3.2.1 Training Set up**

We consider three activation functions. The first one is Sigmoid activation , the second one is Tanh activation and the third one is ReLU activation . These are among the most commonly used activation functions. Note also that Sigmoid and Tanh activations are often used in classification problems, which is closely related to our goals. On the other hand, ReLU differs from these functions in that it is non-analytic, more easy to handle the derivatives, and is less likely to suffer from vanishing-gradient problems.

We would like to combine these activation functions with different training algorithms to figure out the best model. The algorithms we consider are SGD, Adam, Adagrad, ASGD and Adadelta.

The structures of the base model is a four-layer fully-connected neural network:

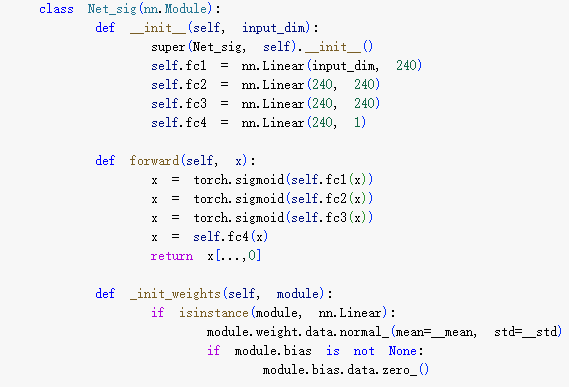


Image 3.2.1.1. Structure of models. Input -> 240 neurons -> 240 neurons -> 240 neurons -> linear combination of 240 neurons.

In the training, we set the learning rate as 0.0001, the iteration number as 1000, and we used a standard Gaussian sampling for initialization of parameters for our models.

**3.2.2 Mini Batch training**

We also perform a mini batch training after training these models. We set the batch size to be 64, and consider all the three activation functions: Sigmoid, Tanh and ReLU.

**3.2.3 Comparison of Optimization methods**

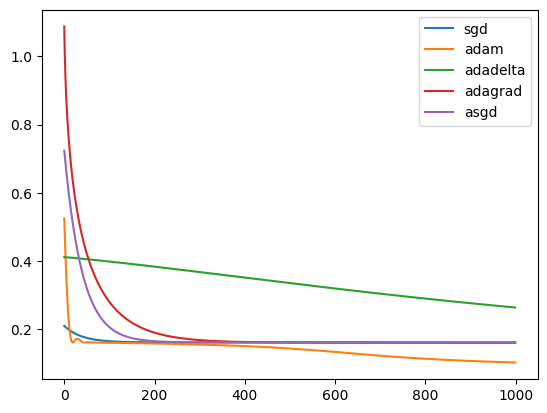
****

Image 3.2.3.1 Loss function vs time steps under sigmoid as activation function

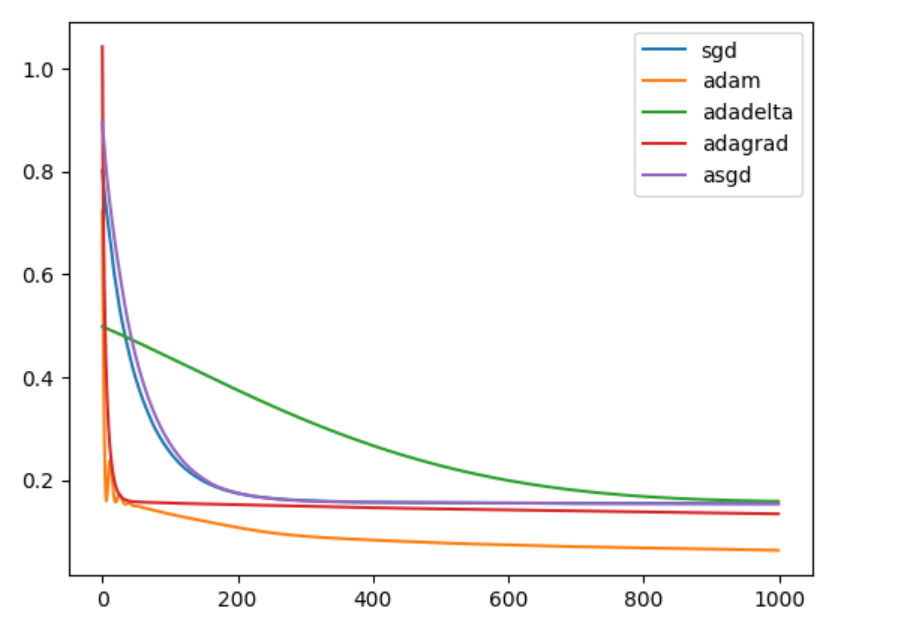
****

Image 3.2.3.2 Loss function vs time steps under sigmoid as activation function

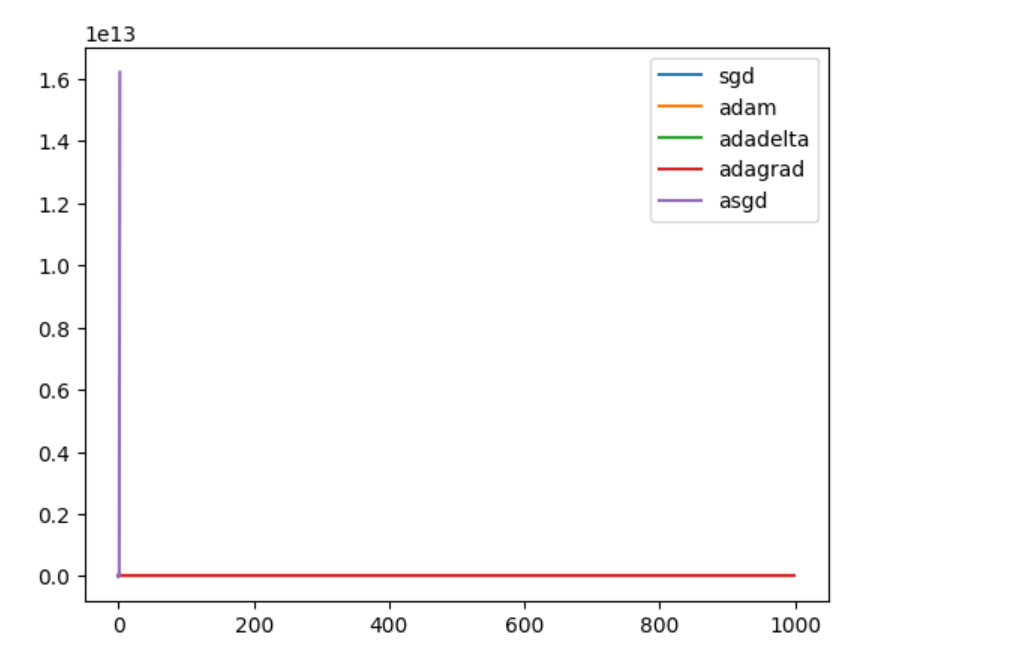
****

Image 3.2.3.3. Loss function vs time steps under sigmoid as activation function

The picture above shows a comparison of various optimization algorithms used in the context of different activation functions. Here are several conclusions:

* Convergence speed: the optimization algorithms converge at different speeds. Adam and Adadelta, for instance, have a rapid decline in loss initially, suggesting that they are quicker to find a direction that reduces loss. ASGD(Average Stochastic Gradient Descent) also decreases quickly but plateau earlier than Adam.
* Stability: SGD (Stochastic Gradient Descent) shows a more stable, but slower convergence. It doesn’t decrease as rapidly as Adam or Adadelta, but it maintains a steady pace. This stability can sometimes be beneficial in avoiding local minima early in training.
* Long-Term Performance: Over a longer period, Adam maintains a lower loss compared to the others, suggesting that it might be more effective for this GPA prediction model and loss landscape.
* Adagrad:The Adagrad algorithm appears to decrease smoothly and then plateau. This could be due to its learning rate adjustment mechanism, which can cause the learning rate to become very small over time, slowing down learning.
* Computational Efficiency: Although Adam and Adadelta converge quickly, they might be more computationally expensive per iteration compared to SGD.

**3.2.4 Hyperparameter training**

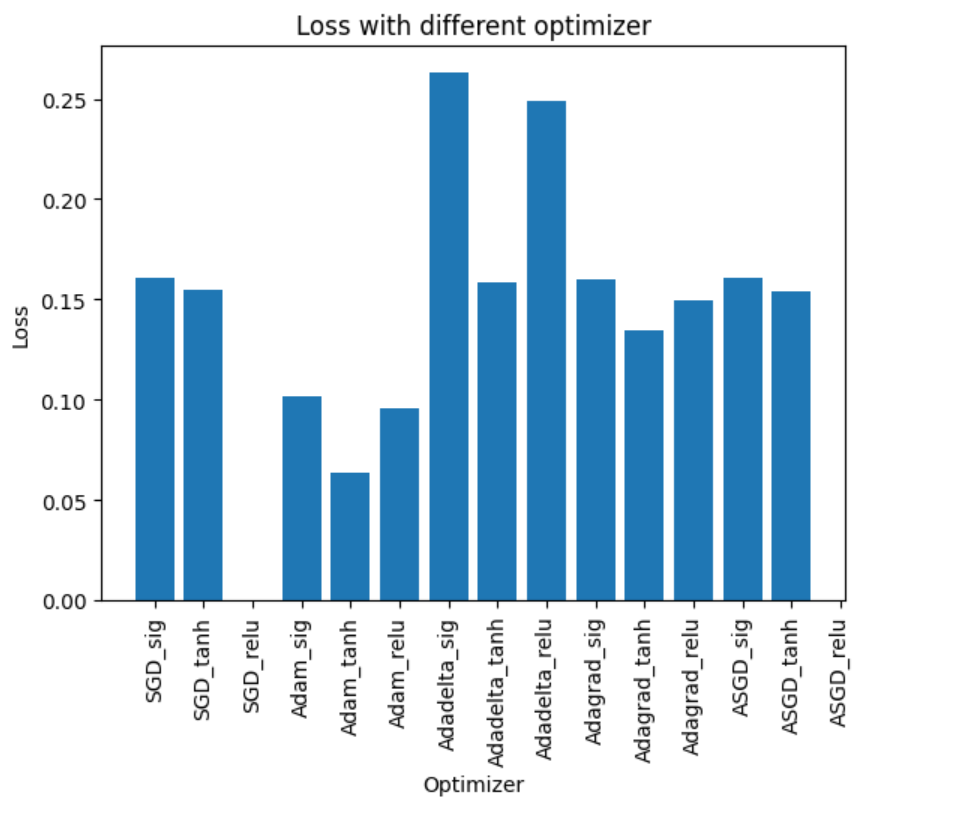
****

Image 3.2.4.1: Loss with different optimizers with respect to different activation function.

The image provides comparison with different activation functions. One important thing is that ReLU is not fitable for this model. Here are some conclusions:

* Activation functions: The choice of activation function has a significant impact on the performance of the optimizer. For example, when comparing the SGD optimizer with the sigmoid activation function to SGD with tanh activation function, the latter shows a lower loss, suggesting that tanh may be performing better with sigmoid as optimization function.
* Tanh’s Overall Performance: Tanh results in lower loss values compared to sigmoid, indicating that it is a more effective activation function for this particular problem or dataset.
* Consistency: Some optimizers, like Adam, are more robust in attaining lower final loss when we change the activation functions, compared to others such as SGD.

**3.2.5 Feature importance**

We first used year, course number, schedule type and scale of grade as variables to predict whether a Math course is easy or not. We define whether a Math course is easy if the percent of A/A+ is greater than 30 percent. Then we train our model and calculate Shapley value for each feature and we get

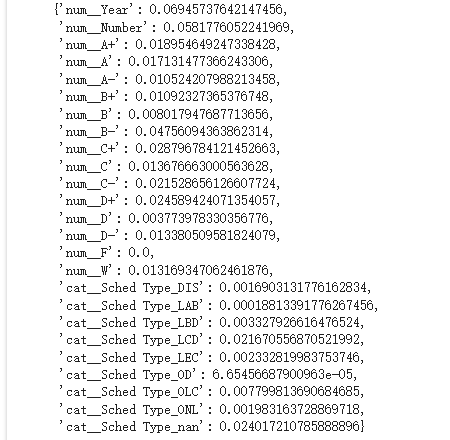
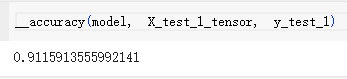


Image 3.2.5.1. Feature importance results.

As we can see, year and number are more important than the grade scale and schedule type. So the feature “num\_Year” and course number have the most effect on predicting whether a Math course is easy or not. And the grade scales and “Sched Type” are least important. We then dropped “Sched Type” and re-trained our model. This allows us to have a simpler dataset, smaller model (fewer parameters), and less training time, meanwhile maintaining a satisfactory accuracy.

**3.3 Model Results**

****

The accuracy of our model on the test set is 91.2%, meaning that our model is able to correctly predict or classify the course as easy or hard given last year’s data. Here are some important points:

* Balance of dataset: Because our dataset is balanced, the accuracy shows that our model has a high performance.
* Generally: This prediction generalizes with UIUC’s math course which holds at least once per two years.

**4 Conclusion**

In this project, we trained our model with different activation functions and algorithms. The best testing accuracy of prediction is 91.2% which is quite satisfactory to us.

We also investigated the feature importance with respect to our model. The most important features are course numbers and years. Some possible explanations are given below. In terms of year, we found that the *average GPA* increased from 2010 to 2022. This might be due to

1. Professors gain practical experience from the past courses. They improve their teaching skills in a way that helps students have a better understanding of the course.
2. Advances in teaching methods and technology may positively impact student learning outcomes. Interactive and engaging teaching methods, as well as the integration of technology, could contribute to better understanding of course material and improved performance.
3. Improved access to educational resources, including libraries, online databases, and academic support services, can enhance students' ability to excel in their studies, potentially leading to higher GPAs.
4. GPA Inflation: Higher grades are awarded for the same level of performance over time, contributing to the increase in average GPA. This may occur due to changing grading standards, pressure to maintain high retention rates, or a desire to boost students' competitiveness in the job market or graduate school admissions.

5. UIUC implements initiatives aimed at improving student success, retention, and graduation rates such as the Student success center at illinois. Support services, tutoring programs, and mentorship opportunities can help students perform better academically, leading to higher average GPAs.

In terms of course numbers, we found that the *average GPA* increases as course number increases. This might be due to

1. Increased specialization: in many academic systems, higher-level courses are often more specialized and tailored to a student's chosen major or field of study. As students progress through their academic journey, they may have a better understanding of their interests and strengths, leading to improved performance in courses aligned with their chosen path. Also, with each passing course, students accumulate a broader set of skills and knowledge. As they build on their foundational understanding, they may find subsequent courses more manageable and perform better academically.
2. Motivation and interest: As students progress in their academic journey, they may have the opportunity to choose courses that align more closely with their interests. This increased alignment can boost motivation, engagement, and overall performance.
3. Selective Enrollment: Higher-level courses may have prerequisites, resulting in a more selective enrollment process. This could lead to a cohort of students who are more committed, academically inclined, and better prepared for the challenges of advanced coursework, contributing to higher average GPAs.
4. Smaller Class Sizes: Advanced courses often have smaller class sizes, allowing for more personalized attention from instructors. This can facilitate better understanding of the material, increased participation, and more targeted feedback, all of which can positively impact student performance.

**5 Discussion**

In this project, we investigate the grade data of UIUC courses offered in recent years. In this part, we make discussions about our results, our methods and future work.

**5.1 Model Training**

Focusing on predicting how hard it is for each course, in Milestone 4 we used different activation functions and algorithms to train the models to fit our data. In particular, we used Sigmoid, ReLU and Tanh as activations, and SGD, Adam, Adadelta, Adagrad and ASGD as our algorithms. We noticed that different choices of activation functions and algorithms significantly affected the final loss value and the generalization of our model. We found that Sigmoid or Tanh paired with Adam works best, i.e., gives the smallest final loss value and highest test accuracy. On the other hand, any activation function paired with SGD or Adagrad works poorly.

**5.2 Practical Implication of Results**

In this project, we trained our models to predict the difficulty of each course based on our selected features. The difficulty of a course is of great importance to the school, the students and the professors. Students may want to know what factors contribute to the difficulty of the course, so that they can decide which course to take, how the teaching style of a course would affect its difficulty, and when to take it. Also note that we only investigated math courses. However, as UIUC is mostly a science & engineering-based school, most students are required to take several, if not many math courses.

This in turn helps schools arrange these courses. For example, the school may add sections to math courses that many students are interested in, while canceling less attractive ones. Besides, it helps the school notice what kind of teaching styles and resources help students most. For professors, the results help them decide what courses to offer, and in what way they should teach them.

**5.3 Future Works**

In this part we discuss several possible ways to extend our work and results.

1. In this project our entire focus is math courses. Even if math courses are a must for most of UIUC students, there are many other courses they take. For example, as UIUC is mostly science & engineering-based, we may also investigate engineering courses.
2. After investigating several different categories of courses, it is interesting and meaningful to see how their difficulties are related to one another. For example, if a math course MATHXXX is hard, and another course ENGYYY uses a lot of knowledge about MATHXXX, does that in general imply ENGYYY is also hard? We may also compare and contrast the difficulty of each category of courses.
3. We may also investigate the difficulty of a series of courses. This would provide students and schools with more information about the course/major structure as a whole. Students are often required to take several courses in a series. By learning about the difficulty of each series of courses, they can better arrange their curriculum. For students who are taking them, it also helps them know when they should put more effort into them.

1. Feature importance:

i) Discuss why we choose our feature. Practical reasons and numerical reasons (e.g. data distribution)?

Ans: For students, GPA is the most important way to measure student performance in this course. Also, GPA is a standardized numerical score, which it is easier to assess and compare the academic qualifications of a large number of students.

Based on the visualization, we find that year and course number are two biggest terms that can affect students’ GPA. Moreover, based on practical experience, scheduled type is also an important factor. Thus, we choose these three factors to be our desired features.

ii) Discuss how we come up with the feature “avg. GPA” or “easy/hard” etc. In particular,

Why the feature we define is of high importance, while the others are not so important.

Why we do not use the letter grades “separately”: “A”, “B-”, … but average them.

Notebook & related discussion:

a) Average gpa is calculated by a letter grade between 0 to 4.

a) Based on our goal, we must ignore the separate letter grades (”A“, “B-”, “C+“, …). Also, as we are focusing on each Math course, we must ignore Subject types -> this reduces to Year, Term, YearTerm, course title, sched. Type and course number.

b) Based on our empirical experience, we can ignore the Course title and sched. Type, even if it shows close relation in the data.

c) So we need to investigate Year, Term, YearTerm, course number.

iii) Data exploration. Show the form of our data.

2. Discuss the result:

i) Theoretical. About the training of models, etc.

Why it is difficult to do validation:

Year & Year term: We only have 13 years, which is inappropriate to do validation on years. Also, for different years, it may have different professors in the same math courses.

Professors: Different professors may have different teaching styles, grading policies, etc. and for EACH course, there are just a few professors teaching it, so it is inappropriate to do validation either.

Course: We are fitting each course. And it is inappropriate to validate the result for one course on another (e.g. how can one use the result for Calc. I to validate Algebra II?). Similar reason for e.g. Sched. type.

ii) Meaning for stakeholders etc. students/profs/boards.

For professors

Make the course more popular

Change their course content based on the model result and their course GPA

For students

Make a choice between scheduled types (e.g. Lecture, Discussion)

Make a decision between similar type courses (e.g. Math 416, Math 415)

Gain higher GPA

For school

Adding sections to popular course to arrange the curriculum

iii) Difference between different activation function.

3. Future works

i) Extend to other subjects.

ii) Add validation to our work.

iii) Analyze difference between different subjects.