IS 670 – Project 2

Andrew Lee

Bernardo Alejandre

Eddy Mendoza

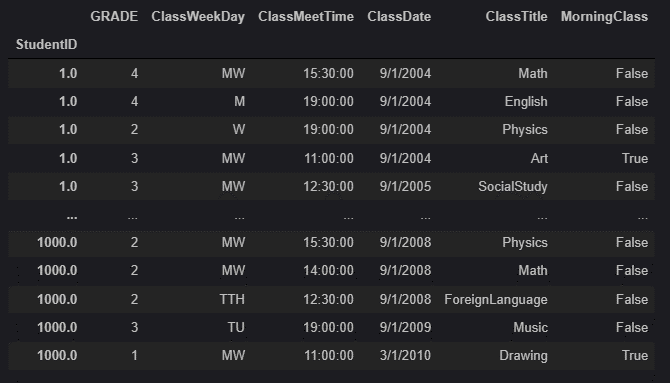
Justin Rinearson

**Question 1**

We want to find out if students perform better in morning classes or afternoon classes, which are after 12:00pm. Since we are comparing two groups of students, we can utilize a t-test to see if the means are equal. We can create the following hypothesis:

μ1 = morning class performance  
 μ2 = afternoon class performance  
 H0 🡪 μ1 = u2  
 HA 🡪 μ1 != μ2

The data needs to be preprocessed in order to get test our hypothesis. All grades need to be converted from alphanumeric to numeric using the scale A = 4, B = 3, C = 2, D = 1, F = 0. Once grades have been converted, we categorized classes as either morning or not morning. This was done by categorizing morning classes as between 00:00:00 and 11:59:59 and all classes after this time are considered not morning. Then we calculated the average GPA by student separated by morning and afternoon classes.



Our sample for each group includes all students since students took both morning and afternoon classes. Once we had the GPA for each student based on morning or not morning classes, we were able to perform the t-test.

First, we checked to see if the variances were equal and if the sample size was equal. Based on the findings below, they were not.

Morning student class variance: **0.464021**

Morning student class variance: **0.316714**

Total number of morning classes taken: **3112**

Total number of afternoon classes taken: **6885**

Morning class averages: **2.76**

Afternoon class averages: **2.78**

We were comparing 3112 classes to 6885 classes. Therefore, we ran a Welch’s t-test to compare the two groups and we resulted in a p-value of 0.49.

*Ttest\_indResult(statistic=-0.6837973690386011, pvalue=****0.4941919621986579****)*

Our p-value indicates there is approximately a 50% chance that we will reject the hypothesis of μ1 = μ2. Because we cannot make a decision off the p-value alone, we calculated the means manually of the two groups. Morning class average GPA is 2.76 and afternoon class average is 2.78. Based on these findings, we can accept the null hypothesis. However, using the t-test alone, it could be based on random chance that the means are so close.

**Question 2**

Physics, math, and statistics are considered quantitative classes. We want to see if students performed better when they took one quantitative class per semester or two or more. To make our world easier to understand, we broke the groups up based on the following criteria. Group 1 contains students who took no more than one quantitative class per semester throughout their entire program. For example, a student in group 1 may have taken four semesters worth of classes, and in each of those semesters, they would have not taken more than one quantitative class. Group 2 contains students who took two or more quantitative classes during at least one semester throughout their entire program. For example, a student in group 2 may have also taken four semesters worth of classes, however, one of those semesters may have had two quantitative classes. Since we are comparing two groups of students, we can utilize a t-test to see if the means are equal. Based on this criterion, we developed the following hypothesis:

μ1 = group 1 (one quantitative class per semester)  
 μ2 = group 2 (two or more quantitative classes during at least one semester)  
 H0 🡪 μ2 >= u1  
 HA 🡪 μ2 < μ1

The data needs to be preprocessed to get test our hypothesis. All grades need to be converted from alphanumeric to numeric using the scale A = 4, B = 3, C = 2, D = 1, F = 0. We also needed to categorize class dates by semesters. We did this by using the method dt.quarter(2) which took the dates from ClassDate, multiplied it by 2, then we assigned it as either semester 1 or semester 2 as shown below.

Graphical user interface, text

Description automatically generated

We then grouped the classes by student ID, semester, and class title then added a new column to signify if the class was a quantitative class or not (true/false).

A screenshot of a computer

Description automatically generated with low confidence

Once we had the quantitative classes identified by student and semester, we dropped all classes that were marked as false since we do not need to consider them in this test.

Text

Description automatically generated with medium confidence

We then counted the number of quantitative classes that were taken per semester by each student. This is how we divided our students into either group 1 or group 2 based on the criterion we defined above.

Text

Description automatically generated with low confidence

We were able to split group 1 and group 2 into two dataframes.

Text

Description automatically generated

Text

Description automatically generated with medium confidence

We then calculated the GPA for each student in each group. Group 1 is on the left and group 2 is on the right.

A picture containing table

Description automatically generatedA picture containing graphical user interface

Description automatically generated

We used Welch’s t-test because the sample sizes are different. Based on the t-test, we found our p-value of 0.0004, which is considered statistically significant.

Ttest\_indResult(statistic=-3.5461771331879848, pvalue=**0.000411516210637409**)

Comparing the two means, we can see that group 1 mean GPA was equal to 2.71 and group 2 mean GPA was equal to 2.86.

Group 1 mean GPA: **2.710573476702508**

Group 2 mean GPA: **2.835453474676089**

Based on this, and our small p value, we can determine that μ2 > μ1 and therefore we do not reject the null hypothesis. Group 2 students, who took two or more quantitative classes during at least one semester throughout the entire program performed better than those students in group 1 who only took at most one quantitative class per semester throughout their entire program.

**Question 3**

We want to find out which major has the highest quantitative GPA. Quantitative classes are math, physics, and statistics. Therefore, we do not want to take other classes into consideration. All grades need to be converted from alphanumeric to numeric using the scale A = 4, B = 3, C = 2, D = 1, F = 0.

We then needed to remove all rows in the dataframe that were not math, physics, and statistics.



We then calculated the mean and grouped by major and GPA. We concluded that the marketing major had the highest GPA when taking quantitative classes only into consideration.

A screenshot of a computer screen

Description automatically generated with low confidence

**Question 4**

We want to see if we can use existing data to predict whether or not students will graduate on time. We cannot use the last two semesters of student data since it will be too late by then to use that data to factor it into whether or not the student will graduate on time. The goal is to use XGBoost for our model, but we must preprocess the data first.

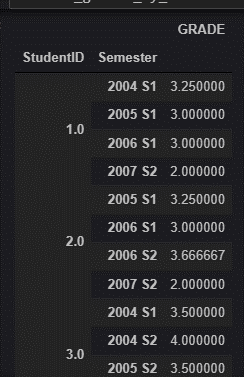
All grades need to be converted from alphanumeric to numeric using the scale A = 4, B = 3, C = 2, D = 1, F = 0. We also needed to categorize class dates by semesters. We did this by using the method dt.quarter(2) which took the dates from ClassDate, multiplied it by 2, then we assigned it as either semester 1 or semester 2 as shown below.

Text

Description automatically generated with medium confidence

(Continue on next page)

We then calculated the student GPAs by semester



We then needed to obtain a dataframe grouped by StudentID and the semesters they took except the last two. This could have probably been done on the existing dataframe, but we chose to use a second one then merge to keep our data organized. We dropped the last two semesters by using .tail(2)

Graphical user interface

Description automatically generated with medium confidence

We then merged the dataframe with the GPAs and the semesters except the last two together.

A screenshot of a computer

Description automatically generated with low confidence

We do not need to have the student grades grouped by semester now and only need their total GPA since the last two semesters are now excluded.

A picture containing chart

Description automatically generated

We then merged the GPA dataframe above to the StudentGraduate dataframe to add CollegeGPA as a new attribute that we can use in our prediction model.

Table

Description automatically generated with medium confidence

Using 25% of the data as testing data, we were able to use XGBoost with gamma=9 to predict graduate on time with an accuracy of 75%, which, when used in the real world, is extremely high.

Text

Description automatically generated