Zhenrong Lao

Professor Robinson

DSBA 6520

August 9, 2021

**Students Performance in Exams**

**Project I—Use Case, Data Model and Projections**

**I. Dataset Use Case**

Nowadays, there are so many resources available to students. For example, there are teachers there to help them in and out of class, their peers can give them a hand, tutors are available for parents to hire, and there are so many free online learning materials. However, not every student can do a good job in high school such as obtaining a high score on their exams even if all of them have chances to do it. In this project, the goal is to understand why they should perform great which still is not the case. To better analyze the influence of various factors on the student’s performance, student data is collected with 8 variables described as follows:

1) gender: the gender of each student (male/female)

2) race: the race of each student (there are three groups: group A, group B, group C)

3) parental level of education: the highest educational qualification of any parent of each student

4) lunch\_type: the type of lunch package selected for each student(standard/reduced)

5) test\_prep: if the test preparation course was completed by the student or not

6) math\_score: score in math (our target variable)

7) reading\_score: score in reading

8) writing\_score: score in writing

More specifically, several questions through graph analytics on this dataset will be answered: How many students pass the test for each test and how many students fail the test for each test? Is there any similarity for students who pass or fail exams? Is test preparation enough for exam pass? What’s the main driver of good test scores? Based on analysis, could we improve the students’ performance in the tests?

**II. Graph Data Model**

Diagram

Description automatically generated

**III. Graph Projections**

Chart, diagram, bubble chart

Description automatically generated

**Project II—Graph DB setup and Application of Algorithms**

**I. Neo4J Database Setup**

Graphical user interface, text, application

Description automatically generated

**II. Cypher Queries**

**Cypher Queries 1:** Those three queries show letter grade distribution of all three exams

* 1. Math letter grade distribution

MATCH (s:MathGroup)

WITH s,SIZE((s)-[:HAS\_MATH\_GRADE]-()) AS math\_degree

RETURN s.math\_letter\_grade AS math\_letter\_grade ,math\_degree

ORDER BY math\_letter\_grade

* 1. Reading letter grade distribution

MATCH (s1:ReadingGroup)

WITH s1,SIZE((s1)-[:HAS\_READING\_GRADE]-()) AS reading\_degree

RETURN s1.reading\_letter\_grade AS reading\_letter\_grade ,reading\_degree

ORDER BY reading\_letter\_grade

* 1. Writing letter grade distribution

MATCH (s2:WritingGroup)

WITH s2,SIZE((s2)-[:HAS\_WRITING\_GRADE]-()) AS writing\_degree

RETURN s2.writing\_letter\_grade AS writing\_letter\_grade ,writing\_degree

ORDER BY writing\_letter\_grade

|  |  |  |  |
| --- | --- | --- | --- |
| **Letter\_grade** | **Math** | **Reading** | **Writing** |
| **A** | 58 | 79 | 78 |
| **B** | 135 | 170 | 157 |
| **C** | 216 | 264 | 254 |
| **D** | 268 | 233 | 230 |
| **F** | 323 | 254 | 281 |

*Table 1. Three course letter grades*

There is total 1000 students in the dataset. We can know from table 1, Out of them, there are 58 students who got A in Math, there are 79 students who got A in reading exam and there are 78 students who got A in Writing exam. In addition, there are more students who failed (got D or F) in Math than those failed in Reading or Writing. Math exam is harder than reading or writing exam.

**Cypher Queries 2**

2.1 Total number of students who completed the test preparation course.

MATCH

(s: Student{test\_preparation\_course:"completed"})

RETURN count(s)

2.2 Number of students who completed the test preparation course and pass the math exam.

MATCH (s1: Student{test\_preparation\_course:"completed"})-[i1: HAS\_MATH\_GRADE]-(m:MathGroup)

WHERE m.math\_letter\_grade="A" OR m.math\_letter\_grade='B' OR m.math\_letter\_grade='C'

RETURN count(s1)

2.3 Number of students who completed the test preparation course and pass the writing exam.

MATCH (s2: Student{test\_preparation\_course:"completed"})-[i2:HAS\_WRITING\_GRADE]-(w:WritingGroup)

WHERE w.writing\_letter\_grade="A" OR w.writing\_letter\_grade ='B' OR w.writing\_letter\_grade='C'

RETURN count(s2)

2.4 Number of students who completed the test preparation course and pass the reading exam.

MATCH (s3: Student{test\_preparation\_course:"completed"})-[i3:HAS\_READING\_GRADE]-(r:ReadingGroup)

WHERE r.reading\_letter\_grade='A' OR r.reading\_letter\_grade='B' OR r.reading\_letter\_grade='C'

RETURN count(s3)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Couse Name | Completed TPR | Math | Reading | Writing |
| Total Student | 358 | 176 | 234 | 238 |

*Table 2. Number of students completed the test preparation course and pass the exam.*

Table2 shows number of students who completed TPR and passed one of exam (got C or above).

There are 358 students who completed test preparation course. Out of them, almost 176 students (49%) passed the math exam. For 642 students who did not complete test preparation course, 233 of them (36%) passed math exams. For reading exam, 65% students who completed test preparation course pass the exam, while only 43% student who did not complete course pass the exam. This indicates that test preparation is quite useful for students to prepare exams.

**Cypher Queries 3:** Number of male students pass the three exams

3.1Number of males who pass the math exam.

MATCH (s1: Student{gender:"male"})- [r1: HAS\_MATH\_GRADE]- (m1: MathGroup)

WHERE m1.math\_letter\_grade<>"D" AND m1.math\_letter\_grade<>"F"

RETURN count(s1) as male\_math\_pass

3.2 Number of males who pass the reading exam.

MATCH (s2: Student{gender:"male"})- [r2: HAS\_READING\_GRADE]- (m2:ReadingGroup)

WHERE m2.reading\_letter\_grade<>"D" AND m2.reading\_letter\_grade<>"F"

RETURN count(s2) as male\_math\_pass

3.3 Number of males who pass the writing exam.

MATCH (s3: Student{gender:"male"})- [r3: HAS\_WRITING\_GRADE]- (m3: WritingGroup)

WHERE m3.writing\_letter\_grade<>"D" AND m3.writing\_letter\_grade<>"F"

RETURN count(s3) as male\_writing\_pass

|  |  |  |  |
| --- | --- | --- | --- |
| Gender | Math | Reading | Writing |
| Male | 231 | 195 | 164 |
| Female | 178 | 318 | 325 |

*Table 3. Number of students about male and females who pass the writing exam*

There are 325 female students who passed writing exam, while only 164 male students who passed writing exam. Also, there are 318 female students who passed reading exam and only 195 male students passed reading exam. In general, female has higher passed rate than male in writing and reading. However, male students perform better than female students in math with higher passing rate.

**III. Graph Algorithms**

**Algorithms 1: Weakly Connected Components: connected components**

As a community detection algorithm, WCC is used to evaluate how groups of nodes are cluster. It can be considered as a preprocessing step for directed graphs, since it helps quickly identify disconnected groups. From componentID, we can easily know which group each node belong to. Let’s use a Cypher projection to build a new graph named student-race-cypher. It will contain people that belong to the same race. There are five race group and largest group has 319 students.

CALL gds.graph.create.cypher(

'student-race-cypher',

'MATCH (n:Student) RETURN id(n) AS id',

'MATCH (p1:Student)-[:BELONG]->(c:Race)<-[:BELONG]-(p2:Student) RETURN id(p1) AS source, id(p2) AS target'

)

CALL gds.wcc.stream('student-race-cypher')

YIELD nodeId, componentId

RETURN componentId AS component, count(nodeId) AS size

ORDER BY size DESC

|  |  |  |
| --- | --- | --- |
|  | **Component** | **Size** |
| **1** | 1 | 319 |
| **2** | 8 | 262 |
| **3** | 0 | 190 |
| **4** | 32 | 140 |
| **5** | 3 | 89 |

*Table 4. The group size about race.*

**Diagram

Description automatically generated**

**Algorithms 2: Louvain**

The Louvain algorithm is a community detection algorithm designed to identify clusters of nodes in a graph. The algorithm is to find student distribution over parental education levels. There are 226 students whose parents have college degree, and only 59 students whose parents have master’s degree.

CALL gds.graph.create('student-PLE','Student','HAS\_SAME\_PLE')

CALL gds.louvain.stream('student-PLE')

YIELD nodeId, communityId

RETURN communityId, COUNT(DISTINCT nodeId) AS size

ORDER BY score DESC

|  |  |  |  |
| --- | --- | --- | --- |
|  | **CommunityId** | **Degree** | **Size** |
| **1** | 638 | some college | 226 |
| **2** | 147 | associate’s degree | 222 |
| **3** | 682 | high school | 196 |
| **4** | 775 | some high school | 179 |
| **5** | 862 | bachelor’s degree | 118 |
| **6** | 567 | master’s degree | 59 |

*Table 5. The size about parental education levels*

Diagram, venn diagram

Description automatically generated

**Algorithms 3: Node Similarity**

The Node Similarity algorithm compares pairs of nodes in a graph based on their connections to other nodes. Two nodes are considered similar if they share many of the same neighbors. Before you run the Node Similarity algorithm, we to create a projected graph that consists of characters and the various entities to which they relate. Table 6 shows students usually have same results in all exams will have highest score in similarity. This allows to identify pair of students who perform similarly.

CALL gds.graph.create('student-grade', ['Student', 'MathGroup', 'ReadingGroup', 'WritingGroup'], '\*')

CALL gds.nodeSimilarity.stream('student-grade',{degreeCutoff: 20})

YIELD node1, node2, similarity

RETURN gds.util.asNode(node1).student\_id AS student1 , gds.util.asNode(node2).student\_id AS student2, similarity

ORDER BY similarity DESC

LIMIT 4

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Student 1** | **Student 2** | **Similarity** |
| **1** | 805 | 284 | 0.9955357142857143 |
| **2** | 284 | 805 | 0.9955357142857143 |
| **3** | 98 | 721 | 0.995475113122172 |
| **4** | 721 | 98 | 0.995475113122172 |

*Table 6. Similarity score for each pair student*

Diagram

Description automatically generated

**Project III – Graph Visualizations**

**Cypher Action 1**

**Search Phrase:** Student who gotwriting\_letter\_grade=$writing\_letter\_grade and reading\_letter\_grade=$reading\_letter\_grade and w3. math\_letter\_grade= $math\_letter\_grade

#### Description: Find students who perform similar in all three exams or students are in the same set of grade letters for all exams

**Code:**

MATCH (p:Student)-[i1:HAS\_WRITING\_GRADE]-(w1:WritingGroup)

MATCH (p:Student)-[i2:HAS\_READING\_GRADE]-(w2:ReadingGroup)

MATCH (p:Student)-[i3:HAS\_MATH\_GRADE]-(w3:MathGroup)

WHERE (w1.writing\_letter\_grade=$writing\_letter\_grade) AND (w2.reading\_letter\_grade=$reading\_letter\_grade) AND (w3.math\_letter\_grade= $math\_letter\_grade)

RETURN p,i1,w1,i2,w2,i3,w3

Diagram

Description automatically generated

*Figure 1. Example about students who got A in writing, B in the reading, and C in math*

**Importance:** This cypher action allows the user to identify students who perform similar in all exams. To improve student’s grades, it’s very important to target student group with similar performance. Usually, different strategies will be used for student groups to improve their grades. Inviting different student groups to let them share learning experience in the area they are good at.

**Cypher Action 2**

**Search Phrase:** Student who took test\_preparation\_course $test\_preparation\_course and got reading\_letter\_grade $reading\_letter\_grade

**Description:** Find Student who perform good at one exam and complete the test preparation course.

**Code:**

MATCH (p:Student)-[i1:HAS\_READING\_GRADE]-(w:ReadingGroup)

WHERE p.test\_preparation\_course= $test\_preparation\_course AND w.reading\_letter\_grade= $reading\_letter\_grade

RETURN p, i1, w

Chart

Description automatically generated

*Figure 2. Student who completed the test preparation course and got A in reading exam*

**Importance**: Test preparation is usually important for students to have a good grade in the exam. This cypher action allows school to find the best students in one exam and complete test preparation. Those students can be invited to share their opinions about if or how test preparation is useful for the course they are good at.

**Cypher Action 3**

**Search Phrase:** Information about student $student\_id

**Description:** Search for student all information given student\_id.

**Code:**

MATCH(r:Race)<-[b:BELONG]-(s:Student)-[b1:HAS]->(p:PrentalLevelOfEducation)

MATCH(m:MathGroup)<-[r1:HAS\_MATH\_GRADE]-(s:Student)-[r2:HAS\_READING\_GRADE]->(m1:ReadingGroup)

MATCH (w:WritingGroup)<-[r3:HAS\_WRITING\_GRADE]-(s:Student)

WHERE s.student\_id=$student\_id

return r,b,s,b1,p,m,r1,r2,m1,w,r3

*Diagram

Description automatically generated*

*Figure 3. Information about student 1.*

**Importance**: From this cypher query, it’s easy to find specific information about one student such as gender, race, parental education level and exam results, etc.

**Cypher Action 4**

**Search Phrase:** Student's parental level of education is $degree and student in $group

**Description:** Find students who has specific parental education level in specific racial group

**Code:**

MATCH (p:PrentalLevelOfEducation{degree:$degree})<-[r2:HAS]-(s:Student)-[r1:BELONG]->(g:Race{group:$group})

RETURN p, r2, s, r1, g

*Diagram

Description automatically generated*

*Figure 4. Students who have high school parental education level are in racial group A*

**Importance**: From this cypher query, it’s easy to check students’ family and racial background.

**Visualizations 1**

**Code:**

MATCH (p:Student)-[i1:HAS\_WRITING\_GRADE]-(w1:WritingGroup)

MATCH (p:Student)-[i2:HAS\_READING\_GRADE]-(w2:ReadingGroup)

MATCH (p:Student)-[i3:HAS\_MATH\_GRADE]-(w3:MathGroup)

WHERE (w1.writing\_letter\_grade="A" AND w2.reading\_letter\_grade="A" AND w3.math\_letter\_grade= "A"AND TEST) OR (w1.writing\_letter\_grade="F" AND w2.reading\_letter\_grade="F" AND w3.math\_letter\_grade= "F")

RETURN p,i1,w1,i2,w2,i3,w3

**A picture containing diagram

Description automatically generatedDiagram

Description automatically generated**

*Figure 5. Comparison of best student group and worst student group*   
Figure 5 shows that there are two screenshots of visualization for size comparison of best student group and worst student group. Left plot is general framework of two groups and right plot show more detail information of one group. In the left plot, the left group contains students who got A in all three exams and right group contains students who got F in all three exams. Those visualization allows to easily identify the best student group and the worst one and to understand size difference between two groups. In general, there are much fewer students in the best group than that in the worst group, which indicates that exams are not easy, and students did not perform well in general.

**Graph Visualizations 2**

**Code:**

MATCH (p:Student) - [a1:HAS]- (p1:PrentalLevelOfEducation)

MATCH (p:Student)-[i1:HAS\_WRITING\_GRADE]->(w1:WritingGroup)

MATCH (p:Student)-[i2:HAS\_READING\_GRADE]-(w2:ReadingGroup)

MATCH (p:Student)-[i3:HAS\_MATH\_GRADE]-(w3:MathGroup)

WHERE

(w1.writing\_letter\_grade="A" AND w2.reading\_letter\_grade="A" AND w3.math\_letter\_grade= "A" AND p1.degree="bachelor's degree") OR

(w1.writing\_letter\_grade="A" AND w2.reading\_letter\_grade="A" AND w3.math\_letter\_grade= "A" AND p1.degree= "master's degree") OR

(w1.writing\_letter\_grade="A" AND w2.reading\_letter\_grade="A" AND w3.math\_letter\_grade= "A" AND p1.degree= "high school")

RETURN p,i1,w1,i2,w2,i3,w3,a1,p1

**Diagram

Description automatically generated**

*Figure 6. Parental education level for the best student group*

Figure 6 shows the screenshot of visualization for checking parental education level for the best student group who got A in all three exams. Basically, there are three parental degrees for students: high school degree, bachelor’s degree, and master’s degree. From this plot, it’s easy to see that there is only student in this group who has parents with high school education and most of best students has parents with bachelor’s degree or higher. Therefore, parental education will directly or indirectly affect students’ performance in the school.

**Graph Visualizations 3**

**Code:**

MATCH (s:Student)-[r1:HAS\_MATH\_GRADE]->(m:MathGroup)

MATCH (s:Student)-[r2:HAS\_WRITING\_GRADE]->(w:WritingGroup)

MATCH (s:Student)-[r3:HAS\_READING\_GRADE]->(r:ReadingGroup)

WHERE m.math\_letter\_grade = w.writing\_letter\_grade=r.reading\_letter\_grade

RETURN \*

**A picture containing accessory, colorful, vector graphics

Description automatically generated**

*Figure 7. Five three exams same grade group (A, B, C, D, F)*

The left graph in Figure 7 shows students who failed in three exams. There are more students in this group than those in other group, which verifies that those exams are very hard, and more students perform poorly in those exams.

**Extra Credit ---“Above and Beyond”**

**Machine Learning Algorithm**

After application of Cypher action to find student group who perform similarly, two groups are identified. One (good student group) includes students who passed all exams (got C or better in all exams), while another (bad student group) includes students who failed all exams (got letter lower than C in all exams). Those two groups are combined and are analyzed to understand which factor will lead student to each group. More specifically, there are three factors: parental level of education, lunch provided, completion of test preparation course; one response variable: which group student belong to (1: student belongs to good student group,0: student belongs to bad student group).

Two machine learning algorithms are applied: One is logistic regression, and another is Random Forest.

Table

Description automatically generated with medium confidence

*Figure 8. Logistic regression results*

Figure 8 shows the results of logistic regression, and it indicates that two variables including lunch\_standard and test\_preparation\_course\_none have significant effect on student’s performance. In detail, provided lunch with positive coefficient will contribute to students’ performance, and no completion of test preparation course with negative coefficient will be harmful to students’ performance, which are consistent with our conjecture.

**Graphical user interface, chart

Description automatically generated**

*Figure 9. Feature importance from random forest*

From results (Figure 9) of Random Forest, it also indicates that two variables including lunch\_standard and test\_preparation\_course\_none are most important factors affecting students’ performance in those exams. Therefore, encouraging students to finish test preparation course and school providing lunch to students are very important to improve students’ performance.

**Summary**

With advent of Internet, students are easily access to many free online learning resources. However, not every student can do a good job in high school such as obtaining a high score on their exams even if all of them have chances to do it. The goal is to understand the influence of the parents’ background, test preparation on students’ performance. The dataset comprises of 1,000 rows and 8 columns, with which we continue to determine what all the features which plays a vital role in affecting the student's performance, and we also try to solve some of the myths for example: 1) Does one gender excel another? 2) Does practice help to excel scores? 3) Does student good at math bad at writing? 4) Is test preparation enough for exam pass? Given answers to those questions, some suggestion will be provided to help student to obtain better results later.

In general, various cipher queries, graph algorithms are applied as tools to answer those questions. More specifically, three cipher queries are used to explore letter grade distribution over, the effect of test preparation on student’s exam performance and the performance of students with different gender. Basically, those exams are not easy for students since many students got F in all of three exams than those got A. Students with preparation usually perform much better than those without any preparation whatever racial group or gender group they belong to. Male students usually are good at math and female students excel at reading and writing areas. In addition, the Weakly Connected Components can be used to find connected components to better understand racial group structure in the network. Another community detection algorithm called Louvain algorithm is designed to identify clusters of nodes in a graph. In this way, the distribution over parental education levels can be easily identified. Through Node Similarity algorithm, it’s easy to figure out pair of students who perform similarly in all of three exams. Those algorithms are allowed to investigate different group sizes about exams performance, race, and prenatal education levels.

Finally, different Cipher actions, graph visualizations are applied to remove irrelevant information from large network. In detail, as a method to create dynamic queries that let user to find specific information they need such as gender, race, parental education level and exam results of one student. Besides that, it’s also crucial to target student group with similar performance by the Cypher action. Based on that, we can apply different strategies to improve student’s performance in the specific group. Finding best students is very important since those students can provide useful learning experience such as role of test preparation course to those who failed one of those exams. The next step is to visualize best student groups, which is subset of the entire graph. It shows parental education level for the best student group who got A in all three exams and verifies that parental education will directly or indirectly affects students’ performance in the school. Two machine learning algorithms are applied to further explore the effect of different factors on students’ performance. Based on results, test preparation course and lunch provided by school are key factors of students’ performance.

In a sum, those exams are not easy and thus need students to work hard on that. Therefore, test preparation course can used to boost student’s performance in those exams. Female students outperform male students in reading and writing exams, but in math scores male students are better than female students. The good parental education will also give students advantage to succeed in those exams. In order to improve students’ performance, it better to invite different student groups to let them share learning experience in the area they are good at. In addition, encouraging students to attend test preparation course and school providing lunch to students are keys for students who would like to succeed in those areas.

**Reference**

Seshapanpu, Jakki. “Students Performance in Exams.” *Kaggle*, 2018, www.kaggle.com/spscientist/students-performance-in-exams.