The Next Best Course Project

DNSC 6279

Team Project

Spring 2015

Prepared by:

WEAAM ALGHERAIBI
RAMAH AL BALAWI
LAWRENCE GADSDEN
VALERIYA MALOBRODSKAYA

TABLE OF CONTENTS

1. PROBLEM STATEMENT

The course registration procedure implemented at GW should aim to minimize time spent on course selection, optimize the use of all available courses at GW, because a substantial amount of valuable University resources go to waste every time a potentially useful and interesting class gets overlooked by a student. This procedure should also be fairly straightforward, so it can be easily understood and correctly applied to each student's customized history.

However, system that is currently implemented at the University is purely "manual" in nature and leaves it completely up to a student to do the course search, which can be time-consuming, especially in case with the course electives. The inefficiency of the current system represents a significant financial burden for the University: GWU averages more than 2300 courses/semester and fair share of them get cancelled due to the lack of registrants.

The concept behind development of the new methodology is similar in nature to the retail industry approach, known as Next Best Offer (NBO). It is widely used for the purpose of referring customers to a subset of products that might be of interest to them, based on their demographics data, as well as shopping history. With the use of data mining techniques, as well as advanced analytics methods, our team will develop an interactive model that will propose relevant courses to the student, based on their attributes.

GWU will benefit substantially from the adoption of the new course registration procedure. Once it is implemented, University administration can use predictions to meet course demand. University will also be able push out the underexposed classes to the right students, thus optimizing its operations. In addition, by being one of the first to adopt this innovative, yet

creative approach, University will gain recognition as an educational pioneer and trend-setter in the academic environment.

Optimization of the current registration system and adoption of new, more efficient practices is crucial for the continued competitiveness of the University. In this proposal, the alternative course registration procedure is analyzed using available data mining techniques.

2. DATA SELECTION

Data for the current project has been obtained from GWU BI division. Our team has received 2 data sets:

- Student and course information
- Prerequisite information

Dataset 1 (Student and course information)

Student info

Name	Description
STUDENT_UNIQUE_IDENTIFIER	Student unique ID
REGISTER_TERM_CODE	Student's registration term code
REGISTER_TERM_DESC	Registration term
STUDENT_COLLEGE_GROUP_CODE	Student's school affiliation code
STUDENT_COLLEGE_GROUP_DESC	Student's school
STUDENT_DEGREE_CODE	Student's degree code
STUDENT_DEGREE_DESC	Student's degree
STUDENT_MAJOR_CODE	Student's major code
STUDENT_MAJOR_DESC	Student's major description
STUDENT_MINOR_CODE	Student's minor code

STUDENT_MINOR_DESC	Student's minor
STUDENT_LEVEL_CODE	Student's level code
STUDENT_LEVEL_DESC	Student's level
STUDENT_TYPE_CODE	Student's type code
STUDENT_TYPE_DESC	Student's type
STUDENT_CAMPUS_GROUP_CODE	Student's campus code
STUDENT_CAMPUS_GROUP_DESC	Student's campus

Course info

Name	Description
COURSE_NUMBER	Course Number
COURSE	Course Name
COURSE_LEVEL	Course level
COURSE_COLLEGE_GROUP_CODE	School code
COURSE_COLLEGE_GROUP_DESC	School name (GWU has 12 different colleges)
COURSE_CAMPUS_GROUP_CODE	Campus code
COURSE_CAMPUS_GROUP_DESC	Campus name
COURSE_STATUS_CODE	Status code
COURSE_STATUS_DESC	Status description
COURSE_DEPARTMENT_CODE	Department code
COURSE_DEPARTMENT_DESC	Department description
COURSE_SCHEDULE_TYPE_CODE	Schedule code
COURSE_SCHEDULE_TYPE_DESC	Schedule description
COURSE_SUBJECT_CODE	Subject code
COURSE_SUBJECT_DESC	Subject Description

CREDIT_HOURS	Number of credit hours

Dataset 2 (Prerequisite information)

PREREQ_EFFECTIVE_TERM	Term that a prerequisite came into effect
COURSE_SUBJECT_CODE	Subject code
COURSE_NUMBER	Course number
PREREQ	Prerequisites for the course

3. EXPLORE AND VISUALIZE

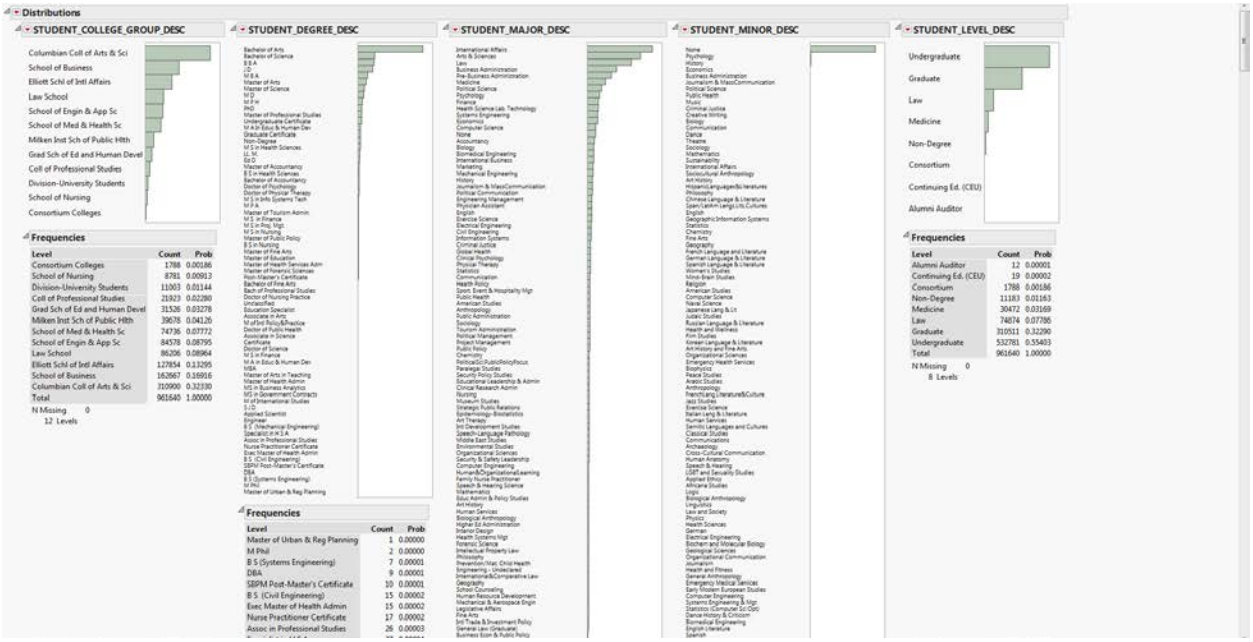
Dataset 1 (Student and course information)

SAS JMP Pro 11 and Tableau 8.0 were used for exploratory analysis of the original dataset. Below are the results of our analysis on GW student population:

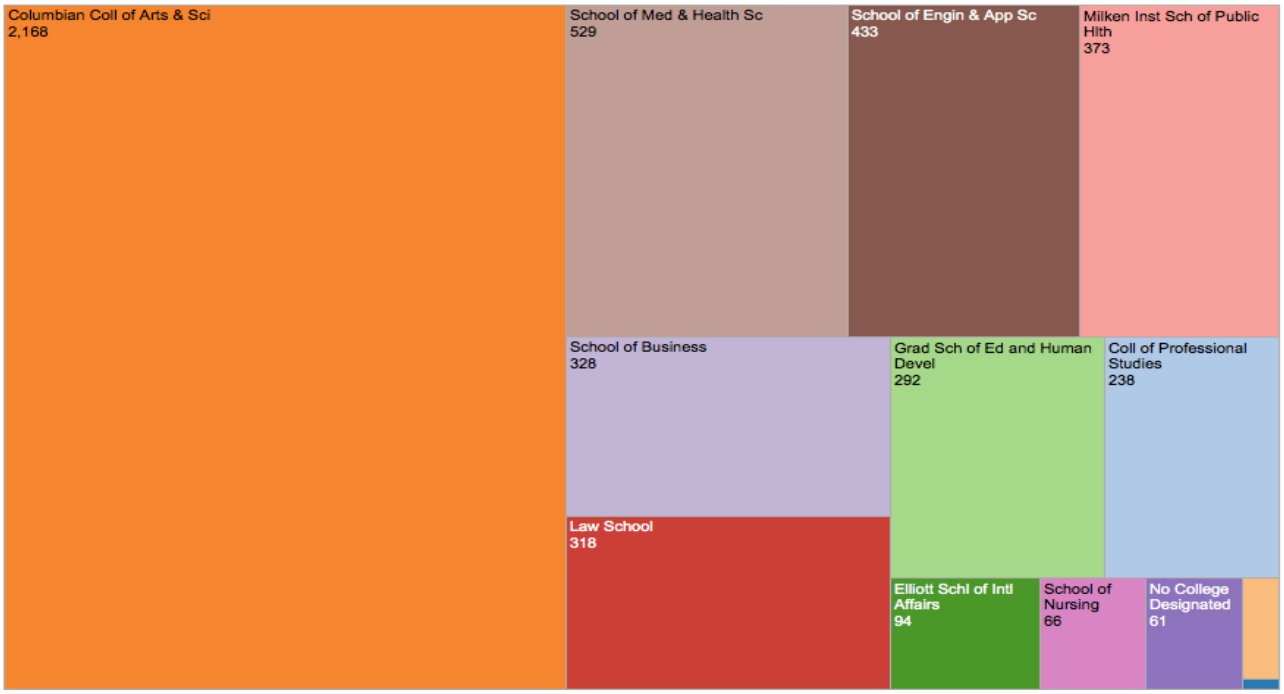
55.4% of the students are pursuing degrees at the undergraduate level, and 32.30% of the students are graduates. The rest of the students are either studying Law or Medicine, with 7.79% and 3.17% respectively.

The majority of students are pursuing a Bachelors of Arts degree and are affiliated with the Columbian College of Arts and Sciences. The most popular major is International Affairs, offered by the Eliot School of International Affairs.

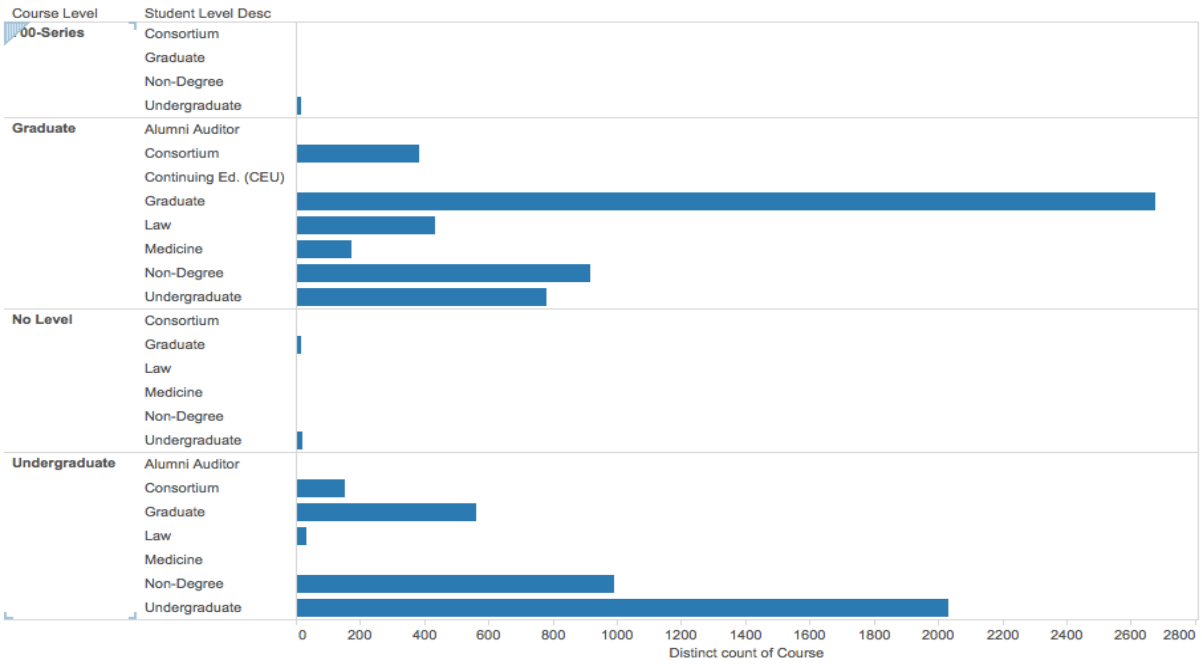
91.5% of the students do not have a minor. Out of 8.5% of the student with a minor Psychology and History are most popular, with 0.06% and 0.05% respectively.



The majority of the students are registered in one of the 2,168 courses offered by the College of Arts and Sciences, which is almost half of the courses offered in this dataset. The number of courses offered by the other schools ranges between 66 in the School of Nursing and 529 courses in the School of Medicine and Health Sciences.

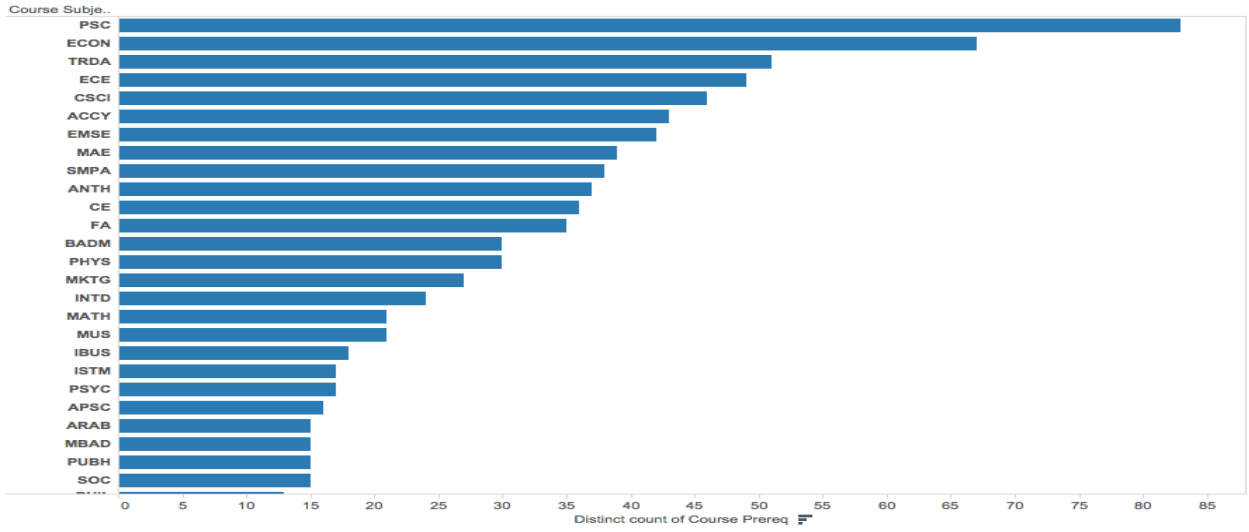


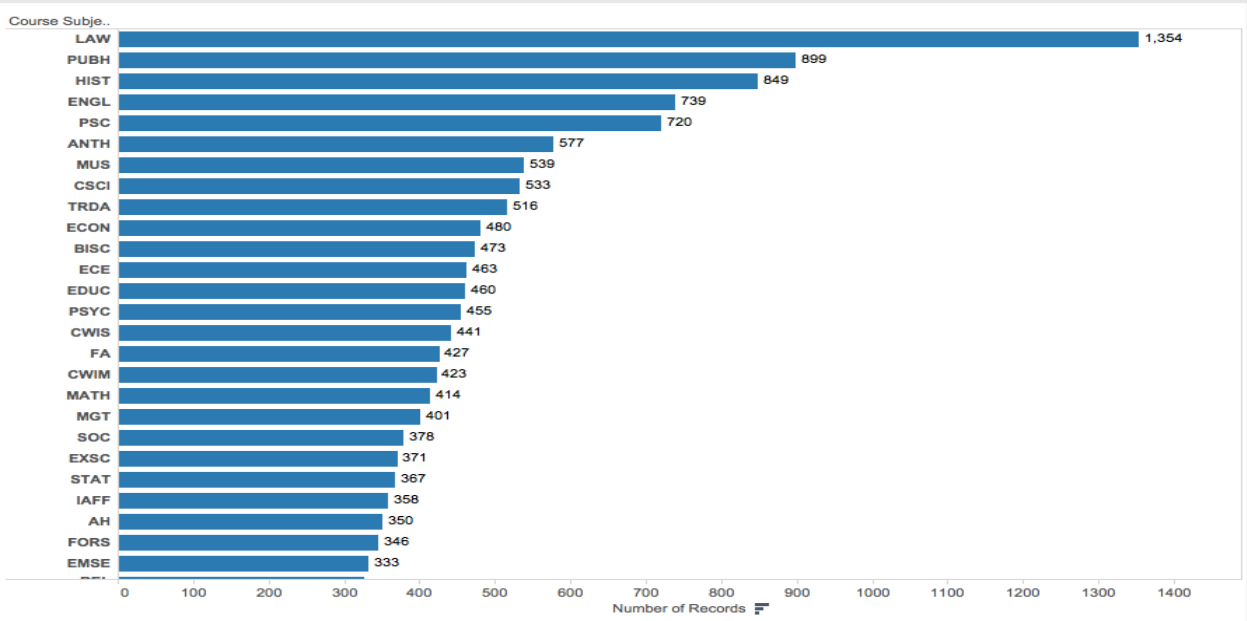
The majority of students are pursuing a degree at the Graduate or Undergraduate level.



Dataset 2 (Prerequisite information)

For the prerequisite information dataset, exploratory analysis shows that not all courses require prerequisite courses. The highest number of courses that do not require courses are those where Law is the subject. On the other hand, the highest number of courses that require prerequisites are those where Psychology and Economics are the subjects.





The Prerequisite data is organized in a way that the course is on the right hand side, and prerequisite courses are on the right hand side. The prerequisite course list can vary from being only one course to being many courses. Many of the prerequisites contained “and” and “or” rules (as seen below).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	PREREQ_EFF	COURSE_S	COURSE_NUMI	COURSE_PREREQ																		
2	201203	ECON	3190	((ECON-1011 or ECON-011) and (ECON-1012 or ECON-012)) and (ECON-2101 or ECON-101 or ECON-2103 or ECON-103) or HONR-2044 or HONR-044 or HONR-2043 or HONR-043																		
3	201203	ECON	3191	((ECON-1011 or ECON-011) and (ECON-1012 or ECON-012)) and (MATH-1221 or MATH-1231 or MATH-1252 or MATH-021 or MATH-031 or MATH-052) or HONR-2044 or HONR-044 or HONR-2043 or HONR-043																		
4	201003	ACCY	2002	(ACCY-051 or ACCY-2001)																		

4. DATA PREPROCESSING

Below is the data preprocessing that took place. The preprocessing has been performed using R software.

- The course number, department abbreviation were merged to create a single new column

- Courses containing the old numbering for the Spring 2010 were found and changed to new course numbers[1]
- Commas were removed from course names because they were mistaken for separators when extracting association rules
- Registration number has been created for all the semesters, in order to be able to organize them in an increasing order
- "MBA" in student_desc changed to "M B A" so that MBA becomes one degree
- Prerequisite course file has been also altered in order for it to be suitable for the analysis

5. MINING TECHNIQUE SELECTION

5. 1 Why Sequence Association Mining?

Sequence Association Mining is the data mining technique selected for analyzing the problem at hand. The decision behind selecting this particular method has been based on the number of facts stemming from the nature of the data provided to us.

Course and student information dataset had a large number of attributes. Inputting this data into a decision tree or logistic regression would cause a large number of rules, as well as large number of attributes, which is suboptimal in this case and would make it hard to interpret the output in case of the decision tree. Timeline was also of importance in this particular project, i.e. chronological order of student's course history. Decision tree and logistic regression, as well as association analysis do not offer this option.

[1] Scraped http://my.gwu.edu/mod/pws/courserenumbering.cfm for course renumbering data then searched for and changed course numbers.

In the search of the best method, our team had also considered a popular technique used in the recommendation analysis, called Collaborative Filtering. This methodology offers one-to-one associations, which is too simple for our purposes. It also disregards the notion of an "item set", i.e. classes that get picked together by a student in one semester. As mentioned above, chronology is vital for this project, and method of Collaborative Filtering doesn't provide us with this opportunity. Considering the requirements of this project and caveats of the aforementioned techniques, our team has adopted Sequence Association Mining as the choice of method for this assignment.

Before proceeding further, it is important to discuss the shortcomings of this approach. Sequence Association is primarily used in an exploratory analysis, and our team to score and extract the rules so that we could make recommendations. Furthermore, our team had run into an issue of automating support and confidence selection for this method since each subset of classes varies by the student's major (more popular major have more classes in the basket, as opposed to the less "popular" major that have fewer classes). While the support levels often fall within a small range, some levels may be too high to create rules and some may be so low that the an everyday computer does not have the power to process it.

5. 2 Why R?

The choice of Sequence Association went hand-in-hand with the choice of the software for this analysis. Our team has made a decision to perform the analysis in R, as opposed to the SAS Enterprise Miner. SAS Miner, being a useful tool, however is poorly suited for our purposes, because of it lacked the flexibility to do many of the things that we envisioned for this project. It allows very few manipulations with the data (data-preprocessing), as well as very rigid filtering

techniques. During our analysis, we needed an extensive use of data sub-setting and pre-processing, and were able to accomplish these steps in R because of the access to its wide range of libraries and methods.

5.3 Code Structure

The logic behind our code is the following:

1. We randomly pick a transaction and extract student's ID associated with it. This will allow us to make a prediction on the randomly selected individual.

2. We randomly pick a semester for which we are going to make the prediction for this student.

3. From the student ID, we extract the following attributes: school, major, as well as student level (graduate/undergraduate).

4. Using the above attributes, we subset our data to have only students who match the attributes of the student we have randomly selected. It makes the most logical sense to predict classes from the pool of the students that have the most in common with the selected student.

5. The nest step in our analysis is to subset the matching pool of student by the students who took the same courses as our random student. After this process, we are left with the list of courses that were taken by the students with same attributes as our student, as well as students who have at least one course in common with our selected individual.

6. We proceed to convert that data into a transactional data with sequence information:

Student Ids = id

Courses = item

Reg Values = sequence values (1-9 correlates with semesters)

7. We input our subset into cspade method, which is an implementation of Sequence Association mining technique.

8. We select support and confidence, as a function of the size of the subset of classes, that varies by the degree of an individual.

9. This method generates rule in chronological format LHS -> RHS.

10. We further subset the LHS of the rules to ensure that all of them contain at least one of our student's courses from previous two semesters.

11. We impose a restriction on RHS, so that it cannot contain courses that have already been taken by our student, as well as the "leave of absence" category.

12. We proceed to remove courses from RHS that are not offered in the current semester, by checking against all the pool of courses in the selected semester.

13. By multiplying the support and confidence values of each rule, we proceed to rank the rules in terms of the highest support and confidence.

14. We choose up to first 5 courses that received the highest rank and check them against the actual courses our randomly selected student in step 1 has taken.

15. We calculate the accuracy measure, as the percentage of the courses that were recommended that also coincided with the actual courses taken by our student.

16. We run the recommended courses against the prerequisite requirement file to determine the courses that require prerequisites

17. We match the prerequisite courses with student history to determine if the student has fulfilled the requirements for the selected courses

18. We determine which courses can't be taken, because a student didn't fulfill all of the requirements

5. 4 Analysis Limitations

In this section we would like to address the limitations that were imposed on the analysis by either the data or the shortcomings of our model.

One of the problems with the model described above is a cold start problem. More specifically, we are only able to make a prediction for a student, who has a course history prior to the semester that has been selected for the prediction. Student population also has to have course history prior to the selected semester.

Another issue is that our model is not able to predict new courses that are offered by GW, because there is not yet a relationship linking the student course historical data and the new courses.

Departments sometimes slightly alter the course name and this can also pose a problem. Our model would treat this course as a completely different class and wouldn't produce matches with the old course name, therefore not producing correct predictions.

Below is the description of possible analysis venues that we could have taken, if the data we have received would contain the following information:

1. Student classification: 1st year undergraduate, 2nd year undergraduate, 2nd year graduate, etc

2. Student grades, as a proxy for course preference. Having this information would have made our predictions more tailored to the student's interests and abilities.

3. Student degree requirements for each major. Our team has attempted to gather data for the major requirements, in order to be able to make more specific recommendations. However, due to the large number of majors at GW, this process would surpass the scope of this project.

6. DATA MINING RESULTS

The data mining results can vary depending on the student and the semester that is chosen. For this demonstration, we allowed R to randomly select a student and a semester to be used for recommendations. The ending results are below.

```
total elapsed time: 5.65s

These courses are recommended for student 812965 for spring 2012:
[1] "PUBH 6014: Practicum"                    "PUBH 6402: Washington Seminar"
[3] "PUBH 6015: Culminating Experience"       "PUBH 6005: Policy Approaches/PublicHealth"
[5] "PUBH 6435: GH Prog Dev & Implementation"

These recommendations were correct for the semester:[1] "PUBH 6402: Washington Seminar"
The recommendations were 0.3333333 accurate for this semester.

The recommendations were 0.6 accurate if semesters up to spring 2013 are included.
These recommendations were correct when inlcuding spring 2013 :
[1] "PUBH 6402: Washington Seminar"     "PUBH 6014: Practicum"
[3] "PUBH 6015: Culminating Experience"
> |
```

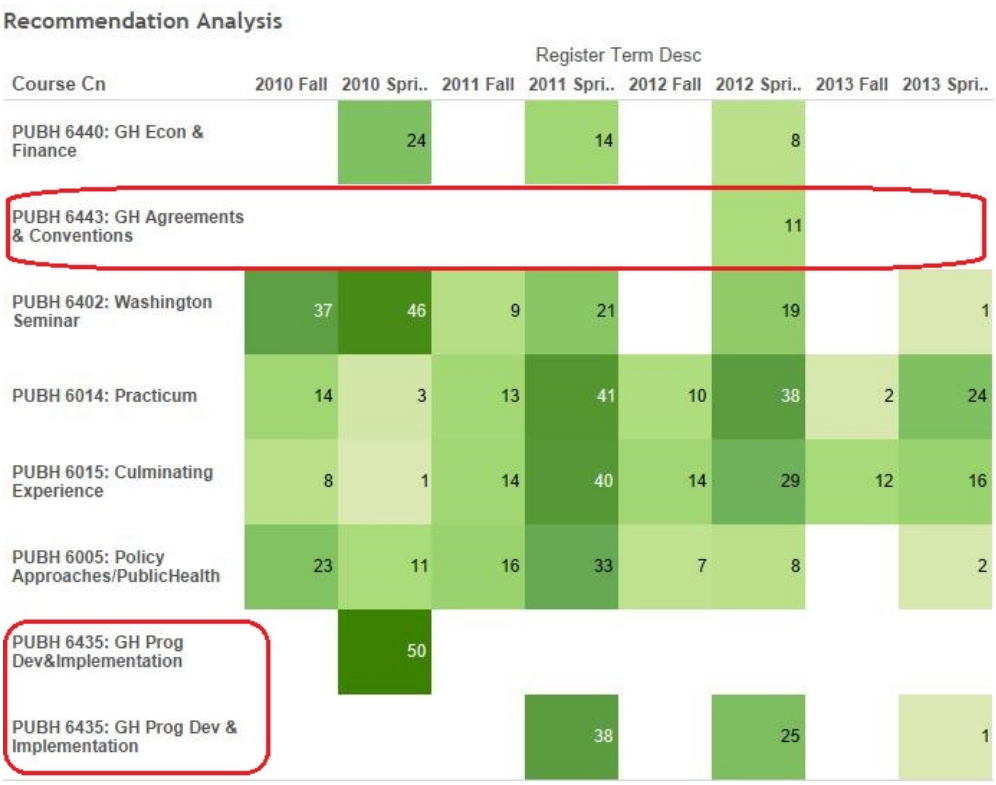
The results show five courses that student 812965 may be interested in for spring 2012 (based on the historical data of other similar students). The student took three courses during the semester, with the student choosing one of the recommended courses. The results show that the *correct* recommendation was PUBH 6402: Washington Seminar. The results also generate an accuracy measure by dividing the amount of correct recommendations by the amount of courses that the student took during the semester. In this demonstration, the accuracy percentage is 0.33.

Below is the result of checking the fulfillment of the prerequisites of student ID 1290936 against the prerequisite document:

```
> source( ~/.active-rstudio-document )
[1] "Checking requirements for:"
[1] "PSYC 3154: Psychology of Crime & Violence" "PSYC 3122: Cognitive Neuroscience"
All requirements are fulfilled for:
PSYC 3154: Psychology of Crime & Violence
All requirements are fulfilled for:
PSYC 3122: Cognitive Neuroscience
>
```

After analyzing the results of numerous students, a pattern was noticed. Many times a student would not select a recommended course for a selected semester, but later select the course in a following semester. Students may have countless reasons to take a course in later semesters. Thus, the results were updated to include the accuracy of the recommended courses for the rest of a student's course history (up to spring 2014). This accuracy is calculated differently in that it measures the amount of recommendations selected by the total amount of recommendations. For the demonstration, the accuracy is .60. The *correct* recommendations include: PUBH 6402: Washington Seminar, PUBH 6014: Practicum, and PUBH 6015: Culminating Experience.

As a result of the data and model chosen, there are some complications with the results.



During spring 2012, the selected student took PUBH 6443: GH Agreements & Conventions. While the results, just shows that the recommendation was not correct, a further look shows that the course is new and that it was offered for the first time that semester. Sequence mining builds its rules based on historical data, so it would not have been for a recommendation for this course to be generated. Thus, it can be argued that the accuracy could be slightly higher to account for the new course.

Furthermore, the engine recommended “PUBH 6435: GH Prog Dev & Implementation”. When examining the recommendation, one notices that there is another instance of the course as “PUBH 6435: GH Prog Dev&Implementation”. To most people the difference may seem trivial, but small changes in course names is a common occurrence in the data. While it is not debilitating, these small changes sometimes trick the engine into thinking that a highly recommended course is not available for the semester (so that the engine skips it), and it also

disregard the course as a match for later semesters if the name change occurred following the semester.

Altogether, the engine seems to have a reasonable amount of accuracy. Given that students take core courses more frequently than electives, the recommendations are often biased towards courses that meet a major or degree requirement. Nonetheless, as the G-PAC requirements show, students can take different paths when fulfilling their core courses so recommendations can be helpful. Additionally, it is able to find rules and generate recommendations for elective courses (especially if there is a strong relationship).

Implementing degree and major information, student year, and grading information can help this product perform better in the future. Also, it may be helpful to split the way that the recommendations work so that one focuses on recommending core courses, while the other focuses on recommending electives.