University of Diploma Printing: Department of Arts Reconstruction

Thomas Jeffrey Berdahl
Department of Data Analytics
Penn State University
University Park, PA 16802
tjb6126@psu.edu

***Abstract*—Uses the CRISP-DM process to restructure the Department of Arts at the University of Diploma Printing. Frequent Itemset Mining is used with course log data to complete objectives presented by departmental executives.**

Keywords—data mining, data cleaning, data pre-processing

I. OBJECTIVES & OVERVIEW

A. OVERVIEW

The College of Arts and Sciences (CAS) at the University of Diploma Printing is experiencing rapid growth. Within this College, the Department of Arts consists of twelve full-time faculty six adjunct (part-time) instructors, and approximately fifteen graduate assistants. The staff is accredited with many well-established, and renowned professional artists and designers with an array of exhibitions, commissions, and awards to their names. The Department currently has 240 undergraduate majors within its Bachelor of Arts, and Bachelor of Fine Arts programs with 46 different courses ranging from Art and Religion to Freshwater Ecology.

B. OBJECTIVES

As a result of stagnant enrollment over the last five years, changes in student demographics and interests, and shifts in disciplinary approaches, the curriculum for the Department of Arts will undergo reconstruction to be more accurately aligned with the goals of undergraduate students. The main tasks of the redesign have been strategically established by department executives, and are listed as follows:

- Reducing number of credit hours from 39 to 33.
- Defining concentration areas for elective courses.
- Reducing the number of offered elective courses.

For the Department of Art reconstruction operation to be a success, the tasks above must be satisfied. Additionally, the following questions must be answered.

- Is there correlation between classes in which a student takes?
- What classes are the most popular & least popular in terms of enrollment frequency?
- What trends can be discovered that are beneficial to the decision-making process?

II. DATA UNDERSTANDING

A. Dataset Overview

Two datasets were acquired by the Department to help solve the task presented by management. The first dataset "Course History" contains records of all student enrollments for the past five years (from 2001 to 2005)[1]. Each entry has the students full name, course(s) taken, and the semester each course was taken. The second dataset "Offered Courses" lists all the currently offered courses including the course name and course number sectioned by the type of course (foundational, core, or elective).

The dataset contains a total of 3 columns and 4918 rows with column headers "Student name", "semester new" and "coursename" in the top row.

B. Data Understanding

To better understand the dataset at hand, I created a few visualizations in Tableau. Using a bar chart with the course name on the x-axis and the course frequency on the y-axis, it's evident that many courses, approximately 1/3, have an enrollment frequency of less than 3 students over the span of 5 years. Additionally, the courses with the top 7 greatest enrollment frequencies had a mean enrollment frequency of 343 over the 5 years the data represents, which is significantly above the mean of 28.9, and represents 48.93% of total enrollments. However, we must note that significant inferences can't be made from this data as data cleaning has not yet been performed. The purpose of these visualizations is to receive a preliminary understanding of our data, and to discern data cleaning needs. These exploratory visualizations will also help identify variables in the data that can be removed due to irrelevance.

Further evaluation of the "Course History" dataset will occur during the pre-processing phase before we commence our modeling.

III. DATA PRE-PROCESSING

A. Pre-Processing Overview

The data preprocessing stage is imperative to getting an accurate output, used to make critical business decisions. I'll be extensively evaluating the data to ensure all necessary pre-processing is completed. This process will begin with data cleaning. The main objectives of data cleaning are to transform the raw data into a dataset free of any error, that can be used in our model for data mining. Any syntax and spelling errors will need to be corrected, irrelevant and duplicated values removed, missing values filled in or removed, and all other problematic outliers or structural errors will have to be fixed. A final validation will be completed to ensure the dataset is ready for our model.

B. Data Cleaning: Syntax Errors

Sorting 'coursename' by ascending alphabetical order in a table reveals multiple variated spellings for some courses. This error completely tarnishes any data mining attempts, such as the brief statistics mentioned in the previous section. Various syntax errors involved with inconsistency in course names and were identified and corrected to ensure frequency measures can later occur without error. For example, the course "Business German: A Micro Perspective" was tabulated in other ways such as: "Business German – A Micro Perspective" and "Business German A Micro Perspective". As is, these entries would be counted as separate courses, skewing future tabulations for course frequency. All instances of inconsistency among course entries were identified and altered to have uniform course names. Another example of the many syntax errors, the course "CELL. BIOL. & BIOCHEM." had many alternative entries including, but not limited to the following:

- CELL and BIO and BIOCHEMISTRY
- CELL BIOL & BIOCHEM
- CEL BIO BIOCHEMISTRY
- CEL and BIO and BIOCHEMISTRY
- CELL BIOLOGY and BIOCHEM

Additionally, inconsistencies with student names were discovered and corrected. For example, student XYZ (student name is disclosed and remained confidential for privacy purposes), has two variations. One with the suffix of "1995" and another with a suffix of "1999". This significance of these suffixes is unknown and unexplained by the dataset. As a result, we will assume this is the same student and both will be changed to "XYZ" without the respective suffix of 1995 or 1999.

Another example error discovered was with student ABC. There were two different spellings of the last name, with consistent spellings of the first name. The correct spelling was identified and corrected.

Additional errors were caught but identified as non-critical errors that will not harm data mining processes. A student's first name was likely entered incorrectly, however, there is not a closely matched first & last name combination that would make this error significant. There were many similar instances of a prefix character in a first name that likely does not belong, but with all instances including the errant character, the data is not negatively impacted. Take for example, John Smith (a pseudonym is used to protect the identity of students). While John Smith is the logical correct name, the "Courses History" dataset may non-intentionally contain the name DJohn Smith, but because there is no closely matched name in the dataset, this error is deemed insignificant.

C. Data Cleaning: Duplicates

Many duplicated data entries (1200+) were discovered and deleted from the "Course History" dataset. Duplicate data inhibits successful data reporting and can have critical impacts on the business decision making process. These duplicates implied that many students took courses more than one time. For example, student XYZ is listed as taking CELL. BIO. & BIOCHEM. twice during the Spring 2005 semester, which is not possible. All duplicates have been removed.

Additionally, there were instances of students taking the same course in different semesters. For example, student XYZ is listed for taking "American Health Policy" in the Summer 2002 & Spring 2004 semesters. There is no indicator in the dataset telling us if the student completed the course. Multiple scenarios exists where a student may have withdrawn from a course without completion, and then later re-enrolling in the same course. Without proper evidence to back up this claim, these instances (45) will also be labeled duplicates and removed from the dataset.

Another possibly problematic error was observed. There were two student names that are identically pronounced but with slightly different spelling, but as the course records and dates differ between the two, I'll assume they are different entities, and not duplicates.

D. Data Cleaning: Blank Values

A small number of missing values (3) were discovered in the "Course History" dataset and were promptly removed.

E. Data Cleaning: Insignificant Courses

Among all the courses listed in the "Course History" dataset, many are courses that are currently not offered by the program. 97 total courses have been identified as courses currently not offered and were promptly removed.

Over 1400 entries were discovered and identified as irrelevant. Additionally, the concentration areas are created based on elective courses, the data for core curriculum and foundational courses will be excluded for the frequent itemset model to limit the frequent item sets to relevant courses. Before

the data cleaning process, there were 4,918 data entries for student enrollments over the past 5 years. After removing all blank entries, duplicates and irrelevant courses, the dataset contained 2,157 data entries, comprised of 37 unique courses taken by 429 unique students.

F. Constructing & Formatting Data

To perform frequent itemset mining, the data needed to be transposed to work with the Apriori Algorithm in Python. Using Microsoft Excel, I used the "transpose", "unique", and "filter" formulas to transpose the courses from columns to rows, with a unique student name value on each row. Additionally, the semester data entries were removed as the order of classes is irrelevant in this application. As the specific student's name does not matter, and each row at this point signifies a different student, the column with student name data has been removed.

The data has now been properly formatted for frequent itemset mining to occur and the csv file will be imported into a Python project.

IV. MODELING

A. Model 1

The main objective of our modeling is to gather enough insight to complete the business tasks of reducing the number of credit hours from 39 to 33, defining concentration areas for elective courses and reducing the number of offered elective courses. To discover patterns in the classes students take, Association Rule Mining, specifically frequent itemset mining will be performed with the Apriori Algorithm. The Apriori Algorithm utilizes candidate generation where frequent subsets are extended one at a time. Frequent itemset mining is a data mining technique that is the first step in association rule mining. It's typically used to perform a market basket analysis (MBA), but can also prove useful for instances like this where the goal is to discover items that are associated with each other, and in our case, courses that are frequently taken together by students at the University of Diploma Printing in the Department of Arts.

This algorithm is simple and effective when used on smaller datasets like in this scenario and will provide us with an unlimited amount of itemsets as long and they meet one condition, the support. Support with our dataset represents the number of students that have enrolled in the course(s) listed in the itemset. Support is typically displayed as a decimal but can be transformed into a percentage by multiplying the decimal by 100.

Utilizing Python, with the "Pandas" & "Mlxtend" extension libraries we can complete the necessary frequent itemset mining. Mlxtend is a Python library containing data science tools, including an Apriori Algorithm library that can be used to discover frequent item sets and association rules [2]. The Pandas library will be used to import our csv file and to perform other data manipulation tasks needed to complete our modeling [3].

The main goals of our modeling are to determine the support of every elective course to establish concentration options, and to terminate irrelevant or unpopular courses. The minimum support threshold will be 0.025 or 2.5%. With 296 unique students on record for enrolling in any one elective course, a 2.5% minimum support threshold equates to 7.4 students. This means that at least 7 students must have taken the course for it to be considered in the model. We round 7.4 down to 14 because having a fraction of a student is not possible. In addition to being excluded from our model, any elective course that does not meet the minimum support threshold will no longer be offered.

B. Model 2

The need for a second model surfaced during the evaluation of the first model. I raised the assumption that the data input into Model 1 was misleading and skewed because all students who took at least 1 elective course were included, meaning a student who only enrolled in 16.7% of the required elective courses were included. I chose to create a second model, mirroring Model 1 with the Apriori Algorithm for frequent itemset mining, except exclusively with students who have enrolled in 4 or more elective courses. Not only does this take the upper half (students who have taken 4 – 6 electives), but more importantly represents the new department requirement of 4 elective courses (12 credit hours). This model is a more closely aligned with the goal of the department. This dataset featured 53 unique student's course records.

V. EVALUATION

A. Limitations

The dataset only containing three columns of data (course name, semester, and student name) limits the potential of the data mining process. Metadata for each course and student would've provided useful insights when creating concentrations with elective courses. Numerous types of metadata could have been included such as the student's major, gender, and student class (freshmen, sophomore, junior and senior).

Using exclusively the supplied data influenced a decision that was very one faceted. There was no use of qualitative data from students or professors including their comments on the courses or the program. The decision was strictly made from quantitative data mining. It may be the case that students disapproving of the course offerings is a popular opinion. While the data implies that a large frequency (39.52%) of all students who took at least 1 elective over the five years the data covers took American Health Policy, this does not tell us that students have a positive perception of the course, it just implies that it's likely the preferred option for many students relative to the other elective course offerings. Access to additional resources, would've been very beneficial to the restructuring of the Department of Arts.

Additionally, the specific majors offered by the Department of Arts were not specified, and subsequently there was no column including data of any students major. This complicated the process of establishing course concentrations. With this data it would've been easier to speculate what concentrations may be of interest to the students correlating to their major. The different majors could've been included in the frequent itemset mining and would've told us if students of a particular major enrolled in a specific class at a high frequency. We could've inferred that a grouping of elective courses as a concentration made sense from the association rules. For example, we may have concluded, if a student is a World History major then there is a 65% chance (support), they will take Early Balcan History/Society and Early Mesopotam History/Society as electives. We could then formulate the concentrations from the association rules of the antecedents (majors) with the highest support.

Another aspect missing absent from the dataset is data regarding the capacity for each course, e.g., if a specific course is taken more frequently than historical data shows as a result of removing other courses, will the Department of Arts have the necessary faculty to instruct the course? Are there faculty members that meet the requirements to teach said course? These questions likely succeed the final decisions by the Department of Arts, e.g., the department might decide to hire or train existing faculty to meet the needs of the program based on their final decision.

As the data involves all 296 unique students who took at least 1 elective course, the students who have not yet completed the current 6 elective course requirement are included in this dataset. The data would be less skewed if a threshold was established for number of electives taken e.g., all students who have taken less than 3 electives could have been excluded from the data mining. 82.1% of 296 students included in the dataset took 1-3 electives. To combat this issue, I'll perform another instance of frequent itemset mining with only the 53 (17.9%) students who took 4 or more electives and will compare the results with the 1st frequent itemset mining. For more information about the data mining process for this second model, refer to "IV. MODELING" section B "Model 2". The comparison of the two frequent itemset algorithms is in the next section "Results & Findings".

B. Results & Findings

The main objectives of the modeling are to reduce the number of credit hours from 39 to 33, defining concentration areas for elective courses and reducing the number of offered elective courses. With the requirement of 39 total credit hours reduced to 33, this would alter the elective course requirement from 6 to 4 courses.

1) Model 1

The results of first model of frequent itemset mining indicates there's no 4-itemsets of courses frequently taken together as they failed to surpass the 2.5% minimum support threshold. However, there was one 3-itemset that reached the threshold with a support of 4.39%, {American Health Policy, Contemporary Pol.Thought, and Freshwater Ecology). This can be interpreted as 4.39% of all students who have taken at least 1 elective course. There was a total of 23 1-itemsets in this model that exceed the 2.5% minimum support threshold.

2) Model 2

As highlighted in the previous section, the evaluation process for Model 1 revealed a need to build another model to test my claim that the students who have enrolled in 1 – 3 courses skewed the data. In Model 2, {'American Health Policy'} received the highest support with 60.38%. There were 24 1-itemsets in this model that exceed the 2.5% minimum support threshold.

3) Comparing Model 1 & Model 2

When comparing Model 1 (all students who have taken at least 1 elective course) with Model 2 (all students who have taken 4 or more elective courses), the first thing that I noticed is the difference in amount of frequent itemsets that surpassed the uniform 2.5% minimum support threshold. Model 1 had 42 total frequent itemsets while model 2 had 174. This tells us that the 82.1% of students in Model 1 who took 1 – 3 elective courses were skewing the support in a negative direction. To further compare the two models, their outputs have been exported as CSV files to be modulated in Excel for further manipulation to create a visualization to compare the two models. Each CSV file was imported into its own sheet, and I added a third column "Model #" to differentiate the models in the next stage. The tables now had three columns: "Model #", "Support %", and "itemsets". This table displayed all itemsets from the two models sorted by support % in descending order. I then copied all the itemsets into a new spreadsheet. Then, I removed all duplicates (27) and was left with 189 unique itemsets. Next, I used the "XLOOKUP" formula and referenced the previous spreadsheet to input the support values for Models 1 & 2 for each itemset into separate columns. My table now had the following three columns: "Itemset", "Model 1 support" and "Model 2 support". To enhance readability, I used conditional formatting on the support values in the two columns that changed the color of each cell depending on the value. Higher support values were green fading down to yellow and then to red as the support values reached the minimum support threshold of 2.5%. If a model did not contain a particular itemset, the IF formula returned "0". This indicated that the item set's support in the model was less than 2.5% and did not surpass the threshold.

The results were as expected that Model 2 would have higher support percentages for a vast majority of courses. The largest difference from Model 1 to Model 2 was the 2-itemset of {'Communications Internship', 'American Healthy Policy'} with Model 2 having a higher support by 0.2830 (28.30%). The closest support on an itemset between the two models was {'Art & Religion'} with Model two having a higher support by 0.004 (0.40%). The range of the differences in support from Model 1

to Model 2 was as follows: [-7.77% to 28.30%] for a range of 36.07%.

There were 3 major differences between the 1-itemsets of each model. Model 1 did not contain either {'Comm & The Presidency'} or {'French Thought Since 1945'}, while Model 2 did not contain {'Early Mesopotam History/Society'}. The complete list of courses (1-itemsets) that exceeded the threshold are listed in the next section "Deployment".

C. Deployment

Upon extensive evaluation and comparison of the two models, I have enough data backed evidence to complete the main objectives of reducing number of credit hours from 39 to 33, defining concentration areas for elective courses, and reducing the number of offered elective courses. I have summarized my recommendations for management below.

1) New Elective Course List

As a result of the frequent itemset models, there should be 25 offered elective courses, as they all proceeded to reach the support threshold. The list of maintained courses is below and is listed in alphabetical order.

1. 21st Century Russian Literature: Fiction and Reality
2. AESTHETICS
3. AFRICAN AMERICAN LIT
4. AMERICAN HEALTH POLICY
5. AMERICAN SOUTH 1861-PRES
6. ART AND RELIGION
7. AUGUSTAN CULTRAL REVOLUTION
8. BECOMING HUMAN
9. BRITISH POETRY 1660-1914
10. Business German: A Micro Perspective
11. CELL. BIOL. & BIOCHEM.
12. COMM & THE PRESIDENCY (Model 2 Only)
13. COMMUNICATIONS INTERSHIP
14. COMPARATIVE POLITICS
15. CONTEMP ART – 1945 TO PRESENT
16. COMTEMPORARY POL.THOUGHT
17. DEVIL'S PACT LIT/FILM
18. EARLY MESOPOTAM HISTORY/SOCIETY (Model 1 Only)
19. ELEMENTARY ARABIC II
20. EUROPE IN A WIDER WORLD
21. EVIDENCE BASED CRIME AND JUSTICE POLICY
22. Environmental Studies Research Seminar Junior Level
23. FRANCE & THE EUROP.UNION
24. FRESHWATER ECOLOGY
25. French Thought Since 1945

2) Removed Courses

The following courses failed to reach the 2.5% minimum support threshold and should no longer be offered. The resources contributed towards these courses can be transferred to improving the list of 25 courses above. *Courses 2 & 7*, "AMERICAN SOCIAL POLICY" and "French Though Till 1945" had a support of 0.00% meaning no student has taken either course in the 5 years the data covers. The following 7 courses should be terminated.

1. 19th CENTURY BRITISH LITERATURE
2. *AMERICAN SOCIAL POLICY (0.00% Support)*
3. ANALYZING THE POL WORLD
4. COMTEMPORARY SOCIO THEORY
5. EARLY BALCAN HIST/SOC
6. ELEMENTARY GERMAN
7. *French Though Till 1945 (0.00% Support)*

3) Concentrations

With a goal to identify concentrations for the Department of Arts, I utilized the 4-itemset results from Model to guide my decision making. As previously mentioned, metadata would have proved useful in making these decisions, specifically course descriptions, insights from students & faculty with involvement in the courses, and the majors offered by the department, including the major(s) for each student.

I used my own vigilance in collaboration with my modeling to create the concentrations listed below. All concentrations are 4-itemsets that surpassed the required 2.5% minimum support threshold.

1. Concentration in International Relations

(Model 2 Support: 5.66%, - *highest supported 4-itemset*)

- COMMUNICATIONS INTERNSHIP
- 21st Century Russian Literature: Fiction and Reality
- AMERICAN HEALTH POLICY
- ELEMENTARY ARABIC II

2. Concentration in U.S. Public Health

(Model 2 4-itemset Support: 3.77% -)

- COMMUNICATIONS INTERNSHIP
- BECOMING HUMAN
- CELL. BIOL. & BIOCHEM
- AMERICAN HEALTH POLICY

3. **Concentration in Environmental Health**

(Model 2 3-itemset Support: 3.77%)

- FRESHWATER ECOLOGY
- Environmental Studies Research Seminar Junior Level
- CONTEMPORARY POL.THOUGHT
- AMERICAN HEALTH POLICY

4. **Concentration in European Studies (Choose at least 4 courses)**

- French Thought Since 1945 (Model 2 Support: 3.8%)
- EUROPE IN A WIDER WORLD (Model 2 Support: 5.7%)
- FRANCE & THE EUROP.UNION (Model 2 Support: 13.2%)
- BRITISH POETRY 1660 – 1914 (Model 2 Support: 9.4%)
- Business German: A Micro Perspective (Model 2 Support: 15.1%)
- AUGUSTAN CULTURAL REVOLUTION (Model 2 Support: 3.78%)

5. **Concentration in World Literature & Art (Choose at least 4 courses)**

- AFRICAN AMERICAN LIT (Model 2 Support: 15.1%)
- ART & RELIGION (Model 2 Support: 9.4%)
- DEVIL'S PACT LIT/FILM (Model 2 Support: 11.3%)
- BRITISH POETRY 1660 – 1914 (Model 2 Support: 9.4%)
- AESTHETICS (Model 2 Support: 7.5%)

6. **Concentration in Western World Studies**

- EARLY MESOPOTAM HISTORY/SOCIETY (Model 1 Support: 3.4%)
- ELEMENTARY ARABIC II (Model 2 Support: 47.2% - *second highest 1-itemset support %*)
- 21st Century Russian Literature: Fiction and Reality (Model 2 Support: 15.1%)
- EUROPE IN A WIDER WORLD (Model 2 Support: 5.67%)

7. **Concentration in U.S. Political Science**

- EVIDENCED BASED CRIME AND JUSTICE POLICY (Model 2 Support: 3.78%)
- COMM & THE PRESIDENCY (Model 2 Support: 3.78%)
- AMERICAN SOUTH 1861-PRES (Model 2 Support: 15.1%)
- COMPARATIVE POLITICS 1660 – 1914 (Model 2 Support: 9.4%)

Concentration 3, "Environmental Health" contains the 3-itemset {'AMERICAN HEALTH POLICY', CONTEMPORARY POL.THOUGHT', 'FRESHWATER ECOLOGY'} with a support of 4.39% and support of 11.32% in Model 2. I added the course "Environmental Studies Research Seminar Junior Level" because it fits the focus of the concentration, even though there was no 4-itemset including the course. Concentrations 4, 5, 6 and 7 were created by individual 1-itemset support and by using qualitative inferences to group the courses. These 4 concentrations do not have the itemset support of the first 3, however, as 21.9% of courses are no longer going to be offered, in addition to the establishment to these concentrations, I expect the enrollment numbers to perform much greater than the historical data may indicate.

Concentration 6, "Middle Eastern Studies" contains the course "EARLY MESOPOTAM HISTORY/SOCIETY" which does not exceed the support threshold in Model 2 but has 3.4% support in Model 1 indicating that it may be trending in a positive direction as students who have taken 1- 3 elective courses and taking the course more frequently.

VI. Summary of Results

To summaries the results of the reconstruction of the University of Diploma Printing: Department of Arts Reconstruction, All objectives were met, including decreasing the required credit hours from 39 to 33, which directly correlates with the next objective, reducing the number of offered elective courses. The final objective was to establish elective course concentrations. Analyzing the findings from frequent itemset mining, I established 7 concentrations: Western World Studies, U.S. Political Science, World Literatrue & Art, European Studies, Environmental Health, U.S. Public Health, and International Relations. The overall course offerings was cut by 21.9%. from 32 to 25 courses. The next step is for the Department Executives to create a strategic implementation plan and timeline for the restructured College of Arts.

REFERENCES

[1] University of Diploma Printing, 2022. Projectdata.csv

[2] Raschka, S (2018). Mlxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack.
J Open Source Softw 3(24)

[3] McKinney, W., & others. (2010). Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference (Vol. 445, pp. 51–56).