Class Review: A Statistical Machine Learning Approach to Penn Course Ratings

David Fan and Rachel Leong
May 1, 2019

Abstract

This project was conducted with the aim of gaining a better understanding of end-of-semester course reviews that Penn students give. By utilizing textual data from course syllabi and descriptions, demographic data on faculty members and logistical information on courses, we (1) build models to predict the ratings that courses receive, (2) test hypotheses on factors influencing and biases in ratings and (3) cluster courses based on their descriptions.

1 Introduction

Penn Course Review has historically been the go-to place for Penn students to find relevant information on the quality of classes and plays a crucial role in course selection by students.

Our problem statements are highly tied to this topic and can be broken down into three main parts. Overall, we aim to tackle not only students behavior when it comes to course ratings but also offer tangible recommendations to the Penn Community on class construction and classification:

The first problem statement is that while end-of-semester ratings are useful, the data is only available after the course is offered. We believe that through the development of an accurate predictive model that utilizes pre-semester course information, we could help the school pick the best classes to offer. This could also allow for strategic decisions such as what classes to prioritize at the peak time (a.k.a. The time where classes are best received). We think that it is in the best interest of the school to offer courses that are going to be well received by the students given the constraints to the amount of classes that can be offered.

The second problem statement is that there may be biases associated with course ratings. We hypothesize that the ratings (course quality in particularly) are highly correlated with the demographic profile of professors, how "nice" the professor is and how convenient the class is (i.e. Less amount of workload, better class time and etc.) and are might not be a good proxy for class quality. We aim to single out these biases underlying the ratings.

The third problem statement is that while classes largely get clustered under departments, classes under each departments or even cross departments can certainly be further broken down into smaller clusters (i.e. STAT 471, 422 and ACCT 270 could largely fall under predictive modelling, while STAT 432 and 442 could be classified as theoretical statistics). Therefore, our goal is to cluster courses based on the course descriptions. Using text analysis, we aim to cluster these classes into smaller more meaningful clusters for students to choose classes at ease across different Penn schools.

We believe the insights from this paper can have a number of applications. Firstly, by garnering perspectives on what exactly Penn students look for in a class, the class curriculum can be better designed to suit certain preferences. Secondly, if clear biases are found in the assessment of course quality associating with the instructor involved in the class, intervention may be needed to correct said problem. Thirdly, by having a better clustering system for classes that go beyond department label, this feature can be incorporated into tools like Penn Course Review to provide people with a better sense of the class itself. Fourthly, this can help Penn strategically launch classes that students may like.

2 Data Preprocessing

2.1 Data Collection and Cleaning

The data collection process was sourced from four streams:

1. *Penn Course Review Data*

   We used the Penn Course Review API to extract course ratings to be used as the target variable in our analysis. We decided to use $rCourseQuality$ as the overall response variable as it incorporated the other categories that students rated at the end of the semester. $rCourseQuality$ is a continuous variable within the range $[0, 4]$. The lowest rated course in the dataset had a value of 0.76 and the highest had a value of 4.0.

2. *Syllabus Data*

   Constrained by resources, we decided to focus exclusively on Wharton classes, as the syllabi are readily available online on the Wharton Syllabus App, a data center for all Wharton syllabi dating back to 2003. We wrote a multi-level web scraping script to gather the AWS links of the syllabus and later used a PDF parser to extract relevant text. We were able to scrape over 5000 syllabi by the end of the process. Once we obtained the texts, we utilized Python's Natural Language Toolkit (NLTK) package to extract relevant terms. This resulted in a document term matrix which we hoped to use to conduct sentiment analysis, establish proxies for the amount of work required in the classes and determine the theme of the course.


```
g\xa0a\xa0much\xa0more\xa0extensive\xa0\r\ntreatment\xa0of\xa0spreadsheet\xa0application\xa0development\xa0(using\xa0
Visual\xa0Basic\xa0for\xa0Applications)\xa0and\xa0\r\ndatabase\xa0management\xa0(using\xa0SQL\xa0as\xa0implemented\xa
0in\xa0the\xa0SQL\xa0Server\xa0database\xa0system\xa0--\xa0transact-\r\nSQL).\xa0In\xa0addition,\xa0the\xa0course\xa0w
ill\xa0cover\xa0related\xa0skills\xa0needed\xa0to\xa0effectively\xa0develop\xa0computer\xa0\r\napplications\xa0includ
ing\xa0specification\xa0development,\xa0interface\xa0design,\xa0and\xa0testing,\xa0as\xa0well\xa0as\xa0more\xa0\r\nsp
ecialized\xa0topics\xa0directly\xa0related\xa0to\xa0the\xa0use\xa0of\xa0databases\xa0such\xa0as\xa0data\xa0structure
s,\xa0database\xa0\r\ndesign,\xa0and\xa0data\xa0security.\xa0\xa0Time\xa0permitting,\xa0we\xa0will\xa0also\xa0explore
\xa0"big\xa0data"\xa0tools\xa0such\xa0as\xa0\r\nHadoop/Hive.\xa0\xa0The\xa0course\xa0is\xa0not\xa0a\xa0substitute\xa0
for\xa0a\xa0computer\xa0science\xa0course\xa0in\xa0programming,\xa0\r\nalgorithm\xa0design\xa0or\xa0databases,\xa0but
\xa0will\xa0draw\xa0on\xa0computer\xa0science\xa0concepts\xa0as\xa0needed\xa0to\xa0provide\xa0\r\na\xa0working\xa0kno
wledge\xa0of\xa0the\xa0necessary\xa0theory\xa0and\xa0necessary\xa0skills.\xa0\r\n\r\nThe\xa0course\xa0is\xa0intended
\xa0for\xa0students\xa0without\xa0prior\xa0experience\xa0in\xa0programming\,\xa0but\xa0students\xa0must\xa0\r\nhave\x
a0familiarity\xa0with\xa0computer-based\xa0tools\xa0as\xa0covered\xa0in\xa0IDD101\xa0or\xa0equivalent,\xa0or\xa0thro
ugh\xa0\r\npersonal\xa0experience.\xa0\xa0\xa0The\xa0course\xa0is\xa0definitely\xa0INTRODUCTORY\xa0in\xa0that\xa0it\x
a0does\xa0not\xa0require\xa0prior\xa0\r\nknowledge\xa0of\xa0the\xa0material.\xa0\xa0That\xa0does\xa0not\xa0mean\xa0it
\xa0will\xa0be\xa0EASY\xa0since\xa0computers\xa0can\xa0be\xa0\r\nunforgiving\xa0when\xa0you\xa0make\xa0a\xa0programmi
ng\xa0mistake\xa0and\xa0some\xa0concepts,\xa0like\xa0object\xa0orientation\xa0\r\nand\xa0set-based\xa0reasoning,\xa0a
re\xa0intellectually\xa0challenging.\xa0We\xa0expect\xa0the\xa0course\xa0to\xa0be\xa0especially\xa0\r\useful\xa0for
\xa0students\xa0seeking\xa0entry-level\xa0analyst\xa0positions\xa0in\xa0data-intensive\xa0firms,\xa0or\xa0those\xa0ge
nerally\xa0\r\nseeking\xa0to\xa0broaden\xa0their\xa0knowledge\xa0and\xa0skills\xa0in\xa0the\xa0construction\xa0and\xa
0use\xa0of\xa0computer-based\xa0\r\nanalytical\xa0tools.\xa0\xa0The\xa0course\xa0counts\xa0toward\xa0the\xa0general\x
a0OPIM/OIDD\xa0concentration\xa0and\xa0the\xa0\r\nInformation\xa0Systems\xa0and\xa0Business\xa0Analytics\xa0tracks.\x
a0\r\n\r\nCourse\xa0Format:\xa0\xa0The\xa0course\xa0will\xa0meet\xa0in\xa0the\xa0JMHH375\xa0Computer\xa0Lab.\xa0\xa0M
ost\xa0sessions\xa0will\xa0include\xa0\r\nat\xa0least\xa0half\xa0the\xa0time\xa0on\xa0in-class\xa0computer\xa0exercis
es.\xa0\xa0\xa0While\xa0this\xa0will\xa0likely\xa0lighten\xa0the\xa0out\xa0of\xa0class\xa0\r\n\r\n\xa0\xa0\xa0\xa0\xa0
0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0
\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\r\n\xa0This
```


Figure 1: Parsed PDF syllabus before cleaning

3. *Course Description Data*

   We scraped the Wharton website for course description data and by the end of the process, had a dictionary of the descriptions of 240 courses offered by the school.

4. *Course Time and Location Data*

   We scraped the Penn Registrar archives for PDF course timetables from previous semesters. Then we cleaned the data by determining patterns in the text and using regular expression (regex) to tease out relevant information. This was challenging as there were many edge cases in the data which needed to be accounted for. Once we completed the data cleaning process, we had a dataframe of term, course, location, day and time, which we hoped to use to better understand how the logistics of a course affect its ratings.

```
OIDD-222  INTERNET LAW & POLICY                1 CU
 401 LEC MW 10:30-12NOON JMHH F85       WERBACH K
     CROSS LISTED: LGST-222
     MAX W/CROSS LIST: 60

OIDD-240  GAMIFICATION FOR BUS                 1 CU
 401 LEC MW 9-10:30AM JMHH F85          WERBACH K
     CROSS LISTED: LGST-240
     MAX W/CROSS LIST: 60

OIDD-261  RISK ANALY & ENV MGMT                1 CU
 401                                  KUNREUTHER H
     LEC MW 10:30-12NOON JMHH F90
     AUDITORS NEED PERMISSION
     ATTENDANCE AT FIRST CLASS MANDATORY
     CROSS LISTED: BEPP-261 BEPP-761 BEPP-961
     MAX W/CROSS LIST: 40
                  ESE -567 OIDD-761
 402 LEC MW 3-4:30PM JMHH G50        KUNREUTHER H
     AUDITORS NEED PERMISSION
     ATTENDANCE AT FIRST CLASS MANDATORY
     CROSS LISTED: BEPP-261 BEPP-761 BEPP-961
     MAX W/CROSS LIST: 40
                  ESE -567 OIDD-761
```

Figure 2: Timetable PDF to obtain course logistics

5. *Faculty Demographic Data*

   We manually labelled faculty members' ethnicity and gender by conducting research on Google, LinkedIn and the Wharton website. This resulted in a dataframe of instructor, race and gender which we hoped to use to determine if there are gender and racial biases in course ratings.

2.2 Data Wrangling and Modification

Aggregating the data on term and course, and dropping rows with nan values produced a joined

dataset of 2587 observations. We then proceeded to process our syllabus text information by extracting the following key elements:

- *The sentiment associated with the Syllabus:* We used the NLTK SentimentIntensityAnalyzer() function to generate a score that reflects how positive or negative the tone of the syllabus is. The syllabi were predominantly neutral, but had either relatively more positive or negative tones. Using this, we encoded the sentiment as 0 if it was more negative and 1 if it was more positive.

- *A document term matrix to reflect the words in each syllabus:* We used the TermDocumentMatrix() function to extract said information. These words served a number of purposes from providing a proxy for the amount of work required in each class to an understanding of central thesis of the class. To avoid sparsity in our dataframe, we only selected words that appeared in at least 100 documents.

The final dataframe had 2587 rows and 1934 columns, consisting of course logistics, instructor demographic information, and the term document matrix.

Given that students are be more interested in the relative quality of a course, we decided to treat this as a classification task, by rounding the ratings to the nearest 0.5 to create 7 factors $1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0$.

3 Implementation and Analysis

3.1 Predictive Model

Using an 80/20 train-test split on the data, we obtained 1916 training observations and 627 test observations. Although we used cross-validation techniques to tune the model parameters, we calculated the test accuracy rate for all models using the same test data for consistency in comparing models.

We tried multiple machine learning models such as Random Forests, K-Nearest Neighbors, Support Vector Machine and Neural Networks to build a classifier that determined the ratings of courses. The baseline model was built using a multinomial distribution to randomly assign ratings to the courses using the proportion of each rating in the training dataset. This yielded a test accuracy rate of 0.23.

3.1.1 Random Forest

Random forests classifier is an ensemble learning method for classification that operates by constructing multiple decision trees at training time and outputting the class that is the mode of the classes of the individual trees. This method corrects for decision trees' habit of overfitting to their training set.

Utilizing grid search on the number of trees and features, we tuned the random forest parameters and found that 300 trees and 31 features at each split yielded the lowest misclassification rate. This resulted in a test accuracy rate of 0.55. Delving deeper into the errors, although 0.45 of the observations were misclassified, this misclassification was within +/- 0.16 of the actual course rating on average.

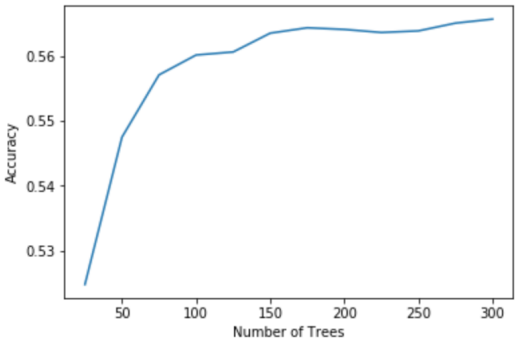


Figure 3: Tuning number of trees

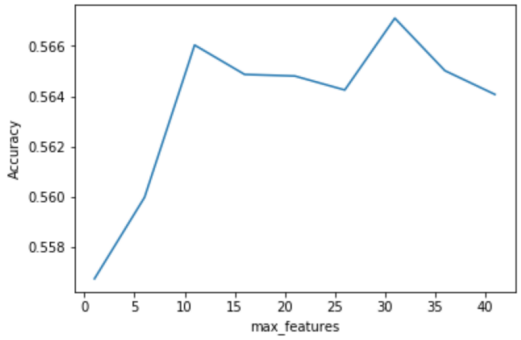


Figure 4: Tuning max depth

3.1.2 K-Nearest Neighbors

The k-nearest neighbors algorithm is a non-parametric method used for classification (or regression) in which the input consists of the k closest training examples in the feature space. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors.

By utilizing 10-fold cross validation, we determined that the optimal number of neighbors was 1, indicating that the objects were simply assigned to the class of their single nearest neighbor. This yielded a test accuracy rate of 0.57.

3.1.3 Support Vector Machine

A support-vector machine model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

By utilizing 10-fold cross-validation, we determined that the best regularization parameter for a non-linear classification model, C was 9. This yielded a test accuracy rate of 0.58.

3.1.4 Final Model: Neural Network

An artificial neural network is a network of simple neurons, which receive input, change their internal state according to that input, and produce output depending on the input and activation. The network forms by forming a directed, weighted graph. The weights as well as the functions that compute the activation can be modified by a process called learning which is governed by a learning rule.

Since there is no systematic method to tune the neural network model, we used trial and error to determine the optimal number of epochs, neurons at each layer, number of layers, and batch size.

The final model had 1916 features as input, 3 hidden layers with [20, 20, 15] neurons using rectified linear unit (ReLU) activation functions and an output layer with 7 categories using a softmax activation function.

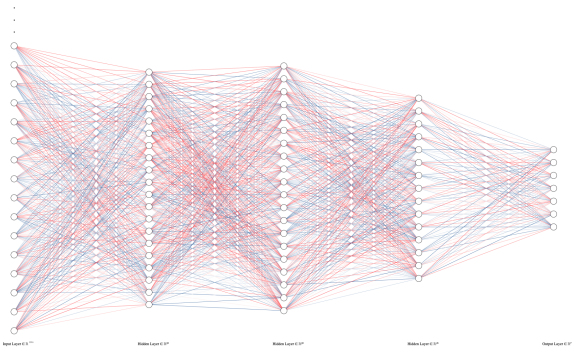


Figure 5: Neural network architecture

There were 100 epochs and a batch size of 10. We used a categorical cross-entropy loss function and Adam optimizer. The test accuracy rate was 0.61.

3.1.5 Model Performance

By comparing the test accuracy performance of the models we trained, the neural network performed marginally the best.

Models	Test Accuracy
Baseline	23%
Random Forest	55%
K-Nearest Neighbors	57%
Support Vector Machine	58%
Neural Network	61%

Table 1: Accuracy rates of models

However, delving deeper, 85% of the misclassified observations were within +/- 1 bucket away from its true value (ie. if the true rating bucket is 2.5, the observation was either predicted to be 2.0 or 3.0).

3.2 Hypothesis testing

In line with our goal to single out potential biases underlying course ratings, we decided to perform a series of statistical hypothesis testing on some of our speculations. Specifically, we tested demographic and logistic factors that relate to a class. Given that our hypothesis tests are primarily on testing anomalies in course rating in relation to these factors, we approached the tests methodologically:

Two-Sample t-Test
Using an unpaired two sample t-test for equal means to test for differences between groups within and out of our factors. We believe our data generally meet the underlying assumptions of said test:

- *The data are continuous:* For this part of the analysis, we will be utilizing the continuous version of our target variable, CourseQuality
- *The data follow the normal probability distribution:* The Course Quality approximately follows a normal probability distribution
- *The variances of the two populations are equal:* Since the students at large came from the same population (Penn students), we don't expect a drastic change in the dispersion of rating in different classes
- *The two samples are independent:* We believe that at large the students rates the class independently from each other, thus each subset group of students will be independent from each other
- *Both samples are simple random samples from their respective populations:* We don't expect students' decision when picking a class to be largely affected by non-course related factors

Chi-Square Goodness of Fit Test
Using a chi-square goodness of fit test to test abnormality in course quality distribution between samples with and without the factor of interest. We believe our data generally suffice the underlying assumption of said test:

- *The sampling method is simple random sampling:* we don't expect students' decision when picking a class to be largely affected non-course related factors. In the event that this assumption may not be complete met (i.e. Time of the class) we assume this characteristic for the ease of this analysis.
- *The variable under study is categorical:*We transformed our continuous data to a categorically binned variable with 8 factors
- *The expected value of the number of sample observations in each level of the variable is at least 5:* We have enough observations under each relevant category

3.2.1 On Race

This segment of our hypothesis test focuses on the racial information of the instructors heading a class. After manipulating the data, we have the distribution below:

rCourseQualityBinned	race	1.0	1.5	2.0	2.5	3.0	3.5	4.0
0	White	11.0	67.0	326.0	554.0	565.0	273.0	34.0
1	Black	0.0	3.0	7.0	18.0	29.0	15.0	0.0
2	Asian	6.0	25.0	118.0	138.0	90.0	18.0	1.0
3	Indian	0.0	3.0	43.0	58.0	74.0	19.0	0.0
4	Others	1.0	8.0	25.0	28.0	20.0	8.0	2.0

Figure 6: Race rating cross table

This section mainly tested on whether or not the course rating distribution and average rating of classes of which the instructors are minorities significantly differs. Specifically, we examined classes headed by East Asian (hence forth refer to as Asian), Black, and Indian professors.

Overall, minority professors seem to be rated lower compared to Caucasian professors (Below Average with a confidence interval of [0.10, 0.19]). When comparing across groups, the result suggests a favor for African American Professors (Above average with a confidence interval of [0.06, 0.314]), a neutral reaction for Indian Professors (Average with a confidence interval of [-0.05, 0.08]), and a significantly low rating for Asian Professors (Below average with a confidence interval of [-0.30,- 0.197]).

We also observed a difference in terms of the distributions of course ratings across minority instructors (see figure 13 in appendix). For instance, the ratings of African American Professors appear less normal, with its mode peaking right above the 3 point mark. The rating of Asian Professors on the other hand shows an over all shift to the left (closer to 0) compared to non-Asian Professors. Ratings of Indian Professors have the mode remain at a similar place compared to its counterparts, however its distribution seems much wider, with a fatter tail. The

seemingly converging consensus that gives a low rating for Asian Professors prompted us to focus the next part of the analysis on that subgroup.

While the low rating of Asian Professors might be a result of Simpson's Paradox (a phenomenon in probability and statistics, in which a trend appears in several different groups of data but disappears or reverses when these groups are combined) that stems from the disproportionate amount of Asian professors in class areas where ratings are significantly lower: Business Economics and Public Policy, Accounting and Statistics (See figure 17 and 18 in appendix).

we found that even after controlling for departments, Asian Professors are still rated lower than their colleagues in the same department (See figure 19 and 20 in appendix). This indicates that the lower rating of Asian Professor is not likely to be by chance, but due to certain factors that actually persist. We believe that this may be a result of an unfamiliar accent that some Asian Professors have. Nonetheless, we believe that is an area that could be further investigated to examine the causality behind observed rating.

3.2.2 On Gender

This section of our paper aims to examine the effect of instructor gender on class ratings.

	gender	ACCT	BEPP	FNCE	HCMG	LGST	MGMT	MKTG	OIDD	REAL	STAT	chi_stats	p_value
0	Male	222.00000	138.000000	377.000000	18.000000	127.000000	313.000000	226.000000	66.000000	52.000000	225.000000	221.384	1.07068e-42
1	Female	80.00000	61.000000	47.000000	13.000000	55.000000	218.000000	224.000000	21.000000	54.000000	50.000000		
2	Expected	205.92501	135.692308	289.113259	21.137998	124.100503	362.073444	306.841902	59.322768	72.278315	187.514496		
3	Expected	96.07499	63.307692	134.886741	9.862002	57.899497	168.926556	143.158098	27.677232	33.721685	87.485504		

Figure 7: Gender Rating Cross Table

By observing the department distribution of Male and Female instructors, we found that in line with common perception, more male instructors tend to appear in quantitative disciplines such as Statistics and Finance, while female instructors are more likely to appear in disciplines such as Management and Marketing (See figure 21 in appendix). From previous analysis, it may seem as though female teacher are likely going to be rated higher as qualitatively disciplines tend to be appreciated more, However, we found that male instructors on average are still more well received than females (Confidence interval of [0.0003, 0.0396])(See figure 22 in appendix).

However, there is only a small difference and it may have been caused by external factors such as the relevance of course content that may have differed between male and female professors. This is another area with a potential underlying causality.

One thing that we particularly looked into is the interaction between Race and Gender. We found that both Black and Asian Female Professors are rated better compared to Black and Asian Male Professors. While the flip is true for all other races (See figure 23 in appendix). This duality constitutes another interesting trend that should be further investigated.

3.2.3 On Class Time

This subsection aims to test whether or not "convenience" plays a role in the rating of course quality. Specifically, this section will dive into examining whether or not the class time affects the course rating. To best assess the aggregate effects, we binned the class into 4 bins. Classes starting from 8-10AM are classified as early classes, classes starting from 11AM -2PM are classified as mid day classes, classes starting from 3PM - 5PM are classified as later afternoon, and classes starting from 6PM onward are classified as night classes.

This generated the data frame below. Excluding night classes which only had 3 observations, we focused our analysis on the first 3 bins of classes

rCourseQualityBinned	start_range	1.0	1.5	2.0	2.5	3.0	3.5	4.0
0	early	7.0	25.0	131.0	258.0	242.0	119.0	10.0
1	late-afteroon	2.0	10.0	78.0	165.0	202.0	102.0	24.0
2	mid-day	9.0	71.0	310.0	373.0	331.0	111.0	3.0
3	night	0.0	0.0	0.0	0.0	3.0	1.0	0.0

Figure 8: Class Time Rating Cross Table

Contrary to our initial speculation that early morning classes are likely going to contribute negatively to course ratings, we found that early morning classes are actually rated higher than the average rating of classes at a different times (Above average with a confidence interval of [0.06, 0.314]). Classes happening during mid-day are rated quite

low (Below average with a confidence interval of [0.06, 0.314]), while the classes happening during late afternoon are rated quite high (Above average with a confidence interval of [0.06, 0.314]). The distribution all the classes look quite similar except for the slight difference in the location of their mode (See figure 14 in appendix).

We could infer a number of things from this observation. First, there is likely a self selection effect for morning classes: individuals who perceive morning class as an added burden will unlikely take those classes to begin with. Second, the reason as to why mid-day classes are rated lowly is likely a result of its direct overlap with typical meal time and nap-time (associated with food coma).

While we can't assume the causality, the ratings that differ across class time may also be affected by the department placement. In fact most classes scheduled in the late afternoon are classes from the Management and Marketing departments, which tend to be rated higher, while those in mid day are typically Accounting or BEPP classes, that tend to have a lower rating.

3.3 Class Clustering

There are many courses at Penn across the different schools which complement each other, but as students, we are often times unaware of classes beyond those that the people around us take. When finding similar courses, search engines will return results that match the name of the course, but that may be misleading. For instance, STAT 474 Modern Regression isn't actually about regression, but rather is an applied machine learning course more akin to STAT 471 Modern Data Mining and ESE 305 Foundations of Data Science and CIS 545 Big Data Analytics.

For this part, we used the course description associated with all Wharton courses and tested our model against non-Wharton courses that we believe share similar content with a group of Wharton classes. We gathered the data from the Wharton Website, resulting in a dataset of 240 observations, each for a Wharton Class that is currently offered, and three columns: Class ID, Class Name and Class Description.

3.3.1 Latent Dirichlet Allocation (LDA)

LDA is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. For example, if observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word's presence is attributable to one of the document's topics. It is a form of topic modelling that is unsupervised.

By tuning the number of topics in our model, we found that 30 topics was the most appropriate given that it presented a reasonable coherence score and was small enough for students to be able to easily understand when choosing courses to take.

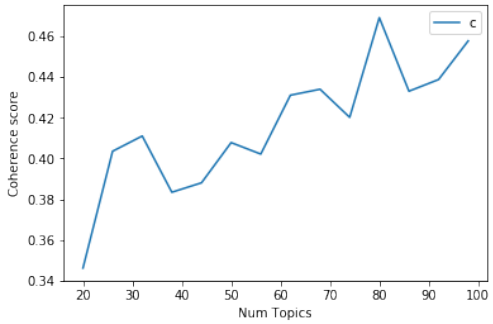


Figure 9: Coherence plot

Once we had determined the topics, we manually labelled them by names we thought were appropriate given the most frequently appeared words and the courses that were assigned to them. An example of a cluster that we labeled was the 'Investment Banking' Cluster. The words associated came with the following weights: *0.063*"situat" + 0.063*"ethic" + 0.043*"statement" + 0.037*"conflict" + 0.034*"valuat" + 0.032*"includ" + 0.031*"account" + 0.028*"techniqu" + 0.027*"analys" + 0.024*"cover.*

Combining this with the classes that were placed into the cluster. We deemed that it is best suited with label of Investing Banking as it taps into necessary financial valuation and negotiations skill needed for the position

Something worth noting is that each class is not clustered into one cluster alone, but instead follows

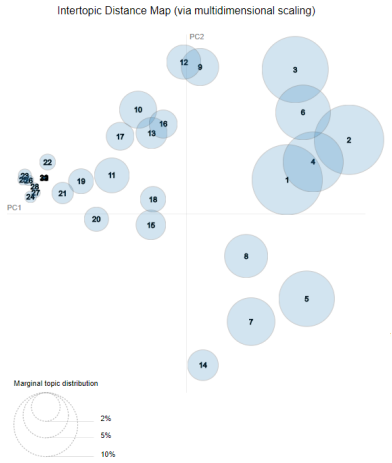


Figure 10: Final clusters in 2-D using PCA

description	course	name	topic	perc
This course includes not only conflict resolut...	MGMT291	Negotiations	10.0	0.516367
The focus of this course is on the valuation o...	FNCE207	Corporate Valuation	10.0	0.550344
This course is an introduction to the basic co...	ACCT101	Acct & Financial Report	10.0	0.217002
This course builds on the knowledge you obtain...	ACCT212	Fin Measurement & Disclo	10.0	0.228754

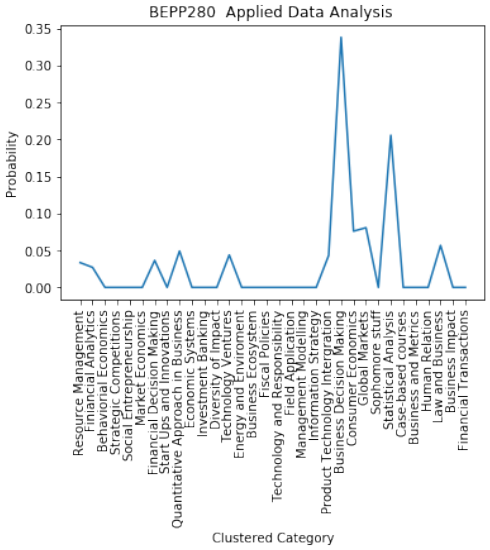
Figure 11: Courses within 'Investment Banking'

a probability distribution across a number of topics/clusters as seen in the example below:

3.3.2 Model Performance

We tested the model on 5 classes, CIS 545, ESE 305, NURS 230, ECON 246 and MATH 530 to see which topics and Wharton courses they were most associated with and found the results to be favorable.

- *CIS 545 Big Data Analytics* **Topics**: Business Decision Making, Market Economics and Statistical Analysis. Similar to Applied Data Analysis and Data Analysis in the Digital Economy
- *ESE 305 Foundation of Data Science* **Topics**: Management Modelling, Business Decision Making and Statistical Analysis. Similar to Forecasting Methods of Management and Introduction to Bayes Data Analysis
- *NURS 230 Statistics for Research and Measurement* **Topics**: Statistical Analysis. Similar to Statistical Inference and Intro to Statistics



- *ECON 246 Money and Banking* **Topics**: Consumer Economics. Similar to Capital Markets and Policy Decision by Central Bank
- *MATH 530 Math of Finance* **Topics**: Quantitative Approach in Business and Global Markets. Similar to Investment Management and Strategic Equity in Finance

As illustrated in the plot below, picking the two most representative topic: Business Decision Making and Statistical Analysis, CIS 545 Big Data Analytics's distribution is closest to that of STAT 474 Modern Regression, ESE 305 Foundations of Data Science, STAT 471 Modern Data Mining, STAT 442 Bayesian Data Analysis, BEPP 280 Applied Data Analysis, STAT 435 Forecasting Methods Management.

4 Conclusion

We started this project with three goals in mind: (1) to predict course ratings using syllabi text, instructor demographics and course logistics, (2) to test hypotheses on factors and biases related to course ratings and (3) to cluster courses based on their descriptions. The data was scraped from multiple sources, namely using the PennCourseReview API, Wharton website, Penn Registrar archives and manual labeling.

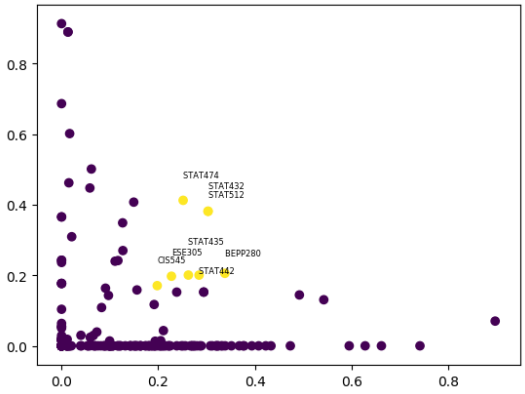


Figure 12: Courses within same cluster of CIS 545

To predict the ratings, we tested out different machine learning models, and settled on a final neural network model that achieved a test accuracy rate of 61%, with 85% of its misclassified observations being within +/- 1 classification bucket away. The other models performed only marginally worse.

To test hypotheses on biases and factors influencing course ratings, we conducted t-tests and Chi-square tests to determine if our null hypotheses could be rejected at the α level. We found that asian professors are rated lower than their colleagues in the same department, Black and Asian female professors are rated higher than their male counterparts but the opposite is true for White professors, and midday classes are generally rated lower but this could also be due to confounding factors such as the class department.

To cluster the courses, we trained an LDA model on course descriptions and used that to predict the probability that a new course description was bucketed into each of the topics. For this task, we compared coherence scores and decided on using 30 topics as it provided a fair trade-off between coherence and understandability for students to be able to easily choose courses.

One of the main challenges we faced was in data cleaning because the data that we scraped was mostly unstructured. For instance, every professor had a differently formatted syllabus and the format of the course timetable from the Registrar archives was not uniform across all courses and years. Therefore, we needed to identify patterns within the text and utilize regex to tease them out. Additionally, we needed to account for edge cases that did not follow the aformentioned patterns.

Moving forward, the models we built could be improved on. The predictive model achieved an accuracy rate of 61% but perhaps by combining different models, a higher accuracy rate could be achieved. There are also many other hypotheses that could be tested using the data, especially since there may be confounding variables affecting the outcomes. The clustering model could also be trained on courses from other schools to obtain a more holistic product that Penn students can use.

References

En.wikipedia.org. (2019). Alternation. *Random forest.* [online] Available at: https://en.wikipedia.org/wiki/Random_forest [Accessed 1 May 2019]

En.wikipedia.org. (2019). Alternation. *K-nearest neighbors algorithm.* [online] Available at: https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm [Accessed 1 May 2019]

En.wikipedia.org. (2019). Alternation. *Support vector machine.* [online] Available at: https://en.wikipedia.org/wiki/Support-vector_machine [Accessed 1 May 2019]

En.wikipedia.org. (2019). Alternation. *Artificial neural network.* [online] Available at: https://en.wikipedia.org/wiki/Artificial_neural_network [Accessed 1 May 2019]

En.wikipedia.org. (2019). Alternation. *Artificial neural network.* [online] Available at: https://en.wikipedia.org/wiki/Student%27s_t-test [Accessed 1 May 2019]

En.wikipedia.org. (2019). Alternation. *Chi square test* [online] Available at: https://en.wikipedia.org/wiki/Chi-squared_test [Accessed 1 May 2019]

En.wikipedia.org. (2019). Alternation. *Goodness of fit.* [online] Available at: https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation [Accessed 1 May 2019]

5 Appendix

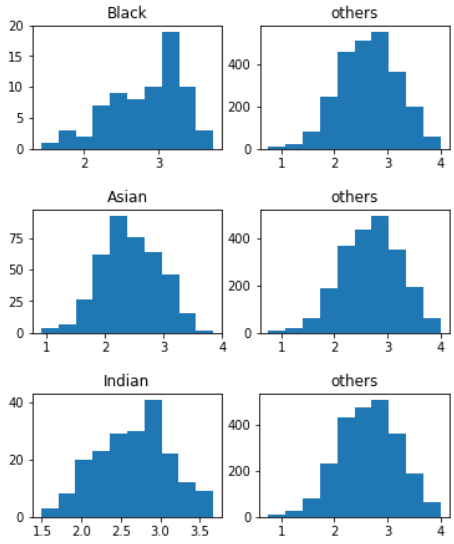


Figure 13: Course Quality Distribution

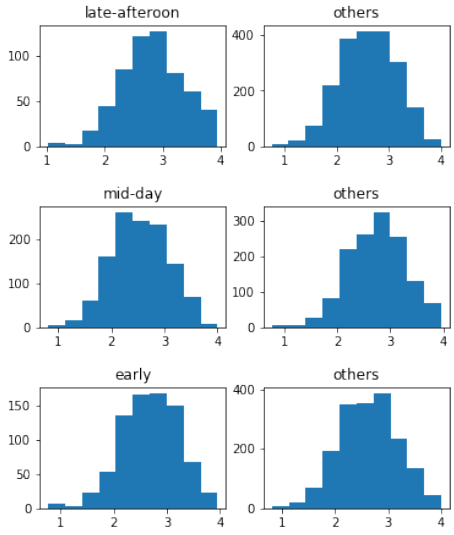


Figure 14: Distribution of ratings by class time

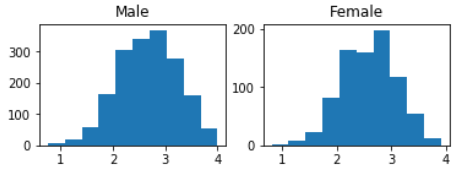


Figure 15: Distribution of Course Quality by Gender

	of_interest	mean	t_stats	p_value	conf_interval_lb	conf_interval_ub
0	Black	2.832083	2.93861	0.00332606	0.0698732	0.314468
1	Others	2.639913				

	of_interest	mean	t_stats	p_value	conf_interval_lb	conf_interval_ub
0	Asian	2.431717	-8.54382	2.18358e-17	-0.306459	-0.19782
1	Others	2.683857				

	of_interest	mean	t_stats	p_value	conf_interval_lb	conf_interval_ub
0	Indian	2.661320	0.427914	0.668749	-0.0508054	0.0855705
1	Others	2.643937				

Figure 16: T-test Racial Information results

	race	ACCT	BEPP	OIDD	REAL	STAT	chi_stats	p_value
0	Asian	65.000000	37.000000	21.000000	24.000000	92.000000	246.867	4.57318e-48
1	Everything Else	237.000000	162.000000	66.000000	82.000000	183.000000		
2	Expected	46.228063	30.461538	13.317356	16.225744	42.095091		
3	Expected	255.771937	168.538462	73.682644	89.774256	232.904909		

Figure 17: Excerpt T-Tests of Specific Departments

rCourseQualityBinned	Department	1.0	1.5	2.0	2.5	3.0	3.5	4.0
0	ACCT	1.0	18.0	157.0	109.0	13.0	4.0	0.0
1	BEPP	3.0	26.0	86.0	49.0	20.0	15.0	0.0
2	FNCE	0.0	13.0	60.0	147.0	160.0	43.0	1.0
3	HCMG	0.0	0.0	0.0	12.0	16.0	3.0	0.0
4	LGST	0.0	0.0	16.0	19.0	61.0	76.0	10.0
5	MGMT	3.0	16.0	76.0	158.0	171.0	89.0	18.0
6	MKTG	4.0	5.0	35.0	139.0	198.0	63.0	6.0
7	OIDD	0.0	0.0	9.0	27.0	35.0	16.0	0.0
8	REAL	0.0	0.0	17.0	41.0	40.0	8.0	0.0
9	STAT	7.0	28.0	63.0	95.0	64.0	16.0	2.0

Figure 18: Course Rating by Department

	of_interest	mean	t_stats	p_value	conf_interval_lb	conf_interval_ub
0	MKTG	2.807689	6.98255	3.67144e-12	0.147776	0.245487
1	Others	2.611058				

	of_interest	mean	t_stats	p_value	conf_interval_lb	conf_interval_ub
0	ACCT	2.213808	-15.1936	5.28522e-50	-0.534097	-0.442856
1	Others	2.702284				

	of_interest	mean	t_stats	p_value	conf_interval_lb	conf_interval_ub
0	STAT	2.443164	-6.52181	8.32566e-11	-0.296593	-0.155679
1	Others	2.669299				

Figure 19: Asian Professors Distribution Across Departments

	of_interest	mean	t_stats	p_value	conf_interval_lb	conf_interval_ub
0	AsianSTAT	2.095652	-9.98198	4.74999e-23	-0.670862	-0.468887
1	OtherSTAT	2.617869				

	of_interest	mean	t_stats	p_value	conf_interval_lb	conf_interval_ub
0	AsianACCT	2.176308	-7.05417	2.22044e-12	-0.55501	-0.407069
1	OtherACCT	2.224093				

	of_interest	mean	t_stats	p_value	conf_interval_lb	conf_interval_ub
0	AsianBEPP	2.157838	-5.48072	4.64567e-08	-0.577363	-0.411628
1	OtherBEPP	2.251728				

Figure 20: Excerpt T-Tests for Asian Professors in Selected Departments

	gender	ACCT	BEPP	FNCE	HCMG	LGST	MGMT	MKTG	OIDD	REAL	STAT	chi_stats	p_value
0	Male	222.00000	138.000000	377.000000	18.000000	127.000000	313.000000	226.000000	66.000000	52.000000	225.000000	221.384	1.07068e-42
1	Female	80.00000	61.000000	47.000000	13.000000	55.000000	218.000000	224.000000	21.000000	54.000000	50.000000		
2	Expected	205.92501	135.692308	289.113259	21.137998	124.100503	362.073444	306.841902	59.322768	72.278315	187.514496		
3	Expected	96.07499	63.307692	134.886741	9.862002	57.899497	168.926556	143.158098	27.677232	33.721685	87.485504		

Figure 21: Distribution Across Departments by Gender

	of_interest	mean	t_stats	p_value	conf_interval_lb	conf_interval_ub
0	Male	2.671627	3.59143	0.000335001	0.0396725	0.126084
1	Female	2.588748				

Figure 22: T-Tests of Course Quality by Gender

	of_interest	mean	t_stats	p_value	conf_interval_lb	conf_interval_ub
0	AsianMale	2.379467	-6.57587	5.83172e-11	-0.367739	-0.201002
1	Others	2.663838				

	of_interest	mean	t_stats	p_value	conf_interval_lb	conf_interval_ub
0	AsianFemale	2.470617	-5.05168	4.68479e-07	-0.256935	-0.12595
1	Others	2.662059				

	of_interest	mean	t_stats	p_value	conf_interval_lb	conf_interval_ub
0	IndianMale	2.711597	1.35236	0.176377	-0.0147516	0.15382
1	Others	2.642062				

	of_interest	mean	t_stats	p_value	conf_interval_lb	conf_interval_ub
0	IndianFemale	2.584615	-0.992591	0.321002	-0.169869	0.0448075
1	Others	2.647146				

	of_interest	mean	t_stats	p_value	conf_interval_lb	conf_interval_ub
0	WhiteMale	2.704714	5.86586	5.03786e-09	0.0840009	0.16752
1	Others	2.578953				

	of_interest	mean	t_stats	p_value	conf_interval_lb	conf_interval_ub
0	WhiteFemale	2.645794	0.0231899	0.981501	-0.0493342	0.0506346
1	Others	2.645144				

	of_interest	mean	t_stats	p_value	conf_interval_lb	conf_interval_ub
0	BlackFemale	2.976429	2.26943	0.0233245	0.00287985	0.663059
1	Others	2.643459				

	of_interest	mean	t_stats	p_value	conf_interval_lb	conf_interval_ub
0	BlackMale	2.797241	2.13797	0.0326123	0.0231999	0.287732
1	Others	2.641775				

	of_interest	mean	t_stats	p_value	conf_interval_lb	conf_interval_ub
0	OthersFemale	2.460526	-2.09515	0.0362547	-0.417103	0.0421262
1	Others	2.648015				

	of_interest	mean	t_stats	p_value	conf_interval_lb	conf_interval_ub
0	OthersMale	2.527222	-1.60033	0.109647	-0.261993	0.0208829
1	Others	2.647777				

Figure 23: T-Test on Gender Race Interaction

	race	1.0	1.5	2.0	2.5	3.0	3.5	4.0	chi_stats	p_value
0	Black	0.000000	3.000000	7.00000	18.000000	29.00000	15.000000	0.000000	12.5343	0.0510562
1	Everything Else	18.000000	103.000000	512.00000	778.000000	749.00000	318.000000	37.000000		
2	Expected	0.500966	2.950135	14.44453	22.153846	21.65288	9.267878	1.029764		
3	Expected	17.499034	103.049865	504.55547	773.846154	756.34712	323.732122	35.970236		

	race	1.0	1.5	2.0	2.5	3.0	3.5	4.0	chi_stats	p_value
0	Asian	6.000000	25.000000	118.000000	138.000000	90.000000	18.000000	1.000000	72.845	1.06527e-13
1	Everything Else	12.000000	81.000000	401.000000	658.000000	688.000000	315.000000	36.000000		
2	Expected	2.755315	16.225744	79.444917	121.846154	119.090839	50.973328	5.663703		
3	Expected	15.244685	89.774256	439.555083	674.153846	658.909161	282.026672	31.336297		

	race	1.0	1.5	2.0	2.5	3.0	3.5	4.0	chi_stats	p_value
0	Indian	0.0000	3.000000	43.00000	58.000000	74.000000	19.000000	0.000000	14.1399	0.0281124
1	Everything Else	18.0000	103.000000	476.00000	738.000000	704.000000	314.000000	37.000000		
2	Expected	1.3707	8.071898	39.52184	60.615385	59.244685	25.357944	2.817549		
3	Expected	16.6293	97.928102	479.47816	735.384615	718.755315	307.642056	34.182451		

Figure 24: Chi-Squared Test results

	gender	1.0	1.5	2.0	2.5	3.0	3.5	4.0	chi_stats	p_value
0	Male	13.000000	77.000000	325.000000	534.000000	520.000000	263.000000	32.000000	33.2071	9.56615e-06
1	Female	5.000000	29.000000	194.000000	262.000000	258.000000	70.000000	5.000000		
2	Expected	12.273676	72.278315	353.890993	542.769231	530.495555	227.063007	25.229223		
3	Expected	5.726324	33.721685	165.109007	253.230769	247.504445	105.936993	11.770777		

Figure 25: Proportion Test

	start_range	1.0	1.5	2.0	2.5	3.0	3.5	4.0	chi_stats	p_value
0	early	7.00000	25.000000	131.000000	258.000000	242.000000	119.000000	10.000000	15.736	0.0152435
1	Everything Else	11.00000	81.000000	388.000000	538.000000	536.000000	214.000000	27.000000		
2	Expected	5.51063	32.451488	158.889834	243.692308	238.181678	101.946656	11.327406		
3	Expected	12.48937	73.548512	360.110166	552.307692	539.818322	231.053344	25.672594		

	start_range	1.0	1.5	2.0	2.5	3.0	3.5	4.0	chi_stats	p_value
0	mid-day	9.000000	71.000000	310.00000	373.000000	331.000000	111.000000	3.000000	104.444	2.96181e-20
1	Everything Else	9.000000	35.000000	209.00000	423.000000	447.000000	222.000000	34.000000		
2	Expected	8.405102	49.496714	242.34712	371.692308	363.287205	155.494395	17.277155		
3	Expected	9.594898	56.503286	276.65288	424.307692	414.712795	177.505605	19.722845		

	start_range	1.0	1.5	2.0	2.5	3.0	3.5	4.0	chi_stats	p_value
0	late-afteroon	2.000000	10.000000	78.000000	165.000000	202.000000	102.000000	24.00000	85.7246	2.33621e-16
1	Everything Else	16.000000	96.000000	441.000000	631.000000	576.000000	231.000000	13.00000		
2	Expected	4.056436	23.887901	116.960572	179.384615	175.328179	75.044066	8.33823		
3	Expected	13.943564	82.112099	402.039428	616.615385	602.671821	257.955934	28.66177		

	start_range	1.0	1.5	2.0	2.5	3.0	3.5	4.0	chi_stats	p_value
0	night	0.000000	0.000000	0.000000	0.000000	3.000000	1.000000	0.000000	5.43228	0.48968
1	Everything Else	18.000000	106.000000	519.000000	796.000000	775.000000	332.000000	37.000000		
2	Expected	0.027831	0.163896	0.802474	1.230769	1.202938	0.514882	0.057209		
3	Expected	17.972169	105.836104	518.197526	794.769231	776.797062	332.485118	36.942791		

Figure 26: Distribution

	of_interest	mean	t_stats	p_value	conf_interval_lb	conf_interval_ub
0	early	2.681843	2.25748	0.0240614	0.00720303	0.0982443
1	Others	2.629120				

	of_interest	mean	t_stats	p_value	conf_interval_lb	conf_interval_ub
0	mid-day	2.536126	-9.64919	1.14346e-21	-0.246304	-0.16317
1	Others	2.740863				

	of_interest	mean	t_stats	p_value	conf_interval_lb	conf_interval_ub
0	late-afteroon	2.818113	8.7808	2.89416e-18	0.173697	0.27258
1	Others	2.594975				

	of_interest	mean	t_stats	p_value	conf_interval_lb	conf_interval_ub
0	night	3.167500	1.90869	0.0564132	-0.108471	1.15457
1	Others	2.644452				

Figure 27: Distribution

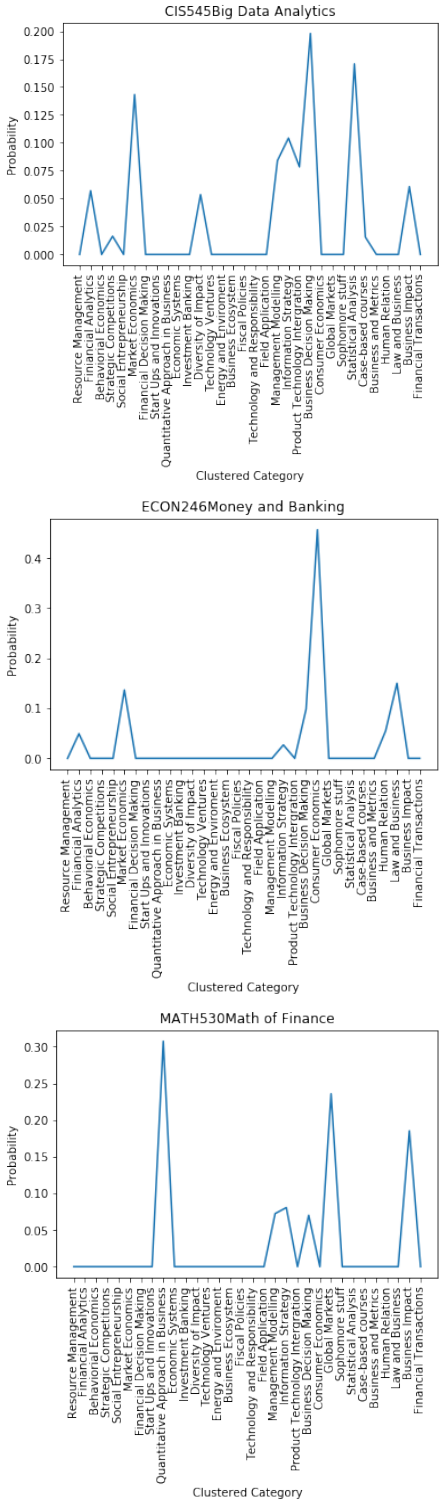


Figure 28: Distribution of topics of courses in test data

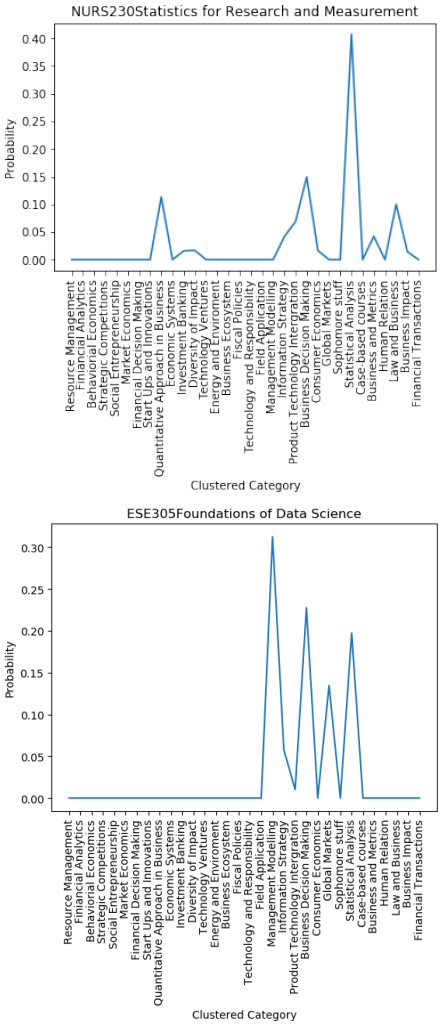


Figure 29: Distribution of topics of courses in test data