BIA 6301 APPLIED DATA MINING

HOMEWORK ASSIGNMENT #3

General Instruction: The homework assignment is due on the assigned date at 5:45 PM. Assignment turned in after the due date and time will lose 2 points for every day late. No assignment will be accepted one week after it is assigned. Here is a breakdown of the point distribution.

|  |  |  |
| --- | --- | --- |
| **Task** | **Points Possible** | **Preferred File Name** |
| Part A: Memorandum on Cluster Analysis  Part B: PCA  Part C: Final Project Proposal | 10 points  10 points  3 points | LastName\_memo.docx  LastName\_PartBC.docx |
| R markdown documentation for Parts A & B | 2 points | LastName.html or LastName.docx or LastName.pdf |
| Extra Credit | 1 point |  |
| ***Total*** | ***25 points*** |  |

To help facilitate the grading process, please use the file naming convention listed in the table above. Please upload the required files onto Blackboard for grading.

The html outputs of your markdown files will not be graded, but they will be checked if necessary to verify your findings and recommendations. Point deductions may occur if there are major discrepancies between your written answers and memorandums and the knitted markdown files.

Please upload your Word documents individually onto Blackboard. Please put your html files in a zipped folder and upload it. Blackboard’s upload feature does not accept html files.

**Part A: Updating Cluster Analysis Exercise**

You have learned that k-means is not the correct model when you have mixed data type. Please review the k-means cluster analysis you did in Homework Assignment #2 and update the analysis using the correct algorithm(s) for your cleaned data set.

Please write a business memorandum to the CEO of General Casualty Kansas City (GCKC) addressing what you found in the (updated) cluster analysis exercise. The memorandum needs to address the following items:

1. What is (are) the business question(s)/problem(s) you were attempting to solve/answer with the cluster analysis exercise?

Separate Document.

1. What did you find conducting the cluster analysis exercise? Can you describe profiles of the crash clusters?

Separate Document.

1. Can you make at least two recommendations for how the cluster analysis exercise can help GCKC inform its decision to expand to new regional markets?

Separate Document.

The memorandum should be between two to three pages. Since you are writing to a general managerial audience who is also your superior, you should write in clear, business professional language. (Please no technobabble.) You should format and spellcheck your memorandum prior to submission.

Please save the memorandum as a separate file and do not include answers for Parts B & C. You should use a separate file for Parts B &C.

**Part B: Principal Components Analysis (PCA) on University Ranking**

The data set **Universities.csv** contain information on 1,302 American colleges and universities offering an undergraduate program. For each university, there are 17 measurements that include continuous measurements (such as tuition and graduation rate) and categorical measurements (such as location by state and whether it is a private or public school). You should remove all categorical variables and missing numeric measurements from the data set prior to doing the tasks below.

1. Conduct a principal component analysis (PCA) of the given data set and comment on the results.

When conducting PCA on a normalized and formatted version of the given data set, I found that three principal components accounted for 64.52% of the variation within the data set. Although the Scree Plot showed three principal components to be where the ‘elbow’ occurs, I decided that a higher percentage of variation needed to be accounted for. Thus, I chose 6 principal components to be an accurate representation since this accounts for 80.14% of variation present in the data set. The table below shows the characteristics that defined each principal component.

|  |  |  |  |
| --- | --- | --- | --- |
| Principal Component Number | Significant Positive Weighting | Significant Negative Weighting | Potential Classification |
| 1 | Student/Faculty Ratio, FT Undergrad | In-State Tuition, Out-of-State Tuition | Community Colleges |
| 2 | Application Received, Application Accepted | None | State Colleges |
| 3 | Estimated Book Costs, Estimated Personal $ | Additional Fees, % Faculty with PHD | Specialized Programs (Trade Schools) |
| 4 | Additional Fees, Room | Estimated Personal $, % New Student from Top 25% | Two Year Junior Colleges |
| 5 | PT Undergrad, Board | Estimated Book Costs, Additional Fees | Non-traditional Colleges |
| 6 | Student/Faculty Ratio, Estimated Book Costs | Estimated Personal $, Additional Fees | Community Colleges |

In this table, significant positive weighting refers to categories where a high value dominated the component. Similarly, significant negative weighting refers to categories where a low value dominated the component. This provides characteristics that explain the variation for each component. Therefore, PCA can be viewed as another method of clustering. As such, I was able to determine types of schools that each principal component may be referring to. For example, Component 5 has a high part-time undergraduate population and high boarding costs, with a low estimated book cost and additional fees. To me this indicates non-traditional students that attend school online, thus non-traditional colleges. The same deductive approach was applied to the other components to produce a potential classification. Although, I found it surprising that Student/Faculty Ratio never appeared in the significant negative weighting column as this would potentially indicate private colleges.

I believe my last statement about never seeing a low student/faculty ratio appear is directly linked to the instructions provided by the question: “remove all categorical variables and missing numeric measurements from the data set prior to doing the tasks below”. The dimensions of the original data set were 1302 observations across 20 variables. After following the instructions, the formatted data set only contained 471 observations across 18 variables. In other words, simply throwing out observations with missing variables removed 64% of the original data set. It is very plausible that this removal of observations removed many private colleges due to their data being harder to obtain and therefore presented as missing. This is one specific case, but I feel confident in saying that a significant amount of variance was removed that may have further explained the data. If I was reporting this to a superior, I would have used imputation techniques to fill in the missing values, in turn preserving observations that held variance.

1. Should the data be normalized? Discuss what characterizes the components you consider key.

Yes, the data within this data set should be normalized. When conducting PCA without normalizing the data set, it appeared that variables such as In-State Tuition, Out-of-State Tuition, Application Received, and Application Accepted etc. had the largest loading weights. After investigating the ranges of every variable in the data set, it was found that these variables also had the largest range. Hence, these variables were ‘dominating’ the PCA technique when searching for the greatest variance. Once the data set was scaled and PCA was conducted, the variables with the largest loading weights changed. And included variables such as Public/Private, Estimated Book Costs, Estimated Personal $, etc. which were not previously included.

When determining the components to consider key, I use two concepts. Firstly, I will create a Scree Plot and observe where the ‘elbow’ occurs. In other words, when plotting the number of components vs. the variance, I am looking for the point where adding more variables does not provide a significant increase in variance. In both scree plots, this occurred at three principal components. Secondly, when summarizing PCA, you will be provided with a cumulative proportion for the variance explained. Depending upon the analyst and the number of variables in the data set, you may decide to choose the number of components that correspond to explaining 90% of the variance, of 80% of the variance within the data set. Notice, that in my analysis above, I chose six principal components which goes against the Scree Plot, because I wanted a higher percentage of the variance being explained. This is simply a judgement call by the analyst which is common for unsupervised learning methods.

**Part C: Final Project Proposal**

Please write one paragraph describing your final project proposal. Please include these items in your project proposal:

* Business question(s) you want to answer or solve
* Data source(s) you plan to use
* Supervised and/or unsupervised learning methods you plan to use to mine the data source(s). A minimum of two methods is required.

Business Problem

. ………………………………………

The goal for every high school principal is to see every student walk out of their doors with a diploma in hand ready to take on the world. Unfortunately, the harsh reality is that not every school can meet this. Many individuals believe that factors such as class size, location and even teacher salaries impact this goal, but I have never found anyone to justify their claims beyond hypothetical reasoning. The goal of my project is to see if there is data to support such claims and whether there are other factors that contribute to graduation rate.

Data Source(s)

. ………………………………………

The Massachusetts Department of Education publicly posts reports on enrollment, class size, teacher salaries, graduation rates, and advanced placement. Nigel Dalziel has compiled all the data into one source located on Kaggle (https://www.kaggle.com/ndalziel/massachusetts-public-schools-data). The data set contains 1,861 observations across 321 variables. After initial filtering it appears that there are 365 observations pertaining to high schools, across 60 variables that are relevant to my business problem.

Learning Methods

. ………………………………………

I will first approach the data set with an unsupervised learning model, cluster analysis, to see if there are any groups of schools that form naturally. After my exploratory findings, this approach will give me further insight into the types of schools that this data set contains and potentially help me to classify groups.

Once I have a better understanding of the records in my data set I will use a supervised learning model, tree models, to target graduation rate. A tree model will effectively provide the variable with the largest impact on graduation rate as the first split. Utilizing training and test sets I can then compare the performance of my model to actual records.

**Extra Credit: Introduction to GitHub**

The purpose of the extra credit assignment is to introduce you to GitHub, a platform to manage and share projects. Programmers, data scientists, and analysts use GitHub to share codes and data. Please work through the tutorial in the Tutorial\_R\_Project\_Git\_No\_Command\_Line.docx file. To earn the extra credit point, you will need to send me the link to a repository with the **crash.csv** data set and **.Rmd** file for Part A. You can include the link in the file for Parts B & C of the homework assignment.