BEAM079J

Coding Analytics for Accounting and Finance

Assignment 2022

30 credits (100% of Final Grade)

Deadline: 12.00pm 2nd December 2022

Word Limit: 7500 (Includes tables, references, appendices)

Submission to include:

(1) Written report
(2) Python code

Answer **one** of the following questions:

EITHER Assignment Option(A)

OR Assignment Option (B)

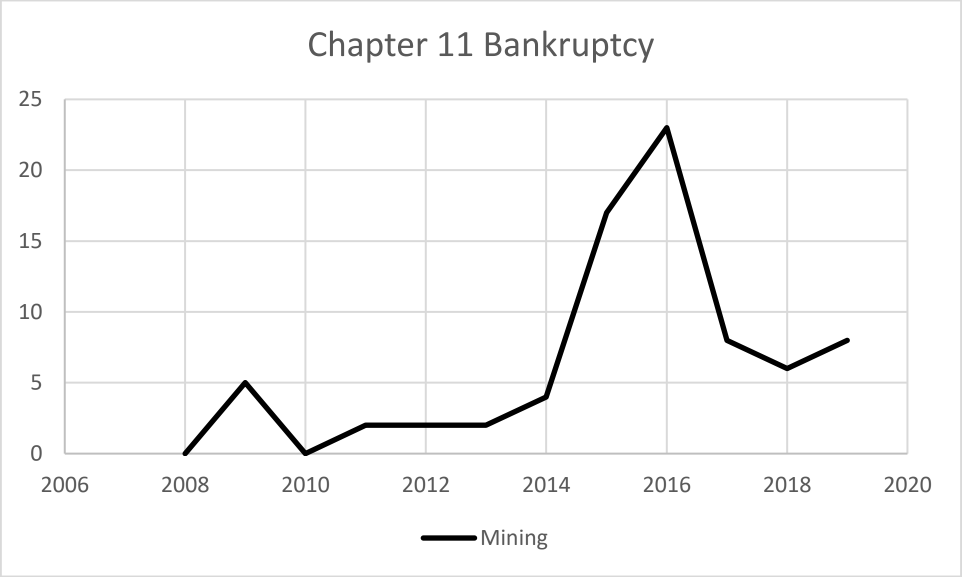
OR Assignment Option (C)

Assignment Option (A) – Neural Networks and Sentiment Analysis

Requirement

This assignment incorporates several elements of the BEAM079 module and requires that you perform an empirical investigation and discussion of the following scenario - building upon the work you may have carried out in BEAM078 Applied Empirical Accounting.

In 2015/2016 the US market saw a record high number of Chapter 11 Bankruptcy filings across the mining industry – unprecedented in recent times.[1, 2]



Using financial data, sentiment analysis, and neural networks, your task is to provide empirical analysis and establish whether these bankruptcies could have been predicted, utilising a model that can be used to predict mining industry bankruptcies one year prior to failure.

Incorporating a mixture of financial and textual data, you are tasked with analysing the annual reports of 27 bankrupt mining companies, and 73 non-bankrupt companies (the training sample) during the years 2015/2016. The list of companies, the annual reports, and some financial data has been provided[3]. Your primary analyses will be based on this "training" sample of 100 firms. I have also provided you with data on an additional 26 (13 bankrupt, 13 non-bankrupt), for validation purposes if required.

Using your python skills acquired within this module (and elsewhere), your report should seek to provide (but not limited to):

[1] https://www.mining.com/warning-of-another-string-of-mining-bankruptcies-in-2016/

[2] https://www.sierraclub.org/sierra/2016-3-may-june/feature/coal-industry-bankrupt

[3] (one year before failure). I have provided pdfs and some basic financial information here. It is perfectly fine to collect more data as you see fit.

An **introduction** – Discussion of the mining industry and recent bankruptcies, what are the potential causes, and what are your research goals?

A **literature review** – You have 3 strands of literature to discuss namely: Bankruptcy prediction literature, sentiment analysis literature, neural network literature. Remember that a literature review should be a cohesive discussion of extant literature, how it relates to your research, and is the main indicator of whether you understand the subject area or not.

Methodology - What statistical tests/models are you going to perform and why?

Data – A description of the data and its origin, including summary statistics of the key variables.

Univariate analysis – The inclusion of (but not limited to) tests of difference (t-tests) between the two groups of firms in terms of both financial and textual data (tone and readability); correlation – ensure/demonstrate that there are no extreme correlations that will affect your multivariate neural network; a comparison of bankrupt and non-bankrupt word clouds.

Multivariate analysis – Neural Network(s). A predictive model that uses both financial and non-financial data. How accurate is it (ROC AUC)? Some extensions might include - Is it better having textual data in the model rather than just having financial variables? Does it retain its accuracy when applied to the validation sample? How much more accurate is a neural network than a logit model – in terms of both training and validation? Weighted vs unweighted sentiment (advanced).

Conclusion – What are your key findings and what are the implications?

Python code – A code (*.py) file which documents each stage of your analysis (uploaded as a separate file).

Notes:

- The provided annual reports were collected from Perfect Filings (UoE database library).
- The provided financial data was collected from WRDS compustat.
- The list of failed companies was collected from the Lopucki bankruptcy database.[4]
- The list of sentiment words from the Loughran McDonald dictionary is provided.

Feel free to collect any additional information that you feel is necessary to your study.

[4] https://lopucki.law.ucla.edu/

Assignment Option (B) – Fraud Detection

Requirement

This assignment surrounds fraud detection and asks you to investigate the application of Benford Law as a tool to detect financial reporting fraud.

A list of 45 companies has been provided for you. These companies have been determined by Audit Analytics (available through WRDS) to have not only misrepresented their accounts; but have done so in a fraudulent manner. The details of each fraud and the account-years which are affected are provided.[5]

Using financial data, your task is to provide empirical analysis and establish whether these fraudulent accounts could have been predicted by Benford Law.

In particular, you should analyse these fraudulent reports and compare the results to companies that are deemed to be clean. The clean companies can be any of your choosing, but the clean sample must be comparable to those which are fraudulent (e.g. in terms of size, industry, profitability). [6]

Your analysis should include a comparison of MAD, KS and Chi-squared test for both groups – including tests of difference.

You are also required to perform separate analyses, not only on the financial data as a whole, but according to type of information i.e. Balance sheet, Income Statement, and Cashflow items. Comparisons should be made to the findings of Amiram et al (2015).

You may also wish to take the type of fraud being committed into account when performing any analysis.

No financial data is provided – you will need to manually collect this yourselves through, for example, Compustat (WRDS).

Using your python skills acquired within this module (and elsewhere), your report should seek to provide (but not limited to):

An **introduction** – Discussion of Benford Law, the potential application to fraudulent account detection, what your research goals are, and what you hope to achieve through your analysis.

A **literature review** – You have two strands of literature to discuss, namely Benford Law, and fraud detection. Remember that a literature review should be a cohesive discussion of extant literature, how it relates to your research, and is the main indicator of whether you understand the subject area or not.

[5] There are potentially 126 total years of fraudulent accounts that were required to be restated. Only 39/45 companies had sufficient accounting data (available through compustat) at the time of writing this and therefore your final analysis may result in fewer firm-years.

[6] You may infer that any company which has not been determined by Audit Analytics to be fraudulent, is a "clean" company.

Methodology - What statistical tests/models are you going to perform and why?

Data – A description of the data and its origin, including summary statistics of the key variables.

Univariate analysis – The inclusion of (but not limited to) tests of difference between the two groups of firms in terms of the financial data acquired, along with separate analysis of balance sheet, income and cashflow statement items.

Multivariate analysis – No multivariate analysis is required but you may wish to investigate Amiram et al’s proposition that misstatements are more likely to occur in smaller, younger, more volatile, growing firms. This may be done by using key independent variables in order to explain your Benford conformity measures by way of a regression. Remember that the companies I have given you represent fraudulent misstatements only, and therefore may not adhere to the findings of Amirim et al (2015). You may wish to uncover evidence as to the type of companies that fraudulent misstatements are more likely to occur.

Conclusion – What are your key findings and what are the implications?

Python code – A code (*.py) file which documents each stage of your analysis (uploaded as a separate file).

You may expand or explore this topic in any way you feel appropriate - provided that the key areas highlighted above are covered.

Assignment Option (C) – Social Network Analysis.

Requirement

This assignment surrounds Social Network Analysis and requires you to evaluate a network of company directors.

Using data from BoardEx (available via WRDS), you are tasked with compiling a network of company directors using your knowledge of Social Network Theory. The network should consist of nodes (the directors) and edges. The edges between each node should be connected only if the directors sit on the same company board in the same year.

Your network should span at least one whole year and include the latest full (calendar) year of data from BoardEx (currently this would be 2021). BoardEx has three databases of companies and their directors, US, UK, and Europe; you may use any **one** of these for your analysis.[7][8]

Your analysis should include a statistical investigation into the network you create. Who are the key, central and influential players in the network? Why are they so influential? Do they often represent a particular type of company or companies, or have particular characteristics?

To answer these questions, you should provide information regarding the network in terms of the centrality measures provided to you in class: Degree, Betweenness, Closeness and Eigenvector centrality.

Using your python skills acquired within this module (and elsewhere), your report should seek to provide (but not limited to):

An **introduction** – A general discussion of Social Network Analysis, the potential application to corporate director networks, what your research goals are, and what you hope to achieve through your analysis.

A **literature review** – You have one main strand of literature to assess, Social Network Theory and in particular, its utilisation in the area of business power and influence. Remember that a literature review should be a cohesive discussion of extant literature, how it relates to your research, and is the main indicator of whether you understand the subject area or not.

Methodology - What statistical tests/models are you going to perform and why?

Data – A description of the data and its origin, including summary statistics of the key variables where appropriate, including the size of the network and other pertaining information.

[7] Using BoardEx is the easier option, as I know that the data exists here. You are welcome to investigate directors of companies within any country – but it will be up the individual to source the necessary data.

[8] Data can be collected from **WRDS - BoardEx - (REGION) - Organization Summary - Analytics**
From this database you can get information on directors. the key things you need to calculate the edges are:
Company ID and/or Board ID, Director ID/name, Annual report date

Analysis – You are required to analyse the network as a whole and discuss the most influential directors across the network. Other options would be to look at individual business sectors independently, and/or use the size (e.g. market value) of each company as edge weights - in order that directors of larger companies are (appropriately?) given more influence than those at the helm of smaller companies. For an additional analysis (time permitting) you may also wish to perform a regression that utilises individual director characteristics (also available on BoardEx) in order to explain those with high centrality scores. For example, are these directors of a certain age compared to their peers? Did they go to a particular School or University? Have they been on the board a long time? Inside or Outside directors?

Visualisation – Visualise your network. This can be done in any way you feel fit. Make sure to explain your visualisation and how it was created. You may wish to colour code the network by community, or by industry for extra visual appeal. The examples and links given to you in class will assist you on this.

Conclusion – What are your key findings and what are the implications?

Python code – A code (*.py) file which documents each stage of your analysis (uploaded as a separate file).

You may expand or explore this topic in any way you feel appropriate - provided that the key areas highlighted above are covered.