MASTER OF SCIENCE IN ACCOUNTING (DATA AND ANALYTICS)
ACCT653 FORECASTING AND FORENSIC ANALYTICS

TEAM PROJECT INSTRUCTIONS

Aim

The group project aims to provide students with an understanding of emerging topics in financial forecasting and forensic accounting. Students should be able to use machine learning algorithms and other analytics tools to critically analyze and evaluate various aspects of a given topic and dataset. The research topics are presented in the Appendix.

Composition of teams

The number of team members depends on the size of your enrolled class section. All group members should be registered for the same class section. I reserve the right in my ultimate discretion to allocate group members.

Deadline

Deadline for ***softcopy*** submission of the project report (including both report and code), all data used, and presentation slides is ***1159pm, Sunday, 17 March 2024***. No extension of the deadline date will be entertained. Early submission is welcomed, although no bonus marks will be awarded.

Online Submission Instruction

You are required to submit electronically the following contents: *project report (including both report and code) in the qmd or Rmd format, the HTML and PDF versions of your report and code (after render or knit from the qmd/Rmd file), all data used*, and *presentation slides* through the official SMU OneDrive.

You are required to use the R programming language and the Quarto or R Markdown file format (.qmd or .Rmd). You are required to send me evidence of consent from all group members if you prefer using Python instead.

The cover page of the report should clearly state the title of the research report, the names (**as your matriculation names and in alphabetical order**) and Student ID

numbers of all the team members, class section, the date of submission (month & year will suffice), and the instructor's name. The file name should follow the prescribed format: groupnumber_topic (e.g., G1_FraudDetection.qmd). It is advised not to use space in file names.

All submissions must be through the official SMU cloud drive OneDrive. You are advised to save all files into one folder (with subfolder structure when necessary) and use relative path in your code, thus I can reproduce your code without changes of any code. No zipped files please.

If any chunk of your code runs for more than 1 minute on your computer, please clearly state it at the beginning of your code chunk.

If all your code runs for more than 3 minutes on your computer, please clearly state the running time at the beginning of your report.

Project outline

Your written report should include, but not be restricted to, the following:

1. An introduction section to present a preview of the main points of the in-depth report. The introduction should contain enough information for a reader to get familiarized with what is discussed in the full report without having to read it.
2. Exploratory data analysis including but not limited to an in-depth look at the main data features (dependent variable and some main independent variables), missing values, outliers, categorical variables, distributions, and correlations of the dataset. If you are using external data, clearly state the source of the data and how you obtain the data. You are expected to share the code if you scrape any data from the internet. All data collection procedures should be clearly documented and reproducible.
3. Model selection, implementation, evaluation, and refinement: explain your reasoning for model selection. Train your models using the training dataset and test your models using the testing dataset. Evaluate your models using the metrics provided for the topic. You may use any statistical algorithms introduced in this course and other courses. Explain the process your group took for refining/iterating on your model.
    a. You are expected to train at least three (more is good) different types of algorithms; suggestions include: linear models such as OLS and Logistic; regularization models such as Lasso, Ridge and Elastic Net; Bagging and boosting models such as XGBoost and Random Forest; Neural Networks such as autoencoder and CNN; Automated machine learning models such as H2O and AutoML.
    b. Ensemble your models using a blending or stacking technique -- this could be as simple as a straight average, a majority vote, or a stacked predictor (linear or other model type).
4. Conclusion: how well did your model do? What lessons did you learn from this project? Make a conclusion to close your project report. It is recommended to end the report with an appropriate, meaningful final sentence that ties the whole point of the report together.

Presentation

Each group is required to give a *20-minute presentation* (including Q&A and class discussion) in class. All group members are required to take part in the presentation. Presentation order will be randomly scheduled in the last session.

Assessment & Evaluation

The team project report carries 30% of the total marks for the course. Credit will be given for good analysis, quality content and rationale, prediction accuracy, creativity, good writing and programming style, insightful report, and an interesting oral presentation.

Academic Integrity

All work is to be performed exclusively by the members of the team and all team members must contribute their fair share to the project. Teams are not to consult with other teams, nor to obtain help from any other persons within or outside of SMU. If outside research is used, sources are to be cited and information discovered via outside research is to be clearly labeled as such (in appropriate footnotes or in a bibliography appendix). If outside research is performed, the product of your research is not to be shared with any student who is not a member of the team.

Unless I hear from you otherwise, I will assume that all members of your team are contributing their fair share to the project and I will award the same marks to all members of your team. If you have strong grounds to contend that a particular member of your team should not receive the same marks as the other member(s) including yourself, you should send an email to me independently under confidential cover explaining clearly the reasons for your allegation with concrete evidence not later than the deadline date for the submission of your team report. I will evaluate each allegation on its own merit, but my ultimate decision with respect to the award of marks to each team member shall be final.

I view very seriously any plagiarism of previous terms' project reports done by other students. The penalty for such plagiarism is an "F" grade for the whole course.

Appendix A Project Topics

A selected project topic from the following will be pre-assigned to your group on eLearn. Check eLearn to confirm your project topic.

Topic 1: Predicting Changes in Annual Reports

This project tasks you with *predicting* how much US companies will alter their annual reports in the year 2012. A set of five years of historical data for the same companies is provided for you. Annual reports cover an extensive set of information about companies, including their history, primary business, investments, industry, outlook, risk factors, financials, and more. However, there is a phenomenon of "boilerplate language" in annual reports – much of these reports are the same year-over-year (Make disclosures more relevant, UK FRC urges (fm-magazine.com)). As such, you will predict how much companies' annual reports will change year-over-year. You will be allowed to use any outside data to build your model.

The project is hosted on Kaggle. It is a private competition, so don't share the following link with outsiders.
https://www.kaggle.com/t/0ece10a279dd4aac96d8e91796d0df97

Topic 2: Predicting the number of Twitter followers

This project tasks you with *predicting* the expected number of Twitter followers for US corporations. A set of daily counts of Twitter followers for select companies during the year 2017 will be provided. You will be tasked with modelling the changes in followers and will need to forecast this out beyond the time frame of the provided data. You will be allowed to use any outside data to build your model.

Please note that some companies have missing financial data in 2017 because they have been delisted from stock exchanges due to mergers and acquisitions. If you want to include their financial data, you may retrieve older financials from Compustat.

The project is hosted on Kaggle. It is a private competition, so don't share the following link with outsiders.
https://www.kaggle.com/t/6aaa3c5cba29456a81fe2f1d75473761

Topic 3: Detecting Intentional Accounting Errors

This project tasks you with *detecting* intentional misstatements by US corporations. These are accounting errors where either the company itself admitted to intentionally misstating their filing, or the US government (SEC or DOJ) or investors (by lawsuit) determined this to be the case. A set of multiple years of intentional misstatements will be provided to you, linked with company identifying information. You will be tasked with detecting intentional misstatements in the year following the provided data. You will be allowed to use any outside data to build your model.

The project is hosted on Kaggle. It is a private competition, so don't share the following link with outsiders.
https://www.kaggle.com/t/fd8cab5fe806404aa91e97c3832b1ba5

Topic 4: Forecasting GDP Growth

This project is inspired by Datar et al. (2020).

This project tasks you with *forecasting* quarterly nominal GDP growth rate in the United States. The dataset contains estimated nominal gross domestic product (GDP) annual growth rate in each quarter from 1990 to 2020, released by the Bureau of Economic Analysis (BEA) of the United States Department of Commerce. The estimate is the “third” or "final" estimate of GDP issued in the third month after the relevant quarter which is based on more complete source data than were available for the "second" estimate issued in the second month and the "advance" estimate issued in the first month after the relevant quarter. You may refer to the BEA website for more information on the estimate of GDP in the United States. You will train models to forecast the BEA's final GDP estimate for a quarter. Specifically, you are required to answer this question: Is accounting information useful for GDP forecast? You will be allowed to use any outside data to build your model.

The project is hosted on Kaggle. It is a private competition, so don’t share the following link with outsiders.
https://www.kaggle.com/t/4576fa1ad29d4aee8b49ebd19ee846e3

Appendix B Outside Data Resources

You may be allowed to use any outside data to build your model for the project. The following is a list of outside data resources you may use. Note that you are welcome to use other data resources which is beyond the following list.

Data available through SMU WRDS subscription

WRDS is a platform for many commercial databases (Browse Data by Concept (upenn.edu)). The following is a list of databases which may be relevant to your project:

- S&P Compustat: annual and quarterly financial information for global public and private companies. You may refer the manual from WRDS here: Compustat (upenn.edu)
    - If you want merge stock price data (CRSP) and financial data (Compustat), you should use the CRSP/Compustat Merged dataset which includes linking identifier with CRSP.
- CRSP: stock prices for US public listed companies. The merging of data from Compustat and CRSP is challenging but WRDS has provided a merged database. The manual for CRSP is at CRSP (upenn.edu)
    - Daily or monthly stock prices is from CRSP/Annual Update/Stock Security Files/Daily or Monthly stock File. If you want annual stock return data, you may sum the monthly stock returns.
    - There are various ways to get market capitalization of a company. The simplest way is to retrieve it directly from Compustat/North America/Fundamental Annual.

    - Risk free rate could be retrieved from Fama-French 3 Factors Plus Momentum - Monthly Frequency.
- AuditAnalytics: information on auditors, audit fees, audit opinions, internal control (SOX 404 and 302), etc for SEC registrants. Audit Analytics (upenn.edu)
- Execucomp: executive compensation data, including profile data of executives. Execucomp (upenn.edu)
- WRDS SEC Analytics Suite: Readability and sentiment scores and other textual analytics for SEC filings. WRDS SEC Analytics Suite (upenn.edu)
- Thomson Reuters I/B/E/S: financial analyst forecasts data. IBES (upenn.edu)
- S&P Credit Ratings: for global issuers. S&P Credit Ratings (upenn.edu)
- ESG data: Sustainalytics (upenn.edu)
- News and social media content analytics: RavenPack
- Mergers and Acquisitions: SDC

If you want to retrieve large volume of data from WRDS, you may use programming to do so. See the programming support page from WRDS at Programming at WRDS (upenn.edu).

Merging of the various data sources is always a challenging task. Fortunately, WRDS has prepared some linking identifiers for subscribers. You may read the guide by Stanford University. In general, financial data and stock prices are indexed by "gvkey", a unique identifier and primary key for each company in the databases. For non-financial data such as the AuditAnalytics, it is indexed by CIK or Ticker, both of which are also available in the Compustat. You may check which one is more accurate for merging non-financial data with the financial data for your project. My experience is that the Ticker may be a better identifier for merging, but it depends on your data.

Annual Report Text from SEC EDGAR Website

Besides the WRDS SEC Analytics Suite data, you may use R or Python to download the text version of annual reports and other filings directly from SEC EDGAR website. It could be used for textual analytics to generate more features for your project.

I tried the R package "edgar" (https://github.com/Gunratan/edgar) and it worked in May 2023. It can generate sentiment measures of SEC filings automatically. Some other measures such as word count and file size are also generated. It can also download text data such as MD&A for additional processing. This may be very useful if you want to incorporate such textual measures into your models.

I have also tried the Python library "sec-edgar" at https://github.com/sec-edgar/sec-edgar which downloads text filings such as 10K and 10Q and it worked in Jan 2022. If you use Jupyter Notebook for this package, you may have to add the following code to avoid a bug.

```
import nest_asyncio
nest_asyncio.apply()
```

There is a good tutorial on scraping EDGAR data using Python: How to Web Scrape the SEC | Part 1 - YouTube. The video was posted few years ago and the process may be different. The main idea should be the same though.

If you want to code yourself, here is a workable scraper in Python: https://github.com/Theling/10-K-scraper. My ex-students tried it in 2021 but I don't know whether it is still working now. You may need to adjust it to fit into your request.

Also note that the SEC EDGAR system limits each user to a total of no more than 10 requests per second, regardless of the number of machines used to submit requests. For more details, refer to https://www.sec.gov/developer.

Older Tweets

As of Jan 2024, Twitter allows 1,500 tweets only for free account.

The Python package "snscrape" (https://github.com/JustAnotherArchivist/snscrape) works in May 2023 and you may be able to download historical tweets. A tutorial is at https://medium.com/better-programming/how-to-scrape-tweets-with-snscrape-90124ed006af.

Alternatively, you may try the Sctweet package in Python https://github.com/Altimis/Scweet. Some students tried the package in March 2022.

You may try the rtweet package at ropensci/rtweet: R client for interacting with Twitter's [stream and REST] APIs (github.com) to get tweets data. You may also try the GetOldTweets3 Python library to retrieve older tweets for analytics. Mottl/GetOldTweets3: A Python 3 library and a corresponding command line utility for accessing old tweets (github.com). There are many other R/Python packages for this purpose.

I did not try the above older tweets packages and there may be new restrictions from Twitter and new packages on GitHub.

Google Trends Data

You may try to use the Google Trends data for forecasting.
https://github.com/PMassicotte/gtrendsR
https://github.com/qztseng/google-trends-daily

The output "interest over time" numbers represent search interest relative to the highest point for the given region and time. A value of 100 is the peak popularity for the term. A value of 50 means that the term is half as popular. A score of 0 means there was not enough data for this term.

As the output is a relative number, you have to search the keyword one by one. You may have to use loop to search for multiple keywords. You may also try the "lapply()" and "purr::map()" instead of the loop iteration. You may adapt the following R code if you use the GTrendsR package.

```
library(gtrendsR)
query <- c("amd", "amex", "dell", "microsoft", "singapore",
           "sia", "kepple", "dhl", "fedex", "grab")

google_data <- data.frame()
for (i in 1:length(query)) {
  temp_list <- gtrends(query[i], gprop = "web", time = "2017-1-1 2017-6-30")
  temp_df <- temp_list[[1]]
  google_data <- rbind(google_data, temp_df)
}
```

Feel free to explore other sources of data.

--END--