Data Science Techniques and Real-World Applications

WS 2025

Case Study 1: SE / CRSP Merge

Introduction

Earnings conference calls are regular quarterly meetings between the senior management teams of a company and their equity analysts. Compared to the firms' other types of disclosures, earnings calls are very informative for market participants, as they contain forward-looking details about the firms' expected performance and direction.

A Hedge Fund approached you to develop a trading strategy based on the information contents of earnings conference calls. They have access to "StreetEvent" (henceforth, SE), a big multi-gigabytes dataset of the earnings conference calls for all the global companies. The biggest challenge they are facing now is merging this dataset with the stock return data provided by the Centre for Research in Security Prices (CRSP).

Case materials

In this task, you must find a link between SE and CRSP datasets. In particular, you must fill the column "MergeComnam" of the file "SEmappings DAFA.csv", which is provided in the Data folder. You should find the corresponding company name from the "CRSPnames.csv" file.

"SEmappingsDAFA.csv" has five columns: *Seid* is the unique file ID for each earnings call in the SE database. *SECompanyName* and *SEticker* are the name and identifiers of the companies in the SE dataset. You are 1 also provided with the column *SEHeadline*, which is the headline of the SE files. Finally, the last column *MergeComnam* is empty (NAN) and needs to be filled by you in this task.

"CRSPnames.csv" has four columns: *DATE* is the first date that the company appears in the CRSP database. *COMNAM* is the company's name in the CRSPdataset. Note that this column is always uppercased. *PERMNO* and *PERMCO* are the unique security and company identifiers in CRSP, respectively. One firm (permco) can have multiple securities (permno). E.g. Alphabet, the parent company of Google, has class A and class C shares outstanding.

The information in the *SEHeadline* column is of special importance for your task. *SEHeadline* is the only historical information in the SE database. For example, if a company changes its name at some point, SEHeadline still shows the correct name at the time of the earnings call. In contrast, SECompanyName and SEticker show todays company name and tickers, respectively. In other words, SE keeps only the most updated name of a company. For example, *Apple Inc.* was named "Apple Computer, Inc." before 2007. Nevertheless, *SECompanyName* shows *Apple Inc.* for all the records of this company. On the other hand, in CRSP, we see historic name changes. For example, we can see that *ENERGY INC* (with the PERMNO 10001) was named *ENERGY WEST INC* before August 2009. Please note that name changes are not only because of corporate actions; other events like M&As also change the names of companies in SE. For example, if company A fully acquires company B, you won't be able to find company B in *SECompanyName* anymore, while the information in the *SEHeadline* remains unchanged.

Therefore, you are expected to use the information in the *SEheadline* for the correct merge of SE and CRSP

Questions

1. How many companies have you managed to merge?
2. What are the main obstacles to a perfect merge?
3. What is special about CRSP names? Do you preprocess CRSP names before merging them? If so, what changes do you make?
4. Which tokens in the company names make your task challenging? How did you deal with them?
5. What should be (theoretically) the number of records per firms' fiscal year? What is the actual number of observations per firms' fiscal year? Is this information helpful in identifying mismatches? Explain!
6. What should be the relation between permco and the company name from the SEHeadline? More specifically, can one SEHeadlines company name be correctly linked to multiple permcos (i.e., to more than one rm in CRSP)? Can one permco be correctly linked to multiple company names from SEHeadline?
7. What additional rm information would be helpful to check the validity of a match?

How to deliver

You must submit the processed file of "SEmappingsDAFA" as well as a Jupyter notebook (converted to HTML) where you describe the steps you followed, the challenges, and a description of your results.

1. Create a Jupyter notebook.
2. Perform all your data cleanings and analyses in one notebook with proper markdown/comments.
3. Start your notebook with your names, student IDs, and the contribution (including the percentage of contribution) of each member.
4. Rename as: "Yourteamnumber_case1.ipynb" (you can find your team number on Canvas)
    - Example: team1_case1.ipynb
5. IMPORTANT: Convert your notebooks to an HTML page using "nbconvert"
6. (optionally) add a README file
7. Add the mapping file "SEmappingsDAFA_YourName_YourMatriculationID.csv"
8. Zip your root folder with the same naming
9. Upload your file to Canvas

Remarks

1. There is no perfect solution! I need to see your approaches to the problems, not your solution. Use your time efficiently!