DBA4761: Tidyverse Principles And Tidymodels

Lecturer: Rafael Nicolas Fermin Cota (bizfcrn@nus.edu.sg)

Session : Semester 1, 2024/2025

Description

This course is for students interested in financial modeling programming. Based on previous experiences, there is a great demand for students who can apply the tidy data modeling techniques covered in class. The tidy data principles is the framework used for collecting, cleaning, organizing and analyzing data using the R Programming Language. I always advise students interested in private markets investing to learn some coding and statistical learning. If they learn to code in a programming language like R, it will open new avenues for their analytical work, will vastly improve their productivity and will protect the future of their career as forces of automation and artificial intelligence rapidly advance in the financial sector.

Course Outline

Module 1: Tidyverse principles

It is becoming increasingly common for investment management firms to collect huge amounts of data over time. The incoming data is diverse – balance sheets, income statements, statements of cash flow, operating metrics, debt schedules, budgets vs. actual, segment data, month/quarterly/annual data points, etc. For private equity firms, for example, the challenge right now is to manage data from operating companies efficiently and make that data ready for analysis, but before a single insight can be derived from any operating company, there are a number of steps that need to be taken. These steps include, but are not limited to: (i) collecting and structuring the data, (ii) cleaning and standardizing the data, (iii) linking the data with other data sets, (iv) generating aggregations or new data attributes, (v) storing it in a distributed database, and (vi) making it available for investment decisions.

The aim for this module is to teach students different tidy data techniques to clean, structure, aggregate, and explore structure and unstructured data to drive insights and make better informed decisions. The foundation for tidy data management is the tidyverse, a collection of libraries, that work in harmony, are built for scalability, and are taught in this module. In class I also introduce students to dbplyr, which is the database backend for dplyr that allows end-users to easily use remote database tables as if they are in-memory data frames by automatically converting dplyr code into SQL. The dplyr package, for example, calls itself a grammar of data manipulation, but it is also an impedance matching circuit that let students write lovely composable R functions and still gain the benefits of SQL query optimization and performance that the database implementation has to offer.

Module 2: Tidymodels ecosystem

For years Python has had a solid advantage in tooling for machine learning (eg., scikit-learn), but thanks to the Max Khun and his team at RStudio the importance of machine learning application with R is expanding at a tremendous rate. The existing integration of the tidyverse and statistical modeling is especially valuable in the investment management industry. These companies are already using tidyverse solutions to track and analyze tremendous amount of data flowing in/out from each portfolio company every month. For each operating company, what is required is to turn data into information, better decisions, and stronger organization. In this module, I introduce a new collection of tidy software in the R programming language for model building, tidymodels, including supervised machine learning for text analysis and feature-based time series analysis.

Prior to tidymodels, a data scientist might have just skipped tuning a parameter, or stopped with the first model type he/she tried, because iterating would be too difficult. With the tidymodels suite of packages the steps needed to apply to a dataset before and after fitting a model can easily be implemented. Tidymodels packages, like parsnip, recipes, and rsample provide a grammar for building, visualizing, testing, and comparing models that are focused on machine learning. Machine learning is the study and application of algorithms that learn from and make predictions on data.

There is a lot of traps students can fall into when doing machine learning, and the tidymodels packages are designed to avoid them. The single most impressive one is avoiding data leakage during data cleaning. The recipes package offers functions for performing feature engineering operations, but that is only part of its value. The brilliant insight of the recipes packages is that each of those feature engineering steps is itself a model that needs to be trained. By combining a recipe with a model specification into a workflow object, students can encapsulate the entire start-to-end process of feature engineering and modeling. These conveniences are not just a matter of saving effort or time; together they add up to a more robust and effective process; allowing students to focus exclusively on scientific and statistical questions about data and machine learning model.

Sample Lectures/Projects

https://www.linkedin.com/pulse/updated-dba4761-material-rafael-nicolas-fermin-cota-qjjpe/ (**A MUST READ**)

Optional Reading List

[1] R for Data Science: https://r4ds.had.co.nz/
[2] Text Mining with R: https://www.tidytextmining.com/
[3] Tidy Modeling with R: https://www.tmwr.org/

Assessment

Continuous Assessment:

Class Contribution	20%
Individual Assignments	40%
Group Project	40%