**FINA 4359: PROJECT 2**

**Module 4, 2021**

**Overriding Objective**

Determine 1) source of data/signal/independent variable; 2) stocks/ETF/dependent variable; 3) Model Set-up (Lagged Time-Series v.s. Non-Lagged Time Series, Multiple Linear v.s. Univariant) such that at least one model produces predictive power and hence profitable trading strategy.

Tasks:

Determine Signal

* Foot Traffic (Consider Safegraph which Prof can share with us a SQL DB)
* Covid data and Covid Vaccination
* News/Social Media Sentiment (w/o NLP to make our life easier) -->? Media
* TSA Air Travel Passenger Number (DONE)
* Road Congestion (DONE)
* Box Office (DONE)
* Web Traffic (DONE)
* Hotel Checkin Data and Casino Data
* Others…..Make sure we can cover as many aspects of an economy as possible Consumer, Industrial, Media Reporting, etc)
* Restaurants
* FootTraffice → ser

Airfare

Determine Securities:

* JETS ETF,
* Individual Airlines
* SP Plus (The Shape of TSA reminds me of the shape of trading performance of this stock)
* AMC/CNK
* Hotels (Hilton, Hyatt)
* Offices
* Casinos
* EPR/Other Leisure & Entertainment REITs
* Revenue of Travel and Leisure Business

Indicator Construction:

Wrap 3 sources into 1 signal

* Wrap all Signals as one single indicator/number/independent variable by equal-weighting the daily return of all individual signals

Determine Model (Easy Stuff)

* Time-Series (ETF\_Tourism(t) ~ indicator(t) + c)
* Lagged Time-Series (ETF\_Tourism(t) ~ indicator(t-10) + c)
* Multiple Linear Regression → Q.E and Interest Rate Change
* Univariate Regression against only one of the signals

Rex to run a regression on TSA wrst JETS and get guys a sense of how the regression looks like

Tasks Done:

* Rex to throw the R Codes on Air-travel, Congestion, Box Office on Drive

**What to provide**

1. **What to submit**

A write-up, including code, documenting what you did. And code.

1. **Instructions**

**Username: fina\_4359\_u3548379**

**Password: 856e7**

Your mission is to

* Identify and collect, strongly preferably via automation, a dataset
* Show me your database diagram, is it logical to store the data in raw form or did you transform it and why?
  + → Data Transformation with respect to frequency(weekly)

Start the dataset from 14th Jan

|  |  |  |
| --- | --- | --- |
| Source/Description | Timeline for Data, Format, Updated or not | What transformations are possible? |
| Web-Traffic Data | Jan 1st 2020 - 10th May 2021, CSVs, No live data due to API calls |  |
| TSA & JETS ETF | Jan 3rd, 2020 - 13 May 2021 (Can Adjust Start Day), Data.Frame (Can export into CSVs), Scraping, Updated everyday |  |
| Trips by Distance  library(RSocrata)  data = read.socrata("https://data.bts.gov/resource/w96p-f2qv.json") | --- March 13th, 2021, CSV, Last updated April 17th 2021 (I presume monthly updates) |  |
| Road Congestion | 1st Jan 2021-17th May 2021, Data.Frame, Scrapping, Updated Daily. |  |
| Box Office | 2000 - 14th May 2021, Data.Frame, Scrapping, Updated Weekly. |  |
| Airline Employment  https://www.transtats.bts.gov/Employment/ |  |  |
| <https://aspm.faa.gov/>  Airline flights | Available within 60 days after the end of the month |  |

* From the data, what indicator will you create? Show me the summary statistics of such an indicator. Do your statistics make sense?

1. **Test all individual signals with the JETS ETF and see the correlation between → 5 data sources 5 X 4 Regression(try to get more significant regression) →**
2. **For each database we try to come up with the regression that has the highest correlation and significant values and select that for indicator.**
3. Build Make a final indicator that is (equally weighted, and ) with above significant signals.

Indicator = Sigma 1 to n (1/n\*highest correlation for each database)

Y = alpha + B(Indicator)

1. Multiple linear regression → use the 5 top correlations Y = alpha + B1X1 + B2X2 + …..

**(As Benchmark)**

* What is the economic purpose of this indicator, how will you apply it? Justify the indicator through economic reasoning, citing papers or businesses that apply this model, etc.
  + To go the extra mile, you can actually show predictive relationships that you claim. If you think that measure X predicts GDP, then show that!

1. **What you are being graded on**

Sophistication of analysis and communication. A key challenge with data science in organizations is that most stakeholders are non-technical. Therefore, if you communicate with jargon, you will confuse people. Don’t show me a bunch of screenshots, give me a powerpoint or writeup that makes sense that you’d be proud to show anyone.