Project Proposal

Machine Learning for Macro Diffusion Indexes

Student Contact Details

Name: Rishi R

Email ID: rishiramesh08@gmail.com

Contact Number: +91 8618738092

GitHub: https://github.com/Rishi0812

LinkedIn: https://www.linkedin.com/in/rishi0812/

Geographic Location: Bangalore, India

Institution: Jain Deemed-to-be University, Bangalore

Program: Bachelor of Technology in CSE with Specialization in Artificial Intelligence & Machine Learning

Stage: 1st Year (2nd Semester)

Student Bio

I am Rishi from Jain Deemed-to-be University, Bangalore, India. I'm currently pursuing my Bachelor's degree in Computer Science & Engineering with a specialization in AI & ML. I had previously participated in the Google Code-In Contest 2019-20 for the organization R Project for Statistical Computing and was the runner-up of the competition.

Throughout my GCI journey, I found R Programming as a very interesting language and was astonished by its computing skills in all the fields. I would love to continue contributing to the R Programming language, and GSoC would be a sublime opportunity for me to go ahead with it, I have been following R and its packages even after the competition. I have also been improving my R Skills along with various other languages such as **C, C++, Python**, etc.

GSoC would enable me to level up my journey which started in the GCI Contest, I have developed a keen interest in AI n ML, Open Source organizations and its tools would help me a lot throughout my journey, it would be my pleasure if I could learn and develop under your precious guidance for an esteemed organization like R Project for Statistical Computing.

I have hands-on experience in R Programming and its basic machine learning capabilities. I have previously worked on Data modification & analysis, data visualization, package testing and management,

creating a vignette for packages, etc. I'm also familiar with popular packages of R such as data.table, dplyr & ggplot2. I went through your project details and found it very insightful and in my skill sets. It would be a great opportunity of learning and developing with you for the project Machine Learning for Macro Diffusion Indexes.

Project Synopsis

The Project Machine Learning for Macro Diffusion Indexes aims on creating series of potentially useful diffusion indexes and the data that may be used to construct them.

Then applying random forest and/or other appropriate machine learning techniques to the data, with the goal of demonstrating the relative performance of those methods.

Project Details

Introduction

Macroeconomics (from the Greek prefix makro- meaning "large" + economics) is a branch of economics dealing with the performance, structure, behavior, and decision-making of an economy as a whole. For example, using interest rates, taxes and government spending to regulate an economy's growth and stability.

The main goal of the project is to help macroeconomists to obtain useful insights from a dataset as a whole with the help of useful Machine Learning Algorithms by creating it's potential diffusion indexes.

FRED-MD is a large, monthly frequency, macroeconomic database that was organized with the goal of establishing a convenient starting point for empirical analysis that requires "big data." It is publicly available, updatable using the FRED database, and convenient in that it manages data changes and revisions on behalf of researchers. The authors of the database acknowledge that such data can be useful for constructing diffusion indexes and studying business cycle chronology.

Creating a R Package

To parse/obtain the FRED-MD database into R, We will be using an existing R package called **'fbi'**. This package should allow us to successfully import our datasets into the R Environment.

Further, for the main part of the project we will be focusing on grouping the different base factors for a particular macroeconomic indicator into **modules**. Since the factors for some macroeconomic indicators may vary upon the economists' choice, we will introduce a function in order to **delete or**

add any factors into the main modules in the users database. Hence by this method every user will have the flexibility to modify their own macroeconomic indicators in their respective databases.

The potential base macroeconomic indicators that we would be adding to our package initially will include but not limited to:

- Interest rates
- GDP growth rates
- The stock market
- Production and manufacturing statistics
- Labour market statistics
- Bond yields

We will be creating separate modules for each indicators in which we will add it's corresponding factors from the FRED-MD Database.

Further, after thorough data processing, users will be able to have direct access to latest macroeconomic databases directly through this package. Users can now use R packages such as 'rpart' or any other means to successfully apply ML Algorithms such as Random Forest or any other meaningful method directly on the dataset obtained/created from our package in order to obtain insightful forms of relative performance of the data.

Later, we will be concluding the project by preparing a thorough vignette of the package including all the details in depth.

Benefits to Community

Macroeconomic indicators are important to any trader because they can have a significant influence on market movements. This is why most fundamental analysis will incorporate macroeconomic indicators.

There is no way to be certain that these indicators are reliable on their own, but they do have a role in shaping the economy. Even if these indicators just influence other traders to open and close positions, this can be enough to create volatility in the market. Market participants will be keeping an eye on analysts' predictions of the data ahead of their release. The bigger the difference between the analysts' predictions and the actual figure, the more volatility can be expected in financial markets – as positions are adjusted to reflect the actual figure.

From the above proposed R Package and methods, interpreting and visualizing a particular macro economic database will be more efficient and easier to all the economists and users.

Timeline

Having an organized schedule is the most important factor when it comes to handling a project as it increases productivity and allows us be consistent.

I have divided the the project timeline into 5 Phases, Each phase contains **2 weeks** each and it's respective deadlines.

I understand the valuable time that the mentor's volunteer for the project, hence the weekly discussions and the evaluations that would be required will be done as and when they are most suited and comfortable according to them.

Community Bonding:

May 17, 2021 - June 7, 2021

print("Hello World to the R Community")

- I would use this time frame to get to know more about the R Community & it's culture.
- Interact with my peers & the mentors and trying to understand different work environments.
- Get ready for the project by bridging up for any required skill gap.

Phase 1:

June 7, 2021 - June 21, 2021

- Creating a base package.
- Creating different modules for different macroeconomic indicators.
- Discussing base factors with the mentors for the modules.

Phase 2:

June 21, 2021 - July 5, 2021

- Writing functions to parse different factors from the FRED-MD database.
- Incorporating the factors into modules.

Phase 3:

July 5, 2021 - July 19, 2021

- Testing the function for adding the factors into the modules.
- Completing about 70-80% of the modules before the deadline of first evaluations.

Phase 4:

July 19, 2021 - Aug 2, 2021

- Complete the remaining modules.

- Writing a function to allow the users to edit the factors in the module.
- Thorough testing of the package.
- Applying ML Algorithms and check the databases.

Phase 5:

Aug 2, 2021 - Aug 16, 2021

- Writing a vignette containing all the info in depth.
- Finalizing the complete code and documentations.
- Applying different tests with series of data varying in the time period.

Wrapping Up:

Aug 16, 2021 - Aug 23, 2021

- Finalizing the Code checks and documentations.
- Getting final reviews and checks from fellow peers.
- Submit to Google team.

Related Work

Since I'm from a Computer Science background, I started researching on unfamiliar terms and concepts related to macroeconomics, I read 3 related research papers and studied some similar projects on the web.

I researched on potential data sources [FRED-MD] along with a R package called **'fbi'** which allows us to easily obtain the monthly data provided by fred into the R workspace and also potentially useful & important areas for diffusion indexes. I also acquired some knowledge on R's Tree based models and Random forest ML Algorithm from the suggested courses. Apart from these I completed all the tests asked for the respected project.

Test Link - https://github.com/Rishi0812/GSoC_2021--ML_for_macro_diffusion_indexes_tests

Why Choose Me?

As mentioned earlier, I have hands-on experience with working in R Programming language which includes but not limited to **Data modification & analysis, data visualization, package testing and**

management, creating a vignette for packages, etc and also familiar with popular packages of R such as **data.table, dplyr & ggplot2**.

I also have good amount of knowledge in the R Markdown and adhere to the good coding standards prescribed by Google's R style guide. Apart from this I am familiar to both Git and Github and have been using it for quite some time. Also I have complete working familiarity in both the windows and the linux kernel Operating Systems such as Ubuntu.

Apart from these, I also do regularly invest and trade in stocks and cryptocurrencies which makes me understand the basic economic terms and factors which may come handy while developing the project.

I have also tried to fulfill all the criteria required for the project and would continue to gain & upskill my knowledge throughout the whole journey.

After GSoC

GSoC would be an initial part of the journey, but I believe open source is a never ending wonderful & insightful journey.

I will continue maintaining and developing what we would have started and also continue upgrading the limited dataset support to different varieties of data which would involve all of the data around the world & it's different categories. I'll also be looking for more efficient machine learning algorithms for upgrading our computing skills over time as it is a rapid developing field.

****

This pdf is entirely created using RMarkdown by Rishi R