Finance and Consumer Behavior Group DMP Outline

Consumer Questions

1. What kinds of things influence how people spend their money and on what. How do these factors connect to financial security and well being

Consumer 1.1 Project Overview

Research objectives: To get a better understanding of the consumer behavior when it comes to financial status and well-being.
Possibly see how they compare to previous times and how it was affected by different global events

Data types and sources:

Data sourced from websites with pre-existing data, eg. Consumer Financial Protection Bureau
As well as some possible api data about consumer behavior if we can access it
Team responsibilities:

Create a finished product together bringing multiple pieces into one, a visual display along with a written report that work in tandem. For the data we use, to become more familiar with it and do some exploratory analysis into the consumer behavior related to financial status

1.2 Data Collection (2pts)

Collection methods and protocols:

Possibly using a pre-existing dataset and explaining where it came from and making sure it is reliable and as unbiased as possible. Potentially using API key to find consumer data from a larger enterprise
Quality assurance procedures:

Checking for missing data and gaps in time, making sure data is representative of the population we want to study
Ethical and legal compliance:

Making sure its public data, no sensitive info, or have licensing to use the data and report on it

1.3 Data Organization

File naming and organization:

Shared drive folder and/or github, using folders to keep things of similar nature together
Version control strategy:

If done in github, push/pull will be used to control the version

Documentation standards:

Using markdown files with lots of comments explaining thoughts and processes throughout the code for editing purposes and reproducibility

1.4 Data Storage and Security

Storage infrastructure:

Github repository

Backup and recovery plans:

Using backups of the repository weekly or biweekly to save our data

Access controls and security:

Limit access to repository to the group and instructor

1.5 Data Sharing and Access

Sharing timeline and platforms:

After class or after finishing work, push to github at least weekly to update the group on work that has been done in regards to the project. Also provide the finished dataset once cleaned and ready

Access restrictions and licensing:

Cite and callout any sources used to acquire data. If public access no licensing issues should arise

Community standards:

Keep code and data readable not just to the group but as if a stranger was looking in. Note anything that helped with progression or research of any sort.

1.6 Data Preservation

Long-term storage plans:

Keep a version history on github. Optionally could also keep flash drive saves for later use. Post project there are no long term storage plans

Format migration strategy:

Keep code in a single language or use chat gpt with assistance translating to other languages. Also keeping all files in csv format

Succession planning:

Keep access to the github repository and upkeep as time goes on

1.7 Budget and Resources

Personnel time allocation:

About 4-6 hours a week, 3 being in class time and any extra time spent outside of class.

Technology and infrastructure costs:

As of now we foresee no cost for tech and infrastructure but that could be subject to change

Long-term preservation expenses:

Github repository

1.8 Compliance and Ethics

Regulatory requirements:

Keeping personal and private data out and following rules and restrictions on API keys if applicable

IRB approvals:

N/A

Privacy protection measures:

Will be void of personal data for all privacy concerns

Technical based questions:

1. How does hedging (e.g. delta or vega) work for options, and what are some example strategies? Given historical daily option prices, can you test the performance of one or more of these strategies?

2. What are other examples of branching processes in finance? How well do these models fit empirical data? This could give us a sense of other places where probability/moment generating functions could be applicable. Finding which branching processes work in finance gives a random but structured set of probabilities, and can allow traders to possibly predict prices or economic movements with a mathematical function.

Went with option number 2, but can branch to option 1 if wanted as will use a similar dataset

Technical 1.1 Project Overview

Research objectives:

- Identify and formalize branching processes that are relevant to finance (ex: order flow clustering can be related to days that maybe a bank defaulted, and then how price movements of equities moved in this cluster)
- Fit these processes against empirical financial data
- Moment generating functions might be possible to be made if these branching processes become relevant
- Examine whether these "cascades of price movements" can be modeled as a branching process (a model that basically builds a structure tree around semi-random events).

Data types and sources:

- High frequency order book data
    - https://polygon.io/pricing
    - https://lobsterdata.com/
    - Yahoo finance (yfinance) python library
- Dates from events that sparked large financial events (bankruptcies, defaults, general world events)
- Can simulate other branching processes to compare to

Team responsibilities:

- Become familiar with stochastic branching models and what the theory is behind them
- Acquire the data and process/ clean it to be ready to model
- Define specific branching models based on research and try to fit these models to the data
    - Fit models using statistical fit tests and MLparameters
    - Goodness of fit test
    - Simulation based comparison using ML
- Implement AI Agents to automate pipelines to process incoming data and run how branching models fit
- Simulate branching model and then compare to actual price values to find how well it fits the actual market.

- Report the what branching models worked and didn't work

1.2 Data Collection (2pts)

Collection methods and protocols:

- Gather financial equity time series data using the API of chosen data source
- Create fake data to make stochastic branching model structure on before training on the actual financial data (sanity check)

Quality assurance procedures:

- Make sure timeframes are aligned and that there are no time gaps in the data.
- Make sure no actual financial data makes its way to the branching model before training.

Ethical and legal compliance:

- Make sure data can be used for academia. Do not use any sensitive information.

1.3 Data Organization

File naming and organization:

- Keep each branching technique to its own folder within github. Use structured naming to keep track of files.
- Keep each assignment to its own folder

Version control strategy:

- Github repository (pull and push)

Documentation standards:

- Write all code in Markdown (better explainability of what each block of code does)
- README goes into more depth about findings and what is being worked on
- Data directory

1.4 Data Storage and Security

Storage infrastructure:

- Github repository; can store data in cloud

Backup and recovery plans:

- Can backup the repository weekly and mirror the repository

Access controls and security:

- Keep repository access to only instructor and group members until time for submission

1.5 Data Sharing and Access

Sharing timeline and platforms:

- Update code after each work session on the Github repository.
- Release clean data set and simulation data once created
- Release repository after submission
- Team members, reviewers and public repo visitors will have access to the project

Access restrictions and licensing:
- Give credit to data sources
- Only use licenses where we are permitted to do so
- Open Access: code, simulations, Yahoo Finance data
- No Access: Polygon.io and LobsterData raw datasets(due to vendor restrictions)

Community standards:
- Follow normal clean code practices to keep project well organized and structured
- Note and cite any notable research that led to findings of caused findings

1.6 Data Preservation

Long-term storage plans:
- Keep copies of each version of data on Github
- Keep an archive of code and data with tags
- Data will be kept until the end of the project
- No long term storage plans
- No triggers for deletion and archival planned

Format migration strategy:
- Keep datasets as CSV files, and code in either Python or R
- No permanent storage repository is planned except Github
- No succession plan for data stewardship is planed

Succession planning:
- Assign group members to roles in order to keep repository running well after the submission deadline

1.7 Budget and Resources

Personnel time allocation:
- 4 hours a week per member of the group (2 hours for each class period time)

Technology and infrastructure costs:
- Use free datasets when possible
- More sophisticated data will most likely cost money. For now can have a budget in the range of $50-100 for API credits and this can be splitted across the group if needed

Long-term preservation expenses:
- Only subscription for data API if needed
- Github repository

1.8 Compliance and Ethics

Regulatory requirements:

- Follow licensing rules based on the data API you are drawing from (academic use policy, etc.)

IRB approvals:

- Not needed. Only need base level market data.

Privacy protection measures:

- No direct personal data will be involved
- Will check the dataset beforehand to make sure no private data of companies that shouldn't be exposed is released.