Plan Overview

A Data Management Plan created using DMPonline

Title: Translating Machine Learning into Finance: The Role of ‘Modelling Entrepreneurs’ in the OTC Derivatives Markets

Creator: Taylor Spears

Principal Investigator: Taylor Spears

Data Manager: Taylor Spears

Affiliation: University of Edinburgh

Template: UoE Data Management Plan

Project abstract:

Interest in machine learning has grown significantly within finance in recent years. With its capacity to automate broad areas of decision-making, machine learning techniques have the potential to improve the efficiency of financial intermediation, an activity that has exhibited relatively poor efficiency gains over the last century. Yet, institutional and regulatory factors are likely to play a critical role in determining whether techniques drawn from machine learning can be successfully adopted within the financial services industry, given the extent to which modelling activities in this industry are shaped by regulation. The aim of this project is to investigate how practitioners working within the markets for derivatives that are traded on an ‘over-the-counter’ (OTC) basis are adapting machine learning techniques to fit within the institutional and regulatory environment in which these markets operate, and the impact of these choices on how banks measure the value and risk of derivatives.

ID: 58232

Last modified: 27-10-2020

Translating Machine Learning into Finance: The Role of ‘Modelling Entrepreneurs’ in the OTC Derivatives Markets

Data Capture

What data will be generated or reused in this research? Outline the volume, type, format etc.

This project will primarily involve the collection of qualitative interview data from two groups of individuals. The first are what this project calls 'modelling entrepreneurs': quantitative specialists who are promoting the development and use of new forms of modelling practice drawn from machine learning within the OTC derivatives markets. The second are individuals who work within banks’ model validation and valuation control functions who are involved in overseeing the developments of new models within banks.

Interviews with individuals who consent to having their interviews recorded will lead to the production of interview recordings (in the form of .wav or .mp3 files), which will later be transcribed and coded for qualitative analysis. Where recording is not possible, a member of the Research Team will take detailed notes of the content of the interview, which will later be analyzed as a part of the project. In total, we expect to produce around 30-40 interviews and associated recordings and notes as a part of this project, with each interview lasting anywhere from 30 minutes to an hour on average.

A secondary source of data will be notes taken by a member of the Research Team at conferences and training events related to the use of machine learning within the derivatives markets. The purpose of these training events is to better understand the technical and regulatory issues facing practitioners within this field, and to identify potential interviewees for the project.

Finally, a third source of data will be technical documents and articles on the use of machine learning in the derivatives markets, which will be analyzed to understand the technical and regulatory issues facing practitioners and to identify potential interviewees. Most of these documents will be public (e.g. drawn from relevant industry journals, such as Risk), but there may be occasional documents provided by interviewees as a part of the data collection process.

How much data will be generated?

- 0 - 5 GB

Interview data will be in the form of audio files, which will later be transcribed by a third-party transcriptionist.

Other data include publicly-available technical documents and notes taken by the Research Team.

Data Management

How will the data be documented to ensure it can be understood?

Interview transcripts will be pseudo-anonymized to protect the identity of participants in this research project. Certain non-personally identifiable meta-data about each interviewee will be collected and maintained by the Research Team to determine the representativeness of the sample of interviewees. These data include the date and location of the interview, the interviewee's gender, general information on the interviewee's functional position within their organization (e.g. 'Works within the Risk Management department'), as well as an anonymized code indicating their employer type. These employer type codes will be maintained in a separate table, which will offer a broad description of the employing organization (e.g. 'a G-16 dealer bank', 'a risk advisory firm', 'a securities regulator', etc.).

Where will the data be stored and backed-up?

Interview audio files will be stored in an encrypted volume controlled by the Principal Investigator, which is secured using AES-256 encryption. Following transcription by a third-party transcriptionist, interview transcripts will be verified for accuracy and pseudonymized.

Pseudonymized interview transcripts and associated metadata will be stored on a private drive on the University of Edinburgh Business School filestore. This is high quality, enterprise-class storage with guaranteed backup and resilience. The data are automatically replicated to an off-site disaster facility and backed up with a 60-day retention period, with 10 days of file history visible online.

Integrity

How will you quality assure your data?

Anonymised interview transcripts and associated metadata will be verified for accuracy by the Research Team prior to the deletion of the original recordings.

Confidentiality and IPR

How will you manage any ethical and IPR issues?

Interviewees will be provided with a Participant Information Sheet that clearly explains the aims and objectives of the research project, any risks/benefits they may incur from participating, and information about how interviewees' data will be used/processed and protected. Interviews will only proceed after prospective interviewees have given explicit oral or written consent to participate in the research project. Interview recordings will need to be transmitted to and processed by a third-party transcription firm. The Research Team will ensure that the contracted firm has adequate data protection policies and procedures in place to protect subject confidentiality. After being transcribed, all interviews will be pseudonymized in order to remove any personally identifiable information about research participants.

The Research Team is currently applying for ethical approval from the CAHSS's Research Ethics, Integrity, and Governance Team. That application discusses these issues in more detail.

Preservation & Sharing

Which data do you plan to keep and for how long? Please note that this data should be recorded in Pure.

The need to protect research participants' anonymity limits our ability to make interview data and metadata produced during the project available to the public. To balance these concerns against the need for adequate data retention, interview data and metadata will be deposited in a secure data repository controlled by the University of Edinburgh.

Can you share your data? If not please clarify where it will be stored and preserved.

- No

Interview data and metadata will be stored via the University of Edinburgh's DataVault service following completion of the project. This is an encrypted archive for non-public research data that is managed by the University of Edinburgh. This data will be held for no less than ten years after the initial collection of the data.

Which data will be shared and how?

Pseudonymized interview data and metadata will be shared within the Research Team via the University's DataSync service. Where members of the Research Team are employed by universities other than the University of Edinburgh, the data will be shared via the University's Microsoft OneDrive service. Portions of pseudonymized interview data will occasionally need to be shared with academic publishers and reviewers as a part of the peer-review process. Transmission of these data will be made in a format determined by the publisher.

Are any restrictions on data sharing required?

Even though the final pseudonymized interview data and metadata will be stripped of specific personally identifiable information, we believe that there remains the possibility of deductive disclosure of research subjects' identities through analysis of the transcripts themselves.

Thus, we will only be able to make the pseudonymized interview data and metadata available to researchers who have signed a data-

sharing agreement that provides for: (1) a commitment to using the data only for research purposes and not to identify any individual participant; (2) a commitment to securing the data using appropriate computer technology; and (3) a commitment to destroying or returning the data after analyses are completed.