Integrity and ethics for researchers using big data and machine learning techniques in derivative pricing

Charlie Browne, Head of Risk, Quant & Market Data Solutions

This essay forms part of the preparation of my PhD thesis, **"The trading book in banking: a study of risk factors in derivative pricing."** I hope the content will provide food for thought and a useful reference for data scientists and chief data officers who seek to ensure ethics and integrity in data mining, model construction, analytics, pricing and the recommendations in research reports to clients.

The literature on derivatives pricing is extensive, too broad and deep to make a systematic review of all of the accumulated literature possible. Google Scholar returns 567,000 results for "derivatives pricing", 205,000 results for "risk factors derivative pricing", 3,800,000 for "market risk" and 265,000 for "market risk derivative pricing". The vast majority of this writing was done in the almost fifty years since the seminal Black-Scholes-Merton model (Black and Scholes, 1973; Merton, 1973) – a model that is central to this study - was published in 1973. While not all of these google search results will be related to the pricing of financial derivatives, the quantity of research available on this single topic nonetheless is indicative of the vast volumes of the data that has been created since the advent of the internet.

Research integrity and ethics in data science

The purpose of this essay is to critically review the issues associated with research integrity and ethics as it applies to data science techniques for derivative pricing models. Han, Pei and Kamber (2011) refer to data mining as knowledge discovery and pattern recognition from data. The data scientists that work in the field of data mining have a responsibility to ensure the data is not misused. The Data Science Code of Professional Conduct (2019) states that if data scientists believe that the data is being misused by a client, for example to communicate a false reality or promote an understanding that is incorrect, they have a responsibility to disclose the misuse, including, if necessary, disclosure to the proper authorities. The code describes the obligation on the data scientist to take reasonable measures to inform and persuade the client to take an alternative approach.

'Data mining techniques should adhere to responsible research practices'

The responsibilities of the data scientist, as outlined in the Data Science Code of Professional Conduct (2019), are consistent with the principles of research integrity. Israel (2014) describes research integrity as the principles that need to be abided by to ensure research is valid, trustworthy and beneficial to society. Moher et al (2020) discuss the Hong Kong Principles. These principles were developed during a conference

on research integrity in Hong Kong in 2019. The principles covered responsible research practices, the value of complete reporting, the benefits of open science, the need to cover a wide range of research activities and the importance of peer review. All of these principles apply to the practice of data mining. Data mining techniques should adhere to responsible research practices. The reporting and analysis of results needs to be complete and interpreted as accurately as possible and the principles of open science should be aspired to when it comes to the knowledge discovered during the data mining process. Steneck (2007) argues that research integrity is closely aligned with good citizenship. Data scientists should treat the recipients of the knowledge and interpretation extracted from the data in the same manner as a responsible citizen would. Judgement should be applied which accounts for the sensitivity of the information or the researcher's interpretation of the information and the potential reaction of recipients to it.

'Due to the complexity of many of the algorithms in data science, the use of data science techniques offers the opportunity to misrepresent contributions in research.'

Han et al (2011) argue that we humans are living in a data age. Millions of giga-bytes of data are transferred around the internet every day and we are immersed in it. The data is stored in various storage devices and machines. The data is created by business, society, science, engineering, medicine and many other sections of the global community. Data mining turns this data into knowledge. Researchers today use data science to create algorithms that translate the data into information used by the public at large. Israel (2014) posits that research integrity and ethics is about understanding the impact of a set of research activities and applying judgement as to which of them are defensible. Researchers that use data science techniques have a similar responsibility. The algorithms they create that mine and interpret the data need to be written with the researcher's integrity in mind. Researchers have an obligation to behave with integrity and honesty in their creations. They need to be conscious that the research they are producing needs to be transparent and auditable. Due to the complexity of many of the algorithms in data science, the use of data science techniques offers the opportunity to misrepresent contributions in research. Clearly this must be avoided and researchers with access to advanced data science techniques should use their privileges judiciously and with integrity.

Despite the increased acceptance of qualitative research designs in the field of quantitative finance and in particular in research on derivative pricing models, most studies still use the quantitative method. To a certain extent this is to be expected. A study of derivative models and their risk factors is an unlikely place for a qualitative approach whose objective is to capture and analyse the thoughts, words and other unquantifiable aspects of the information in the minds of research participants that the qualitative research method attempts to reveal. Israel (2014) discusses fairness in the peer review processes. A question presents itself here: will researchers who use sophisticated data science techniques be favoured over those who don't in the peer review process? Data science techniques allow quantitative researchers to include data volumes that are orders of magnitude higher than researchers could use even as recently as twenty years ago. Quantitative research is also the traditional form of research. Creswell (2009) discusses the difficulties that researchers have had in the past persuading faculty and other research audience as to the merits of research that uses qualitative methods. He discusses how this has changed in more recent years leading to an increase in the perceived legitimacy of both qualitative research and mixed methods approaches, i.e., methods that employ both qualitative and quantitative designs. A potential area for future research relates to the emergence of data science and its impact on the bias of academics towards quantitative methods.

One exception to the prevalence of the use of the quantitative method in research on financial derivatives is the study of Bezzina and Grima (2011). They performed a study of attitudes towards risk management controls of derivative trading in banking. A qualitative approach was used which involved 420 interviews using an on-line survey. The interview results were aggregated using an exploratory factor analysis approach across four demographic variables: gender, experience, education and position held. The resulting factor analysis offered insight into five hypothesised dimensions of trading book controls: risk management controls, misuse, expertise, perception and benefits. Israel (2014) discusses collegiality in scientific interactions and the need to take account of the potential sensitivities or viewpoints of the participants of the study. In Bezzina and Grima (2011), because the study necessitated the grouping of research participants into cohorts based on gender, experience, education and position held, the authors clearly would have needed to have exercised care in ensuring that the participants felt they were not being stereotyped in any way.

'Bezzina and Grima (2011) performed a study of attitudes towards risk management controls of derivative trading in banking'

Oancea (2014) discusses the meaning of ethics. He argues that ethics is the analysis of actions. Actions can be categorized as those that are good or virtuous courses of action and those that are not. Ethical behaviour is about protecting those that could potentially be impacted by those actions. This includes individual people, companies, social groupings of various kinds, as well as the environment. Ethical actions can also be considered those actions that have the potential to increase the amount of good for those impacted by the actions. Oancea (2014) posits that researchers should as a matter of course act in an ethical way by ensuring that they are always trying to have a positive impact on those are impacted by that research. Any adverse impact of the research should be minimised or avoided.

Weinbaum et al (2018) discuss the role that ethics plays in scientific research. They conclude that the role that ethics plays in research varies by discipline and by country. The material they reviewed included academic literature across a variety of scientific fields. The research included interviews with knowledgeable participants in the US, China and several countries in Europe. They argue that, while there is commonality across both geographical region and the different scientific disciplines on perceptions of what constitutes ethical behaviour in research, there is also much variation. The authors review several emerging topics that are seen to be at the forefront of discussions on ethics in the future. The important themes were the ethics of research in the sciences, the research methods, and how the research was conducted. How the research is applied is deemed less relevant.

Robson and McCartan (2016) describe ethics as the principles that guide people's behaviour. These principles can be social norms or can be codified in the forms of written down rules. Israel (2014) argues that there are four main approaches to ethics in research: agent-centred, principleism, rule-centred, and critical. For agent-centred research it is the character of the researcher and the moral quality of the research that should inform decisions of research ethics. The consequences of the researchers' actions are less relevant. Sinnott-Armstrong (2003) states that the agent-centred researcher can cite Kant's (Kant, 1785) location of the moral quality of acts within the principles of the acts. The impact of those acts on others, again, is less relevant in this philosophy. For Kant, the only thing that was unquestionably good was good will itself. Characteristics such as courage, humility, bravery and integrity are central in this philosophy.

Israel (2014) contrasts agent-centred ethical philosophy to the philosophy of consequentialism. Consequentialism is the view that the higher the quantity of good outcomes that come from an action the more ethical that action can be deemed. Sheffler (1988) states that the most familiar version of consequentialism is utilitarianism. Utilitarianism is the view that the morally right course of action to take is the one that produces the highest quantity of good outcomes. The classical utilitarians of the 19th century, Jeremy Bentham and John Stuart Mill (Mill, 1859), argued that the overall objective in the actions taken by society should be the maximization of the quantity of good outcomes. That is, to bring about the greatest amount of good for the greatest number. Utilitarianism is associated with the concepts of impartiality and agent-neutrality. Agent-neutrality is the concept that a good outcome for one individual cannot be deemed to be more important than a good outcome for another individual.

'The Data Science Code of Professional Conduct (2019) describes a data scientist as a professional who uses scientific methods to liberate and create meaning from raw data'

In the Data Science Code of Professional Conduct (2019) a data scientist is described as a professional who uses scientific methods to liberate and create meaning from raw data. The agent-centred and utilitarian ethical philosophies of Kant and Mill should be adhered to by the researcher who uses data science techniques to analyse data. For example, Han et al (2011) describe clustering as a technique in data mining that groups a set of data objects into multiple groups or clusters so that objects within a cluster have similar features. The similarities of the data are assessed based on the attribute values. They often involve distance measures. Different clustering methods exist: partitioning methods, hierarchical methods, density-based methods and grid-based methods. Clustering is one of the techniques that allows researchers to draw meaning from the data. As they do so they should bear in mind the ethical principles and guidelines associated with the acts that they are performing. Using clustering techniques to draw meaning from the data should be done only where the researcher's objectives are aligned with the principles of integrity and honesty. Whether it is clustering, data reduction, gradient descent or any other data science technique that is being used to support the extraction of meaning from high volumes of often unstructured data, the researcher's intentions should remain consistent. The results of the analysis and the objective of the data science should always be the greater good of the public at large. Data science techniques are becoming increasingly important in research today. With the huge quantities of data available in databases across the internet and the increasingly sophisticated techniques that allow data scientists to draw inferences from that data, there is obvious scope for unethical behaviour.

The fabrication and falsification of data is discussed by Sterba (2006) and Israel (2014). Data can be falsified in many different ways. Streiner (2002) discusses the technique of dichotomizing continuous data. He offers several reasons why this approach is flawed and should not be used. The use of data for dual purposes is also a potential method for data fabrication or falsification. For example, data that is used for exploratory purposes should not be re-used as part of the same research for confirmation purposes. Care must be taken where several models fit the data but offer different conclusions. These models need to be tested and inconsistencies explained. Conclusions should not be drawn and used in the research if models do not consistently fit the data.

'Alexander (2001) describes a principal component analysis (PCA) approach as a specific data reduction technique. She argues that financial markets are characterised by a high degree of collinearity'

An example of a data science technique that could be used for falsification and fabrication in the area of finance is the use of data reduction in machine learning. In Alexander (2001) the author describes a principal component analysis (PCA) approach as a specific data reduction technique. She argues that financial markets are characterised by a high degree of collinearity. The collinearity occurs because while there may be an innumerable amount of data points available, there are often only a few key sources of information in the data. The paper uses a standard approach for extracting the key data points. These data points are the uncorrelated sources of variation in a multivariate PCA system. The author states that PCA is often associated with the analysis of interest rate curves in financial markets. The assumed interpretation of this interest rate analysis is that the first principal component represents the general level of interest rates, the second principal component represents the slope of the interest rate curve and the third principal component is an indicator of the amount of curvature in the interest rate curve. Alexander (2001), however, does not focus on interest rates. She instead presents a principal component model of traded volatility

'It is easy to see how researchers could take advantage of the complexity embedded in these models, i.e., take advantage of chance patterns observed in the data to mislead or arrive at deliberately false conclusions'

smiles incorporating fixed strike volatility deviations from ATM volatility. While machine learning tools such as PCA offer powerful solutions for data reduction and machine learning, from an ethical perspective it is easy to see how researchers could take advantage of the complexity embedded in these models, i.e., take advantage of chance patterns observed in the data to mislead or arrive at deliberately false conclusions. It is obvious that the researcher's integrity is paramount in data science and as techniques evolve and become more prevalent, ethical considerations will become even more important.

Strauss and Corbin (1994) describe grounded theory in research as a methodology that is grounded in the data. In contrast to traditional research, it can sometimes even contradict scientific method. The UK data ethics framework (UK Data Ethics Framework, 2018) states that data users must ensure data used in a project is accurate, and represents as best as possible the true picture underlying the data. The data should be used in proportion to the relevance of the information being conveyed and it must be of sufficient quality to prevent question marks about its accuracy. If there are any limitations in the ability of the data to inform users, these must be clearly explained. These ethical considerations should be considered where grounded theory has been used as the research philosophy. Grounded theory uses an inductive methodology that categorises information and builds theory and a conceptual framework from anchors identified within the data. Where data science techniques are used for such approaches, care must be taken by researchers to ensure that data that is used in such research is not fabricated or falsified in any way. The systematic analysis of that data supports bottom-up construction type approach used in grounded theory and as always there is opportunity for falsification.

The UK data ethics framework (UK Data Ethics Framework, 2018) states that users of data need to understand their legal position with respect to use of data, The requirement to be aware of the laws and legislation associated with the use of a particular data source is clearly applicable also to researchers who are using data for their studies. A well-known example of regulation that governs the use of data is the European Union General Data Protection Legislation, known as GDPR. GDPR legislates that the subjects of data have the right to request access to the personal information that an organisation holds about them. They also have the right to demand that an organisation destroys their personal information. GDPR applies to most types of organisations including banks and financial services.

It is often the case that data for research is taken from open sources platforms. Lerner and Tirole (2002) discuss the legal foundations of open-source licensing. Prior to the 1980s, software tended to be produced and distributed without a license specifying how it can be used. During the 1980s, however, software developers started to become concerned by behaviours of users of the software that they deemed to be unethical. Richard Stallman was a Massachusetts Institute of Technology (MIT) software developer who at the time was concerned about unethical use of open-source software. In particular, he was concerned that software that was initially developed for public use was subsequently being modified by opportunistic developers for commercial gain. He devised the idea of a software license that required that the source code behind the software programmes remain free for public use. The resulting license was called the GNU Public license. Lerner and Tirole (2002) argue that despite the uncertain legal framework that open-source licenses sit within, the developers of open-source software continue to remain concerned about the choice of open-source license that their software is published under. Today there are several different types of open-source license – each of which offers licensees specific terms of usage and restrictions. According to opensource.org the most common open-source licenses are Apache Licenses, Berkley Software Distribution (BSD) licenses, GNU general public licenses (GPL), GNU Lesser General Public License (LGPL), MIT (Massachusetts Institute of technology) licenses, Mozilla Public Licenses (MPL), Common Development and Distribution Licenses (CDDL), and the Eclipse Public Licenses (EPL).

To aid with the conclusion of this paper it is useful to refer to the Responsible Conduct of Research (RCR). It contains a codified set of principles that offer an overview of the ethical standards that researchers should aspire to where data science techniques are used in their research. The objective of the RCR is the promotion of responsible scientific inquiry. It seeks to facilitate collaborative research environments that

'The Responsible Conduct of Research (RCR) contains a codified set of principles that offer an overview of the ethical standards'

promote research for the good of the public. The underlying assumption is that the public will trust in the research if there is a general belief that they will benefit from it. Data scientists will likewise benefit from adhering to the principles included in the RCR. These include ensuring that the reputation of both researchers and their institution are protected, the avoidance of behaviours that would discredit the research, and the identification of mechanisms to allow observers to respond to questionable practices.

References

Alexander, C., 2001. Principal component analysis of volatility smiles and skews. Available at SSRN 248128.

Amilon, H., 2003. A neural network versus Black–Scholes: a comparison of pricing and hedging performances. Journal of Forecasting, 22(4), pp.317-335.

Black, F. and Scholes, M., 1973. The Pricing of Options and Corporate Liabilities. The Journal of Political Economy, 81(3), pp.637-654.

Bezzina, F.H. and Grima, S., 2012. Exploring factors affecting the proper use of derivatives: An empirical study with active users and controllers

Brogaard, J. and Zareei, A., 2021. Machine learning and the stock market. In Proceedings of Paris December 2020 Finance Meeting EUROFIDAI-ESSEC.

Culkin, R. and Das, S.R., 2017. Machine learning in finance: the case of deep learning for option pricing. Journal of Investment Management, 15(4), pp.92-100.

Feng, G., Giglio, S. and Xiu, D., 2020. Taming the factor zoo: A test of new factors. The Journal of Finance, 75(3), pp.1327-1370.

Gigerenzer, G., 2018. The heuristics revolution: Rethinking the role of uncertainty in finance. In The Behavioural Finance Revolution. Edward Elgar Publishing.

Han, J., Pei, J. and Kamber, M., 2011. Data mining: concepts and techniques. Elsevier.

Ince, H. and Trafalis, T.B., 2004, July. Kernel principal component analysis and support vector machines for stock price prediction. In 2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No. 04CH37541) (Vol. 3, pp. 2053-2058). IEEE.

Israel, M., 2014. Research ethics and integrity for social scientists: Beyond regulatory compliance. Sage.

Kant, I., 1785. Lectures on Ethics.(L. Infield, Trans.) New York. NY: Harper & Row.

Lerner, I., Goldblum, A., Rayan, A., Vardi, A. and Michaeli, A., 2017. From finance to molecular modeling algorithms: The risk and return heuristic. Curr. Top. Pept. Protein Res, 18, pp.117-131.

MacIntyre, A., 2003. A Short History of Ethics: a history of moral philosophy from the Homeric age to the 20th century. Routledge.

Manners, I., 2008. The normative ethics of the European Union. International affairs, pp.45-60.

Mill, J.S., 1859. Utilitarianism (1863). Utilitarianism, Liberty, Representative Government, pp.7-9.

Moher, D., Bouter, L., Kleinert, S., Glasziou, P., Sham, M.H., Barbour, V., Coriat, A.M., Foeger, N. and Dirnagl, U., 2020. The Hong Kong Principles for assessing researchers: Fostering research integrity. PLoS Biology, 18(7), p.e3000737.

Scheffler, S. ed., 1988. Consequentialism and its Critics. Oxford University Press on Demand.

Sinnott-Armstrong, W., 2003. Consequentialism.

Streiner, D.L., 2002. Breaking up is hard to do: the heartbreak of dichotomizing continuous data. The Canadian Journal of Psychiatry, 47(3), pp.262-266.

Weinbaum, C., Landree, E., Blumenthal, M.S., Piquado, T. and Gutierrez, C.I., 2018. Ethics in Scientific Research: An Examination of Ethical Principles and Emerging Topics.

Ye, T. and Zhang, L., 2019. Derivatives pricing via machine learning. Boston University Questrom School of Business Research Paper, (3352688).