CAPCO

THE CAPCO INSTITUTE JOURNAL OF FINANCIAL TRANSFORMATION

DATA ANALYTICS

Machine learning for advanced data analytics: Challenges, use-cases and best practices to maximize business value

NADIR BASMA | MAXIMILLIAN PHIPPS
PAUL HENRY | HELEN WEBB

DATA ANALYTICS

50TH EDITION | NOVEMBER 2019

THE CAPCO INSTITUTE

JOURNAL OF FINANCIAL TRANSFORMATION

RECIPIENT OF THE APEX AWARD FOR PUBLICATION EXCELLENCE

CONTENTS

DATA MANAGEMENT

DATA ANALYTICS

DATA INTELLIGENCE

DEAR READER,

Welcome to the milestone 50th edition of the Capco Institute Journal of Financial Transformation.

Launched in 2001, the Journal has covered topics which have charted the evolution of the financial services sector and recorded the fundamental transformation of the industry. Its pages have been filled with invaluable insights covering everything from risk, wealth, and pricing, to digitization, design thinking, automation, and much more.

The Journal has also been privileged to include contributions from some of the world's foremost thinkers from academia and the industry, including 20 Nobel Laureates, and over 200 senior financial executives and regulators, and has been co-published with some of the most prestigious business schools from around the world.

I am proud to celebrate reaching 50 editions of the Journal, and today, the underlying principle of the Journal remains unchanged: to deliver thinking to advance the field of applied finance, looking forward to how we can meet the important challenges of the future.

Data is playing a crucial role in informing decision-making to drive financial institutions forward, and organizations are unlocking hidden value through harvesting, analyzing and managing their data. The papers in this edition demonstrate a growing emphasis on this field, examining such topics as machine learning and AI, regulatory compliance, program implementation, and strategy.

As ever, you can expect the highest caliber of research and practical guidance from our distinguished contributors, and I trust that this will prove useful to your own thinking and decision making. I look forward to sharing future editions of the Journal with you.

Lance Levy, **Capco CEO**

FOREWORD

Since the launch of the Journal of Financial Transformation nearly 20 years ago, we have witnessed a global financial crisis, the re-emergence of regulation as a dominant engine of change, a monumental increase in computer processing power, the emergence of the cloud and other disruptive technologies, and a significant shift in consumer habits and expectations.

Throughout, there has been one constant: the immense volume of data that financial services institutions accumulate through their interactions with their clients and risk management activities. Today, the scale, processing power and opportunities to gather, analyze and deploy that data has grown beyond all recognition.

That is why we are dedicating the 50th issue of the Journal of Financial Transformation to the topic of data, which has the power to change the financial industry just as profoundly over the coming 20 years and 50 issues. The articles gathered in this issue cover a broad spectrum of data-related topics, ranging from the opportunities presented by data analytics to enhance business performance to the challenges inherent in wrestling with legacy information architectures. In many cases, achieving the former is held back by shortcomings around the quality of, and access to, data arising from the latter.

It is these twin pillars of opportunity and challenge that inform the current inflection point at which the financial industry now stands. Whilst there is opportunity to improve user experiences through better customer segmentation or artificial intelligence, for example, there are also fundamental challenges around how organizations achieve this – and if they can, whether they should.

The expanding field of data ethics will consume a great deal of senior executive time as organizations find their feet as they slowly progress forward into this new territory. In my view, it is critical that organizations use this time wisely, and do not just focus on short-term opportunities but rather ground themselves in the practical challenges they face. Financial institutions must invest in the core building blocks of data architecture and management, so that as they innovate, they are not held back, but set up for long-term success.

I hope that you enjoy reading this edition of the Journal and that it helps you in your endeavours to tackle the challenges of today's data environment.

Guest Editor
Chris Probert, **Partner, Capco**

MACHINE LEARNING FOR ADVANCED DATA ANALYTICS: CHALLENGES, USE-CASES AND BEST PRACTICES TO MAXIMIZE BUSINESS VALUE

NADIR BASMA | Associate Consultant, Capco
MAXIMILLIAN PHIPPS | Associate Consultant, Capco
PAUL HENRY | Associate Consultant, Capco
HELEN WEBB | Associate Consultant, Capco

ABSTRACT

As the amount of data produced and stored by organizations increases, the need for advanced analytics in order to turn this data into meaningful business insights becomes crucial. One such technique is machine learning, a wide set of tools that builds mathematical models with minimal human decision making. Although machine learning has the potential to be immensely powerful, it requires well-considered planning and the engagement of key business stakeholders. The type of machine learning used will be determined by the business question the organization is trying to answer, as well as the type and quality of data available. Throughout the development process, ethical considerations and explainability need to be considered by all team members. In this paper, we present some of the challenges of implementing a machine learning project and the best practices to mitigate these challenges.

1. INTRODUCTION

Across almost all industries, an unprecedented amount of data is currently being generated, with an estimated 2.5 quintillion (10^{18}) bytes of data created across the globe each day.[1] The financial services industry is no exception, with the New York Stock Exchange capturing 1 TB of trade data each trading session, for instance. Not only is the amount of data being produced increasing, but so too is the variety of formats in which it is being produced and the structural complexity of this data. As well as highlighting the importance of controls required to ensure that data quality standards are met,[2] the amount and complexity of data an organization handles on a daily basis brings into focus the need for advanced analytics to generate actionable insights.

Machine learning (ML) is a field with strong relevance to advanced analytics. At the most basic level, machine learning describes the use of computational algorithms more advanced than traditional analytics methods (for example, SQL queries and data mining approaches) that are employed to gain insight into large datasets. While the term itself is relatively new, its core concept of learning from data without relying on rules-based programming is not. Machine learning techniques have foundations firmly in the science of statistics, with these concepts built on and refined into the rich and diverse set of tools at our disposal today.

Machine learning has valuable applications in a diverse range of fields, including financial services. These applications include detecting financial crime, predicting loan repayment defaults, and providing personalized customer engagement.

[1] Quintero, D., L. Bolinches, A. G. Sutandyo, N. Joly, R. T. Katahira, 2016, "IBM data engine for hadoop and spark" IBM Redbooks, https://ibm.co/2kgrCoT
[2] Please refer to "Data quality imperatives for data migration initiatives: A guide for data practitioners" in this edition of the Journal.

One of the key enablers of growth for machine learning has been the availability of cloud-based infrastructure.

Figure 1 provides a brief summary of our recommended approach, and further explanation of concepts is provided throughout the article.

2. ASSESSING WHETHER MACHINE LEARNING IS SUITABLE FOR THE BUSINESS CASE

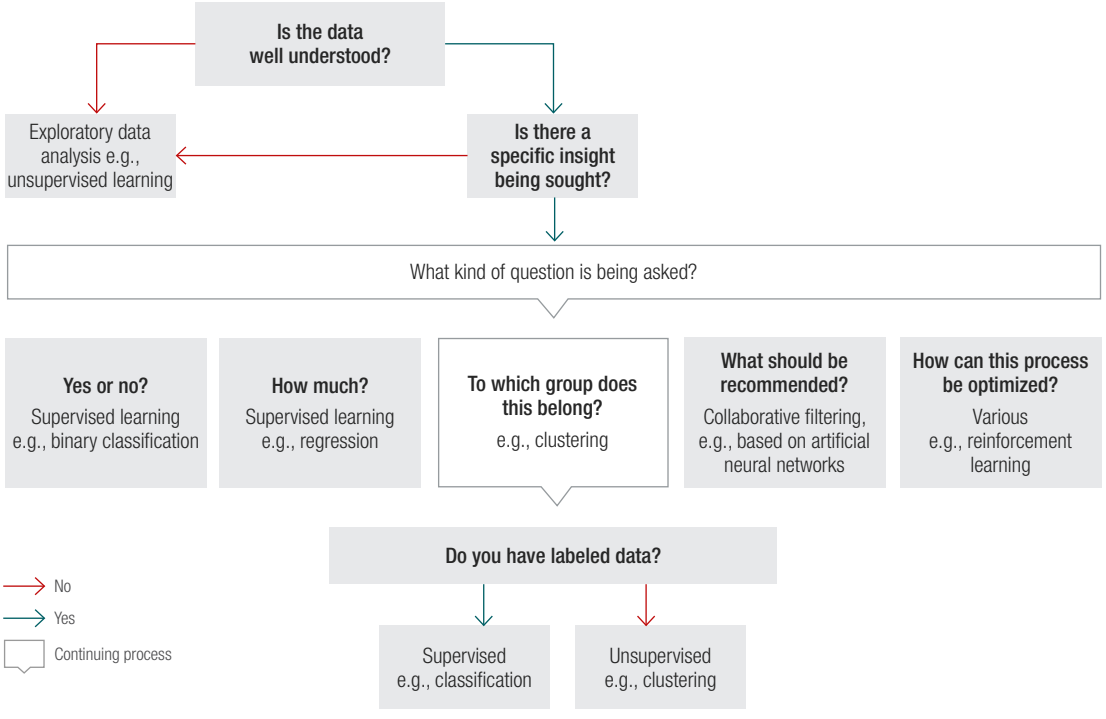
While machine learning is a powerful data analytics tool, machine learning projects can yield disappointing results without a well-considered business case. For a truly successful analytics project delivering insights with value, both data scientists and stakeholders need to speak the same language. This requires both sides to regularly communicate their understanding of the work and help each other understand their expectations of the project. It is important that business managers know that it is their responsibility to ask important business questions of their data scientists and that the data scientists know that they need to be able to answer these questions in ways that are understandable to the business. These questions are not related to the inner workings of every algorithm of interest, such as whether they will be using cosine or cartesian distance to determine which observation should belong to which cluster. The questions asked should refer to the nature of the results, and which tools are most appropriate. Later, we will discuss what some of these key questions are.

While it is tempting to think that machine learning can solve any business question, for many projects more traditional and less technically challenging analytical methods, such as statistical modeling and descriptive statistics, can provide a comparable level of insight. Some problems are not analytics problems at all, but rather automation problems that require process engineering and robotic process automation.[3] By discussing the most appropriate techniques for the desired outcome, both data scientists and stakeholders can develop a clearer understanding of the expected timescales and outcomes for the work.

2.1 Does this problem require exploratory analysis or actionable insights to be derived?

Machine learning can be used to directly gain actionable insights, or to explore large datasets (exploratory data analysis). Exploratory analysis is often used as an initial step in order to search for trends and patterns that could lead to

Figure 1: Our data commercialization framework



[3] For further discussion of automation trends in the financial services industry, see the Journal of Financial Transformation 46, https://bit.ly/2IKS0HM

actionable insights. This approach is, therefore, the solution to the problem that discovering actionable insights requires prior knowledge of what to search for. The discussion of which approach will be taken is one of the most important conversations that any stakeholder can have with their data science team, and as a result, can be the biggest source of misaligned expectations between the two groups if it is not had at, or before, the onset of a data science project.

2.2 What is the nature of the insight being sought?

Two fundamentally different learning approaches, supervised learning (SL) and unsupervised learning (UL), suit different datasets. Here, we discuss them in relation to the insight being sought. SL algorithms require labels to be included with the dataset; these labels are simply the true values that the algorithm learns to reproduce. An example of supervised learning is a basic fraud detection model that takes in data relating to the transaction and returns a prediction of whether this transaction is fraudulent or not. This input data, termed the 'features', might be the date, time, location, and amount relating to the transaction. Developing this SL fraud detection model begins with building a dataset to train the model; this involves taking historical examples of both fraudulent and genuine transactions and storing their features in a dataset. An additional column is included with this dataset that describes whether each particular example relates to a fraudulent transaction or not; these are termed our 'labels' and this is the key ingredient that differentiates SL from UL. The model is then taught to detect future fraudulent transactions; this is performed by repeatedly feeding the model with rows of data relating to a particular transaction, from which the model seeks to reproduce the transaction's label values, that is 'fraudulent' or 'genuine', given the input data consisting of such data as location and date. With every new transaction the machine learning algorithm receives, the model's ability to reproduce this label's value improves. This training is performed in such a way that the model is general and avoids overfitting to the training data, thereby avoiding making spurious predictions of future events.

UL, on the other hand, does not require labeled outputs or inputs. At the simplest level, UL seeks to characterize data by considering their relationship to one another, a common example being clustering data by partitioning and associating datapoints into groups with similar properties. As a simple example, let us consider a dataset containing information about a bowl of fruit including features such as color, weight, size/dimensions, and time to ripen. We may apply a clustering algorithm to this dataset that contains no fruit name information (i.e., missing labels such as 'apple' and 'banana'). A successful clustering run will group the fruits by their features. It is likely that the algorithm will have implicitly learned to group the data such that each cluster contains a high proportion of a particular fruit. Since this is an unsupervised learning approach, however, the algorithm achieves this without knowledge of the labels (i.e., names) of each particular fruit; these labels are therefore implied by the final clusters.

There is an element of human intervention and decision making required when adopting an SL method; this is due to the selection and categorization of data for training typically being performed manually, and the requirement to tag unlabeled data with labels. In UL, however, there is less human intervention as the algorithm learns from data that is not labeled; the preparation of training datasets forms the most time-consuming step. This is also the stage at which errors and biases are most likely to be introduced.

2.3 What is the necessary level of certainty?

At first sight, it may be expected that all analyses should be performed to the highest level of accuracy. This, however, is not the case, with the difficulty lying in the balance of cost and outcome value. Specific criteria for meeting an outcome should be directly related to the original business problem being solved. Stakeholders and data scientists should fully commit to regularly assessing results and findings in order to determine whether these meet the success criteria, and decide whether conducting further work is necessary. Alternatively, the business may be at a stage where it can confidently take a decision based on the basis of the current insights gained.

3. CHOOSING A MACHINE LEARNING APPROACH ACCORDING TO THE BUSINESS QUESTION BEING ASKED

In many cases, it is constructive to rephrase the question asked to enable actionable insights to be obtained more efficiently without impacting the overall business objective. For instance, if the objective is to detect risky investments, asking the question "how risky is this investment?" is a complex question that requires the risk to be measured and quantified. Reframing this question as "is this a risky investment?" allows the question to be answered more easily (since this requires a yes/no-type answer), while likely meeting the same business objective. Questions can be considered equivalent if the business outcome remains unchanged and the number of assumptions made does not increase. Changing the question asked should always be driven by the computational complexity

Table 1: Summary of the types of business questions and machine learning approaches that can be used to answer those questions

TYPE OF QUESTION	EXAMPLE	EXAMPLE MACHINE LEARNING APPROACH	DATA REQUIRED
YES OR NO?	Should this customer be granted a loan of £5000?	Supervised classification	Labeled data
HOW MUCH…?	What will the stock price of Apple be in 2025?	Supervised regression	Large amounts of labeled data
TO WHICH GROUP DOES THIS BELONG?	Is this transaction normal or potentially fraudulent?	Unsupervised clustering	Large amounts of unlabeled data
WHAT SHOULD BE RECOMMENDED?	What type of investment portfolio should I recommend to this customer?	Unsupervised collaborative filtering	Large amounts of user preference data
HOW CAN THIS BE OPTIMIZED?	How can the investment in marketing be optimized for maximum ROI?	Reinforcement learning	Small amounts of unlabeled data but a strong data science team and large amounts of computational power are required

in answering different types of questions, rather than asking questions based on the types of data that are easily available. Each of the types of questions mentioned can be answered by different types of SL algorithms and may require differing amounts of training data. Machine learning is very well suited to answering five general types of questions (Table 1).

3.1 Yes or no? – classification

Decision tree-based algorithms are used as the basis for many classification and regression problems. The generalized classification and regression tree (CART) is the catch-all term used to describe the general use of decision trees for such problems. These methods typically have the benefit of allowing for the importance of a variable to be measured and the decision tree to be visually charted, making decisions explainable. This explainability of machine learning models is of key importance to the data science team for solving business challenges, as it allows the machine learning model to be easily understood from a non-technical perspective. Machine learning models that can be intuitively explained are key to ensuring that data science teams and stakeholders can work in parallel, rather than diverging. The implications of explainability in machine learning are discussed further below.

3.1.1 EXAMPLE: CLASSIFICATION IN FINANCIAL SERVICES: MONEY LAUNDERING DETECTION

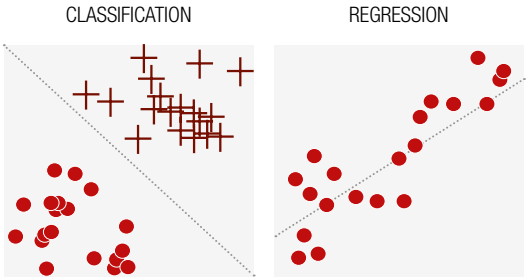
Money laundering is a global problem that provides the means for criminals to conceal their illegal financial profits. While reducing and eliminating money laundering is a key focus for many governments, it remains a difficult task that is often approached by assessing the prior transactional histories of individuals. Typically, however, money laundering is not an individual matter but rather involves groups of people working together as a collective. Savage et al. (2017)[4] recently proposed an SL approach to this problem. This approach uses a combination of network analysis (i.e., analysis of the groups of people and their accounts that are potentially involved in money laundering activities) alongside a classification algorithm. The evaluation of this combined approach indicated that the method is able to correctly detect suspicious activity with a low rate of false positives. This method is, therefore, shown to have high potential for deployment in a real-world environment.

3.2 How much? – regression

Regression, on the other hand, predicts a numerical value. The simplest type of regression is linear regression, which seeks to relate two variables by using the value of one variable to predict the value of the other variable. An example of a simple linear regression might be predicting the salary of an employee based on his/her age. In this type of regression, training data would include historical data containing values of both variables. Although regression can be performed with relatively little data, the quality of the output will generally increase with the amount of training data available. It is best practice to use regression when large training datasets are available. A problem with multiple known variables is called a multivariate regression problem, and a regression problem where input variables are ordered by time and a future prediction is sought is a time series forecasting problem. It is possible when analysis is begun that it is not clear which variables contribute to changes in the value being predicted.

[4] Savage, D., Q. Wang, P. Chou, X. J. Zhang, and X. Yu, 2016, "Detection of money laundering groups: supervised learning on small networks," The AAAI-17 Workshop on AI and Operations Research for Social Good, WS-17-01, https://bit.ly/2kIDDUf

Figure 2: Classification versus regression



Source: Korbut (2017)[7]

The decision of variables to use in linear regression can be steered through critically assessing the relevance of each variable one-by-one, referred to as manual feature engineering. This makes use of the fact that variables in a dataset are often redundant and able to be described to a large degree by other variables in the dataset, or perhaps may not possess any relation to the values sought to be predicted. Performing feature engineering is an important step towards explainable machine learning by providing a sound reasoning for features being included in the dataset.

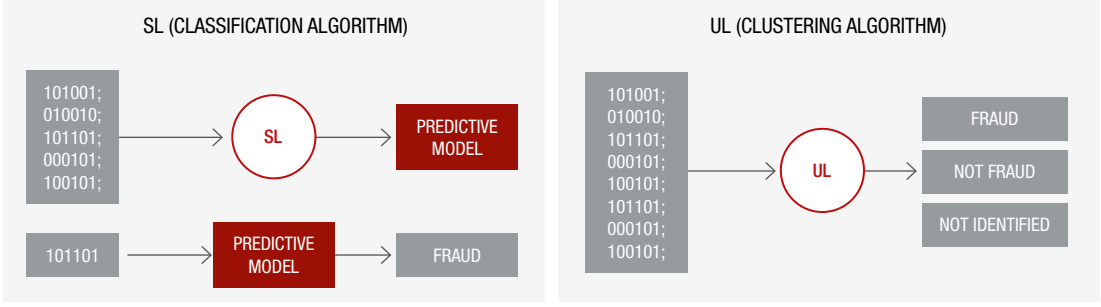
3.2.1 EXAMPLE: REGRESSION IN STRESS-TESTING

Regression has been used extensively for applied stress-testing in financial institutions. Regression is often well suited to analyses where the ability to extrapolate to unknown situations is required. In terms of stress-testing, this means predicting how financial institutions will cope under extreme market conditions. Specifically, regression analysis allows historical data during non-extreme market conditions to be extrapolated to extreme conditions. An example of the use of regression for stress-testing is by the Federal Reserve Bank of New York. In order to estimate possible future capital shortfalls, linear regression models are used. These calculations form part of the bank's balance sheet projections that feed into a wider banking stress-test.[5] Models that can be used to produce such forecasts are not limited to simple linear regression models, with more sophisticated approaches available that better capture complex trends.[6] While often used by quantitative analysts, these more advanced models are generally less preferred for the purposes of stress-testing. In the case of stress-testing specifically, the ability to communicate the results, methods, and assumptions of a model to senior business stakeholders is of paramount importance. With the increased difficulty associated with building persuasive arguments using these more sophisticated models, linear models are typically preferred.

3.3 To which group does this belong? – clustering

Clustering algorithms seek to group data points into distinct groups based on their features, with a successful clustering run providing support for hypotheses such as the data being separable into high or low risk groups. Clustering is useful in exploratory analysis because it can automatically identify structure in data. In situations where it is either impossible or impractical for a human to identify trends in the data, UL can provide initial insights that can then be used to test individual hypotheses. For instance, clustering methods can

Figure 3: The differences between the SL and UL algorithms with relation to the fraud detection example



Here, 0s and 1s are used to represent features that are input into the SL, UL and predictive models, with classes (fraudulent, non-fraudulent and not identifiable) and unlabeled clusters output by the model.

Source: modified version of Zhou (2018)[8]

[5] Angeloni, I., 2014, "Stress-testing banks: are econometric models growing young again?" Speech by Ignazio Angeloni, Member of the Supervisory Board of the ECB, at the Inaugural Conference for the Program on Financial Stability, School of Management, Yale University, August 1, https://bit.ly/2lKekBq
[6] Chan-Lau, J. A., 2017, "Lasso regressions and forecasting models in applied stress testing," IMF Working Paper WP/17/108, https://bit.ly/2mcljTR
[7] Korbut, I., 2017, "Machine learning algorithms: which one to choose for your problem," Medium, October 26, https://bit.ly/2iFkZeF
[8] Zhou, L., 2018, "Simplify machine learning pipeline analysis with object storage,"Western Digital Blog, May 3, https://bit.ly/2lLt3Mm

be straightforwardly applied to group customers into sets, which can often be an insightful start for further analysis.

3.3.1 EXAMPLE: CLUSTERING IN FRAUD DETECTION

Clustering is an effective technique for anomaly detection. In financial services, this is useful in anti-money laundering to identify unusual or fraudulent transactions. An example of this is Citibank, who have entered into a strategic partnership with Feedzai,[9] a machine learning solutions business, to provide real time fraud risk management using machine learning. Feedzai's solution[10] transforms data streams to create risk profiles for fraud detection, using machine learning to process client transactions automatically. Feedzai is able to do this in millisecond timescales, providing Citibank with a highly rapid and powerful fraud detection product.

3.4 What should be recommended? – collaborative filtering

Recommendation engines are all around us in the form of Amazon telling us what we might be interested in buying, through to Facebook finding people we may know. Recommendation engines have been slow to take off in the financial services sector but have the potential to change the way that portfolios are optimized and how products are cross-sold. We will look at collaborative filtering, which is the most common machine learning method underlying recommendation engines. The name is derived from the idea that the data from many similar users can collaborate to recommend products to a customer in the way that friends would collaborate and recommend purchases to one another in the real world. Two algorithms that can be used for collaborative filtering are 'nearest neighbors' and 'matrix factorization'.

The 'nearest neighbors' technique is a type of classification algorithm used in collaborative filtering that uses historical data of users' ratings for products as training data for predictions about which other products specific users are likely to buy. This data can use either implicit or explicit ratings of products. Implicit ratings are ones where the user has given a numeric rating to a product and explicit ratings are inferred from things like page views or purchases and returns. This type of algorithm finds a user's "nearest neighbors" in terms of taste, based on ratings and recommends products that those customers have bought. One of the biggest problems with nearest neighbors in collaborative filtering is that data may be very sparse because users have not rated enough products, or a product has not been rated by enough users. For this reason, it is best practice to use this method if you have large amounts of user data.

Unlike nearest neighbors, matrix factorization creates latent features that are not present in the historical data but are created from underlying patterns in this data. For example, a youth account and an educational loan may all have a "relating to children" feature. Although the algorithm does not know what the feature represents because it has no human knowledge, it knows that such a feature exists and relates specific products to one another. This means that products can be recommended based on users' historical data even if they have not rated the same products. It upgrades the recommendation engine from "people who bought this product also bought another product" to "people who buy these types of products also buy another type of product". This type of algorithm typically needs less user data to get started than nearest neighbors.

3.4.1 EXAMPLE: COLLABORATIVE FILTERING IN PRIVATE BANKING

InCube is a company that is developing bespoke recommender engines for private banks. These engines use several AI techniques, including collaborative filtering, to recommend products for clients to add to their existing portfolios. One aspect that makes these recommendations successful is that before any AI algorithms are utilized, business rules are applied to ensure that regulatory requirements are met and that conditions which are obvious to humans but are not necessarily taken into account by algorithms are met. For instance, a product will not be recommended to a client if that product is already part of their portfolio.

3.5 How can this be optimized? – reinforcement learning

Reinforcement learning (RL) is the third basic machine learning paradigm, alongside SL and UL, and is best suited to problems that involve many complex variables and is based on a method of trial-and-improvement, iteratively testing, and refining models to give the 'best' outcome. This final 'best' solution can be highly varied in form. Possibilities for these types of machine learning applications include finding

[9] Feedzai, 2018, "Citi partners with Feedzai to provide machine learning payment solutions," press release, December 19, https://bit.ly/2INlogv
[10] Feedzai.com

optimal strategies for playing casual games, such as chess, or optimizing states, such as a headcount number, maximizing employee utility while maintaining an acceptable risk level of under-resourcing at times of heightened workload.

RL is a computationally-intensive procedure that requires many iterations in comparison to SL and UL methods. For this reason, an RL approach should only be considered when the solution model cannot be cast in the frame of SL or UL approaches. RL works best where the problem can be considered as behavior driven, the algorithm can be one that learns how to act in a certain environment to maximize reward ('reward-based learning'), or the decision-making processes being modeled can be considered to be partly random and partly determined by the actions of a decision maker.

Central to developing RL models is the setup of a simulation environment. This environment provides the means for training the model by providing a simulated response to the model. Let us consider the example of RL for teaching a self-driving car (Figure 4). Here, the simulated environment enables teaching of the algorithm to keep to lanes and avoid simulated humans, all in a safe and protected manner. When we speak about an algorithm performing RL tasks, we refer to it as an agent.

3.5.1 EXAMPLE: REINFORCEMENT LEARNING IN ALGORITHMIC TRADING

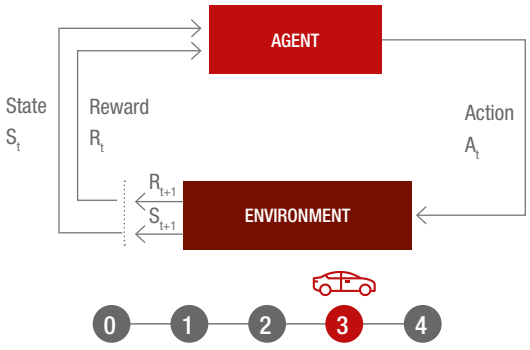
In this example, the simulated environment in which the agent learns is a simulation of market conditions. In algorithmic trading that is based on traditional statistics, there are typically two components of a trade. The first is known as the policy, which would be dictated by the traders, and the second part is known as the mechanism, which would be executed by a computer. In machine learning-driven algorithmic trading, the policy would be learned from training data of past trades using RL and a trader would not be involved.

4. TOOLING

Traditionally, the success of machine learning has relied on human ML experts to perform tasks such as data pre-processing and cleaning, feature selection and model construction, parameter optimization, model postprocessing, and analysis. However, today, new machine learning algorithms can autonomously identify patterns, analyze data, and even interpret data by producing reports and data visualizations. There is now an ever-growing array of tools and services designed to facilitate big data analytics outside of the technology lab, and across the organization as a whole. Not only that, these tools come boxed and wrapped up with an easy-to-use platform, providing an agility unlike that of the coding-heavy, statistical world of traditional machine learning methods. Much of the technical analysis work is now delegated to the machine.

These developments have extended their reach to tools that can incorporate a machine learning capability to automate data preparation, insight discovery, and data science. Large technology players have capitalized on this trajectory through their 'machine learning as a service' offerings. Google, AWS (Amazon Web Services), and Microsoft are expediting this trend. Google launched BigQuery, a tool designed to make it easy to access and manipulate large datasets, requiring knowledge of SQL only as opposed to traditional data science languages such as R and Python. More recently, Google added a new capability to BigQuery by introducing BigQuery ML, a tool to build and deploy machine learning models through simple, broadly understandable SQL statements. Analysts can build and operationalize machine learning models on large-scale structured or semi-structured data directly inside BigQuery, using simple SQL in a fraction of the time.

Figure 4: Training a self-driving algorithm through reinforcement learning



In this example environment, there exists an agent (the car) that can perform two actions: move forwards or backwards. This environment features four states, with the default state being 1. If the car takes the backward action randomly, the environment assigns it a new state of 0. The goal of the car is to reach state 3. The car applies random actions until it reaches its goal, and then terminates. If the car chooses to move backwards when at position 0 or forward at position 4, the car remains in place.

Source: Modified version of Huang (2018)[11]

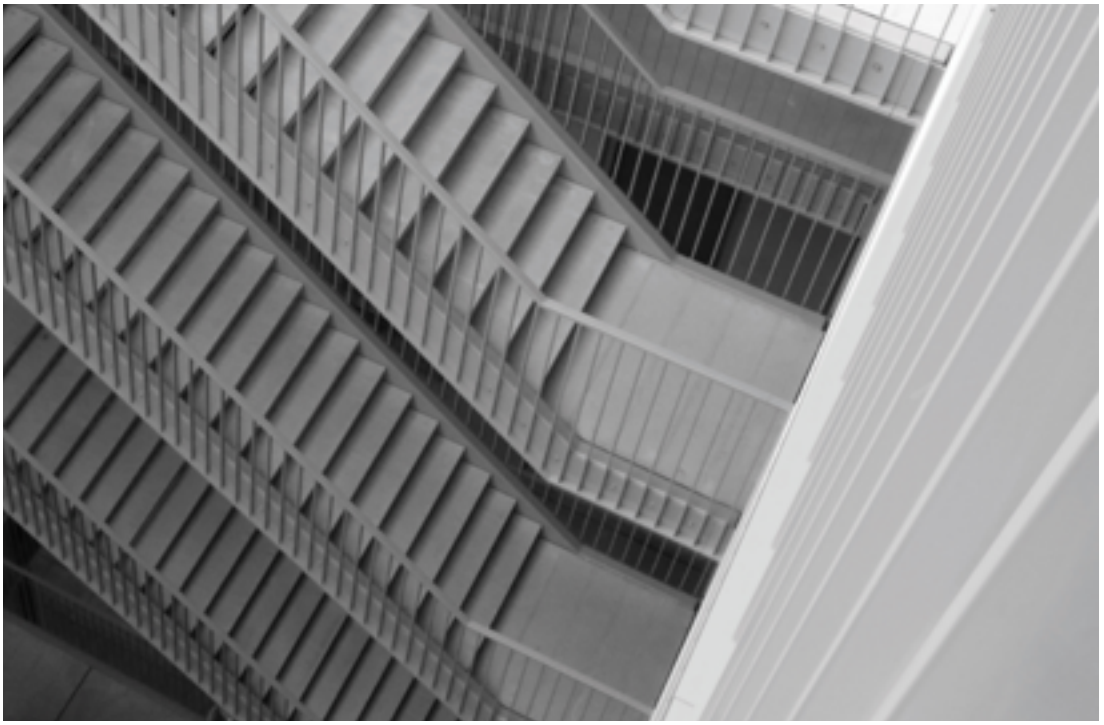
[11] Huang, S., 2018, "Introduction to various reinforcement learning algorithm. Part I (Q-learning, SARSA, DQN, DDPG)," Medium, January 12, https://bit.ly/2K66Tje

AWS has launched SageMaker, a fully-managed platform that enables data scientists to quickly and easily build, train, and deploy machine learning models at scale. The solution removes barriers that typically slow down developers who want to use machine learning by promoting a visual-centric approach to model development that integrates with other Amazon tools. Similarly, Microsoft Azure ML Studio offers a cloud-based environment that can be accessed from the web browser and used to create machine learning-based models on any dataset. Following the general trend of accessible machine learning, Azure's ML platform gives data scientists, without any prior machine learning experience, the ability to experiment with machine learning on datasets.

Other tools also exist to allow those who know the theory of deep learning but have no coding experience to create a deep learning model within minutes without a single line of code. Lobe (recently acquired by Microsoft) is an example of this. Lobe offers users a clean drag-and-drop interface for building deep learning algorithms from scratch, without having to know the ins and outs of machine learning libraries such as TensorFlow, Keras, or PyTorch.

4.1 Example: Financial services machine learning platform

A user-friendly machine learning platform being used in the financial services industry is Kensho, which is a company focused on the development of tailored machine learning-based early warning systems. On voting to leave the European Union in June 2016, traders with access to Kensho's product had at their disposal powerful insights to their advantage.[12] The platform provided access to research-quality machine learning-driven information, informing them that, historically, a populist vote such as Brexit tends to result in a long-term drop in local currency. Traders were then able to exploit this information and profit from changes in the currency rate. Early-warning models are not limited to trading and have uses for predicting credit risk and due-diligence in order to predict future bankruptcy risk in the supply chain.



[12] Gara, A., 2017, "Kensho's AI for investors just got valued at over $500 million in funding round From Wall Street," Forbes, February 28, https://bit.ly/2IPGEm0

"One approach to considering if the way the data is used is ethical is to question whether customers are being empowered with greater choice, or disempowered by restricting choice."

5. ETHICAL CONSIDERATIONS

Machine learning has suddenly shifted from relative academic obscurity to being in common usage; however, this has not always led to good outcomes. In recent years, some well-designed machine learning projects have led to the unintended reputational damage of institutions. This is because their work has been considered "unethical" due to use of sensitive personal data or the automation of key personal decisions without the necessary level of human oversight. An example of this latter case is Microsoft's twitter chatbot, Tay,[13] released in 2016. Tay was designed to mimic speech patterns of a typical millennial, and was equipped with the ability to learn from the Twitter conversations she engaged in. This ability proved to be to Tay's detriment, as it was soon observed making derogatory comments, in one tweet[14] proclaiming "bush did 9/11 and Hitler would have done a better job than the monkey we have now. Donald trump is the only hope we've got." [sic]

With advanced analytics techniques becoming a key competitive advantage in the digital age, many businesses must quickly learn to adapt to new ethical concerns. The best defense against ethical breaches is to ensure that all members of a machine learning project are aware of their responsibility to understand what is being done, and how to raise concerns. It is no longer acceptable for data scientists to behave like the naïve, impartial computers they instruct; it is imperative that they take time to consider the impacts of adding sensitive information to their models. Likewise, stakeholders and managers cannot treat their data science team as a black box that will return with an insight a few weeks after the initial project brief. They must take an active interest in what data is being used, and how this could be viewed by an external audience. Ill-conceived machine learning projects are given little leeway in public discourse, and must, therefore, have clear ethical guidelines, because while key insights can help elevate a market leading business, they can also result in irreparable loss of trust.

These ethical concerns extend, not only to the models themselves but the data upon which they have been trained. A good example of a breach of consumer trust around data collection is the scandal surrounding the 'Unroll.me' application, which scans email inboxes to flag subscriptions from which users may wish to unsubscribe. It came to light that the application was gathering information from users' inboxes and selling it to companies such as Uber, with the revelation resulting in public outcry. One approach to considering whether the way the data is used is ethical is to question whether customers are being empowered with greater choice, or disempowered by restricting choice. These choices can relate to customers' power to decide whether they would like their data to be used or to the number of products and services to which the customer has access when their data is used.

It is important in projects which use machine learning to know who is responsible for decisions made by algorithms; this should be someone within the team with experience developing the model. These decisions should not be left to legal teams after development has taken place but rather be taken into consideration throughout the development lifecycle. In addition, the General Data Protection Regulation (GDPR) requires that organizations are able to provide an explanation of any decisions made using their data. We may not be able to fully explain complex models, for example artificial neural networks (a computational brain-like network of synapses which typically include hidden interconnected layers), but we can explain what data was included in building the model and how important each "feature" or variable is in the results that are generated. We can do this through feature attribution, which determines how much influence changing a particular variable has on changing the results generated by the model. Data scientists must be willing to take responsibility for the decisions that models they have trained make.

Using feature attribution and other explainable AI techniques to check models also allows human subject matter experts (SMEs) to pick up errors in logic that may have originated from anomalies within training data. A famous example of a 'model gone wrong'

[13] Price, R., 2016, "Microsoft is deleting its AI chatbot's incredibly racist tweets," Business Insider, March 24, https://bit.ly/2IJE1Cb
[14] Hunt, E., 2016, "Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter," The Guardian, March 24, https://bit.ly/23B1uAG

is the algorithm designed at the University of Pittsburgh to predict outcomes in pneumonia patients in the mid-1990s. The model correctly learned that, for the data on which it had been trained, people who also had asthma were less likely to die than other patients and so recommended that these asthmatic patients be sent home with antibiotics rather than admitted to hospital. However, the real underlying reason people with asthma were less likely to die in this particular dataset was because they were almost immediately placed in intensive care units where they received such a high level of care that they rarely have negative outcomes. Sending patients with asthma home with antibiotics could potentially have led to fatal outcomes for these patients.

6. CONCLUSION

Machine learning can be a powerful addition to any data analytics tool kit but requires careful planning and a high-level understanding of the techniques involved by all stakeholders. It is also important that data scientists and organizational stakeholders keep ethical considerations in mind. Table 2 summarizes the challenges that may be faced when implementing machine learning projects and best practices that can mitigate these challenges and harness the power of machine learning to generate value for the business.

Table 2: Challenges and best practices in machine learning

CHALLENGES	BEST PRACTICES
Misconceiving machine learning as a 'black box' that can be used to solve all problems with a one-size-fits-all approach	Set simple and concise objectives based on a single specific task or use-case
Machine learning may not be the correct solution to the problem, and it might be solved with simpler alternatives	Set clearly defined success criteria to monitor the performance of the model or system, for example, reduce full time equivalent (FTE) by x% or reduce customer wait times by x amount
Cognitive systems improve over time and results and benefits may not be immediately realized	Manage stakeholders' expectations by informing them of the limitations as well as abilities of predictive models and automation systems
Defining processes for obtaining and maintaining high quality clean data	Educate staff that not all data is predictive, and that machine learning cannot be relied on to solve all problems associated with operational inefficiency
Obtaining long-term stakeholder buy-in to gain the real benefits of these systems	Clearly define how the machine learning process works, what the inputs and outputs are, and how these integrate with the existing processes and methodologies
Proof of concepts (PoCs) and production systems have very different build and deployment requirements; migration from a PoC to a production system is likely to require an entirely new architecture and rebuild	PoCs and prototypes should be built to test and demonstrate functionality and win stakeholder buy-in before production model build and deployment

ABOUT CAPCO

Capco is a global technology and management consultancy dedicated to the financial services industry. Our professionals combine innovative thinking with unrivalled industry knowledge to offer our clients consulting expertise, complex technology and package integration, transformation delivery, and managed services, to move their organizations forward.

Through our collaborative and efficient approach, we help our clients successfully innovate, increase revenue, manage risk and regulatory change, reduce costs, and enhance controls. We specialize primarily in banking, capital markets, wealth and asset management and insurance. We also have an energy consulting practice in the US. We serve our clients from offices in leading financial centers across the Americas, Europe, and Asia Pacific.

WORLDWIDE OFFICES

APAC
Bangalore
Bangkok
Hong Kong
Kuala Lumpur
Pune
Singapore

EUROPE
Bratislava
Brussels
Dusseldorf
Edinburgh
Frankfurt
Geneva
London
Paris
Vienna
Warsaw
Zurich

NORTH AMERICA
Charlotte
Chicago
Dallas
Houston
New York
Orlando
Toronto
Tysons Corner
Washington, DC

SOUTH AMERICA
São Paulo

WWW.CAPCO.COM

CAPCO