Explaining AI in Finance: Past, Present, Prospects

Barry Quinn

```
```{r}
library(reticulate)
# install_miniconda()
# conda_create(envname = "XAI")
# conda_install(envname = "XAI",packages = c("shap"))
#conda_install(envname = "XAI",channel="numba",packages = c("llvmlite"))
#py_install(packages = c("scikit-learn","shap"),envname = "XAI")
#py_install(packages = c("llvmlite"),envname = "XAI")
use_condaenv(condaenv = "XAI")
```
```

Introduction

The rapid digital transformation of the finance industry over the past few decades has been predominantly driven by the integration of Artificial Intelligence (AI) and machine learning technologies. These technologies have heralded a new era in finance, catalyzing innovations in trading, risk management, fraud detection, customer service, and a plethora of other areas, bringing significant changes to business models, operations, and services (Arner, Barberis, and Buckley 2020). Today, financial institutions leverage these advanced technologies to generate insights, automate processes, and improve decision-making.

Despite the revolutionary potential of AI, its application in finance is not devoid of challenges. A significant issue arises around the transparency and interpretability of AI decision-making, often described as the 'black box' problem. This term refers to the difficulty in understanding how complex AI and machine learning models arrive at their decisions (Dhar 2018). This opacity presents substantial ethical, legal, and practical challenges, especially in an industry as regulated and risk-averse as finance (Bhatt, Xiang, and Sharma 2020).

In response to these concerns, the field of Explainable Artificial Intelligence (XAI) has emerged with an aim to make AI's decision-making process more transparent and comprehensible to human users (Molnar 2020). This development is particularly crucial in the financial sector, where understanding the rationale behind decisions can have enormous implications for trust, compliance, risk management, and customer satisfaction.

This paper aims to critically analyze the role of XAI in finance, tracing its historical development, examining its current applications, and exploring its future prospects. By providing a comprehensive review of XAI in the context of finance, we hope to shed light on its importance and potential for the industry while addressing ongoing challenges and areas for future research.

Literature review

Past: Early Application of AI in Finance

The integration of AI in finance has a history that dates back to the latter half of the 20th century. Initially, financial institutions deployed rule-based systems for a variety of functions, such as automated trading and risk analysis. These were the earliest forms of AI applied to finance and were quite simplistic compared to today's advanced systems.

These early AI systems were based on sets of pre-programmed rules, often developed by human experts, and were typically deterministic in nature. That is, given the same input, these systems would always provide the same output. They were transparent and easily interpretable because they followed clearly defined, pre-set rules (Gomber et al. 2018).

However, the effectiveness of rule-based systems was limited by their rigid, inflexible design. These systems were not designed to learn from new data or adapt to changing conditions, making them less useful in the dynamic world of finance, which is characterised by evolving markets, changing regulatory landscapes, and unpredictable economic conditions (Dhar 2018).

The desire for more adaptive, responsive systems led to the development of machine learning algorithms. Machine learning represented a significant advancement in the field of AI. Unlike rule-based systems, machine learning models could learn from data, identify patterns, and make predictions or decisions based on those patterns. In finance, machine learning algorithms found applications in a range of areas, from predicting stock prices to identifying fraudulent transactions (Athey 2021).

The transition from rule-based systems to machine learning models marked a pivotal shift in AI's role in finance. However, this transition was not without its challenges. Machine learning models, particularly more complex ones like neural networks, introduced a level of opacity and complexity that made it difficult for human users to understand how they made decisions (Rudin 2019). This lack of transparency and interpretability in machine learning systems

became known as the 'black box' problem and forms the backdrop against which the field of explainable AI has emerged.

In summary, the past of AI in finance was marked by the transition from transparent but inflexible rule-based systems to powerful but opaque machine learning models. The need to address the transparency issues introduced by machine learning has led to the development of explainable AI, the current and future state of which will be discussed in the following sections.

Present: The Rise of Machine Learning and XAI

Certainly, here's an expanded version of the "Present" section:

In the present landscape, machine learning models, such as neural networks, decision trees, and support vector machines, have become integral parts of financial institutions. These models perform a variety of tasks, including credit scoring, fraud detection, algorithmic trading, portfolio optimization, and customer segmentation, among others (Athey 2021).

Despite their efficiency and sophistication, these models often work as 'black boxes,' where the internal decision-making process is obscured from the users. This opaqueness can pose considerable challenges. On a practical level, it hinders human users, such as loan officers or portfolio managers, from understanding and trusting the model's decisions. On a regulatory level, it poses problems for accountability and compliance, especially in jurisdictions where decisions affecting individuals must be explainable (Rudin 2019).

Enter the field of Explainable Artificial Intelligence (XAI), which seeks to make AI's decision-making process more transparent and interpretable. A variety of techniques fall under the umbrella of XAI, and these can be broadly classified into two categories: model-specific methods and model-agnostic methods (Molnar 2020).

Model-specific methods, such as coefficient interpretation in linear regression or feature importance in decision trees, provide insights into how these specific models operate. However, their application is limited to the particular model types for which they were developed (Molnar 2020).

Model-agnostic methods, on the other hand, can be applied to any machine learning model. They seek to provide explanations for individual predictions regardless of the complexity or type of the underlying model. Examples of these techniques include Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP). LIME offers explanations by approximating the prediction of a complex model with a simpler, locally-fitted model around the prediction point (Ribeiro, Singh, & Guestrin, 2016). SHAP, meanwhile, allocates the contribution of each feature to the prediction for individual data points, based on concepts from cooperative game theory (Lundberg and Lee 2017).

Yet, despite these advancements, achieving true explainability in AI remains a significant challenge. Many of these methods provide post-hoc explanations, which attempt to interpret the model's behavior after it has been trained. This process often involves a trade-off between accuracy and interpretability, with more complex models offering greater accuracy but less interpretability (Bhatt, Xiang, and Sharma 2020).

Moreover, explainability is not just a technical problem but also a human-centered one. The effectiveness of an explanation largely depends on the recipient's perspective and the context in which it is given (**Mill?**) (Miller, 2019). What may be a satisfactory explanation to a data scientist might be incomprehensible to a loan officer or a customer, indicating that the development of XAI needs to take into account the human factors of understandability and trust.

In conclusion, the present state of XAI in finance is marked by considerable advancements, but also by ongoing challenges. The shift towards more transparent and interpretable AI models is underway, with various methods being developed and applied. However, achieving a balance between the complexity (and thereby, the performance) of models and their interpretability remains a significant hurdle. As we look towards the future, it is crucial that these challenges are addressed, and XAI continues to evolve to meet the demands of transparency and interpretability in the financial industry.

Prospects: The Future of XAI in Finance

As we move into the future, the demand for transparency and interpretability in AI systems within the finance sector is expected to grow. The prospective advancements and challenges in the field of XAI reflect this.

One significant direction for future research and development is integrating explainability directly into the model-building process, rather than treating it as an afterthought. This approach, often called intrinsic explainability, involves building models that are naturally interpretable, such as explainable boosting machines or interpretable decision sets (Lakkaraju, Bach, and Leskovec 2016; Lou, Caruana, and Gehrke 2012). The development of such models can help mitigate the trade-off between accuracy and interpretability that characterizes post-hoc explanation methods.

Furthermore, as the field of XAI evolves, it will be essential to focus on the users' perspective. What constitutes a 'good' explanation can vary based on the recipient and the context. Therefore, future XAI methods should consider tailoring explanations to different users' needs and capabilities. They should also address how to best communicate these explanations to ensure they are understandable and useful (Miller, 2019).

Moreover, as AI and XAI become more commonplace in finance, it's likely that regulations will evolve to address the new challenges they present. The European Union's General Data Protection Regulation (GDPR) has already introduced a 'right to explanation', where individuals

can ask for explanations of decisions made by automated systems that affect them (Goodman and Flaxman 2017). In the future, we might see more regulations like these, requiring financial institutions to provide clear, understandable explanations for AI-based decisions.

However, implementing such regulations comes with its own challenges. Regulators will need to define what constitutes an 'explanation' and a 'decision' in the context of AI. They will also need to set standards for how detailed and understandable these explanations must be (Edwards and Veale 2017).

In summary, the future of XAI in finance is ripe with opportunities for making AI decision-making more transparent and accountable. However, it also presents challenges that need to be addressed through continued research, development, and thoughtful regulation.

Methodology

Econometrics and XAI

Explainable AI (XAI) could certainly play a significant role in improving the field of econometrics, which is the application of statistical methods to economic data in order to give empirical content to economic relations.

Traditionally, econometric models have been designed to be inherently interpretable, as they often depend on linear relationships and other simplistic assumptions to ensure that the parameters of the model can be easily interpreted. However, these assumptions can be limiting, as they might not fully capture the complexities of real-world economic phenomena.

With the advent of machine learning, econometricians have been able to create models that can learn complex patterns from data, leading to more accurate predictions. But the drawback is that these models are often 'black boxes,' making it difficult to understand how they make decisions.

This is where XAI comes in. By using techniques developed under the umbrella of XAI, econometricians could make their machine learning models more transparent, allowing them to understand how each input variable contributes to the model's predictions. This increased transparency could make these models more acceptable to economists and policymakers, who need to understand the decision-making process to make informed decisions.

For instance, SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) are two such techniques that could be applied to make machine learning models more interpretable. These techniques provide explanations for individual predictions, helping to understand the contributions of different features to the model's output.

In addition to improving model transparency, XAI could also help address some of the statistical challenges in econometrics. For example, XAI could help econometricians understand

the variable importance in their models, which could be useful in addressing issues related to multicollinearity, where independent variables in a regression model are highly correlated.

In conclusion, XAI holds considerable potential in improving econometrics, particularly as the field increasingly incorporates machine learning models. By making these models more transparent, XAI can help econometricians and policymakers better understand and trust the predictions derived from these models, which in turn can lead to better decision-making.

Shapley values versus OLS regression coefficients

Absolutely, let's explore the analogy between Shapley values and the coefficients in a linear regression model.

In a linear regression model, each predictor variable is assigned a coefficient that represents its partial effect on the outcome variable, controlling for all other predictors. The coefficient can be interpreted as the expected change in the outcome variable for a one-unit change in the predictor, holding all other predictors constant.

Analogously, the Shapley value for a player in a cooperative game represents the average contribution of the player to the worth of all possible coalitions that the player can be a part of. The Shapley value takes into account all possible ways in which the coalition can be formed and averages over them.

In both cases, the goal is to fairly distribute some total quantity (the total worth of the grand coalition in a cooperative game, or the total variance of the outcome variable in a linear regression model) among different contributors (the players in a cooperative game, or the predictor variables in a linear regression model).

However, there are important differences as well. While linear regression assumes a specific linear and additive form for the relationship between predictors and the outcome, Shapley values make no such assumption. Shapley values can handle any type of game, including non-cooperative games and games with complex interactions between players.

Also, the computation of Shapley values takes into account all possible orders in which players can join the coalition, reflecting the idea that the contribution of a player may depend on which other players are already in the coalition. In contrast, linear regression coefficients are typically computed using a method (like ordinary least squares) that does not consider different orders of entering the predictors into the model.

In the context of explainable AI, the Shapley value concept has been applied to machine learning models to compute the contribution of each feature to the prediction for a particular instance. This can provide more nuanced and reliable interpretations than simply looking at the coefficients of a linear model, especially for complex models that capture non-linear and interactive effects.

Coalition game and shapley values

To fix ideas, let's consider a simple cooperative game involving three players: A, B, and C. The worth of each coalition of players is given by a characteristic function v:

- v({}) = 0 (worth of the empty coalition)
- v({A}) = 100
- v({B}) = 200
- v({C}) = 300
- v({A, B}) = 400
- v({A, C}) = 500
- v({B, C}) = 600
- v({A, B, C}) = 800

We want to distribute the total worth of the grand coalition (v({A, B, C}) = 800) among the players in a way that reflects their contribution to the coalition.

The Shapley value is one way to do this. For each player, it computes the average marginal contribution of the player to all possible coalitions. This is done by considering all permutations of the players and for each permutation, adding up the marginal contributions of the player when they join the coalition.

Confronting regression coefficients using XAI

In a linear regression model, each predictor variable is assigned a coefficient that represents its partial effect on the outcome variable, controlling for all other predictors. The coefficient can be interpreted as the expected change in the outcome variable for a one-unit change in the predictor, holding all other predictors constant.

Analogously, the Shapley value for a player in a cooperative game represents the average contribution of the player to the worth of all possible coalitions that the player can be a part of. The Shapley value takes into account all possible ways in which the coalition can be formed and averages over them.

In both cases, the goal is to fairly distribute some total quantity (the total worth of the grand coalition in a cooperative game, or the total variance of the outcome variable in a linear regression model) among different contributors (the players in a cooperative game, or the predictor variables in a linear regression model).

However, there are important differences as well. While linear regression assumes a specific linear and additive form for the relationship between predictors and the outcome, Shapley values make no such assumption. Shapley values can handle any type of game, including non-cooperative games and games with complex interactions between players.

Also, the computation of Shapley values takes into account all possible orders in which players can join the coalition, reflecting the idea that the contribution of a player may depend on which other players are already in the coalition. In contrast, linear regression coefficients are typically computed using a method (like ordinary least squares) that does not consider different orders of entering the predictors into the model.

In the context of explainable AI, the Shapley value concept has been applied to machine learning models to compute the contribution of each feature to the prediction for a particular instance. This can provide more nuanced and reliable interpretations than simply looking at the coefficients of a linear model, especially for complex models that capture non-linear and interactive effects.

A finance example

To steelman the case here is a simulated example. We'll use the `sklearn` and `shap` libraries in Python to fit a linear regression model and a more complex model (Random Forest) to the simulated data. Then we'll use the shap library to compute SHAP values for the Random Forest model.

Let's consider a simple three-factor model (size, value, and momentum factors) to predict asset returns.

First, let's install the necessary packages (if not already installed):

Now let's import the necessary packages and generate some simulated data: g

```
```{python}
import numpy as np
import pandas as pd
from sklearn.linear_model import LinearRegression
from sklearn.ensemble import RandomForestRegressor
import shap
import warnings
warnings.filterwarnings("ignore")

# Set a seed for reproducibility
np.random.seed(0)

# Generate simulated factor values
n = 1000  # number of assets
size = np.random.normal(0, 1, n)
value = np.random.normal(0, 1, n)
momentum = np.random.normal(0, 1, n)
```
```