

How to Fight Financial Crime with AI

Data Intensive Science CDT Seminar 22nd November 2023





F E A T U R E S P A C E



Fraud is a big problem

Fraud is everywhere

40% of reported crime in UK

https://publications.parliament.uk/pa/cm5803/cmselect/cmjust/12/report.html



Fraud is damaging





Fraudsters are sophisticated

Tech

AI clones child's voice in fake kidnapping scam

'I never doubted for one second it was her,' mother says

Anthony Cuthbertson • Thursday 13 April 2023 17:05 BST • [

FraudGPT: The Villain Avatar of ChatGPT



Rakesh Krishnan: Tue, Jul 25, 2023 @ 08:03 AM









Threat Actor

FraudGPT





Featurespace fights fraud

We use machine learning to fight fraud

F E A T U R E S P A C E

OUTSMART RISK



Our work has global reach



We're based in Cambridge







Work with us!

We have a cool team

We are the **Innovation Lab** @ Featurespace:



David SuttonChief Innovation Officer

Ex: Research Associate in Cosmology @ Cambridge



Iker Perez
Principal Research Scientist

Ex: Assistant Professor in Statistics and Data Science @ Nottingham



Jason Wong
Lead Research Engineer

Ex: MSc in ML @ Edinburgh, Maths @ Oxford



Piotrek Skalski Senior Research Scientist

Ex: PhD in Physics @ Cambridge



Stuart Burrell
Senior Research Engineer

Ex: Research Associate in Mathematics @ St Andrews, MPhil in MLMI @ Cambridge



Kamal Parameswaran

Research Scientist

Ex: Senior Research Associate in Safe & Ethical AI @ Turing



We work on cool research

Recent publications:

Attribution of Predictive Uncertainties in Classification Models

Featurespace Research, Cambridge, United Kingdom

Predictive uncertainties in classification tasks are often a consequence of model inadequacy or insufficient training data. In popular applications, such as image processing, we are often required to scrutinise these uncertainties by meaningfully attributing them to input features. This helps to improve interpretability assessments. However, there exist few effective frameworks for this purpose. Vanilla forms of popular methods for the provision of saliency masks, such as SHAP or integrated gradients, adapt poorly to target measures of uncertainty. Thus, state-of-the-art tools instead proceed by creating counterfactual or adversarial feature vectors, and assign attributions by direct comparison to original images. In this paper, we present a novel framework that combines path integrals. counterfactual explanations and generative models, in order to procure attributions that contain few observable artefacts or noise. We evidence that this outperforms existing alternatives through quantitative evaluations with popular benchmarking methods and data sets of varying complexity.

language processing [Xiao and Wang, 2019], network analysis [Perez and Casale, 2021] or image processing [Kendall and Gal, 2017], to name only a few.

Thus, there exists a growing interest in methods for uncertainty estimation [e.g. Depeweg et al., 2018, Smith and Gal. 2018. Van Amersfoort et al., 2020. Tuna et al., 20211 for purposes such as procuring adversarial examples, active learning or out-of-distribution detection. Recent work has proposed mechanisms for the attribution of predictive uncertainties to input features, such as pixels in an image [Van Looveren and Klaise, 2019, Antoran et al., 2021, Schu et al., 2021], with the goal of complementing interpretability tools disproportionately centred on explaining model scores. and to improve transparency in deployments of predictive models. These methods proceed by identifying counterfactual (in-distribution) or adversarial (out-of-distribution) explanations, i.e. small variations in the value of input features which output new model scores with minimal uncertainty This has helped understand the strengths and weaknesses of various models. However, the relative contribution of individual pixels to poor model performance is up to human guesswork, or assigned by plain comparisons between an image and its altered representation. We report that uncertainty attributions derived following these approaches perform poorly, when measured by popular quantitative evaluations of image saliency maps.

Locally Differentially Private Embedding Models in Distributed Fraud Prevention Systems

Iker Perez*, Jason Wong*, Piotr Skalski*, Stuart Burrell*, Richard Mortier†, Derek McAuley‡ and David Sutton* *Featurespace, United Kingdom [†]University of Cambridge, United Kingdom

[‡]University of Nottingham, United Kingdom {iker.perez, jason.wong, piorr.skalski, stuart.burrell, david.sutton}@featurespace.co.uk rmm1002@cam.ac.uk, derek.mcauley@nottingham.ac.uk

machine learning solutions in fraud prevention. However, preven-settings 4-6 and often rely on Recurrent Neural Nettion systems are commonly serviced to financial institutions in isolation, and few provisions exist for data sharing due to fears of unintentional leaks and adversarial attacks. Collaborative learn ing advances in finance are rare, and it is hard to find real-world insights derived from privacy-preserving data processing systems. In this paper, we present a collaborative deep learning framework for fraud prevention, designed from a privacy standpoint, and awarded at the recent PETs Prize Challenges. We leverage latent embedded representations of varied-length transaction sequences. along with local differential privacy, in order to construct a data release mechanism which can securely inform externally hosted fraud and anomaly detection models. We assess our contribution vo distributed data sets donated by large payment networks, and demonstrate robustness to popular inference-time attacks. along with utility-privacy trade-offs analogous to published work in alternative application domains.

Index Terms—deep learning, privacy, financial systems

I INTRODUCTION

Innovation in Machine Learning (ML) methodologies has made advances to provision financial system actors with tools for fraud prevention and anti-money laundering, which account for losses estimated at 2-5% of the global Gross Domestic Product [1]. There currently exists an ecosystem of technology partners and risk management vendors that service ML solu-

Abstract—Global financial crime activity is driving demand for increasingly researched and deployed in controlled production works [7]-[9] or Attention mechanisms [10]. Integrating DL within fraud prevention reduces the need for extensive feature engineering [TT] and facilitates the retrieval of embedded representations for financial transactions, which capture fraud typologies in a manner reusable across systems. However, DL model deployments may memorise and unintentionally leak personal or commercially sensitive information when exposed to partnering institutions and external parties [12]-[16].

Privacy leaks may compromise both training data and inference-time inputs. Commonly, attackers target public interfaces that expose inference-time functionalities for distributed models, however, these may also be susceptible to trainingtime attacks by malicious parties in a consortium [16], [17]. Attack typologies are diverse, and popular choices include membership, attribute inference and model inversions. In all cases, a model becomes susceptible to attacks due to external over-exposure to its logic, architecture, gradients, predictive confidence or latent representations [18]. Thus, there exists a growing body of work exploring model design, training and deployment from privacy standpoints across a variety of domains including text, vision or speech, and considering multiple threat models and adversarial exploitation strategies

Towards a Foundation Purchasing Model: Pretrained Generative Autoregression on Transaction Sequences

Piotr Skalski David Sutton Stuart Burrell Innovation Lab Featurespace Featurespace Cambridge, UK Cambridge, UK Cambridge, UK piotr.skalski@featurespace.co.uk david.sutton@featurespace.co.uk stuart.burrell@featurespace.co.uk

> Jason Wong Innovation Lab Innovation Lab Featurespace Featurespace Cambridge, UK iker.perez@featurespace.co.ul

Cambridge, UK jason.wong@featurespace.co.uk ACM Reference Format:

ABSTRACT

Machine learning models underpin many modern financial systems for use cases such as fraud detection and churn prediction. Most are based on supervised learning with hand-engineered features, which relies heavily on the availability of labelled data. Large selfsupervised generative models have shown tremendous success in natural language processing and computer vision, yet so far they haven't been adapted to multivariate time series of financial transactions. In this paper we present a generative pretraining method that can be used to obtain contextualised embeddings of financial transactions. Benchmarks on public datasets demonstrate that it outperforms state-of-the-art self-supervised methods on a range of downstream tasks. We additionally perform large-scale pretraining of an embedding model using a corpus of data from 180 issuing banks containing 5.1 billion transactions and apply it to the card fraud detection problem on hold-out datasets. The embedding model significantly improves value detection rate at high precision thresholds and transfers well to out-of-domain distributions.

CCS CONCEPTS

 Applied computing → Online banking;
 Computing methodologies → Unsupervised learning; Learning latent represen-

New York, NY, USA, 9 pages, https://doi.org/10.1145/3604237.3626850 1 INTRODUCTION

Foundation models have seen tremendous success and wide adoption within the past couple of years. They have proven their ability to leverage large corners of data and scale to hundreds of hillions of parameters. On textual data, these models can be used not only to generate human-level text but also to produce contextualised embeddings of individual tokens, sentences, and even whole doc uments that can be fed as inputs to downstream models. Their rapid success has been in no small part due to the development of self-supervised learning (SSL) methods such as autoregressive [27] and masked [13] language modelling which have allowed models to learn contextual representations of input tokens without relying

Piotr Skalski, David Sutton, Stuart Burrell, Iker Perez, and Jason Wong. 2023.

Towards a Foundation Purchasing Model: Pretrained Generative Autore

ression on Transaction Sequences. In 4th ACM International Conference of

AI in Finance (ICAIF '23), November 27-29, 2023, Brooklyn, NY, USA. ACM,

While these methods have already been successfully used with different modalities such as natural language [4, 11, 22, 27, 28] computer vision [26, 30], audio [3, 12], and tabular data [1, 20, 31] there has been little work to adapt them to the case of multivariate

UAT 2022 ICDM 2023 **ICAIF** 2023



We will have open roles + internships

In 2024, we will hiring

- Research scientists
- Research engineers
- Interns

Register your interest @ <u>featurespace.com/innovation-roles</u>





Collaborative AI

Data Intensive Science CDT Seminar 22nd November 2023

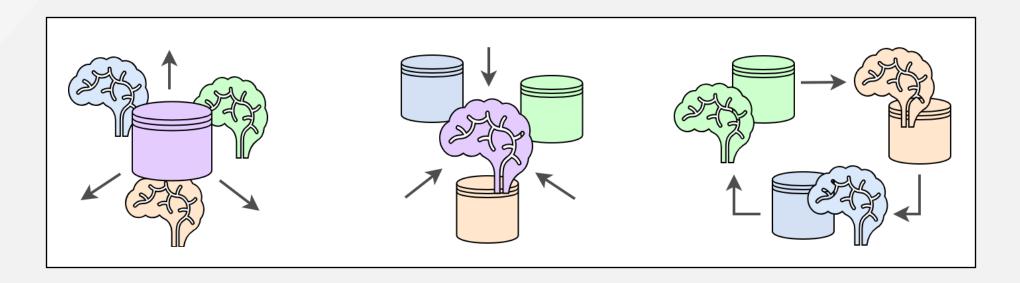




F E A T U R E S P A C E

Prevention systems are commonly serviced to financial institutions in isolation.

Few provisions exist for data sharing -> Fears of unintentional leaks and adversarial attacks.



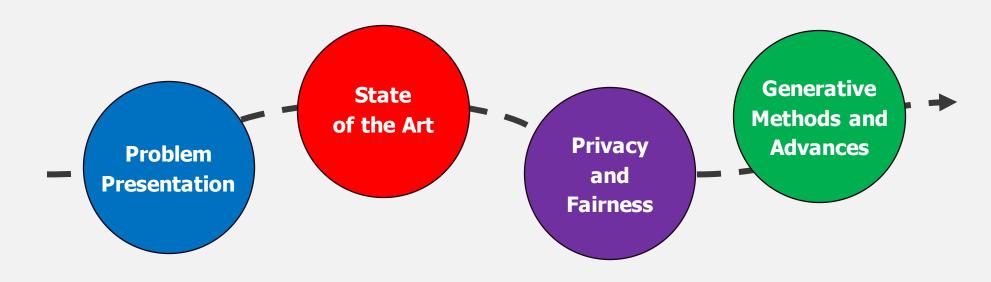
However, there exists evidence that financial data sharing can yield benefits.



Institutions produce incentives for collaborative solutions to financial crime.



Let us go through the following...





Problem Presentation

Crime detection systems monitor sequences of transaction data.

Commonly, formulated as a supervised learning task for binary classification.

A transaction x_t recorded at time t > 0 is endowed with a crime label $y \in \{0, 1\}$, and conditional responses are considered Bernoulli distributed, s.t.

$$y|\boldsymbol{x}_t \sim \text{Ber}(\mathbb{P}(y=1|\boldsymbol{X}_{\leq t}))$$

where $oldsymbol{X}_{< t}$ are transactions preceding and including $oldsymbol{x}_t$.

Unbalanced Labels

Originator to Beneficiary

Tabular Data

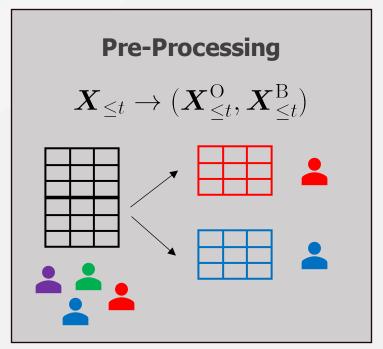
Throughput

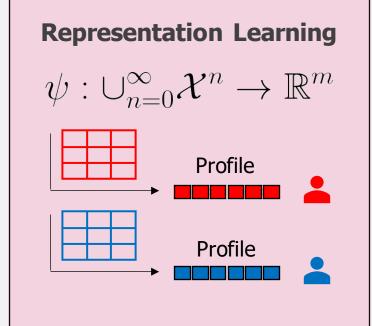


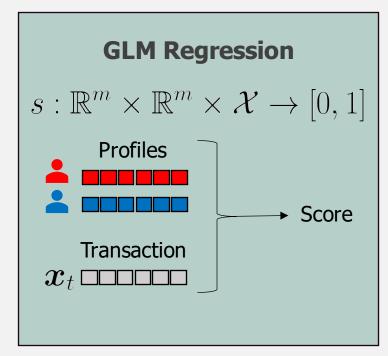
Problem Presentation

A function $f: \bigcup_{n=0}^{\infty} \mathcal{X}^n \to [0,1]$ estimates $\mathbb{P}(y=1|\boldsymbol{X}_{< t})$.

Decomposed...







Project Idea. We produce scores, observe actions and record causal effects. How to automate algorithms for optimal decision making?



Problem Presentation

Why Collaborate?

Reliable feature profiles can only be generated from complete transacting histories.

• Available only to their **managing** Financial Institutions.

Profiles for externally managed accounts are approximated.



State of the Art

Industry



Here is database:

- **Option A**: Contains **all** personal data for entities associated in a consortia.
- Option B: Contains minimal details for a few known criminals accounts.

Here is a **centralised** algorithm:

 Delegate full responsibility to a network or processor, no one is happy.

Academia



Sophisticated Federated algorithm:

- It just needs 25 GPUs to run.
- Must pass gradients around 10 million times, maximum latency of 0.001 milliseconds.
- Everyone ensure datasets conform to the same schema and are distributionally equivalent.
- To make sure it works, please let me centralise all data just this time.
- The **orchestrator** rules, period.

Concerns: Heterogeneity, Adaptability, Scalability, Efficiency and Regulatory Compliance.



Why is this gap not bridged?

Privacy, Fairness and Compliance.

Collaborative models **convoluted** and **leak information** when subjected to adversarial attacks.

Privacy enhancing technologies are researched in model designs:

Differential Privacy

Homomorphic Encryption

Secure Multi-Party Computation Zero Knowledge Proofs

Project Idea. How to design attacks against public model APIs? What about Federated Learning models?

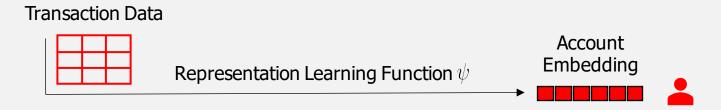


Generative Advances

Simplify: Do not train joint models, share the profiles.

Transaction Embeddings: Numerical representations of variable-length transaction data.

• A representation learning function $\psi: \cup_{n=0}^{\infty} \mathcal{X}^n \to \mathbb{R}^m$ is an embedding function.



Obfuscated profiles. **Is sharing safe?**

What do recipients need? How to make them comprehensive? How protect sensitive attributes?

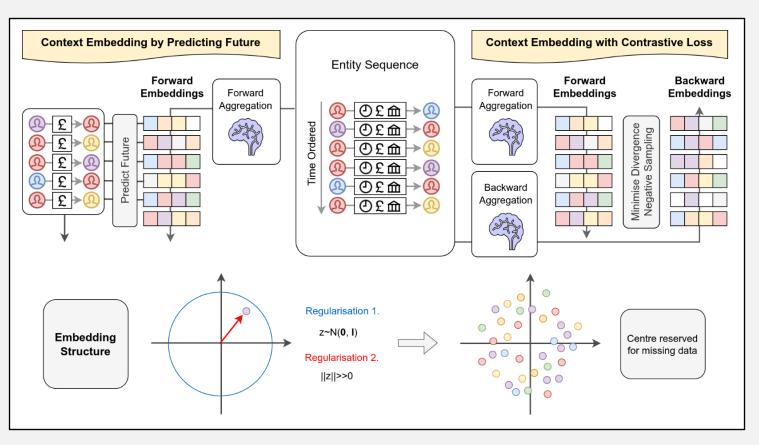


Generative Advances: Training

Collaborative Financial Crime Detection

- Language Approach? Next-event prediction, dual encoding...
- Vision Approach? Contrastive Learning, Adversarial Training, ...

Ideas...

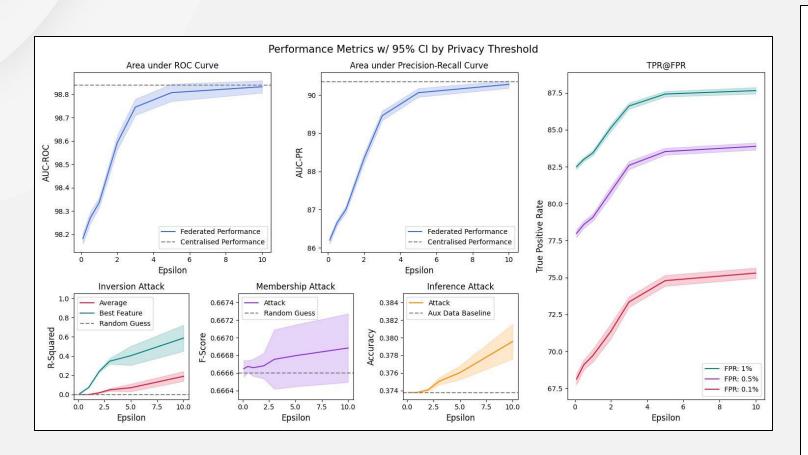


Project Idea. How design such a method properly? How accommodate multimodal outputs? How leverage embeddings for Generative data synthesis? Other purposes?



Generative Advances

Transfer Learning with **Differentially-Private** Generative Embeddings...



Algorithm 1: SGD step. Bank
$$\beta=1$$
. Peer-to-peer.

Input: Mini-batch $\{(\boldsymbol{x}_{i,t_i},\boldsymbol{z}_{i,t_i}^{\mathrm{o}},y_i)\}_{i=1,\dots,N}$.

Loss function \mathcal{L} and learning rate $\lambda>0$.

Slack term $\gamma\to0^+$.

Output: Updated s_1 and $g_{1,\beta},\,\beta=2,\dots,B$.

for $i=1,\dots,N$ do

$$\begin{array}{c} \rho_i\leftarrow getBeneficiary(\boldsymbol{x}_{i,t_i})\\ \hookrightarrow & \mathbf{Beneficiary}\ \mathbf{Bank}\\ \rho_i\\ \text{Publish}\ \boldsymbol{z}_{i,t_i}^{\mathrm{b}}=\mathcal{M}_{\rho_i}(\boldsymbol{X}_{\leq t_i}^{\mathrm{B}})\\ \text{Pre-process}\ \boldsymbol{r}_{i,t_i}^{\mathrm{b}}=g_{1,\rho_i}(\boldsymbol{z}_{i,t_i}^{\mathrm{b}})\\ \text{Predict}\ \hat{y}_i=s_1(\boldsymbol{z}_{i,t_i}^{\mathrm{o}},\boldsymbol{r}_{i,t_i}^{\mathrm{b}},\boldsymbol{x}_{i,t_i})\\ \mathbf{end}\\ \text{Update weights}\ \omega_s\ \text{of}\ s_1: \\ \omega_s\leftarrow\omega_s-\frac{\lambda}{N}\sum_{i=1}^N\nabla_{\omega_s}\mathcal{L}(y_i,\hat{y}_i)\\ \mathbf{for}\ \beta=2,\dots,B\ \mathbf{do}\ \text{update}\ \text{weights}\ \omega_g\ \text{of}\ g_{1,\beta}: \\ \omega_g\leftarrow\omega_g-\frac{\lambda}{N_\beta+\gamma}\sum_{i=1}^N\mathbb{I}_{\beta=\rho_i}\cdot\nabla_{\omega_g}\mathcal{L}(y_i,\hat{y}_i)\\ \text{where}\ N_\beta=\sum_{i=1}^N\mathbb{I}_{\beta=\rho_i}.\\ \mathbf{end}\\ \mathbf{end} \end{array}$$





Natural language interfaces

Data Intensive Science CDT Seminar 22nd November 2023





F E A T U R E S P A C E

Structure

- Motivation for natural language interfaces
- Discuss three open problems:
 - 1. Develop trust-worthy, fair, and robust agents for user support
 - 2. Generate natural language explanations of automated decision making
 - 3. Build a financial crime detective to automate complex investigations
- Q&A



Motivation

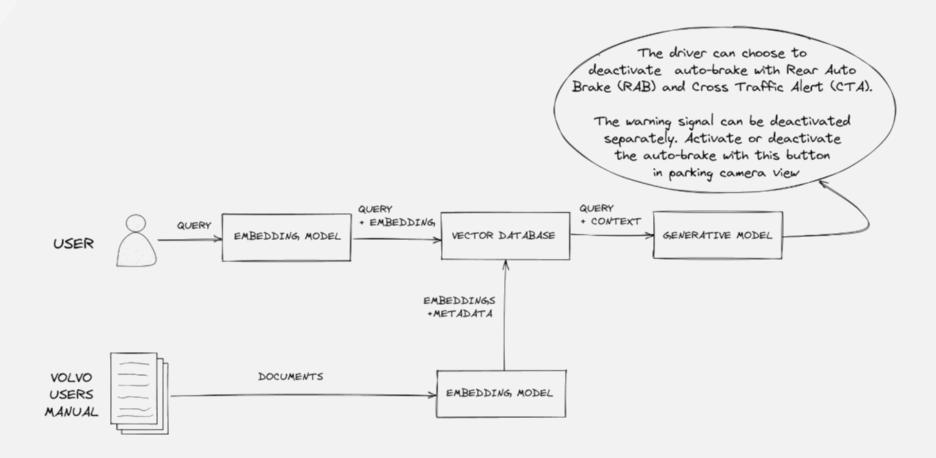


- Recent advances in LLM technology (GPT-4/Turbo, GPTs, LLama2) have the capacity to transform how users interact with software:
 - i. Can you do action X?
 - ii. How do I achieve Y?
 - iii. Why did you produce output Z?

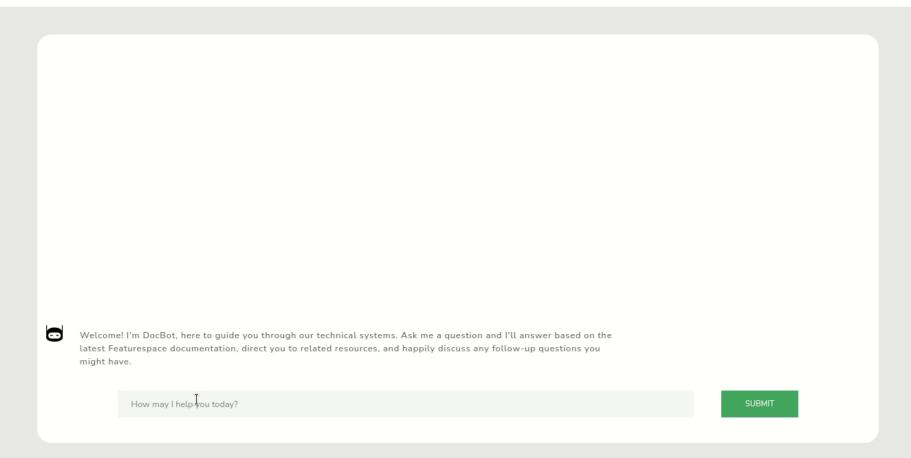


Problem 1. Agents for user support

- Typical approach uses Retrieval Augmented Generation – embedding relevant information within LLM prompts obtained via efficient similarity search over vector databases (image credit: Pinecone).







Powered by Llama 2

Knowledge bases: ARIC Documentation

Research projects



- 1. How do we **robustly** and **automatically** evaluate performance?
- 2. How best to incorporate **user feedback** and/or **AI feedback** for continuous fine-tuning?
- 3. Integrating diverse knowledge bases (Slack, technical documentation, ...) and construct efficient retrieval regimes (recursive retrieval, reranking, ...)
- 4. Building an coding-assistant for a **proprietary coding language** with only sparse datasets?
- 5. Engineering the above given latency, hardware and API constraints



Problem 2. Natural language augmented explainable AI

- Explainable AI (XAI) is a huge field trying to pry open the black box of neural networks and explain why they produce the outputs they do.
- Simple approach is to limit hypothesis space to inherently interpretable models: Bayesian Rule Lists,
 Sparse Linear Models, et al.
- Post-hoc explainability measures attempt to explain arbitrarily complex models after-the-fact.
 - i. **Feature importance**: which dimensions of the feature vector contributed most to the prediction? (Saliency, Lime, Integrated Gradients, Shap, DeepLift, and Perturbation Masks, ...)
 - ii. **Example importance**: which training examples contributed most to the prediction? (Influence Function, Deep K-Nearest, Neighbours, TraceIn, SimplEx, ...)
- Caveat these methods are far from perfect, but are continuously improving.



Research project

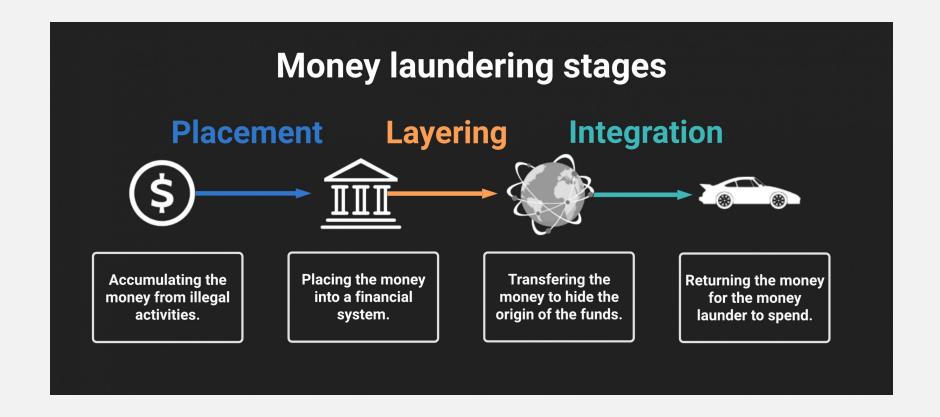
- Creating natural language model explanations from post-hoc XAI outputs
 - i. Can we use an LLM to reliably convert heatmaps of feature importance scores into natural language explanations?
 - ii. Can we use an LLM to explain a fraud prediction based on natural language summaries of relevant training examples?

Event time	Time since previous	POS entry mode	Merchant country	Amount	AVS result	CVV result	Merchant fraud rate	MCC fraud rate	Merchant country fraud rate	Score
5 2017-11-06 14:29:30	10 days 15:08:00	e-commerce	UK	67.54	no match	match	-6.00	-6.00	-7.00	0.95
4 2017-10-26 23:21:17	16 days 14:16:00	e-commerce	UK	60.43	no match	match	-6.00	-6.00	-7.00	0.96
3 2017-10-10 09:05:41	28 days 16:13:00	keyed	UK	375.94	full match	match	-8.00	-8.00	-7.00	0.04
2 2017-09-11 16:53:00	4 days 02:53:00	e-commerce	Luxembourg	34.26	full match		-12.00	-11.00	-11.00	0.16
2017-09-07 14:00:26	21 days 23:04:00	keyed	UK	25.47	full match	match	-8.00	-15.00	-18.00	0.18
0 2017-08-16 14:56:02	0 days 00:00:00	keyed	UK	50.14		match	-7.00	-13.00	-16.00	0.33



Problem 3. Build a financial crime detective to automate complex antimoney laundering investigations

- The capacity of governments and financial institutions to crack-down on crimes such as anti-money laundering (AML) is severely bottlenecked by manual labour.





Research project

- Can we fine-tune a LLM to **accurately** and **verifiably** assess and summarise suspicious behaviour in a transaction history suspected of AML?
- How can we evaluate the resulting LLM to assess whether it is reporting in a fair and unbiased way
 with respect to protected attributes? (Learning from human preferences, OpenAI 2017, Constitutional AI,
 Anthropic 2022).
- Can we build out a fully-fledged investigator agent; equipped with **tools** (API access, web-search) to aid an investigation?







Thanks for listening!

Feel free to add my on **LinkedIn** or email <u>stuart.burrell@feauturespace.co.uk</u> if you'd like to talk more.

References

- Bai, Yuntao, et al. "Constitutional AI: Harmlessness from AI feedback." arXiv preprint arXiv:2212.08073 (2022).
- Christiano, Paul F., et al. "**Deep reinforcement learning from human preferences.**" *Advances in neural information processing systems* 30 (2017).
- Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in neural information processing systems 30 (2017).
- Adadi, Amina and Mohammed Berrada. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)." IEEE Access 6 (2018): 52138-52160.
- LlamaIndex, Liu, J., https://github.com/run-llama/llama_index



Work with us!

F E A T U R E S P A C E

featurespace.com/innovation-roles

