
# O Uso de ML e IA no Processo de Investimento

*White Paper — Arctan Research*

## Sumário Executivo

> *A Arctan é uma família disciplinada e teoricamente fundamentada de estratégias quantitativas de renda variável e alocação de ativos que utiliza o aprendizado de máquina como um aproximador funcional flexível sobre um conjunto reduzido de variáveis economicamente motivadas, envolvido por um overlay de risco condicional ao regime — e não como um mecanismo de predição de caixa-preta.*

Na Arctan Research, identificamos três desafios centrais para a formulação e o uso de ferramentas de ML em finanças quantitativas:

- **Fatores em excesso:** Existe um conjunto muito amplo — e provavelmente dependente do regime — de fatores que determinam os preços e retornos dos ativos.
- **Dados insuficientes:** Dispomos de apenas uma "linha do tempo" para análise: a história econômica efetivamente ocorrida. Métodos como validação cruzada ou simulações de Monte Carlo têm utilidade limitada em finanças, pois não contamos com um modelo coerente do processo gerador dos dados.
- **Não-estacionariedade:** Mesmo que os fatores relevantes sejam identificados, as relações quantitativas entre fatores e retornos provavelmente sofrem tanto de mudanças lentas e graduais quanto de quebras mais abruptas e episódicas.

Esses desafios não podem ser eliminados — apenas gerenciados por meio de escolhas de design cuidadosamente calibradas. As respostas arquitetônicas da Arctan são três:

- **Features curadas:** Pré-processamos intensamente os sinais financeiros e econômicos com base em expertise de domínio e décadas de experiência empírica, oferecendo ao ML um espaço restrito e interpretável para aprender.
- **Condicionalidade ao regime:** Particionamos a linha do tempo em regimes econômicos, aplicamos ponderação de features e construção de carteira dependentes do regime, e utilizamos alavancagem variável para reduzir bruscamente o risco nas fronteiras perigosas de transição.

- **Monitoramento contínuo:** Realizamos uma bateria de testes estatísticos de qualidade do sinal a cada rebalanceamento, para detectar deterioração e acionar re-estimação ou redução de risco antes que os danos se acumulem.

## Vivemos uma Era Revolucionária

As possibilidades abertas pelo desenvolvimento acelerado da inteligência artificial já estão remodelando todos os campos da atividade humana. Mas ainda não dispomos de inteligência artificial geral (AGI). A tecnologia atual é simultaneamente muito eficaz em determinadas tarefas e muito limitada — quando não francamente prejudicial — em outras. Enquanto não tivermos AGI plena, qualquer aplicação de Machine Learning (ML) e IA precisa ser adaptada ao domínio específico em questão.

Este é um ponto crucial, pois a invenção de "chat bots" capazes de discutir qualquer assunto com aparente fluência pode sugerir que **a expertise de domínio** — a compreensão profunda que um especialista tem do seu campo — deixou de ter valor. Em geral, e para o processo de investimento em particular, isso é obviamente falso, razão pela qual não é possível "vibe codar" o caminho para se tornar um trader milionário, apesar das inúmeras afirmações nesse sentido no X ou no YouTube.

A natureza dos mercados financeiros ilustra esse ponto de forma contundente. O sucesso do ML em diversas áreas — e cabe lembrar que os Large Language Models são apenas uma instanciação das redes neurais, tecnologia cujas raízes conceituais remontam às décadas de 1940 e 1950[1] mas cujo poder atual deriva da arquitetura transformer introduzida em 2017[2] — surgiu da combinação de enormes aumentos na capacidade computacional com grandes volumes de dados. São necessários ambos: capacidade de processamento e dados.[3]

## Os Três Desafios do ML em Finanças

Mas o que são "os dados" nos mercados financeiros? Em primeiro lugar: *há simultaneamente dados demais e de menos.*

Há dados *demais* porque não existe um conjunto fixo de fatores que determine os preços dos ativos. Eles mudam o tempo todo e dependendo do momento. A teoria e os modelos empíricos atuais podem mapear o que *deveria* importar, mas isso não limita o que *de fato* importa. Nossos modelos de precificação de ativos de ponta podem ser expressos em termos de um kernel estocástico de precificação, mas na prática aquilo que determina esse fator é potencialmente ilimitado — a literatura acadêmica já catalogou mais de 300 sinais publicados de previsão de retornos.[4]

Há também *dados de menos.* Para pensar sistematicamente sobre os mercados, dispomos de apenas uma "linha do tempo" histórica — o que efetivamente ocorreu. A impossibilidade de realizar experimentos (do tipo: "se o Lehman Brothers tivesse sido socorrido pelo governo americano, teríamos tido uma crise em 2008?") restringe severamente nossa capacidade de mapear fatores em resultados. Não conseguimos gerar linhas do tempo sintéticas porque não dispomos de um modelo subjacente coerente do mundo real.

E na medida em que conseguimos identificar regularidades empíricas nos dados — como, por exemplo, que maior lucratividade tende a elevar o valor das ações — a relação quantitativa exata entre essas variáveis muda ao longo do tempo. Os mercados

financeiros são *não-estacionários* no sentido estatístico, em parte porque a economia subjacente evolui e lidamos com seres humanos altamente adaptativos.

Em resumo, temos um domínio em que:

- *Não é possível afirmar com certeza quais fatores determinarão os retornos futuros;*
- *Estamos severamente limitados pela impossibilidade de executar linhas do tempo alternativas;*
- *As relações entre variáveis se deslocam ao longo do tempo — não-estacionariedade pervasiva.*

*Dados demais, dados de menos…* Motivo para desespero? Certamente não. Cada desafio pode ser enfrentado com **a aplicação correta das ferramentas adequadas, orientada por escolhas de design feitas à luz da experiência de domínio.** É isso que a Arctan Research faz.

## Desafio 1: Quais São os Fatores Relevantes?

O ML trata fundamentalmente de aprender formas funcionais. Dado um conjunto de variáveis de entrada (features) e resultados, o modelo aprende o melhor mapeamento entre eles e consegue, então, prever fora da amostra o que deve ocorrer dado um conjunto específico de valores das features.

Havendo dados suficientes, o ML consegue identificar tanto quais variáveis importam quanto a natureza da relação entre elas e o resultado — mesmo na presença de ruído substancial. Mas em um domínio com restrição de dados, incerteza genuína sobre o que importa e apenas um conjunto experimental, o ML terá grande dificuldade para identificar features relevantes e encontrar formas funcionais confiáveis. É por isso que as tentativas de alimentar modelos de deep learning com dados financeiros brutos não têm produzido resultados consistentes em horizontes mensais e trimestrais.

> *A referência acadêmica canônica é Gu, Kelly & Xiu (2020),* cujo estudo de larga escala sobre métodos de ML em precificação de ativos constatou que o aprendizado profundo supera abordagens clássicas apenas quando aplicado a características elaboradas por especialistas — e não a dados brutos de preços. Esse resultado é precisamente o fundamento empírico da filosofia de design baseada em features curadas da Arctan.[5]

**Todo processo de investimento sistemático começa com a identificação de um sinal de investimento estatisticamente significativo.** Denominamos isso de "função analista".

Como isso se dá na prática? Primeiro, examinar os priors teóricos e empíricos sobre o que deveria importar. Segundo, pré-processar os dados de modo a gerar um sinal útil — o que pode exigir regularização intensa e validação cruzada cuidadosa que respeite a estrutura temporal dos dados. Ambos os processos requerem expertise de domínio profunda: décadas de experiência real nos mercados e conhecimento sólido da literatura econômica e financeira relevante.

Nossos modelos de renda variável definem um conjunto de features com aproximadamente 25 dimensões, cobrindo dados fundamentais de empresas, sensibilidade a choques macroeconômicos e características estatísticas dos retornos acionários. Cada feature é escolhida por uma razão e, quando necessário, pré-processada para separar sinal de ruído.

Mantemos intencionalmente nosso conjunto de features *pequeno e interpretável.* Acreditamos que o verdadeiro sinal para os retornos acionários reside em um espaço de baixa dimensionalidade, porém complexo. Esses sinais são então utilizados por *um aproximador funcional não-linear e flexível sobre as features.* No desafio de "escolher os dados e descobrir a forma funcional", optamos por selecionar os dados com base no conhecimento de domínio para que o ML possa encontrar a melhor forma funcional que mapeia features em retornos.

## Desafio 2: Superando a Limitação de Uma Única Linha do Tempo

Nos mercados financeiros, não é possível realizar experimentos. Nem mesmo simulações eficazes, pois não contamos com um modelo subjacente coerente do mundo real.

Mas se não podemos, sem acesso ao multiverso, gerar linhas do tempo alternativas, podemos tomar a linha do tempo existente e **definir diferentes regimes.** Nosso trabalho demonstra que os fatores relevantes, os dados importantes e a forma como esses dados se mapeiam em retornos futuros mudam dependendo do regime financeiro e macroeconômico. Nada funciona de forma incondicional.

Boa parte da literatura quantitativa trata a seleção de fatores como estacionária: por exemplo, empresas de "qualidade" com lucratividade consistente deveriam auferir um prêmio de risco positivo. O que constatamos é que isso NÃO é o caso: o prêmio de qualidade existe, mas apenas em determinados regimes.

Os modelos da Arctan **alternam qual direção latente do corte transversal acompanhar com base no regime.** A construção de carteira é baseada em regime — não em um mapeamento único e definitivo entre features e resultados.

Como exemplo, há uma distinção clara no mercado acionário brasileiro entre empresas de "qualidade" — geradoras de alto ROE estável, como os grandes bancos — e empresas de maior volatilidade, voltadas para commodities e consumo. A Arctan utiliza o espaço de features e os sinais de regime para inclinar suavemente a carteira em direção às empresas com maior retorno esperado ajustado ao risco, dado o seu entendimento do regime vigente.

A determinação dos regimes relevantes também é amplamente orientada pela expertise de domínio. Nosso trabalho demonstra que **cada mercado acionário tem uma estrutura de regime muito específica**, baseada em sua estrutura econômica, institucional e política e em como se insere na economia global mais ampla. Não existe uma variável "simples" que determine efetivamente qual é o regime relevante.

Na Arctan, a especificação do regime não molda apenas a construção da carteira, mas também a exposição bruta. Todos os nossos modelos de renda variável incorporam **uma feature de alavancagem variável que gerencia a exposição bruta em função do regime e do risco de transições de regime.** Ao determinar em tempo real o regime corrente e monitorar o risco de transição — especialmente transições para regimes mais arriscados e negativos —, a Arctan gera o que chamamos de **"alpha protetivo":** desempenho cross-seccional mais consistente, composto por alavancagem variável que reduz bruscamente o risco quando necessário. Os modelos apresentam suas características de desempenho mais fortes nos extremos do "barbell" de regimes — quando identificam booms confirmados ou crises em desenvolvimento.

## Desafio 3: Gerenciando a Não-Estacionariedade Pervasiva

Os dados financeiros e econômicos são ao mesmo tempo muito ruidosos e instáveis: sua relação sinal-ruído é baixa. Encontre um fator que aparentemente funciona e, como a literatura tem demonstrado, essa "descoberta" frequentemente levará aquele fator a deixar de funcionar — seja por arbitragem de investidores que leram o mesmo artigo, seja por mudanças estruturais na economia.

O desafio está em implementar tanto **procedimentos diagnósticos quanto prescritivos** para determinar se as relações mensuradas ainda se sustentam.

*Do ponto de vista diagnóstico,* é preciso examinar os dados e exercer julgamento sobre o comprimento da amostra e outras formas de regularização. É justamente nesses julgamentos que o conhecimento de domínio e a experiência especializada são mais valiosos.

*Do ponto de vista prescritivo,* é necessário aplicar uma bateria de testes estatísticos para medir continuamente a qualidade do sinal. Embora a natureza de curto prazo das amostras financeiras frequentemente limite o poder desses testes, quando combinados com o conhecimento de domínio, uma probabilidade crescente de quebra estrutural pode ser gerenciada por meio da re-estimação das relações estatísticas relevantes ou da redução do nível geral de risco assumido.

Um ponto importante merece ênfase: embora cada modelo da Arctan opere sem intervenções discricionárias no nível das decisões — os parâmetros são definidos durante a especificação do modelo e mantidos fixos entre revisões programadas —, a necessidade de monitorar e avaliar proativamente o desempenho do modelo em tempo real significa que os modelos nunca operam verdadeiramente sem supervisão humana. **O monitoramento ativo e contínuo é parte central da arquitetura, não um complemento.**

## A Arquitetura da Arctan: Uma Síntese Neurossimbólica

Uma geração anterior de modelos de IA — a era da *IA simbólica* ou sistemas especialistas, que floresceu dos anos 1960 até os 1980 com sistemas marcantes como MYCIN e DENDRAL[6] — utilizava uma combinação de lógica simbólica e opinião de especialistas para codificar regras de decisão em domínios específicos. Essa abordagem, apesar de promissora inicialmente, foi em grande parte abandonada quando os pesquisadores constataram que muito do julgamento especializado não poderia, de fato, ser conscientemente articulado e codificado. O campo entrou em um inverno da IA, seguido pela ascensão do machine learning estatístico (SVMs, florestas aleatórias, boosting) ao longo dos anos 1990 e 2000, e então pela revolução do aprendizado profundo, convencionalmente datada de 2012.

A ascensão do ML, e especialmente do aprendizado profundo (*deep learning*), foi na direção oposta: deixar o modelo aprender tudo a partir dos dados, com estrutura prévia mínima.

Uma forma de entender a Arctan é como uma junção dessas duas abordagens — o que a literatura contemporânea de pesquisa em IA denomina *IA neurossimbólica:*[7]

conhecimento de domínio especializado e codificado orientando escolhas de design onde agrega valor, combinado com aprendizado dinâmico a partir dos dados tanto para extração de sinais quanto para gestão de risco. Enquanto não tivermos AGI plena, essa combinação — estrutura simbólica onde a temos, aprendizado estatístico onde precisamos — é o caminho mais robusto a seguir.

## Conclusões

A aplicação de IA e ML a finanças abre muitas oportunidades. O ciclo de pesquisa é vastamente acelerado: ir da tese ao código e ao teste — algo que antes demandava semanas ou meses — pode agora ser feito em uma fração do tempo. Os LLMs podem ser utilizados de forma eficaz para comunicação com investidores prospectivos e existentes. Os ganhos em eficiência e redução de custos são expressivos, especialmente para organizações que já nascem centradas em IA, sem precisar adaptar processos legados.

A Arctan Research é 100% centrada em IA. Nada é feito sem ferramentas e agentes de IA.

Mas a IA e o ML precisam ser adaptados aos problemas específicos em questão. Os mercados financeiros são não apenas complexos, mas altamente mutáveis. Não basta encontrar estratégias lucrátivas — elas precisam *ser robustas.* Quando houver trade-off entre lucratividade e robustez, opte sempre pela robustez. Apenas a robustez maximiza a probabilidade de que o processo de investimento codificado consiga sobreviver às frequentes pequenas mudanças no ambiente e se adaptar quando mudanças de grande magnitude ocorrerem.

Este white paper apresentou a filosofia de design subjacente à família de produtos Arctan — as estratégias de renda variável Arctan Long-Only e Market-Neutral — demonstrando como cada escolha arquitetônica é uma resposta direta a um dos três desafios fundamentais do ML em finanças: fatores em excesso (resposta: features curadas), apenas uma linha do tempo (resposta: condicionalidade ao regime) e não-estacionariedade pervasiva (resposta: monitoramento contínuo e re-estimação adaptativa).

**O que a Arctan não está afirmando**

Não afirmamos ter AGI. Não afirmamos ter resolvido o problema da não-estacionariedade. Não afirmamos extrair alpha de fatores inéditos, desconhecidos pela literatura acadêmica. Nossa afirmação é mais modesta e mais durável: um framework disciplinado e sistematicamente monitorado que utiliza o ML da forma como ele efetivamente funciona em ambientes com escassez de dados e não-estacionários — orientado pela expertise de domínio, delimitado pela estrutura de regime e monitorado continuamente para identificar sinais de deterioração.

### Referências

Cochrane, J.H. (2011). Discount Rates. Journal of Finance, 66(4), 1047–1108.

Garcez, A. d'A. & Lamb, L.C. (2020). Neurosymbolic AI: The 3rd Wave. arXiv:2012.05876.

Gu, S., Kelly, B. & Xiu, D. (2020). Empirical Asset Pricing via Machine Learning. Review of Financial Studies, 33(5), 2223–2273.

Harvey, C.R., Liu, Y. & Zhu, H. (2016). …and the Cross-Section of Expected Returns. Review of Financial Studies, 29(1), 5–68.

Hoffmann, J. et al. (2022). Training Compute-Optimal Large Language Models. arXiv:2203.15556.

Hou, K., Xue, C. & Zhang, L. (2020). Replicating Anomalies. Review of Financial Studies, 33(5), 2019–2133.

Kaplan, J. et al. (2020). Scaling Laws for Neural Language Models. arXiv:2001.08361.

Krizhevsky, A., Sutskever, I. & Hinton, G.E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. NeurIPS 2012.

McCulloch, W.S. & Pitts, W. (1943). A Logical Calculus of the Ideas Immanent in Nervous Activity. Bulletin of Mathematical Biophysics, 5, 115–133.

Rosenblatt, F. (1957). The Perceptron: A Perceiving and Recognizing Automaton. Cornell Aeronautical Laboratory Report.

Sutton, R. (2019). The Bitter Lesson. http://www.incompleteideas.net/IncIdeas/BitterLesson.html

Vaswani, A. et al. (2017). Attention Is All You Need. NeurIPS 2017. arXiv:1706.03762.