

Desafio Equifax

Desvendando o mundo do crédito

Sobre a Equifax

A Equifax é um dos maiores bureaus de crédito e empresas de inteligência de dados do mundo. Com uma operação sólida e abrangente, a companhia atua em mais de 24 países e registra um faturamento anual global superior a US\$6 bilhões. No Brasil, a marca consolidou sua presença recentemente ao adquirir a Boa Vista, unindo sua infraestrutura global e capacidade de investimento ao profundo conhecimento da dinâmica financeira do mercado brasileiro.

Para muito além da atuação tradicional de um bureau, a Equifax funciona hoje como uma verdadeira potência tecnológica mundial. O que move o nosso negócio é a capacidade de transformar petabytes de informações brutas em decisões inteligentes e seguras. Fazemos isso colocando a Inteligência Artificial, Machine Learning e estatística avançada no centro da nossa estratégia. A Equifax possui mais de 300 patentes sobre IA no mundo e desenvolve só no Brasil mais de 150 modelos de machine learning por ano.

O objetivo da Equifax é ajudar as pessoas a viverem sua melhor vida financeira e fazemos isso através de soluções inteligentes que tornam todo o ciclo do crédito mais eficiente. Com mais de 150 mil clientes no Brasil, as soluções da Equifax tornam o crédito mais justo e eficiente, permitindo que as pessoas realizem seus sonhos de forma sustentável.

Para quem estuda exatas, dados e tecnologia, a Equifax representa um dos ecossistemas mais avançados do mercado para se construir uma carreira em Dados e Analytics, oferecendo um ambiente de alta complexidade analítica e de amplo aprendizado:

- **Pioneirismo Global em IA:** Não apenas consumimos tecnologia; nós a criamos. A Equifax desenvolveu e patenteou mais de 300 soluções de IA ao redor do mundo. Uma de nossas diretrizes estratégicas é acelerar a utilização de IA de forma eficiente e, aqui no Brasil, estamos assumindo um papel de referência, desenvolvendo soluções que servem de exemplo para todas as operações da Equifax no mundo, trabalhando com o que há de mais moderno.
- **Poder Computacional e Cloud-Native:** Nossa operação inteira roda na Equifax Cloud, uma infraestrutura bilionária projetada para processar grande volume de dados, sem limites de gargalos para rodar hiperparâmetros ou arquiteturas complexas de modelagem. Isso permite alavancar nossas soluções e produtos, sem gargalos de tecnologia.
- **O Motor do Sistema Financeiro:** Apenas na operação brasileira, nosso time de Data Science desenvolve e gerencia mais de 150 modelos estatísticos por ano — atuando em frentes que vão de risco e propensão à cobrança e fraude. Nossos códigos são o motor que define as estratégias de concessão de crédito dos maiores bancos, varejistas e fintechs do país.

- **Equipe de ponta:** Nossas equipes são formadas por profissionais referências no mercado e trabalhamos com uma cultura de colaboração mútua, proporcionando um ambiente fértil de aprendizado.

Na Equifax, buscamos talentos inquietos, que queiram estar na fronteira entre o rigor acadêmico e a velocidade do mercado. Aqui, você não apenas aperta botões; você investiga, desenha novas soluções e usa a matemática para criar impacto real na sociedade e na economia.

O Datathon é a sua chance de vivenciar a nossa rotina, testar o seu raciocínio contra dados reais e provar que você está pronto para construir o futuro da inteligência de crédito. **Bem-vindos ao desafio!**



Desafio Equifax: desvendando o mundo do crédito

Imagine que você é um Cientista de Dados recém-contratado pela Equifax. Sua primeira grande missão é construir um modelo de concessão de crédito para um grande cliente da companhia. Para isso, disponibilizaremos uma base de dados anonimizada contendo o histórico de milhares de concessões de crédito.

Esta base possui diversas variáveis preditoras (as características no momento da concessão), a "safra" (o mês/ano em que o crédito foi dado) e a variável target (se o cliente pagou ou não a dívida). O seu objetivo é prever o futuro da forma mais precisa possível: usar o passado para ensinar um modelo de Machine Learning a identificar os bons e os maus pagadores.

O desafio é dividido em **três entregas incrementais, onde cada uma delas deverá gerar pelo menos um modelo de concessão de crédito diferente:**

1. **O Modelo Base (Apenas Dados Disponibilizados):** Você deverá desenvolver o melhor modelo preditivo possível de *Credit Scoring* utilizando estritamente as colunas que fornecemos. O grande teste aqui será a validação: nós avaliaremos o seu modelo em uma base **OOT (Out-of-Time)**, ou seja, uma base de dados de um período futuro em que você não terá a resposta (*target*).
2. **O Modelo Tunado (Feature Engineering):** Utilizando as informações fornecidas, como você pode melhorar a performance do seu modelo utilizando Feature Engineering? Utilize agregações, transformações, ou o que considerar necessário para trazer ganho de performance, interpretabilidade ou estabilidade ao seu modelo. Exemplo: Sabendo a idade e região de um cliente, é possível inferir se a renda de um cliente é superior/inferior à renda média de sua região.
3. **O Modelo Macro (Dados Externos e Públicos):** O mundo real não acontece no vácuo; a economia afeta o bolso das pessoas. Nesta etapa, você deverá enriquecer a base original com dados macroeconômicos reais extraídos de fontes públicas (ex: inflação, juros, desemprego, selic) e verificar se a adição dessas informações ao modelo consegue melhorar sua performance.

Atenção aos detalhes do mundo real: Não basta apenas "jogar" os dados no algoritmo. Queremos ver uma **Análise Exploratória e Descritiva** de altíssima qualidade. Além disso, para a Entrega 3, você deverá agir como um verdadeiro Engenheiro de *Machine Learning*: terá que explicar a fonte do dado macroeconômico, a metodologia usada para cruzar isso com o indivíduo e, principalmente, como esse modelo rodaria em produção. Lembre-se de que dados do governo têm atraso (*delay*) de divulgação. Como o seu modelo em produção lidaria com uma variável que só é atualizada a cada 30 ou 60 dias?

Além da performance técnica dos modelos, a capacidade de comunicar os achados é fundamental. Portanto, **você deverá preparar uma apresentação executiva que conecte as decisões de modelagem aos resultados de negócio**, explicando não apenas o "como", mas o "porquê" das estratégias adotadas.

Conceitos Importantes

Para navegar com sucesso neste Datathon, você precisa dominar o dialeto do mundo do crédito. Aqui estão os conceitos fundamentais que guiarão o seu trabalho:

- **Credit Scoring (Pontuação de Crédito):** É o modelo estatístico que avalia o risco de um cliente. Ele transforma o perfil e o histórico de uma pessoa em uma probabilidade de inadimplência (ou em uma pontuação). Quanto maior o risco, menor o score.
- **Adimplência vs. Inadimplência (Good vs. Bad):** No jargão de risco, o cliente adimplente ("Good" ou "Bom") é aquele que honra seus pagamentos no prazo. O cliente inadimplente ("Bad" ou "Mau") é aquele que atrasa ou não paga. Essa será a sua classe a ser prevista pelo modelo.
- **Target (A Variável Resposta):** É a definição exata do que consideramos ser um "Mau" pagador para o algoritmo. No mercado, usamos nomenclaturas específicas para definir essa janela de observação e severidade.
 - Exemplo clássico - m6ever60: Significa que, observando os primeiros 6 meses de vida do empréstimo (m6), o cliente atingiu pelo menos uma vez a marca de 60 dias de atraso (ever 60).
 - **Dica:** como o modelo de crédito precisa ser se comportar de forma decrescente de acordo com o risco (quanto maior o score, menor o risco), o evento de interesse deve ser a adimplência e não a inadimplência.
- **Safra (Vintage):** Refere-se à data (geralmente mês e ano) em que o crédito foi concedido ou a conta foi aberta. A análise por safra é crucial porque o comportamento econômico muda com o tempo. A análise por safras é essencial para avaliar a evolução do comportamento da carteira de clientes e a estabilidade dos modelos ao longo do tempo. Além disso, respeitar o efeito temporal (utilizar apenas informações dos indivíduos anteriormente a data de concessão) é essencial para não utilizar nenhuma informação futura.
- **Amostra OOT (Out-of-Time):** A base OOT (Out-of-Time) é a porção de dados com safras posteriores às da Base de Treino e Teste. Normalmente ela é utilizada para avaliar se a performance e a distribuição do modelo são estáveis em datas posteriores às de desenvolvimento dos modelos.. O objetivo é garantir que o seu modelo não "decorou" o passado (overfitting), mas aprendeu os padrões reais e consegue prever meses futuros.
 - **Distribuição de Scores:** Visualize a distribuição dos escores do modelo para diferentes safras da base OOT. Se a forma da distribuição mudar drasticamente, pode indicar instabilidade.
 - **Monitoramento de KS1 e KS2:** No contexto da estabilidade, ao invés de apenas um valor de KS, é útil monitorar o KS em diferentes safras (o que você poderia chamar de KS por safra, ou KS1 para uma safra e KS2 para outra). A ideia é observar se a capacidade de separação do modelo se mantém consistente. Se o KS variar significativamente entre safras, pode indicar problemas de estabilidade.
- **Feature Engineering (Engenharia de Variáveis):** É a arte de usar o conhecimento de negócios para criar novas variáveis preditoras a partir de dados brutos. Frequentemente, a combinação de duas variáveis simples gera uma nova variável com poder preditivo importante..
- **Dados Macroeconômicos (e seu Delay):** Variáveis que descrevem a economia em geral (Taxa Selic, IPCA, PNAD). O grande desafio no crédito é o "delay" (atraso de divulgação). A inflação de março, por exemplo, só é divulgada em meados de abril. Um modelo em produção em 01 de abril não terá a inflação de março disponível. Você precisa arquitetar a solução considerando esse atraso.

- **Métricas de Performance (Gini, KS, AUC):** São métricas padrão para avaliar modelos de classificação de risco de crédito. Elas medem a capacidade do modelo de discriminar entre "bons" e "maus" pagadores. Valores mais altos indicam um modelo com maior poder de separação.
 - **Base de Treino e Teste:** A **base de treino** é a porção dos dados usada para ensinar e ajustar os parâmetros do modelo. A **base de teste** é uma amostra separada, usada para avaliar a performance do modelo em dados não vistos.
-

Ferramentas

O desafio poderá ser feito em Python (via scripts e Jupyter Notebooks) ou R (scripts e RMarkdown).

Bibliotecas

Abaixo, há a sugestão de pacotes que você pode utilizar. Fique a vontade para usar pacotes adicionais, além de construir seu próprio código.

Etapa / Categoria	Python	R
Manipulação de Dados	pandas, numpy	dplyr, data.table
Visualização	matplotlib, plotly	ggplot2, plotly
Algoritmos de Boosting	lightgbm, xgboost, catboost, sklearn	lightgbm, xgboost, catboost
Algoritmos de Redes Neurais	tensorflow, pytorch	torch, keras
Pipeline e Validação	scikit-learn	tidymodels (ou caret)
Métricas de Crédito (KS/AUC)	scipy, scikit-learn	pROC, ROCR
Interpretabilidade	shap	treeshap, DALEX
Entrega / Notebook	Jupyter Notebook (.ipynb)	RMarkdown (.Rmd)

Conformidade dos Entregáveis

Para a conclusão bem-sucedida do desafio, os participantes deverão realizar as seguintes entregas obrigatórias, que serão avaliadas quanto à completude e conformidade:

A. Código-Fonte Completo e Documentado:

- Jupyter Notebook(s) ou scripts (.py / .Rmd) limpos, organizados e comentados.
- O código deve cobrir todo o pipeline: Limpeza/EDA, Engenharia de Features, Integração Macro, Treinamento dos 3 Modelos, Avaliação de Métricas (KS, AUC, Gini) e Explicabilidade (ex: SHAP).
- Seção de Transparência: Caso tenha utilizado IA, deixe um comentário nas funções e etapas onde isso ocorreu. O uso de IA não é proibido, mas é importante ter transparência quanto a isso.

B. Base OOT Escorada:

- O arquivo `base_oot.csv` preenchido com a coluna contendo o score predito pelo seu modelo vencedor escolhido.

C. Relatório Técnico do Desenvolvimento:

- Um relatório técnico detalhando o desenvolvimento da sua solução. O documento deve apresentar a metodologia adotada do início ao fim - cobrindo desde a Análise Exploratória de Dados, passando por Feature Engineering e Feature Selection, até a escolha do modelo final. Para cada etapa, descreva com clareza os critérios utilizados, o seu racional de tomada de decisão e os resultados obtidos. Não se esqueça de descrever como você lidaria com os indicadores macroeconômicos em produção.

D. Apresentação Executiva (Pitch de Negócios):

- Deck de slides (PDF ou PPTX) traduzindo os achados técnicos em linguagem de negócios.
 - A apresentação deve explicar as decisões de modelagem, a justificativa da escolha do modelo vencedor, o impacto de negócio esperado e o "porquê" de cada estratégia adotada.
-

Critérios de Avaliação

O seu projeto não será avaliado apenas pela métrica de acerto final do algoritmo. Na Equifax, valorizamos a visão ponta a ponta do Cientista de Dados: desde a limpeza do dado até a viabilidade de implantar a solução no mundo real.

Sua equipe será avaliada nos seguintes pilares:

1. Conformidade dos Entregáveis

- Entrega A: Código-Fonte Completo e Documentado.
- Entrega B: Base de Dados Escorada (OOT).
- Entrega C: Relatório Técnico.
- Entrega D: Apresentação Executiva.

Estes itens devem ser avaliados quanto à completude, clareza e aderência aos requisitos técnicos estabelecidos.

2. Qualidade da Análise Descritiva e Exploratória (EDA)

*Esperamos encontrar essas informações no relatório técnico, e no notebook/Rmarkdown se achar necessário.

- Profundidade no entendimento da base de dados;
- Identificação, tratamento e justificativa para possíveis tratamentos de dados;
- Geração de *insights* visuais e de negócios antes da modelagem;
- Análise da inadimplência e suas possíveis análises segmentadas e regionalizadas.
- Metodologias de seleção de variáveis e redução de dimensionalidade

3. Poder de Discriminação do Modelo (Performance Estatística)

- O desempenho preditivo dos modelos desenvolvidos, com peso especial para o resultado na nossa base **OOT cega** (na qual vocês não possuem o *target*).
- Serão avaliadas métricas clássicas de crédito, como **KS (Kolmogorov-Smirnov)**, **AUC (ROC)** e estabilidade para garantir que não houve *overfitting*.

4. Criatividade e Lógica no *Feature Engineering* (Entrega 2)

*Esperamos encontrar essas informações no relatório técnico.

- Quais variáveis novas foram criadas?
- A lógica de negócios por trás da criação dessas variáveis faz sentido? Elas trouxeram ganho real de performance em relação ao Modelo Base (Entrega 1)?

5. Metodologia de Uso dos Dados Macroeconômicos (Entrega 3)

*Esperamos encontrar essas informações no relatório técnico.

- Qualidade, relevância e rastreabilidade das fontes públicas escolhidas.
- A metodologia matemática e lógica utilizada para associar o dado macro (que é global/regional) à linha do indivíduo (a concessão de crédito específica).

6. Modelo Vencedor e Interpretabilidade

*Esperamos encontrar essas informações no relatório técnico.

- Os critérios para escolha do modelo vencedor foram adequados?
- Identifique as variáveis mais importantes do modelo e mostre o como e o quanto elas afetam as previsões do modelo vencedor.

7. Visão de Negócios e Produtização

*Esperamos encontrar essas informações no relatório técnico.

- Quão bem a equipe explicou como o modelo da Entrega 3 seria implantado em produção?
- A equipe levou em consideração problemas do mundo real, como a periodicidade de captura dos dados externos, os *delays* de divulgação (latência da informação) e a viabilidade sistêmica do modelo?

8. Organização, Código e Documentação

- Clareza do raciocínio, documentação do código (Jupyter Notebook / scripts limpos e bem comentados) e cumprimento exato de todos os itens solicitados no regulamento.
 - Apresentação comercial: clareza na exposição dos resultados, capacidade de traduzir termos técnicos para o público de negócios e qualidade visual do material.
 - Identificação das etapas e códigos que tiveram o auxílio e/ou foram gerados por IA.
-