Universidade de São Paulo (USP)
Faculdade de Economia, Administração, Contabilidade e Atuária (FEA)
Departamento de Administração

EAD0830 - IA e ML Aplicados a Finanças

Atividade Computacional 2

Objetivo: desenvolver uma competição voltada à construção de modelos de classificação, no estilo de um *case* de análise de dados. Cada equipe deverá selecionar um modelo de classificação, que pode ser um dos abordados em aula (como regressão logística ou k-NN) ou qualquer outro modelo apropriado. Além disso, todos os grupos deverão aplicar também o modelo XGBoost, que é amplamente considerado um *benchmark* em tarefas de classificação por apresentar excelentes resultados em diferentes contextos. Essa comparação permitirá que os grupos identifiquem, eventualmente, um modelo superior ao XGBoost, ao mesmo tempo em que desenvolvem competências práticas na aplicação dessa técnica – um conhecimento essencial para o mercado.

Orientações:

- a base de dados contém informações de transações financeiras. Dentre as transações, algumas delas são fraudes ($Fraude = 1$), enquanto as demais são transações lícitas ($Fraude = 0$);
- os dados são desbalanceados, de forma que a classe "Fraude" corresponde a uma parcela pequena da amostra – muito comum em aplicações reais;
- os atributos (8 no total) são numéricos, sendo apenas 1 deles categórico;
- a variável resposta é o atributo *Fraude*, que assume valor 1 em caso de transação fraudulenta, e 0 caso contrário (transação lícita);
- são disponibilizadas duas amostras. A amostra treino apresenta a classe dos objetos, que deverá ser usada para construção dos classificadores. Essa amostra deve ser dividida em sub-amostras (treinamento e teste) para escolher o melhor modelo da equipe – estimação e validação. Em seguida, o modelo selecionado deve ser aplicado na amostra teste dispobilizada no moodle. Para esses dados, a classe real não é fornecida. Vocês devem aplicar o classificador para essa amostra e entregar as classes previstas para avaliação;
- o objetivo de cada grupo/equipe é construir o melhor classificador para tais dados e fornecer as classes previstas na amostra teste. Como a classe é desbalanceada, sugere-se considerar não apenas a matriz de confusão, mas também o indicador área sob a curva (*area under the curve* – AUC), medida que mensura a acuidade de classificadores (inúmeras bibliotecas de classificação calculam essa medida automaticamente).

Relatório Empírico: redigir um texto contendo a descrição do classificador escolhido (qual método, suas vantagens e desvantagens); as justificativas de todas as decisões metodológicas (divisão da amostra, estrutura selecionada e atributos considerados); e os resultados (matriz de confusão e AUC para os dados da planilha treino, nas subdivisões definidas pelas equipe para gerar o modelo selecionado) e sua comparação com o XGBoost (seu modelo foi melhor que o XGBoost?[1]).

[1] Bibliotecas que implementam o XGBoost em R (`https://xgboost.readthedocs.io/en/stable/R-package/xgboost_introduction.html`) e em Python (`https://xgboost.readthedocs.io/en/stable/python/python_intro.htm`).

Material e Avaliação: será avaliada a adequação das decisões metodológicas e a discussão apropriada dos resultados. O texto deverá ser entregue em arquivo com extensão .pdf em fonte Times New Roman, tamanho 12, texto justificado, e espaçamento simples. Dividir o texto em duas seções: i) metodologia; ii) resultados e discussão. Limite máximo de 3 páginas. Entregar, em conjunto, o código elaborado para obter os resultados (na linguagem que preferir), e uma **planilha com as classes previstas para os dados da amostra de teste** (conforme exemplo disponibilizado no moodle).

Data de entrega: **até às 23h59 de 26 de maio de 2025** - via moodle. Não são aceitas entregas fora do prazo.

Instruções finais: os grupos devem ser compostos por 6 a 8 integrantes - sem exceções.