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 para a construção de melhores modelos para classificação (estilo *case* de análise de dados). Cada equipe deverá selecionar um modelo de classificação dentre os avaliados em aula (regressão logística ou k-NN), assim como considerar qualquer método de sua preferência.

Orientações:

- a base de dados contém informações de transações financeiras com a utilização de cartões de crédito. Dentre as transações algumas delas são fraudes ($class = 1$), enquanto as demais são transações lícitas ($class = 0$);
- os dados são desbalanceados, de forma que a classe "fraude" corresponde a uma parcela pequena da amostra ($\approx 0,21\%$) - muito comum em aplicações reais;
- os atributos (21 no total) são numéricos e as interpretações não são divulgadas por questões de confidencialidade dos dados - nosso objetivo nesta atividade é desenvolver melhores classificadores, e não entender as relações entre as variáveis;
- a variável resposta é o atributo *class*, 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 suas respectivas discussões.

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 5 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 21 de Junho de 2023** - via moodle.

Instruções finais: os discentes deverão desenvolver o trabalho em grupos de até 8 integrantes (não há distinção entre as turmas do diurno e noturno). A entrega comporá parte da avaliação na disciplina (1/3 da média final). Os integrantes do grupo vencedor da competição terão um acréscimo de 1 ponto na média final.