SCC0277 - Competições de Ciências de Dados
Segundo Projeto Prático

Neste projeto vamos nos aprofundar no desenvolvimento do problema **IEEE-CIS Fraud Detection** do Kaggle. Utilize a biblioteca scikit-learn para criar os modelos pedidos. Você pode usar outras bibliotecas para fazer as demais análises =)

Todas as respostas devem ser justificadas com base em:

1. **Código Python mostrando a(s) análise(s) e/ou o(s) modelo(s) feitos;**
2. **O resultado da(s) análise(s) e/ou do(s) modelo(s) e**
3. **Uma explicação textual (pode ser breve) da conclusão obtida.**

Em caso de plágio (mesmo que parcial) o trabalho de todos os alunos envolvidos receberá nota ZERO.

Desorganização excessiva do código resultará em redução da nota do projeto. Exemplos:

- **Códigos que devem ser rodados de forma não sequencial;**
- **Projeto entregue em vários arquivos sem um README;**
- **...**

Bom projeto,
Tiago.

Questões

Embora as questões a seguir estejam numeradas - caso ache necessário - você pode desenvolver os código em outra ordem. Só se lembre de deixar clara a resposta de cada uma das questões.

Questão 1 (valor 2,0 pontos)

a) Compare as métricas acurácia e AUC para essa base de dados simulando os seguintes modelos:
   - Modelo que classifica aleatoriamente (50% chance de dizer que é fraude e 50% de dizer que não é);
   - Modelo que classifica todos os casos como fraude;
   - Modelo que classifica todos os casos como não fraude.

b) Discuta os resultados.

Questão 2 (valor 2,0 pontos)

a) Proponha 3 novas variáveis explicativas que façam sentido do ponto de vista prático do problema;
   - Pelo menos duas delas devem ser criadas usando as informações dos arquivos sobre os clientes (arquivos com nome **`identity`**);
   - Dica: leia sobre a função merge do Pandas.

b) Faça uma análise ilustrando a (possível) qualidade das variáveis propostas.

Questão 3 (valor 2,0 pontos)

a) Aprimore o pipeline desenvolvido em aula para que seja possível comparar mais de uma técnica de modelagem usando um mesmo grid-search;
   - **Obs.: não use cross-validation!!!** Separe a base em treino e validação usando o campo referente a data da transação;

b) Faça uso de uma técnica de modelagem ainda não falada na disciplina. Essa técnica não precisa ser uma modelo de regressão, ela pode ser uma técnica

de agrupamento para ajudar na criação de novas variáveis. **Considere como técnicas já vistas: KNN, árvores, ensembles, Naive Bayes e SVM.**

Questão 4 (valor 2,0 pontos)

a) Comparar vários modelos candidatos e escolher o melhor de acordo com a AUC;
b) Gerar os resultados do melhor modelo na base de teste e enviar para o Kaggle. Colocar alguma evidência do envio no projeto entregue;

Questão 5 (valor 2,0 pontos)

a) Ao executar o `predict_proba` podemos obter os scores ("probabilidades") de cada transação ser uma fraude. Numa situação prática deve-se escolher um ponto de corte a partir do qual as transações serão consideradas como fraude e barradas/bloqueadas. Escolha um bom valor de ponto de corte considerando:
    - O valor monetário perdido em uma fraude (prejuízo);
    - O valor bloqueado numa transação que não era fraude (ele provavelmente gera algum prejuízo, dado que o cliente pode migrar para o concorrente)
    - E uma taxa aceitável de falsos positivos, dado que cada falso positivo causa um transtorno para um cliente e pode levar a empresa a perdê-lo.

    Obs.: Justifique muito bem a sua resposta!