Prof. M.Sc. Howard Roatti – FAESA Centro Universitário

Poderia prever quem estaria interessado em comprar uma apólice de seguro e explicar por quê?

Data Analysis and Machine Learning Hackathon

Avaliação C3 (8,0 Pontos) – 2022/1

Chegamos a nossa última avaliação da disciplina de **Análise de Dados Aplicada à Computação**, e agora devemos utilizar tudo o que aprendemos para essa tarefa que está sendo proposta e algo mais. Vocês deverão se reunir em grupos de no ***mínimo 3 e no máximo 4 alunos***. Os ***nomes dos componentes*** devem ser enviados para howard.cruz@faesa.br ***até o dia 05/06*** com o título **ADAC – AVALIAÇÃO C3 - GRUPO**. *Caso o grupo não envie o número de membros mínimo ou alguém não esteja fazendo parte de nenhum dos grupos, eu irei distribuí-los pelos grupos aleatoriamente*.

Datas Importantes:

- **Envio** de e-mail com os **nomes dos componentes** dos grupos: 05/06
- Divulgação da **ordem de apresentação**: 07/06(5HC) ou 09/06(5SC)
- **Envio do Jupyter Notebook (Python ou R)**: 18/06
    - Todos os grupos deverão enviar o Jupyter Notebook (com as análises, criação e avaliação dos modelos).
- **Apresentações**: 5HC 21/06 e 28/06; 5SC 23/06 e 30/06

A TAREFA de vocês envolve (*mas não se restringe a*):

- Acessar o link Insurance Company Benchmark (COIL 2000) Data Set e:
    - Compreender sobre o que esses dados tratam através das informações disponíveis no link.
    - Compreender o significado de cada atributo através do arquivo **dictionary.txt** (use um tradutor, caso tenham dificuldade em compreender o significado de cada atributo).
    - Realizar o download do *dataset*, composto de: **ticdata2000.txt** que são dados de treino com o último atributo sendo o rótulo de classe de cada linha do conjunto de dados, onde 0 é a ausência de uma apólice e 1 é a presença da apólice de seguros a ser predita. **ticeval2000.txt** e **tictgts2000.txt** que são dados de validação e rótulos de classificação para cada linha dos dados de validação, respectivamente.
    - Mantenham os dados de validação separados, eles **serão utilizados no final do processo para previsão e explicação dos resultados previstos pelo modelo gerado**. Mas lembre-se que o mesmo tratamento que aplicarem nos dados de treino e teste, deve ser aplicado nos dados de validação.

- Antes de iniciar o processo de criação do modelo, vocês deverão **Explorar os Dados**, isso irá auxiliá-los na interpretação das previsões feitas pelo modelo.
    - A interpretação dos resultados das previsões deve conter uma análise descritiva informando o porquê o modelo previu determinada situação para as amostras de **validação**. Vocês poderão agrupá-las (*clustering*) e descrever o comportamento do modelo para cada grupo, ao invés de individualmente.
- Para treinar o modelo separe os dados de *treino e teste* e apliquem uma das técnicas de validação cruzada: **hold-out crossvalidation**, ou **one-leave-out crossvalidation** ou **k-fold crossvalidation**. Lembrem-se de gerar as métricas de avaliação e comentar para verificarmos o desempenho do modelo ao final da validação cruzada. Experimentem diversas configurações dos modelos, não parem no primeiro resultado!
- Quando estiverem satisfeitos com os resultados do modelo, aplique-o aos **dados de validação**. Como temos os rótulos dos dados de validação, podemos comparar e gerar as métricas de avaliação para verificar o desempenho do modelo nesse conjunto de dados.
- Mantenha tudo que foi feito no Jupyter Notebook, inclusive as bibliotecas que precisaram instalar, expliquem cada passo feito no Jupyter Notebook (dica: usem células ***markdown*** para ficar mais elegante e compreensível).
- Criem uma ordem que faça sentido para vocês no processo de desenvolvimento do trabalho.
- Apresentação/Demonstração: sua apresentação deve conter a análise exploratória dos dados, onde vocês deverão fazer *data storytelling* (pesquisem sobre o assunto), explicação dos modelos selecionados (o porquê o selecionaram cada modelo. E não vale dizer “porque sim”!), explicação dos resultados obtidos em geral. Explicação geral do que levou o modelo a errar nas validações. Sejam criteriosos na seleção dos modelos, não incluam todos, nem escolham qualquer um. Lembrem-se como eles se comportam e quais características eles têm de interessante para serem utilizados.
    - Experimentem mais de um modelo, mostrem o comparativo entre eles e utilize, além da forma como o modelo se comporta sobre os dados, as métricas avaliativas obtidas por eles para explicar a sua seleção.
    - As métricas apresentadas que deverão estar no trabalho (mas não se restringem) são: *recall*, *specificity*, *precision*, Macro-F1, Micro-F1, ROC e AUC.
- Defina uma ou mais pessoas para apresentar o trabalho.
- Apenas um componente precisará enviar o Jupyter Notebook com os nomes de todos os membros do grupo até a data prevista.
- **Não será possível mudar de grupo.**

- **Não será possível alterar a data de apresentação. Programem-se para estarem presentes!**
- Todos os componentes devem estar presentes no dia da apresentação. A falta sem motivo implica no **desconto de 0,5 pontos** do componente ausente.
- Atrasos na entrega serão tolerados, serão **descontados 0,5 pontos** do grupo por dia de atraso.
- Durante a apresentação será avaliado a clareza nas explicações, o domínio do assunto e a construção da apresentação como um fluxo simples de entendimento. **A falta de algum dos itens do Notebook fará com que haja desconto na apresentação**. **Valor 4,0 pontos**.
- O Jupyter Notebook deverá conter todos os itens discriminados anteriormente, cito: (1) Análise Exploratória e Explicativa dos Dados, (2) Tratamento dos Dados Quando Houver Necessidade, (3) Explicação Passo-a-Passo de Cada Tarefa Executada, (4) *Crossvalidation*, (5) Métricas Avaliativas de Desempenho de Cada Modelo Experimentado, (6) Referências de Pesquisa Externas Utilizadas. **Valor 4,0 pontos**.
- **Não serão aceitos trabalhos idênticos.**
- **Não serão aceitos cópias parciais ou completas de trabalhos de terceiros.**
- **Em caso de dúvidas, procurem:** Google, Stack Overflow, Kaggle, Fóruns, Blogs, Medium, Towards Data Science, Analytics Vidhya, Documentação Oficial das Bibliotecas, Colegas de Turma mais Experientes ou me procurem!