Bootcamp: Arquiteto de Machine Learning

Desafio Final

Módulo 5	Desafio Final

Objetivos

Exercitar os seguintes conceitos trabalhados no Bootcamp:

✓ Análise exploratória dos dados (EDA - *Exploratory Data Analysis)*.

✓ Preparação dos dados.

✓ Comparação e ajuste de modelos de classificação.

Enunciado

Neste desafio, serão abordados, em grande parte, os conceitos trabalhados durante todo o Bootcamp Arquiteto de Machine Learning. Esses conceitos serão explorados por meio da detecção de fraudes em cartões de crédito. A análise do dataset e as aplicações de técnicas como subamostragem e deep learning serão utilizadas na solução deste desafio final.

Para este desafio, será utilizado o dataset "*Credit Card Fraud Detection*", disponível no ***Kaggle*** (https://www.kaggle.com/mlg-ulb/creditcardfraud?select=creditcard.csv). Esse dataset contém um conjunto de transações realizadas durante o mês de setembro de 2013. Os dados dessas transações foram coletados durante 2 dias.

Como os dados presentes nesse dataset apresentam transações reais, foi necessária a aplicação do PCA (*Principal Component Analysis*) sobre os dados originais a fim de garantir a privacidade dos dados. Desse modo, os valores presentes nas colunas "V1" a "V28" apresentam as componentes principais dessas transações, a coluna "Time" apresenta a diferença de tempo, em segundos, entre o momento de ocorrência da transação e a primeira transação presente no dataset, a coluna "Amount" apresenta o

valor da transação e a coluna “Class” mostra a classificação da transação como fraude (1) ou não fraudulenta (0). Essa coluna “Class” será utilizada como saída para os algoritmos supervisionados utilizados.

Atividades

Os alunos deverão desempenhar as seguintes atividades:

Acessar o Google Collaboratory.

1. Realizar o upload do dataset “creditcard.csv” presente no link: https://drive.google.com/file/d/1s_bSWBT_e5RzfzD7-x-Db0AZByyD3Xsv/view?usp=sharing

3. Construa um código utilizando como base os trabalhos práticos, os códigos apresentados nos módulos e o enunciado das questões presentes neste desafio.

4. Para a implementação dos algoritmos, utilize as definições abaixo:

Divisão entre treinamento e teste

X_train, X_test, y_train, y_test = train_test_split(entrada_normalizada, dados[Class], test_size = 0.3,random_state=42)

Algoritmo Regressão logística:

lr = LogisticRegression(max_iter=1000,random_state=42)

Algoritmo MLP

mlp = MLPClassifier(alpha=0.001,hidden_layer_sizes=(10,), activation='relu',solver='adam' ,random_state=1)

Obs.:

1. Utilize a normalização dos dados por meio do StandardScaler() para todos os algoritmos.

2. Para a divisão dos dados de treinamento e teste dos algoritmos, utilize o valor de “**random_state=42**” e a proporção de **70**% para **treinamento** e **30**% para **teste**.

3. Aplique primeiro a normalização e, depois, aplique a divisão dos dados entre treinamento e teste. Para a aplicação de todos os modelos, utilize essa sequência de passos.

4. Utilize a variável “Class” como saída e as demais como entrada do modelo.

5. Para as questões de subamostragem, utilize os comandos a seguir:

```
#encontrando o número de instâncias da classe 1

n_fraude = len (df_cartoes_ajustado[df_cartoes_ajustado.Class==1])

indices_fraude = np.array (df_cartoes_ajustado[df_cartoes_ajustado.Class==1].index)

indices_sem_fraude=np.array (df_cartoes_ajustado[df_cartoes_ajustado.Class==0].index)

#escolhendo indices aleatórios para os dados normais. Selecionando a mesma quantidade de dados para as transações fraudulentas

np.random.seed(0)

escolha_sem_fraude = np.random.choice (indices_sem_fraude, n_fraude, replace = False )

#selecionando a quantidade de dados por meio dos indices

indices_subamostragem=np.concatenate([indices_fraude,escolha_sem_fraude],axis=None)

#escolhendo os dados por meio dos indices escolhidos

dados_subamostrados = df_cartoes_ajustado.iloc[indices_subamostragem,:]

#identificando os valores de entradas e saida para a subamostragem

entradas=dados_subamostrados[['V1', 'V2', 'V3', 'V4', 'V5', 'V6', 'V7', 'V8', 'V9', 'V10', 'V11',

    'V12', 'V13', 'V14', 'V15', 'V16', 'V17', 'V18', 'V19', 'V20', 'V21',

    'V22', 'V23', 'V24', 'V25', 'V26', 'V27', 'V28',

    'scaled_amount', 'scaled_time']]

saida=dados_subamostrados[['Class']]
```

Respostas Finais

Os alunos deverão desenvolver a prática e, depois, responder às seguintes questões objetivas: