data expert.

Data Science & Machine Learning

ASSIGNMENT

Portfólio Individual

CLUSTERING E REGRESSÃO

#19

dnc

INSTRUÇÕES ASSIGNMENT 19

INSTRUÇÕES

Este Assignment #19 é um mix de conteúdos de Clusterização e Regressão.
As questões respondidas deste assignment valem nota normalmente como nos anteriores. O desenvolvimento da resolução que será postado no seu github pessoal e servirá como um portfólio individual de Clusterização e Regressão

Logo, você deverá:

- Desenvolver o modelo e colocar no **Github**
- Responder as **questões** referentes a solução deste Assignment
- Link do Github: na última questão você deverá inserir o link do seu github contendo o desenvolvimento deste assignment.

OBJETIVO DO ASSIGNMENT COMO PORTFÓLIO INDIVIDUAL

1. Portfólio individual serve para **demonstrar o seu conhecimento** adquirido no curso através de **projetos intermediários** publicados em seu github pessoal. ** Podem ser apresentados a empresas em entrevistas de emprego.*

2. No final do curso, vocês terão no total de 3 projetos intermediários como portfólio. Cada um deles valerá entregas de Assignments distribuídos nos slots finais 6, 7 e 8.

3. Os conteúdos podem se misturar com módulos anteriores. O importante é mostrar sua ideia, ter um portfólio estruturado.

dnc

INSTRUÇÕES PARA O ASSIGNMENT 19

CONTEXTO

Você é um cientista de dados e precisa desenvolver um projeto para o banco digital em expansão. Este banco irá anunciar a função crédito para os novos clientes que entrarem, baseando-se nos dados de movimentações deles de outros bancos e dos nossos clientes. Tem um conjunto de dados com 18 colunas, descritas abaixo. São aproximadamente dados de 9000 clientes e seus comportamentos financeiros como dinheiro em conta, frequência de movimentações, pagamentos com cartões, compras parceladas e limite de crédito.

Link: https://www.kaggle.com/arjunbhasin2013/ccdata

O PROJETO:

Os dados que você recebeu são dados brutos de quase 9000 clientes de outros bancos que também podem se tornar novos clientes. Uma forma de cativar e adquirir novos clientes é oferecendo um crédito justo, baseado nestes dados que você coletou.

Você deve então realizar **duas modelagens**, uma para criar rótulos (labels) para os clientes agrupando eles por comportamentos e em seguida, utilizar dessa nova feature para criar um modelo de regressão que irá propor um crédito para o cliente baseado-se nestes dados. Abaixo, estão expostas algumas etapas SUGERIDAS para você realizar no projeto.

OBS: O desenvolvimento e a decisão do modelo é totalmente sua, portanto se achar que deve utilizar um valor/ procedimento diferente, sinta-se livre para fazer os testes e validar suas hipóteses para achar o resultado coerente.

dnc

INSTRUÇÕES PARA O ASSIGNMENT 19

LIMPEZA DOS DADOS

- Remova dados não importantes para análise (CUST_ID)
- Para o caso de valores (OQUE?) preencher os valores ausentes com o mínimo da coluna.
- Remover os outliers: selecionar os dados que estejam **abaixo** dos 95% das colunas ['BALANCE','PURCHASES','MINIMUM_PAYMENTS', 'PAYMENTS', 'CREDIT_LIMIT','PURCHASES_TRX','ONEOFF_PURCHASES', 'CASH_ADVANCE','CASH_ADVANCE_TRX']

EDA

Realize uma EDA da forma como preferir, explore os dados, levante ideias, avalie correlações.

1° ETAPA: CLUSTERIZAÇÃO

Modelo de agrupamento (Clusterização) é algo essencial para correlacionar os semelhantes. Avalie os métodos Elbow (Cotovelo) e Silhouette (Silhueta), por exemplo, para encontrar o número ótimo de Clusteres. Utilizando o GaussianMixture, faça a clusterização agrupando os clientes mais similares. Utilize esses labels como variáveis categóricas no modelo de Regressão.

2° ETAPA: REGRESSÃO

Você deve então criar um modelo de Machine Learning de Regressão para prever o **CREDIT_LIMIT** utilizando os dados bancários do usuário e o tipo deles (calculado na Primeira Parte). Para isso, utilize o modelo de Redes Neurais para Regressão do Sklearn (Para fins de testes, é interessante realizar uma otimização)

Documentação:
https://scikit-learn.org/stable/modules/neural_networks_supervised.html

dnc

INSTRUÇÕES PARA O ASSIGNMENT 19

FEATURE DICT

- CUSTID : Identification of Credit Card holder (Categorical)
- BALANCE : Balance amount left in their account to make purchases
- BALANCEFREQUENCY : How frequently the Balance is updated, score between 0 and 1 (1 = frequently updated, 0 = not frequently updated)
- PURCHASES : Amount of purchases made from account
- ONEOFFPURCHASES : Maximum purchase amount done in one-go
- INSTALLMENTSPURCHASES : Amount of purchase done in installment
- CASHADVANCE : Cash in advance given by the user
- PURCHASESFREQUENCY : How frequently the Purchases are being made, score between 0 and 1 (1 = frequently purchased, 0 = not frequently purchased)
- ONEOFFPURCHASESFREQUENCY : How frequently Purchases are happening in one-go (1 = frequently purchased, 0 = not frequently purchased)
- PURCHASESINSTALLMENTSFREQUENCY : How frequently purchases in installments are being done (1 = frequently done, 0 = not frequently done)
- CASHADVANCEFREQUENCY : How frequently the cash in advance being paid
- CASHADVANCETRX : Number of Transactions made with "Cash in Advanced"
- PURCHASESTRX : Numbe of purchase transactions made
- CREDITLIMIT : Limit of Credit Card for user
- PAYMENTS : Amount of Payment done by user
- MINIMUM_PAYMENTS : Minimum amount of payments made by user
- PRCFULLPAYMENT : Percent of full payment paid by user
- TENURE : Tenure of credit card service for user

dnc

INSTRUÇÕES PARA O ASSIGNMENT 19

BOAS PRÁTICAS DE UM PORTFÓLIO

Lembrando que ao longo da sua análise, introdução, EDA, transformações e modelagem, é importante ressaltar a ideia por trás da sua ação para que o leitor entenda o raciocínio usado. Essa prática de documentação enriquece ainda mais seu projeto e garante um entendimento melhor do público interessado.

Links como referência:

1. Guia para aplicação dos métodos
https://www.kdnuggets.com/2019/10/clustering-metrics-better-elbow-method.html

2. Elbow Method vs Silhouette Score
https://vitalflux.com/elbow-method-silhouette-score-which-better/

dnc

dnc

#HARDWORK

>