[EN] Introduction

Once a model goes from laboratory conditions to being deployed in the real world, the problem of changing conditions and tendencies may cause performance degradation. This phenomenon can be explained by the presence of concept drift. Additionally, one should make sure (as much as possible) that no bias exists in data to avoid any legal actions or unintended consequences for your business.

For this exercise, an adaptation from a Kaggle dataset will be used focusing on predicting the probability of a certain Instagram account following your own account. The datasets have a target feature that states if a certain account has, or has not, followed your own account, named “Class”. Our dataset is composed of the following 16 input features:

1. A boolean feature on whether the account is a business account or not;
2. A boolean feature that describes if the user is from the second most popular country the majority of the testing business's account's followers are from;
3. A numerical feature of the length of the username: (Note we choose this seemingly arbitrary feature since some engagement strategies suggest including the descriptive niche in the username to improve search results);
4. A boolean feature of whether the user is from the same country as the testing account in question;
5. A boolean feature of whether the account is a professional account or not: (Note a business account and a professional account are two separate modes on the Instagram platform);
6. A numerical feature that describes the number of posted videos aka InstaReels;
7. A boolean feature that describes gender of the user;
8. A boolean feature that describes if the user has recently joined Instagram;
9. A boolean feature that describes if the user's identity is verified or not by Instagram;
10. A boolean feature that describes whether the user's account is private or not;
11. A numerical feature that describes the length of the user's biography;
12. A numerical feature that describes the mean number of likes the user gets on their posts;
13. A numerical feature that describes the number of mutual followers between the testing account and the user's account;
14. A numerical feature that describes the user's number of followers;
15. A numerical feature that describes the user's number of posts;
16. A numerical feature that describes the percentage of the user's followers that the user itself follows;

Link to the official dataset:

https://www.kaggle.com/datasets/gabriellacolletti/concept-drifted-data

Exercises

1) Assuming you have a running model in a deployed server in Google Cloud Platform for a while, you want to understand if recent data is drifting from the initial data used for training. The main issue is that no ground truth is available for comparison, so you can only rely on input data for drift detection. Since you want to know if a particular feature has drifted, you are advised to use univariate analysis techniques. Based on both given datasets (historical.csv + new_data.csv), can you tell if there's any drift? If yes, can you please state what are the features that drifted and explain why you reached such a conclusion?

2) Imagining that some of the features suffered a drift, you still want to make sure that the drift is significant enough for model retraining. Since some of the drifts are smaller than others, you want to perform a broader analysis of the drift, so you decide to go for a multivariate analysis. For this, please state your choice of approach and explain if a significant drift exists and justifies retraining an existing model.

3) By using the provided dataset (training_data.csv) train a classifier that can predict the probability of an account following your Instagram account. Once you trained the model, the business specialist wants to better understand how the input features are impacting the predictions of the model. Hence, you need to come up with a solution to increase explainability. Based on this, the business specialist can devise a new strategy to increase the number of followers in Instagram accounts. Please indicate what techniques might be used, what technique have you chosen and why, together with the top 3 features that maximize someone to follow a certain Instagram account.

4) Due to a newly EU regulation that foresees high penalties for companies where discrimination for any societal sectors or groups exist, you are asked by your Ethical Department to check if there's any bias towards a new follower being a man or a woman. This bias is really important due to digital marketing strategies, where both women and men should be treated equally and target in the same ways. From the given training data, can you reach to a conclusion on the existence of a gender bias? Please justify.

[PT] Introdução

Assim que um modelo é passado da fase de desenvolvimento para a fase de deployment em ambiente real, o problema de mudanças de condições e tendências pode provocar uma degradação de performance. Este fenómeno pode ser explicado com a presença de "*concept drift*". Adicionalmente, é necessário garantir (tanto quanto possível) que o modelo treinado ou os dados em si não são tendenciosos (biased) para evitar quaisquer ações legais ou consequências não previstas para o negócio.

Para este exercício, uma adaptação de um dataset do Kaggle será utilizado focando principalmente na previsão de probabilidade de uma determinada conta de Instagram seguir a nossa própria conta. O dataset contém uma variável "target" que indica se uma determinada conta seguiu, ou não, a nossa conta de Instagram, chamada "Class". O dataset é composto pelas seguintes 16 variáveis de entrada:

1. Uma variável booleana que indica se uma conta é empresarial ou não;
2. Uma variável booleana que indica se o utilizador é proveniente do segundo país mais popular relativamente aos seguidores
3. Uma variável numérica que indica a dimensão do username;
4. Uma variável booleana que indica se a potencial conta é do mesmo país que a nossa própria conta.
5. Uma variável booleana que indica se a conta é profissional ou não (pf notem que contas empresariais e profissionais têm representações diferentes na Instagram);
6. Uma variável booleana que indica a número de vídeos postados, mais conhecidos como InstaReels;
7. Uma variável booleana que indica o género do utilizador da conta;
8. Uma variável booleana que indica se um utilizador se juntou recentemente ao Instagram;
9. Uma variável booleana que indica se a identidade de um utilizador foi verificada pelo Instagram;
10. Uma variável booleana que indica se uma conta é privada ou não;
11. Uma variável numérica que indica a dimensão da biografia de um utilizador;
12. Uma variável numérica que indica a média de "likes" que um utilizador tem em todas as publicações;
13. Uma variável numérica que indica o número de seguidores mútuos entre a conta em questão e nossa própria conta;
14. Uma variável numérica que indica o número de seguidores;
15. Uma variável numérica que indica o número de publicações;
16. Uma variável numérica que indica a percentagem de seguidores que a conta tem relativamente às contas que a conta segue. E.g. Se uma conta tem 100 seguidores e essa mesma conta segue esses 100 seguidores, então a percentagem é de 100%. E conta não segue nenhum dos seus seguidores, então a percentagem é de 0%.

Link para o dataset oficial:

https://www.kaggle.com/datasets/gabriellacolletti/concept-drifted-data

Exercícios

1) Assumindo que temos um modelo treinado e a executar num servidor na Google Cloud Platform já há algum tempo, é necessário perceber se existe algum "drift" dos dados recolhidos atualmente sobre os dados usados para treino. O principal problema é que não existem dados recentes classificados (não existe "ground truth") para comparação, por isso apenas nos podemos guiar pelos dados recentes de entrada do modelo. Visto querermos saber se uma variável específica sofreu algum "drift", ´´e aconselhada a utilização de técnicas mono-variável. Baseados nos dados fornecidos (historical.csv + new_data.csv), pf indique se existe algum "drift", e se sim, indique quais as variáveis que sofreram esse "drift" e explique o porquê de as ter considerado?

2) Imaginando que algumas das variáveis sofreram um "drift", queremos saber se o impacto desse "drift" é realmente significativo para justificar o retreino do modelo existente. Visto que os "drifts" de algumas variáveis são menores do que outros, considere agora uma análise multi-variável. Para isso, pf indique que abordagem utilizou e explique se o "drift" é significativo o suficiente que justifique o retreino do modelo.

3) Utilizando os dados fornecidos (training_data.csv), pf treine um classificador que preveja a probabilidade de uma conta seguir a nossa própria conta. Uma vez treinado o modelo, o especialista do negócio afeto ao projeto quer melhor entender como é que a variáveis de entrada têm influência nas previsões realizadas. Posto isto, é necessário desenvolver uma solução para aumentar a explicabilidade do modelo. Com esta informação, o especialista conseguirá criar uma nova estratégia que aumente o número de seguidores da nossa conta de Instagram. Pf indique quais a técnicas que podem ser utilizadas para este fim, qual a técnica utilizada e porquê, juntamente com o top 3 das variáveis que maximizam uma determinada conta seguir a nossa própria.

4) Devido a uma nova norma europeia que prevê altas penalizações para empresas onde exista alguma tipo de discriminação sobre algum setor social ou grupo, o departamento de ética pede-lhe para verificar se os dados ou modelo são tendenciosos relativamente a homens e mulheres. Esta verificação é bastante importante visto que novas estratégias de Marketing Digital estão a ser criadas e é necessário garantir que tanto homens ou mulheres são tratados de maneira igualitária. Atendendo aos dados de treino (training_data.csv), pf indique se existe ou não algum tipo de comportamento tendencioso do modelo ou se os dados são tendenciosos. Pf justifique.