

Programa Avançado em Data Science e Decisão

Computação para a Ciência dos Dados

Projeto Integrador

Contextualização

O projeto integrador de 2024 consiste em construir um pipeline completo de pré-processamento de dados e construção de modelo preditivo de classificação, voltado para prever se uma empresa irá deixar de operar em até dois anos.

Os dados foram coletados e curados pela Bisnode, uma empresa europeia do ramo de *business information*.

São dados entre os anos de 2005 e 2016, com empresas em alguns segmentos da economia (como produtos eletrônicos, equipamentos elétricos, motores, etc.) e serviços (alimentação, bebidas e hospedagem). Empresas que possuem receita acima de 100 milhões de Euro foram remotivadas para fins de minimizar as possibilidades de identificação.

Atividades de Pré-processamento

Todo o pipeline de pré-processamento de dados para a construção dos modelos preditivos deve ser feito em Python.

Isto é, seu grupo deve produzir um notebook em Python e o output desse pipeline deve ser um arquivo já pronto para ser usado para treinar o modelo.

Obs: vocês podem deixar o split dos dados para o R, se preferir.

Mas atenção, essa é a única exceção 😊.

Requisitos

Aqui vamos listar alguns requisitos que seu grupo deve executar em Python.

- . O pré-processamento deve usar as bibliotecas Pandas e Numpy. Será permitido também usar dfply ★. Ah, e os gráficos podem ser feitos com a biblioteca de sua preferência! (Quem sabe Altair ...)

. Remova as colunas ['COGS', 'finished_prod', 'net_dom_sales', 'net_exp_sales', 'wages', 'D'] pois elas apresentam um percentual considerável de *missing data*

- Dica: use o pacote missingno para verificar o percentual de missing das features. Seu grupo pode optar por remover outras variáveis ou, até mesmo, tratá-las! 😊

Requisitos

- . Remova de seus dados os registros do ano de 2016
- . Será preciso criar uma coluna da variável dependente que será objeto da predição. Para isso, use o conceito de que uma empresa deixou de operar se ela esteve ativa no ano X , mas não apresentou vendas em $X + 2$ anos.
 - 🧑‍🔬 dica: sugerimos que você use as funções `stack`, `unstack`, `groupby` e `shift` para isso.
 - 😎 Melhor dica ainda é não perder a aula em que faremos isso juntos!

Requisitos

- . Filtre para trabalhar apenas com empresas do ano de 2012
- . Agora é o momento de olhar por inconsistências nos dados. Por exemplo, veja a coluna Sales . Há volumes de venda negativos, isso não faz sentido!
 - Aproveite para usar np.where para ajustar isso. De modo que onde Sales < 0 você já pode substituir por 0
 - Essa variável é bastante assimétrica, concorda? Será que vale criar novas colunas que representem o valor emlog dessa coluna?
 - Será que isso também se aplica para as demais?

Requisitos

- . Crie novas colunas, como idade da empresa (faça isso pela subtração de `founded_year` e `year`). Ah, cuide bem dos missing values. `np.where` pode ajudar bastante!
- . Filtre seus dados para ter empresas que possuem receita (*revenue*) abaixo de 10 milhões de euros e acima de 1000 euros.
- . Busque sempre embasar qualquer decisão de tratamento das variáveis. Faça isso com o auxílio de estatísticas descritivas e também de gráficos de apoio.
- . E lembre-se que Data Science é uma atividade exploratória! Super normal você voltar e rever seus passos de pre-processamento após ter os resultados iniciais de seus modelos. 😊

Formato de entrega

- . Um jupyter notebook com todo o código
 - bem documentado, use **Markdown** para explicar as idéias, código, gráficos, etc...
 - com uso de estatísticas descritivas e gráficos
 - enviar arquivo `.ipynb` e `.html`
- . Apresentação com os principais pontos, usando o próprio notebook (10 min)
 - todos os grupos estão cientes desse pipeline, então foque na apresentação nos diferenciais do seu grupo e nos pontos que foram mais desafiadores