Turing Academy - 2023 - Entregável III

Análise de Dados/Machine Learning

Chegou o momento! Os entregáveis que você, trainee, deverá realizar consistem em um conjunto de 3 exercícios que avaliarão o seu domínio nos tópicos estudados durante o Turing Academy até o momento, dentre os quais:

- Manipulação de DataFrame com pandas;
- Visualização;
- Limpeza;
- Análise;
- Pré-Processamento de dados;
- Machine Learning;

Para não sobrecarregá-los, soltaremos um exercício por vez (1 por semana). Recomendamos fortemente que vocês tentem fazer os exercícios à medida que eles saiam (não deixem acumular!!!).

Para a entrega dos exercícios, pediremos para que vocês criem um repositório no GitHub (ele deverá ser público) e adicionem o link para ele no espaço apropriado dessa planilha: https://docs.google.com/spreadsheets. Submetam a esse repositório 1 arquivo no formato .ipynp para cada entregável. O prazo para entrega de todos os exercícios é até dia 21/07 (não necessariamente eles precisam ser entregues na ordem).

Para a realização dos exercícios, não se limitem aos conhecimentos vistos durante as aulas ou ao material do TA - incentivamos que busquem conhecimento em outras fontes, como no DataCamp ou no Youtube. Não hesitem em chamar os seus mentores ou qualquer um dos professores do TA para tirar dúvidas. Estaremos sempre à disposição! Bom trabalho!

EX3 - Promoções na TurinGucci

Nos últimos anos, o mercado de trabalho tem mudado muito, e muitas profissões tiveram um crescimento muito grande; com isso, uma grande empresa que fabrica roupas, a *TurinGucci*, precisa saber quais de seus funcionários devem ser promovidos para que a produtividade da empresa cresça!

Por esse motivo, foi contratada uma segunda empresa especializada em serviços de RH, o *Grupo Touring*, para analisar dados a respeito de cientistas de dados. O *Grupo Touring* busca ajuda na construção de um modelo para predizer se um determinado funcionário deve ser promovido ou não.

Por isso, sabendo do seu sucesso nas suas análises anteriores, o *Grupo Touring* te propõe um desafio a mais: além de limpar e analisar alguns dados para eles, você terá que **construir um modelo** que prediz se um determinado funcionário deve ou não ser promovido. Portanto, você deverá **limpar os dados** fornecidos e realizar uma **análise rápida** a fim de obter insights interessantes, como por exemplo quais as relações entre as colunas, e quais colunas mais se relacionam com a target (se o funcionário deve ou não ser promovido). Por fim, você deverá treinar um modelo que tenha bons resultados para realizar a predição necessária.

Você pode baixar o dataset para o projeto aqui: dataset3.psv. As colunas presentes nele são as seguintes:

- **employee_id:** ID único do funcionário;
- **department:** departamento do funcionário;
- **region:** região de emprego (não ordenada);
- **education:** nível educacional do funcionário;
- **gender:** gênero do funcionário
- **recruitment_channel:** canal de recrutamento para o funcionário;
- **no_ of_ trainings:** número de treinamentos completados em anos passados, sejam em soft skills, technical skills, etc.
- **age:** idade do funcionário;

- **previous_ year_ rating:** avaliação do funcionário para o ano anterior;
- **length_ of_ service:** duração do serviço em anos;
- **awards_ won?:** se ganhou algum prêmio em anos anteriores, recebe 1, do contrário 0;
- **avg_ training_ score:** pontuação média nas avaliações de treinamento atuais;
- **is_promoted: (Target)** recomendado ou não para promoção.

Apresentado o problema, faça o seguinte:

a) Utilize diferentes técnicas aprendidas para realizar a limpeza do dataset.
b) Realize uma breve análise do conjunto de dados. Não é necessário fazer algo muito detalhado.
c) Realize o Categorical Encoding das features catégóricas do dataset.
d) Promova a normalização das colunas com variáveis contínuas.
e) Divida seus dados entre treino e teste. Utilizando a biblioteca sklearn treine um modelo KNN. Determine o menor número de vizinhos (K) para o qual a acurácia do modelo para os dados de teste é a maior possível.
f) Você percebe algum erro no modelo que acabou de treinar? Pesquise o que são Dados Desbalanceados e como isso pode afetar no desempenho de modelos de Machine Learning.
g) Pesqueise sobre Precisão e Recall, métricas de avaliação de modelos mais adequadas para lidar com dados desbalanceados. Calcule a precisão e o recall do modelo que você treinou. O que essas métricas indicam?
h) Opcional: Pesquise por um modelo ou por técnicas de Machine Learning capazes de minimizar o efeito de classes desbalanceadas. Tente aplicar. Esse Turing Talks pode te ajudar: Dados Desbalanceados — O que são e como lidar com eles | by Felipe Azank | Turing Talks | Medium.