Projeto de análise, limpeza e predição de dados

Enunciado:

Chegou o momento de mais um projetinho! No primeiro projeto, você realizou uma limpeza extensa e depois uma análise para obter insights significativos. Agora, você deve realizar novamente essas tarefas em outro dataset e treinar também os conceitos que aprendeu sobre predição de dados. Para isso, você utilizará todos os conhecimentos que adquiriu com o Turing Academy e com os cursos vistos para construir seu projeto, mas claro, sempre com ajuda e apoio de seus mentores!

Problema:

Nos últimos anos, o mercado de trabalho tem mudado muito, e muitas profissões tiveram um crescimento muito grande, como por exemplo empregos na área da tecnologia, gerando oportunidades cada vez melhores para os profissionais.

Por esse motivo, uma empresa do ramo de tecnologia, a *Turing Alpha*, foi contratada por uma segunda empresa especializada em serviços de RH, a *Nelsonson's*, para analisar dados a respeito de cientistas de dados. A *Nelsonson's* busca ajuda na **construção de um modelo para predizer se um determinado cientista de dados deseja mudar de emprego ou não**.

Por isso, sabendo do seu sucesso na análise que fez sobre as olimpíadas, a *Turing Alpha* te propõe um desafio a mais: **além de limpar e analisar alguns dados para eles, você terá que construir um modelo que prediz se uma determinada pessoa está ou não procurando mudar de emprego.** 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 a pessoa está buscando ou não mudar de emprego). Por fim, você **deverá treinar um modelo que tenha bons resultados para realizar a predição necessária**. Para isso, utilize os conhecimentos adquiridos nas aulas a respeito das bibliotecas de manipulação e análise de dados e sobre aprendizado supervisionado e métricas.

Informações sobre o dataset

Você pode baixar os datasets para o projeto aqui. Na pasta, terá dois arquivos, um para ser usado na hora do treino (“train.csv”) e outro na hora teste (“test.csv”).

Esses datasets possuem informações a respeito de candidatos para um determinado processo seletivo. As colunas presentes são:

- **enrollee_id:** ID exclusivo para o candidato
- **city:** Código da cidade
- **city_ development _index:** Índice de desenvolvimento da cidade (em escala)
- **gender:** Gênero do candidato
- **relevent_experience:** Experiência relevante do candidato
- **enrolled_university:** Tipo de curso universitário matriculado, se houver
- **education_level:** Nível de educação do candidato
- **major_discipline:** Major principal de educação do candidato
    - *STEM:* Science, Technology, Engineering, and Math Degree
    - *Business Major:* Administração de negócios
    - *Humanities:* Humanidades
    - *Arts:* Artes
    - *No major:* Não faz/possui um major
    - *Other:* Outro
- **experience:** Experiência total do candidato em anos
- **company_size:** Nº de funcionários na empresa do empregador atual
- **company_type:** Tipo de empregador atual
- **lastnewjob:** Diferença em anos entre o emprego anterior e o emprego atual
- **training_hours:** Horas de treinamento concluidas
- **target:** 0 - Não está procurando uma mudança de emprego ou 1 - Procurando uma mudança de emprego (essa coluna está presente apenas no dataset de treino)

Materiais de apoio

Caso seja necessário, não hesite em rever as aulas dadas no Turing Academy ou os cursos do Datacamp sobre os assuntos, e também não se esqueça de recorrer ao seu mentor sempre que tiver dúvidas ou precisar de ajuda!

Dicas

- ➔ Caso prefira, você pode juntar os dois datasets para a limpeza e análise e depois separá-los de novo
- ➔ Na limpeza, sempre se atente se as colunas estão com os data types corretos e se os dados dentro delas fazem sentido (Para colunas categóricas verifique se os valores únicos fazem sentido e verifique também sempre se os valores mínimos e máximos das variáveis numéricas fazem sentido para cada coluna)
- ➔ Os NaNs devem ser tratados apropriadamente, e nem sempre removê-los é a melhor solução
- ➔ Em uma análise é sempre necessário comentar todos os passos e descobertas, mesmo as que parecem "óbvias", pois é importante para o leitor entender cada etapa, e não se esqueça também de realizar uma conclusão
- ➔ Caso precise de inspiração, veja análises de outras pessoas (o Kaggle é um bom lugar para isso)
- ➔ Não misture os dados de treino e teste, lembre-se que dados vazados podem prejudicar muito sua predição - Se você optar por juntar os dois dataframes para a limpeza e análise, separe bem depois
- ➔ Como o dataset possui diversas features (colunas), para a predição use somente as que irão fazer diferença na hora de predizer a target, nem todas precisam ser utilizadas
- ➔ Fique à vontade para testar diferentes modelos, até os que não foram passados durante as aulas
- ➔ Teste diferentes métricas, não apenas a acurácia