Projeto

Turma focada em Data Science

Proposta

O resultado de um projeto de machine learning depende fundamentalmente da qualidade dos dados utilizados - porém no dia a dia os dados quase nunca estarão como precisamos. É comum encontrarmos dados faltando, imputações incorretas, entre outros. Parte do trabalho de um cientista de dados é lidar com a realidade, ou seja, ajustar os dados da melhor forma. Muitas vezes uma boa analise prévia também salva muito tempo na hora de implementar um modelo de machine learning.

Seu trabalho nesse projeto é, portanto, utilizar as técnicas e abordagens aprendidas em python para lidar com uma base de dados relativamente problemática e analisar as entradas, tirando o máximo de *insigths* possível. Isso é comumente chamado de EDA (Exploratory Data Analysis) no mundo de Ciência de Dados.

Entregas

Suas entregas serão a base final em csv, o código desenvolvido (100% de sua autoria), imagens que sejam resultado da análise e conclusões gerais em um arquivo pdf.

Os Dados

Estamos fornecendo uma base que contém dados relacionados a sessões de "speed dating". Essa base será *futuramente* retomada por você quando tivermos trabalhado mais na parte de machine learning! A ideia é: considerando esses dados, é possível dizer se duas pessoas irão dar "match"?

Os dados foram coletados de participantes em eventos experimentais de 2002-2004. Durante os eventos, os participantes tiveram um "primeiro encontro" de quatro minutos com todos os outros participantes do sexo oposto. Ao final de seus quatro minutos, os participantes foram questionados se gostariam de ver a pessoa novamente e avaliaram o parceiro em seis atributos: Atratividade, Sinceridade, Inteligência, Diversão, Ambição e Interesses Compartilhados. O conjunto de dados também inclui dados de questionário coletados de participantes em diferentes pontos do processo. Esses campos incluem: dados demográficos, hábitos de namoro, autopercepção em atributos-chave, crenças sobre o que os outros consideram valioso em um parceiro e informações sobre estilo de vida. A base foi alterada para fins do projeto.

As variáveis são:

- genero
- idade
- idade_parceiro
- raça
- raça_parceiro
- Mesma raça?
- importancia_mesma_raca - o quão importante é ter um parceiro da mesma raça?
- importancia_mesma_religiao - o quão importante é ter um parceiro da mesma

religião?

- formação - área de formação do participante
- pref_parceiro_atratividade - o quão importante é atratividade para o parceiro?
- pref_parceiro_sinceridade - o quão importante é sinceridade para o parceiro?
- pref_parceiro_inteligencia - o quão importante é inteligência para o parceiro?
- pref_parceiro_engracado - o quão importante é ser engraçado para o parceiro?
- pref_parceiro_ambicao - o quão importante é ser ambicioso para o parceiro?
- pref_parceiro_interessescomuns - o quão importante é ter interesses comuns para o parceiro?
- avaliacao_atratividade - avaliação do parceiro sobre minha atratividade no dia do evento
- avaliacao_sinceridade - avaliação do parceiro sobre minha sinceridade no dia do evento
- avaliacao_inteligencia - avaliação do parceiro sobre minha inteligência no dia do evento
- avaliacao_engracado - avaliação do parceiro sobre eu ser engraçado no dia do evento
- avaliacao_ambicao - avaliação do parceiro sobre eu ser ambicioso no dia do evento
- avaliacao_interessescomuns - avaliação do parceiro sobre termos interesses comuns no dia do evento
- busco_atratividade - o quanto busco atratividade?
- busco_sinceridade - o quanto busco sinceridade?
- busco_inteligencia - o quanto busco inteligência?
- busco_engracado - o quanto busco ter um parceiro engraçado?
- busco_ambicao - o quanto busco ambição?
- busco_interessescomuns - o quanto busco interesses comuns?
- sou_atrativo - minha autoavaliação
- sou_sincero - minha autoavaliação
- sou_inteligente - minha autoavaliação
- sou_engracado - minha autoavaliação
- sou_ambicioso - minha autoavaliação
- parceiro_atrativo - minha avaliação sobre o parceiro
- parceiro_sincero - minha avaliação sobre o parceiro
- parceiro_inteligente - minha avaliação sobre o parceiro
- parceiro_engracado - minha avaliação sobre o parceiro
- parceiro_ambicioso - minha avaliação sobre o parceiro
- interessescomunsparceiro - minha avaliação sobre o parceiro
- esportes - gosto?
- esportes_tv - gosto?
- exercicio - gosto?
- jantares - gosto?
- museus - gosto?

- arte - gosto?
- trilhas - gosto?
- videogames - gosto?
- leitura - gosto?
- tv - gosto?
- teatro - gosto?
- filmes - gosto?
- shows - gosto?
- musica - gosto?
- shopping - gosto?
- yoga - gosto?
- correlacao_interesses
- expectativas - Quão feliz você espera estar com as pessoas que conhecer durante o evento de encontros rápidos?
- quantos_acho_vao_gostar_demim - de 20 pessoas, quantas você acha que vão gostar de você?
- quantos_matches_acho_terei
- gostou_parceiro - gostou do seu parceiro?
- parceiro_gostou_demim_acho - acho que o parceiro gostou de mim?
- conhecia_parceiro - já conhecia o parceiro previamente?
- decisao - quero mais um encontro?
- decisao_parceiro - ele/ela quer mais um encontro comigo?
- match

Seu objetivo agora é aplicar os recursos aprendidos para preparar os dados, gerar uma análise (lembra da aula sobre EDA?) e tirar suas conclusões sobre esses dados. Se quiser, também é possível ir além e relacionar variáveis a fim de criar novas variáveis (chamadas de dummies) que serão também consideradas pelo modelo.

Informações Gerais

A entrega deve ser 100% de sua autoria e resultado de trabalho individual. O projeto foi disponibilizado no dia 29/03 e a data limite para entrega é o dia 04/04. Todos terão 2h para trabalhar durante o dia 31/03, e poderão tirar dúvidas com a Ana nesse período.

Dica: para importar a base use encoding= 'unicode_escape'