Teste para candidatos a Estatístico e/ou Engenheiro de Machine Learning na Appus

26 de março de 2016

INSTRUÇÕES

Para a solução das questões, o candidato pode escolher qualquer tecnologia de seu interesse.[1] As questões foram especificadas de forma que o candidato apresente soluções de acordo com seu entendimento, pois a capacidade de lidar com problemas diversos e propor soluções será um ponto avaliado. Os resultados podem ser enviados para o e-mail hudson@appus.technology em duas opções:

- Resumo em pdf da estratégia de solução adotada, resultados obtidos e códigos utilizados em cada questão;
- Disponibilizar acesso a repositório de código onde os resultados estão armazenados.

A partir das respostas enviadas, a Appus selecionará os candidatos para a fase de entrevistas, onde este apresentará para nossa equipe os resultados obtidos.

TAREFAS

1. Um Consultor de Recursos Humanos na função de parceiro de negócio e responsável por munir os gerentes com *insights* sobre a força de trabalho está preocupado com o alto índice de turnover na sua empresa, principalmente para públicos específicos (trainee, talentos, posição crítica, gerentes de vendas). Em função disto, busca uma forma de predizer quais colaboradores são mais suscetíveis a deixar a empresa, de forma voluntária, ao longo do próximo ano. Para tanto, dispõe dos seguintes dados:
    - Desligamento: variável binária (0 = Não; 1 = Sim) indicando se o colaborador foi desligado no último ano (voluntariamente)

[1] Algumas tecnologias que recomendamos: python, Jupyter, scikit-learn, pandas, R, RStudio, R Markdown, docker, GitHub, Bitbucket, AWS

- Ex-trainee: variável binária (0 = Não; 1 = Sim)
- Data de nascimento: algumas estão com formato DD/MM/AAAA equanto outras estão com formato MM/DD/AA.
- Sexo: variável binária (M = Masculino; F = Feminino)
- Data de admissão: algumas estão com formato DD/MM/AAAA, outras estão com formato MM/DD/AA
- Cargo: variável categórica que indica o cargo ocupado no último ano
- Área: variável categórica que indica a área que trabalhou no último ano
- Salário médio mensal: variável numérica que representa o salário médio recebido no último ano
- Posição crítica: variável binária (0 = Colaborador não ocupa uma posição crítica; 1 = Colaborador ocupa uma posição crítica)
- Gestor: variável categórica
- Avaliação de desempenho: variável categórica (ordinal) que representa o conceito recebido pelo colaborador na última avaliação de desempenho
- Distância residência trabalho (Km): variável numérica que indica a distância, em Km, entre a residência e o trabalho do colaborador
- Tempo deslocamento (min): variável numérica que indica o tempo de deslocamento entre residência e trabalho, em minutos
- Turnover mercado: variável numérica que corresponde ao valor do turnover de mercado, considerando o cargo do colaborador

A partir dessas informações e com os dados disponibilizados (**questao1.csv**) faça as seguintes atividades (Lembre-se de descrever sua proposta de solução e comentar se você acha que possíveis preditores não foram contemplados):

- Explore o conjunto de dados utilizando estatística univariada e bivariada, bem como gráficos e testes;
- Realize tarefas de aprendizagem de máquina com o objetivo de predizer colaboradores que podem deixar a empresa no próximo ano: variável *target*: **desligamento**.

2. É sabido que o modelo de regressão linear possui determinadas suposições estatísticas: homocedasticidade, autocorrelação, endogeneidade, multicolinearidade e normalidade. Assim, o objetivo desta tarefa é verificar o impacto da violação de tais suposições tanto nos parâmetros quanto na variável respota/target como descrito abaixo:
    - Homocedasticidade, autocorrelação e endogeneidade: simule 10.000 modelos de regressão linear simples ($Y = \beta_0 + \beta_1 X_1 + \varepsilon$), verifique em quantos deles os intervalos de confiança não contém os verdadeiros valores dos parâmetros (definidos como input da simulação). Além disso, teste se a quantidade de intervalos de confiança que não contém os verdadeiros valores dos parâmetros para cada modelo simulado (considere 5% de nível de significância) é significativa.
    - Multicolinearidade: simule duas covariáveis altamente correlacionadas, estime um modelo de regressão linear múltipla com tais variáveis ($Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1 + \varepsilon$). Compare e discuta os resultados obtidos na regressão anterior com os obtidos a partir de duas regressões simples ($Y = \beta_0 + \beta_1 X_1 + \varepsilon$ e $Y = \beta_0 + \beta_2 X_2 + \varepsilon$).
    - Normalidade: simule 10.000 modelos de regressão linear simples ($Y = \beta_0 + \beta_1 X_1 + \varepsilon$) considerando uma amostra de tamanho pequeno e faças as mesmas análises do primeiro item desta questão.