Bootcamp: Analista de Dados para o Mercado Financeiro

Módulo 5: Desafio Final

O principal propósito deste desafio consiste em conduzir todas as etapas fundamentais de um projeto de ciência de dados. Isso engloba desde a coleta até a implementação do algoritmo Random Forest, aplicando-o aos conjuntos de dados relacionados de funcionários deixando a empresa. A meta central dessa iniciativa é realizar a classificação dos funcionários em permaneceu ou não permaneceu na empresa com base em seus atributos, objetivando a obtenção de insights essenciais para a compreensão detalhada do perfil do funcionário.

Objetivos geral do desafio

Exercitar os seguintes conceitos trabalhados no curso:

1. Coleta de dados estruturados;
2. Tratamento, limpeza e processamento de dados;
3. Análise de dados;
4. Visualização de dados;
5. Desenvolvimento de algoritmos de *Machine Learning:*
    a. Random Forest.
6. Práticas de manipulação de dados.

Enunciado

A retenção de talentos é um desafio crucial enfrentado por muitas empresas, especialmente em organizações multinacionais, onde a competição por profissionais qualificados é acirrada. A capacidade de prever a probabilidade de um funcionário deixar a empresa pode proporcionar uma vantagem estratégica significativa, permitindo que os departamentos de recursos humanos identifiquem e implementem medidas proativas para reter os talentos-chave.

Para abordar essa necessidade, propomos o desenvolvimento de um algoritmo baseado em Random Forest, uma técnica poderosa de aprendizado de máquina, para classificar funcionários com base em sua probabilidade de deixar a empresa. Para isso, utilizaremos o conjunto de dados Human Resource Dataset, que contém uma variedade de atributos relevantes para essa análise.

Entre os atributos coletados, destacam-se fatores como nível de satisfação, tempo desde a última avaliação, número de projetos, horas médias trabalhadas mensalmente e tempo dedicado à empresa. Além disso, consideramos se o funcionário esteve envolvido em acidentes de trabalho, se recebeu promoções nos últimos cinco anos, o departamento em que trabalha e sua categoria salarial.

Reconhecer e reter funcionários qualificados é de suma importância para o sucesso de uma organização. Funcionários talentosos não apenas contribuem significativamente para os resultados financeiros da empresa, mas também desempenham um papel crucial na cultura organizacional e no sucesso a longo prazo. A perda desses profissionais pode resultar em custos substanciais, tanto financeiros quanto intangíveis, como perda de conhecimento e experiência, interrupção da equipe e redução da moral.

Para mitigar a saída de funcionários qualificados, é essencial adotar medidas proativas, como oferecer oportunidades de desenvolvimento profissional, reconhecimento e recompensas adequadas, garantir um ambiente de trabalho saudável e promover uma cultura de transparência e comunicação aberta. Além disso, a análise preditiva fornecida pelo algoritmo de Random Forest pode servir como uma ferramenta valiosa para identificar funcionários em risco e implementar estratégias direcionadas de retenção.

Descrição das colunas

Nome da Coluna	Descrição
satisfaction_level	Nível de satisfação do funcionário em porcentagem. 100% ou 1 significa muito satisfeito. 0% ou 0 significa não satisfeito.
last_evaluation	Tempo desde a última avaliação em anos.
number_project	Número de projetos em que o funcionário está trabalhando.
average_montly_hours	Média de horas trabalhadas pelo funcionário nos últimos 3 meses.
time_spend_company	Tempo que o funcionário passa viajando para o escritório.
Work_accident	Se o funcionário esteve envolvido em um acidente de trabalho.
promotion_last_5years	Se o funcionário recebeu uma promoção nos últimos 5 anos.
Department	Departamento em que o funcionário está trabalhando.
salary	Categoria salarial: baixa, média ou alta.
left	Se o funcionário deixou a empresa. 1 significa deixou a empresa. 0 significa permaneceu na empres

Tabela para transformação dos dados para serem utilizadas no algoritmo de Machine Learning.

Departamento	Código
sales	0
accounting	1
hr	2
technical	3
support	4
management	5
IT	6
product_mng	7
marketing	8
RandD	9

Categoria Salarial	Código
low	0
medium	1
high	2

Atividades do enunciado

Os alunos deverão desempenhar as seguintes atividades:

1. Coletar dados do dataset:
    a. Human Resource Dataset.
2. Analisar os dados coletados;
3. Tratar os dados coletados;
4. Avaliar dados ausentes e duplicados;
5. Criar algoritmo de Random Forest;
6. Responder às questões práticas do desafio.

Dicas e Orientações do Professor

1. Analise os dados com cuidado.
2. Elimine os dados duplicados (se necessário).

3. Antes de enviar as respostas, verifique se o gabarito está correto.
4. Atenção no momento de filtrar e corrigir dados (se necessário).
5. Tenha atenção no que pede cada questão.
6. Para o balanceamento dos dados, utilize:
    a. RandomUnderSampler(random_state = 42)
    b. TomekLinks(sampling_strategy='all')
7. Separe 70% dos dados para treinar o algoritmo e 30% para testar.
    a. Utilize o parâmetro de random_state=42.
8. Utilize os seguintes parâmetros para o algoritmo de Random Forest random_state=42, n_estimators=100, max_depth=100
9. O dataset utilizado no trabalho pode ser obtido no link:
    a. https://leandrolessa.com.br/datasets/
10. Abaixo segue as versões utilizadas das bibliotecas nesse trabalho

Para garantir a obtenção dos mesmos resultados do projeto, é recomendável o uso das mesmas versões das bibliotecas

```
VERSÕES UTILIZADAS PARA O DESENVOLVIMENTO DO TRABALHO

PANDAS: 1.5.2
SEABORN: 0.12.1
SKLEARN: 1.2.0
```

É crucial reconhecer que a linguagem de programação Python e suas bibliotecas associadas estão em constante evolução. Como resultado, pode ocorrer que funções ou métodos específicos, que costumavam estar disponíveis em versões anteriores, deixem de existir ou passem a ser implementados de maneira diferente em versões mais recentes.

Essas atualizações são realizadas para melhorar a eficiência, corrigir erros e fornecer novos recursos aos desenvolvedores. No entanto, essa dinâmica de mudança também pode criar desafios, especialmente quando se trabalha com código legado ou ao compartilhar código com outros membros da equipe. Portanto, é de extrema importância que os alunos estejam cientes dessas mudanças e estejam dispostos a se adaptar a elas.

ATENÇÃO PARA TRATAMENTO DE DADOS

Avalie se será necessário realizar tratamento de dados ausentes nos datasets disponibilizados.

Instruções para correção de dados ausentes

1. Dados categóricos:
    a. Utilize a estratégia de correção pela moda referente àquela coluna especifica.
2. Dados numéricos:
    a. Média da coluna analisada.

Acredito no potencial de todos vocês!

Bom desafio a todos!