Bootcamp: Analista de dados com ênfase no Mercado financeiro

Desafio

Módulo 2: Fundamentos em análise e ciência de dados

Objetivos geral do desafio

Exercitar os seguintes conceitos trabalhados no curso:

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

Enunciado

Você é um cientista de dados recém-contratado por uma respeitável corretora de investimentos, recebeu uma tarefa crucial: desenvolver um modelo de Machine Learning para classificar os investidores em diferentes perfis: Arrojado, Moderado e Conservador. Essa classificação é fundamental para o sucesso das operações financeiras, uma vez que permite à corretora compreender as preferências, aversões ao risco e objetivos específicos de cada cliente.

Ao entender o perfil de cada investidor, a corretora pode personalizar suas recomendações e estratégias de investimento de acordo com as necessidades individuais de cada cliente. Os investidores Arrojados, por exemplo, podem estar dispostos a assumir riscos mais elevados em busca de retornos mais altos, enquanto os Conservadores preferem estratégias de baixo risco e estabilidade. Já os Moderados buscam um equilíbrio entre risco e retorno.

Além disso, a classificação do perfil do investidor desempenha um papel crucial na gestão de riscos. Compreender o apetite ao risco de cada cliente permite à corretora ajustar suas estratégias para garantir uma carteira de investimentos equilibrada e alinhada aos objetivos individuais, mitigando possíveis impactos negativos em caso de flutuações no mercado.

Por meio da implementação de um algoritmo de Machine Learning, como o Random Forest, João poderá analisar uma ampla gama de variáveis e padrões de comportamento dos investidores, proporcionando uma visão mais abrangente e precisa do perfil de cada cliente. Isso permitirá à corretora tomar decisões mais informadas, oferecendo produtos e serviços financeiros mais adequados às expectativas e necessidades de seus clientes, fortalecendo assim o relacionamento e a confiança entre a corretora e seus investidores.

Atividades do enunciado

Os alunos deverão desempenhar as seguintes atividades:

1. Coletar dados dos datasets:
    a. dados_clientes_financeira.csv (dados de investidores)
    b. investimentos_clientes.csv (dados de investidores)
    c. estado_regiao.csv (dados de Estados)

2. Analisar os dados coletados;
3. Tratar os dados coletados;
4. Realizar as junções entre as tabelas;
5. Avaliar dados ausentes nas colunas;
6. Criar algoritmo de Random Forest;
7. Responder às questões práticas do desafio.

Dicas e Orientações do Professor

1. Analise com cuidado os dados.
    a. Dados duplicados e ausentes
2. Antes de enviar as respostas, verifique se o gabarito está correto.
3. Realize todas as manipulações e junções dos dados antes de responder às questões do desafio.
    a. Após a junções dos datasets, realize a ordenação baseado no ID do cliente.
4. Atenção no momento de filtrar e corrigir dados (se necessário).
5. Tenha atenção no que pede cada questão.
6. Os dados disponibilizados nos datasets são fictícios. Ou seja, não têm relação com o mundo real.
7. Para o balanceamento dos dados:
    a. RandomUnderSampler(random_state = 42)
    b. TomekLinks(sampling_strategy='all')

8. Separe 80% dos dados para treinar o algoritmo e 20% para testar.
    a. Utilize o parâmetro de random_state=42.
9. Utilize os seguintes parâmetros para o algoritmo de Random Forest random_state=42, n_estimators=100, max_depth=100
10. O dataset utilizado no trabalho pode ser obtido no link:
    a. https://leandrolessa.com.br/datasets/
11. Abaixo segue as versões utilizadas das bibliotecas neste trabalho.
12. Abaixo segue algumas dicas para realizar o trabalho.
    a. https://leandrolessa.com.br/tutoriais/integracao-de-dados-descubra-os-4-tipos-de-joins-essenciais/
    b. https://leandrolessa.com.br/tutoriais/automatizando-coleta-e-extracao-de-arquivos-zip-na-web-com-python/

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.

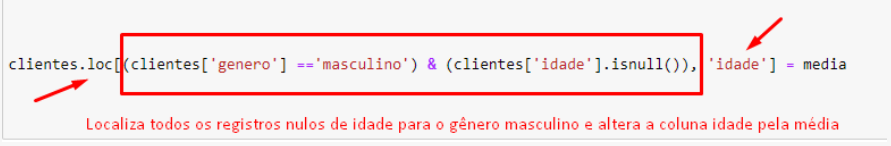
Instruções para correção de dados ausentes

1. Dados categóricos:
    a. Utilize a estratégia de correção pela moda referente aquela coluna especifica.
2. Dados numéricos:
    a. Realize a correção da variável utilizando a média por gênero. Ou seja, realize o filtro do dataset antes de aplicar a média nos dados. Com o resultado obtido, transforme a arredondando a média utilizando o comando round.

```
mediaM.round()
```

Faça a correção de gênero um de cada vez.

Exemplo



Ao lado contém a relação de todos atributos utilizados no modelo, bem como a codificação para as variáveis categóricas.

Tabela dos códigos para os Atributos Categóricos

Coluna	Atributo	Código
genero	Feminino	0
genero	Masculino	1
estado_civil	Casado	0
estado_civil	Divorciado	1
estado_civil	Solteiro	2
estado_civil	Viúvo	3
escolaridade	Ensino Médio	0
escolaridade	Graduação	1
escolaridade	Pós-graduação	2
regiao	Centro-Oeste	0
regiao	Nordeste	1
regiao	Norte	2
regiao	Sudeste	3
regiao	Sul	4

```
atributos_modelo = ['genero_codificado','estado_civil_codificado', 'escolaridade_codificado',
                    'regiao_codificado', 'idade', 'valor_investido', 'valor_investido_acoes',
                    'valor_investido_fii', 'valor_investido_dolar','valor_investido_renda_fixa', 'perfil']
```

Por favor, siga rigorosamente todos os passos e orientações descritos neste documento. O resultado final está diretamente vinculado à execução precisa de cada etapa.

Acredito no potencial de todos vocês!

Bom desafio a todos!