Iniciado em	sábado, 12 dez 2020, 14:03
Estado	Finalizada
Concluída em	sábado, 12 dez 2020, 18:00
Tempo empregado	3 horas 56 minutos

Questão **1**

Completo

Vale 0,50 ponto(s).

Questão 1.1 - O que é overfitting e underfitting? Quando ocorrem?

Overfitting: é quando o modelo treinado apresenta uma alta variância e uma grande performance nos resultados de treinamento, mas performance ruim em novos dados. Esse problema ocorre quando o modelo treinado "decora" todo o espaço de distribuição do conjunto de treinamento, mas quando é avaliado em um novo conjunto de dados o modelo tenta aplicar os mesmos pesos treinados sobre o novo conjunto de dados, tendo assim uma performance ruim, algo que pode ser causado por ter sido utilizado uma quantidade insuficiente de exemplos durante o treinamento do modelo.

Underfitting: é quando o modelo apresenta um alto viés e uma performance ruim nos resultados de treinamento. Esse problema ocorre quando o modelo não é capaz de aprender as nuances dos dados de forma a generalizar um modelo que apresente uma predição confiável, algo que pode ter sido causado por uma escolha incorreta dos dados a serem utilizados no treinamento do modelo.

Questão **2**

Completo

Vale 1,00 ponto(s).

Questão 1.2 - Referente à questão 1.1, responda:

Explique como evitar a ocorrência desses fenômenos nos dados. Descreva um exemplo.

Overfitting: Caso haja poucos dados a serem utilizados no treinamento do modelo uma alternativa é adquirir novos dados ou aplicar métodos de amostragem, de forma a gerar novos dados aleatórios que possam complementar os dados do treinamento. Dentre os métodos disponíveis podemos mencionar k-Fold Cross-Validation que irá selecionar amostras diferentes durante o treinamento do modelo.

Underfitting: Caso os dados disponíveis possuam muitas variáveis diferentes se faz necessário escolher corretamente quais variáveis são realmente importantes para o modelo, dessa forma pode ser aplicado o método Principal Component Analysis - PCA para a identificação das variáveis que apresentam uma alta correlação e importância para o modelo.

Questão **3**

Completo

Vale 1,00 ponto(s).

Questão 1.3 - Imagine que você tem um conjunto de dados com 90% das observações na classe A e 10% na classe B. Explique como você faria a classificação de forma a evitar a predominância da classe A nos resultados.

Dentre as opções disponíveis para o pré-processamento dos dados é possível aplicar as técnicas:

- Undersampling: que consiste em reduzir a classe A excluindo dados de forma aleatória, de forma a diminuir a discrepância entre as classes;

- Oversampling: que consiste em gerar novos casos para a classe B a partir dos dados existentes, de forma a diminuir a discrepância entre as classes.

Ambas as técnicas podem ser aplicadas através do método SMOTE que apresenta formas claras de como realizar essas operações.

Questão **4**

Completo

Vale 1,50 ponto(s).

Questão 2.1 - O arquivo contratacaocorona-27-07-acertado.csv contém informações de compras emergenciais ligadas a COVID 19.

1. Leia o arquivo e considere apenas as colunas 'QUANTIDADE', 'VALOR_UNITM-ARIO' e 'VALOR_TOTAL'. Verifique quantos dados faltantes existem no DataFrame resultante.

2. Remova as linhas do data frame que contenham dados faltantes. Verifique quantas linhas foram removidas.

3. Prepare a coluna 'VALOR_TOTAL' para ser processada como numérica, e a seguir busque por outliers presentes nesta coluna. Para isso use o método do desvio padrão com $\sigma = 3$. Verifique o número de outliers encontrados.

Com base nos itens acima, assinale a alternativa correta:

Escolha uma opção:

- a. Dados faltantes: 30, Linhas removidas: 30, Outliers encontrados: 2
- b. Dados faltantes: 12, Linhas removidas: 12, Outliers encontrados: 4
- c. Dados faltantes: 30, Linhas removidas: 24, Outliers encontrados: 6
- d. Dados faltantes: 30, Linhas removidas: 12, Outliers encontrados: 4
- e. Dados faltantes: 12, Linhas removidas: 12, Outliers encontrados: 6

Questão **5**

Completo

Vale 1,00 ponto(s).

Questão 2.2 -

Ao concluir a questão e assinalar a resposta que julgue correta no Moodle, você ainda precisa realizar os seguintes passos para concluir a submissão da resposta:

1) Exportar o notebook que utilizou para resolver a questão da prova em formato .py e fazer upload no Moodle. Atenção: você **não** deve fazer upload de um arquivo notebook (.ipynb), mas sim um arquivo texto .py contendo os códigos python que utilizou para resolver as questões. O arquivo .py pode ser gerado através da opção:

File --> Download as --> Python (.py) disponível no Jupyter Notebook.
ou
File --> Download .py no Google Colab

Caso não esteja utilizando o Jupyter, copie e cole seu código em um arquivo ASCII (Texto) salvando com a extensão .py

2) O arquivo deve ser nomeado com seu nome e sobrenome, SEM ESPAÇO entre as partes de seu nome. Exemplo: moacirponti.py

3) É OBRIGATÓRIO conter no cabeçalho (início) do arquivo um comentário com o seu nome completo. Por exemplo, # Moacir Ponti

O arquivo submetido será verificado por plágio.

guilhermelourenco.py

Questão **6**

Completo

Vale 2,50 ponto(s).

Questão 3.1 - Instruções gerais: Para esta questão, utilize os comandos disponíveis no Jupyter notebook neste **link** e os dados deste **link**. Faça uma cópia do notebook na sua máquina, você deverá fazer o upload do arquivo .ipynb com os comandos utilizados ao final da questão. Use duas casas decimais para as respostas.

No aniversário de 20 anos do programa de estímulo à produção de grãos, realizou-se um estudo para quantificar seu impacto.

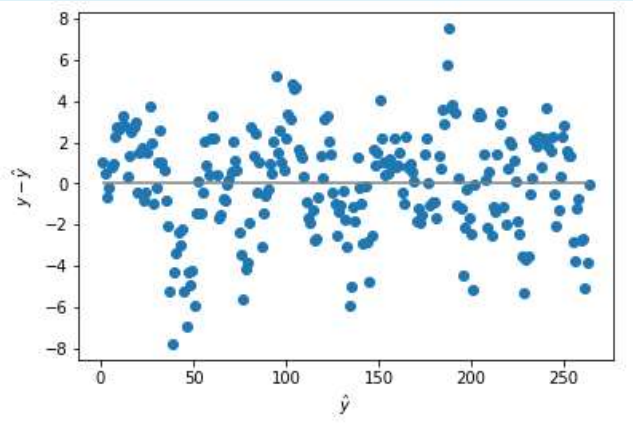
Uma comparação das 10 primeiras décadas do programa com as 10 últimas décadas detectou um aumento importante da produção, com média mensal de [66.464798] toneladas, para a década de 80, e [198.144759] toneladas, para a década de 90 (teste t de Student para amostras independentes, supondo-se variâncias equivalentes: p-valor<0,001). Nota-se que a variabilidade na produção mensal, porém, manteve-se equivalente em ambas as décadas (teste de Levene: p-valor=[0.796719]).

Graficamente, observa-se um aumento da produtividade y ao longo dos anos que, numa análise preliminar, pode ser evidenciada pela reta de regressão linear simples, dada por:

y.chapeu= [1.1618] +[1.0974] X.

Resumidamente, pode-se dizer que houve um aumento de aproximadamente [9,74] % na produção a cada mês. Por exemplo, no mês de [JANEIRO] de 1990, a produção predita é de 132,85 toneladas.

Porém, a análise de regressão linear simples não é adequada para tal estudo, o que pode ser evidenciado pelo gráfico de [resíduos] abaixo:



Note que as observações não estão aleatoriamente dispersas em torno de zero, refletindo um padrão que sugere a existência de [CORRELAÇÃO] positiva entre observações consecutivas.

Ao escolher um modelo ARIMA para os resíduos, observou-se a ordem [1],[0],[0], o que [CONFIRMA] (confirma/contraria) os gráficos de autocorrelação e autocorrelação parcial dos resíduos.

A base de dados foi dividida em bases de treino e [TESTE], com 80% das observações para a base de treino, ou seja, [192] observações. Com a base de treino, ajustou-se um modelo ARIMA para a produção, e o melhor modelo obtido via stepwise foi o ARIMA [0],[1],[1]. Isso significa que o modelo escolhido possui componentes [INTEGRADOS] e de média móvel.

Calculando-se então as previsões para a base de teste, a raiz quadrática do erro quadrático médio, RMSE, obtém-se [31.90975].

Questão **7**

Completo

Não avaliada

Questão 3.2 - **Obrigatório** - Faça *upload* do arquivo de notebook.ipynb para conferência posterior, nos casos em que as correções automatizadas configuradas no sistema não levarem em consideração diferentes versões do mesmo método estatístico implementadas em Python.

guilhermelourenco.ipynb

Questão **8**

Completo

Vale 1,50 ponto(s).

Questão 4.1 -

1. Projete um **autoencoder**.
2. Compile e treine esse **autoencoder** com dados de 2016.
3. Compute o **erro quadrático da reconstrução** dos dados de 2017 e, a partir desse, obtenha a chave primária do funcionário (ou seja, **funcPK**) cujos dados possuem o maior erro. O funcionário a ser inspecionado é o funcionário identificado por essa **funcPK**.
4. Faça uma consulta que tem como objetivo investigar **os meses e os valores dos salários recebidos no ano de 2017 pelo funcionário que está sendo inspecionado**.

Selecione a alternativa que identifica qual a **chave primária** obtida após o Item 3, qual o **nome do(a) funcionário(a)**, qual o **maior valor de salário** recebido pelo(a) funcionário(a) e qual o **mês** do ano de 2017 no qual o(a) funcionário(a) recebeu esse maior salário.

Escolha uma opção:

- a. funcPK = 64, nome = ABILIO BARBOSA; maior salário = 4420.97; mês = 1
- b. funcPK = 147, nome = ABILIO BARBOSA; maior salário = 250000.00; mês = 7
- c. funcPK = 147, nome = ABIMAELE BORGES; maior salário = 250000.00; mês = 7
- d. funcPK = 147, nome = ABIMAELE BORGES; maior salário = 1559.94; mês = 1
- e. funcPK = 64, nome = ABILIO BARBOSA; maior salário = 250000.00; mês = 9

Questão **9**

Não respondido

Vale 1,00 ponto(s).

Questão 4.2 - Além de responder a questão no Moodle você deve:

exportar esse notebook com sua solução para as questões da prova em formato .py e fazer upload no Moodle. Atenção: você **não** deve fazer upload de um arquivo notebook
(.ipynb), mas sim um arquivo texto .py contendo os códigos python que utilizou para resolver as questões.

O arquivo .py pode ser gerado através da opção:

File --> Download as --> Python (.py) disponível no Jupyter Notebook.
ou File --> Download .py no Google Colab

Caso não esteja utilizando o Jupyter, copie e cole seu código em um arquivo ASCII (Texto) salvando com a extensão .py

2) Você deve salvar esse notebook com sua solução para as questões da prova em formato .pdf e fazer upload no Moodle

3) Os arquivos devem ser nomeados com seu nome e sobrenome, sem espaços. Exemplo: moacirponti.py e moacirponti.pdf

4) É OBRIGATÓRIO conter no cabeçalho (início) do arquivo um comentário / texto com o seu nome completo

Desejamos uma boa prova!

◄ Material de apoio

Seguir para...