**Projeto da disciplina**

**Ciência de Dados e Inteligência Artificial**

**Fase 2**

|  |  |
| --- | --- |
| **Nome do estudante** | **Daiane Deponti Bolzan** |

Desenvolva um processo de ciência de dados no Orange Data Mining, cobrindo os elementos abaixo. **Para cada um dos itens solicitados é necessário inserir imagens que evidenciem o trabalho realizado**.

|  |
| --- |
| Exploração dos dados |
| *Que tipo de experimentos você fez na exploração dos dados (verificação de outliers, cálculos de médias, dados inválidos etc.).* |
| A proposta de trabalho é verificar se **a pandemia de covid** teve algum impacto significativo nas taxas de câmbio brasileiro. Para isso, criei uma variável target chamada Pandemia com valores binários (0 e 1) para identificar o período considerando este como de **11 de março de 2020 a 05 de maio de 2023** que é o período em que a OMS classificou como Emergência de Saúde Pública de Importância Internacional (ESPII).  Os dados estão divididos em dois arquivos, nos quais foi preciso usar o Edit Domain para definir os dados, especialmente a coluna com as datas. |
| Depois disso, os arquivos com os dados foram mergeados com base na data:    Para verificar as estatísticas utilizou-se Feature Statistics:    E para visualização dos dados usamos Data Info:    Para analisar outlier, utilizou-se o Box Plot:    Também utilizei o Select Rows para tratar dados faltantes, desta forma, eles foram removidos da tabela. |

|  |
| --- |
| Escolha de, ao menos, três algoritmos de aprendizado para a modelagem |
| *Apresente os algoritmos utilizados e justifique a escolha.* |
| Vamos testar três modelos para prever a taxa de câmbio futura com base nos dados históricos:   * **Regressão Logística** → Para identificar modificações entre as taxas de câmbio no período. * **Random Forest** → Para identificar quais moedas mais contribuíram para diferenciar os períodos. * **Naive Bayes** → Classificador probabilístico baseado em **probabilidades condicionais** |

|  |
| --- |
| Preparação dos dados de acordo com as características dos algoritmos de aprendizado escolhidos |
| *Descreva o processo realizado para essa etapa.* |
| Além de adicionar as linhas e colunas corretas inclui um pre-processador para remover missing data. A variável target foi definida como a variável pandemia que é uma variável que assume valores binários de 0 e 1, sendo 1 para o período da pandemia, buscando assim identificar se houve ou não alguma alteração significativa nas taxas de câmbio durante esse período. |
| Na seleção das rows, apenas especifiquei que os valores deveriam ser definidos, excluindo assim valores faltantes:  E acrescentei um pre-processador para caso ainda assim algum valor faltante tenha passado, que fosse ajustado antes de chegar aos modelos de regressão logística e Random forest.    No caso do Naive Bayes, não é necessário utilizar pre-processador. |

|  |
| --- |
| Execução dos experimentos de aprendizado e coleta das métricas |
| *Descreva o processo e os resultados obtidos.*  O primeiro algoritmo aplicado foi o de regressão logarítmica que apresentou os seguintes coeficientes: |
| Apesar de eu ter tido bastante dificuldade com a ferramenta, ao final consegui conectar corretamente o data sampler com os algoritmos e os learners do test and score e obter os resultados: |
| Gerei também a matriz de confusão e a análise ROC: |
| Para explorar os padrões presentes nos dados e observar possíveis mudanças no comportamento das variáveis durante o período da pandemia, foi utilizado o algoritmo **Naive Bayes**. O modelo apresentou um desempenho **satisfatório**, com boa taxa de acerto geral, especialmente na classificação da classe majoritária (classe 0).  A matriz de confusão revelou os seguintes resultados:   * **Verdadeiros Negativos (classe 0 corretamente classificada):** 3519 * **Falsos Positivos (classe 0 classificada como 1):** 532 * **Falsos Negativos (classe 1 classificada como 0):** 0 * **Verdadeiros Positivos (classe 1 corretamente classificada):** 553   Esses números indicam que o modelo foi bastante eficaz ao classificar corretamente a classe 1 (presumivelmente associada ao período **durante a pandemia**), com **nenhum falso negativo** — ou seja, **todas as instâncias reais da classe 1 foram corretamente identificadas**. Por outro lado, houve uma taxa de erro mais expressiva ao identificar a classe 0, com 532 casos incorretamente atribuídos como sendo da classe 1.  Esse comportamento pode ser interpretado como uma **tendência do modelo em superestimar a classe 1**, o que sugere que características associadas ao período da pandemia tiveram um peso considerável nas distribuições probabilísticas utilizadas pelo Naive Bayes. Essa diferença nos padrões reforça a hipótese de que **houve, sim, uma mudança significativa no comportamento das variáveis analisadas durante a pandemia**, especialmente em relação às taxas de câmbio.  Além disso, como o modelo fornece as probabilidades associadas a cada classe, essa abordagem também pode ser útil para analisar **o grau de certeza das previsões**, permitindo uma análise mais refinada em aplicações futuras. |

|  |
| --- |
| Relato dos experimentos e lições aprendidas |
| *Apresente uma reflexão acerca dos resultados obtidos com este projeto.* |
| Os resultados obtidos no **Test and Score** foram os seguintes:   |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | Modelo | AUC | CA (Accuracy) | F1 | Precision | Recall | MCC | | **Random Forest** | 1.000 | 0.997 | 0.997 | 0.997 | 0.997 | 0.985 | | **Logistic Regression** | 0.997 | 0.992 | 0.992 | 0.992 | 0.992 | 0.961 | | **Naive Bayes** | 0.963 | 0.884 | 0.899 | 0.884 | 0.884 | 0.665 |   Comparando-se os modelos utilizando **AUC**, o modelo de Random Forest é o melhor para descrever os dados. Mas o desempenho dos três modelos foi bastante significativo indicando que os modelos conseguem prever com alta precisão se um registro pertence ao período da pandemia com base nas taxas de câmbio.  **O primeiro achado importante é que houve sim mudança nas taxas de câmbio durante o período da pandemia.**  Analisando os coeficientes da **regressão logística**, observa-se os seguintes resultados:   |  |  |  | | --- | --- | --- | | Moeda / Cotação | Coef. | Interpretação | | Intercepto | -23.91 | Base do modelo; desloca a curva de decisão. | | Dólar americano (venda) | -0.087 | Impacto negativo leve – aumento dessa taxa reduz levemente a chance da classe 1. | | Euro (ask) | 2.87 | Forte impacto positivo – aumento favorece classe 1. | | Euro (bid) | 2.85 | Similar ao ask – alta influência positiva. | | Iene japonês (ask) | 0.53 | Pequeno impacto positivo. | | Iene japonês (bid) | 0.53 | Mesmo impacto do ask – consistente. | | Libra esterlina (ask) | -4.70 | Forte impacto negativo – aumento reduz chance da classe 1. | | Libra esterlina (bid) | -4.70 | Consistente com o ask – alta influência negativa. | | Dólar australiano (ask) | 3.91 | Muito forte impacto positivo – aumento favorece classe 1. | | Dólar australiano (bid) | 3.91 | Igual ao ask – influência muito forte. | | Dólar canadense (ask) | 4.55 | Impacto muito alto positivo – grande influência. | | Dólar canadense (bid) | 4.55 | Também muito relevante. | | Dólar americano (compra) | -0.0917 | Impacto negativo leve, similar ao dólar venda. |   De um modo geral, as moedas como maior influência positiva foram o dólar canadense, dólar australiano e euro, indicando que o R$ se valorizou frente a essas moedas e as moedas com maior influência negativa foram o dólar americano e a libra esterlina indicando que o R$ se desvalorizou frente a estas moedas.  Com o resultado do **Random Forest** é possível identificar quais as variáveis mais importantes para o modelo.     1. **Exchange rate - Free - Japanese Yen (ask)** → **0.181** 2. **Exchange rate - Free - Australian Dollar (bid)** → **0.309** 3. **Exchange rate - Free - Japanese Yen (bid)** → **0.116** 4. **Exchange rate - Free - Euro (bid/ask)** → também contribuíram, mas com valores menores. 5. **Exchange rate - Free - Canadian Dollar (ask/bid)** → entre 0.075 e 0.010. 6. **Exchange rate - Free - U.S. Dollar (buy/sale)** → contribuíram **menos**, especialmente o de compra (buy): **0.092**, e o de venda: **0.026**. 7. **Libra esterlina (British Pound)** → teve **baixa influência**, abaixo de 0.05.   Isso confirma a hipótese anterior de que as moedas com maior influência positiva foram o dólar canadense, dólar australiano e euro. Por outro lado, o dólar americano teve menor importância no modelo – isso pode indicar que o modelo considerou essa moeda menos útil para prever o desfecho, talvez por já estar embutida no comportamento geral do câmbio (já que o dólar é moeda de referência global).  Os resultados obtidos com o modelo de Random Forest sugerem que **houve, sim, mudanças relevantes na dinâmica cambial durante o período da pandemia de COVID-19**. A alteração na importância relativa das moedas na explicação da variação da taxa de câmbio brasileira pode indicar que diferentes moedas responderam de formas distintas aos efeitos econômicos e políticos provocados pela pandemia.  Por exemplo, moedas como o **dólar australiano, euro e iene japonês** apresentaram maior influência sobre o comportamento da taxa de câmbio brasileira, o que pode refletir **mudanças nos fluxos de comércio, investimentos internacionais e políticas monetárias específicas desses países**. Isso contrasta com o comportamento de moedas tradicionalmente mais influentes, como o **dólar americano e a libra esterlina**, que tiveram menor importância no modelo, possivelmente indicando uma **diminuição momentânea de sua correlação com a moeda brasileira**.  Assim, os dados analisados e a distribuição da importância das variáveis sugerem que **a pandemia teve impacto na estrutura de relações cambiais do Brasil com o exterior**, refletindo uma **mudança de padrão na sensibilidade do real frente a diferentes moedas estrangeiras**. |

|  |
| --- |
| LINK PARA O ARQUIVO DO PROJETO DO ORANGE E DOS DADOS UTILIZADOS |
| *Insira os links para os arquivos.* |
| O arquivo está neste arquivo zipado |