Documentação do Processo de Construção e Treinamento do Modelo

Introdução

Este documento fornece uma visão detalhada do processo de construção e treinamento do modelo de aprendizado de máquina para previsão de demanda de produtos. Descreve as etapas, parâmetros selecionados e os resultados obtidos durante o desenvolvimento do modelo.

Objetivo

O objetivo principal deste modelo é prever a demanda futura de produtos com base em dados históricos de vendas e características do produto. Isso permitirá uma melhor gestão de estoque, garantindo que a empresa possa atender à demanda sem excesso de estoque.

Etapas do Processo

1. Exploração de Dados e Pré-processamento

Coleta de Dados

- **Fontes de Dados**: Os dados foram coletados do banco de dados MySQL da empresa, especificamente das tabelas **`estoque`** e **`vendas`**.
- **Variáveis/Features Incluídas**:
  - **`produto_id`**: Identificador do produto.
  - **`quantidade_em_estoque`**: Quantidade de produtos disponíveis em estoque.
  - **`data_venda`**: Data da venda.
  - **`quantidade_vendida`**: Quantidade de produtos vendidos.

Limpeza e Pré-processamento

- **Identificação e Tratamento de Valores Ausentes**: Não foram encontrados valores ausentes significativos nos dados coletados.
- **Tratamento de Outliers**: Outliers foram identificados e analisados, mas não foram removidos inicialmente devido ao impacto potencial sobre a precisão das previsões.
- **Transformações Aplicadas aos Dados**:
  - Conversão da coluna **`data_venda`** para o tipo datetime.
  - Extração de características adicionais a partir da data da venda: **`dia_da_semana`**, **`mes`** e **`ano`**.

2. Implementação de Modelos de Aprendizado de Máquina

Escolha de Algoritmos

- **Justificativa**: O algoritmo Random Forest foi escolhido devido à sua robustez e capacidade de lidar com datasets com muitas variáveis, além de ser menos propenso ao overfitting em comparação com outros modelos.

Implementação

- **Detalhes da Implementação**: O modelo foi implementado utilizando a biblioteca Scikit-learn.
- **Bibliotecas Utilizadas**:
    - Pandas para manipulação de dados.
    - Numpy para operações matemáticas.
    - Scikit-learn para construção e avaliação do modelo.
    - SQLAlchemy e MySQL Connector para integração com o banco de dados.

3. Otimização e Validação do Modelo

Otimização de Hiperparâmetros

- **Processo de Otimização**: Utilização do GridSearchCV para otimização dos hiperparâmetros do modelo.
- **Hiperparâmetros Ajustados**:
    - `n_estimators`: Número de árvores na floresta.
    - `max_features`: Número de características a serem consideradas para encontrar a melhor divisão.
    - `max_depth`: Profundidade máxima das árvores.
    - `min_samples_split`: Número mínimo de amostras necessárias para dividir um nó interno.
    - `min_samples_leaf`: Número mínimo de amostras necessárias em um nó folha.

Validação Cruzada

- **Método de Validação**: Validação cruzada com 5 folds foi utilizada para avaliar o desempenho do modelo.
- **Resultados da Validação Cruzada**:
    - A média do R2 Score nas validações foi usada para selecionar o melhor conjunto de hiperparâmetros.

Parâmetros do Modelo

- **Hiperparâmetros Finais**:
    - `n_estimators`: 200
    - `max_features`: 'sqrt'
    - `max_depth`: 20

    - `min_samples_split`: 5
    - `min_samples_leaf`: 2
- **Outros Parâmetros Relevantes**: Random state foi definido como 42 para garantir a reprodutibilidade dos resultados.

Métricas de Avaliação

- **Descrição das Métricas Utilizadas**:
    - `Mean Absolute Error (MAE)`: Média dos erros absolutos entre as previsões e os valores reais.
    - `Mean Squared Error (MSE)`: Média dos quadrados dos erros entre as previsões e os valores reais.
    - `R2 Score`: Coeficiente de determinação que indica a proporção da variância nos dados de saída que é previsível a partir dos dados de entrada.
- **Resultados Específicos**:
    - Erro Médio Absoluto (MAE): 3.45
    - Erro Médio Quadrático (MSE): 18.76
    - R2 Score: 0.82

Resultados e Conclusão

O modelo de Random Forest otimizado apresentou um bom desempenho com um R2 Score de 0.82, indicando que 82% da variação na demanda de produtos pode ser explicada pelas características incluídas no modelo. A previsão da demanda futura para um produto específico mostrou uma média de demanda prevista de 5.67 unidades por dia, com um estoque sugerido de 39.69 unidades para uma semana de segurança. Esses resultados mostram que o modelo pode ser uma ferramenta útil para a gestão de estoque, ajudando a empresa a manter níveis adequados de produtos para atender à demanda sem excessos.