CET0621 – Aprendizado de Máquina na Análise de Dados

Proposta de projeto

Nome do Aluno 1: Gustavo Ferreira Lima RA: 2023611300

Nome do Aluno 2: Mateus de Almeida Frigo RA: 2023611431

Proposta de Projeto: Previsão de Demanda para Categorias Chave de Produtos no Varejo Alimentício

1. **Descrição do Problema:**

O setor de varejo alimentício opera com margens frequentemente apertadas e lida com o desafio constante da gestão de estoques, especialmente para produtos perecíveis ou de alta rotatividade. Uma previsão de demanda inadequada pode levar a dois problemas principais: excesso de estoque, resultando em custos de armazenagem, aumento de perdas por vencimento e necessidade de promoções agressivas para escoamento; e falta de estoque (ruptura), que implica em vendas perdidas, insatisfação dos clientes e potencial migração para concorrentes. A capacidade de antecipar com precisão a quantidade de produtos que os consumidores irão adquirir em um período futuro é, portanto, crucial para a eficiência operacional e rentabilidade do negócio.

Este projeto se propõe a investigar e aplicar modelos de aprendizado de máquina para realizar a previsão de demanda para categorias chave de produtos em um contexto de varejo alimentício. Esta tarefa é um problema clássico de estimação (regressão), um dos pilares do aprendizado supervisionado estudado na disciplina.

2. **Objetivos do Trabalho:** O presente trabalho tem como objetivos principais:

Seleção e Preparação de Dados: Identificar e selecionar um dataset público adequado (e.g., Kaggle) que contenha dados históricos de vendas de produtos em um ambiente de varejo, preferencialmente com granularidade que permita a análise por categorias de produtos alimentícios. Realizar as etapas de pré-processamento necessárias, como limpeza de dados, tratamento de valores ausentes, transformação de variáveis e agregação dos dados para o nível de análise desejado (ex: demanda diária ou semanal por categoria).

Engenharia de Atributos (Features): Desenvolver um conjunto de atributos informativos a partir dos dados brutos, que possam influenciar a demanda. Isso incluirá atributos temporais (dia da semana, mês, ano, indicadores de feriados ou datas especiais), atributos baseados em lags de vendas (demanda em períodos anteriores) e, possivelmente, estatísticas de janelas móveis (média de vendas recentes).

Desenvolvimento e Treinamento de Modelos Preditivos: Implementar, treinar e avaliar o desempenho de diferentes algoritmos de aprendizado de máquina para a tarefa de estimação. Serão explorados modelos como:

Regressão Linear (como baseline).

Modelos baseados em árvores de decisão (ex: Decision Tree Regressor, Random Forest Regressor).

Modelos de ensemble mais robustos (ex: Gradient Boosting Regressor).

Se pertinente e o tempo permitir, Redes Neurais Artificiais (MLP) para regressão.

Avaliação e Comparação de Modelos: Utilizar métricas de avaliação apropriadas para problemas de regressão, como Erro Médio Absoluto (MAE), Raiz do Erro Quadrático Médio (RMSE) e Coeficiente de Determinação (R^2), para comparar o desempenho preditivo dos diferentes modelos.

Análise e Interpretação dos Resultados: Analisar criticamente os resultados obtidos, identificar os modelos com melhor desempenho, discutir a importância relativa dos diferentes atributos para a previsão da demanda e apresentar as conclusões do estudo, incluindo limitações e possíveis trabalhos futuros.

3. **Justificativa para sua Realização:**

A previsão de demanda é um problema de grande relevância prática e econômica para o setor de varejo alimentício. Uma solução eficaz baseada em aprendizado de máquina pode

auxiliar as empresas a otimizarem a gestão de seus estoques, reduzir desperdícios (especialmente de produtos perecíveis), melhorar a disponibilidade de produtos nas gôndolas, e, consequentemente, aumentar a satisfação dos clientes e a lucratividade.

As técnicas de estimação vistas na disciplina (Tópico03a) são diretamente aplicáveis. A escolha dos modelos se baseia em:

Regressão Linear: Fornece um modelo simples e interpretável, útil como ponto de partida e comparação.

Modelos baseados em Árvores (Decision Trees, Random Forest): São capazes de capturar relações não-lineares complexas nos dados e são robustos a outliers. Random Forest, sendo um método de ensemble (conceito visto na aula 2 - Ensembles), frequentemente oferece melhor desempenho e generalização do que árvores individuais ao combinar múltiplas árvores treinadas em subconjuntos dos dados.

Gradient Boosting: Outra técnica de ensemble poderosa que constrói modelos de forma sequencial, onde cada novo modelo tenta corrigir os erros dos anteriores, geralmente levando a alta precisão.

Redes Neurais (MLP): Conforme visto ena aula 2 - Perceptrons Multicamadas e aula 3 - Estimação de Dados, MLPs são aproximadores universais de funções e podem modelar padrões complexos, sendo uma alternativa não-paramétrica interessante para problemas de estimação.

Este projeto permitirá ao grupo aplicar de forma integrada os conceitos de pré-processamento de dados, engenharia de features, seleção e treinamento de modelos, e avaliação de desempenho, cobrindo as principais etapas do processo de descoberta de conhecimento em dados (KDD) conforme discutido na disciplina.

4. Referências (Preliminares) Datasets Potenciais (Kaggle):

Store Item Demand Forecasting Challenge: https://www.kaggle.com/competitions/store-item-demand-forecasting

Corporación Favorita Grocery Sales Forecasting: https://www.kaggle.com/competitions/favorita-grocery-sales-forecasting

Walmart Recruiting - Store Sales Forecasting: https://www.kaggle.com/competitions/walmart-recruiting-store-sales-forecasting

Materiais da Disciplina CET0621 (Prof. Guilherme Palermo Coelho):

Coelho, G. P. (2025). Tópico02b: Classificação de Dados - Parte 2 (Perceptrons Multicamadas). Unicamp.

Coelho, G. P. (2025). Tópico02c: Classificação de Dados - Parte 3 (Ensembles). Unicamp.

Coelho, G. P. (2025). Tópico03a: Estimação de Dados. Unicamp.