Artigo de conclusão da Formação em Cientista de Dados

Carlos Eduardo Lopes

[Formação em Cientista de Dados-T07-Mai](http://harveworkplace.com.br/course/view.php?id=33" \o "Formação em Cientista de Dados-T07-Mai)

# Resumo

A proposta de trabalho final para conclusão da formação em cientista de dados é criar um modelo preditivo com o Python, e realizar a previsão para os próximos períodos e visualizá-lo através de ferramentas de Business Inteligence.

# Objetivo

Desenvolver modelos preditivos analisando várias hipóteses de modelos escolhendo aquele com melhor desempenho para utilização dentro do BI.

# Execução do projeto

Para esse projeto inicialmente foi realizada uma busca na base de dados da [www.kaggle.com](http://www.kaggle.com), a intensão deste projeto desde o início era realizar um modelo preditivo de vendas. Encontrada uma base de dados de uma empresa de varejo foi possível iniciar o processo de modelagem.

Foi identificada que a base de dados possuía os seguintes parâmetros: data da venda, preço, volume de estoque do dia e volume de vendas. Após esse processo iniciou-se o processo de ETL. A base de dados foi carregada através da biblioteca pandas, para a modelagem foi realizado os seguintes ajustes, conversão do campo ‘data’ que estava como objeto para data, e retirada os campos que era string para o treinamento do modelo.

Com a biblioteca scikit-learn foi separado os dados entre treino e teste para realizarmos a execução dos modelos preditivos e testes dos mesmos para avaliarmos sua performance.

Ainda usando a biblioteca citada realizamos três análises com os dados, através de regressão linear, regressão não linear (polinomial de graus), arvore de decisão de regressão, Random forest regression e redes neurais MPL. Após avaliar os 5 modelos vimos que nenhum deles obteve score satisfatório para resolver o problema.

Ainda com o resultado ruim para fins didáticos, após esse processo foi carregado um data frame com duas semana simulando os parâmetros para previsão de duas semana. Com esses dados foi consolidado os dados históricos com os dados da previsão e exportado em csv finalizando o processo de ETL.

Essa nova base de dados alimentou um dashboard criado no power bi finalizando o projeto gerando a visualização desses dados.

# Conclusões

Com os dados foi possível realizar todo o processo para os modelos preditivos e criticá-los para a entrega final. Creio que em um processo no dia a dia seria interessante agregar novas variáveis para deixar o modelo melhor como por exemplo desconto, quantidade de itens promocionados, datas comemorativas como black fryday, agregariam ao processo. Testar também o dataset em outros modelos como os de series temporais.

Ao final carregar os dados diretamente em um banco de dados SQL deixariam o processo mais robusto.

# Referências e Notebooks.

Informe os links de notebooks, blogs e livros que foram utilizados. Exemplos abaixo.

*Notebook do Projeto, disponível em* [*https://raw.githubusercontent.com/CarlosLopes88/projeto\_prev\_vendas/main/%5BHarve%5D\_FDS\_MODELO\_PROJETO\_FINAL\_ARTIGO.ipynb*](https://raw.githubusercontent.com/CarlosLopes88/projeto_prev_vendas/main/%5BHarve%5D_FDS_MODELO_PROJETO_FINAL_ARTIGO.ipynb)

*Base de dados utilizada disponível www.site.com.brhttps://www.kaggle.com/tevecsystems/retail-sales-forecasting*