

Atualização do Projeto: Modelagem

No código, além da parte de treinamento do modelo de regressão, algumas etapas importantes foram feitas para garantir que os dados fossem adequados e específicos para o que estamos tentando prever: o valor de exportação (VL_FOB) de produtos, com foco no agronegócio.

1. Dicionários de Expansão

Primeiramente, criamos três dicionários de expansão para facilitar a interpretação dos dados e tornar a análise mais rica. Esses dicionários funcionam como uma forma de traduzir os códigos presentes nos dados em informações mais compreensíveis:

- **dict_country.csv**: Este arquivo contém um mapeamento entre os códigos dos países (como os códigos numéricos usados no sistema de dados) e seus respectivos nomes. A ideia aqui é que, ao invés de trabalhar diretamente com códigos numéricos, a gente consiga associar um nome de país, o que ajuda muito na hora de interpretar os resultados ou gerar relatórios mais amigáveis.
- **dict_ncm_product.csv**: Similar ao dicionário de países, esse arquivo faz o mapeamento entre os códigos NCM (Nomenclatura Comum do Mercosul) e a descrição dos produtos. O código NCM é como uma "identidade" para cada produto no comércio exterior, então, traduzir isso para um nome de produto torna a análise mais clara e fácil de entender.
- **dict_urf.csv**: Esse dicionário mapeia os códigos de URF (Unidade da Receita Federal) com seus respectivos nomes. Isso é importante, porque as URFs são responsáveis pela fiscalização das exportações e importações, e associar os códigos das URFs com seus nomes ajuda a entender de onde os produtos estão sendo enviados, caso esse dado seja relevante.

2. Filtro de Produtos Relevantes ao Agronegócio

Além disso, criamos um **filtro** específico para isolar produtos do **agronegócio**. Isso foi feito como uma espécie de **pós-processamento dos dados**. Em vez de usar todos os produtos disponíveis, que podem incluir itens não tão relevantes

para o setor, esse filtro ajuda a focar nos produtos que realmente fazem sentido dentro do contexto agrícola e agroindustrial. Esse passo é importante porque, ao trabalhar com grandes volumes de dados, às vezes é necessário aplicar esse tipo de filtragem para que o modelo de regressão foque nos produtos que realmente vão ser exportados pelo setor agrícola.

O filtro permite que apenas produtos relevantes para o agronegócio sejam considerados nas análises, o que deixa os dados mais consistentes e as previsões mais precisas, já que estamos removendo informações que podem "bagunçar" o modelo. Esse pós-processamento ajuda o modelo a se concentrar naquilo que realmente importa e elimina "ruídos" que poderiam prejudicar a qualidade da previsão.

Modelo de Regressão

A função criada realiza uma **regressão linear** para prever o valor de exportação (VL_FOB) de um produto com base em várias características, como ano, mês, peso, país de origem, e outras variáveis. O produto em questão é filtrado pelo seu **id_product**, neste caso, estamos testando a regressão para o produto **CARNES**.

O modelo foi treinado com dados históricos, onde o objetivo é que, a partir dessas variáveis, ele consiga prever o valor de exportação (VL_FOB) para novos dados. O processo é feito em duas etapas principais:

1. **Pré-processamento dos dados:** As variáveis categóricas (como país e estado) são transformadas para um formato que o modelo possa entender, utilizando um método chamado **OneHotEncoder**.
2. **Treinamento e previsão:** O modelo de **regressão linear** é treinado com os dados e, em seguida, gera previsões para os valores de exportação (VL_FOB) a partir dos dados de teste.

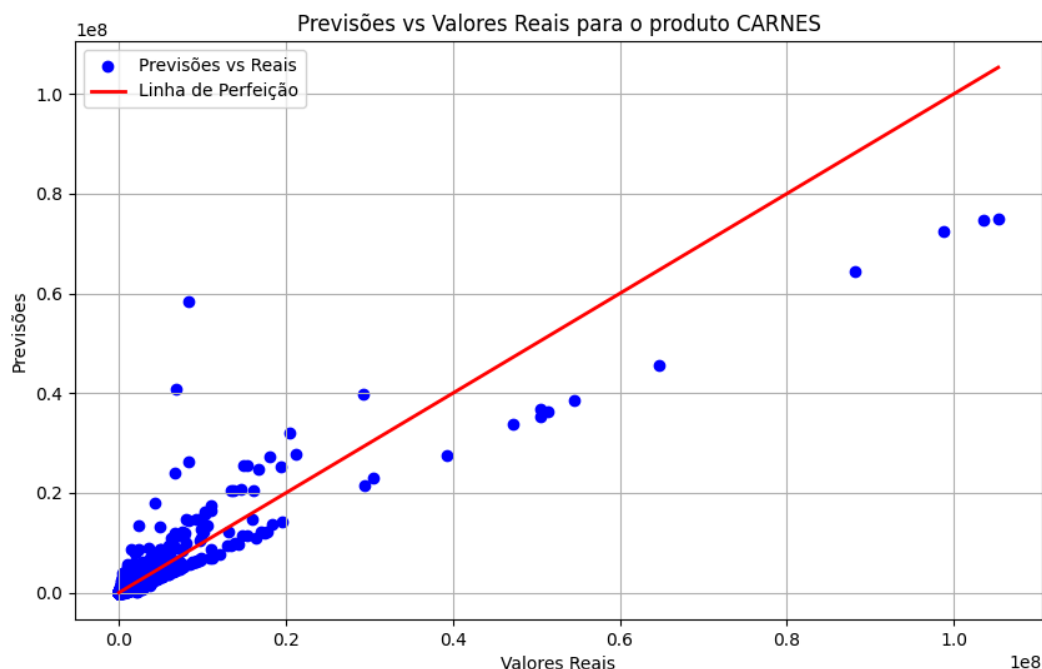
O resultado da execução da função é um conjunto de **previsões** feitas pelo modelo para o produto CARNES, comparado com os **valores reais** de exportação. O modelo é avaliado usando duas métricas de erro: **Erro Quadrático Médio (MSE)** e **Erro Médio Absoluto (MAE)**.

Resultados e Interpretação:

- **Previsões:** O modelo gerou valores como 996.673,18, 15.479,95, etc. Essas são as previsões feitas pelo modelo para os valores de exportação.
- **Valores reais:** Os valores reais de exportação para esses produtos são muito menores, como 190.213, 74.903, etc.

O **Erro Quadrático Médio (MSE)** foi calculado como **928.952.130.102**, indicando que, em média, as previsões estão significativamente distantes dos valores reais. O **Erro Médio Absoluto (MAE)**, que foi de **159.938,15**, mostra que o modelo erra, em média, por cerca de 159.938 unidades nas suas previsões.

Esses resultados sugerem que o modelo de regressão linear precisa de ajustes, como a inclusão de mais variáveis ou o uso de outro tipo de modelo, já que as previsões estão muito distantes da realidade observada.



Conclusão

Ao longo deste processo, desenvolvemos um fluxo de trabalho robusto que envolve desde o tratamento e pré-processamento dos dados até a construção de um modelo de regressão linear para prever o **valor das exportações (VL_FOB)** de produtos específicos, com foco no setor do agronegócio,

especialmente produtos como **Carnes**. Nosso objetivo inicial era criar uma análise preditiva capaz de auxiliar no entendimento dos fatores que influenciam o valor de exportação, e, ao mesmo tempo, melhorar a tomada de decisão nas estratégias comerciais, especialmente em relação à exportação de produtos agrícolas.

1. Tratamento e Expansão dos Dados

A primeira parte crucial desse processo envolveu a preparação dos dados. Para garantir que os dados fossem compreensíveis e úteis, criamos três **dicionários de expansão** (para países, códigos NCM de produtos e URFs). Esses dicionários permitiram que traduzíssemos os códigos numéricos presentes no conjunto de dados em informações mais legíveis e interpretáveis. Isso facilitou tanto a análise quanto a visualização dos dados, já que, ao invés de trabalhar com códigos técnicos, passamos a associar nomes de países, tipos de produtos e unidades fiscais.

Além disso, foi criado um **filtro pós-processamento** para focar exclusivamente em produtos do **agronegócio**, eliminando qualquer dado que não fosse relevante para o modelo. Isso foi especialmente importante, pois garantiu que o modelo de regressão estivesse lidando apenas com produtos que fazem parte do setor agrícola e agroindustrial, melhorando a precisão da previsão e tornando a análise mais alinhada com os objetivos da empresa.

2. Modelo de Regressão Linear

Com os dados devidamente tratados e focados no agronegócio, construímos um modelo de **regressão linear**, uma técnica de aprendizado supervisionado que tenta encontrar a melhor relação entre as variáveis independentes (como ano, mês, país, etc.) e a variável dependente (VL_FOB). Utilizando um **pipeline** de pré-processamento com **OneHotEncoder** para tratar variáveis categóricas (como país e estado), conseguimos treinar o modelo para prever os valores de exportação para o produto específico escolhido, no caso, as carnes.

O desempenho do modelo foi avaliado usando métricas como **Erro Quadrático Médio (MSE)** e **Erro Médio Absoluto (MAE)**, que nos mostraram o quanto as previsões estavam distantes dos valores reais. Embora o modelo tenha apresentado uma grande diferença entre previsões e valores reais, o objetivo

aqui foi principalmente entender a relação entre as variáveis e como isso pode ser melhorado no futuro, seja ajustando o modelo ou explorando outras técnicas.