# Proposta de Desafio Técnico: Logística e Custo de Frete

## Objetivo

Desenvolver um modelo de regressão para prever o custo de envio (providerShippingCost) utilizando análises descritiva, diagnóstica, preditiva e prescritiva. Os candidatos devem trabalhar com um conjunto de dados representando solicitações de envio, explorando, modelando e apresentando suas descobertas de forma detalhada. O objetivo é alcançar as seguintes métricas de desempenho:

* **MAE**: 1,18
* **RMSE**: 3,451
* **RMSLE**: 0,132
* **R²**: 0,889
* **MAPE**: 8,506

## Conjuntos de Dados

O conjunto de dados utilizado é synthetic\_delivery\_requests.csv, que contém informações detalhadas sobre solicitações de envio.

### Identificação e Origem

1. **idShop**: Identificador único da loja.
2. **distinctOrigins**: Número de origens distintas do envio.
3. **originBranchOfficeId**: Identificador da filial de origem.
4. **origin**: Localização de origem do envio.
5. **destination**: Localização de destino do envio.

### Produto

1. **sku**: Identificador único do produto.
2. **quantity**: Quantidade do produto.
3. **productCategory**: Categoria do produto.
4. **weight**: Peso do produto.
5. **costOfGoods**: Custo dos produtos.

### Datas e Prazos

1. **formatted\_date**: Data formatada do pedido.
2. **estimatedDeliveryTimeValue**: Valor estimado do tempo de entrega.
3. **estimatedDeliveryTimeUnit**: Unidade de tempo estimada (ex: dias, horas).
4. **deliveryEstimateBusinessDays**: Estimativa de entrega em dias úteis.

### Método e Custo

1. **shippingMethod**: Método de envio (ex: padrão, expresso).
2. **fulfillmentMethod**: Método de atendimento (ex: dropshipping, estoque próprio).
3. **freightCostCurrency**: Moeda do custo do frete.
4. **providerShippingCost**: Custo do frete fornecido pelo provedor.
5. **deliveryMethodId**: Identificador do método de entrega.
6. **deliveryMethodType**: Tipo de método de entrega.

### Opções e Logística

1. **isRecommendation**: Indica se é uma recomendação (booleano).
2. **isPickupEnable**: Indica se a coleta está habilitada (booleano).
3. **isSchedulingEnable**: Indica se o agendamento está habilitado (booleano).
4. **descriptionLogistic**: Descrição logística.
5. **description**: Descrição adicional do frete.

### Informações Adicionais

1. **state**: Estado do frete (SC, SP, RJ, ...).
2. **displayName**: Nome de exibição do produto.
3. **deliveryMethodType**: Tipo de método de entrega.
4. **deliveryEstimateBusinessDays**: Estimativa de dias úteis para entrega.

## Tarefas e Critérios de Avaliação

### Análise Descritiva (1 ponto)

1. **Exploração Inicial:**
   * Realize uma análise descritiva para entender as características gerais dos dados. (0,5 ponto)
   * Utilize estatísticas resumidas e visualizações para explorar a distribuição dos dados e identificar padrões. (0,5 ponto)

### Análise Diagnóstica (1 ponto)

1. **Identificação de Fatores Influentes:**
   * Identifique e analise os fatores que influenciam o custo de envio. (0,5 ponto)
   * Realize testes estatísticos e análises para validar hipóteses sobre as causas subjacentes. (0,5 ponto)

### Exploração Adicional (0 pontos)

1. **Feature Engineering:**
   * Crie novas variáveis que possam melhorar a análise, com base nas colunas existentes.
   * Considere técnicas de redução dimensional, se apropriado.
   * Utilize técnicas de classificação ou clusterização para segmentar os dados, se isso adicionar valor à análise.
   * Adicione variáveis externas ou construa novas variáveis com base nos dados iniciais.

### Análise Preditiva (1,5 pontos)

1. **Desenvolvimento do Modelo:**
   * Separe o conjunto de dados em três partes: treino, validação e teste. (0,25 ponto)
   * Desenvolva um modelo de regressão para prever o custo de envio (providerShippingCost). (0,5 ponto)
   * Otimize os hiperparâmetros utilizando apenas os conjuntos de treino e teste. (0,25 ponto)
   * Avalie a performance do modelo utilizando as métricas alvo: MAE, MAPE, RMSE, RMSLE, R². As métricas devem ser comparadas utilizando uma média de 5 escolhas aleatórias do grupo de validação. (0,5 ponto)

### Análise Prescritiva (1 ponto)

1. **Propostas de Ação:**
   * 1. Proponha ações baseadas nos resultados preditivos para otimizar o custo de envio. (0,5 ponto)
     2. Sugira melhorias no processo logístico para reduzir custos e melhorar a eficiência. (0,5 ponto)

### Apresentação em PowerPoint (5 pontos)

Crie uma apresentação, incluindo:

1. **Introdução (0,5 pontos):**
   * Contexto do problema e objetivo do desafio.
2. **Metodologia (0,5 pontos):**
   * Descrição dos dados utilizados, preparação dos dados e metodologia de análise.
3. **Análise Descritiva (0,5 pontos):**
   * Resultados e visualizações das análises descritivas.
4. **Análise Diagnóstica (0,5 pontos):**
   * Hipóteses testadas, métodos utilizados e principais descobertas.
5. **Exploração Adicional (2 pontos):**
   * Feature engineering, redução dimensional, classificação ou clusterização, e quaisquer variáveis externas ou novas variáveis criadas que possam ajudar a melhorar o modelo.
6. **Análise Preditiva (0 pontos):**
   * Modelo desenvolvido, métricas de avaliação e resultados preditivos.
7. **Análise Prescritiva (0,5 pontos):**
   * Ações propostas e justificativas.
8. **Conclusão (1 ponto):**
   * Resumo das descobertas e impacto esperado das ações propostas.

Boa tarde - primeiramente quero agradecer a oportunidade de apresentar este projeto e espero que de forma sucinta e efetiva este atenda as expectativas e possa mostrar o potencial que podemos gerar através de dados e ferramentas aprendizado de maquina em uma área muito estratégica de negócios que é a PRECIFICAÇÃO.

Gostaria de Comecar contextualizando o que vamos fazer ou propor:

Extrair insights destes dados, talvez propor ações para empresa, e criar ferramentas de automatização de preço de envio ou até de monitoramento deste preço para identificar possíveis fraudes ou anomalias durante as transações, ou otimizar estes preços.

Olhando para os dados de forma mais detalhada, podemos dividir em 6 subgrupos de informações para cada instância deste dataset tabular:

- Id e origem: Qual a loja, filial, qual a origem (que pode ser o centro de distrib.) e destino dos produtos.

-Datas e Prazos: embora dadas ñ fornec. Temos os prazos em DIAS

- Metodo e Custo: Metodo de env: Express, normal, transp.., Atendim.: Estoque próprio ou dropshipping, e Custo desta entrega em REAIS.

- Inform logística: se foi recomendação, Agendamento, Coleta Dispo.

- Inf adicionais: Estado ex. SP,RJ e método de de entrega, tempo estimado de entre.

A gente pode ver que essas informações são redund. E podem ser removidas

Vamos começar analise Descrit dados alvo (preço envi)

3 tipos e gráficos: Grágico de caixa: Nem da pra distinguir valor mediana, interquartis, apenas os possíveis outliers. O mesmo com gráfico de violino.

Escala logaritimica pq ele resolve esta assimetria na distribuição dos dados.

Muitos dados concentrados em valores pequenos e alguns dados em valores maiores.

Para prosseguir com a analise tive que fazer algumas transformações das variáveis cat. Para numérica. Elas tinham padrão de: nome da coluna \_ ID.

Podemos ver que 25% tem preço 0. Até o terceiro quartil os dados se concentram em valores de 1 decimal.

EM ANALISE DIAGNOSTICA vamos focar no que pode influenciar a performance do modelo de regressão do custo de envio, e a correlação das variáveis com o mesmo.

E pra isso continuei removendo dados que não agregam informações para o nosso desafio:

Remoção de variáveis constantes como unidades de medidas: Reais e Dias;

Origem e destino vazio ou nulos poderiam ser erro de engenharia de dados neste caso ou da regra de negócio que valida as transações (campo de analise prescritiva)

Remoção de IsSchedule = 2 Sendo que até então era esperado uma variável booleana mas tinham 3 valores distintos.

Utilizei um gráfico de calor e a própria tabela de correla.

Variaveis que se destacam por alta correlação com outras variáveis. Poderia encorrer em multicolinearidade ou até dados duplicados.

Utilizei Variance Inflactor Factor (fator de inflação de variância): Ele testa relação linear entre variáveis: mede se a relação segue em escala linear ou proporcionalidade CONSTANTE.

Consegui destacar que estas colunas tem o mesmo numero de linhas com valores iguais.

Terminar:

Correlação das variáveis para cada faixa de preço

E o peso de cada variável muda conforme estas faixas.

Faixas menores: modalidade de envio, origem e principalmente o destino de envio (63%).

Faixas maiores é o peso do produto que é 97% de correlação. 97 significa que é proporcional quase perfeito.direção positiva.

ANALISE PREDITIA:

Antes dividi os dados em Treino teste e validação;

Dividi dados categóricos e quantitativos para testar um encoder dos dados categóricos.

Treinamento preliminar mostra que o target encoder não performou melhor do que os dados normais.

O Catboost performou melhor que outros modelos de regressão.

Catboost utiliza arvores de decisão de forma sequencial, diferente de um random forest. Cada arvore é otimizada com o resíduo da arvore anterior e otimizado com gradiente (uma variável parcial e multidimensional), serve para minimizar uma função que neste caso é a função de perda. Neste caso o CATBoost utiliza a função MSE

Em seguida utilizei a ferramenta de GridSeachCV que treina o modelo em cada combinação dos parâmetros que você passar dependendo do modelo.

Ele utilizou a métrica MSE que penaliza erros maiores. Devido os valores de resíduo serem elevados ao quadrado.