INSTITUTO TECNOLÓGICO DE AERONÁUTICA

Lucas Melo de Oliveira

PREDIÇÃO DE CONSUMO DE COMBUSTÍVEL POR VOO EM AVIAÇÃO COMERCIAL

Trabalho de Graduação
2024

Curso de Engenheria Civil-Aeronáutica

CDU 656.7

Lucas Melo de Oliveira

PREDIÇÃO DE CONSUMO DE COMBUSTÍVEL POR VOO EM AVIAÇÃO COMERCIAL

Orientador

Prof. Marcelo Xavier Guterres (ITA)

ENGENHERIA CIVIL-AERONÁUTICA

São José dos Campos
Instituto Tecnológico de Aeronáutica

2024

Dados Internacionais de Catalogação-na-Publicação (CIP)
Divisão de Informação e Documentação

de Oliveira, Lucas Melo
Predição de consumo de combustível por voo em aviação comercial / Lucas Melo de Oliveira. São José dos Campos, 2024.
43f.

Trabalho de Graduação – Curso de Engenheria Civil-Aeronáutica– Instituto Tecnológico de Aeronáutica, 2024. Orientador: Prof. Marcelo Xavier Guterres.

1. Machine Learning. 2. Aviação. I. Instituto Tecnológico de Aeronáutica. II. Título.

REFERÊNCIA BIBLIOGRÁFICA

DE OLIVEIRA, Lucas Melo. **Predição de consumo de combustível por voo em aviação comercial**. 2024. 43f. Trabalho de Conclusão de Curso (Graduação) – Instituto Tecnológico de Aeronáutica, São José dos Campos.

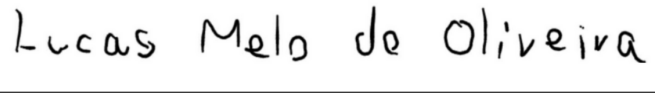
CESSÃO DE DIREITOS

NOME DO AUTOR: Lucas Melo de Oliveira
TITULO DO TRABALHO: Predição de consumo de combustível por voo em aviação comercial.
TIPO DO TRABALHO/ANO: Trabalho de Conclusão de Curso (Graduação) / 2024

Lucas Melo de Oliveira
Rua H8B, 234
12.228-461 – São José dos Campos–SP

PREDIÇÃO DE CONSUMO DE COMBUSTÍVEL POR VOO EM AVIAÇÃO COMERCIAL

Essa publicação foi aceita como Relatório Final de Trabalho de Graduação

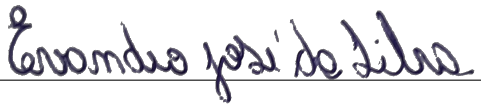


Lucas Melo de Oliveira

Autor

Marcelo Xavier Guterres (ITA)

Orientador



Prof. Evandro José da Silva
Coordenador do Curso de Engenheria Civil-Aeronáutica

São José dos Campos, 23 de novembro de 2024.

Agradecimentos

Primeiramente, queria agrader minha vó Therezinha por ter confiado em mim e investido na minha educação desde sempre. Sua confiança em mim me fez confiar em mim também. Sem você eu nem tentaria!

Ao meu avô Edvaldo, carrego muita saudade e jamais esquecerei todas as vezes que o senhor se esforçou pelo meu melhor, fazendo o que podia e não podia por mim, me levando a rodoviária, garantindo que eu ia acordar no horário da prova...

Aos meus pais, por terem me criado com amor, carinho e respeito mesmo com tantas dificuldades. Obrigado por me desafiarem quando preciso, pelo esforço, por tudo.

Alexandre e família: vocês me acolheram como alguém da família no meu pior momento e eu sempre estarei em dívida com vocês. Sem vocês meu sonho teria morrido e eu não tenho ideia do que seria minha vida... muito obrigado.

Aos amigos do 316, 234 e da família B+ agradeço o acolhimento. Cheguei no ITA não conhecendo quase ninguém e saio lotado de amigos, me sinto imensamente grato por isso! Vou sentir saudades demais da convivência e da resenha.

Por fim, a todos que passaram pela minha vida, obrigado por deixarem as partes boas.

Resumo

O consumo de combustível representa uma parcela significativa dos custos operacionais na aviação comercial brasileira. Esta pesquisa desenvolve modelos preditivos para estimar o consumo de combustível por voo utilizando técnicas de aprendizado de máquina, com foco em diferentes horizontes de planejamento operacional. O estudo utiliza uma base de dados com mais de um milhão de registros de voos comerciais brasileiros, fornecida pela Agência Nacional de Aviação Civil (ANAC). A metodologia empregada inclui preprocessamento robusto dos dados, engenharia de features específica para o contexto aeronáutico e avaliação sistemática de diferentes algoritmos de machine learning. O estudo também propõe diretrizes práticas para definição de margens de planejamento baseadas em evidências quantitativas, contribuindo para a otimização do planejamento operacional das companhias aéreas brasileiras.

Abstract

Fuel consumption represents a significant portion of operational costs in Brazilian commercial aviation. This research develops predictive models to estimate fuel consumption per flight using machine learning techniques, focusing on different operational planning horizons. The study utilizes a database with over one million Brazilian commercial flight records, provided by the National Civil Aviation Agency (ANAC). The methodology includes robust data preprocessing, feature engineering specific to the aeronautical context, and systematic evaluation of different machine learning algorithms. The study also proposes practical guidelines for defining planning margins based on quantitative evidence, contributing to the optimization of Brazilian airlines' operational planning.

Sumário

1 Introdução

1.1 Introdução

A indústria da aviação desempenha um papel crucial no desenvolvimento econômico e social do Brasil. Com sua vasta extensão territorial e diversidade geográfica, o país depende amplamente do transporte aéreo para conectar regiões distantes, facilitando o fluxo de pessoas e mercadorias. Segundo dados da Agência Nacional de Aviação Civil(ANAC, 2023), mais de 100 milhões de passageiros foram transportados em voos domésticos e internacionais em 2022, evidenciando a relevância desse modal na matriz de transportes nacional.

Apesar de sua importância estratégica, as companhias aéreas enfrentam desafios operacionais e financeiros significativos para manter suas atividades de forma sustentável. Entre os principais obstáculos está o elevado consumo de combustível, que representa uma parcela substancial dos custos operacionais. De acordo com a ANAC(ANAC, 2020), em seus Anuários do Transporte Aéreo, o combustível correspondeu, em 2019, a aproximadamente 26,5% dos custos operacionais das empresas aéreas brasileiras, fração crescente, que atingiu incríveis 41% em 2022. Esse alto percentual é influenciado pela volatilidade dos preços do petróleo no mercado internacional e pelas flutuações cambiais, que impactam diretamente o custo do querosene de aviação (QAV).

O elevado consumo de combustível na aviação comercial tem implicações econômicas e ambientais significativas. Do ponto de vista econômico, o aumento nos custos operacionais pode ser repassado para as tarifas aéreas, tornando as viagens mais onerosas para os passageiros e potencialmente limitando o acesso ao transporte aéreo. Isso é particularmente crítico em um país de dimensões continentais, onde alternativas de transporte podem ser limitadas ou inexistentes. Ambientalmente, a aviação comercial é responsável por cerca de 2,5% das emissões globais de gases de efeito estufa. Embora esse percentual possa parecer modesto, a crescente demanda por transporte aéreo e as projeções de aumento no número de voos tornam urgente a busca por soluções que mitiguem o impacto ambiental do setor.

Segundo Anderson(ANDERSON, 2017) , diversos fatores influenciam o consumo de combustível das aeronaves, como o peso da aeronave, condições meteorológicas, rotas e eficiência dos motores. Naturalmente, o peso é um elemento crítico: quanto maior a massa total durante o voo, maior será a quantidade de combustível necessária para mantê-la em altitude de cruzeiro. As condições atmosféricas, como ventos contrários ou favoráveis e turbulências, também afetam o consumo. Além disso, o design aerodinâmico da aeronave e a eficiência dos motores desempenham papéis fundamentais na otimização do consumo de combustível.

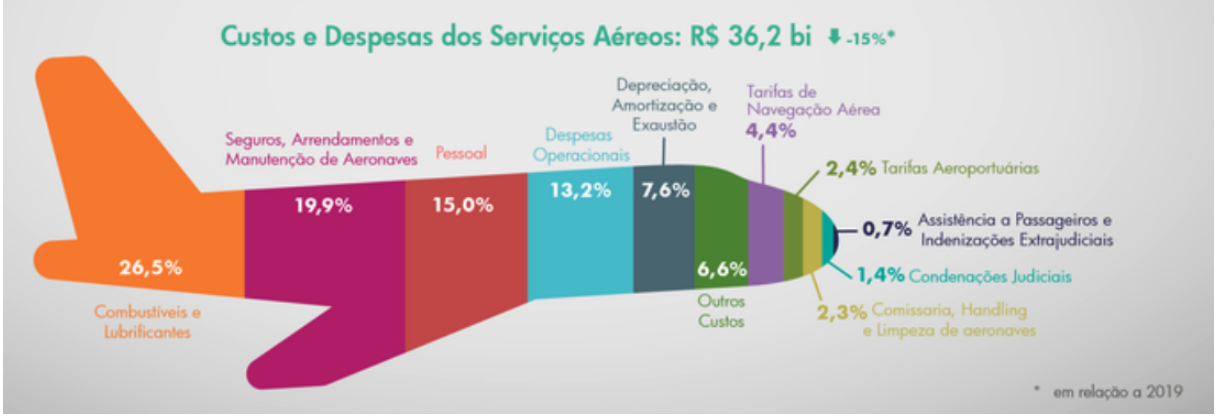


FIGURA 1.1 – Em 2019, combustíveis e lubrificantes corresponderam à 26,5% dos custos das empresas aéreas. (ANAC, 2020)

Em resposta a esses desafios, a indústria tem buscado soluções para reduzir o consumo de combustível e as emissões associadas. Iniciativas como a modernização da frota com aeronaves mais eficientes, otimização de rotas e procedimentos operacionais, e a implementação de tecnologias emergentes são algumas das estratégias adotadas. Com efeito, destaca-se o programa APU Zero da Azul Linhas Aéreas que, com a minimização do uso da APU (motor auxiliar da aeronave), pôde economizar 50 milhões de litros de querosene de aviação em dois anos.

A questão também tem suscitado interesse na academia, onde a aplicação de técnicas de aprendizado de máquina tem ganhado destaque por oferecer potencial para aprimorar a eficiência operacional, prever demandas, otimizar rotas e, especificamente, prever o consumo de combustível. Nesse contexto, destacam-se trabalhos que estimam o consumo de combustível utilizando informações fornecidas pelos registradores de acesso rápido (HONG *et al.*, 2014) e o desenvolvimento de modelos explícitos de desempenho para aeronaves (STOLZER, 2002).

No entanto, persiste uma lacuna significativa na previsão precisa do consumo de combustível por voo, o que poderia auxiliar as companhias aéreas no planejamento operacional e na tomada de decisões estratégicas. O desenvolvimento de modelos mais precisos e robustos nesse âmbito é essencial para promover a sustentabilidade econômica e ambiental

do setor aéreo.

1.2 Problema de Pesquisa

É possível desenvolver um modelo preditivo preciso para o consumo de combustível de voos comerciais utilizando dados operacionais disponíveis no momento do planejamento do voo?

1.3 Objetivo Geral

Desenvolver modelos preditivos e classificatórios utilizando técnicas de aprendizado de máquina para estimar o consumo de combustível em operações aeroportuárias, com base em variáveis operacionais dos voos.

1.4 Objetivos Específicos

- Obter e preparar um conjunto de dados detalhado sobre operações aeroportuárias fornecido pela Agência Nacional de Aviação Civil (ANAC).
- Realizar análise exploratória dos dados para compreender as características e identificar padrões relevantes.
- Tratar variáveis categóricas de alta cardinalidade e dados faltantes, garantindo a qualidade dos dados para a modelagem.
- Selecionar as variáveis mais relevantes que influenciam o consumo de combustível.
- Implementar modelos preditivos com alvo no consumo de combustível
- Avaliar o desempenho dos modelos por meio de métricas adequadas e comparar os resultados obtidos.
- Fornecer insights que possam auxiliar na tomada de decisões estratégicas pelas companhias aéreas e autoridades aeroportuárias.

1.5 Justificativa

A relevância deste estudo reside na potencial contribuição para a eficiência operacional das companhias aéreas e para a sustentabilidade ambiental. Ao desenvolver modelos

capazes de prever com precisão o consumo de combustível, é possível:

- Reduzir os custos operacionais associados ao consumo de combustível, que representam uma parcela significativa das despesas das companhias aéreas.
- Otimizar o planejamento de rotas e a alocação de aeronaves, melhorando a eficiência das operações.
- Minimizar as emissões de gases de efeito estufa, contribuindo para a redução do impacto ambiental da aviação.
- Fornecer uma ferramenta estratégica para as autoridades aeroportuárias na gestão do tráfego aéreo e infraestrutura.
- Avançar o conhecimento na aplicação de técnicas de aprendizado de máquina em problemas complexos envolvendo grandes volumes de dados e variáveis categóricas de alta cardinalidade.

Este trabalho busca, portanto, preencher uma lacuna na previsão do consumo de combustível na aviação, oferecendo soluções práticas e insights valiosos para o setor.

2 Referencial Teórico

2.1 Revisão Bibliográfica

A previsão de consumo de combustível em operações aeronáuticas tem sido objeto de estudo na literatura, com abordagens variando desde modelos estatísticos tradicionais até técnicas modernas de *Machine Learning*. Esta seção apresenta uma análise das principais contribuições na área, com foco nas aplicações de aprendizado de máquina para este problema específico.

2.1.1 Evolução das Abordagens de Machine Learning

As primeiras aplicações de *Machine Learning* na previsão de consumo de combustível em aviação comercial foram documentadas por Li (LI, 2010), que apresentou uma revisão abrangente focada principalmente em redes neurais e dados simulados. O autor destacou a importância do tema ao demonstrar que o combustível representa entre 39% e 47% do peso total da aeronave, dependendo do modelo. O estudo utilizou o software SIMMOD para gerar dados simulados, uma vez que dados reais de companhias aéreas eram raramente disponibilizados devido a questões de confidencialidade.

A abordagem inicial de Li concentrou-se em *Multilayer Perceptrons* (MLPs) e considerava principalmente parâmetros técnicos como temperatura, pressão, densidade, viscosidade e velocidade sônica. Embora pioneiro, o trabalho limitava-se a simulações e não considerava variáveis operacionais reais do dia a dia das companhias aéreas.

2.1.2 Aplicações em Companhias Low-Cost

Um avanço significativo na área foi apresentado por Horiguchi *et al.* (HORIGUCHI *et al.*, 2017), que realizaram um dos primeiros estudos com dados reais de uma companhia aérea low-cost asiática. Utilizando uma base de dados com 54.000 voos e aproximadamente 9,9 milhões de passageiros, os autores aplicaram técnicas como Random Forests, XGBoost e Deep Neural Networks, alcançando um RMSE relativo de 8,8% na previsão de consumo

de combustível.

O trabalho de Horiguchi et al. é particularmente relevante por demonstrar a viabilidade de técnicas de *Machine Learning* em dados reais, além de considerar diferentes horizontes temporais para previsão (1 dia, 1 semana e 5 meses antes do voo). Os autores também identificaram variáveis-chave para a previsão, como quantidade de combustível a bordo, número de passageiros e horário programado de partida.

2.2 Aprendizado de Máquina

O aprendizado de máquina é uma subdisciplina da inteligência artificial que se concentra no desenvolvimento de algoritmos capazes de aprender padrões e relações a partir de dados empíricos. Formalmente, dados de entrada $\mathbf{X} \in \mathbb{R}^{n \times p}$ são utilizados para estimar uma função $f : \mathbb{R}^p \to \mathbb{R}$ que mapeia as variáveis preditoras para uma variável alvo $\mathbf{y} \in \mathbb{R}^n$. O objetivo é encontrar uma função f que minimize uma função de perda $\mathcal{L}$:

$$\mathcal{L}(\mathbf{y}, f(\mathbf{X})) \tag{2.1}$$

Nos modelos supervisionados, o algoritmo aprende uma relação entre as entradas e saídas a partir de um conjunto de treinamento, onde cada observação é um ponto no espaço $\mathbb{R}^p$. A natureza matricial do problema é evidenciada na representação das variáveis preditoras como uma matriz $\mathbf{X}$ e das respostas como um vetor $\mathbf{y}$. Assim, o problema de aprendizado pode ser formulado como a minimização de uma função de custo em termos dos parâmetros $\boldsymbol{\theta}$ do modelo:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\mathbf{y}, f(\mathbf{X}; \boldsymbol{\theta})) \tag{2.2}$$

Na aviação comercial, o aprendizado de máquina tem sido aplicado para diversas finalidades, como manutenção preditiva, otimização de rotas e predição de consumo de combustível. Dada a natureza da operação aérea, grandes volumes de dados operacionais são utilizados para modelar o comportamento complexo do consumo em função de múltiplos fatores.

2.3 Conceitos Fundamentais

2.3.1 Generalização e Validação

No aprendizado de máquina, a capacidade de generalização do modelo é fundamental. Dois fenômenos importantes neste contexto são o *overfitting* e o *underfitting*. O *overfitting* ocorre quando o modelo se ajusta excessivamente aos dados de treino, capturando inclusive o ruído, resultando em:

$$\mathbb{E}[\mathcal{L}(\mathbf{y}_{\text{teste}}, f(\mathbf{X}_{\text{teste}})] \gg \mathbb{E}[\mathcal{L}(\mathbf{y}_{\text{treino}}, f(\mathbf{X}_{\text{treino}})] \tag{2.3}$$

O *underfitting*, por outro lado, ocorre quando o modelo é muito simples para capturar a complexidade dos dados, resultando em alto erro tanto no treino quanto no teste.

Para avaliar adequadamente a capacidade de generalização, utiliza-se a validação cruzada k-fold, onde os dados são divididos em k partes, e o modelo é treinado k vezes, cada vez usando uma parte diferente como teste:

$$\text{Score}_{\text{CV}} = \frac{1}{k} \sum_{i=1}^{k} \mathcal{L}(\mathbf{y}_i, f_i(\mathbf{X}_i)) \tag{2.4}$$

2.3.2 Preprocessamento de Dados

2.3.2.1 Codificação One-Hot

A codificação one-hot transforma uma variável categórica com k categorias em k variáveis binárias:

$$\text{categoria} \rightarrow [0, \ldots, 1, \ldots, 0] \tag{2.5}$$

Por exemplo, uma feature categórica ”feature”com dois valores distintos A e B, após esse processo seria dividida em duas features distintas, ”feature_a”sinalizando a presença ou não de A na feature original e ”feature_b”, fazendo o mesmo para B. Dessa maneira, caso o valor original da feature fosse A, teríamos uma coluna com

Este processo é essencial para converter dados categóricos em um formato numérico que os algoritmos possam processar.

2.3.2.2 Normalização e Padronização

A padronização (StandardScaler) transforma as features para terem média zero e variância unitária:

$$x'_i = \frac{x_i - \mu}{\sigma} \tag{2.6}$$

Já a normalização (MinMaxScaler) escala os valores para um intervalo específico, geralmente [0,1]:

$$x'_i = \frac{x_i - \min(x)}{\max(x) - \min(x)} \tag{2.7}$$

2.3.2.3 Feature Engineering

A engenharia de features é o processo de criar novas variáveis a partir das existentes, visando melhorar a performance do modelo. Por exemplo, para uma feature temporal t, podemos criar componentes cíclicas:

$$\begin{aligned} \sin_t &= \sin(2\pi t/T) \\ \cos_t &= \cos(2\pi t/T) \end{aligned} \tag{2.8}$$

onde T é o período do ciclo.

2.4 Modelos de Aprendizado de Máquina

2.4.1 Modelos de Regressão

Nos problemas de regressão, busca-se prever um valor numérico contínuo com base em variáveis independentes. Um modelo clássico é a regressão linear múltipla, onde se assume uma relação linear entre as variáveis independentes e a variável dependente:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.9}$$

onde $\boldsymbol{\beta} \in \mathbb{R}^p$ é o vetor de coeficientes e $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ é o termo de erro aleatório.

2.4.1.1 Métricas de Avaliação

Para avaliar a performance de modelos de regressão, utilizam-se métricas que quantificam a discrepância entre os valores preditos $\hat{\mathbf{y}}$ e os observados $\mathbf{y}$. As principais métricas são:

2.4.1.1.1 Erro Quadrático Médio (RMSE) O RMSE é a raiz quadrada da média dos quadrados dos erros:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2.10}$$

Esta métrica penaliza fortemente erros grandes devido ao termo quadrático e é expressa nas mesmas unidades da variável alvo.

2.4.1.1.2 Erro Médio Absoluto (MAE) O MAE calcula a média das diferenças absolutas entre valores preditos e observados:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2.11}$$

É uma métrica mais robusta a outliers, pois não eleva os erros ao quadrado.

2.4.1.1.3 Coeficiente de Determinação (R^2) O R^2 mede a proporção da variância total da variável dependente explicada pelo modelo:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2.12}$$

onde $\bar{y}$ é a média dos valores observados. O R^2 varia entre 0 e 1, sendo 1 a representação de um modelo perfeito.

2.4.2 Árvores de Decisão e Ensembles

2.4.2.1 Árvores de Decisão

Uma árvore de decisão particiona recursivamente o espaço de features usando regras do tipo "if-then". Em cada nó, a árvore escolhe a feature j e o threshold t que maximizam a redução na impureza:

$$\Delta I(j,t) = I(\text{pai}) - \frac{N_{\text{esq}}}{N} I(\text{esq}) - \frac{N_{\text{dir}}}{N} I(\text{dir}) \tag{2.13}$$

onde I é uma medida de impureza (como variância para regressão ou entropia para classificação), e N representa o número de amostras.

2.4.2.2 Ensemble Learning

Métodos ensemble combinam múltiplos modelos base para criar um modelo mais robusto. As duas principais abordagens são:

- **Bagging:** Treina modelos independentemente em diferentes amostras bootstrap dos dados:
$$f_{\text{ensemble}}(x) = \frac{1}{M} \sum_{m=1}^{M} f_m(x) \tag{2.14}$$
- **Boosting:** Treina modelos sequencialmente, cada um focando nos erros dos anteriores:
$$f_{\text{ensemble}}(x) = \sum_{m=1}^{M} \alpha_m f_m(x) \tag{2.15}$$
onde α_m são os pesos de cada modelo.

2.4.2.3 Gradient Boosting

O Gradient Boosting constrói um modelo aditivo:

$$f_m(x) = f_{m-1}(x) + \eta h_m(x) \tag{2.16}$$

onde h_m é treinado nos resíduos negativos do gradiente da função de perda:

$$r_{im} = - \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right]_{f=f_{m-1}} \tag{2.17}$$

e η é a taxa de aprendizado.

2.4.3 Algoritmos Implementados

2.4.3.1 Algoritmos Baseados em Gradient Boosting

2.4.3.1.1 XGBoost O *XGBoost* (eXtreme Gradient Boosting) é um algoritmo de *boosting* baseado em árvores de decisão que utiliza técnicas de otimização para melhorar

a velocidade e o desempenho. Ele constrói modelos sequencialmente, onde cada novo modelo tenta corrigir os erros dos modelos anteriores.

A função objetivo regularizada a ser minimizada é:

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^{t} \Omega(f_k) \tag{2.18}$$

onde l é a função de perda, $\hat{y}_i^{(t)}$ é a predição na iteração t, e $\Omega(f_k)$ é o termo de regularização do modelo f_k.

2.4.3.1.2 CatBoost O *CatBoost* é um algoritmo de *gradient boosting* que lida de forma eficiente com variáveis categóricas, evitando a necessidade de pré-processamento como codificação *one-hot*. Utiliza técnicas como *Ordered Boosting* para reduzir o *overfitting* e melhorar o desempenho.

Matematicamente, segue o procedimento padrão do *gradient boosting*, mas incorpora métodos específicos para tratamento de dados categóricos.

2.4.3.1.3 LightGBM O *LightGBM* é um framework de *gradient boosting* que utiliza algoritmos baseados em árvores de decisão. É projetado para ser eficiente em termos de memória e velocidade, usando técnicas como *Gradient-based One-Side Sampling* (GOSS) e *Exclusive Feature Bundling* (EFB).

A função objetivo é semelhante à do *XGBoost*, minimizando uma função de perda com termos de regularização.

2.4.3.1.4 HistGradientBoostingRegressor O *HistGradientBoostingRegressor* é uma implementação de *gradient boosting* que utiliza histogramas para acelerar o treinamento, sendo especialmente eficaz com grandes conjuntos de dados.

2.4.3.2 Modelos Lineares

2.4.3.2.1 ElasticNet O *ElasticNet* é um modelo linear que combina as penalizações L1 (*Lasso*) e L2 (*Ridge*) para regularização, controladas pelos parâmetros λ_1 e λ_2.

A função custo a ser minimizada é:

$$\min_{\beta} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \beta)^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2 \right\} \tag{2.19}$$

2.4.3.2.2 Lasso O *Lasso* (*Least Absolute Shrinkage and Selection Operator*) é um modelo linear com penalização L1, que pode reduzir coeficientes a zero, efetivamente selecionando características.

A função custo é:

$$\min_{\beta} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \beta)^2 + \lambda \|\beta\|_1 \right\} \tag{2.20}$$

2.4.3.2.3 SGDRegressor O *SGDRegressor* é um modelo linear que otimiza uma função de perda regularizada usando o método do gradiente descendente estocástico.

A função custo genérica é:

$$\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} L(y_i, \mathbf{x}_i^\top \beta) + \lambda R(\beta) \right\} \tag{2.21}$$

onde L é a função de perda (por exemplo, perda quadrática) e R é o termo de regularização.

2.4.3.3 Outros Métodos

2.4.3.3.1 KNeighborsRegressor O *K-Nearest Neighbors Regressor* (KNN Regressor) prevê o valor de uma nova amostra com base na média dos valores das k amostras mais próximas no espaço de características.

A predição é dada por:

$$\hat{y} = \frac{1}{k} \sum_{i \in \mathcal{N}_k(x)} y_i \tag{2.22}$$

onde $\mathcal{N}_k(x)$ é o conjunto das k amostras mais próximas de x.

2.4.3.3.2 Stacking Regressor O *Stacking Regressor* combina múltiplos modelos de regressão (aprendizes base) e utiliza um modelo meta-aprendiz para integrar as predições dos aprendizes base.

Matematicamente, as predições dos modelos base $h_1(x), h_2(x), \ldots, h_m(x)$ são usadas como entradas para o meta-modelo H:

$$\hat{y} = H(h_1(x), h_2(x), \ldots, h_m(x)) \tag{2.23}$$

3 Metodologia

Este capítulo apresenta a metodologia empregada no desenvolvimento do estudo de predição de consumo de combustível em aviação comercial. São descritos os dados utilizados, os processos de tratamento e preparação dos dados, bem como os modelos e métricas empregados na análise.

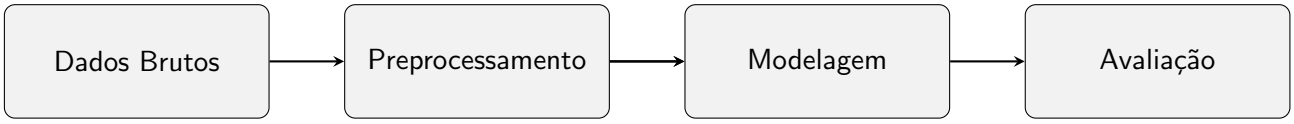


FIGURA 3.1 – Visão macro do processo de predição de consumo de combustível

3.1 Base de Dados

3.1.1 Fonte e Escopo dos Dados

Os dados utilizados neste estudo foram obtidos através do Portal de Dados Abertos da Agência Nacional de Aviação Civil (ANAC), que disponibiliza informações detalhadas sobre voos comerciais realizados no Brasil. O conjunto de dados abrange operações realizadas entre os anos de 2000 e 2024, constituindo uma base abrangente do cenário da aviação comercial brasileira.

O conjunto de dados inicial contém 1.025.131 registros de voos, cada um representando uma operação aérea realizada no período. Estes registros contemplam informações de todas as companhias aéreas que operaram voos comerciais no Brasil durante o período analisado, oferecendo uma visão completa do setor.

3.1.2 Variáveis do Estudo

A variável alvo do estudo é o consumo de combustível em litros (COMBUSTIVEL_LITROS), que representa o volume total de combustível utilizado em cada operação. As variáveis preditoras foram divididas em dois grupos: categóricas e numéricas, conforme apresentado nas Tabelas 3.1 e 3.2.

TABELA 3.1 – Variáveis Categóricas

Variável	Descrição
AEROPORTO_DE_ORIGEM_SIGLA	Código IATA do aeroporto de origem
AEROPORTO_DE_DESTINO_SIGLA	Código IATA do aeroporto de destino
GRUPO_DE_VOO	Classificação do tipo de voo
NATUREZA	Natureza da operação
EMPRESA_NOME	Nome da empresa aérea operadora

TABELA 3.2 – Variáveis Numéricas

Variável	Descrição
ANO	Ano de realização do voo
MES	Mês de realização do voo
PASSAGEIROS_PAGOS	Número de passageiros pagantes
PASSAGEIROS_GRATIS	Número de passageiros não pagantes
CARGA_PAGA_KG	Peso da carga paga em quilogramas
CARGA_GRATIS_KG	Peso da carga não paga em quilogramas
CORREIO_KG	Peso de correio transportado em quilogramas
ASK	Available Seat Kilometers
RPK	Revenue Passenger Kilometers
ATK	Available Tonne Kilometers
RTK	Revenue Tonne Kilometers
DISTANCIA_VOADA_KM	Distância voada em quilômetros
DECOLAGENS	Número de decolagens
ASSENTOS	Número de assentos disponíveis
PAYLOAD	Carga paga máxima em quilogramas
HORAS_VOADAS	Duração do voo em horas
BAGAGEM_KG	Peso total de bagagem em quilogramas

3.2 Processamento dos Dados

3.2.1 Tratamento Inicial

O processo de tratamento dos dados foi implementado através da classe `DataPreprocessor`, que executa uma série de operações de limpeza e validação. Inicialmente, foram removidos os registros que apresentavam valores faltantes em qualquer uma das variáveis, uma decisão tomada devido ao grande volume de dados disponíveis e à impossibilidade de imputação segura destes valores no contexto de consumo de combustível.

Em seguida, foram removidos registros que apresentavam valores inválidos, incluindo:

- Consumo de combustível negativo ou zero
- Registros com distância voada igual a zero mas consumo de combustível diferente de zero

- Valores inconsistentes na relação entre distância e consumo

Para o tratamento de outliers, foi implementada uma abordagem baseada no consumo específico de combustível, calculado pela razão entre o volume de combustível consumido e a distância voada (litros/km). Foram removidos os registros que apresentavam valores de consumo específico além de três desvios padrão da média, por serem considerados potencialmente errôneos ou não representativos das operações normais.

3.2.2 Engenharia de Features

Durante o processamento dos dados, foi implementada uma estratégia abrangente de engenharia de features, organizada em três categorias principais: básicas, temporais e avançadas. Essa estruturação permitiu uma abordagem sistemática para capturar diferentes aspectos do consumo de combustível.

3.2.2.1 Features Básicas

Inicialmente, foram criadas variáveis que representam aspectos fundamentais da operação:

- `PESO_TOTAL_KG`: Agregação de todas as cargas transportadas (paga, grátis, correio e bagagem), fornecendo uma medida única do peso total transportado
- `PASSAGEIROS_TOTAL`: Soma dos passageiros pagos e gratuitos
- `TAXA_OCUPACAO`: Razão entre o total de passageiros e o número de assentos disponíveis, medindo a eficiência no uso da capacidade da aeronave

3.2.2.2 Features Temporais

Para capturar padrões temporais e sazonais no consumo, foram desenvolvidas:

- `EFICIENCIA_TREND_[EMPRESA]`: Média móvel de 12 meses da eficiência de combustível por empresa, calculada como a razão entre consumo e (distância $\times$ peso)
- `SECTOR_EFFICIENCY`: Eficiência média do setor para cada período
- `RELATIVE_EFFICIENCY`: Comparação da eficiência individual com a média do setor
- `MONTH_EFFICIENCY`: Padrão de eficiência mensal por empresa
- Features cíclicas (`MES_SIN`, `MES_COS`): Transformações trigonométricas do mês para capturar sazonalidade

3.2.2.3 Features Avançadas

Foram criadas features complexas para capturar relações não lineares e interações:

- **Métricas de Eficiência:**
  - `EFICIENCIA_KM`: Consumo por quilômetro voado
  - `EFICIENCIA_PESO`: Consumo por quilograma transportado
  - `EFICIENCIA_COMBINADA`: Consumo normalizado por distância e peso
- **Features de Interação:** Multiplicações e divisões entre pares de variáveis numéricas principais (distância, peso total e taxa de ocupação)
- **Features Agregadas:** Estatísticas por dimensões operacionais:
  - Consumo médio e desvio padrão por empresa
  - Consumo médio e desvio padrão por aeroporto de origem
  - Consumo médio e desvio padrão por aeroporto de destino

Um sistema robusto de validação foi implementado para garantir a qualidade das features geradas. Cada nova feature passou por análises de:

- Presença de valores faltantes (limite de 30% de missing values)
- Presença de valores infinitos (limite de 10%)
- Tratamento automático de outliers e valores inválidos

Após a criação das features, foi realizada uma análise de correlação entre todas as variáveis numéricas, utilizando um limiar de 0,95 para identificar e remover variáveis altamente correlacionadas, visando reduzir a multicolinearidade e melhorar a estabilidade dos modelos. Este processo garantiu um conjunto de features informativo e não redundante para o treinamento dos modelos.

3.3 Implementação

3.3.1 Estrutura de Classes

O desenvolvimento do sistema de predição foi organizado em duas classes principais:

1. `DataPreprocessor`: Responsável por todo o processamento inicial dos dados, incluindo:
    - Remoção de valores faltantes e inválidos
    - Tratamento de outliers
    - Criação de novas features
    - Análise e tratamento de correlações
2. `ModelsRunner`: Gerencia o treinamento e avaliação dos modelos, incluindo:
    - Preparação dos pipelines de processamento
    - Otimização de hiperparâmetros
    - Validação cruzada
    - Persistência dos modelos e resultados

3.3.2 Modelos Utilizados

Para a predição do consumo de combustível, foram implementados diversos modelos de machine learning, abrangendo diferentes paradigmas e complexidades:

- **Modelos Lineares:**
    - **Elastic Net**: Combina regularizações L1 e L2, sendo eficaz para dados com multicolinearidade
    - **Lasso Regression**: Utiliza regularização L1 para seleção automática de features
    - **SGDRegressor**: Implementação eficiente para grandes volumes de dados usando gradiente descendente estocástico
- **Modelos Baseados em Árvores:**
    - **XGBoost**: Framework de gradient boosting otimizado, conhecido por sua performance superior em competições
    - **CatBoost**: Especializado no tratamento eficiente de variáveis categóricas
    - **LightGBM**: Implementação leve e rápida de gradient boosting usando técnicas GOSS e EFB
    - **Hist Gradient Boosting**: Implementação do scikit-learn otimizada para grandes datasets
- **Modelos Baseados em Instância:**

  - **KNeighbors**: Modelo não-paramétrico que prediz baseado nas k amostras mais próximas

- **Meta-Modelos:**

  - **Stacking**: Combina predições de múltiplos modelos base usando um meta-modelo

A escolha deste conjunto diversificado de modelos visa explorar diferentes aspectos do problema:

- **Linearidade vs. Não-linearidade**: Os modelos lineares (Elastic Net, Lasso) capturam relações lineares básicas, enquanto os baseados em árvores podem modelar interações complexas
- **Tratamento de Categóricas**: CatBoost oferece tratamento especializado para variáveis categóricas, importante devido à natureza dos dados aeronáuticos
- **Escalabilidade**: LightGBM e Hist Gradient Boosting são otimizados para grandes volumes de dados
- **Ensemble Learning**: O modelo Stacking combina as predições dos demais modelos, potencialmente capturando diferentes aspectos do problema

3.3.3 Otimização de Hiperparâmetros

Para cada modelo, foi implementada uma busca em grade (*Grid Search*) com validação cruzada para otimização dos hiperparâmetros.

3.3.4 Persistência dos Modelos

Para garantir a reprodutibilidade e facilitar o uso posterior dos modelos treinados, foi implementado um sistema robusto de persistência. Cada modelo treinado é salvo em um diretório específico, seguindo a estrutura:

```
models/
   [Nome_do_Modelo]_[Timestamp]/
       model.joblib
       metadata.joblib
```

O arquivo `model.joblib` contém o modelo treinado com seus melhores hiperparâmetros, enquanto o arquivo `metadata.joblib` armazena informações importantes sobre o treinamento, incluindo:

- Nome e versão do modelo
- Timestamp do treinamento
- Melhores hiperparâmetros encontrados
- Métricas de performance em cada fold da validação cruzada
- Lista de features utilizadas (numéricas e categóricas)

3.4 Avaliação dos Modelos

3.4.1 Processo de Validação

A validação dos modelos foi realizada utilizando K-Fold Cross Validation com K = 5, escolhido para:

- Fornecer uma estimativa robusta do desempenho dos modelos
- Reduzir o risco de overfitting
- Garantir que todos os dados sejam utilizados tanto para treino quanto para teste

O processo de validação seguiu as seguintes etapas:

1. Divisão dos dados em 5 folds
2. Para cada fold:
    - Treinamento do modelo com 4 folds
    - Avaliação no fold restante
    - Cálculo das métricas de performance
3. Cálculo das médias e desvios padrão das métricas
4. Treinamento final com todos os dados utilizando os melhores hiperparâmetros

Todo o processo é automatizado através da classe `ModelsRunner`, que gerencia o treinamento, validação e armazenamento dos resultados. Os resultados específicos obtidos para cada modelo serão apresentados e discutidos no Capítulo 4.

4 Resultados

4.1 Análise Exploratória dos Dados

4.1.1 Distribuição do Consumo de Combustível

A análise das estatísticas descritivas do consumo de combustível revelou características importantes para a compreensão do perfil operacional dos voos. A diferença significativa entre média e mediana, combinada com o elevado desvio padrão, indica uma distribuição fortemente assimétrica à direita.

TABELA 4.1 – Estatísticas descritivas do consumo de combustível

Estatística	**Valor (Litros)**
Média	129.429,48
Mediana	29.360,00
Desvio Padrão	238.691,46
Mínimo	1,00
Máximo	5.566.920,00

Esta distribuição assimétrica é reforçada pela análise da distribuição dos voos por faixas de consumo, que mostra uma concentração maior nas faixas de menor consumo:

TABELA 4.2 – Distribuição dos voos por faixa de consumo

Faixa de Consumo	**Quantidade de Voos**	**Percentual**
0-5000L	132.947	23,44%
5001-30000L	152.331	26,85%
30001-150000L	136.835	24,12%
150001-300000L	73.281	12,92%
>300000L	71.898	12,67%

4.1.2 Correlações entre Variáveis

A análise de correlações revelou relações significativas entre o consumo de combustível e outras variáveis operacionais, conforme visualizado na Figura 4.1.

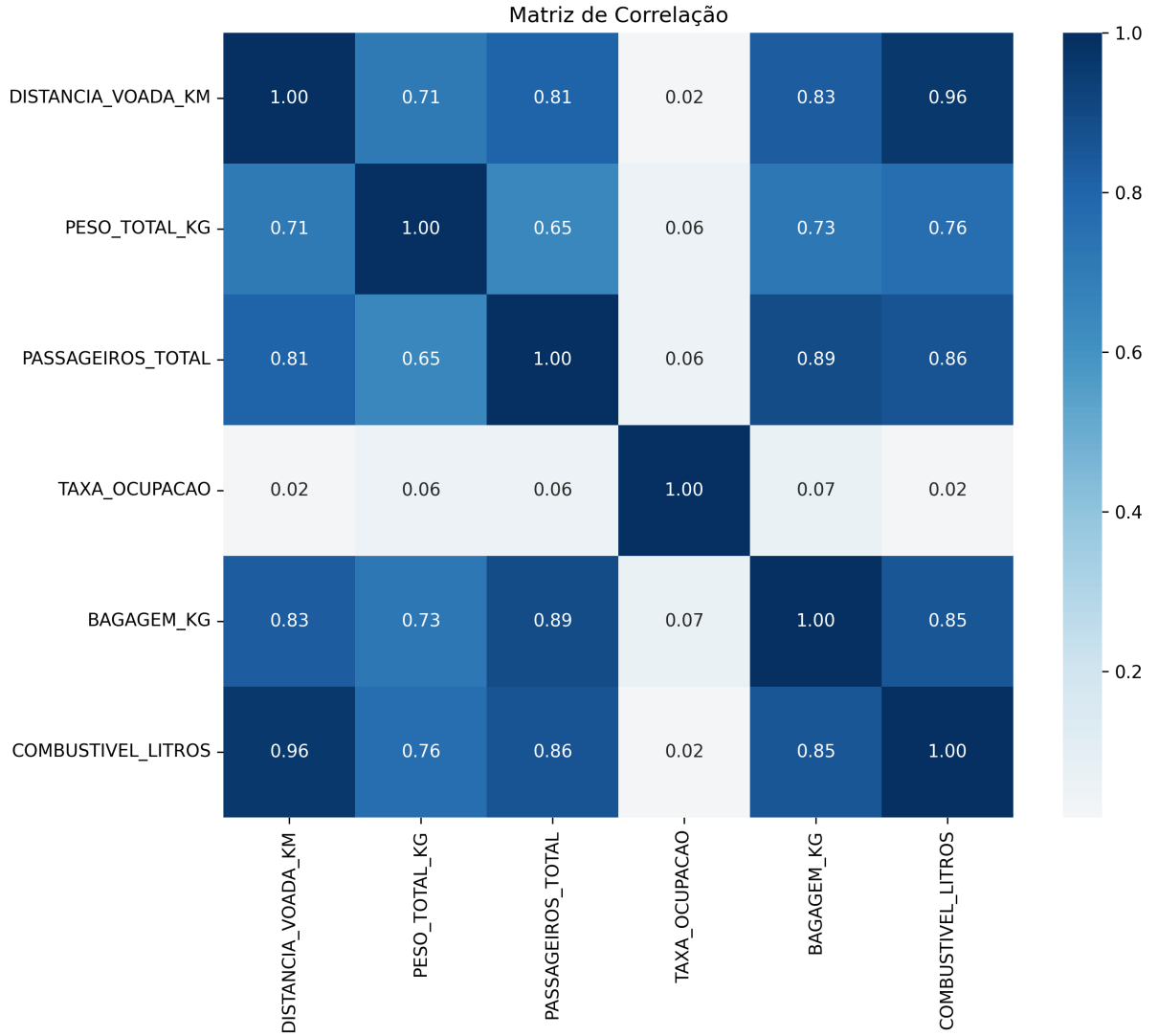


FIGURA 4.1 – Matriz de correlação entre variáveis operacionais

A forte correlação com a distância voada (0,9628) confirma este como principal fator determinante do consumo, como evidenciado também na Figura 4.2. Particularmente interessante é a baixa correlação com a taxa de ocupação (0,0173), sugerindo que a eficiência no preenchimento das aeronaves tem impacto limitado no consumo de combustível.

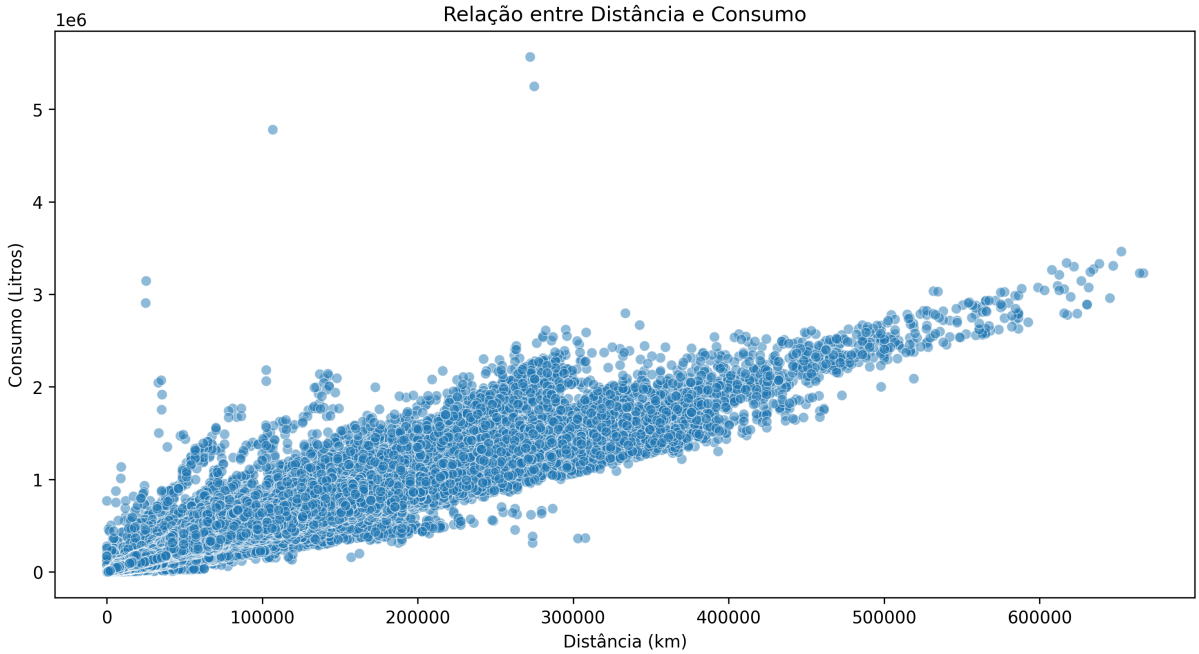


FIGURA 4.2 – Relação entre distância voada e consumo de combustível

4.1.3 Eficiência Operacional por Segmento

A análise da eficiência operacional por faixas de distância revela padrões distintos e uma evolução temporal significativa. Como mostrado na Figura 4.3, voos de curta distância (0-500km) apresentam consistentemente maior consumo por quilômetro, variando entre 7,5 e 15 L/km. Este comportamento é esperado devido à maior influência das fases de decolagem e pouso nestas operações, que são inerentemente mais intensivas no consumo de combustível.

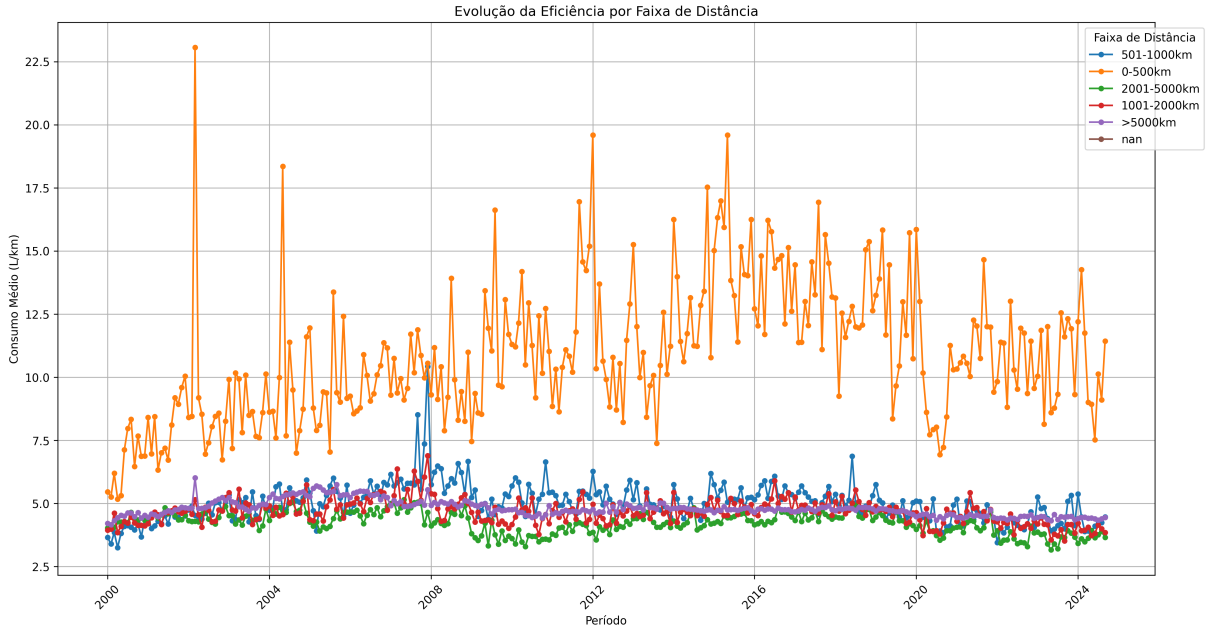


FIGURA 4.3 – Evolução da eficiência por faixa de distância

Em contraste, voos de média e longa distância (acima de 1000km) demonstram maior estabilidade na eficiência, com consumo médio entre 4 e 5 L/km. Esta estabilidade pode ser atribuída à maior proporção do voo em altitude de cruzeiro, fase mais eficiente da operação.

A análise por faixas de consumo (Figura 4.4) evidencia uma tendência geral de melhoria na eficiência operacional ao longo do período estudado. Todas as faixas de consumo apresentaram redução gradual no consumo por quilômetro, com convergência mais acentuada após 2016.

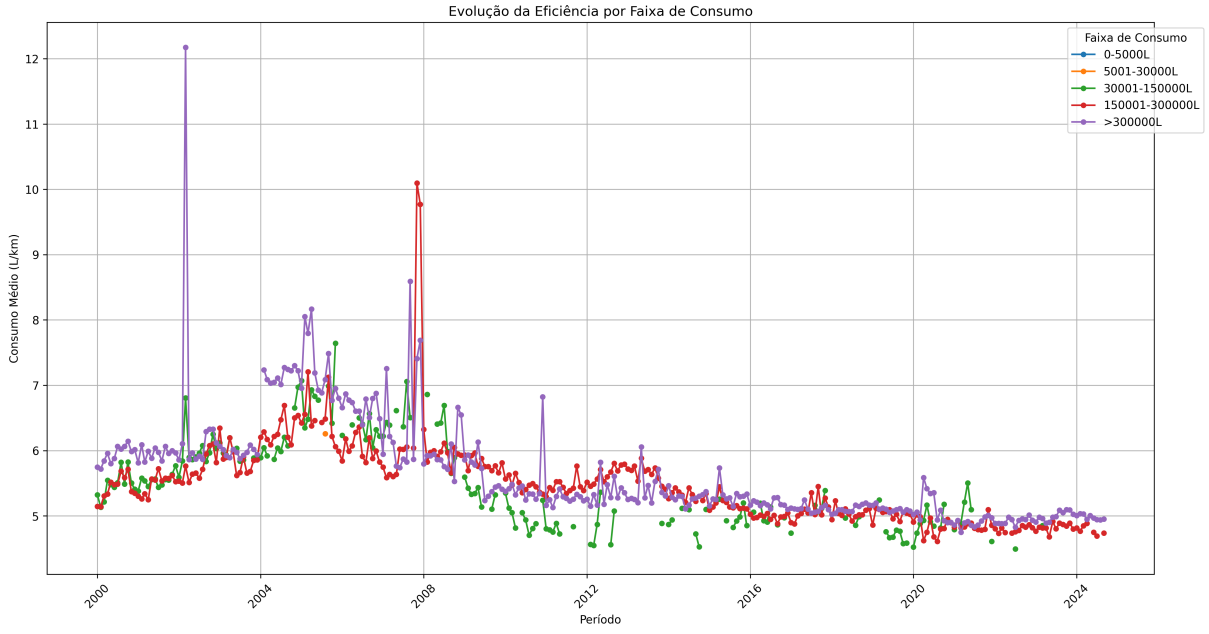


FIGURA 4.4 – Evolução da eficiência por faixa de consumo

4.1.4 Evolução Temporal das Operações

O volume operacional apresenta variações significativas ao longo do período analisado (Figura 4.5). Destacam-se três períodos distintos: uma redução gradual entre 2000 e 2005; estabilização em torno de 24.000 voos anuais entre 2010 e 2019; e uma queda abrupta em 2020, coincidindo com a pandemia de COVID-19, seguida por recuperação expressiva nos anos subsequentes.

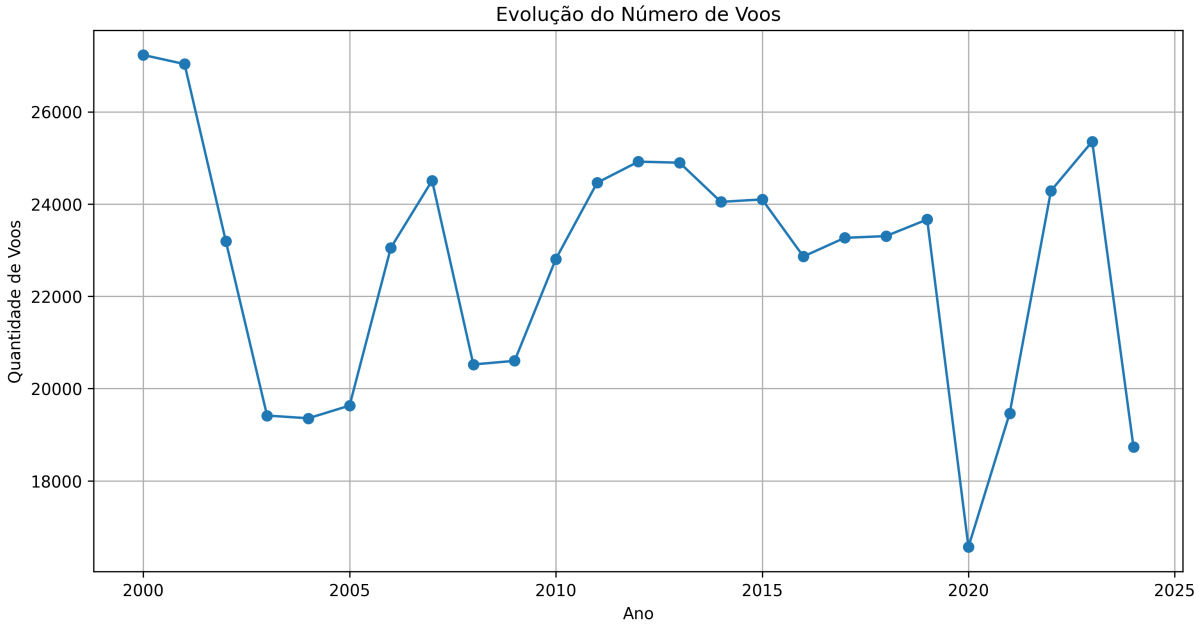


FIGURA 4.5 – Evolução do número de voos ao longo do tempo

4.2 Desempenho Geral dos Modelos

4.3 Comparação das Métricas Principais

Os modelos apresentaram os seguintes desempenhos:

TABELA 4.3 – Comparação de métricas entre modelos

Modelo	**MAE**	**MAPE (%)**	**RMSE**	**R^2**
XGBoost	5.779,19	21,99	22.961,56	0,9908
Stacking	7.272,25	23,68	25.061,18	0,9891
Hist Gradient Boosting	9.366,15	42,27	26.807,76	0,9875
LightGBM	11.501,65	49,15	28.708,71	0,9856
KNeighbors	12.356,84	108,45	28.031,73	0,9863
CatBoost	11.683,54	181,06	27.841,60	0,9865
SGD Regressor	127.378,01	123,26	269.228,05	-0,2623

4.4 Análise por Horizonte de Previsão

O desempenho de alguns dos modelos foi avaliado em diferentes horizontes temporais, de modo a simular a performance do modelo em casos reais de predição, onde temos um número limitado de features. A escolha dos modelos a serem analisados ocorreu de mode que os modelos que performaram mal inicialmente (como o *SGD Regressor*) e modelos que não performam bem com features parciais (como o *KNeighbors*).

TABELA 4.4 – Performance dos modelos em diferentes cenários de predição

Modelo	**Cenário**	**Features**	**MAPE (%)**	**RMSE (L)**	**R^2**
XGBoost	Anual	6	1.544,81	113.885,99	0,774
	Mensal	8	1.572,92	116.250,42	0,765
	Semanal	10	1.552,19	115.951,49	0,766
Stacking	Anual	6	1.801,59	125.962,31	0,724
	Mensal	8	1.827,41	126.277,13	0,722
	Semanal	10	1.810,64	125.669,16	0,725
LightGBM	Anual	6	2.632,71	137.362,58	0,671
	Mensal	8	2.640,46	137.201,14	0,672
	Semanal	10	2.638,55	136.678,02	0,675
CatBoost	Anual	6	2.632,93	140.093,91	0,658
	Mensal	8	2.685,31	137.720,35	0,670
	Semanal	10	2.719,67	138.367,64	0,667

A análise da performance dos modelos em diferentes cenários de predição, apresentada na Tabela 4.4, revela aspectos importantes sobre a capacidade preditiva dos modelos

com informações parciais. Os cenários representam diferentes momentos do planejamento operacional, variando de anual (mais distante do voo) a semanal (mais próximo), com quantidade crescente de informações disponíveis.

O XGBoost demonstrou superioridade consistente em todos os cenários, alcançando o melhor R^2 (0,774) e menor RMSE (113.885,99 litros) no planejamento anual. Curiosamente, sua performance é ligeiramente melhor com menos informações, sugerindo maior robustez na extração de padrões a partir de dados limitados.

O modelo Stacking, ocupando a segunda posição em performance geral, destaca-se pela estabilidade entre cenários. Com variações mínimas no R^2 (0,722-0,725) e RMSE (126.000 litros), demonstra consistência independente da quantidade de informações disponíveis, característica valiosa para aplicações práticas.

LightGBM e CatBoost apresentaram comportamentos similares, com R^2 na faixa de 0,65-0,67, porém com erros absolutos significativamente maiores. Ambos mostram leve tendência de melhora com o aumento de informações disponíveis, evidenciada pelo crescimento do R^2 no cenário mensal.

Um aspecto notável é que todos os modelos mantêm performances relativamente estáveis entre os cenários, com variações máximas de R^2 inferiores a 0,01 unidade. Isso sugere que, uma vez estabelecidos os padrões básicos de consumo (aeroportos, empresa, distância), informações adicionais têm impacto marginal na capacidade preditiva dos modelos.

4.5 Seleção do Melhor Modelo

A seleção do XGBoost como modelo mais adequado para a predição de consumo de combustível baseou-se em uma análise multifatorial que considerou aspectos quantitativos e qualitativos. O modelo destacou-se consistentemente em todas as métricas de avaliação e demonstrou robustez superior em diferentes cenários de aplicação.

4.5.1 Performance Superior em Métricas Principais

O XGBoost apresentou resultados superiores em todas as métricas de avaliação principais:

- MAE de 5.779,19 litros, representando o menor erro absoluto médio entre todos os modelos avaliados
- MAPE de 21,99%, indicando a melhor performance em termos percentuais
- RMSE de 22.961,56 litros, demonstrando maior precisão em termos absolutos

- R^2 de 0,9908, evidenciando excelente capacidade explicativa do modelo

4.5.2 Estabilidade entre Cenários de Predição

Um diferencial significativo do XGBoost foi sua capacidade de manter performance consistente em diferentes horizontes de previsão, como evidenciado na Tabela 4.4. O modelo demonstrou:

- Menor variação de R^2 entre cenários (0,774 a 0,765)
- Performance superior mesmo com conjunto reduzido de features
- Maior estabilidade em predições de longo prazo

4.6 Análise Detalhada do Modelo XGBoost

4.6.1 Comportamento Geral dos Erros

A análise detalhada do XGBoost revela um padrão de erros bem definido, com erro médio de 3.675,76 litros e desvio padrão de 22.665,44 litros. O erro percentual médio de 21,99% indica uma precisão adequada para aplicações práticas, especialmente considerando a complexidade inerente à previsão de consumo de combustível em operações aéreas.

A distribuição assimétrica dos erros, evidenciada pela diferença significativa entre a média (3.675,76 litros) e a mediana (1.000,51 litros) dos erros absolutos, revela uma característica importante do modelo: enquanto a maioria das predições apresenta erros relativamente baixos (como indicado pela mediana), existem casos específicos onde o erro é substancialmente maior. Esta assimetria se mostra mais pronunciada em voos de curta distância, como evidenciado pela análise por faixas, sugerindo que fatores operacionais têm maior impacto relativo nestas operações.

4.6.2 Performance por Faixa de Distância

A análise por faixas de distância revela um padrão claro de melhoria na precisão conforme aumenta a distância do voo, como detalhado na Tabela 4.5.

TABELA 4.5 – Análise Detalhada de Erros por Faixa de Distância

Faixa	**Intervalo (km)**	**Observações**	**Erro (%)**	**Desvio (%)**	**Intervalo de Confiança 95%**
Q1	4-1.112	22.266	56,48	293,43	[29,20 - 83,76]
Q2	1.113-4.491	22.256	25,48	272,83	[13,20 - 37,76]
Q3	4.494-14.880	22.261	19,66	428,97	[6,11 - 33,21]
Q4	14.881-41.044	22.263	4,75	9,11	[3,30 - 6,20]
Q5	41.059-664.201	22.258	3,57	3,56	[2,75 - 4,39]

A análise revela um aspecto fundamental do modelo: sua precisão aumenta significativamente com a distância do voo. Voos no primeiro quintil (até 1.112 km) apresentam não apenas o maior erro médio (56,48%), mas também a maior variabilidade relativa ao tamanho do intervalo de distância. Em contraste, voos acima de 41.059 km (Q5) mostram notável estabilidade, com erro médio de apenas 3,57% e desvio padrão equivalente.

4.7 Padrões de Erro e Outliers

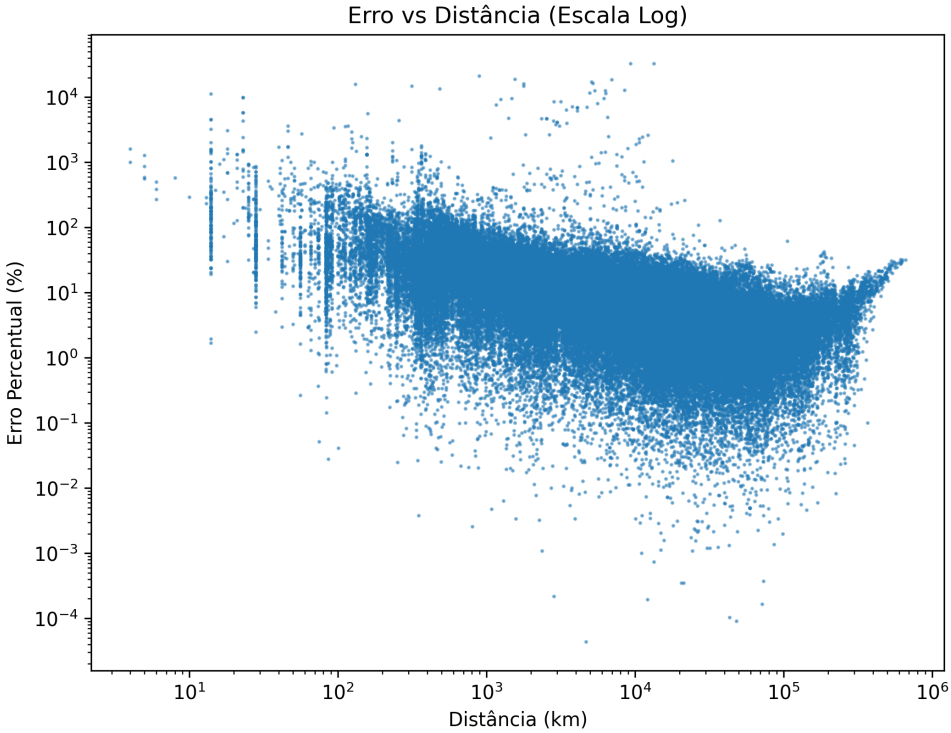


FIGURA 4.6 – Relação entre Erro Percentual e Distância em Escala Logarítmica

A análise da relação entre erro e distância em escala logarítmica (Figura 4.6) revela um padrão claro de convergência: quanto maior a distância do voo, menor e mais consistente

é o erro percentual de predição. Este comportamento forma um padrão característico de ”funil”, onde a dispersão dos erros diminui significativamente para voos mais longos.

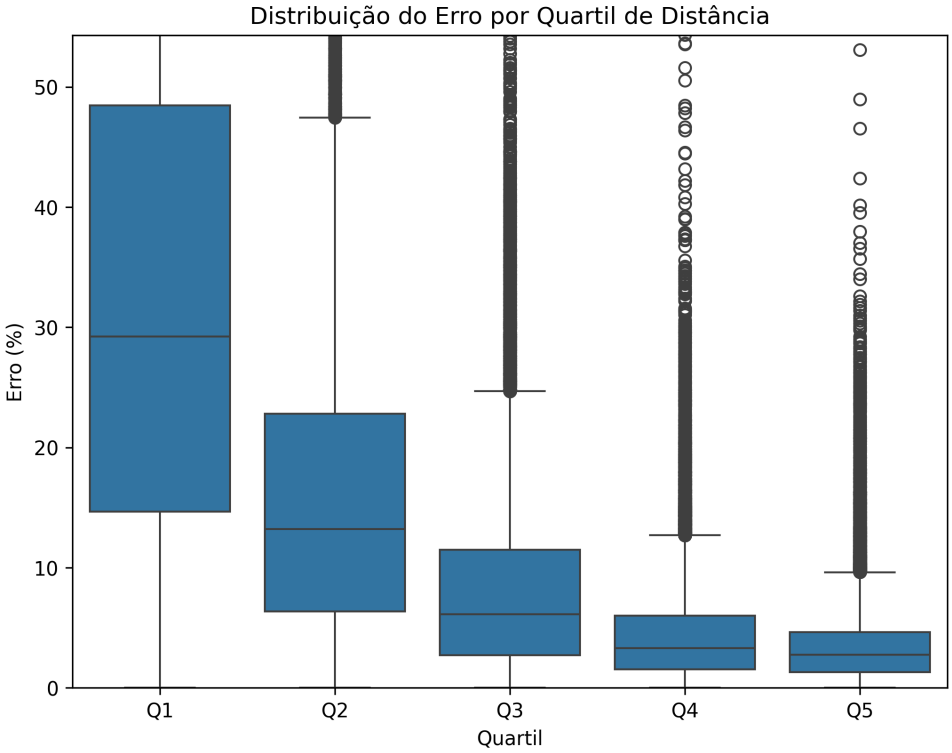


FIGURA 4.7 – Distribuição do Erro por Quartil de Distância

A distribuição dos erros por quartil de distância (Figura 4.7) confirma quantitativamente esta observação. Os outliers, representados pelos pontos além das hastes dos boxplots, são significativamente mais frequentes e extremos nos primeiros quartis de distância.

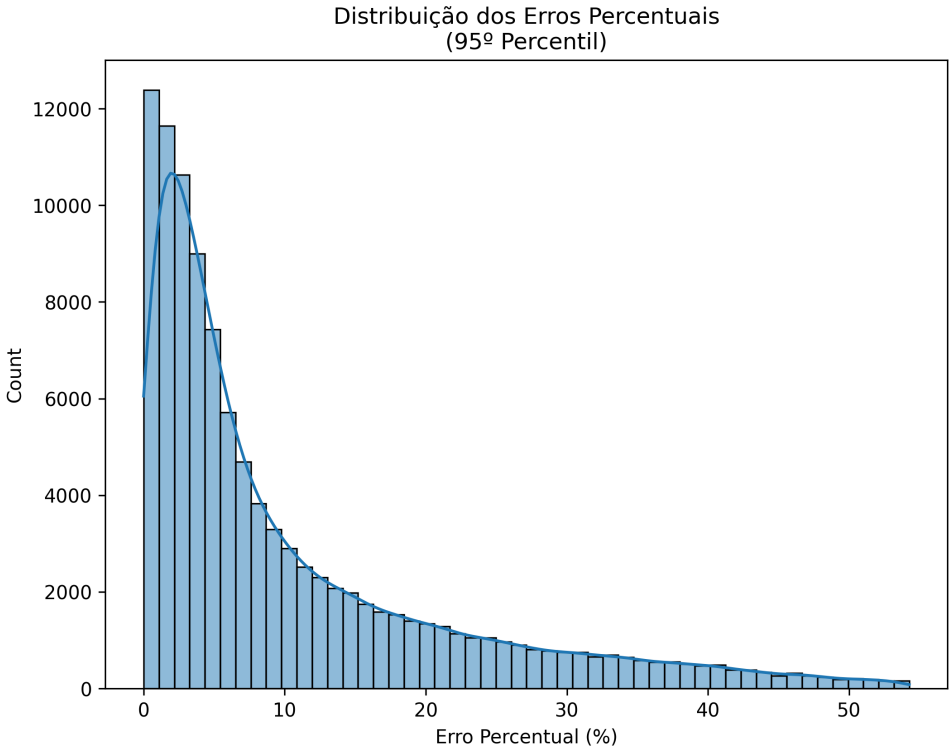


FIGURA 4.8 – Distribuição dos Erros Percentuais

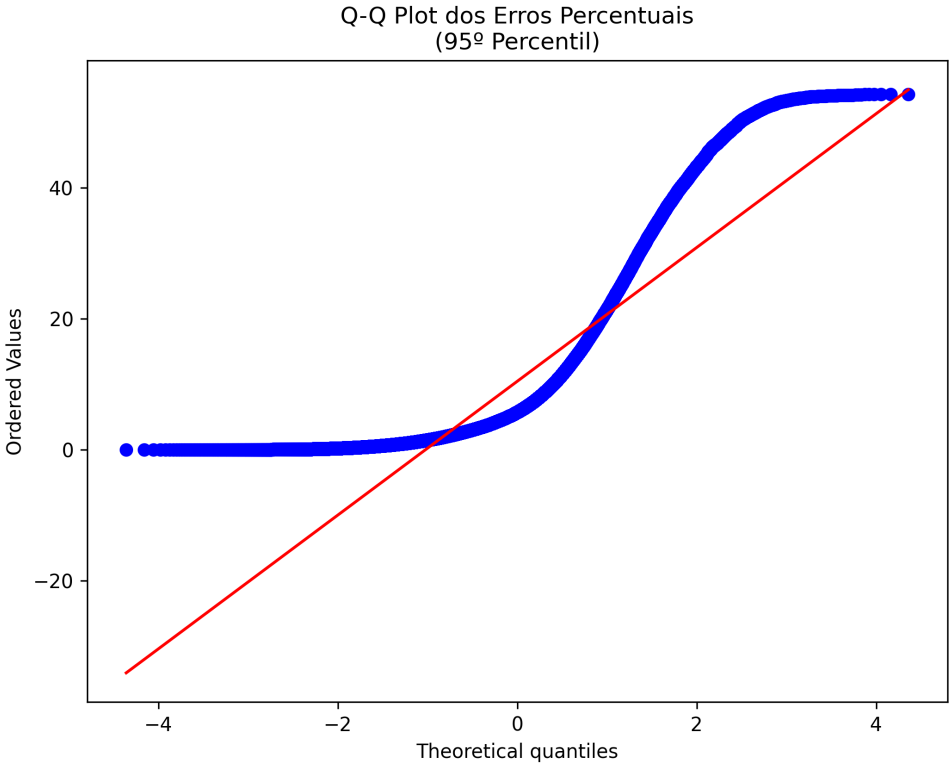


FIGURA 4.9 – Q-Q Plot dos Erros Percentuais

O histograma dos erros percentuais (Figura 4.8) e o Q-Q plot (Figura 4.9) evidenciam a natureza não-gaussiana dos erros de predição, com uma distribuição assimétrica e caudas

pesadas. Esta característica é típica de fenômenos operacionais complexos, onde eventos extremos ocorrem com frequência maior do que seria esperado em uma distribuição normal.

5 Conclusões

5.1 Síntese dos Resultados

5.1.1 Performance dos Modelos

Este trabalho desenvolveu e avaliou diversos modelos para predição de consumo de combustível em operações aéreas comerciais, com destaque para o XGBoost, que alcançou resultados superiores em todas as métricas avaliadas ($R^2 = 0,9908$, MAPE = 21,99%). A análise detalhada da performance revelou padrões importantes na precisão das predições:

- Maior acurácia em voos de longa distância (erro médio de 3,57% para voos acima de 41.000 km)
- Maior variabilidade em voos curtos (erro médio de 56,48% para voos até 1.112 km)
- Comportamento consistente em diferentes horizontes de planejamento

5.1.2 Contribuições Metodológicas

O estudo apresentou contribuições metodológicas significativas para a área de predição de consumo de combustível na aviação comercial. Primeiramente, foi desenvolvida uma estrutura robusta de preprocessamento específica para dados aeronáuticos, que inclui tratamento adequado de variáveis categóricas de alta cardinalidade e dados faltantes. Esta estrutura foi implementada através da classe `DataPreprocessor`, que estabelece um pipeline reprodutível para futuros estudos na área.

Em segundo lugar, foi implementado um sistema inovador de avaliação por horizontes temporais, permitindo análise da degradação de performance com informações parciais. Este sistema, materializado na classe `ModelsRunner`, possibilita a avaliação realista do desempenho dos modelos em diferentes cenários de planejamento operacional.

Por fim, foram criadas features derivadas que capturam aspectos operacionais relevantes, como a taxa de ocupação e eficiência operacional por segmento. Estas features

demonstraram valor significativo na melhoria da precisão dos modelos, como evidenciado pela análise de importância de variáveis.

5.1.3 Implicações Práticas

Os resultados obtidos têm implicações diretas para o planejamento operacional das companhias aéreas. A partir da análise estatística detalhada, foi possível estabelecer margens de segurança baseadas em evidências quantitativas, específicas para diferentes faixas de distância.

TABELA 5.1 – Margens de Segurança Recomendadas por Faixa de Distância

Faixa de Distância (km)	Erro Médio (%)	Margem Mínima (%)	Margem Recomendada (%)
0-1.112	56,48	29,20	83,76
1.113-4.491	25,48	13,20	37,76
4.494-14.880	19,66	6,11	33,21
14.881-41.044	4,75	3,30	6,20
41.059+	3,57	2,75	4,39

A identificação de padrões de consumo por faixa de distância permite um planejamento mais preciso e eficiente das operações. Particularmente, a constatação de que voos de longa distância apresentam maior previsibilidade (erro médio de 3,57%) possibilita um planejamento mais ajustado para estas operações.

5.2 Limitações do Estudo

5.2.1 Limitações dos Dados

O estudo apresenta algumas limitações importantes relacionadas aos dados utilizados:

- **Ausência de Informações Meteorológicas:** A falta de dados sobre condições climáticas, como velocidade e direção do vento, temperatura e pressão atmosférica, que são fatores conhecidos por impactar significativamente o consumo de combustível.
- **Granularidade Temporal:** Os dados disponíveis possuem granularidade mensal, não permitindo a captura de variações diárias ou horárias no consumo de combustível.
- **Informações das Aeronaves:** Ausência de detalhes específicos sobre as características das aeronaves, como idade, configuração específica e estado de manutenção.

5.2.2 Limitações Metodológicas

As principais limitações metodológicas identificadas são:

- **Foco em Regressão:** O estudo concentrou-se em modelos de regressão, não explorando completamente o potencial de abordagens classificatórias que poderiam fornecer insights complementares.
- **Simplificações Operacionais:** Alguns fatores operacionais complexos, como procedimentos específicos de decolagem e pouso, foram necessariamente simplificados no modelo.
- **Validação em Cenários Extremos:** Necessidade de maior validação do modelo em condições operacionais extremas ou não usuais.

5.3 Direções Futuras

5.3.1 Aprimoramentos Técnicos

Para desenvolvimento futuro do trabalho, sugere-se:

- Incorporação de variáveis meteorológicas e fatores operacionais adicionais
- Desenvolvimento de modelos híbridos combinando regressão e classificação
- Implementação de técnicas de análise de séries temporais para capturar padrões sazonais
- Exploração de arquiteturas de redes neurais profundas para captura de padrões mais complexos

5.3.2 Expansões do Escopo

O escopo do trabalho pode ser expandido para incluir:

1. **Análise de Impacto Ambiental:** Incorporação de métricas de emissões de CO_2 e outros gases de efeito estufa.
2. **Otimização de Frota:** Desenvolvimento de módulos para otimização da alocação de aeronaves com base nas previsões de consumo.
3. **Integração Operacional:** Criação de interfaces com sistemas existentes de planejamento operacional.

5.3.3 Aplicações Práticas

Para implementação prática dos resultados, propõe-se:

1. Desenvolvimento de uma interface de usuário para planejamento operacional
2. Criação de um sistema de alertas para desvios significativos no consumo
3. Integração com sistemas existentes de gestão de combustível

5.4 Considerações Finais

Este trabalho demonstrou a viabilidade de utilizar técnicas avançadas de aprendizado de máquina para prever com precisão o consumo de combustível em operações aéreas comerciais. O modelo XGBoost, em particular, alcançou níveis de precisão adequados para aplicações práticas ($R^2 = 0,9908$), especialmente em operações de média e longa distância.

A estrutura metodológica desenvolvida, incluindo o preprocessamento robusto dos dados e a avaliação por horizontes temporais, estabelece uma base sólida para futuros desenvolvimentos na área. As limitações identificadas, longe de diminuírem a relevância dos resultados, apontam caminhos claros para aprimoramentos futuros.

A aplicação prática das descobertas deste estudo tem o potencial de contribuir significativamente para a eficiência operacional do setor aéreo brasileiro. Com uma previsão mais precisa do consumo de combustível, as companhias aéreas podem otimizar suas operações, reduzir custos e minimizar seu impacto ambiental. A metodologia desenvolvida pode ser adaptada e expandida para outros contextos operacionais, contribuindo para o avanço contínuo da gestão eficiente de recursos na aviação comercial.

Referências

ANAC. **Anuário do Transporte Aéreo 2019**. Brasília, 2020.

ANAC. **Anuário do Transporte Aéreo 2022**. Brasília, 2023.

ANDERSON, J. D. **Aircraft Performance and Design**. New York: McGraw-Hill Education, 2017.

HONG, B.; GAN-XIANG, S.; XIAO-DONG, W. The prediction of aircraft fuel consumption based on flight quick access recorder data. *In*: **International Conference on Mechanical Engineering and Control Systems**. **Proceedings** [...]. [*S.l.*]: IEEE, 2014. p. 1–5.

HORIGUCHI, Y. *et al.* Predicting fuel consumption and flight delays for low-cost airlines. *In*: **Proceedings of the AAAI Conference on Artificial Intelligence**. **Proceedings** [...]. [*S.l.*]: AAAI Press, 2017. p. 4686–4693.

LI, Y. Machine learning applications in aircraft fuel consumption prediction. **Journal of Aircraft**, v. 47, n. 6, p. 2085–2096, 2010.

STOLZER, A. J. Fuel consumption modeling of a transport category aircraft using flight operations quality assurance data: A literature review. **Journal of Air Transportation**, v. 7, n. 1, p. 93–102, 2002.

FOLHA DE REGISTRO DO DOCUMENTO

1. CLASSIFICAÇÃO/TIPO	2. DATA	3. REGISTRO N°	4. N° DE PÁGINAS
TC	26 de novembro de 2024	DCTA/ITA/TC-133/2024	45

5. TÍTULO E SUBTÍTULO:

Predição de consumo de combustível por voo em aviação comercial.

6. AUTOR(ES):

Lucas Melo de Oliveira

7. INSTITUIÇÃO(ÕES)/ÓRGÃO(S) INTERNO(S)/DIVISÃO(ÕES):

Instituto Tecnológico de Aeronáutica – ITA

8. PALAVRAS-CHAVE SUGERIDAS PELO AUTOR:

1.Machine Learning 2.Aviação

9.PALAVRAS-CHAVE RESULTANTES DE INDEXAÇÃO:

Consumo de combustível; Predição; Aprendizado (inteligência artificial); Aviação comercial; Transporte aéreo; Transportes.

10. APRESENTAÇÃO: **(X) Nacional ( ) Internacional**

ITA, São José dos Campos. Curso de Graduação em Engenharia Civil-Aeronáutica. Orientador: Evandro José da Silva; Apresentação em 21/11/2021. Publicada em 2024

11. RESUMO:

Este trabalho tem como objetivo desenvolver modelos preditivos para estimar o consumo de combustível em voos comerciais, utilizando técnicas de machine learning. Para tal, trabalhou-se com dados operacionais detalhados, realizando a preparação, tratamento de variáveis e seleção das mais relevantes para a modelagem. Empreguei diferentes modelos de aprendizado de máquina, avaliando o desempenho com métricas adequadas.

12. GRAU DE SIGILO:

(X) OSTENSIVO ( ) RESERVADO ( ) SECRETO