

# Precificação de Passagens Aéreas

Lhayana V. de Macedo<sup>1</sup>

<sup>1</sup>Instituto Metr pole Digital – Universidade Federal do Rio Grande do Norte (UFRN)

lhayana.vieira.073@ufrn.edu.br

**Abstract.** *This study investigates the application of machine learning algorithms for predicting airfare prices, aiming to identify the model that best fits the data and provides the most accurate forecasts. Using data collected over 50 days of flights between major cities in India, three tree-based algorithms (XGBoost, LightGBM, and CatBoost) and one linear model (Ridge Regression) were tested. The model evaluation relied on error metrics such as the Mean Absolute Percentage Error (MAPE), which facilitates relative performance interpretation. The results indicate that XGBoost achieved the lowest MAPE and the highest stability in predictions, making it the most suitable model for pricing issues, with potential benefits for both the airline industry and consumers.*

**Resumo.** *Este estudo investiga a aplica  o de algoritmos de aprendizado de m quina na previs o de pre os de passagens a reas, buscando identificar o modelo que melhor se adequa aos dados e oferece previs es mais precisas. Utilizando dados coletados durante 50 dias de voos entre principais cidades da  ndia, foram testados tr s algoritmos baseados em  rvores de decis o (XGBoost, LightGBM e CatBoost) e um modelo linear (Regress o Ridge). A avalia  o dos modelos baseou-se em m tricas de erro como o Mean Absolute Percentage Error (MAPE), que facilita a interpreta  o relativa do desempenho. Os resultados indicam que o XGBoost apresentou o menor MAPE e maior estabilidade nas previs es, sendo o modelo mais adequado para o problema de precifica  o, com potencial de beneficiar o setor a reo e consumidores.*

## 1. Introdu  o

A precifica  o de passagens a reas   um processo din mico e complexo, influenciado por uma multiplicidade de fatores, como demanda, disponibilidade, sazonalidade e caracter sticas espec ficas de cada voo. A volatilidade dos pre os torna a previs o desse mercado desafiadora e essencial para companhias a reas e consumidores que buscam otimizar suas decis es financeiras. Com a ascens o de tecnologias de aprendizado de m quina, novas abordagens v m sendo exploradas para melhorar a previs o de pre os, aproveitando a capacidade desses modelos de capturar padr es complexos e n o-lineares nos dados.

Neste estudo, utilizamos t cnicas de aprendizado de m quina para desenvolver modelos preditivos que visam melhorar a acur cia na previs o de pre os de passagens a reas. Foram analisados dados coletados de uma plataforma de reservas durante um per odo de 50 dias, abrangendo voos nas classes econ mica e executiva entre as principais cidades metropolitanas da  ndia. Os dados inclu ram informa  es detalhadas sobre as caracter sticas dos voos e das rotas, que foram tratadas e estruturadas para maximizar o desempenho dos modelos.

Foram implementados e comparados quatro algoritmos de aprendizado de máquina: XGBoost, LightGBM, CatBoost (modelos baseados em árvores de decisão) e Regressão Ridge (modelo linear). Estes algoritmos foram escolhidos por suas distintas abordagens de modelagem e características específicas, como a eficiência em lidar com variáveis categóricas e a capacidade de evitar sobreajuste. A análise de desempenho foi baseada em múltiplas métricas de erro, incluindo o Mean Absolute Percentage Error (MAPE), que permite uma interpretação intuitiva e relativa do erro em relação aos valores reais.

Ao integrar técnicas de pré-processamento, como padronização e codificação de variáveis categóricas, em um pipeline de modelagem, buscamos assegurar a consistência e robustez dos modelos. Este estudo contribui para o entendimento de como diferentes algoritmos de aprendizado de máquina podem ser aplicados à precificação de passagens aéreas.

## 2. Base de Dados

### 2.1. Fonte dos Dados

Os dados utilizados neste trabalho foram coletados da plataforma de viagens *EaseMyTrip*, utilizando a ferramenta de raspagem *Octoparse*. A coleta de dados abrangeu um período de 50 dias, de 11 de fevereiro a 31 de março de 2022. A base de dados final, utilizada após limpeza, conta com 300.261 registros únicos de opções de voos entre as principais seis cidades metropolitanas da Índia.

### 2.2. Características dos Dados

O conjunto de dados inclui 11 variáveis, descritas a seguir, que representam diferentes aspectos das opções de voos:

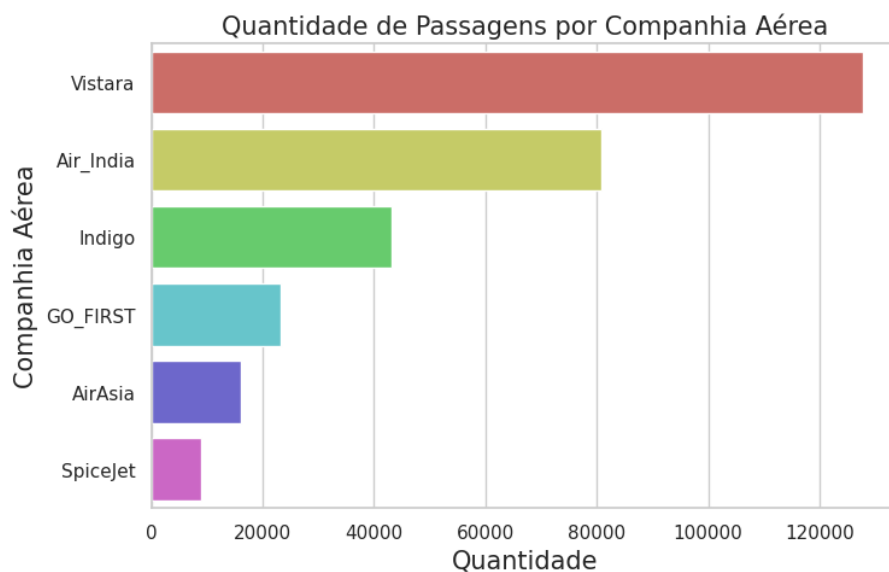
- **Airline:** Nome da companhia aérea. É uma variável categórica com 6 categorias distintas de companhias aéreas.
- **Flight:** Código de identificação do voo, fornecendo informações sobre o código específico da aeronave.
- **Source City:** Cidade de origem do voo (a cidade de partida). Esta variável categórica possui 6 cidades únicas representadas.
- **Departure Time:** Horário de partida categorizado em períodos específicos do dia, agrupados em 6 categorias.
- **Stops:** Número de escalas entre as cidades de origem e destino, sendo uma variável categórica com 3 valores distintos.
- **Arrival Time:** Horário de chegada categorizado em intervalos específicos, similar à variável de partida, com 6 rótulos distintos.
- **Destination City:** Cidade de destino do voo (cidade de chegada). Esta variável categórica também possui 6 cidades únicas.
- **Class:** Classe do assento, representada de forma categórica com duas categorias: Classe Executiva (Business) e Classe Econômica (Economy).
- **Duration:** Tempo total de duração do voo, em horas.
- **Days Left:** Número de dias restantes para a viagem, calculado como a diferença entre a data da viagem e a data da reserva.

- **Price:** Variável alvo (target), que armazena o preço da passagem, em Rupias Indianas (INR).

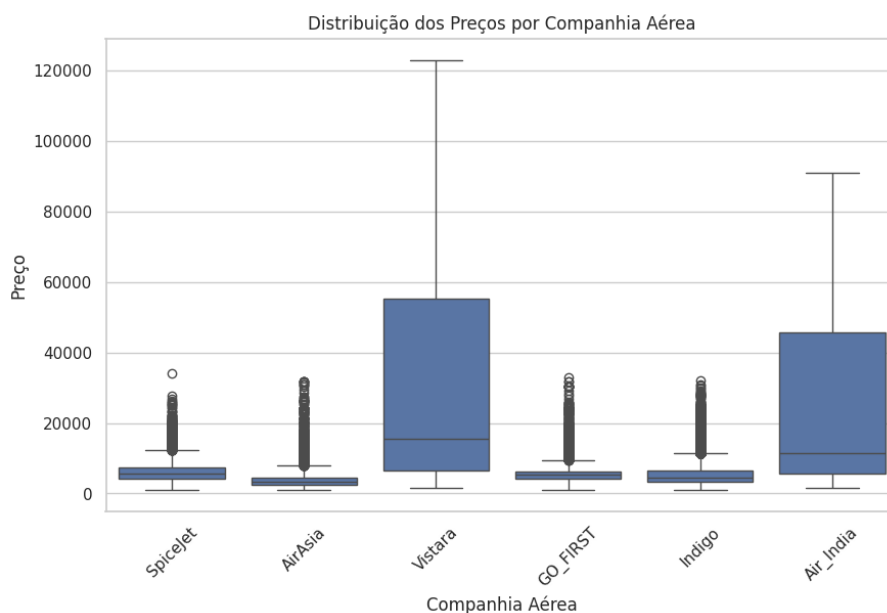
O pré-processamento inicial dos dados envolveu a limpeza e a criação de variáveis adicionais para enriquecer a análise. A coluna route foi criada ao combinar as informações de cidade de origem (source city) e cidade de destino (destination city), permitindo uma análise mais direta das rotas de voo. A classificação dos horários de partida e chegada em períodos específicos (categorizados) ajudou a simplificar a análise temporal dos preços. Todos os valores ausentes foram tratados, e as variáveis categóricas foram transformadas utilizando métodos de codificação específicos para cada modelo, garantindo que os dados estivessem adequados para o processo de modelagem e análise preditiva.

### 3. Análise Exploratória dos Dados

A análise exploratória inicial focou nas companhias aéreas e seu impacto no mercado e nos preços das passagens. O gráfico de quantidade de passageiros por companhia aérea revelou a liderança de mercado da Vistara, que transporta um número significativamente maior de passageiros em comparação com as demais companhias. A Air India ocupa a segunda posição, seguida pela Indigo, enquanto as companhias Go First, AirAsia e SpiceJet apresentam volumes menores de passageiros.

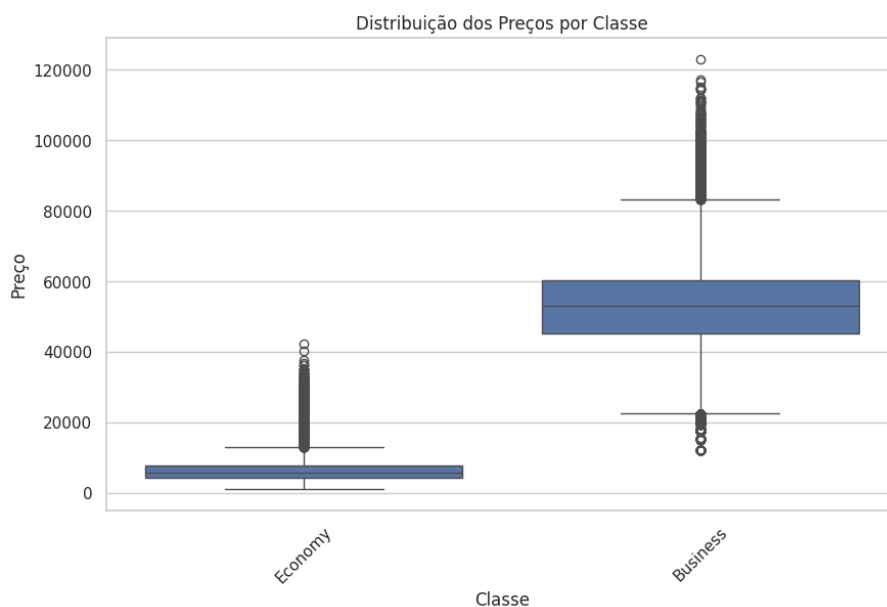


Para entender melhor a relação entre a quantidade de passageiros e os preços praticados, analisamos a distribuição dos preços por companhia aérea.



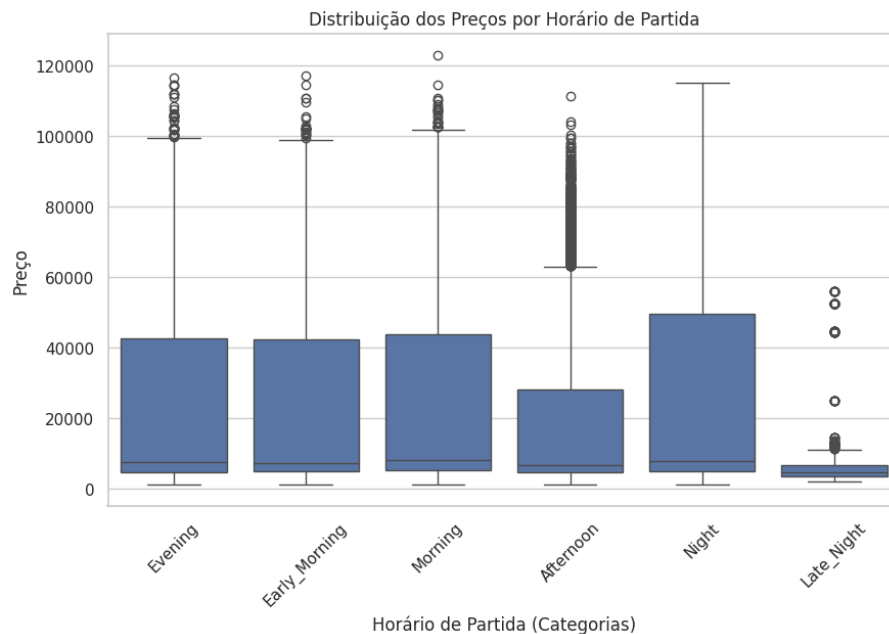
O gráfico revela que a Vistara não só lidera em número de passageiros, mas também apresenta o preço médio mais elevado, embora com alta variabilidade. Esse comportamento sugere uma estratégia de preços diversificada. Em contraste, a SpiceJet se caracteriza por oferecer os preços mais baixos com menor variabilidade, sugerindo uma política de preços mais uniforme.

A análise seguiu com a investigação da influência da classe do assento sobre os preços das passagens. Observou-se que as passagens da classe executiva, que incluem serviços adicionais e maior conforto, apresentavam valores significativamente superiores aos da classe econômica.

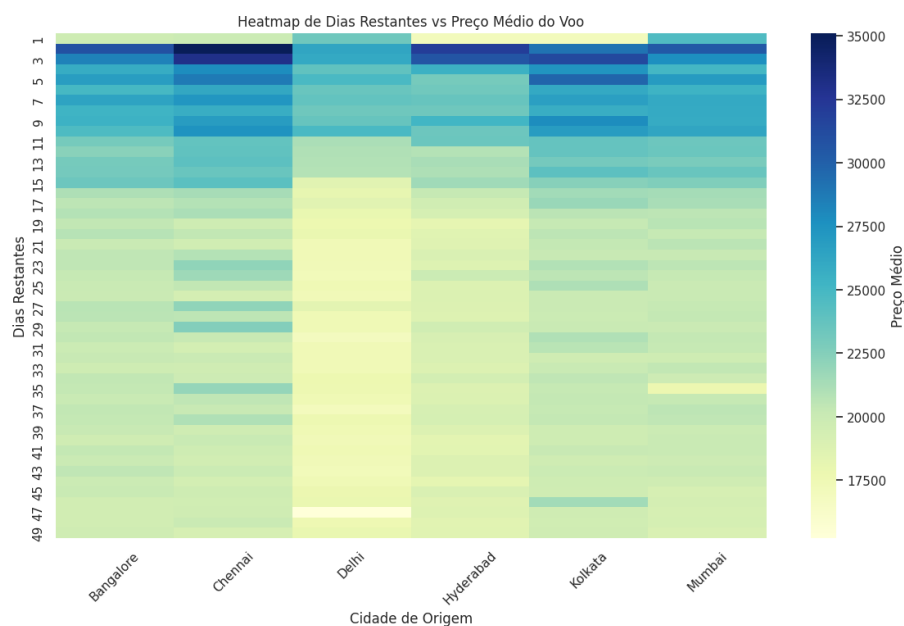


Essa diferença de preços, ainda que previsível, pôde ser quantificada, o que proporciona uma compreensão detalhada da segmentação de mercado entre as classes de assento e evidencia a estrutura de valor percebido pelos consumidores.

Na sequência, foi examinado o efeito do horário de partida sobre os preços das passagens. Os resultados indicaram que voos programados para o final da tarde e madrugada apresentaram preços médios mais baixos e menor variabilidade, possivelmente devido à menor demanda nesses horários. Essa característica sugere uma estratégia de precificação mais competitiva por parte das companhias aéreas para horários menos populares.



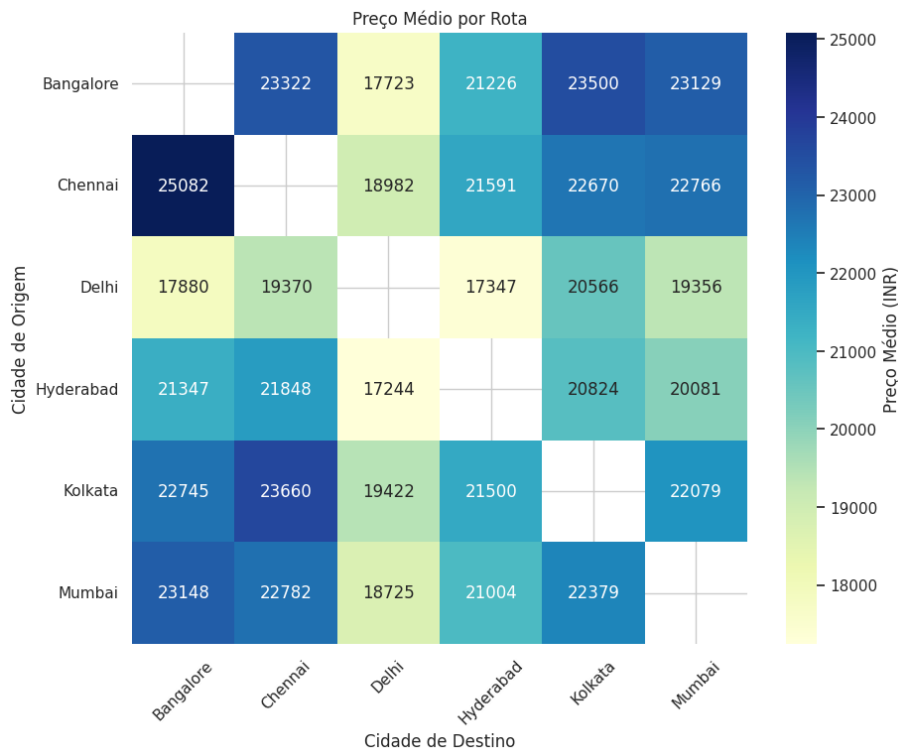
Além disso, a relação entre a proximidade da data da viagem e o preço médio foi examinada por meio de um heatmap.



A análise indicou que a flexibilidade na escolha da data da viagem pode resultar em preços mais acessíveis, uma vez que os valores tendem a aumentar conforme a data de embarque se aproxima. Observou-se também que essa tendência varia de acordo com

a cidade de origem, o que reforça a importância de uma análise individualizada das rotas.

Por fim, um heatmap comparativo dos preços médios entre diferentes rotas evidenciou uma ampla variação de preços, refletindo tanto a demanda quanto a competitividade específica de cada trajeto.



Essas análises constituem uma base sólida para entender as variáveis que influenciam a precificação de passagens aéreas, oferecendo suporte ao desenvolvimento de modelos preditivos.

#### 4. Preparação dos Dados

A etapa de preparação dos dados é fundamental para garantir a qualidade e a eficácia do modelo de machine learning. Nesta fase, as variáveis foram divididas em dois grupos: variáveis numéricas e variáveis categóricas. Essa divisão é essencial, pois diferentes tipos de variáveis demandam tratamentos distintos para serem utilizadas em algoritmos de machine learning. O uso adequado de transformadores para cada tipo de variável ajuda a garantir que o modelo seja capaz de capturar a complexidade dos dados e que produza previsões mais precisas.

Variáveis numéricas são variáveis que representam quantidades contínuas ou discretas, como duration (duração do voo) e days left (dias restantes para a viagem). Modelos de machine learning geralmente utilizam as variáveis numéricas como valores diretos para cálculos internos, mas a escala dessas variáveis pode afetar o desempenho do modelo. Por exemplo, variáveis com grandes diferenças de escala (como horas e dias) podem causar vieses no modelo, favorecendo variáveis com valores mais altos. Para contornar isso, técnicas de normalização ou padronização, como o StandardScaler, são aplicadas.

Variáveis categóricas são variáveis que representam categorias ou classes, como

airline (companhia aérea), source city (cidade de origem) e class (classe do assento). Essas variáveis precisam ser codificadas numericamente para serem compreendidas pelo modelo, pois a maioria dos algoritmos de machine learning não pode lidar com dados de texto diretamente. No entanto, ao transformar variáveis categóricas, é importante preservar a relação entre as categorias e a variável alvo (price), o que é feito usando técnicas de codificação direcionadas ao objetivo, como o TargetEncoder.

#### **4.1. Transformadores Utilizados na Preparação dos Dados**

O StandardScaler é uma técnica de padronização que transforma as variáveis numéricas para que tenham média zero e desvio padrão igual a um, esse processo ajuda a alinhar variáveis com diferentes escalas. Padronizar os dados facilita o aprendizado do modelo, reduzindo a sensibilidade a valores extremos e permitindo uma convergência mais rápida durante o treinamento.

Para codificar as variáveis categóricas, foi utilizado o TargetEncoder, uma técnica que atribui um valor numérico a cada categoria com base em uma média da variável alvo (price) para essa categoria específica. O TargetEncoder é particularmente eficaz em modelos baseados em árvores, como XGBoost, LightGBM e CatBoost, pois ele ajuda o modelo a capturar as relações entre as categorias e o preço. Em vez de transformar uma variável categórica em várias colunas (como no OneHotEncoder), o TargetEncoder mantém uma única coluna, o que reduz a dimensionalidade dos dados e previne problemas de esparsidade.

### **5. Modelagem**

#### **5.1. Seleção e Implementação de Modelos**

No estudo do problema de precificação de passagens aéreas, foram testados diversos algoritmos de aprendizado de máquina com o objetivo de identificar o modelo que melhor se adequa aos dados e oferece previsões precisas. Entre os modelos analisados, incluem-se três modelos baseados em árvores, o XGBoost, o LightGBM e o CatBoost, além de um modelo linear, a Regressão Ridge. A seguir, exploramos brevemente as principais características de cada um desses modelos e sua implementação.

O XGBoost (Extreme Gradient Boosting) é uma implementação avançada do método de boosting de gradiente, que constrói um modelo robusto a partir de uma sequência de árvores de decisão, corrigindo iterativamente os erros de previsões anteriores. Esse modelo é conhecido por sua eficiência e rapidez, pois utiliza otimizações como processamento paralelo e regularização L1 e L2, o que ajuda a evitar o sobreajuste. No caso do XGBoost, cada nova árvore é treinada para ajustar os erros das árvores anteriores, e o modelo final consiste na soma ponderada das previsões de todas as árvores. O XGBoost destaca-se por sua alta precisão e capacidade de generalização, além de sua flexibilidade para ajuste de hiperparâmetros, sendo implementado como uma etapa do pipeline que também inclui o pré-processamento dos dados, integrando a transformação e o treinamento de forma eficiente.

O LightGBM (Light Gradient Boosting Machine) é outro modelo de boosting de gradiente, otimizado para lidar com grandes conjuntos de dados. Diferente do XGBoost, o LightGBM usa uma abordagem chamada leaf-wise, que expande folhas com o maior ganho, ao invés de adotar a abordagem level-wise. Esse método permite ao modelo

alcançar maior precisão com menor tempo de processamento e uso de memória, criando árvores assimétricas que se adaptam melhor aos dados. O LightGBM é vantajoso por sua alta velocidade, baixo uso de memória e excelente desempenho com dados de alta dimensionalidade, além de lidar bem com valores ausentes [Larose 2015]. Para este estudo, o LightGBM foi testado dentro do mesmo pipeline de pré-processamento, garantindo consistência na comparação com outros modelos.

O CatBoost (Categorical Boosting) é um modelo de boosting de gradiente desenvolvido especificamente para lidar com variáveis categóricas, reduzindo a necessidade de pré-processamento extenso, como a aplicação de codificadores complexos. Utilizando uma técnica de “árvore simétrica”, o CatBoost constrói árvores de decisão onde todas as folhas estão no mesmo nível, o que melhora a eficiência do treinamento e reduz o risco de overfitting. [Prokhorenkova et al. 2018]

Por fim, a Regressão Ridge foi utilizada como modelo linear de comparação, sendo uma extensão da regressão linear que inclui penalização L2 para reduzir o sobreajuste e melhorar a generalização do modelo. Na Regressão Ridge, o objetivo é minimizar o erro quadrático médio enquanto adiciona um termo de regularização, o que impede que o modelo se ajuste excessivamente aos dados de treinamento, especialmente quando há multicolinearidade entre os recursos. Essa regularização reduz a magnitude dos coeficientes, promovendo um ajuste mais controlado [Hastie et al. 2009]. Para garantir consistência no pré-processamento, foi utilizado o OneHotEncoder para variáveis categóricas, visto que o modelo linear não pode utilizar codificadores como o TargetEncoder.

Todos os modelos foram integrados em um pipeline que incluiu o pré-processamento das variáveis (uso do StandardScaler para variáveis numéricas e do TargetEncoder ou OneHotEncoder para variáveis categóricas, conforme o modelo) e o treinamento. Esse pipeline permitiu modularizar o processo, garantindo as transformações necessárias para cada tipo de variável e mantendo a consistência ao comparar os resultados de cada modelo. Após o treinamento, os modelos foram avaliados com validação cruzada e métricas de erro (MAE, MSE, RMSE e MAPE), possibilitando uma análise objetiva do desempenho de cada abordagem no problema de precificação de passagens aéreas, e facilitando a identificação do modelo mais adequado.

## 5.2. Avaliação de Desempenho dos Modelos

Para avaliar a precisão dos modelos na previsão dos preços das passagens aéreas, foram utilizadas quatro métricas de erro: *Mean Absolute Error* (MAE), *Mean Squared Error* (MSE), *Root Mean Squared Error* (RMSE) e *Mean Absolute Percentage Error* (MAPE). Cada métrica oferece uma perspectiva diferente sobre o desempenho do modelo, auxiliando na análise detalhada da eficácia de cada algoritmo. Abaixo estão descrições de cada uma das métricas e como elas funcionam.

O Mean Absolute Error (MAE), ou erro absoluto médio, mede a média das diferenças absolutas entre os valores reais e as previsões do modelo. O MAE representa a magnitude média dos erros sem considerar a direção (positiva ou negativa), sendo uma métrica interpretável e fácil de entender. A fórmula do MAE é dada por

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



onde  $y_i$  é o valor real,  $\hat{y}_i$  é o valor predito e  $n$  é o número de amostras. Um MAE menor indica que as previsões estão mais próximas dos valores reais, em média. Como o MAE mede diretamente o valor médio do erro, ele é útil para entender o impacto real dos erros no contexto dos preços das passagens, sem ser influenciado por erros extremos.

O Mean Squared Error (MSE), ou erro quadrático médio, é a média dos quadrados dos erros entre os valores reais e as previsões do modelo. Por elevar os erros ao quadrado, o MSE penaliza erros grandes de forma mais intensa do que o MAE, o que o torna mais sensível a outliers. Sua fórmula é dada por

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Um MSE menor indica que o modelo comete menos erros grandes em média. Essa métrica é útil para modelos que precisam minimizar grandes desvios, sendo mais sensível a erros extremos. No entanto, o MSE não é interpretável diretamente no mesmo intervalo dos valores reais, devido ao efeito de elevação ao quadrado.

O Root Mean Squared Error (RMSE), ou raiz do erro quadrático médio, é a raiz quadrada do MSE. Essa métrica transforma o MSE de volta à mesma unidade dos valores reais, facilitando a interpretação dos erros no contexto dos dados originais. Sua fórmula é expressa por

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

Como o RMSE penaliza erros grandes devido ao uso da média dos quadrados, ele é útil quando queremos minimizar tanto o número quanto o impacto dos erros grandes, mas em uma escala comparável aos preços reais das passagens. Um RMSE menor indica previsões mais precisas e com menos erros extremos. [James et al. 2013]

Por fim, o Mean Absolute Percentage Error (MAPE), ou erro percentual absoluto médio, mede a média do erro absoluto como uma porcentagem dos valores reais, fornecendo uma perspectiva relativa do erro em relação ao valor real. A fórmula do MAPE é dada por

$$MAPE = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

O MAPE é útil para entender o erro em termos percentuais, o que facilita a interpretação de quão grande foi o erro em relação ao preço real. Um MAPE menor indica que as previsões estão mais próximas dos valores reais em termos percentuais, sendo ideal para comparar previsões em dados onde a escala do valor real pode variar muito. No entanto, o MAPE pode ser sensível a valores reais muito próximos de zero, o que pode gerar percentuais exageradamente altos. [Hyndman 2018]

Essas quatro métricas permitem uma análise detalhada do desempenho dos modelos, cada uma oferecendo uma perspectiva diferente. O MAE é fácil de interpretar e é

menos influenciado por outliers. O MSE e o RMSE penalizam mais os erros grandes, sendo úteis quando queremos minimizar o impacto de previsões extremamente distantes. O MAPE facilita a compreensão dos erros em termos percentuais, tornando a análise relativa. Ao utilizar essas métricas em conjunto, podemos avaliar melhor a precisão geral dos modelos e identificar qual abordagem fornece o melhor equilíbrio entre precisão e robustez em diferentes aspectos do problema de precificação de passagens aéreas.

5.3. Comparação dos Modelos e Discussão dos Resultados

Para comparar os modelos de forma eficaz, foi utilizada a métrica *Mean Absolute Percentage Error* (MAPE). O MAPE foi escolhido como a métrica de avaliação principal pela facilidade de interpretação, uma vez que representa o erro em termos percentuais. Esse tipo de métrica é particularmente útil quando o valor da variável alvo, neste caso o preço das passagens aéreas, pode ser de difícil interpretação. Neste caso, o preço está na moeda Rupia Indiana (INR), e como o valor em INR pode ser de difícil interpretação para pessoas fora da Índia, o MAPE oferece uma forma de análise independente da unidade monetária, facilitando a compreensão da precisão do modelo em relação aos preços reais.

Os resultados dos modelos testados são apresentados na Tabela 1. A média do MAPE (Mean MAPE) representa o erro percentual médio de cada modelo, enquanto o desvio padrão (Std MAPE) indica a variação dos erros entre as diferentes execuções da validação cruzada, ajudando a entender a estabilidade do modelo.

Modelo	Mean MAPE	Std MAPE
XGBoost	-0.1439	0.0010
LightGBM	-0.1721	0.0009
CatBoost	-0.1481	0.0006
Ridge	-0.4270	0.0019

Table 1. Resultados dos Modelos

A análise dos resultados indica que o modelo XGBoost apresentou o menor MAPE médio, de 14.39%, seguido pelo CatBoost, com um MAPE médio de 14.81%. Ambos os modelos demonstraram estabilidade nas previsões, com baixos desvios padrão (*Std MAPE*), indicando consistência na precisão do modelo ao longo das diferentes iterações da validação cruzada. Esse baixo desvio padrão sugere que esses modelos são robustos e menos suscetíveis a flutuações nos dados de entrada.

O modelo LightGBM obteve um MAPE médio de 17.21%, levemente superior ao dos modelos XGBoost e CatBoost, mas ainda dentro de um limite aceitável de erro percentual. Assim como os modelos anteriores, o LightGBM também apresentou um desvio padrão baixo, indicando boa estabilidade.

Por outro lado, o modelo de Regressão Ridge apresentou um desempenho significativamente inferior, com um MAPE médio de 42.70%, mais que o dobro dos outros modelos. Esse resultado indica que a Regressão Ridge não conseguiu capturar adequadamente as relações não-lineares nos dados, essenciais para prever o preço das passagens com precisão. Além disso, o desvio padrão mais elevado do Ridge (1.93%) sugere uma menor estabilidade do modelo, refletindo variações maiores nas previsões ao longo das validações cruzadas.

Dessa forma, o XGBoost foi selecionado como o melhor modelo para o problema de precificação das passagens aéreas, pois obteve o menor MAPE, com estabilidade nas previsões e baixa variação dos erros. O uso do MAPE permite que os resultados sejam facilmente interpretados em qualquer unidade monetária, proporcionando uma visão clara da precisão do modelo para diferentes públicos. A combinação de baixo MAPE e estabilidade torna o XGBoost a escolha mais apropriada para prever preços de passagens aéreas com precisão.

## 6. Conclusão

Este projeto teve como objetivo desenvolver um modelo preditivo para precificação de passagens aéreas, utilizando técnicas de machine learning para capturar e prever a variação de preços com base em diversos fatores. A partir de um conjunto de dados coletado ao longo de 50 dias, com informações sobre voos entre as principais cidades metropolitanas da Índia, foram realizadas as etapas de análise exploratória, preparação de dados, seleção e avaliação de diferentes modelos.

Na fase de análise, observou-se que variáveis como companhia aérea, cidade de origem e destino, tempo de antecedência para a viagem, duração do voo e número de escalas têm uma influência significativa sobre o preço final das passagens. Para modelagem, foram testados quatro algoritmos: XGBoost, LightGBM, CatBoost e Regressão Ridge.

O modelo XGBoost destacou-se como a melhor opção, apresentando o menor erro percentual médio (*Mean Absolute Percentage Error* - MAPE) de 14,39%, com um desvio padrão baixo, indicando estabilidade nas previsões. O MAPE foi utilizado como métrica principal devido à facilidade de interpretação, especialmente porque o valor da variável alvo está em Rupias Indianas, unidade monetária que pode ser de difícil interpretação para públicos fora da Índia. O uso do MAPE permitiu avaliar o desempenho dos modelos em termos percentuais, facilitando a compreensão do erro médio das previsões em relação aos valores reais.

Os resultados deste projeto mostram que o modelo XGBoost é capaz de prever os preços das passagens aéreas com precisão, proporcionando um erro percentual reduzido e consistência nas previsões. Futuras melhorias podem incluir a incorporação de variáveis adicionais, ajuste mais fino de hiperparâmetros e o uso de técnicas de ensemble para potencializar ainda mais a precisão do modelo.

## References

- Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.
- Hyndman, R. (2018). *Forecasting: principles and practice*. OTexts.
- James, G., Witten, D., Hastie, T., Tibshirani, R., et al. (2013). *An introduction to statistical learning*, volume 112. Springer.
- Larose, D. T. (2015). *Data mining and predictive analytics*. John Wiley & Sons.
- Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. (2018). Catboost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31.