Análise de Dados para Hotéis: Estratégias de Previsão para Redução de Cancelamentos e Prejuízos

Luís Felipe Barros Pacheco, Rodrigo Moreira Marinho, Erick Keven da Silva Alves, Másio César

[1]Instituto de Computação – Universidade Federal de Alagoas (UFAL)

{lfbp@ic.ufal.br, rmm@ic.ufal.br, eksa@ic.ufal.br, mccm@ic.ufal.br}

Resumo

No setor hoteleiro, a análise de dados de reservas é uma ferramenta estratégica para entender e prever a demanda dos hóspedes, permitindo uma gestão mais eficiente e alinhada às necessidades dos clientes. A aplicação de ciência de dados oferece a oportunidade de transformar informações históricas em insights valiosos, que auxiliam na otimização de tarifas, no planejamento de recursos e na melhoria da experiência dos hóspedes.

Abstract

In the hotel sector, reservation data analysis is a strategic tool for understanding and predicting guest demand, allowing for more efficient management aligned with customer needs. The application of data science offers the opportunity to transform historical information into valuable insights that help optimize rates, resource planning and improve the guest experience.

1. Introdução

A análise de dados de reservas é essencial no setor hoteleiro para entender e prever a demanda, possibilitando uma gestão mais eficiente e focada nas necessidades dos clientes. Neste estudo, analisamos dados de reservas em um hotel urbano é um resort, explorando variáveis como tempo de reserva, duração da estadia e perfil dos hóspedes. Usamos técnicas de aprendizado de máquina e métodos estatísticos para identificar padrões de demanda e prever

picos de ocupação. O objetivo é mostrar como a ciência de dados pode melhorar a precisão das previsões e fundamentar decisões estratégicas para otimizar a operação e a experiência dos hóspedes.

2. Problema de Negócio

A renda de um hotel depende da demanda dos hóspedes, e reduzir o cancelamento de reservas é crucial para otimizar essa renda. Analisar os fatores que levam ao cancelamento permite minimizar prejuízos financeiros e implementar estratégias preventivas, como políticas de cancelamento mais rígidas ou incentivos para manter reservas. Duas questões principais orientam essa análise:

2.1 Qual a probabilidade de um cliente cancelar? Isso ajuda o hotel a adotar medidas para reduzir cancelamentos.

2.2 Quais fatores contribuem para os cancelamentos? Compreender esses fatores permite ao hotel ajustar seu serviço, aumentando a satisfação e a fidelidade dos hóspedes.

3. Descrição dos Dados

	count	mean	std	min	25%	50%	75%	max
is_canceled	119390.0	0.370416	0.482918	0.00	0.00	0.000	1.0	1.0
lead_time	119390.0	104.011416	106.863097	0.00	18.00	69.000	160.0	737.0
arrival_date_year	119390.0	2016.156554	0.707476	2015.00	2016.00	2016.000	2017.0	2017.0
arrival_date_week_number	119390.0	27.165173	13.605138	1.00	16.00	28.000	38.0	53.0
arrival_date_day_of_month	119390.0	15.798241	8.780829	1.00	8.00	16.000	23.0	31.0
stays_in_weekend_nights	119390.0	0.927599	0.998613	0.00	0.00	1.000	2.0	19.0
stays_in_week_nights	119390.0	2.500302	1.908286	0.00	1.00	2.000	3.0	50.0
adults	119390.0	1.856403	0.579261	0.00	2.00	2.000	2.0	55.0
children	119386.0	0.103890	0.398561	0.00	0.00	0.000	0.0	10.0
babies	119390.0	0.007949	0.097436	0.00	0.00	0.000	0.0	10.0
is_repeated_guest	119390.0	0.031912	0.175767	0.00	0.00	0.000	0.0	1.0
previous_cancellations	119390.0	0.087118	0.844336	0.00	0.00	0.000	0.0	26.0
previous_bookings_not_canceled	119390.0	0.137097	1.497437	0.00	0.00	0.000	0.0	72.0
booking_changes	119390.0	0.221124	0.652306	0.00	0.00	0.000	0.0	21.0
agent	103050.0	86.693382	110.774548	1.00	9.00	14.000	229.0	535.0
company	6797.0	189.266735	131.655015	6.00	62.00	179.000	270.0	543.0
days_in_waiting_list	119390.0	2.321149	17.594721	0.00	0.00	0.000	0.0	391.0
adr	119390.0	101.831122	50.535790	-6.38	69.29	94.575	126.0	5400.0
required_car_parking_spaces	119390.0	0.062518	0.245291	0.00	0.00	0.000	0.0	8.0
total_of_special_requests	119390.0	0.571363	0.792798	0.00	0.00	0.000	1.0	5.0

Os dados descrevem diversas características das reservas de hotel, como o número de adultos, crianças e bebês, tempos de espera, solicitações especiais e cancelamentos. Esses dados são essenciais para análises e tomada de decisões no gerenciamento de hotéis, auxiliando na otimização de recursos,

previsão de demanda e melhoria dos serviços oferecidos.

4. Limpeza de Dados

O conjunto de dados de reservas de hotel apresentava valores ausentes em várias colunas, conforme listado abaixo:

- children: 4 valores ausentes
- country: 488 valores ausentes
- agent: 16.340 valores ausentes
- company: 112.593 valores ausentes

Esses valores ausentes podem afetar a qualidade das análises e modelos, pois dados incompletos podem levar a resultados imprecisos e enviesados. Portanto, a identificação e tratamento desses valores é uma etapa essencial na preparação do conjunto de dados para análises subsequentes.

4.1 Estratégias de Limpeza

Para tratar esses valores ausentes, adotamos as seguintes abordagens:

- **Colunas Numéricas:** Valores ausentes foram preenchidos com a mediana da respectiva coluna. A mediana é uma medida robusta e menos sensível a outliers.
- **Colunas Categóricas:** Valores ausentes foram preenchidos com a moda da respectiva coluna, ou seja, o valor mais frequente.

4.2 Resultados da Limpeza

Após a aplicação das estratégias de limpeza, todas as colunas ficaram sem valores ausentes. A transformação específica pode ser vista nos exemplos a seguir:

- A coluna children passou de 4 valores ausentes para 0.
- A coluna country passou de 488 valores ausentes para 0.
- A coluna agent passou de 16.340 valores ausentes para 0.
- A coluna company passou de 112.593 valores ausentes para 0.

4.3 Impacto da Limpeza

A limpeza de dados teve os seguintes impactos:

1. **Melhoria na Qualidade dos Dados:** A ausência de valores nulos garante que todas as operações e análises sejam baseadas em dados completos, aumentando a confiabilidade dos resultados.
2. **Facilitação de Análises Estatísticas:** Com todos os dados presentes, análises estatísticas se tornam mais robustas e representativas da realidade.
3. **Preparação para Modelos de Machine Learning:** A limpeza dos dados prepara o conjunto de dados para ser utilizado em algoritmos de machine learning, que geralmente não aceitam valores nulos ou funcionam melhor com dados completos.
4. **Evitando Erros:** A ausência de valores nulos evita erros e exceções que poderiam ocorrer durante a execução de scripts de análise de dados.

5. Visualização dos dados

A visualização dos dados é crucial para entender padrões e relações entre variáveis. A seguir, discutiremos a matriz de correlação e o gráfico de cancelamento dos dois tipos de hotéis.

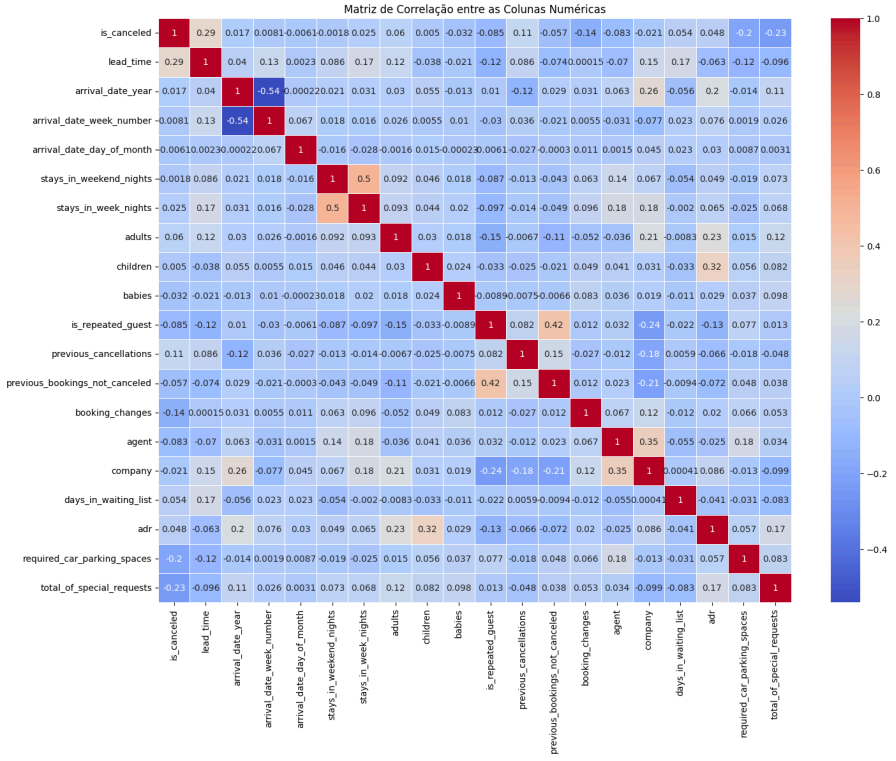
5.1 Matriz de Correlação

A matriz de correlação entre as colunas numéricas apresenta informações importantes sobre como diferentes variáveis se relacionam entre si. Alguns pontos a serem observados incluem:

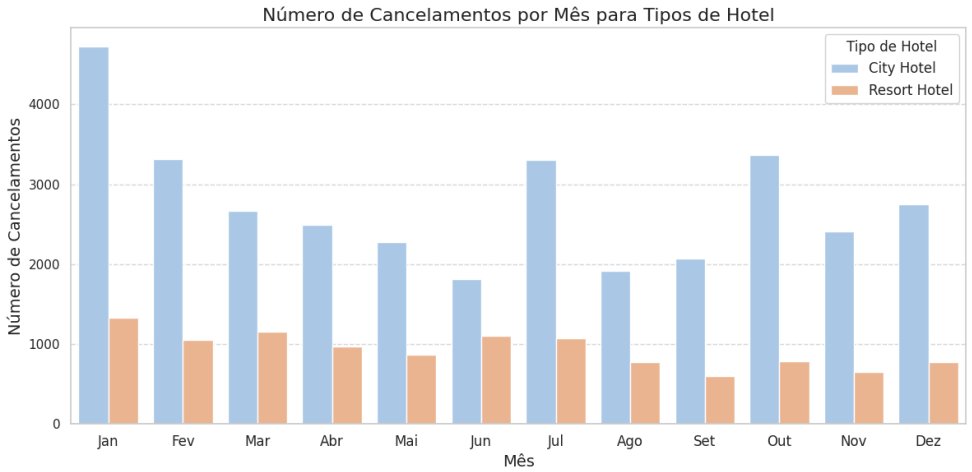
- Lead Time: Esta coluna tem uma correlação positiva com o cancelamento (is_canceled), sugerindo que quanto maior o tempo de antecedência da reserva, maior a probabilidade de cancelamento.
- Required Car Parking Spaces: Esta variável mostra uma correlação negativa com o cancelamento, indicando que reservas que necessitam de estacionamento são menos propensas a serem canceladas.

- Previous Bookings Not Canceled: Existe uma correlação negativa significativa com o cancelamento, sugerindo que hóspedes com histórico de reservas não canceladas são menos propensos a cancelar.
- Previous Cancellations: Apresenta uma correlação positiva com o cancelamento, indicando que hóspedes com histórico de cancelamentos anteriores são mais propensos a cancelar novamente.

Essas correlações ajudam a identificar fatores que podem influenciar o comportamento dos hóspedes em relação ao cancelamento das reservas.



5.2 Gráfico de Cancelamento por Mês para Tipos de Hotel



O gráfico de cancelamento mostra o número de cancelamentos por mês para os dois tipos de hotéis: City Hotel e Resort Hotel. As seguintes observações podem ser feitas:

- City Hotel: Apresenta um número consistentemente maior de cancelamentos em comparação com o Resort Hotel. Janeiro é o mês com o maior número de cancelamentos, seguido por outubro e julho.
- Resort Hotel: Tem um padrão mais estável de cancelamentos ao longo do ano, com picos menores em meses específicos. Janeiro e julho são os meses com os maiores números de cancelamentos.

Essas informações podem ser utilizadas para ajustar estratégias de marketing e políticas de reserva, visando reduzir a taxa de cancelamento em períodos críticos e melhorar a ocupação dos hotéis.

6. Métricas Usadas para Avaliar os Métodos

6.1 F1 Score

O F1 Score é uma métrica que combina a precisão (precision) e a revocação (recall) em um único número. É especialmente útil em situações onde há um

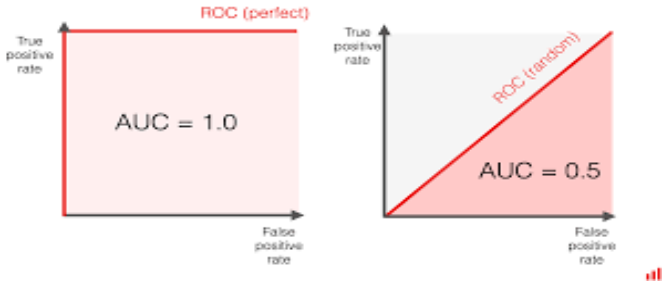
desbalanceamento entre as classes, pois considera tanto os falsos positivos quanto os falsos negativos.

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Um F1 Score próximo de 1 indica um modelo que é eficaz em prever a classe positiva, enquanto um valor próximo de 0 sugere que o modelo não está capturando bem os exemplos da classe positiva.

6.2 AUC (Area Under the ROC Curve)

A AUC mede a capacidade do modelo de discriminar entre as classes. A curva ROC (Receiver Operating Characteristic) plota a taxa de verdadeiros positivos (TPR) contra a taxa de falsos positivos (FPR) em diferentes limiares de decisão.



A AUC é a área sob essa curva. Um valor de AUC de 0,5 indica que o modelo não tem poder discriminatório (equivalente a uma escolha aleatória), enquanto um valor de 1,0 indica um modelo perfeito.
Quanto maior o valor da AUC, melhor o modelo é capaz de classificar corretamente as classes

6.3 Accuracy

A precisão (accuracy) é a proporção de previsões corretas em relação ao total de previsões. É uma métrica simples e frequentemente usada, mas pode ser enganosa em conjuntos de dados desbalanceados.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Onde:

TP = Verdadeiros Positivos
TN = Verdadeiros Negativos
FP = Falsos Positivos
FN = Falsos Negativos

Um valor de precisão próximo de 1 indica que a maioria das previsões está correta, enquanto valores muito baixos indicam que o modelo falha em prever corretamente a maioria das instâncias.

7. Métodos de Aprendizagem de Máquina

7.1 Árvore de Decisão (Decision Tree)

- **Descrição:** Utiliza uma estrutura em forma de árvore para modelar decisões, dividindo os dados em subconjuntos baseados em perguntas sobre os atributos. Cada nó interno representa uma pergunta sobre um atributo, e cada folha representa um resultado.

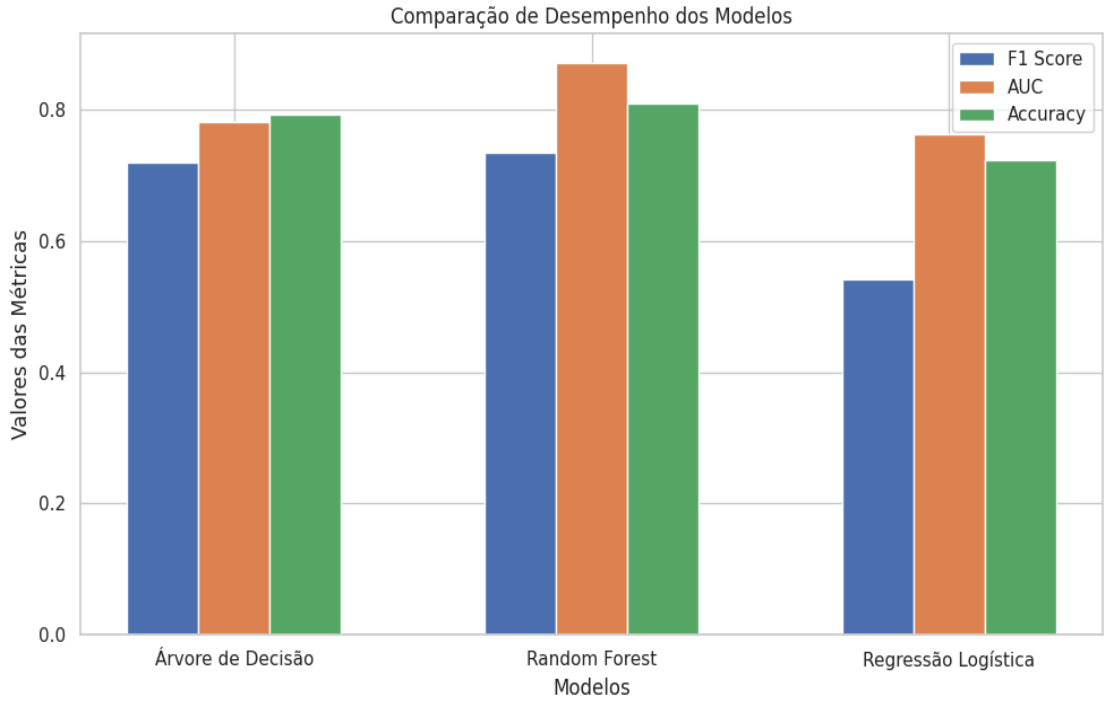
7.2 Random Forest

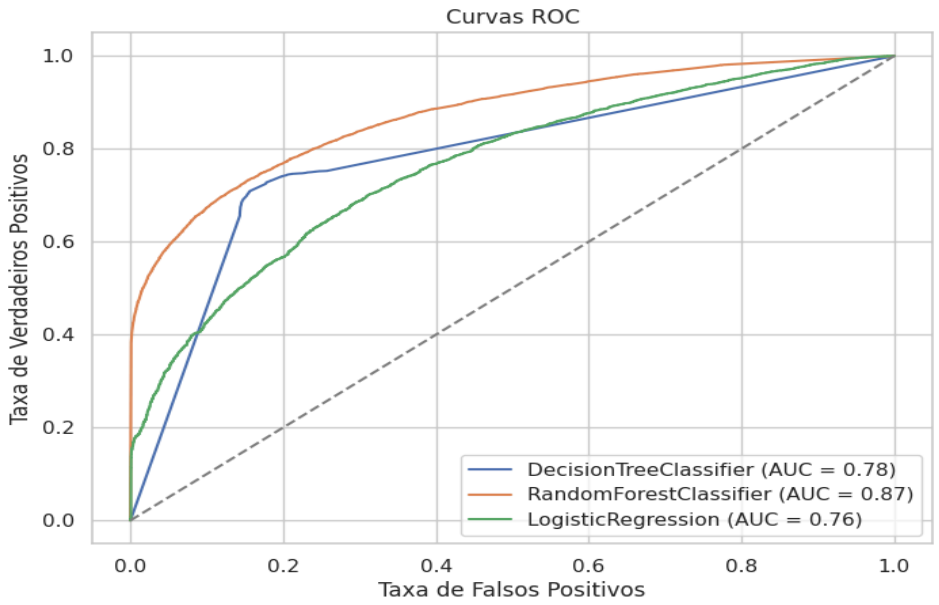
- **Descrição:** Um ensemble (conjunto) de múltiplas árvores de decisão, onde cada árvore é treinada em uma amostra aleatória do conjunto de dados. As previsões são feitas através da média das previsões de todas as árvores, melhorando a robustez e a precisão do modelo.

7.3 Regressão Logística

- **Descrição:** Um modelo estatístico que usa uma função logística para modelar a probabilidade de uma classe ou evento, como a classificação binária (neste caso, se uma reserva será cancelada ou não).

8. Resultados dos Modelos





Com base nos resultados, **o modelo Random Forest é o melhor a ser escolhido**. Ele apresentou os melhores desempenhos em todas as métricas avaliadas (F1 Score, AUC e Accuracy), sugerindo que é o mais eficaz para a tarefa de previsão de cancelamento de reservas no hotel. Além disso, devido à sua natureza de ensemble, o Random Forest tende a ser mais robusto e menos suscetível ao overfitting, especialmente em conjuntos de dados complexos.

8.1 Uso de Amostragem para Avaliar os Modelos

A avaliação de modelos de aprendizado de máquina é uma etapa crítica para garantir que eles apresentem um bom desempenho em dados não vistos. Entre as técnicas de validação mais utilizadas, destacam-se o k-fold cross-validation, leave-one-out e bootstrap. Cada uma dessas abordagens possui características específicas que influenciam sua aplicação e eficácia.

8.2 K-Fold Cross-Validation

O método de k-fold cross-validation envolve a divisão do conjunto de dados em **k** subconjuntos, também conhecidos como "folds". O modelo é treinado utilizando **k-1** folds, enquanto o fold restante é utilizado para teste. Este processo é repetido **k** vezes, de forma que cada fold é empregado como conjunto de teste uma vez. Essa técnica proporciona uma estimativa robusta do desempenho do modelo, minimizando a variabilidade associada a uma única divisão de dados.

8.3 Leave-One-Out Cross-Validation (LOOCV)

O leave-one-out é uma forma extrema de k-fold cross-validation, onde o número de folds é igual ao número total de observações no conjunto de dados. Para cada iteração, o modelo é treinado com todos os dados, exceto um único ponto, que é usado como conjunto de teste. Embora essa abordagem possa fornecer avaliações detalhadas, ela pode ser computacionalmente intensiva, especialmente em conjuntos de dados grandes.

8.4 Bootstrap

A técnica de bootstrap envolve a amostragem com reposição, permitindo a criação de múltiplas amostras a partir do conjunto de dados original. O modelo é treinado em cada uma dessas amostras, e as previsões são combinadas para avaliar o desempenho geral do modelo. O bootstrap é particularmente eficaz para quantificar a incerteza e a variabilidade, oferecendo uma perspectiva abrangente sobre a estabilidade do modelo.

8.5 Teste de Hipótese

Para validar estatisticamente as diferenças observadas entre os modelos, foi realizado um teste t para comparar os F1 Scores do Random Forest e da Árvore de Decisão. Os resultados do teste revelaram:

- **T-statistic**: 8.93
- **P-value**: 4.95×10^{-8}

Um p-value tão baixo, significativamente inferior ao limiar convencional de 0.05, indica que a diferença nos F1 Scores entre os modelos é estatisticamente significativa. Portanto, podemos afirmar que o Random Forest não apenas se destaca em desempenho, mas também que essa superioridade é respaldada por evidências estatísticas, consolidando sua escolha como a melhor opção para a tarefa em questão.

9. Conclusão

A análise realizada demonstra que o uso de k-fold cross-validation se mostrou uma abordagem eficaz para a avaliação dos modelos, proporcionando uma estimativa robusta de seu desempenho. As evidências apresentadas sustentam a escolha do modelo Random Forest como o mais apropriado para prever cancelamentos, dadas suas métricas superiores e a validação estatística positiva obtida.

10. Referência bibliográfica

Evidently AI. (n.d.). **Accuracy, Precision, Recall**. Disponível em: https://www.evidentlyai.com/classification-metrics/accuracy-precision-recall

Encord. (n.d.). **F1 Score Definition**. Disponível em: https://encord.com/glossary/f1-score-definition/

Evidently AI. (n.d.). **Explain ROC Curve**. Disponível em: https://www.evidentlyai.com/classification-metrics/explain-roc-curve

IBM. Machine Learning. Disponível em: https://www.ibm.com/br-pt/topics/machine-learning.

RASCHKA, S.; MIRJALILI, V. *Python Machine Learning: Machine Learning and Deep Learning with Python, Scikit-Learn, and TensorFlow 2.* Packt Publishing, 2019.

CHAVES, M. S.; GOMES, R. M. "Predicting Hotel Booking Cancellations Using Machine Learning Techniques." *Tourism & Management Studies*, v. 17, n. 2, p. 45-53, 2021.