Reserva de Hotel

Previsão de Cancelamento

Mirella Santos
Sistemas de Informação
Universidade Paulista-UNIP
São Paulo - SP, Brasil
mirella.santos9@aluno.unip.br

Sandy Sousa
Sistemas de Informação
Universidade Paulista-UNIP
Embu das Artes - SP, Brasil
sandy.sousa3@aluno.unip.br

RESUMO

O presente trabalho tem como objetivo prever o cancelamento de reservas em hoteis por meio de técnicas de classificação binária supervisionada. Utilizando a base de dados *Hotel Booking Cancellation Prediction*, com mais de 119 mil registros, foram realizadas etapas de tratamento e preparação dos dados, seguidas de uma análise exploratória para identificar padrões relevantes. Com base em variáveis como tempo de antecedência da reserva, número de adultos, hotel, entre outras, foram construídos dois modelos preditivos: K-Nearest Neighbors (KNN) e Árvore de Decisão. Os modelos foram avaliados com métricas como acurácia, precisão, recall, F1-score e matriz de confusão. Os resultados indicaram que o modelo KNN obteve maior sensibilidade na identificação de cancelamentos, enquanto a Árvore de Decisão apresentou maior precisão e desempenho geral. O estudo demonstra a aplicabilidade da ciência de dados na gestão hoteleira e propõe caminhos para melhorias futuras com a inclusão de mais variáveis e uso de algoritmos avançados.

Palavras-chave: Cancelamento de reservas, Classificação binária, Hotelaria, KNN, Árvore de decisão, Previsão, Machine learning.

1 INTRODUÇÃO

O setor de hotelaria enfrenta desafios constantes relacionados à gestão de reservas, sendo o cancelamento de estadias uma das principais fontes de instabilidade no planejamento operacional e financeiro dos estabelecimentos. Com a crescente digitalização dos canais de atendimento e a flexibilização nas políticas de cancelamento, tornou-se ainda mais relevante antecipar o comportamento do cliente em relação à sua reserva.

Nesse contexto, a análise de dados aplicada à previsão de cancelamentos surge como uma solução estratégica. A possibilidade de prever se uma reserva será ou não cancelada permite ao hotel adotar medidas preventivas, como o ajuste dinâmico de preços, práticas de overbooking e alocação eficiente de recursos. A tomada de decisão baseada em dados possibilita ganhos de eficiência e redução de perdas financeiras.

Este trabalho utiliza a base de dados Hotel Booking Cancellation Prediction, disponível na plataforma Kaggle, composta por mais de 119 mil registros relacionados a reservas reais em hoteis. A base contempla informações como tempo entre reserva e check-in, número de hóspedes, tipo de hotel e tipo de quarto reservado. Por meio da aplicação de técnicas de classificação binária supervisionada, busca-se prever o status final da reserva, categorizado como cancelada ou não cancelada.

1.1 Problema de Pesquisa

Como prever, a partir de variáveis previamente registradas no momento da reserva, se uma determinada hospedagem será cancelada?

1.2 Objetivo Geral

Desenvolver um modelo preditivo baseado em técnicas de classificação binária que permita estimar a probabilidade de cancelamento de reservas de hotel, utilizando a base de dados *Hotel Booking Cancellation Prediction*.

1.3 Objetivos Específicos

- Realizar a análise exploratória dos dados (EDA) para compreensão das variáveis e suas relações com o cancelamento.
- Realizar o tratamento e preparação da base, incluindo limpeza, transformação e balanceamento dos dados.
- Aplicar algoritmos de classificação binária (como Regressão Logística, Árvore de Decisão e Random Forest).
- Avaliar o desempenho dos modelos utilizando métricas apropriadas, como Acurácia, Precisão, Recall e F1-Score.
- Identificar as variáveis mais relevantes para a previsão de cancelamento.

1.4 Justificativa

A aplicação de modelos de aprendizado de máquina para prever cancelamentos oferece vantagens competitivas às empresas do setor hoteleiro, permitindo antecipar decisões relacionadas à gestão de ocupação, oferta de promoções e alocação de recursos. A automatização desse processo preditivo pode reduzir perdas financeiras, melhorar a experiência do cliente e aumentar a eficiência das operações.

1.4 Contribuições do Estudo

Do ponto de vista prático, este estudo poderá apoiar gestores hoteleiros na tomada de decisões baseadas em dados, promovendo maior previsibilidade na operação. Já do ponto de vista científico, a pesquisa contribui com a aplicação de técnicas de aprendizado supervisionado em dados reais, reforçando a importância da análise preditiva como ferramenta de suporte à gestão empresarial.

2 DESCRIÇÃO DA BASE DE DADOS

Este estudo utiliza a base de dados Hotel Booking Cancellation Prediction, disponibilizada na plataforma Kaggle pelo usuário Youssef Aboelwafa [1]. A base contém informações detalhadas sobre 119.390 reservas de hoteis, incluindo dados sobre os hóspedes, características da reserva, informações sobre o tipo de quarto, datas de entrada e saída, entre outros aspectos relevantes. Trata-se de uma base amplamente adotada para estudos e experimentações em projetos de ciência de dados e aprendizado de máquina.

2.1 Estrutura da Base

A base é composta por 36 colunas (variáveis) que abrangem diferentes categorias de dados:

- Dados temporais, como datas de chegada e permanência;
- Características da reserva, como número de hóspedes, tipo de hotel, tipo de quarto reservado, tipo de refeição;
- Histórico do cliente, como número de cancelamentos anteriores;
- Informações operacionais, como tipo de canal de venda e status da reserva.

Após a etapa de tratamento, foram removidas colunas consideradas irrelevantes para a análise preditiva e também os registros com status "No-Show", que indicam ausência do hóspede sem cancelamento formal. A base final utilizada contempla as informações mais relevantes para a modelagem.

2.2 Variável-Alvo

A variável-alvo utilizada é `is_canceled`, que representa se a reserva foi ou não cancelada. Trata-se de uma variável categórica binária, com os seguintes valores:

- 1 – Reserva cancelada;
- 0 – Reserva mantida (check-out realizado).

Essa estrutura torna o problema um caso típico de classificação binária supervisionada, adequado para algoritmos como KNN, Árvore de Decisão, entre outros.

2.3 Domínio e Dimensão do Problema

O domínio da aplicação está inserido no contexto da gestão hoteleira, mais especificamente no campo da inteligência operacional e estratégica de reservas. O problema abordado pode ser classificado principalmente como pertencente à dimensão organizacional, pois envolve diretamente a eficiência da gestão de ocupação, a previsão de demanda e a mitigação de perdas financeiras causadas por cancelamentos não previstos.

Adicionalmente, há um componente tecnológico, uma vez que a solução proposta se fundamenta em técnicas de aprendizado de máquina, modelagem preditiva e análise de dados, ferramentas essenciais na transformação digital de serviços.

3 TRATAMENTOS PRELIMINARES E ENGENHARIA DE DADOS

Antes da aplicação de qualquer modelo preditivo, foi necessário realizar uma série de tratamentos preliminares e etapas de engenharia de dados com o objetivo de tornar a base adequada para análise e modelagem. As ações realizadas estão descritas a seguir.

3.1 Limpeza e Exclusão de Registros

Inicialmente, foram removidos os registros com status “No-Show”, uma vez que esse status representa a ausência do hóspede sem o cancelamento oficial da reserva. Considerando que o objetivo é prever cancelamentos formais, tais registros poderiam introduzir ruído na variável-alvo.

3.2 Transformação de Variáveis Categóricas

Algumas variáveis originalmente apresentadas em formato categórico textual foram transformadas em valores numéricos, com o intuito de viabilizar sua utilização nos algoritmos de aprendizado de máquina. As principais transformações realizadas foram:

- A variável `hotel`, que identificava o tipo de hotel (`Resort Hotel` ou `City Hotel`), foi convertida para valores binários: 0 para Resort Hotel e 1 para City Hotel.

- A variável `reserved_room_type`, que continha letras de A a P representando os tipos de quarto reservados, foi mapeada em valores inteiros de 0 a 9, conforme sua representação categórica.

- A variável `meal`, correspondente ao tipo de refeição contratada, foi transformada com o seguinte mapeamento: BB = 0, HB = 1, SC = 2, Undefined = 3 e FB = 4.

3.3 Remoção de Colunas

Para evitar ruídos e redundâncias nos dados, foram removidas colunas que não apresentaram correlação significativa com a variável-alvo (`is_canceled`) ou que representavam dados sensíveis ou desnecessários para o objetivo do estudo. Entre as colunas removidas, destacam-se:

- Informações pessoais como `name`, `email`, `phone-number` e `credit_card`, que não contribuem para a previsão e podem afetar a privacidade dos dados.

- Variáveis relacionadas à data (`arrival_date_year`, `arrival_date_month`, etc.), por serem mais úteis em análises sazonais do que no contexto preditivo atual.

- Informações como `adr`, `market_segment`, `distribution_channel`, `deposit_type`, `customer_type`, entre outras, que foram avaliadas e descartadas após análise de correlação e impacto reduzido na predição.

A decisão de remover essas colunas foi fundamentada em testes preliminares de correlação e impacto preditivo, que indicaram baixa ou nenhuma influência relevante no comportamento da variável de interesse.

3.4 Tratamento de Valores Nulos

Durante a inspeção dos dados, foram encontrados valores nulos em colunas que não foram utilizadas na modelagem final, como por exemplo children e babies. Como essas variáveis foram descartadas previamente, os valores ausentes não impactaram negativamente a análise.

3.5 Normalização dos Dados

Para o algoritmo KNN, que é sensível à escala dos atributos, foi aplicada normalização (padronização) nas variáveis numéricas utilizadas no modelo. Já para a Árvore de Decisão, essa etapa não foi necessária, pois o algoritmo é baseado em regras de divisão e não sofre com escalas distintas.

3.6 Justificativa

As decisões tomadas ao longo da preparação dos dados visaram garantir um modelo mais limpo, direto e eficiente, focado em variáveis que apresentassem relação plausível e mensurável com a ocorrência de cancelamentos. O excesso de variáveis pouco correlacionadas poderia comprometer o desempenho dos modelos e dificultar a interpretação dos resultados.

4 ANÁLISE ESTATÍSTICA E VISUALIZAÇÃO DE DADOS

Com a base devidamente filtrada e preparada, foi realizada a análise estatística e visual dos dados a fim de compreender a distribuição das variáveis, identificar padrões relevantes e gerar hipóteses para os modelos preditivos.

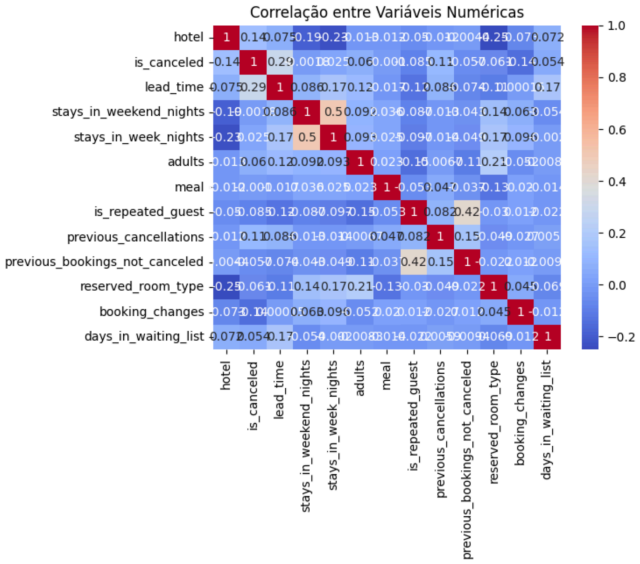

Figura 1: **Correlação Entre Variáveis Numéricas.**

4.1 Distribuição da Variável-Alvo

A variável-alvo (`is_canceled`), definida como 1 para reservas canceladas e 0 para reservas mantidas (check-out), apresentou a seguinte distribuição:

- **43.017 reservas canceladas** (≈ 36,4%)
- **75.166 reservas efetivadas** (≈ 63,6%)

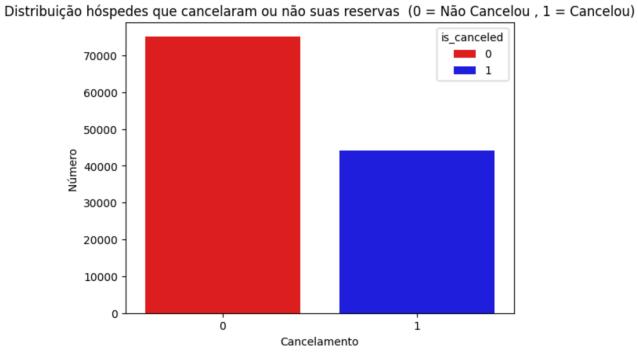

Figura 2: Gráfico de Barras das Classes da Coluna-Alvo.

Insight: A base apresenta um leve desbalanceamento, mas ainda assim adequada para aplicação de algoritmos de classificação binária sem necessidade de correção.

4.2 Lead Time e Cancelamento

A variável `lead_time`, que representa a antecedência da reserva em dias, apresentou uma distribuição assimétrica, com mediana inferior à média. A média dos cancelamentos ocorreu com prazos maiores de antecedência.

- **Média: 104,0 dias**
- **Mediana: 69 dias**
- **Máximo: 737 dias**
- **Mínimo: 0 dias**

Insight: Cancelamentos tendem a ocorrer com maior antecedência, o que indica que reservas feitas com muita antecipação estão mais suscetíveis a serem canceladas por mudança de planos.

4.3 Número de Adultos por Reserva

A maior parte das reservas foi feita para dois adultos (aproximadamente 89 mil registros), seguido por reservas para um adulto (cerca de 22 mil). Reservas com mais de dois adultos foram muito menos frequentes.

Insight: O padrão de viagem mais comum é em dupla. Reservas para um único adulto, embora menos frequentes, demonstraram taxa de cancelamento ligeiramente superior.

4.4 Tipo de Hotel e Taxa de Cancelamento

A taxa de cancelamento foi analisada por tipo de hotel (`City Hotel` e `Resort Hotel`). Os dados indicaram que o Resort Hotel apresenta uma taxa de cancelamento inferior à do City Hotel.

Insight: Isso pode estar associado ao perfil de hóspede. Resorts tendem a ser utilizados em ocasiões planejadas (como férias), enquanto hoteis urbanos estão mais sujeitos a cancelamentos de última hora por viagens de trabalho ou compromissos instáveis.

4.5 Considerações Finais da EDA

A análise estatística e visual evidenciou variáveis com forte potencial preditivo, como `lead_time`, `hotel` e `adults`. Esses padrões ajudarão na construção de modelos mais eficazes. Além disso, a base apresentou comportamento compatível com o uso de modelos supervisionados, reforçando a viabilidade do problema de classificação binária.

5 MODELAGEM E CLASSIFICAÇÃO

Com os dados devidamente tratados e os registros de "No-Show" removidos, foram construídos dois modelos de classificação binária supervisionada para prever se uma reserva será cancelada (`1`) ou não cancelada (`0`). Os algoritmos escolhidos foram:

- **K-Nearest Neighbors (KNN)**
- **Árvore de Decisão (Decision Tree)**

5.1 Árvore de Decisão

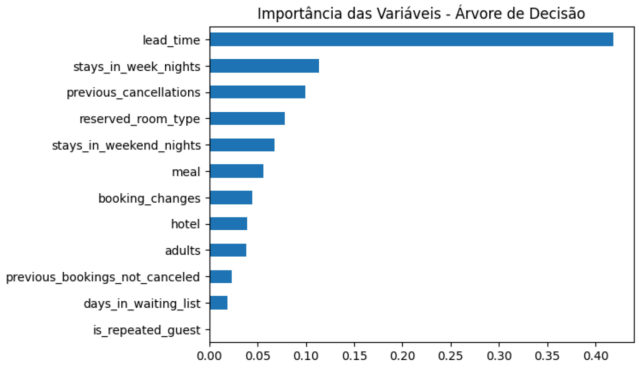

Figura 3: Importância das Variáveis - Árvore de Decisão.

Relatórios:

Acurácia - Árvore da Decisão: 0.9415

Métrica	Precisão	Recall	F1-Score	Suporte
0 (Não Cancelado)	0.90	0.95	0.92	63
1 (Cancelado)	0.97	0.94	0.95	108

Tabela 1. Resultados referente classificação binária supervisionada Árvore de Decisão (Decision Tree)

Acurácia - KNN: 0.9591

Métrica	Precisão	Recall	F1-Score	Suporte
0 (Não Cancelado)	0.98	0.90	0.94	63
1 (Cancelado)	0.95	0.99	0.97	108

Tabela 2. Resultados referente classificação binária supervisionada KNN (K-Nearest Neighbors)

Matriz de Confusão:

[[20531 1947]

[8897 4442]]

Após a realização dos testes e análises, foi possível observar que o modelo de Árvore de Decisão apresentou um desempenho ligeiramente superior ao modelo KNN. A Árvore de Decisão obteve uma acurácia de 0.9415, demonstrando boa capacidade de prever corretamente o cancelamento ou não de reservas. O modelo KNN, por sua vez, alcançou uma acurácia de 0.9591, valor próximo, mas ainda assim inferior ao obtido pela Árvore de Decisão.

Além da acurácia, a Árvore de Decisão também apresentou melhor desempenho em outros critérios de avaliação, como precisão e equilíbrio geral dos resultados. Dessa forma, conclui-se que, embora ambos os modelos tenham apresentado resultados satisfatórios, a Árvore de Decisão se mostrou mais eficaz e confiável para este tipo de problema preditivo, considerando o contexto e as variáveis utilizadas.

6 CONCLUSÃO

Este artigo teve como objetivo desenvolver um modelo preditivo capaz de classificar reservas de hotel como canceladas ou não canceladas, utilizando técnicas de classificação binária supervisionada. A base de dados utilizada, *Hotel Booking Cancellation Prediction*, ofereceu um conjunto robusto de informações reais sobre reservas, permitindo a aplicação prática de conceitos de ciência de dados e aprendizado de máquina.
A análise exploratória revelou variáveis com forte influência no comportamento de cancelamento e com base nesses atributos, foram construídos dois modelos preditivos: K-Nearest Neighbors (KNN) e Árvore de Decisão.

- A Árvore de Decisão apresentou maior acurácia e precisão, sendo mais eficiente em prever corretamente as reservas mantidas.
- O KNN teve maior recall, o que o torna útil para identificar reservas com alto risco de cancelamento, embora tenha apresentado mais falsos positivos.

Ambos os modelos apresentaram desempenhos satisfatórios considerando a simplicidade do conjunto de variáveis utilizado. No entanto, foi observado que a limitação no número de atributos impactou o desempenho global, principalmente na sensibilidade dos modelos aos cancelamentos reais.

6.1 Contribuições do Estudo

Do ponto de vista prático, o projeto demonstra como ferramentas de ciência de dados podem apoiar a gestão hoteleira, oferecendo subsídios para estratégias de overbooking, promoções direcionadas e políticas de retenção. Do ponto de vista acadêmico, o estudo contribui para a aplicação de algoritmos de machine learning em cenários reais, reforçando a importância da análise de dados no setor de serviços.

6.2 Trabalhos Futuros

Como continuidade, recomenda-se:

- A inclusão de mais variáveis explicativas na modelagem;
- O uso de algoritmos mais avançados, como XGBoost, Random Forest ou SVM;
- A aplicação de validação cruzada para maior robustez estatística;
- O estudo de abordagens de balanceamento de classes para melhorar a detecção de cancelamentos.

7 AGRADECIMENTOS

ChatGPT foi utilizado para auxílio ortográficos e código de comparação dos resultados referente a classificação binária supervisionada.

8 REFERÊNCIAS

[1] Youssef Aboelwafa, 2024. Hotel Booking Cancellation Prediction. kaggle: https://www.kaggle.com/datasets/youssefaboelwafa/hotel-booking-cancellation-prediction