INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Modelos de Previsão

Aluno
Pedro Lourenço, aluno n.º 133931

Mestrado: Ciência de Dados

Unidade curricular: Modelos de Previsão

Professora, Anabela Costa
ISCTE – Instituto Universitário de Lisboa

Janeiro 2026

Índice

Introdução

A crescente digitalização do setor do turismo tem reforçado o papel das plataformas online na formação da perceção de qualidade dos serviços hoteleiros. Em particular, plataformas como o TripAdvisor agregam avaliações de hóspedes e sintetizam essa informação num score médio (escala 1–5), que funciona como indicador estatístico da satisfação global e influencia procura, reputação e posicionamento competitivo. Neste enquadramento, interessa analisar quantitativamente até que ponto características **observáveis** dos hotéis (p. ex., selos, patrocínio e *amenities*) se associam à variação sistemática do score, reforçando que a evidência aqui produzida é preditiva e não causal.

O presente trabalho tem como objetivo desenvolver um modelo preditivo para estimar a variável-alvo *score_review* de hotéis no Dubai, a partir do dataset *Dubai_data*. As variáveis explicativas incluem indicadores binários (p. ex., *CarimboTripAdvisor*, *Patrocinado*, *Breakfast*, *WiFi_gratuito*, *Piscina*, *Restaurante*, *Servico_quartos*, *Praia*, *Bar_lounge*, *Tomar_medidas_seguranca*, *Visitar_website_hotel*) e a variável numérica *logavaliacoes* (logaritmo natural do nº de avaliações). Do ponto de vista estatístico, o problema é tratado como **regressão supervisionada**, visando *score_review*. A coerência básica do target é verificada antes da modelação (range observado [2.5,5], sem valores fora de [1,5]), o que entrga consistência com a escala interpretativa do fenómeno.

O desenvolvimento do estudo segue a metodologia **CRISP-DM**, estruturando o processo desde a formulação do problema e inventário de variáveis, até à preparação estatística dos dados (tratamento de omissos, controlo de extremos e binarização), análise exploratória e fase de modelação. A comparação entre modelos é feita por **validação cruzada k-fold apenas no treino**, reservando um conjunto de teste (*holdout*) para avaliação final única do erro de generalização. Todo o pré-processamento que aprende parâmetros é ajustado no treino e aplicado de forma idêntica em validação/teste, reduzindo risco de *data leakage*; adicionalmente, o pipeline guarda outputs por capítulo, assegurando rastreabilidade e auditoria.

O desempenho preditivo é quantificada através de métricas padrão de regressão (RMSE, MAE e (R^2)), complementadas por uma métrica de interpretação direta: percentagem de previsões com erro absoluto ≤ 0,5 pontos na escala do score.

Capítulo 1 - Formulação do Problema

1.1 *Business Understanding*

As plataformas digitais de turismo (e.g., TripAdvisor) agregam milhares de avaliações individuais num *score* médio (escala 1–5) que influencia diretamente a reputação e a procura de um hotel. Do ponto de vista de Ciência de Dados, este *score* pode ser tratado como uma variável quantitativa de satisfação agregada, cujo valor esperado pode (parcialmente) ser explicado por características observáveis do hotel (selos, sinais de marketing, *amenities* e proxies de popularidade).

Embora o *score_review* esteja limitado ao intervalo [1,5] e apresente discretização, assume-se uma abordagem de regressão para modelar a **média condicional** $\mathbb{E}(Y|X)$ e produzir previsões em novos hotéis, tratando o score como uma variável numérica *quasi*-contínua (com incrementos decimais). A interpretação é **associativa/preditiva**, não causal.

Objetivo do trabalho: Pretende-se construir um modelo preditivo que estime ***score_review*** para hotéis no Dubai, a partir das variáveis disponibilizadas no *dataset Dubai_data*, conforme o enunciado.

Enquadramento do problema: Como *score_review* assume valores numéricos na escala [1,5] (incluindo incrementos decimais e não só inteiros), o problema é formulado como **regressão supervisionada**, visando estimar o valor esperado do *score* em observações futuras (hotéis "não vistos").

Questões analíticas orientadoras: O modelo e a análise subsequente procuram responder, em termos preditivos e interpretativos, a questões como:

- Que características/*amenities* (*flags*/indicadores) estão mais associadas a *scores* mais elevados?
- É possível melhorar de forma relevante uma baseline simples (prever a média) usando modelos mais flexíveis (árvores, Random Forest, GBM)?
- Que variáveis parecem ter maior impacto e em que direção (associação positiva/negativa), mantendo a ressalva de que não se trata de inferência causal?

Para garantir que a modelação assenta em evidência objetiva (e não em pressupostos), o pipeline regista métricas de qualidade essenciais: dimensões do dataset, duplicados, *missingness* e coerência do target na escala do enunciado.

1.2 Definição formal do problema

Unidade de análise: Cada observação corresponde a um hotel, descrito por atributos sobretudo binários (*amenities/flags*) e uma variável numérica de popularidade (*logavaliacoes*).

Variável resposta (target): $Y_i \in [1,5]$, onde Y_i é o *score* médio de avaliações do hotel i.

Variáveis explicativas: para cada hotel i, $\mathrm{x}_i = (x_{i1}, \dots, x_{ip})^\top$, onde $x_{ij} \in \{0,1\}$ para os preditores binários (*amenities/flags*) e uma covariável contínua de popularidade dada por $x_{ik} =$ log (nº de avaliações$_i$), isto é, o logaritmo natural do número de *reviews* do hotel i.

Objetivo estatístico: O objetivo é estimar a média condicional $m(\mathrm{x}) = \mathbb{E}(Y \mid \mathrm{X} = \mathrm{x})$, isto é, o **score médio esperado** para hotéis com características x. Como $m(x)$é desconhecida, ajusta-se um modelo aos dados para obter $\hat{m}(x)$. Para um novo hotel com covariáveis x_{new}, a previsão é: $\hat{Y}_{\text{new}} = \hat{m}(\mathrm{x}_{\text{new}}) \approx \mathbb{E}(Y_{\text{new}} \mid \mathrm{X}_{\text{new}} = \mathrm{x}_{\text{new}})$.

1.3 Variáveis disponíveis e hipóteses iniciais

O conjunto de preditores é o definido no enunciado: *CarimboTripAdvisor*, *Patrocinado*, *Breakfast*, *WiFi_gratuito*, Estacionamento_gratuito, *Piscina*, *Restaurante*, *Servico_quartos*, *Praia*, *Bar_lounge*, *Tomar_medidas_seguranca*, *Visitar_website_hotel* e *logavaliacoes*.

Racional estatístico: A predominância de variáveis binárias (0/1) favorece duas leituras/modelos complementares:

- **Modelos lineares:** interpretam o coeficiente como variação média esperada no *score* (condicional às restantes variáveis), úteis para explicação e comparação.
- **Modelos não lineares:** capturam interações e não linearidades (por exemplo, o efeito do selo pode variar com *logavaliacoes*), podendo melhorar desempenho via *trade-off* viés–variância.

Coerência e escala de *logavaliacoes*: A variável apresenta valores compatíveis com uma transformação logarítmica do nº de avaliações (inclui 0, quando o nº de reviews é 1, dado (\log(1)=0)).

1.4 Métricas

O desempenho do modelo deve ser avaliado como **erro de generalização** no conjunto de teste (*holdout*), após seleção/tuning por validação cruzada no treino, sendo usado uma única vez, apenas para estimar desempenho fora da amostra após a seleção

Métricas principais:

- **RMSE**: penaliza mais erros grandes; adequado quando desvios elevados na escala 1–5 são particularmente indesejáveis.
- **MAE (*Mean Absolute Error*)**: interpretação direta em "pontos de score" (erro absoluto médio).
- **R^2**: proporção de variância explicada usado medida global de ajuste (em teste pode ser negativo se o modelo for pior do que prever a média).
- **PCT_0,5**: percentagem de previsões com (|erro|\le 0,5), que é uma tolerância natural dada a discretização do alvo.

O modelo final será o que apresenta **menor RMSE médio em validação cruzada** no treino, usando as restantes métricas como apoio interpretativo (Ver **Anexo C5-1**).

---

2. Processo de ETL e Preparação de Dados

2.1 Leitura do dataset e organização do pipeline

O dataset **Dubai_data.csv** é carregado de forma robusta, os dados são armazenados como **df_raw**. A execução do trabalho é **orquestrada por um pipeline reprodutível**, estruturado por capítulos (CRISP-DM), onde cada etapa escreve **artefactos objetivos** numa pasta dedicada.

2.2 Limpeza base e controlo de qualidade

Antes de qualquer modelação, é aplicada uma limpeza mínima com **registo explícito de métricas QC** (*Quality Control*). Esta etapa é estatisticamente necessária porque a existência de **valores omissos na variável resposta inviabilizam a aprendizagem supervisionada**, **duplicados exatos** aumentam artificialmente o "peso" de algumas observações e valores fora do intervalo do target violariam as regras do problema.

Verificação	Contagem	Percentagem	Ação
n bruto (n_raw)	546	100,0%	—
Duplicados exatos	17	3,1%	Removidos
Missing no target (*score_review*)	0	0,0%	—
Valores fora de [1,5] no target	0	0,0%	—
n final (após deduplicação)	529	96,9%	Usado para split

Isto significa que o dataset do ponto de vista do *target* não apresenta *missings,* mas tem **duplicação estrutural** suficiente para justificar deduplicação. A deduplicação é feita antes do particionamento para evitar que observações idênticas apareçam simultaneamente em treino e teste, o que inflacionaria artificialmente o desempenho (leakage por replicação)

2.3 Binarização padronizada das variáveis explicativas

As variáveis explicativas (exceto ***logavaliacoes***) são convertidas para codificação **0/1**. Esta padronização é importante porque vários modelos assumem entradas numéricas, tornando a interpretação posterior mais direta quando a semântica é uniforme: **1 = atributo presente / 0 = ausente**.

2.4 Divisão estratificada em treino/validação/teste

O dataset limpo (n=529) é dividido de forma **estratificada pelo target** em três subconjuntos com seed = 20260119 com o objetivo de poder replicar os resultados e a consistência do processo aleatório e divisão 70%/10%/20% (treino/validação/teste), com estratificação por níveis de ***score_review***.

Usa-se a estratificação pois reduz a variância das métricas, garantindo representação de níveis raros do alvo em cada subconjunto

---

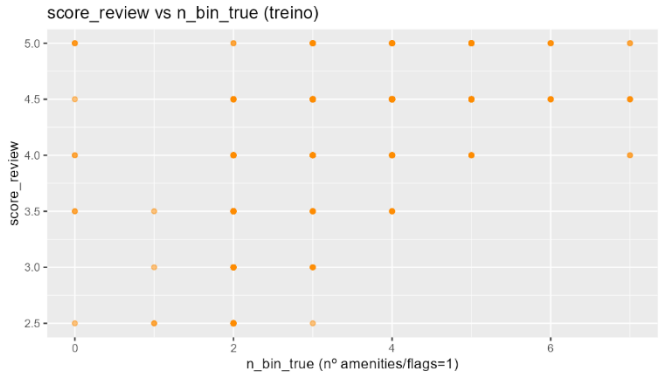
2.5 Pré-processamento com princípio anti-*leakage*

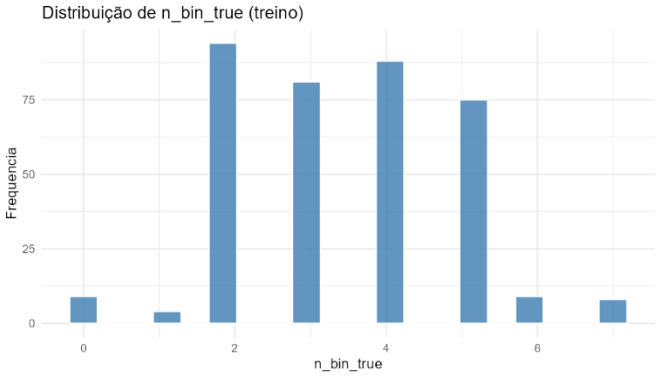
Tudo o que envolve **"aprender parâmetros"** a partir dos dados é ajustado **exclusivamente no treino** (**preprocess_fit**) e depois aplicado **sem recalcular** em validação e teste (**preprocess_apply**). Este é um princípio central para uma estimativa menos enviesada: se parâmetros como médias, desvios-padrão, mediana ou limites de winsorização fossem calculados no dataset completo, estaríamos a introduzir informação do teste no treino (data *leakage*), subestimando artificialmente o erro.

2.6 *Feature engineering*

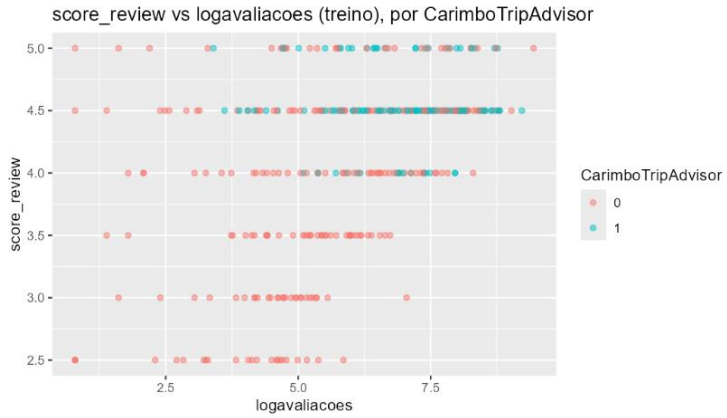
Para aumentar capacidade preditiva foram construídas duas *features* derivadas:

n_bin_true: soma das *flags* binárias ativas por hotel (após remoção das raras). Interpretação estatística; atua como um índice de 'riqueza de amenities', permitindo capturar um efeito agregado e reduzir dependência de coeficientes individuais muito instáveis em amostras pequenas.





Interação *CarimboTripAdvisor* × *logavaliacoes*: A interação permite modelar não-aditividade: o efeito marginal de *CarimboTripAdvisor* pode depender do nível de *logavaliacoes*.



Capítulo 3 - *Data Understanding*

3.1 Objetivos e estratégia de análise exploratória

A etapa de *Data Understanding* tem como objetivo caracterizar estatisticamente o dataset ***Dubai_data*** antes de qualquer transformação, avaliando: (i) **estrutura e tipologia das variáveis**, (ii) **qualidade dos dados** (duplicados e omissos), (iii) **coerência do target** na escala definida, (iv) **distribuições marginais**

(target e variável contínua), (v) **prevalência/raridade** das variáveis binárias, e (vi) **associações bivariadas e correlações** com a variável-alvo. Embora o alvo seja limitado e discretizado, adota-se uma abordagem de regressão para modelar a média condicional $\mathbb{E}(Y|X)$, com foco estritamente preditivo.

As estatísticas seguintes são reportadas após limpeza base e deduplicação (n=529), salvo indicação em contrário.

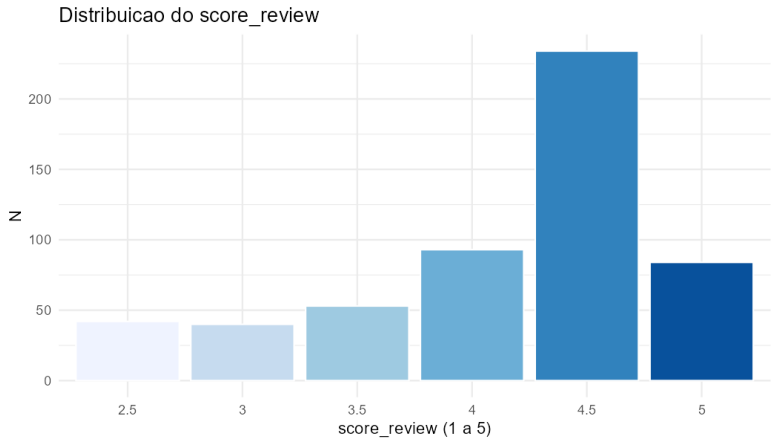
3.2 Duplicados exatos e implicações estatísticas

Foram detetadas **17 linhas duplicadas exatas**. Em termos estatísticos, a presença de duplicados viola a suposição prática de "observações independentes" e pode enviesar medidas descritivas e ajuste de modelos (sobretudo em amostras moderadas), ao **representar padrões repetidos**. Assim, na preparação de dados, justifica-se a **remoção de duplicados** antes da divisão treino/validação/teste.

3.3 Estatística descritiva do target e concentração em valores elevados

Mínimo	1.º quartil	Mediana	Média	3.º quartil	Máximo	Desvio-padrão
2,5	4,0	4,5	4,131	4,5	5,0	≈ 0,718

Isto indica concentração junto ao limite superior e antecipa um fenómeno típico em regressão com variável-alvo limitada: tendência a previsões "centradas" (redução de extremos), sobretudo em modelos que minimizam erro quadrático, pelo que a distribuição por frequências confirma o desbalanceamento moderado: 4,5 é o nível mais frequente: **234/546 = 42,9%**

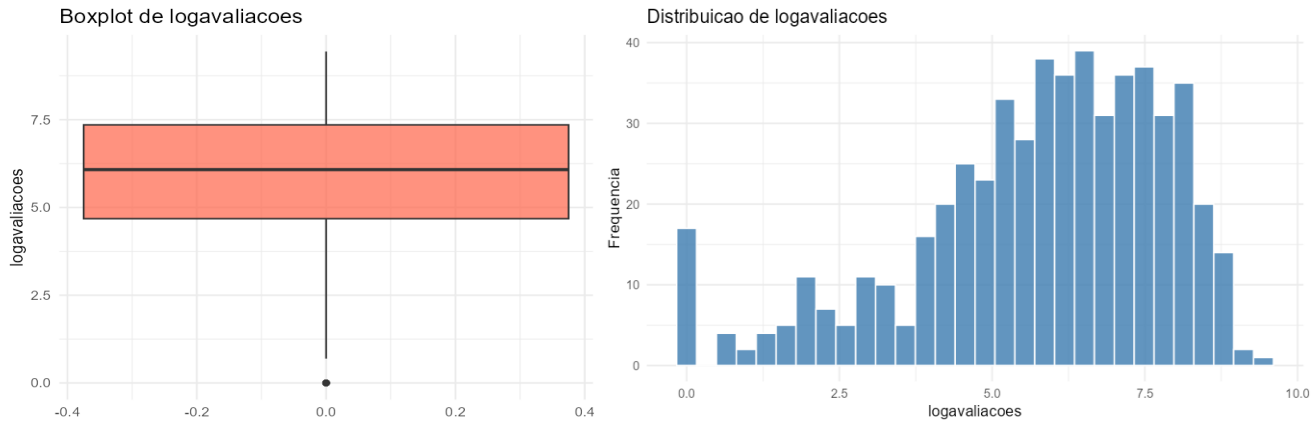


3.4 Distribuição e escala de *logavaliacoes* (popularidade/volume de reviews)

A variável *logavaliacoes* (logaritmo natural do número de reviews) apresenta amplitude elevada e assimetria positiva (cauda à direita), com:

Métrica	Valor	Reviews (≈ exp (Métrica)
Mínimo	0,000	1
1.º quartil	4,680	108
Mediana	6,078	436
3.º quartil	7,352	1559
Máximo	9,442	12 607
Média	5,753	
DP	2,089	

Do ponto de vista estatístico, esta dispersão sugere que a variável pode introduzir ***leverage*** em modelos lineares (valores extremos influenciam o ajuste). A inspeção por *boxplot* e o critério IQR indicam existência de observações na cauda inferior (hotéis com poucas reviews), o que justifica posteriormente uma estratégia robusta como **winsorização aprendida no treino**.



Hotéis com poucas reviews tendem a ter score médio mais instável (maior variância amostral), pelo que *logavaliacoes* pode capturar simultaneamente popularidade e estabilidade do score.

3.5 Variáveis binárias: prevalência e raridade (NZV)

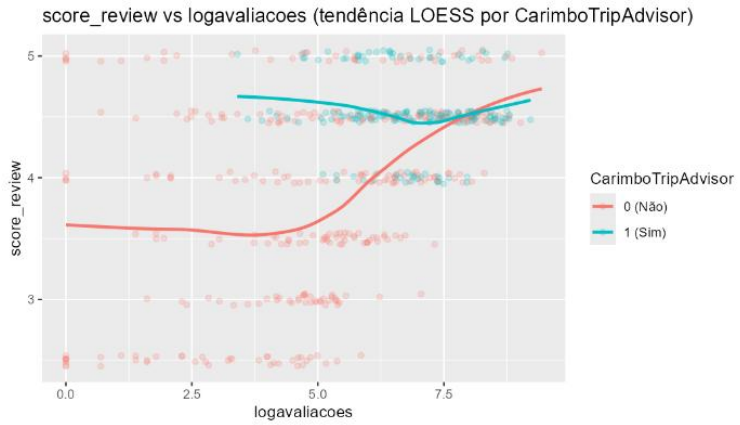
A análise de proporções das 12 variáveis binárias mostra heterogeneidade relevante: existem variáveis muito comuns (pouco discriminativas) e variáveis muito raras (quase constantes) (Ver **Anexo C3-2**).

Do ponto de vista estatístico, variáveis com **variância muito baixa** (quase constantes) contribuem pouco para previsão, podem introduzir ruído e instabilidade e são candidatas naturais a remoção por critérios do

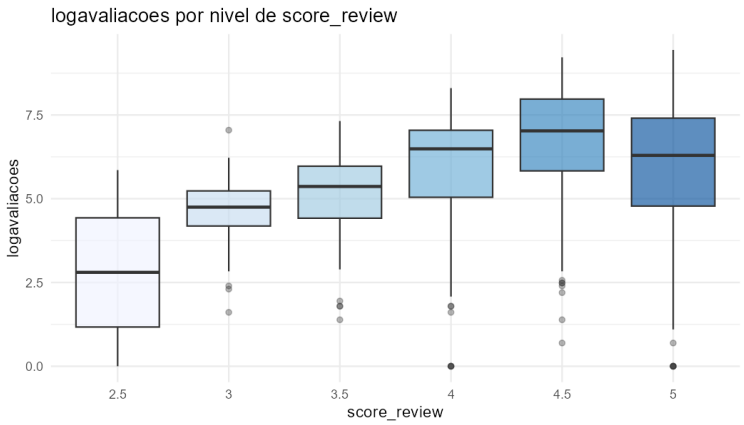
tipo ***near-zero variance*** **(NZV)**. Com limiar de raridade de 2%, foram identificadas as variáveis ***Praia*** (min_prop ≈ 0,004) ***Bar_lounge*** (min_prop ≈ 0,0055) (Desenvolvido no ponto 4).

3.6 Relação bivariada entre *score_review* e *logavaliacoes*

A relação entre o ***score_review*** e *logavaliacoes* é positiva em média, com correlação de *Pearson* moderada (r ≈ 0,450). A leitura substantiva é compatível com a interpretação de *logavaliacoes* como proxy de popularidade/volume de informação: hotéis com mais reviews tendem a apresentar scores médios mais elevados. Esta associação pode ser observada no **diagrama de dispersão com curva de tendência LOESS**, representada abaixo, onde a nuvem de pontos evidencia um aumento médio do ***score_review*** à medida que ***logavaliacoes*** cresce, sendo útil para avaliar a forma global da relação.



Para complementar a leitura do *scatter* usa-se**, imagem** apresentada em seguida, o **boxplot** de ***logavaliacoes*** por nível de ***score_review***. Esta visualização reduz o "ruído" dos pontos individuais e permite comparar, para cada nível do target, a distribuição de ***logavaliacoes***. Em geral, observa-se um deslocamento das medianas para valores mais elevados de ***logavaliacoes*** à medida que o ***score_review*** aumenta, embora exista sobreposição relevante entre níveis, o que sugere que o volume de reviews não é, por si só, suficiente para separar completamente os níveis do target.



3.7 Correlações: força de associação com o target e colinearidade entre preditores

Foi calculada uma **matriz de correlações de Pearson** após codificação numérica. Para variáveis binárias, este coeficiente equivale a uma forma de **correlação ponto-bisserial**, útil como medida exploratória de associação linear, mas sem interpretação causal.

Correlações com *score_review*:

Variável	r
Tomar_medidas_seguranca	≈ 0,620
Estacionamento_gratuito	≈ 0,494
logavaliacoes	≈ 0,450
CarimboTripAdvisor	≈ 0,396
Visitar_website_hotel	≈ 0,362
Restaurante	≈ −0,278
Piscina	≈ −0,178
Servico_quartos	≈ −0,148
Bar_lounge	≈ −0,100

Adicionalmente, observam-se correlações consideráveis entre preditores, sugerindo **colinearidade** (p. ex., *Tomar_medidas_seguranca* com *logavaliacoes* r ≈ 0,620).

Capítulo 4 - Data Preparation

4.1 Objetivo e princípios metodológicos

Nesta fase procede-se à transformação do dataset bruto num conjunto de dados **pronto para modelação**, garantindo simultaneamente: **(i) qualidade e coerência do alvo**, **(ii) reprodutibilidade**, **(iii) prevenção**

de *data leakage* e **(iv) compatibilidade com diferentes famílias de modelos** (lineares/regularizados vs métodos baseados em árvores).

O pipeline segue explicitamente o princípio ***fit* no treino / *apply* em validação e teste** para todos os passos que "aprendem" parâmetros estatísticos (p. ex., limites de winsorização, regras de eliminação de variáveis raras e parâmetros de padronização). Isto é uma condição necessária para que a estimativa do erro de generalização se aproxime do **não enviesamento**.

O pipeline segue o princípio *fit no treino / apply nos restantes* para mitigar data leakage e reduzir viés de otimismo na estimativa do erro de generalização

---

4.2 Limpeza base e controlo de qualidade (QC) antes do *split*

Antes de qualquer particionamento em treino/validação/teste, foi executada uma etapa de **limpeza básica** (*prep_basic*) composta apenas por operações **seguras pré-*split*** (isto é, que não usam informação estatística do conjunto completo e, portanto, não induzem *data leakage*). (Ver **Anexo C4-1**).

4.3 Binarização padronizada das variáveis explicativas (0/1)

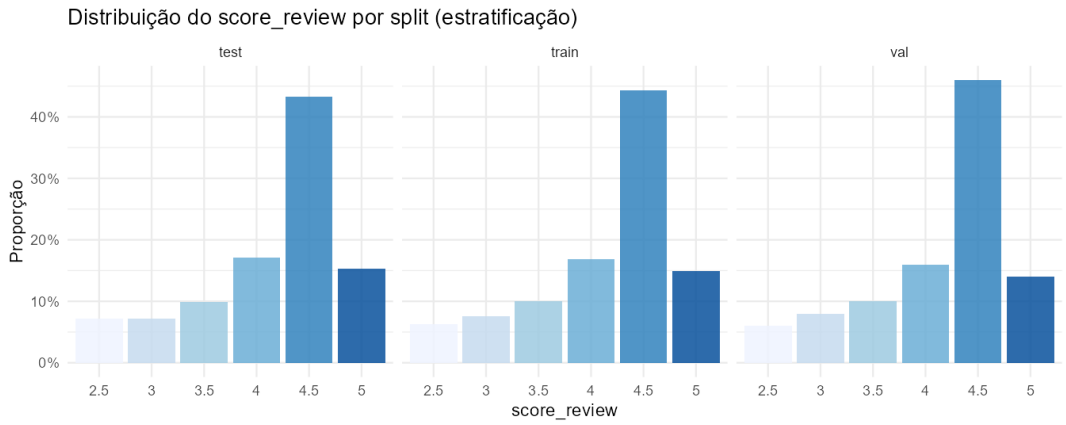
Dado que a maioria das variáveis explicativas são *flags* (*amenities*/selos), todas as preditoras (exceto *logavaliacoes*) são convertidas para codificação **0/1**, através de uma função de coerção robusta, garantindo consistência. O que assegura que medidas como correlações, importâncias e coeficientes são comparáveis e interpretáveis sob a mesma escala.

4.4 Particionamento estratificado em treino/validação/teste

Após a limpeza base, o dataset (n=529) é dividido em três subconjuntos com estratificação por níveis de *score_review*, preservando a distribuição empírica do alvo em cada partição. Este detalhe é particularmente relevante porque *score_review* é **discreto** e está **concentrado em valores altos** (como mostrado no Cap. 3), pelo que um split não estratificado poderia gerar subconjuntos com poucos exemplos de níveis raros, aumentando a variância das métricas.

Conjunto	n	Média	DP	Mediana
Treino	368	≈ 4,151	≈ 0,693	4,5
Validação	50	≈ 4,150	≈ 0,694	4,5
Teste	111	≈ 4,140	≈ 0,711	4,5

A divisão treino/validação/teste foi realizada por **amostragem estratificada** no *score_review*, de forma a preservar a distribuição do target entre subconjuntos. A imagem seguinte mostra que a distribuição de ***score_review*** se mantém semelhante em treino/validação/teste, suportando a comparabilidade entre conjuntos. (Estatísticas descritivas por split em Anexo C4-4.)



4.5 Pré-processamento com desenho anti-*leakage* (*fit* no treino, *apply* nos restantes)

Todos os passos que dependem de estatísticas da amostra são ajustados **exclusivamente no treino** (*preprocess_fit*) e aplicados sem recalibração em validação e teste (*preprocess_apply*). Esta decisão elimina *data leakage* e frisa que as métricas reportadas no teste são uma aproximação credível do desempenho fora da amostra.

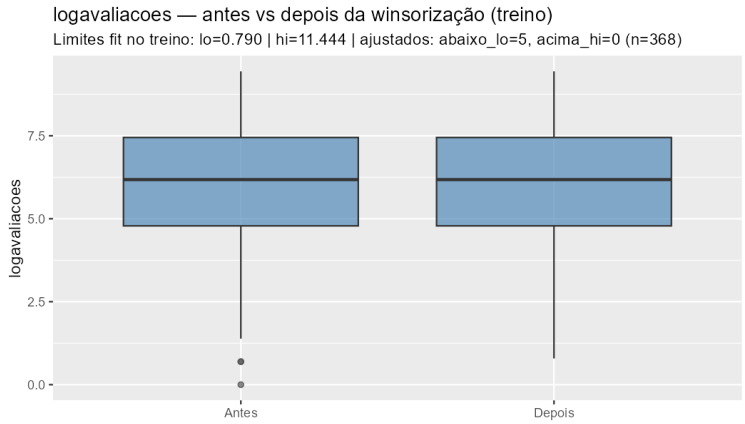
4.5.1 Imputação (robustez do pipeline)

Neste dataset, após a limpeza base, o número de omissos é **zero em todas as variáveis**, pelo que a imputação funciona sobretudo como mecanismo de segurança para reprodutibilidade e extensibilidade do pipeline.

4.5.2 Winsorização de *logavaliacoes* (IQR, k=1.5)

A variável ***logavaliacoes*** apresenta assimetria e assimetria positiva, podendo aumentar a influência de observações extremas (potencial **alavancagem/leverage**) no ajuste de modelos lineares. Para reduzir a sensibilidade a valores extremos **sem eliminar observações**, aplica-se **winsorização** com limites estimados **exclusivamente no conjunto de treino** e posteriormente aplicados a validação e teste (evitando *data leakage*)

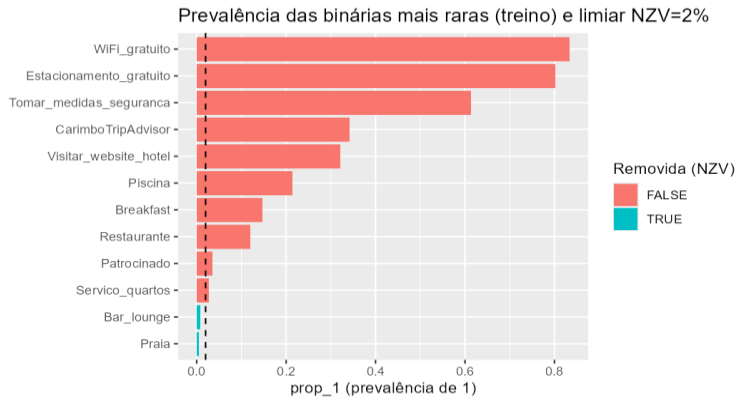
Statistic	Value
Q1	≈ 4,785
Q3	≈ 7,449
IQR = Q3 − Q1	≈ 2,664
Limite inferior: Q1 − 1,5·IQR	≈ 0,790
Limite superior: Q3 + 1,5·IQR	≈ 11,444



Observa-se que o máximo de ***logavaliacoes*** é inferior a 11,444, pelo que **não ocorre truncamento na cauda superior**. Assim, a winsorização atua **apenas na cauda inferior**, substituindo valores abaixo de 0,790 por 0,790. Importa notar que valores baixos de número de avaliações (incluindo zero) são **plausíveis e frequentes** neste contexto; deste modo, esta transformação deve ser interpretada como uma **medida de robustez** para limitar a influência de valores muito pequenos no ajuste, e não como uma "remoção de outliers" no sentido substantivo. Após a aplicação, verifica-se que o mínimo de ***logavaliacoes*** em treino/validação/teste coincide com 0,790, em consistência com a utilização dos parâmetros fixados no treino.

4.5.3 Remoção de variáveis binárias raras (*near-zero variance*)

Variáveis quase constantes são pouco informativas e podem aumentar instabilidade. A regra aplicada calcula, para cada binária, **min (p, 1−p)** e remove se for menor que o limiar de raridade de **0,02**. Foram removidas as variáveis ***Praia*** e ***Bar_lounge*.**



4.6 *Feature engineering*

Para aumentar a capacidade explicativa do modelo sem comprometer a interpretabilidade, foram construídas duas covariáveis derivadas a partir das variáveis originais: (i) uma medida agregada da "intensidade de *amenities*" e (ii) um termo de interação para permitir efeitos condicionais.

1) Contagem de *flags*/*amenities* ativas - n_bin_true: Seja $\mathcal{B}$ o conjunto de índices das variáveis binárias (*amenities*/*flags*). Para cada hotel i, define-se: $\boldsymbol{n_{\text{bin\_true},i}} = \sum_{j \in \mathcal{B}} \mathbb{1}(\boldsymbol{x_{ij}} = \mathbf{1})$, onde 1 é a função indicadora.

Isto funciona como uma **proxy agregada** da oferta de *amenities*, permitindo ao modelo capturar um efeito global (p.ex., "mais *amenities* → maior score") mesmo quando algumas variáveis individuais são raras e, portanto, menos informativas isoladamente.

2) Termo de interação - *CarimboTripAdvisor* × *logavaliacoes*: Para capturar a possibilidade de o impacto do selo variar com a popularidade (volume de avaliações), foi criado o termo: $\boldsymbol{CarimboTripAdvisor\_x\_logavaliacoes_i} = \boldsymbol{C_i} \cdot \boldsymbol{L_i}$ onde $C_i \in \{0,1\}$ representa *CarimboTripAdvisor* e $L_i$representa *logavaliacoes*.

Introduz a possibilidade de **moderação**: o "peso" do "*CarimboTripAdvisor*" pode não ser constante, podendo diferir entre hotéis com poucas avaliações e hotéis com muitas avaliações.

4.7 Padronização (*standardization*) para modelos sensíveis à escala

Para cada covariável numérica X_j, procede-se à **padronização por z-score** estimando os parâmetros **apenas no conjunto de treino**. Seja i observação e j variável: $\mu_{j,\text{tr}} = \frac{1}{n_{\text{tr}}} \sum_{i \in \text{treino}} x_{ij}$, $\sigma_{j,\text{tr}} = \sqrt{\frac{1}{n_{\text{tr}}-1} \sum_{i \in \text{treino}} (x_{ij} - \mu_{j,\text{tr}})^2}$.

A transformação aplicada a qualquer observação (treino/validação/teste) é então: $z_{ij} = \frac{x_{ij} - \mu_{j,\text{tr}}}{\sigma_{j,\text{tr}}}$.

Isto garante que a variável transformada tem média 0 e desvio-padrão 1 no treino, e que validação/teste são transformados com os mesmos parâmetros, evitando *data leakage* e mantendo comparabilidade entre conjuntos.

A evidência da correta aplicação da estandardização observa-se no treino estandardizado: para as variáveis contínuas escaladas, a média aproxima-se de 0 e o desvio-padrão de 1.

variável	mean_train_scaled	sd_train_scaled
logavaliacoes	-1.86E-17	1
n_bin_true	1.36E-16	1
*CarimboTripAdvisor*_x_*logavaliacoes*	-1.43E-18	1

Capítulo 5 - Modelação

5.1 Objetivo

O objetivo desta etapa é comparar um conjunto de modelos preditivos para estimar *score_review*, selecionando a solução com melhor capacidade de generalização através da **separação explícita entre seleção e teste**:

- **Seleção de modelos/hiperparâmetros:** realizada **apenas no conjunto de treino** (n = 368), usando **validação cruzada estratificada** com **7 *folds***.
- **Treino final:** após seleção do melhor modelo por ***Cross-Validation*** (CV), o modelo escolhido é reajustado com **Treino + Validação** (n = 368 + 50 = 418), para maximizar informação disponível antes do teste.
- **Avaliação final (holdout):** desempenho reportado no **Teste** (n = 111), usado uma única vez e não envolvido na seleção.

5.2 Modelos considerados e justificação estatística

5.2.1 Baseline (média do treino)

Como referência mínima, considera-se um modelo ingênuo que prevê sempre a média de *score_review* no conjunto de treino. Esta baseline é essencial para quantificar o valor acrescentado: um modelo útil deve reduzir o erro face a uma regra trivial (ver comparação global em 5.4 e Anexo C5-1).

5.2.2 Regressão Linear (OLS)

A regressão linear (OLS) é incluída como *benchmark* paramétrico e interpretável, estimando um efeito médio aditivo de cada preditor no *score_review*. Apesar de o alvo ser limitado e discretizado, OLS funciona como aproximação de $\mathbb{E}(Y \mid X)$ sob a hipótese de linearidade média e ruído aproximadamente homocedástico. (ver comparação global em 5.4 e Anexo C5-1).

5.2.3 Ridge e Lasso (regularização via glmnet)

Embora o número de preditores seja moderado, existe correlação entre variáveis (colinearidade), o que pode aumentar a variância dos coeficientes em modelos lineares. Por isso, consideram-se modelos regularizados: (i) **Ridge (L2):** reduz a variância dos coeficientes sob colinearidade, estabilizando o ajuste (*trade-off* viés–variância); (ii) **Lasso (L1):** além de regularizar, pode induzir parcimónia ao impor coeficientes exatamente nulos, funcionando como seleção de variáveis. O desempenho comparativo face aos restantes modelos é apresentado em 5.4 e no **Anexo C5-1**.

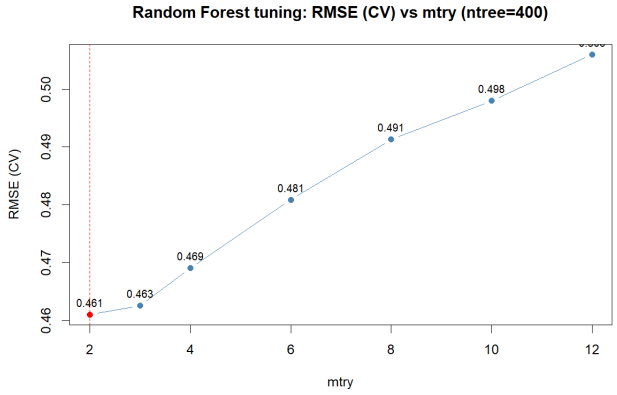
5.2.4 Árvore de Regressão

Uma árvore de decisão é incluída como modelo não linear e interpretável, capaz de capturar interações e efeitos de limiar sem necessidade de especificação explícita. Contudo, árvores individuais tendem a apresentar maior variância e podem sobreajustar, pelo que o seu desempenho serve também de referência para motivar métodos *ensemble* (ver comparação global em 5.4 e Anexo C5-1).

5.2.5 Random Forest (RF)

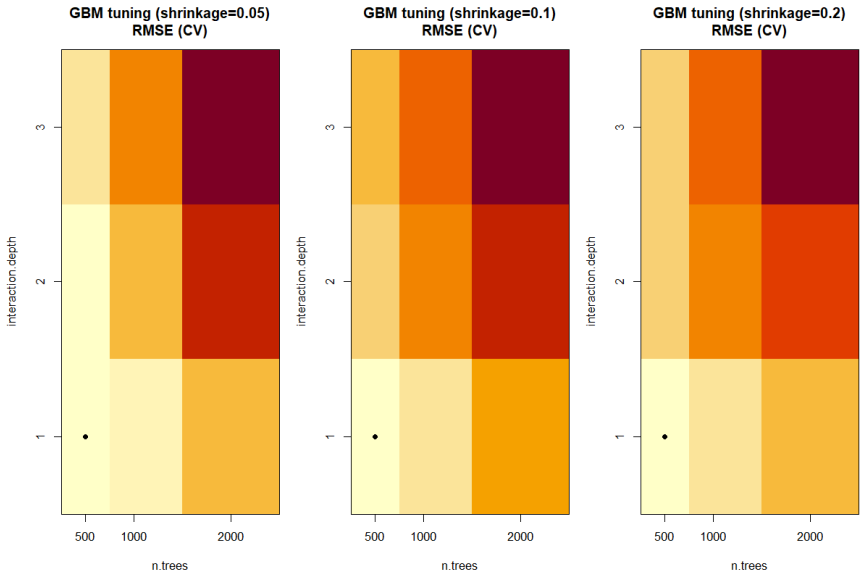
O Random Forest é considerado por combinar múltiplas árvores com amostragem e aleatoriedade em preditores, reduzindo variância e capturando não linearidades e interações. O hiperparâmetro **mtry** (número de variáveis candidatas em cada divisão) foi ajustado por validação cruzada no treino, mantendo **ntree = 400**. O gráfico de **gráfico de linhas,** representada abaixo**,** apresenta o RMSE médio em CV para

diferentes valores de *mtry*, suportando a escolha da configuração com menor erro (mtry=2), sem necessidade de repetir métricas detalhadas nesta subsecção (ver cap 5.4 e Anexo C5-1).



5.2.6 *Gradient Boosting Machine* (GBM)

O GBM é incluído como método *ensemble* baseado em *boosting*, onde árvores fracas são combinadas sequencialmente para reduzir erro, permitindo modelar relações complexas. Foi explorada uma grelha simples de hiperparâmetros - **interaction.depth**, **shrinkage** e **n.trees** - avaliada por validação cruzada no treino. O **mapa de calor**, representado em baixo, mostra um *heatmap* do RMSE médio em CV (por depth × n.trees, em painéis por *shrinkage*), evidenciando a região de melhor desempenho e sustentando a escolha final de hiperparâmetros (ver também Anexo C5-4 e comparação global em 5.4).



5.3 Comparação global por validação cruzada (treino)

A tabela seguinte resume os resultados médios por validação cruzada (ordenados por RMSE):

Modelo	RMSE	MAE	R^2	PCT_0,5
Random Forest (mtry=2)	0,461	0,338	0,554	71,2%
Regressão Linear	0,478	0,360	0,521	70,4%
GBM (depth=1; shrinkage=0,05; 500 árvores)	0,481	0,352	0,512	72,5%
Árvore (cp=0,005; maxdepth=5)	0,482	0,336	0,512	74,2%
Lasso λ_{1se}	0,495	0,368	0,484	69,8%
Ridge λ_{1se}	0,513	0,397	0,447	67,9%
Baseline (média do treino)	0,692	0,563	-0,001	61,2%

Conclusões principais:

- **Todos os modelos superam claramente a baseline**, confirmando sinal preditivo nas variáveis.
- O **Random Forest** obtém o **melhor RMSE e maior R^2**, sendo escolhido como modelo final por critério principal de seleção.
- A **árvore** apresenta PCT_0,5 elevado, mas com RMSE ligeiramente pior (e, estruturalmente, maior variância fora de amostra).
- O **OLS** é competitivo, sugerindo que existe um componente aditivo forte; porém, o RF captura ganhos adicionais via não linearidades/interações.

5.4 Ajuste final e avaliação no conjunto de teste (holdout)

Após seleção do **Random Forest** como melhor modelo por CV, este é reajustado com **Treino + Validação** e avaliado no **Teste** (n = 111), produzindo:

Modelo	Avaliação	RMSE	MAE	R2	PCT_{0,5}
FINAL - Random Forest (mtry=2, ntree=400)	CV (treino, media k-fold)	0.461005	0.338347	0.554298	0.711938
FINAL - Random Forest (mtry=2, ntree=400)	Teste (holdout, n=111)	0.47745	0.330901	0.544784	0.747748

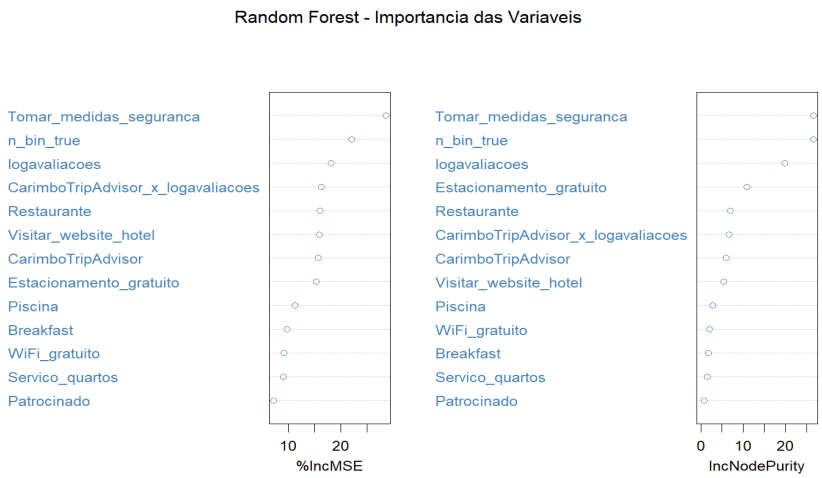
Ou seja, cerca de **3 em cada 4 previsões** ficam a menos de **0,5 pontos** do valor observado. A diferença entre RMSE em CV (0,461) e no teste (0,477) é compatível com variabilidade amostral e **não sugere** ganhos artificiais por *leakage*.

Respeito pela escala [1,5]. As previsões ficaram integralmente dentro do intervalo admissível (3,028 a 4,694), pelo que a operação de "*clipping*" a [1,5] não altera as métricas finais. Observa-se ainda uma

regressão à média, aspeto que será discutido no capítulo de validação/diagnóstico.

5.5 Interpretação do modelo selecionado (importância de variáveis no RF)

Para interpretar os fatores mais relevantes no modelo selecionado, analisou-se a importância de variáveis no Random Forest através de **%IncMSE** (aumento percentual do erro quando a variável é permutada), apresentado no *Cleveland dot plot***,** representado abaixo.



Em termos de ranking, destacam-se como mais influentes:

Variável	%IncMSE
Tomar_medidas_seguranca	28,61
n_bin_true	22,08
logavaliacoes	18,10
CarimboTripAdvisor_x_logavaliacoes	16,25
Restaurante	16,03
Visitar_website_hotel	15,90
CarimboTripAdvisor	15,64
Estacionamento_gratuito	15,34

Estes resultados são coerentes com a análise exploratória do Cap. 3: (i) **Segurança e credibilidade** aparecem como fatores com forte poder discriminativo; (ii) **Popularidade/volume de reviews** (*logavaliacoes*) e a interação com o selo sugerem que o efeito reputacional pode variar com o volume de informação disponível; (iii) **n_bin_true** capta um efeito agregado de "riqueza de *amenities*", indicando que o modelo explora combinações de atributos para melhorar a previsão.

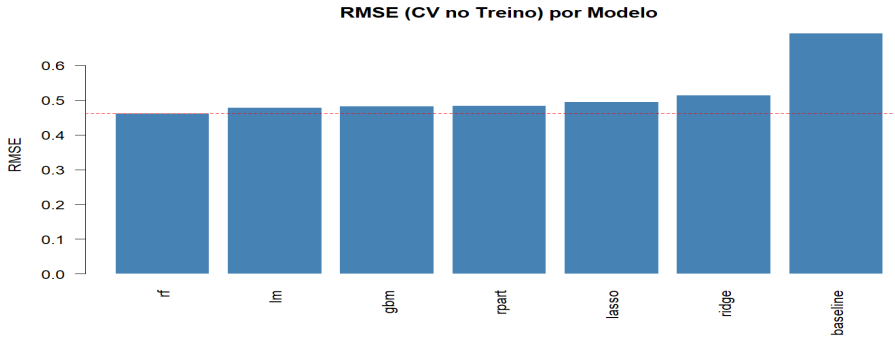
Capítulo 6 - Validação e Avaliação

6.1 Objetivo do capítulo

Este capítulo valida o desempenho do modelo selecionado e avalia a sua capacidade de generalização num conjunto **holdout** (Teste) que não foi utilizado nem na estimação de parâmetros nem na seleção de hiperparâmetros. Para além das métricas agregadas (RMSE, MAE, (R^2) e (PCT_{0,5}), são analisados **diagnósticos gráficos** (observado vs previsto e resíduos) e **padrões de erro** por nível do *score_review*, com foco em: (i) viés sistemático, (ii) *regressão à média* e (iii) limitações associadas à discretização e ao desbalanceamento do alvo.

6.2 Consistência entre validação cruzada e teste (generalização)

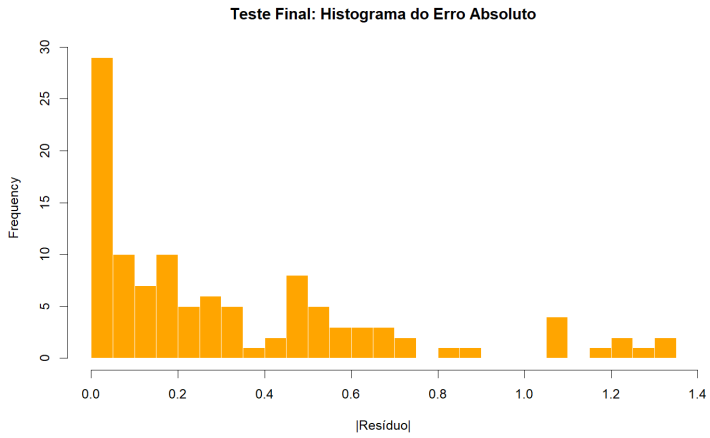
O critério principal de seleção foi o **RMSE médio em validação cruzada (CV) no treino**, com estratificação por níveis do alvo. O modelo com melhor RMSE em CV foi a **Random Forest** com mtry = 2 e ntree = 400. O **gráfico de barras** representado abaixo**,** apresenta a comparação do RMSE em CV por modelo, evidenciando uma vantagem clara da RF face à baseline e ganhos marginais face aos restantes modelos.



Comparando com a CV no treino (RMSE = 0,461), a degradação no teste é de cerca de **0,016** pontos de RMSE. Esta diferença é pequena e é compatível com variabilidade amostral, sugerindo **ausência de sobreajuste severo** e coerência do desenho experimental (anti-*leakage*). Assim, o desempenho reportado no teste pode ser interpretado como uma estimativa credível do erro de generalização, dentro das limitações do conjunto de variáveis disponível (Ver **Anexo C6-2**).

6.3 Métricas no teste e interpretação substantiva

A métrica **MAE = 0,331** significa que a previsão falha cerca de **0,33 pontos** na escala de 1 a 5. A métrica **(PCT_{0,5} = 74,8%)** indica que **aproximadamente 3 em cada 4 hotéis** têm previsão a menos de **meio ponto** do valor observado, o que é particularmente relevante por o alvo variar em incrementos de 0,5.

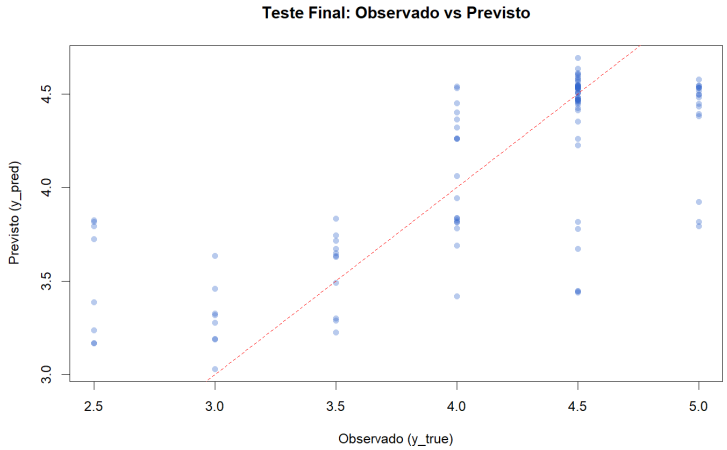


A distribuição dos erros absolutos reforça esta leitura:

Métrica	Valor
mediana(|erro|)	≈ 0,198
P75(|erro|)	≈ 0,492
P90(|erro|)	≈ 0,829
P95(|erro|)	≈ 1,130
máximo(|erro|)	≈ 1,324

6.4 Diagnóstico: observado vs previsto

O **diagrama de dispersão** evidencia características na regressão quando o alvo é limitado e discretizado:



1. **Contração da amplitude das previsões (regressão à média):**
    - No teste, as previsões variam aproximadamente entre **3,03 e 4,69**, enquanto o alvo observado assume os valores discretos **2,5; 3,0; 3,5; 4,0; 4,5; 5,0** .
    - O desvio-padrão do observado no teste é ≈ **0,711**, ao passo que o desvio-padrão do previsto é ≈ **0,482**, indicando previsões mais "centradas".
2. **Dificuldade em prever extremos (2,5 e 5,0):**

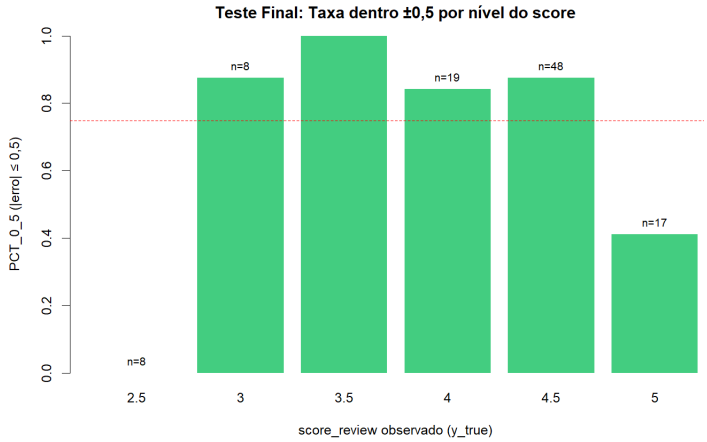
    A análise por nível do *score_review* no teste (consultar **gráfico de barras** em baixo) mostra que os maiores desvios médios ocorrem nos extremos:

    - Para **2,5 (n=8)**, a média prevista é ≈ **3,51** (erro médio ≈ **+1,01**; há sobre-previsão).
    - Para **5,0 (n=17)**, a média prevista é ≈ **4,38** (erro médio absoluto ≈ **0,62**; há sub-previsão).

    Consequentemente, a taxa de acerto dentro de **±0,5** é:

    - **0%** para *score_review* = **2,5;**
    - **41,2%** para *score_review*=**5,0**

    Em contraste, nos níveis intermédios, o desempenho é substancialmente melhor (por exemplo, *score_review* = **3,5** apresenta **100%** dentro de ±0,5 no teste), conforme sintetizado no gráfico em baixo.



Este padrão é coerente com: (i) **escassez de exemplos extremos** (amostra menor → maior incerteza), (ii) **perda de amplitude induzida pela otimização por erro quadrático** (o modelo minimiza erro global e tende a aproximar-se da região de maior densidade do alvo), e (iii) **ausência de variáveis possivelmente**

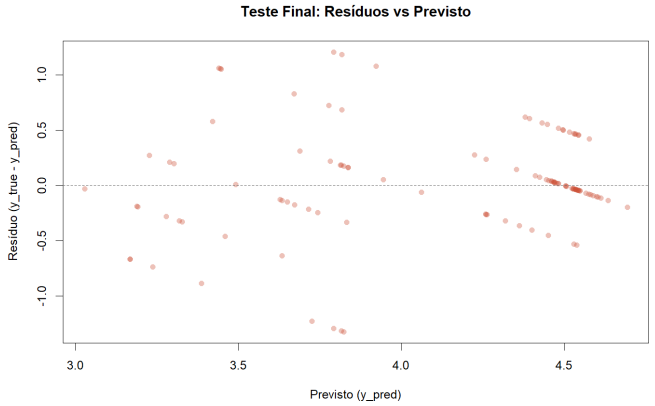
determinantes para distinguir hotéis verdadeiramente “excelentes” ou “fracos” (p.ex., localização, preço, estrelas oficiais, tipo de hotel, segmento, proximidade de atrações), o que limita a capacidade de separar os extremos mesmo quando o desempenho médio é bom.

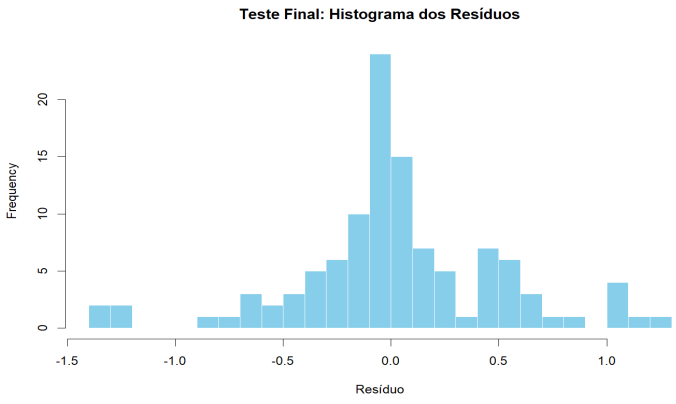
6.5 Diagnóstico: resíduos (viés e dispersão)

Define-se resíduo como (e = y_{true} - y_{pred}). No teste, o resíduo médio é ≈ **0,027**, o que sugere **viés global reduzido** (ligeira tendência a sub-prever, em média). A dispersão dos resíduos é consistente com a escala do problema e com a discretização do alvo:

Percentil	Value (≈)
P5	-0,70
P50	-0,02
P95	0,94

O gráfico “resíduos vs previsto” não evidencia uma heterocedasticidade pronunciada; o padrão visível está fortemente condicionado pela discretização do alvo. Ainda assim, observa-se que previsões elevadas (≈4,5 - 4,7) tendem a gerar resíduos positivos quando o verdadeiro é 5,0 (sub-previsão), e previsões mais baixas (≈3,3–3,9) podem gerar resíduos negativos quando o verdadeiro é 2,5 (sobre-previsão). O gráfico “histograma dos resíduos” confirma uma massa central junto de 0 com caudas assimétricas moderadas, refletindo precisamente estes casos extremos.





6.6 Validação da coerência de escala e efeito do “clipping”

Por construção, *score_review* deve pertencer ao intervalo **[1,5]**. No teste, as previsões da Random Forest ficaram **integralmente dentro deste intervalo** (mínimo ≈ 3,03; máximo ≈ 4,69). Assim, a operação de

"clipping" das previsões para [1,5] **não altera** RMSE, MAE, (R^2) nem (PCT_{0,5}). Este comportamento é consistente com a contração das previsões discutida anteriormente: o modelo raramente se aproxima dos limites 1 ou 5, o que reforça a interpretação de *regressão à média* (ver **Anexo C6-3**).

6.7 Interpretação do modelo final: importância das variáveis (consistência com EDA)

Como apoio à interpretação preditiva do modelo final (Random Forest), analisou-se a **importância por permutação** (aumento do erro quando a variável é embaralhada). O ranking completo encontra-se no **Anexo C6-4**.

Os resultados são **consistentes com a análise exploratória (Capítulo 3)** e com a interpretação apresentada no **Capítulo 5.6**: variáveis associadas a **confiança/segurança**, **popularidade/volume de reviews** (*logavaliacoes*) e um efeito agregado de ***amenities*** (*n_bin_true*) surgem como contribuintes relevantes. A presença da interação ***CarimboTripAdvisor*** × ***logavaliacoes*** entre as variáveis mais importantes reforça a motivação para a sua inclusão como feature no pré-processamento (ver **Cap. 5.5** e **Anexo C6-4**).

Capítulo 7 - Conclusão

A comparação entre modelos supervisionados evidenciou que a **Random Forest** (com *mtry* = 2) apresentou a melhor capacidade de generalização segundo o critério principal (RMSE médio em validação cruzada), sendo por isso selecionada como modelo final. No conjunto de teste (holdout), o modelo obteve **RMSE = 0,477**, **MAE = 0,331**, **R^2 = 0,545** e **PCT_{0,5} = 74,8%**, isto é, cerca de **3 em cada 4 previsões** ficaram a menos de **0,5 pontos** do valor observado na escala [1,5]. Em comparação com a baseline (previsão pela média), estes resultados traduzem uma melhoria substancial, indicando que as variáveis disponibilizadas contêm **sinal preditivo** relevante e que o modelo consegue explorá-lo de forma consistente.

Os diagnósticos do modelo final evidenciam, contudo, um padrão típico de regressão com alvo **limitado e discretizado**: as previsões exibem **contração de amplitude** ("regressão à média"), concentrando-se na zona central da escala, e verifica-se maior dificuldade em reproduzir valores **extremos** de *score_review* (2,5 e 5,0). Assim, embora o desempenho global seja robusto, a precisão degrada-se nos casos raros/extremos, onde existe menor suporte amostral e maior incerteza.

As limitações mais relevantes decorrem de aspetos estruturais do problema e do conjunto de dados:

1. **Alvo discreto e limitado:** *score_review* assume apenas 6 níveis (passos de 0,5). Esta discretização reduz a resolução do problema e, quando se otimiza uma função de perda quadrática (RMSE), favorece previsões mais centradas, penalizando menos a incapacidade de atingir os extremos.
2. **Desbalanceamento e poucos extremos:** os níveis extremos do score têm baixa frequência, o que aumenta a variância das estimativas nesses grupos e tende a aumentar o erro precisamente onde o modelo tem menos informação para aprender padrões diferenciadores.
3. **Conjunto de variáveis incompleto:** o dataset não inclui covariáveis potencialmente determinantes do rating (p.ex., preço, estrelas oficiais, localização e distância a atrações, tipologia/categoria do hotel, dimensão). A ausência destes fatores limita a capacidade de distinguir hotéis verdadeiramente "excecionais" de hotéis "medianos" apenas com base em *amenities*/*flags* e volume de reviews.

Em síntese, o estudo produziu um modelo preditivo **consistente e validado**, com desempenho robusto no holdout e interpretação coerente com a análise exploratória. A **Random Forest** revelou-se uma escolha adequada no compromisso entre flexibilidade e generalização, captando não linearidades e interações que modelos estritamente lineares dificilmente representariam com a mesma eficácia, ainda que permaneça a limitação esperada de menor precisão nos valores extremos do *score_review*.

ANEXOS

Anexos do Capítulo 1 - Formulação do Problema

Anexo C1-1 - Dicionário de variáveis

Variável	Tipo	Função	Hipóteses
score_review	Numérica	target	NA
CarimboTripAdvisor	Booleana	preditor	Espera-se associação positiva (selo/credibilidade pode sinalizar qualidade).
Patrocinado	Booleana	preditor	Efeito incerto: pode refletir marketing (sem relação direta com qualidade) ou segmentação de hotéis.
Breakfast	Booleana	preditor	Espera-se associação positiva (amenity/serviço adicional tende a melhorar experiência).
WiFi_gratuito	Booleana	preditor	Espera-se associação positiva (amenity/serviço adicional tende a melhorar experiência).
Estacionamento_gratuito	Booleana	preditor	Espera-se associação positiva (amenity/serviço adicional tende a melhorar experiência).
Piscina	Booleana	preditor	Espera-se associação positiva (amenity/serviço adicional tende a melhorar experiência).
Restaurante	Booleana	preditor	Espera-se associação positiva (amenity/serviço adicional tende a melhorar experiência).
Servico_quartos	Booleana	preditor	Espera-se associação positiva (amenity/serviço adicional tende a melhorar experiência).
Praia	Booleana	preditor	Espera-se associação positiva (amenity/serviço adicional tende a melhorar experiência).
Bar_lounge	Booleana	preditor	Espera-se associação positiva (amenity/serviço adicional tende a melhorar experiência).
Tomar_medidas_seguranca	Booleana	preditor	Espera-se associação positiva (confiança/qualidade do serviço).
Visitar_website_hotel	Booleana	preditor	Efeito possivelmente positivo (hotéis mais estruturados/atrativos), mas pode ser proxy de procura.
logavaliacoes	Numérica (contínua)	preditor	Efeito incerto: mais avaliações pode estabilizar rating; pode também refletir maior heterogeneidade.

Conteúdo: definição das variáveis, tipo (binária/numérica), papel (target/feature) e hipóteses iniciais.

Anexos do Capítulo 3 - Data Understanding

Anexo C3-2 - Prevalência das variáveis binárias (prop_1)

Variável	Prevalência (prop_1)
WiFi_gratuito	0,806
Estacionamento_gratuito	0,749
Tomar_medidas_seguranca	0,584
CarimboTripAdvisor	0,333
Visitar_website_hotel	0,297
Patrocinado	0,035
Servico_quartos	0,027
Bar_lounge	0,0055
Praia	0,004

Conteúdo: nº de valores distintos por variável (mostra *score_review* discretizado e *logavaliacoes* contínua).

Anexos do Capítulo 4 - Data Preparation

Anexo C4-1 - QC da limpeza base

Descrição	Quantidade
Nº observações no dataset bruto	546
Nº duplicados exatos identificados	17
Nº observações com target em falta	0
Nº observações após remoção de missing no target	546
Nº observações após deduplicação	529
Nº valores do target inválidos após conversão numérica	0
Nº observações removidas por target não numérico	0
Nº valores do target fora de ([1,5])	0

Conteúdo: n bruto, duplicados, n após deduplicação, validações do target.

Anexo C4-4 - Dataset de treino final (CSV)

Ficheiro: result_dir_extracted/outputs_cap4/train.csv

Conteúdo: treino após preparação, com *features* derivadas.

Anexos do Capítulo 5 - Modelação

Anexo C5-1 - Métricas de validação cruzada

model_id	modelo	RMSE	MAE	R2	PCT_{0,5}
rf	Random Forest (mtry=2, ntree=400)	0.461005	0.338347	0.554298	0.711938
lm	Regressão Linear (lm) [dados escalados]	0.477686	0.360496	0.521259	0.703666
gbm	GBM (depth=1, shrinkage=0.05, n.trees=500)	0.481214	0.352303	0.512313	0.725265
rpart	Arvore (rpart) cp=0.005 maxdepth=5	0.482453	0.335955	0.51152	0.742048
lasso	Lasso (glmnet) [lambda.1se, nested-CV]	0.494615	0.367846	0.484193	0.698322
ridge	Ridge (glmnet) [lambda.1se, nested-CV]	0.512629	0.3966	0.446507	0.679131
baseline	Baseline (media treino)	0.691989	0.562978	-0.0007	0.611648

Conteúdo: RMSE, MAE, R^2 e (PCT_{0,5}) por modelo em CV (base de comparação).

Anexo C5-4 - Grelha CV do GBM

interaction.depth	shrinkage	n.trees	RMSE	MAE	R2	PCT_{0,5}
1	0.05	500	0.481214	0.352303	0.512313	0.725265
2	0.05	500	0.487032	0.348891	0.50056	0.747502
3	0.05	500	0.49548	0.353141	0.48285	0.76038
1	0.1	500	0.48811	0.355166	0.498285	0.722517
2	0.1	500	0.514675	0.362336	0.442183	0.744046
3	0.1	500	0.528188	0.373341	0.41261	0.741149
1	0.2	500	0.510872	0.36705	0.449274	0.735694

2	0.2	500	0.542596	0.38135	0.377824	0.717323
3	0.2	500	0.549202	0.388504	0.361741	0.717542
1	0.05	1000	0.488342	0.354695	0.497634	0.733467
2	0.05	1000	0.507174	0.360074	0.457868	0.746753
3	0.05	1000	0.521832	0.368817	0.426442	0.757892
1	0.1	1000	0.507729	0.36357	0.456294	0.735954
2	0.1	1000	0.54462	0.379749	0.37313	0.741598
3	0.1	1000	0.553703	0.381671	0.351636	0.736324
1	0.2	1000	0.531937	0.376957	0.403148	0.741598
2	0.2	1000	0.57722	0.395692	0.292385	0.71984
3	0.2	1000	0.588086	0.402255	0.265979	0.711718
1	0.05	2000	0.505931	0.36246	0.460571	0.733506
2	0.05	2000	0.536693	0.376401	0.3915	0.760639
3	0.05	2000	0.551788	0.384947	0.355834	0.755405
1	0.1	2000	0.535617	0.378348	0.394345	0.739001
2	0.1	2000	0.573645	0.395975	0.303716	0.717433
3	0.1	2000	0.592871	0.407042	0.257281	0.725265
1	0.2	2000	0.555185	0.392857	0.349357	0.73035
2	0.2	2000	0.603788	0.413134	0.223511	0.711868
3	0.2	2000	0.635195	0.434802	0.141699	0.703257

Conteúdo: resultados por depth, shrinkage e n.trees.

Anexos do Capítulo 6 - Validação e Avaliação

Anexo C6-2 - Ranking final de modelos (CV)

rank	model_id	modelo	RMSE	MAE	R2	PCT_{0,5}
1	rf	Random Forest (mtry=2, ntree=400)	0.461004671	0.338346858	0.554297767	0.711938062
2	lm	Regressao Linear (lm) [dados escalados]	0.477686155	0.360496306	0.521258915	0.703666334
3	gbm	GBM (depth=1, shrinkage=0.05, n.trees=500)	0.481214299	0.35230284	0.512313305	0.725264735
4	rpart	Arvore (rpart) cp=0.005 maxdepth=5	0.482453441	0.33595486	0.51151986	0.742047952
5	lasso	Lasso (glmnet) [lambda.1se, nested-CV]	0.494614923	0.367845956	0.484193108	0.698321678
6	ridge	Ridge (glmnet) [lambda.1se, nested-CV]	0.512628986	0.396600207	0.446506818	0.679130869
7	baseline	Baseline (media treino)	0.691988559	0.562977525	-0.000700899	0.611648352

Conteúdo: ranking e resumo comparativo por modelo.
Ficheiro: result_dir_extracted/outputs_cap6/cap6_predicoes_com_diagnosticos.csv
Conteúdo: observado, previsto, resíduos, erro absoluto e estatísticas auxiliares (base para os gráficos).

Anexo C6-4 - Top importâncias por modelo

Item	Score (≈)
Tomar_medidas_seguranca	28,61
n_bin_true	22,08
logavaliacoes	18,10
CarimboTripAdvisor × *logavaliacoes*	16,25
Restaurante	16,03
Visitar_website_hotel	15,90
CarimboTripAdvisor	15,64
Estacionamento_gratuito	15,34

Conteúdo: síntese das variáveis mais importantes (foco no modelo final).